12-2021

Socio-Economic Variables and their Relationships with the Passenger to Population Ratio

Jack Matthew Westbury

Socio-Economic Variables and their Relationships with the Passenger to Population Ratio

By

Jack Matthew Westbury

Bachelor of Science
Aeronautical Science with Flight
Florida Institute of Technology
2017

A thesis submitted to the College of Aeronautics at
Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Aviation Human Factors

Melbourne, Florida
December 2021

The undersigned committee, having examined the attached thesis "Socio-economic Variables and their Relationships with the Passenger to Population Ratio" by Jack Matthew Westbury hereby indicates its unanimous approval

Debbie Carstens, Ph.D.
Professor
College of Aeronautics
Major Advisor

John Deaton, Ph.D.
Professor
College of Aeronautics

Ryan White, Ph.D.
Assistant Professor
Mathematical Sciences

Ulreen Jones-McKinney, Ph.D.
Assistant Professor and Dean
College of Aeronautics

Abstract

Title: Socio-economic Variables and their Relationships with the Passenger to Population Ratio

Author: Jack Matthew Westbury

Major Advisor: Dr. Debbie Carstens

As the aviation industry expands and becomes commonplace in the modern world, it would behoove us to understand how the industry develops and passenger to population ratio could be a measure of that development. If called upon to identify how the aviation industry develops, one could only make suggestions as to which variables might be key in the development of the aviation industry. This study used existing socio-economic and population data from the United Nations and passenger information from the International Air Transport Association, with the aim to deepen the understanding of the relationship between socio-economic variables and the passenger to population ratio of a developed country. Both linear and multiple linear regression were used to analyze fourteen independent variables, both individually and combined, to reveal relationships and potentially form a predictive model. Fifteen countries were initially selected for this study, after cleansing and parsing of available data only fourteen were deemed appropriate. Results were measured using both R^2 and p values where appropriate, with five-fold cross-validation being used on the predictive models to prevent overfitting. Overall, the results of the individual variable analyses varied from insignificant to significant and the best predictive model when the independent variables were combined had an R^2 score of 0.75.

At a glance, variables based on economic factors tend to yield better chance of a relationship with the dependent variable than the social factors, however, both were used in the construction of the best predictive model. Recommendations were made consisting of employing a similar study on countries in different stages of development, all countries data is available for combined, and deeper evaluation of the most significant independent variables.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank each of the individual committee members for their time, encouragement, and analysis. Dr. Carstens as committee chair and my advisor, specifically, played a pivotal role in this study by being both tireless in her work ethic and unquestionably available and open for any questions, both dumb and smart, that I may have had. I am sure she will assure you there were not many of the latter! The time and effort that Dr. Carstens has put in has helped develop a plethora of skills that will play a role in my life in and outside of education.

Second, I would like to extend my sincerest gratitude to Dr. White for his support, especially with some of the trickier statistical analyses and any questions that related to statistics in general. The amount of understanding that I gained from Dr. White's support is exponential, as were the questions from myself to him over time, because just like a math book I had a lot of problems. Dr. White was also able to guide me in the use of cross-validation when the idea of creating a predictive model came around, without his knowledge of statistics, machine-learning, probability, and programming languages likely the predictive model part of this study would have been abandoned.

Last, but not least, of the committee is Dr. Deaton, who was the first professor to light the metaphorical statistical flame in my mind. Dr. Deaton's statistics class was the most enjoyable and practical learning experience I have had at FIT (Don't tell the other professors!) and what was learned in that class paid plentiful dividends during this study. Dr. Deaton's passion for statistics and teaching combined has allowed me to think critically about research that we are presented in daily life and make better judgements

about that research as a result. Dr. Deaton's feedback on the thesis proposal was invaluable to the completion of this study, and although I foolishly ignored his idea of keeping it simple, I greatly appreciate the advice and will never forget the difference between problem and purpose statements!

Lastly, I would like to thank everybody involved in the collection and dissemination of data that serve the UN and IATA. Without these people working tirelessly to collect, input, and store data, this research would have been too difficult to complete. I would also like to take the time to appreciate the people that have come before me who did not have modern technology to complete theses and dissertations, I recognize that we have it a lot easier in the modern era and respect the great works that have happened in our wake.

Dedication

This thesis is dedicated to my family Beth Westbury, Adam Westbury, and Kainan Li, but most specifically my parents Ronald and Gillian Westbury, who adopted me from an abusive father, have loved and supported me unconditionally since my introduction into this world. Arriving at F.I.T in 2014 was a huge leap in my life, leaving friends, family, opportunities, and more back home has been more challenging than I had imagined and without their love and support it would not have been possible, and I would not be the person I am today.

I would also like to dedicate this thesis to two other people. My deeply missed grandfather Ronald Westbury. Granddad not only got me interested in aviation with his stories of being drafted as a paratrooper in the Korean war and the pilots having two mattresses, but he showed me how to be a gentleman. I am so happy that he got to see me on my first flight from Farnborough airport in England and I'm sure he would be very proud to see where I am today. My niece Caitlin Paige Webb, whose childhood I have unforgivably and undesirably missed out on. I will live with the regret of not seeing you grow from closer, but hope that in the future I can make amends for this!

Chapter 1

Introduction

Purpose Statement

The purpose of this study is to examine the relationships between the ratio of airline passengers to population and various socio-economic variables of the sample countries.

Operational Definitions

In the context of this study, the ratio of passengers to population is the number of passengers on all scheduled airline operations, comprising of Part 91, 121, and 135 operations, worldwide to, from, or within the respective country compared to that country's population in 2018. This data will be retrieved from the International Air Transport Association (2019) database. Economic variables are expected to have significant effect when it comes to the ratio of passengers to population based on pre-testing and literature review. For this study, the socio-economic variables are defined as:

- GDP, or the gross domestic product, of a country is defined as gross domestic product per capita, which is the GDP of a country divided by its population, in U.S. dollars from the year 2018 that will be retrieved from the United Nations (2019) database.
- Consumer Price Index is defined as the general level of prices of goods and services purchased by the households for their own final consumption that will be retrieved from the United Nations (2019) database.

- Patents in Force by number is defined as the number of patents that are currently enforced for the representative country and will be retrieved from the United Nations (2019) database.
- Gross Domestic Expenditure on R&D as a percentage of GDP is defined by the United Nations (2019) as "Gross domestic expenditure on scientific research and experimental development (R&D) expressed as a percentage of Gross Domestic Product (GDP)." and will be retrieved from the United Nations database.
- Unemployment Rate Total is defined as the number of unemployed people of working age that are available but without work as a percentage of the labor force and will be retrieved from the United Nations (2019) database.
- Labor Force Participation Total is defined as the percentage of the labor force versus the total working age population between the ages of 15 and 64 and will be retrieved from the United Nations (2019) database.
- Tourist/Visitor Arrivals is measured by the thousands and defined by the United Nations (2019) as "A visitor is a traveler taking a trip to a main destination outside his/her usual environment, for less than a year, for any purpose (business, leisure or other personal purposes) other than to be employed by a resident entity in the country or place visited." and will be retrieved from the United Nations database.
- Tourism Expenditure is measured in millions of U.S. dollars and is defined by the United Nations (2019) as "as the total consumption expenditure

made by a visitor or on behalf of a visitor for and during his/her trip and stay at destination" and will be retrieved from the United Nations database.

- Balance of Payments Current Account is measured in millions of U.S. dollars, collected from the United Nations (2019) database, and defined in two parts. Balance of Payments is defined as "a systematic record of all the economic transactions between residents of that country and the rest of the world." And Current Account is defined as a record of "transactions covering inflows and outflows of goods and services, investment income and current transfers."
- Balance Imports/Exports is also measured in millions of U.S. dollars will be collected from the United Nations (2019) database and is defined as the total inbound and outbound movement of goods through the designated country and will be retrieved from the United Nations database.
- Public Expenditure on Education Government is defined as the percentage of government expenditure on education compared to the total government expenditure and will be retrieved form the United Nations database.
- Public Expenditure on Education GDP is defined as the percentage of government expenditure on education as a percentage of the country's GDP and will be retrieved form the United Nations database.
- Total Population of Concern to the UNHCR is defined as "refugees, returnees, stateless people, the internally displaced and asylum-seekers."

and is measured by number of those entities in a country and will be retrieved from the United Nations database.

The terms highly developed and less developed countries are present throughout this study; these terms are defined by the rankings of each country on the United Nations Human Development Index (2019) report. Figure 1.1 is provided by the United Nations that shows how each country is measured by the Human Development Index.

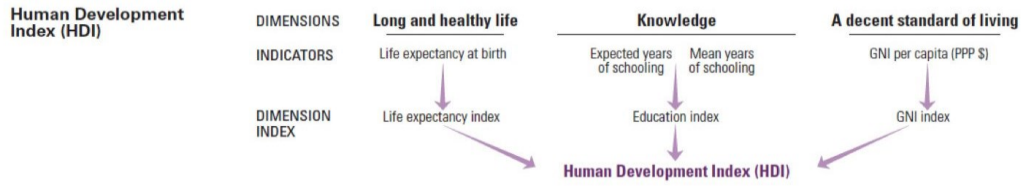

Figure 1.1. Human Development Index Process

Background

When you ask the general public, what impacts the passenger to population ratio, it is improbable that you will get an accurate response. Despite the numerous and thorough machine learning and regression tools available to predict passenger numbers for a particular set of countries, there remains no simple and effective tool for predicting passenger to population ratio.

Many attempts at creating complex prediction tools for passenger demand do exist, Srisaeng et al.'s (2015) genetic algorithm approach tested both linear and quadratic models that examined 74 training datasets alongside 13 out-of-sample datasets to see how robust the genetic algorithm approach is. Based on the findings by Srisaeng et al., the linear and quadratic models gave mean absolute percentage errors of 2.55% and 2.23% respectively,

which shows that the quadratic model had the best predictive power of Australia's domestic passenger demand and were very robust for the purpose.

Passenger travel is known to be a less-than-smooth operation when it comes to estimated or actual passenger numbers, there are many occasions where tickets sold does not match passengers flown and even instances where travel to a destination is disturbed by a diversion of the aircraft. Carbonneau et al.'s (2008) study used the supply chain industry to try to build a tool for supply demand forecasting, which matches the passenger number variance well due to all of the factors that influence supply chain fluctuation. Carbonneau et al. compared traditional methods like linear regression and trend analysis with machine-learning techniques to examine any difference in accuracy of forecast. Testing the different methods with both simulated and archival data, Carbonneau et al. showed that, in general, the more advanced machine-learning techniques did provide better accuracy than traditional methods apart from linear regression. The resultant findings from Carbonneau et al.'s study showed the linear regression is likely the optimal solution for forecasting supply chain demand, not only because it provides similar accuracy to more advanced models but, because of the computational and conceptual simplicity.

The majority of studies have focused on predicting passenger load or demand using different techniques of varying complexity. This study aims to examine relationships between various socio-economic variables and the passenger to population ratio, using regression methods that the literature recommends implicitly, and provide the use of those relationships to potentially build a predictive equation.

Research Question 1

The primary research question that will guide this study is: “What is the relationship between various socio-economic variables and the passenger to population ratio of a country?”

Research Question 2

The secondary research question that will guide this study is: “Can a predictive model using various socio-economic variables be used to predict the passenger to population ratio of a country?”

Hypotheses

This study's hypotheses are directional in nature as the examined literature suggests that some of the selected socio-economic variables are successfully used in many of the current algorithms as variables for predictive purposes. The hypotheses are as follows:

- There will not be a significant relationship between socio-economic variables and the passenger to population ratio of a country.
- The socio-economic variables of a country will not create a predictive model of the passenger to population ratio of a country.
- There will not be a significant relationship between Labour Force Participation and the passenger to population ratio of a country.
- There will not be a significant relationship between GDP per capita and the passenger to population ratio of a country.
- There will not be a significant relationship between Public Expenditure on Education as a % of Government Expenditure and the passenger to population ratio of a country.

Generalizability and Potential Significance

This study will be conducted using highly developed countries, according to United Nations Human Development Index (2019) reporting, and thus can be generalized to similar populations; the applicability to a less developed country is unassumed but gives some room for future testing in that region. The results of the study could yield the ability for agencies on an international level to predict the passenger to population ratio of a country without the complexity of existing methods. The ease of use and accessibility of the results of this study could improve efficiency, in terms of economics and logistics, for a multitude of entities on the airport and those serving the passengers of said airport. Consider a rental car business that has to keep stock for both local business and leisure business, having a simple predictor of passenger to population ratio of the country they are operating in allows them to allocate their stock efficiently between local and leisure business. Many other variables that are typically used in the complex prediction of passenger counts for specific airports or regions are not present during this study; therefore, future testing could include these for stronger generalizability. The goal of the study is to use any potential, and predictive relationship discovered to provide a simple metric for predicting the passenger to population ratio of a country in the hope that it may add in planning purposes for the multitude of entities dependent on the aviation environment.

Limitations and Delimitations

Limitations are the bane of the researcher, the conditions or influences that are out of the researcher's control that place restrictions on a study methodology, results, and

conclusion. This study's first limitation is the accuracy of the data collection by both the United Nations (2019) and IATA (2020) databases. The second limitation of this study is the small sample size of the countries examined in this study; this may result in difficulty finding a significant relationship between variables. Delimitations are the researcher's choices that describe the study boundaries to include the population to be studied. The first delimitation is the choice of using only highly developed countries to examine the relationship; if a variety of development stages were used, then the researcher predicts the results would be less accurate. Another delimitation for this study is the use of accessible and likely well-constructed databases from large entities that have been gathering and analyzing data for a significant period (i.e., IATA and the United Nations).

Chapter 2

Review of Literature

Introduction

This research study aims to identify the amount of passenger traffic in relation to the country's population and make this information available to those who might benefit from its use. Examination of which variables have a relationship with the ratio of passengers to population of a country likely has far-reaching consequences for prediction and planning of a vast array of passenger-based challenges from general passenger flow to the capacity of terminals or security checkpoint areas. This chapter investigates current literature in passenger count prediction, air-travel demand forecasting, and relationships between socio-economic variables and economic conditions. Current research presented in this literature review focuses on programming neural networks and various statistical prediction models to estimate passenger numbers in specific environments such as terminal or security checkpoints. While undoubtedly useful, a broader investigation of what determines the ratio of passengers to the population will provide many industries with easily accessible and prospectively accurate information on expected passenger numbers compared to a country's population. Examining available literature will help in identifying confounding or extraneous variables that may interfere with the independent variable's effect on the dependent variable. Throughout this chapter, the potential impact of a significant relationship between specific variables consisting of GDP per capita, personal or household income, CO2 emissions, quality of life, and passengers' ratio to population are discussed.

Passenger Count Prediction

Predicting passenger count is a complex calculation, and when performed correctly, is useful to airports and a multitude of facilities and operations affected by passenger counts. The various methods that passenger count has attempted to predict will be discussed below, with a genetic algorithm, decision trees, and statistical analyses methodology taking center stage. Laik, et al. (2014) first identify their main problem of how various entities within the airport plan for passenger count or flow. Their research suggests that these entities use a flat figure based on a percent of the passenger load, to complete daily planning from check-in staff required to gate agent availability. Another challenge is the difficulty in estimating the variance in the number of tickets sold and number of ticketed passengers that arrive on time for their flight, their study asserts two solutions to achieve a higher percentage of passengers who arrive for their flight. The first solution is to improve air traffic efficiency, including the most pertinent phases of flight, such as takeoff and landing. The second solution is the optimization of airport passenger capacity and subsequent improvement of airport design because there is a gap in the literature regarding passengers' comfort or convenience. Laik et al. (2014) suggest that as the airline industry gradually grows, passenger-oriented issues become increasingly challenging, as terminals reach peak capacity during the holiday season and peak hours. Any delay at the check-in counter is exponential, significantly further down the line. The aforementioned problems and steady growth of the airline industry emphasize the current

literature gap in needing accurate prediction or forecasting tools to enable greater efficiency in operations.

Laik et al. (2014) research methodology was the decision tree method that is typically used to mine information to predict target variables. According to Song and Lu (2015), the decision tree methodology is a robust, easy-to-use tool. It provides precise results without ambiguity. These attributes make decision trees an excellent tool for the current research, even more so when the data being mined is a large dataset because you can divide the data into learning and validation datasets. This is what Laik et al. did using actual passenger numbers from March 2013 to test their model against historical data, to test the prediction model realistically. The Laik et al. study was conducted using 1 year of data from June 2011 to May 2012 and four independent variables; day of the week, destination cities, time of day, and month of the year to make predictions on the dependent variable, passenger count, for that specific period.

The Laik, et al. (2014) study has one main weakness: the study's limitation is in using solely Asian market data. Using Asian market data means that the population generalizability may only apply to the Asian aviation markets and not further afield. The difference between Asia and the United States of America, for example, when it comes to how the aviation market operates is likely different in many aspects. Overall, Laik et al. concluded that the variance in passenger numbers is based on three factors, destination, day of the week, and month of the year. The prediction model then was tested against past passenger numbers provided slid capability with a root mean square error of 3%-12% for

each of the airlines at the airport that has enabled operators at airports to use the model to predict the number of available staff throughout the day.

Srisaeng, et al. (2015) take a slightly different angle to predict passenger count. Srisaeng et al.'s approach was trying to answer the question of whether we can use genetic algorithm optimization models as a separate predictive tool to the status quo to accurately predict passenger count? Delving into the current methodology and the base materials of those methodologies, the study notes that multiple linear regression models are historically used to predict passenger count. The information available on predicting passenger count in the International Civil Aviation Organization (ICAO) Manual on Air Traffic Forecasting uses best practices from 1985. The study discovered that there may be a more accurate way to predict passenger count given the age of ICAO's ICAO's Manual on Air Traffic Forecasting and the more naturalistic approach that genetic algorithms take to multiple linear regression. This study uses a surprising methodology in the form of a genetic-algorithm optimization model that might be familiar to most as the process of natural selection. Akgüngör and Doğan (2009) best explain the idea behind using genetic algorithms as "Genetic algorithms differ substantially from traditional optimization methods because they search using a population of points in parallel rather than a single point in order to obtain the best solution" (p. 477). This applied research appears to have no controversy in the source of funding or interest that would influence the findings. It appears to try and theoretically add value to the existing literature on predicting passenger count. The study is quantitative, and after extensive literature review, 11 variables were chosen as independent variables to be tested. The 11 variables used are mostly related to

the financial growth, which makes sense as Wensveen (2011) explains the influence of economic growth on air travel demand as the prime mover. When talking specifically about leisure travel demand, Doganis (2009) informs us that the most crucial socio-economic variable is personal or household income when it comes to leisure travel, which further adds validity to using financial-based variables to predict passenger count. The methodology used in this study empirically examined both linear and quadratic GAPAXDE and GARPKSDE models with the dataset of the population from Q4 1992 to Q2 2014. The results were that the quadratic form provided better accuracy, reliability, and predictive power than the linear form. To aid the GAPAXDE and GARPKSDE models' predictive power, two datasets were created, one for testing and the other for training the models. After training had been completed, the testing phase put the prediction models of the linear and quadratic forms of the GAPAXDE and GARPKSDE against the actual dataset. It returned results of 4.89%, and 2.23% mean absolute percentage error values, respectively. The quadratic form of the GAPAXDE and GARPKSDE models proves to be more accurate as a predictive model. When the hypothesis was tested using the T-test, the quadratic form of the GAPAXDE and GARPKSDE models gave evidence of statistically significant differences from the linear form regarding the average forecasting error. In order to make the study more robust, the study could have extended the focus from just Australia to an international focus. The Australian market's focus raises further questions on the applicability or, specifically, population generalizability of this study on different economic systems. For example, Australia is a fairly well-developed country, ranked sixth place by the United Nations Human Development Index (2019) measure. A less-developed

country may have variance in the chosen variables. A less-developed country may have variance in the chosen variables. The study's findings by Srisaeng et al. are that quadratic genetic algorithm models were the most accurate and provided the greatest predictive capability with mean absolute percentage errors being 2.55% and 2.23% for the quadratic GAPAXDE and GARPKSDE models, respectively.

The articles presented above provide strong evidence to the use of socio-economic variables to predict passenger to population ratio of countries. The literature suggests that there may be some relationship between financial variables such as GDP and personal available expenditure and the ratio of passengers to population. Many of the research studies examined choose to include GDP per capita in the group of variables that could influence passenger numbers, both Australia's real GDP per capita and real GDP were chosen as variables to be tested and included. These research studies validate the GDP being included by showing greater accuracy in models when a GDP variable is being used. For example, Srisaeng et al. state that "the inclusion of both real GDP and Australia's population size provided more robust and accurate model forecasting capability; that is, the models utilizing real GDP and population gave better errors" (p. 483), which leads to the conclusion that GDP as a socio-economic variable may have potential as a predictor.

Comparison of Demand Forecasting Methodologies

For air-travel demand forecasting, the currently available methodologies range from complex regression analyses to simpler econometric models. The study by Karlaftis et al. (1996) addresses whether the complexity of a methodology influences the forecasting accuracy and predicting the ability of that methodology. Furthermore, Karlaftis et al. hope

to use the information learned from the analysis to create an analytical framework to develop econometric models that will be accurately tested using past factual data. In the review of the currently available literature, Karlaftis et al. started right at the beginning of air-travel demand prediction by looking at the seminal work from the late 1940s to the early 1950s by Harvey (1951) and Mayhill (1953). Harvey and Mayhill were innovative in using the gravity model for city pair air-travel demand that suggests that the strength of the interaction between two places is a product of their populations divided by the square of their distance apart. The modern focus of the Karlaftis et al. study is on time series models that use passenger and aircraft movement data combined with socio-economic and economic indices. Bartlett (1965) made an attempt at deductively predicting revenue passenger miles per capita (RPM) for the entirety of the United States, whereas Jacobson (1970) tried to predict demand of a specific airport in Virginia. Frankfurt, Germany, and Miami, United States of America airports are the two test benches of the Karlaftis et al. prediction model and uses time-series design to create an accurate predictive model. Time-series design models are a great fit for this type of research as airport data is widely available, and time and date are all logged. The choice of two countries adds robustness to this study as they are geographically distant from one another, eliminating or exposing some internal threats to validity like a history threat. This distance also helps eliminate any outlying extraneous variables like a local short-term economic influence, i.e., local recession due to a bad harvest year. This study identified variables that fit existing preliminary models, i.e., a priori specification, and came up with population, income, GNP, and price of travel as explanatory variables. The conclusions of the study indicate

that the simple models' performance was adequate with an R^2 value of between .72-.94, but more; complex models in the existing literature had higher R^2 values, yet R^2 should not be used as an objective measure of quality. To combat these conclusions, Karlaftis et al. state that a high R^2 does not have a causal relationship with the accuracy of a prediction model, and to avoid any multicollinearity issues simplicity is key. Multicollinearity is addressed in the conclusion, where two variables are so highly correlated it can skew findings, by stating that models with more variables generally suffer from multicollinearity issues, thus, emphasizing the use of a simpler model. Karlaftis et al. raise the following question: If it were possible to inductively deduce the variables that most influence each specific area, then could an airport-specific approach be applied to create a more solid prediction tool in the future?

Carbonneau et al.'s (2008) study perform a comparative analysis for existing prediction models with the absence of information for participants' demands inside an entire supply chain. Several methodologies were tested with distorted demand conditions, including advanced machine learning, neural networks, linear regression, trend, and moving average, among others. The methodologies are separated into two categories, with traditional methods and machine-learning methods compared. Carbonneau et al. uses correlational research methodology to compare two sets of data, one simulated and one actual, with each type of methodology to examine the differences in accuracy and predictive power of each. One of the strengths of this study is the use of archival data, which eliminates any issues with choosing or modifying existing instrumentation to ensure reliability and validity. This comes with its risks of having to assume the dataset has been

gathered and cleaned accurately by the owner. Using both simulated and actual data sets also adds some validity to the research. On the weakness side, this study uses existing data and estimated data from the Canadian Foundries and has a lack of population generalizability outside of Canada. The simulated demand distortion of the supply chain could also be seen as a weakness because each of the four simulated supplier entities have the same structure and behavior, whereas, in real life, this would likely not be the case. Carbonneau et al. found that the Recurrent Neural Network and the Least-Squares Support Vector Machine obtained the best results, especially in the actual demand data set over the simulated demand data set. The Support Vector Machine had the best overall accuracy between both data sets and provided the greatest generalizability. Coming in the last place were the Trend and Naïve forecasting models with the highest level of mean average error. Carbonneau et al. note that despite the differences in accuracy, there were no statistically significant differences in terms of accuracy between Recurrent Neural Networks, Neural Networks, Support Vector Machine, and Multiple Linear Regression. Carbonneau et al. conclude that while advanced forecasting techniques do provide better accuracy, lowering cost and raising efficiency, the difference was not large enough to switch from the traditional forecasting models for the simulated data set. However, in the real data sets, the advanced forecasting methods did provide a large enough improvement in accuracy over simpler models like the trend, naïve, and moving average. The more advanced forecasting models did not significantly, benefit the multiple linear regression model. Thus, using the conceptual and computational simplicity of MLR over the advanced forecasting models is questionable.

The overall conclusion of these studies is that the dynamic nature of demand makes it very difficult to predict, especially when considering the multiple entities that operate within an airport or supply chain process. The studies bring light to what forecasting methods may be worth further investigation and whether the forecasting model's complexity plays a role in the practical accuracy of the outcome. It should be considered that social sciences are a complex beast that likely does not have a one-size-fits-all style model.

Economic Growth and Convergence

Passenger demand and economic growth should be logically related when taking into account a growing country that may be transitioning its primary form of industry, which could bring more passengers. This makes the economic growth of a country relevant when we are discussing the areas of research for the ratio of passengers to population prediction. As such, economic growth will be discussed alongside relationships between economic growth and carbon emissions and the convergence of GDP per capita across a spectrum of countries.

Guvercin's (2019) correlational research addresses the effects of GDP per capita and the crime rate on carbon emissions. It uses the Granger causality test to try and draw more than correlation from the variables. These three variables share somewhat of a triumvirate relationship amongst the existing literature, and the aim is to understand their potential interdependency further. Previous research on the relationship between economic growth and carbon emissions was firstly established by Grossman and Krueger (1993; 1995) that Guvercin explains showed an inverted-U shaped relation between economic

growth and carbon emission level called the Environmental Kuznets Curve (EKC). Guvercin's study used data from 21 European countries over 18 years from 1996 to 2014. Noteworthy ideas from Guvercin include the idea of once a population is adequately wealthy, they will demand that the air quality, or indeed the quality of life, become higher, and thus the carbon emissions would likely fall as noisy and unclean factories will be removed at this request. Guvercin states that another implication of the EKC's inverted-U shape is the relationship between economic development and the refinement and rigorousness of environmental laws as the wealthier a country becomes more aware of of the environmental impact both the population and governing bodies become. A simple example would be a country that is mainly primary industry releasing more emissions whereby a country that is mainly secondary industry would be taking those products from the primary industry and manufacturing them, likely being cleaner in terms of carbon emissions. The relationship between economic growth and carbon emissions is no one-way street; however, carbon emissions can have negative effects on economic growth by changing the quality of air and/or water (Jacobson, 2008; 1-5) that affects the price a willing-purchaser might pay to obtain goods from the agricultural and raw materials markets. Watson et al. (2005, 836-838) observed a reduction in the workforce's productivity and a resultant decline in GDP per capita as a result of an intensification of currently existing medical conditions among the workforces. Although the number of passengers a country has in relation to the country's population is not the be all and end all when it comes to carbon emissions, the aviation industry is certainly one of the inputs when it comes to carbon emissions.

The results of Guvercin's (2019) publication indicate that a reduction in carbon emissions from carbon-intensive production leads to a decline in both crime rate and GDP per capita. More specifically, carbon emission and homicide crimes cause GDP per capita, and carbon emissions and GDP per capita cause homicide. Carbon emissions account for 40% of the variance in GDP per capita, but there is no causal evidence for GDP per capita or homicide, causing carbon emissions. Counter to Guvercin's research is the research of Holtz-Eakin, and Selden (1995) and Azam (2016) that discovers the impact of economic growth on carbon emissions having a positive impact as opposed to concave. Meanwhile, Agras and Chapman (1999) and Richmond and Kaufmann (2006) state the there is no relationship between economic growth and carbon emissions. Given the statistical evidence provided by Guvercin's publication and the opposing research, Guvercin's publication adds to a fair wealth of information on the subject of GDP per capita and carbon emissions being related, but it is the first in the field to add the variable crime rate to the mixture and thus is the first empirical investigation of all three variables.

The greatest strength of Guvercin's (2019) study is that the three variables are being analyzed for the first time together as previous attempts had only looked at two and that the results find a medium effect size of carbon emissions on GDP per capita variance. This is also Guvercin's greatest weakness as attempting to ascertain some causal evidence for three interacting variables is likely to be very difficult as the three variables show some relationship but are found to respond to each other differently given a one-time standard deviation change in any variable. When concluding the study, Guvercin noted that there is some interdependency between variables and that Granger causality tests show some cause

evidence between two of the three pairs of variables on one another. This aids with the selection of socio-economic variables to be tested in future research and guides to better understand how these three variables interrelate and whether multicollinearity will be a large threat when two or more of these variables are used.

Boyle and McCarthy (1999) set out to investigate the statistical evidence of GDP per capita convergence using 84 countries over the period 1960 to 1992. Of the 84 countries, the World Bank typology was used to group the countries into four distinct groups low income, lower middle income, upper middle income, and high income. Examining the convergence of GDP per capita across different wealth zones should further concrete the reliability of the GDP per capita variable and perhaps eliminate the potential for extraneous variables in the potential relationship between GDP per capita and the passenger to population ratio of a country. Sala-i-Martin (1996, p. 1326) gives us a simple and valuable insight into the general trend of convergence by quoting the mnemonic rule of thumb "economies converge at a speed of two percent per year" that serves to help understand the implications of Boyle and McCarthy's (1999) study a little more easily. Young et al. (2007) define the two growth empirics examined in the Boyle and McCarthy study, σ convergence and β-convergence, as "When the dispersion of real per capita income across a group of economies falls over time, there is σ convergence. When the partial correlation between growth in income over time and its initial level is negative, there is β-convergence" (p.1). In order to amplify understanding of these two principles, graphical and more simple definitions are provided below. Sigma convergence relies on the concept that a less-developed country can grow at a faster rate than a more-developed

country and although one may start at 60,000 GDP per capita and the other at 10,000 GDP per capita, they will eventually converge. Figure 2.1 below is a visual example of that where the big country starts at 60,000 GDP per capita and grows at 2% per year, and the small country starts at 10,000 GDP per capita and grows at 10% per year. In the beginning, the gap is fairly large at 50,000 GDP per capita, but the largest gap is at year 4 at 50362 GDP per capita, and as you can see, the graph starts to converge after year 5.

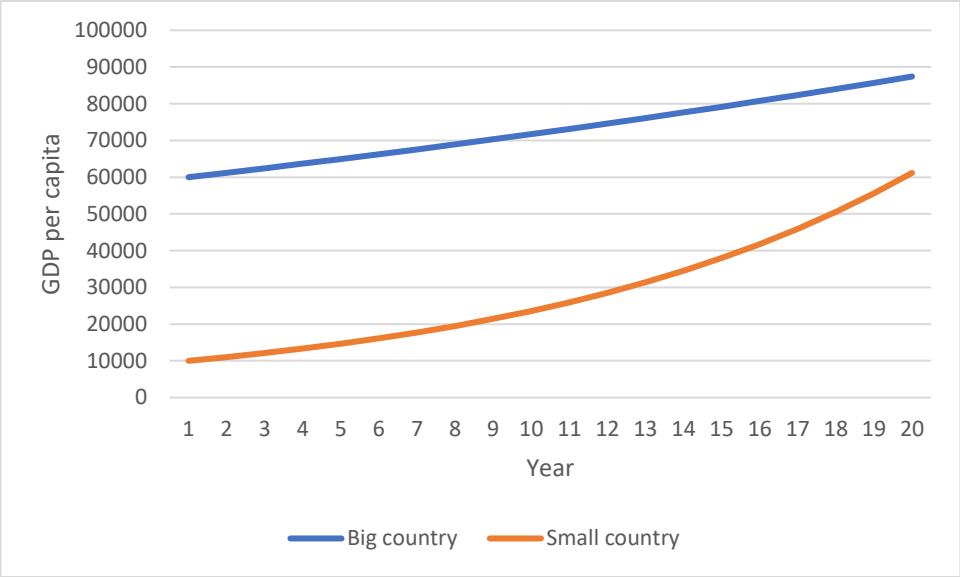

Figure 2.1. *Sigma Convergence*

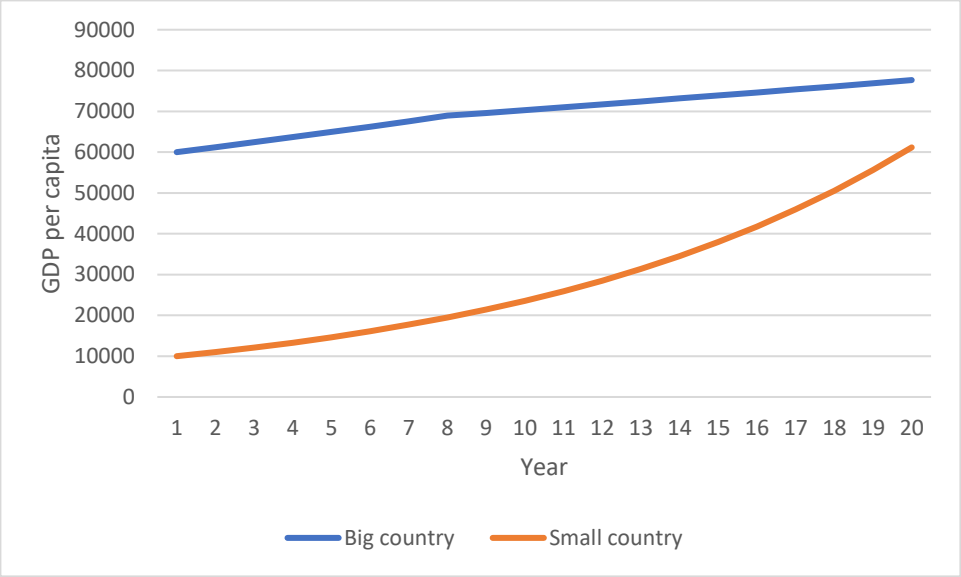

Figure 2.2. *Beta Convergence*

Beta convergence is a term for when the growth of a country slows down. It approaches its steady-state that typically occurs more often for more-developed economies as there is less room for opportunity, and improvements made in those areas are less impactful to an economy. From Figure 2.2 above, it is visible that in the year 20, the big country ends at below 80,000 GDP per capita while the sigma convergence graph ends at around 90,000 GDP per capita at the same point. The big country's growth rate slowed from 2% per annum to 1% per annum and caused convergence in its own right to call beta convergence. Figure 2.3 shows the difference between the two countries in both sigma and beta convergence with the small country remaining the same in both examples and the big country showing negative change due to beta convergence.

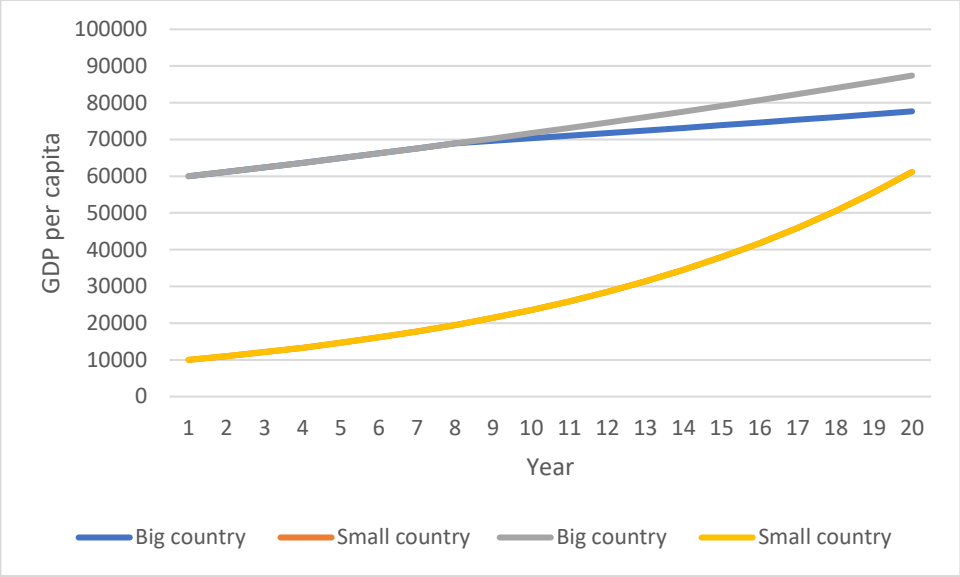

Figure 2.3. *Sigma Convergence vs. Beta Convergence*

Boyle and McCarthy (1999) give us some insight into how the GDP per capita variable may naturally fluctuate that provides further understanding of how GDP per capita and the ratio of passengers to a country's population may interact. The results that Boyle and McCarthy suggest that high income and upper-middle-income countries show sigma convergence. Still, lower-middle-income countries showed essentially negligible sigma convergence with low-income countries diverging. Beta convergence was demonstrated not to have occurred throughout the period preceding the late 1980s and thereafter only occurred in the low-income countries starting 1981. The strengths of Boyle and McCarthy's research comes in the study of multiple income groups of countries over a fairly large date range and the fact that that this study adds to an already sizeable database on economic convergence, the breadth of the research lends itself to good population and

ecological generalizability. The weaknesses of Boyle and McCarthy's research would be the lack of data available or used for the last 20 years or so and any historical effects that may have occurred during that time period may not have gone noticed. If this data had been available or studied, then it would give the reader much more of a modern base for the implications of the research as the economy is likely different from 20 years ago.

In conclusion, for this study, Boyle and McCarthy (1999) suggest that the inverse correlation between income group and degree of sigma convergence provides some thought-provoking ideas. The nature of economic convergence is roughly understood but still unpredictable leading to leapfrogging in GDP per capita rankings, especially in lower-income countries, at certain annual benchmarks. This not only provides some information for delimitations, like choosing certain income-grouped countries to research but provides insight into the predictive power of financial variables alone that will help guide future research variable selection.

Chapter 3

Methodology

Population and Sample

Target Population

This study's target population consists of highly developed countries that were high on the United Nations Human Development Index (2019) scale. The main reason for using only highly developed countries was to limit the influence of extraneous variables on the dependent variable. The reasoning mentioned above is based on the assumption that highly developed countries are all roughly at the same stage of socio-economic development. Thus, the socio-economic variables should be more predictive, assuming some correlation exists. Previous literature suggests that developed countries' environmental and financial desires change in a uniform way from a desire for economic growth to a desire for health and social. For example, as more wealth is available to a country, less of the budget goes into decreasing poverty levels and more of the budget goes into research and development or improving environmental conditions. Research and development budget and tourism expenditure are both socio-economic variables whose relationship with the ratio of passengers to the population will be examined by this study.

Sample

The sample for this study was gathered in what can best be described as convenience sampling, this nonprobability-based sampling method lacks randomness but is appropriate to use considering the purposeful nature of delimiting the dataset in the hope of finding a pattern within boundaries of the United Nations Human Development Index

(2019). The following countries were part of the sample: the United Kingdom, the United States of America, Australia, Belgium, Hong Kong (China), France, Germany, Italy, Japan, Netherlands, Norway, Singapore, Sweden, Switzerland, and Iceland. This study's sample was all of the developed countries that had the chosen socio-economic variable data contained within the United Nations (2019) database and passenger numbers on the IATA (2020) database. The sample countries' mean population was 26,676,343, with a median of 21,150,150, and the mean passenger count was 91,928,550, with a median of 75,833,340.

Statistical Analysis

Assumptions

Multiple Linear Regression and filter-based feature selection will be used to analyze the data collected for this study statistically. The assumptions required to use the MLR are as follows: the data must have a linear relationship, multivariate normality, no multivariate collinearity, homoscedasticity, and at least two independent variables. Microsoft Excel will be used as the source of statistical analysis for this study.

Statistical Test. This study will use Multiple Linear Regression to analyze the data collected from the archival data and then filter-based feature selection to narrow down the best predictive model. The MLR has to meet a number of assumptions to be classified as satisfactory.

1. A linear relationship must exist between the independent and dependent variables, typically a scatterplot graph is used to establish a linear relationship between two or more datasets.

2. The residuals of the regression analysis should be normally distributed, which means that the differences between predicted and observed values should be normally distributed. A histogram can be visually inspected to ensure that the second assumption is met or, alternatively, a goodness-of-fit test can be conducted to ensure normality.
3. Multicollinearity is the third assumption that must be checked for Multiple Linear Regression. Multicollinearity can provide an issue whereby two of the independent variables are highly linearly related and thus change at the same rate when one of the collinear independent variables is adjusted. This is a huge problem when you are trying to identify how each independent variable affects the dependent variable. Multicollinearity can be checked in a number of ways, in this case a Pearson's bivariate correlation matrix was created with Microsoft Excel and a threshold of 0.80 was chosen to not be exceeded. There were 3 instances of correlation over 0.80 but none that were involved in the successful iterations of the predictive independent variable group.
4. Homoscedasticity is the assumption that the variance of the residuals is roughly the same at all points through the data. This can be visually inspected by observing the distance from the regression line of a set of predicted values and a set of standardized residuals.

Procedures

Research Design/Methodology

This study's design was chosen to compare several socio-economic variables' potential relationships and predictive power with the ratio of passengers to the population of a country. After an extensive literature review, socio-economic archival data was selected, obtained, and analyzed from the United Nations (2019) database. The independent variables are each of the socio-economic variables. A correlational research design is used to aid statistical analysis of the independent and dependent variable relationships. Correlational research is the appropriate choice for this study as there is only one group, and the study is analyzing the relationship between that group and multiple other variables statistically.

The analysis will begin with checking if there is any large multicollinearity between the independent variables by plotting a Pearson's bivariate correlation matrix. If larger than 0.80 of correlation is found between the variables and the variables both contribute significantly to the predictive power, the least significant will be removed from the study. Each independent variable will then be examined using Multiple Linear Regression analysis against the passenger to population dataset to determine if any relationship exists. Then Multiple Linear Regression analyses will be run for every variation of independent variable groups possible and the dependent variable to find the best combination of independent variables for modelling the dependent variable. Every number of independent variables will be modelled, i.e., 1 variable model, 2 variable model, 3 variable model and so on, with the best models then tested against each other to leave the

best of the best models. Filter-based feature selection will be completed using each model's R^2 value, which, as previously mentioned, is not the ultimate value for the completeness in a predictive model as we do not want overfitting to occur but is a good general indicator of a potential relationship. As previously mentioned R^2 value is not the ultimate test and to counteract potential overfitting issues, five-fold cross-validation will be used on the best model to ensure overfitting is avoiding and the model is more useful in general than just being able to predict this certain set of data. Cross-validation is a process whereby the data is split into random datasets, in our case five sets, and the model is trained on datasets 2,3,4, and 5 and then tests the trained prediction on dataset 1, which is then repeated for every variation of the five datasets. Pending the hypothesized results, the best model will be used to predict passenger to population ratio of a country, be accurate and generalizable enough to be useful.

Time Schedule and Budget

Study proposal approval – December 2020

Data collection – November 2020

Data analysis – January and February 2021

Study conclusion and final report – April and May 2021

Total cost for the study: The study did not require a budget because all data was gathered online and freely available to the public domain. Data analysis software to be used is freely available to the researcher.

Chapter 4

Results

Introduction

The purpose of this study was to investigate any relationships and display any potential predictive models between socio-economic variables and the passenger to population ratio of a country. This chapter presents the correlational research results using MLR and filter-based feature selection, as discussed in Chapter 3. For this study, MLR provides the primary method to determine statistical significance in the form of a coefficient of determination (R^2), a value that measures the regression line's correlation with the dataset. However, merely using an R^2 value is not stringent enough to create a valuable and generalizable model. Thus, five-fold cross-validation was introduced at the end of the analytical process to ensure overfitting is not a problem. Descriptive statistics and inferential statistics from the MLR provide insight into the sample demographic characteristics and representativeness will be presented in this chapter.

As identified in Chapter 1, the research questions and their respective hypotheses are as follows:

Research Question 1

What is the relationship between various socio-economic variables and the passenger-to-population ratio of a country?

- $H_{01:}$ There will not be a significant relationship between socio-economic variables and the passenger-to-population ratio of a country.

- $H_{A1:}$ There will be a significant relationship between socio-economic variables and the passenger-to-population ratio of a country.
- $H_{02:}$ There will not be a significant relationship between Labour Force Participation and the passenger-to-population ratio of a country.
- $H_{A2:}$ There will be a significant relationship between Labour Force Participation and the passenger-to-population ratio of a country.
- $H_{03:}$ There will not be a significant relationship between GDP per capita and the passenger-to-population ratio of a country.
- $H_{A3:}$ There will be a significant relationship between GDP per capita and the passenger-to-population ratio of a country.
- H04: There will not be a significant relationship between Public Expenditure on Education as a % of Government Expenditure and the passenger to population ratio of a country.
- $H_{A4:}$ There will be a significant relationship between Public Expenditure on Education as a % of Government Expenditure and the passenger to population ratio of a country.

Research Question 2

Can a predictive model using various socio-economic variables predict the passenger-to-population ratio of a country?

- $H_{05:}$ The socio-economic variables of a country will not create a predictive model of the passenger-to-population ratio of a country.

- $H_{A5:}$ The socio-economic variables of a country will create a predictive model of the passenger-to-population ratio of a country.

Descriptive Statistics

This research used a sample size of 14 participants (N = 14). The data collected from the databases are considered complete, and thus, handling missing data was not an issue that required a procedure to process. Data was input into Microsoft Excel to identify demographic and descriptive information. There was one outlier in the data collection from Iceland. Iceland has a population of 33,900 but received over 7 million tourists and visitors during the data period, which is substantially different from the other 14 sample countries' general trend. For example, the passenger to population ratio range is from 1.48 to 8.48, and Iceland's passenger to population ratio is 227.27 and was removed from the study as an outlier. The sample countries' mean population was 26,676,343, with a median of 21,150,150. The mean passenger count was 91,928,550, with a median of 75,833,340. Of the 14 countries that comprised the sample, nine (64%) of the countries are located in the European continent, three (21%) of the countries are located in the Asian continent, one (7%) of the countries is located in the North American continent, and one (7%) of the countries is located in the Australian continent. The countries' gender distributions and age demographics will be displayed and discussed below. The population numbers are from the latest available estimates available in the United Nations statistical database (2019). They are different from those used in the analysis and vary in the year the data was collected but serve to provide insight into gender distribution in each country. Using the latest population information provides a more up-to-date picture of gender and age distribution

without being at the mercy of the complete availability of our group of independent variables. Displayed below is Table 4.1 that provides information on each country, its population, and a breakdown of the gender demographic within that population.

Table 4.1

Countries, their Population, and Gender Demographics

Country	Population	Male (%)	Female (%)
United States of America	308,745,538	151,781,326 (49)	156,964,212 (51)
United Kingdom	63,379,787	31,126,054 (49)	32,253,733 (51)
Australia	23,717,421	11,686,665 (49)	12,030,751 (51)
Belgium	11,000,638	5,401,718 (49)	5,598,920 (51)
Hong Kong	7,336,585	3,375,362 (46)	3,961,223 (54)
France	64,300,821	31,138,550 (48)	33,162,271 (52)
Germany	80,219,695	39,145,941 (49)	41,073,754 (51)
Italy	59,433,744	28,745,507 (48)	30,688,237 (52)
Japan	127,094,745	61,841,738 (49)	65,253,007 (51)
Netherlands	16,655,799	8,243,482 (49)	8,412,317 (51)
Norway	4,979,955	2,495,777 (50)	2,484,178 (50)
Singapore	3,771,721	1,861,133 (49)	1,910,588 (51)
Sweden	9,482,855	4,726,834 (50)	4,756,021 (50)
Switzerland	8,035,391	3,973,280 (49)	4,062,111 (51)

Inferential Statistics

The purpose of this research was to investigate relationships between socio-economic variables and the passenger to population ratio of a country, then take those relationships and attempt to create a valuable and generalizable model for predicting the passenger to population ratio of a country. Data were selected, treated, and then analyzed by the R^2 value to confidently answer the research questions, identifying possible trends or relationships amongst the independent variables. Two primary methods of data analysis were used, consisting of MLR and filter-based feature selection. MLR is relevant to this

study as regression analysis is commonly used for forecasting and prediction purposes. It is appropriate for this study because it presents how variation in the independent variable(s) is connected to variation in the dependent variable. Filter-based feature selection can be described as a method of scoring each subset of data by whatever features the user has instructed the software to filter by, which is very appropriate in this research. It searches for the best relationship predictive model adjudicated primarily by the R^2 value. Also, cross-validation works in tandem with filter-based feature selection to provide control on where the cut-off should be for the scoring of features. All assumptions for the appropriate usage of data analysis methods were met. The assumptions required to use the MLR are as follows: the data must have a linear relationship, multivariate normality, no multivariate collinearity, homoscedasticity, and at least two independent variables.

Variables as Individuals

The initial testing involved testing each variable and possible combination of variables being adjudicated by R^2 value. Table 4.2 provides a list of the variables and the respective identifier assigned during data analysis. The public expenditure on education as a percentage of government expenditure variable appeared to be the best as an individual variable and features throughout all models. As an individual variable, public expenditure on education as a percentage of government expenditure scored $R^2 = 0.80$, $p = <0.00001$ and in our most successful cross-validated model scored a mean cross-validated $R^2 = 0.99$, $p = <0.000003$. A graphical representation and a short textual description of each individual variable are provided below in Figures 4.1 through 4.14.

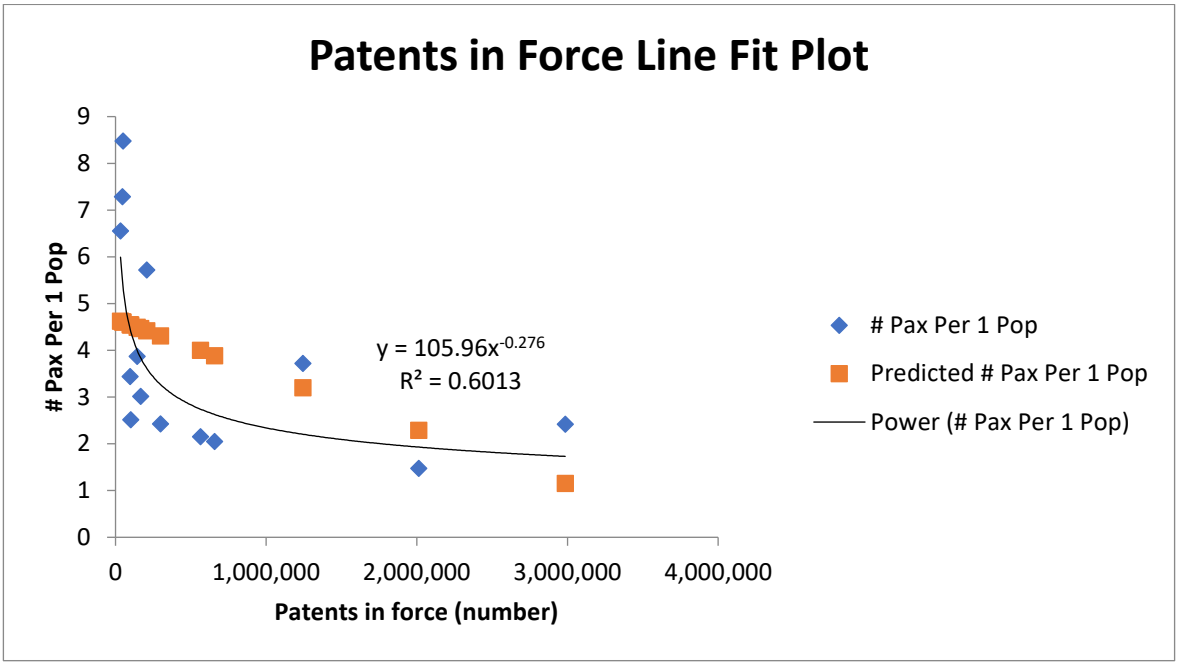

Figure 4.1. *Patents in Force Univariate Regression Line Fit*

Figure 4.1 displays Patents in Force, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between Patents in Force and the number of passengers per population could be due to an assumed relationship between the number of patents a country has and the overall level of technology. Perhaps, with a greater level of technology, a country could have more aviation activity and general traffic in and out of the country, leading to more passengers per population. When a linear regression line is used, Patents in Force, the orange points on the scatterplot, holds an R^2 value of 0.23 that infers a relatively weak explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 105.96x^{-0.276}$, and holds an R^2 value of 0.60 that

shows the highest R^2 value available of the five trendline types. The power trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Patents in Force.

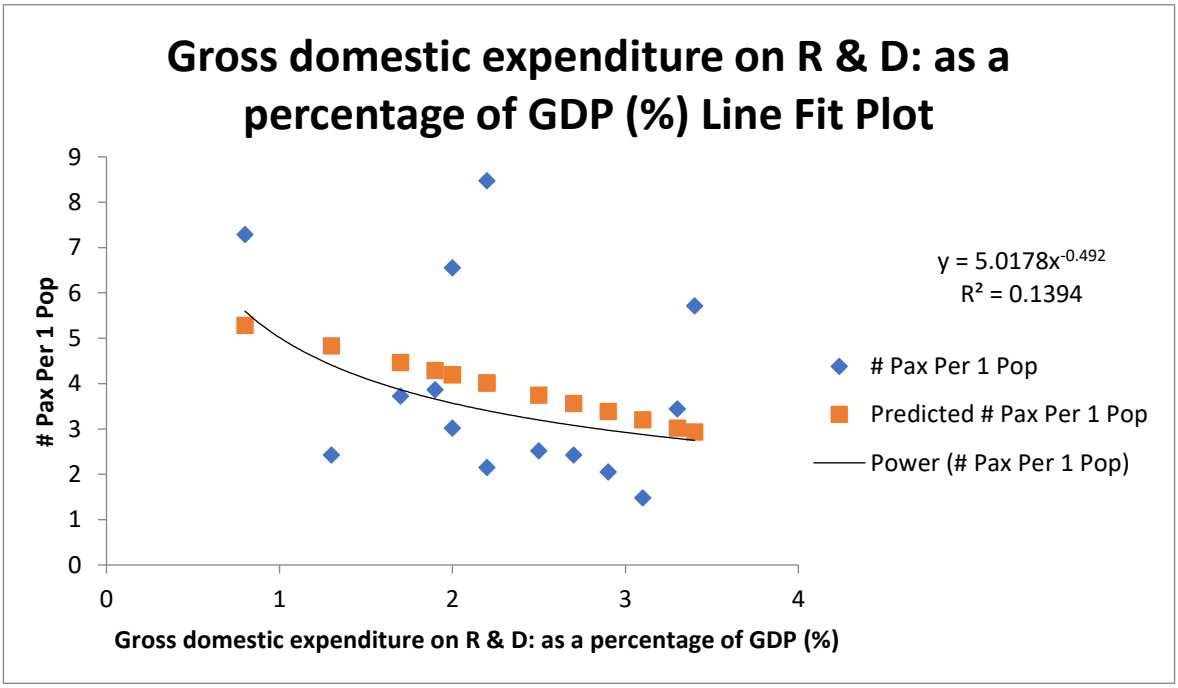

Figure 4.2. *Gross Domestic Expenditure on R & D as % of GDP Univariate Regression Line Fit*

Figure 4.2 displays Gross Domestic Expenditure on R & D as % of GDP, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between Gross Domestic Expenditure on R & D as % of GDP and number of passengers per population could be due to an assumed relationship between the amount of money a country has to spend on research and development and the desirability of that country to be frequented by visitors from tourists to scientists. Also, if a country has excess money to spend more on R & D,

then it is probable that the country is developed in other aspects and possibly more desirable to travel or work in/with as a result. Gross Domestic Expenditure on R & D as % of GDP when a linear regression line is used, the orange points on the scatterplot, holds an R^2 value of 0.10 that infers a weak explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 5.0178x^{-0.492}$, and holds an R^2 value of 0.14 that shows the highest R^2 value available of the five trendline types. The power trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Gross Domestic Expenditure on R & D as % of GDP.

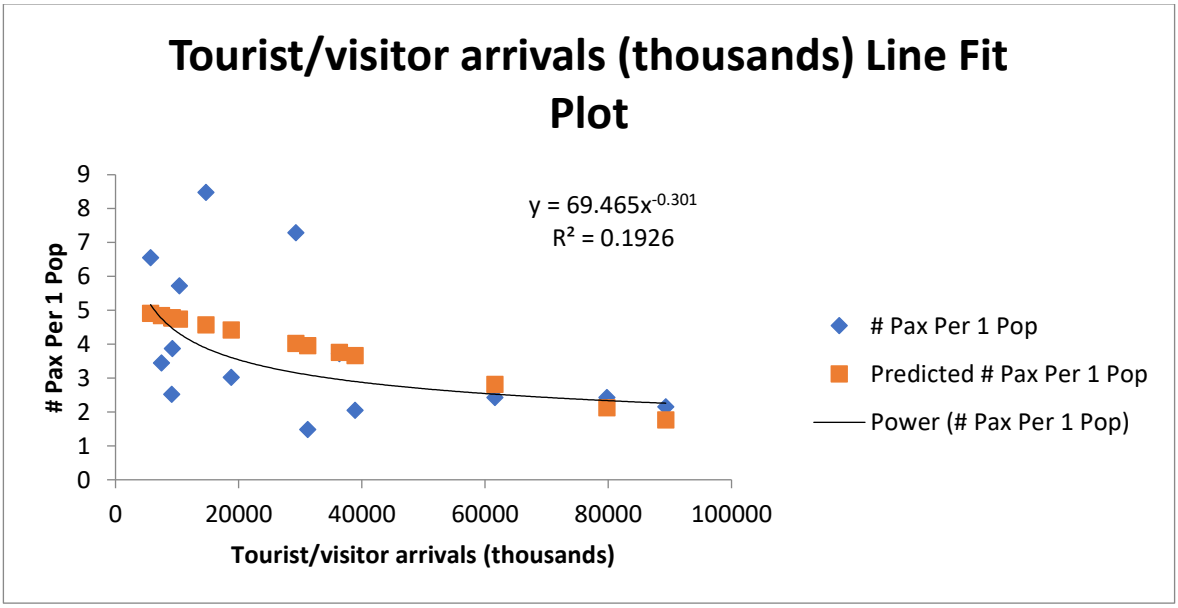

Figure 4.3. *Tourist/Visitor Arrivals Univariate Regression Line Fit*

Figure 4.3 displays Tourist/visitor arrivals, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between Tourist/visitor arrivals and the number of passengers per population could be due to the country's desirability to be visited for leisure purposes. We can then make a logical observation that the more desirable a country is, the more tourists and visitors would arrive, and the passenger to population ratio would then increase. Tourist/visitor arrivals, when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.22 that infers a relatively weak explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 69.465x^{-0.301}$, and holds an R^2 value of 0.20 that shows the highest R^2 value available of the five trendline types. The power trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Tourist/visitor arrivals.

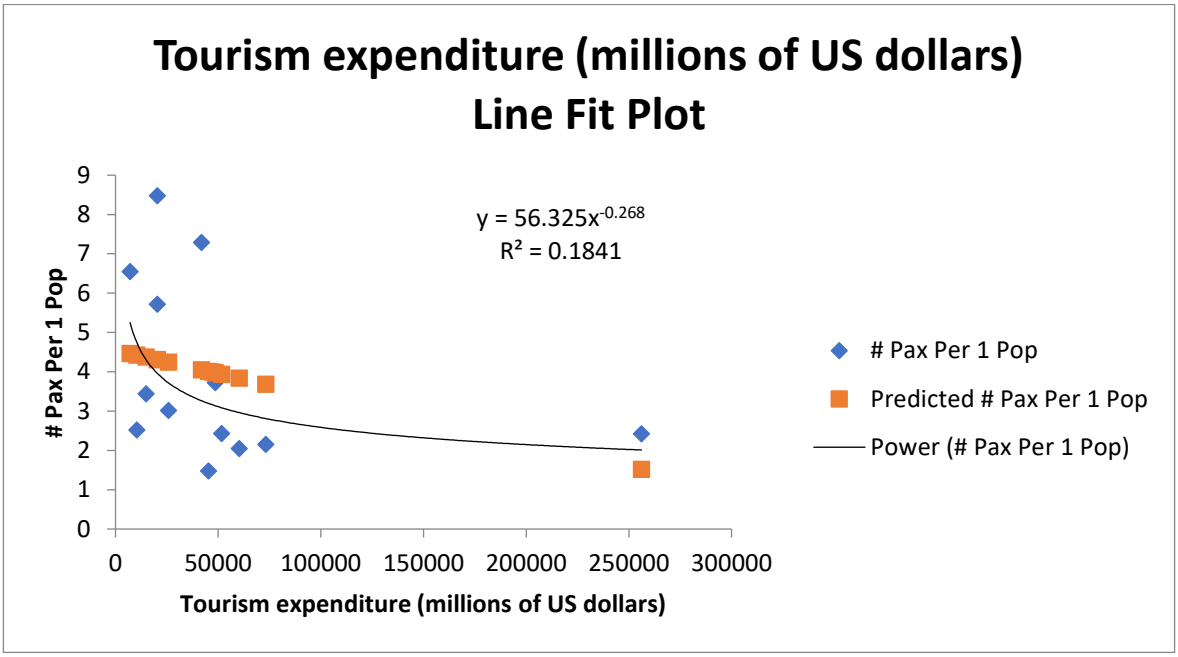

Figure 4.4. *Tourism Expenditure Univariate Regression Line Fit*

Figure 4.4 displays Tourism expenditure, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between Tourism expenditure and the number of passengers per population could be logical as tourism expenditure measures how much money each tourist spends on a visit to the said country; the more money spent as a tourist in a county could indicate how desirable that country is to visit from a tourism perspective an example of which could be Monaco. Tourism expenditure when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.11 that infers a weak explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 56.325x^{-0.268}$, and holds an R^2 value of 0.18 that shows the highest R^2 value available of the five trendline types. The power trendline is used here for two

reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Tourism expenditure.

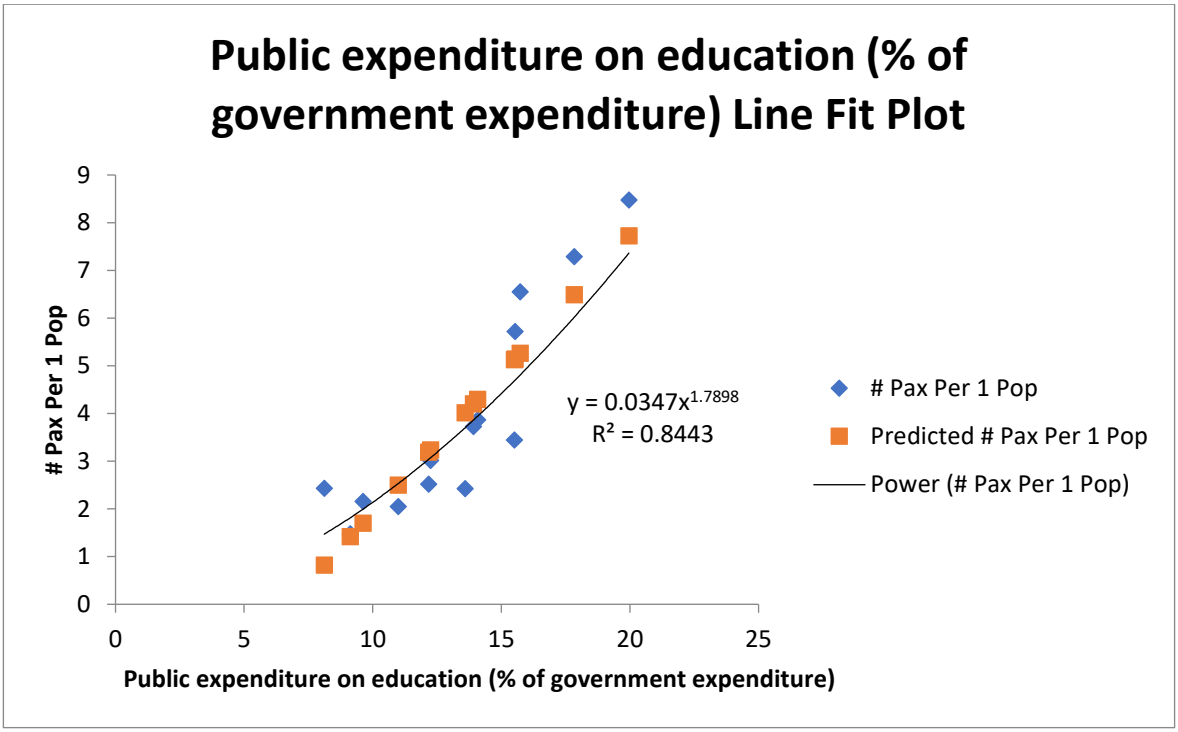

Figure 4.5. *Public Expenditure on Education as % of G'ment Expenditure Univariate Regression Line Fit*

Figure 4.5 displays Public expenditure on education as % of government expenditure, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between Public expenditure on education as % of government expenditure and the number of passengers per population could be due to an assumed relationship between the amount of money that the government chooses to spend on education and the country is more desirable to visit; making a country and the country's population more educated could lead to technological

and societal advancement that warrant visits from scholars, tourists, and prospective migrants. Public expenditure on education as % of government expenditure when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.80 that infers a pretty strong explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 0.0347x^{1.7898}$, and holds an R^2 value of 0.84 that shows the highest R^2 value available of the five trendline types. The power trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Public expenditure on education as % of government expenditure.

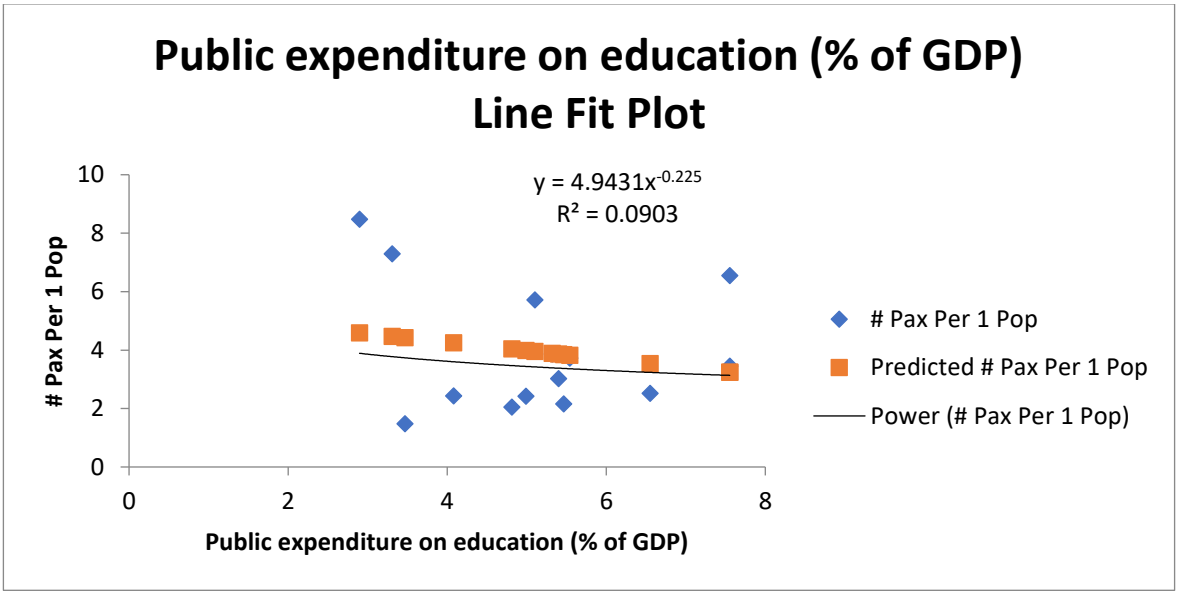

Figure 4.6. *Public Expenditure on Education as % of GDP Univariate Regression Line Fit*

Figure 4.6 displays Public expenditure on education as % of GDP, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per

population. Hypothetically, the relationship between Public expenditure on education as % of GDP and the number of passengers per population could be due to an assumed relationship between the amount of money that a country chooses to spend on education in relation to the gross domestic product of the country and the country is more desirable to visit; making a country and the country's population more educated could lead to technological and societal advancement that warrant visits from scholars, tourists, and prospective migrants. Public expenditure on education as % of GDP when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.04 that infers a very weak explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 4.9431x^{-0.225}$, and holds an R^2 value of 0.09 that shows the highest R^2 value available of the five trendline types. The power trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Public expenditure on education as % of GDP.

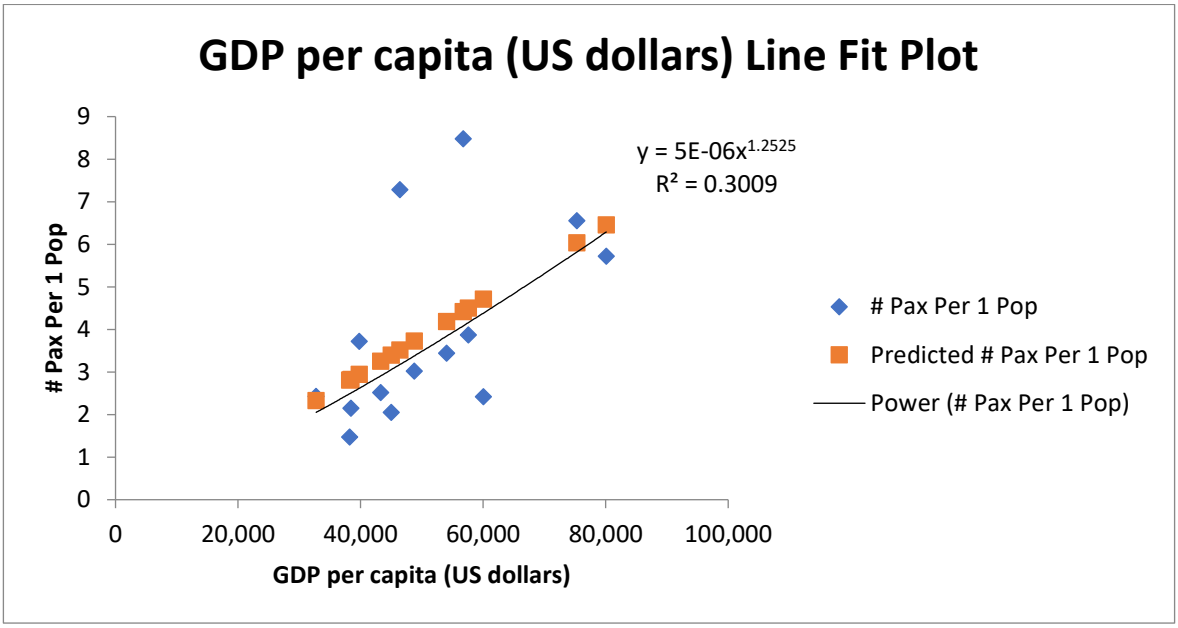

Figure 4.7. *GDP Per Capita in $ Univariate Regression Line Fit*

Figure 4.7 displays GDP per capita, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between GDP per capita and the number of passengers per population could be due to an assumed relationship between the development of products and services in a country and how desirable that country might be to different entities from a business trying to start and grow its trade to an individual that may not have the same opportunities back home; we see an example of the latter in people that travel to different countries and come home with Apple products for example as they may be cheaper or more available there. GDP per capita when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.30 that infers a fair explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 5E\text{-}06x^{1.2525}$, and holds an R^2 value of 0.30

that shows the highest R^2 value available of the five trendline types. The power trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable GDP per capita.

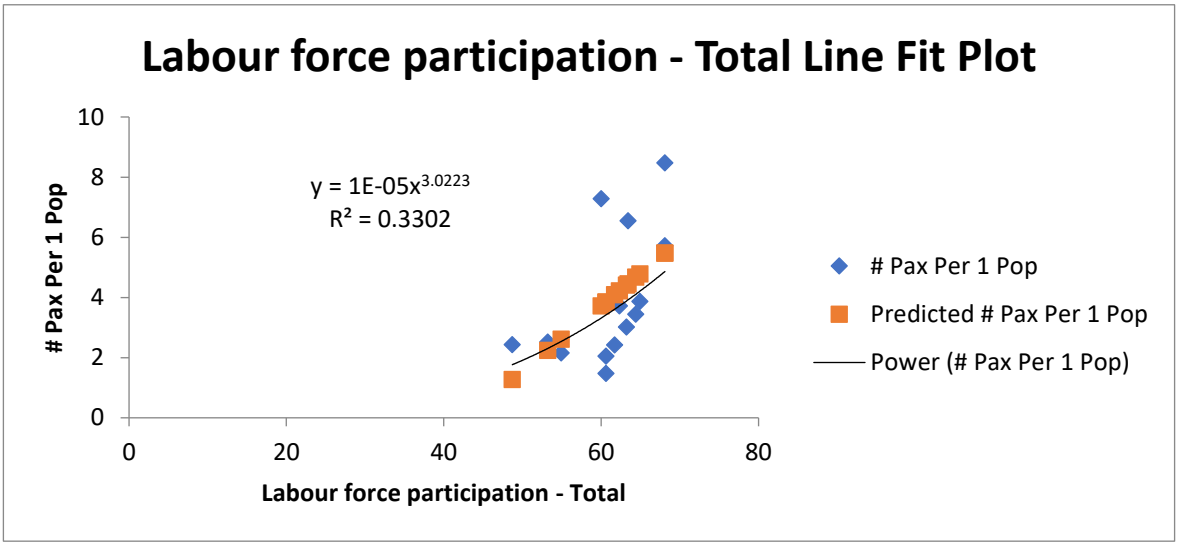

Figure 4.8. Labour Force Participation Total Univariate Regression Line Fit

Figure 4.8 displays Labour force participation, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between Labour force participation and the number of passengers per population could be due to an assumed relationship between the Labour force participation and the overall condition of the economy of a country; from there, we can make an assumption that the better the condition of the economy within a country the more desirable it becomes for entities from tourists to businesses. Labour force participation when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.29 that infers a fair explanation of variance between the actual and predicted

number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 1E\text{-}05x^{3.0223}$, and holds an R^2 value of 0.33 that shows the highest R^2 value available of the five trendline types. The power trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Labour force participation.

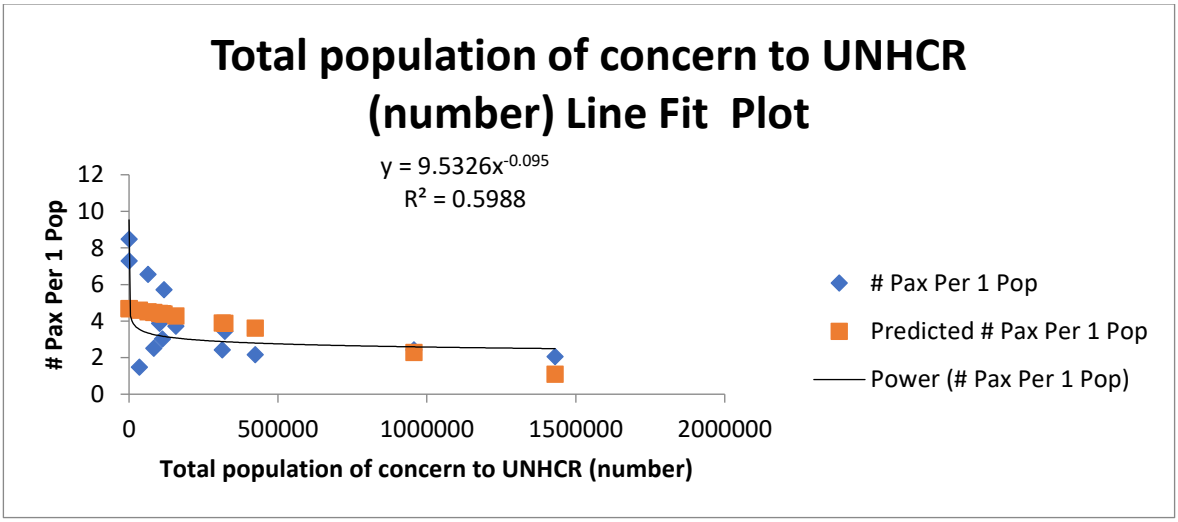

Figure 4.9. Total Population of Concern to UNHCR Univariate Regression Line Fit

Figure 4.9 displays the Total population of concern to UNHCR, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between the Total population of concern to UNHCR and the number of passengers per population could be due to the Total population of concern to UNHCR being an indicator of how well a country is doing both socially and economically; we can say that the population of concern to UNHCR includes people like refugees and that the more refugees that a country can hold would be indicative of the excess of wealth and welfare services that a country has that could make it attractive to a variety of

people. Total population of concern to UNHCR when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.22 that infers a weak to a fair explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 9.5326x^{-0.095}$, and holds an R^2 value of 0.60 that shows the highest R^2 value available of the five trendline types. The power trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Total population of concern to UNHCR.

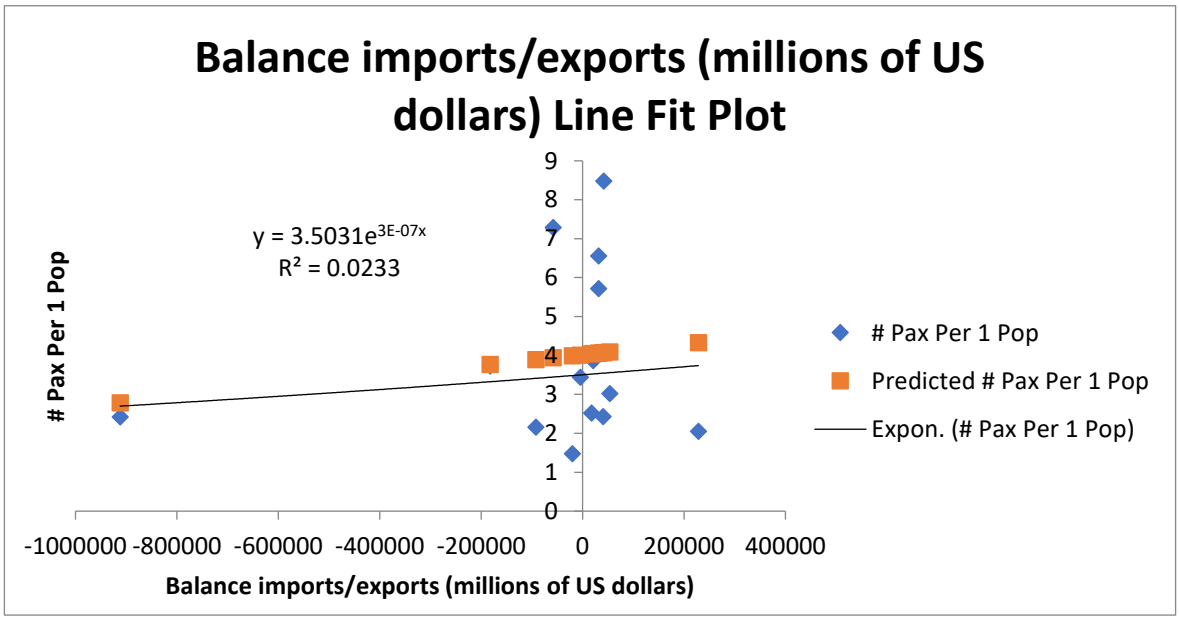

Figure 4.10. Balance Imports/Exports in $ Univariate Regression Line Fit

Figure 4.10 displays Balance imports/exports, with a line of fit based on exponential, in a scatterplot graph as the only variable against the number of passengers per population; power was not used here as power cannot be used when negative values occur. Hypothetically, the

relationship between Balance imports/exports and the number of passengers per population could be due to the Balance of imports and exports projecting the overall condition of the economy of the country as we have previously discussed the overall condition of the economy of a country could be used as an indicator of desirable a country is from an air travel perspective. Balance imports/exports when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.02 that infers a very weak explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 3.5031e^{3E-07x}$, and holds an R^2 value of 0.02 that shows the highest R^2 value available of the five trendline types. The exponential trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, the power trendline could not be used as some of the numbers were negative, and some were positive.

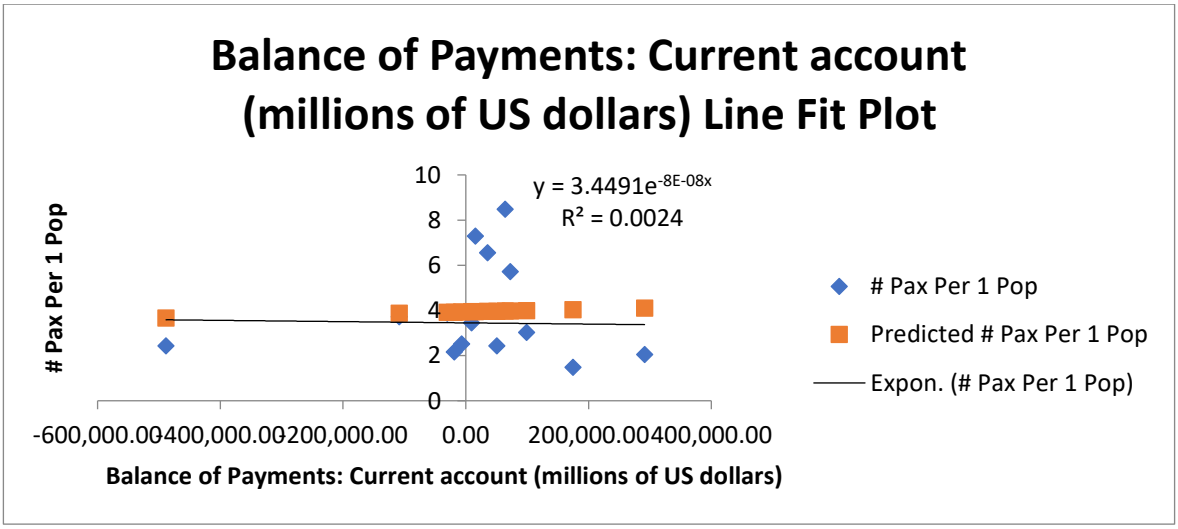

Figure 4.11. Balance of Payments: Current Account in $ Univariate Regression Line Fit

Figure 4.11 displays Balance of payments: Current account, with a line of fit based on exponential, in a scatterplot graph as the only variable against the number of passengers per population; power was not used here as power cannot be used when negative values occur. Hypothetically, the relationship between Balance of payments: Current account and number of passengers per population could be due to similar reasons as Balance imports/exports, as this figure is included in a country's current account with the addition of foreign incoming and outgoing, the current account figure of a country may provide some information on the economic health of the country. Balance of payments: Current account, when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.00 that infers almost zero explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 3.4491e^{-8E-08x}$, and holds an R^2 value of 0.00 that shows the highest R^2 value available of the five trendline types. The exponential trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, the power trendline could not be used as some of the numbers were negative, and some were positive.

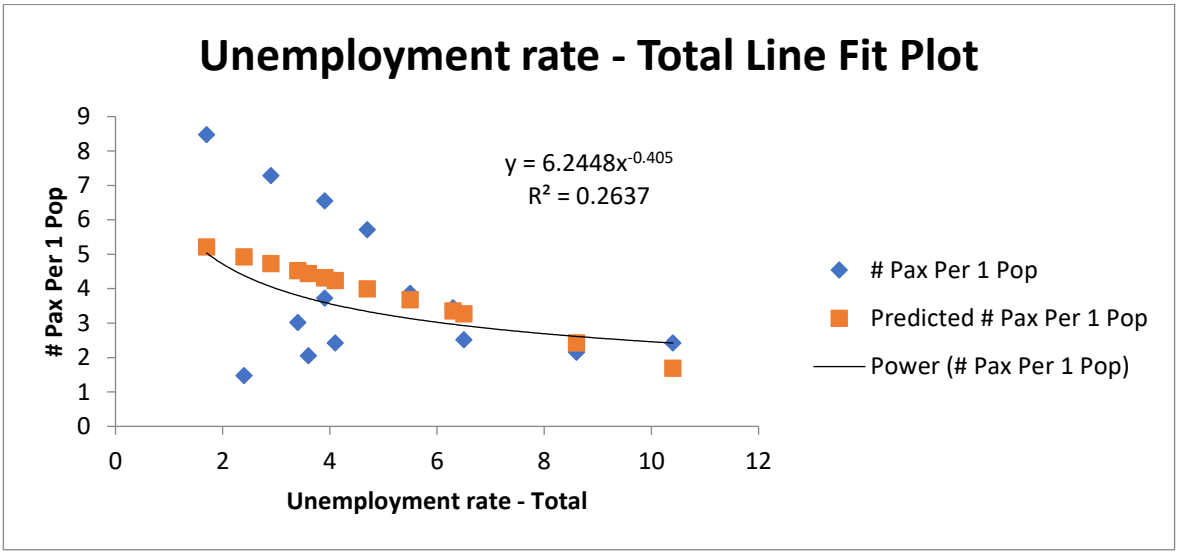

Figure 4.12. Unemployment Rate Total Univariate Regression Line Fit

Figure 4.12 displays Unemployment rate – Total, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between Unemployment rate – Total and number of passengers per population could be due to the total unemployment rate of a country being an indicator of how well a country is doing both socially and economically; we can make the assumption that the lower the unemployment rate, the more desirable a country might be to visit or grow a business in as the infrastructure required to keep your population employed likely requires a good amount of development. Unemployment rate – Total when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.20 that infers a fair explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 6.2448x^{-0.405}$, and holds an R^2 value of 0.26 that shows the highest R^2 value available of the five trendline types. The power trendline is used here for

two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Unemployment rate – Total.

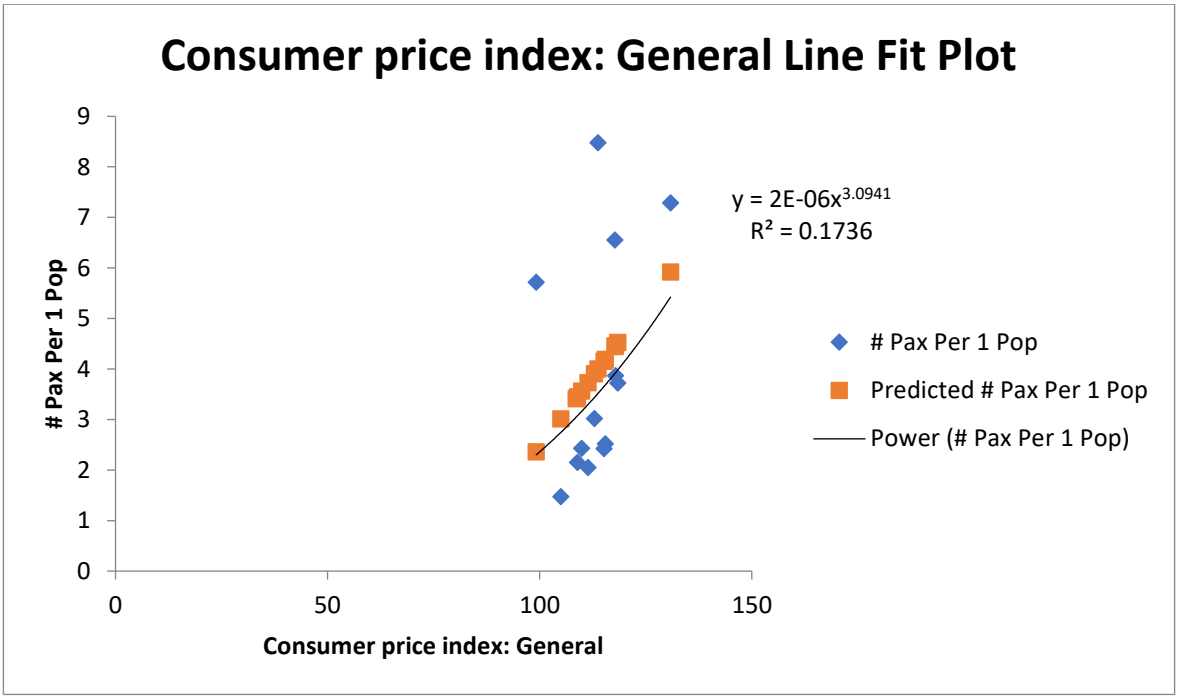

Figure 4.13. Consumer Price Index Univariate Regression Line Fit

Figure 4.13 displays the Consumer price index: General, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between Consumer price index: General and number of passengers per population could be due to Consumer price index: General being an indicator of how expensive or inexpensive a country might be to live in the day-to-day that can definitely be a factor when a tourist is considering where to holiday, assuming that the majority of the population of the majority of countries are in the lower-to-middle class, certain countries that are more pricey may well be more prohibitive to frequent for normally

wealthy people. Consumer price index: General when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.14 that infers a weak to a fair explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 2E\text{-}06x^{3.0941}$, and holds an R^2 value of 0.17 that shows the highest R^2 value available of the five trendline types. The power trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Consumer price index: General.

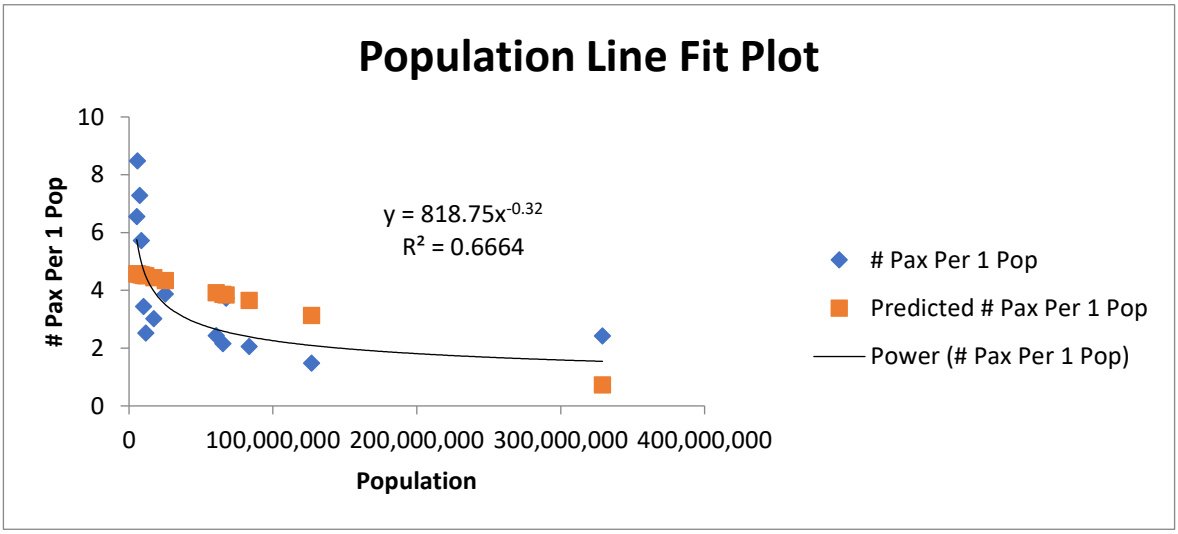

Figure 4.14. Population Univariate Regression Line Fit

Figure 4.14 displays the Population of each country, with a line of fit based on power, in a scatterplot graph as the only variable against the number of passengers per population. Hypothetically, the relationship between the population of each country and the number of passengers per population could be due to the population of each country requiring transport,

both internally and externally, and simply put, the more people you have to move, the more passengers you will have. The population of each country when a linear regression line is used, which are the orange points on the scatterplot, holds an R^2 value of 0.22 that infers a fair explanation of variance between the actual and predicted number of passengers per population. The curve that is visible on the scatterplot is the power trendline, which uses the equation $y = 818.75x^{-0.32}$, and holds an R^2 value of 0.66 that shows the highest R^2 value available of the five trendline types. The power trendline is used here for two reasons. First, to provide a better visual representation of the variation in the actual and predicted number of passengers per population. Second, to provide the greatest possible R^2 value for the variable Population of each country.

Variables Combined in Models

After testing each of the variables individually against the passenger to population ratio of the given country, it was time to tackle research question two and see if any predictive models of use could be created. After extensive testing with MLR, filter-based feature selection, and five-fold cross-validation, the best of the prediction models was presented with a grouping of independent variables consisting of nine variables. There was a second contender for best prediction model and that was the ten variable model which had an R^2 score of 0.61. This score is respectable but considerably lower than the best predictive models R^2 score of 0.75. Table 4.2 provides a list of the variables and the respective identifier assigned during data analysis. The nine variables that completed the best prediction models are: Population ($\beta = 0.00$, $p = 0.001$), Patents in Force ($\beta = 0.00$, $p = 0.02$), Gross Domestic Expenditure on R & D: as a % of GDP ($\beta = -0.84$, $p = <0.001$),

Tourist/Visitor Arrivals (β = 0, p = <0.001), Tourism Expenditure (β = 0.00, p = <0.001), Public expenditure on education: as a % of Government expenditure (β = 0.57, p = <0.001), Public expenditure on education: as a % of GDP (β = -0.33, p = <0.001) , GDP per capita (β = 0.00, p = <0.001), and Labour Force Participation (β = -0.04, p = <0.02). This nine-variable model scored very highly in correlation with the actual dataset, with an R^2 value of 0.99. Many models scored very highly on the R^2 scores, which raised some concerns that will be identified in the discussion section of this study, around overfitting and underfitting the various models. Table 4.3 provides the best models from each number of variables and shows that R^2 scores are above .90 from the three-variable model upwards. As many models scored substantially high R^2 scores, the introduction of five-fold cross-validation was selected in order to make sure that we were not just trying to fit one line to another. This resulted in being able to go from simply identifying relationships, and answering our first research question, to the possibility of identifying a useful predictive model, thus answering the second research question. After the five-fold cross-validation was enacted, the nine-variable model scored the highest of all the possible models with a mean cross-validated R^2 value of 0.75. This score is substantial but would likely be higher pending a few changes in the study. Thus, this study and its results could be improved in the future, and those improvements will be identified in chapter 5.

Table 4.2

Variable Identification Table

Variable #	Variable Name
0	Population
1	Consumer Price Index: General
2	Patents in force (number)
3	Gross domestic expenditure on R & D: as a percentage of GDP (%)
4	Unemployment rate - Total
5	Tourist/visitor arrivals (thousands)
6	Tourism expenditure (millions of US dollars)
7	Balance of Payments: Current account (millions of US dollars)
8	Balance imports/exports (millions of US dollars)
9	Public expenditure on education (% of government expenditure)
10	Public expenditure on education (% of GDP)
11	Total population of concern to UNHCR (number)
12	GDP per capita (US dollars)
13	Labour force participation - Total

Table 4.3

Best Model for Each Number of Variables Before Cross-Validation

# of Variables	Variable #/s	R^2 score
1	9	0.799
2	3,9	0.851
3	1,3,9	0.923
4	1,3,7,9	0.944
5	1,3,7,9,12	0.959
6	3,5,6,9,10,12	0.980
7	0,3,5,6,9,10,12	0.990
8	0,2,3,5,6,9,10,12	0.997
9	0,2,3,5,6,9,10,12,13	0.999
10	0,1,2,3,5,6,7,9,10,12	0.999
11	0,1,2,3,5,6,7,9,10,11,12	0.999
12	0,1,3,4,5,6,8,9,10,11,12,13	0.999

Decisions on Hypotheses

Research Question 1 asked, "What is the relationship between various socio-economic variables and the passenger-to-population ratio of a country?". Based on the results of the statistical analyses, the null hypothesis (H_{01}) states that there will not be a significant relationship between socio-economic variables, and the passenger-to-population ratio of a country can be rejected, and the alternative hypothesis is accepted. The alternative hypothesis (H_{A1}) states that there will be a significant relationship between socio-economic variables and the passenger-to-population ratio of a country.

The second null hypothesis (H_{02}) states that there will not be a significant relationship between Labour Force Participation and the passenger-to-population ratio of a country, this can be rejected, and the alternative hypothesis accepted because the study showed that Labour Force Participation and the passenger-to-population ratio of a country have a significant relationship ($p = 0.04$). The alternative hypothesis (H_{A2}) states that there will be a significant relationship between Labour Force Participation and the passenger-to-population ratio of a country.

The third null hypothesis (H_{03}) states that there will not be a significant relationship between GDP per capita and the passenger-to-population ratio of a country, this can be rejected, and the alternative hypothesis is accepted because the study showed that GDP per capita and the passenger-to-population ratio of a country have a significant relationship ($p = 0.04$). The alternative hypothesis (H_{A3}) states that there will be a significant relationship between GDP per capita and the passenger-to-population ratio of a country.

The fourth null hypothesis (H_{04}) states that there will not be a significant relationship between Public Expenditure on Education as a % of Government Expenditure and the passenger-to-population ratio of a country, this can be rejected, and the alternative hypothesis is accepted because the study showed that Public Expenditure on Education as a % of Government Expenditure and the passenger-to-population ratio of a country have a significant relationship ($p = 0.00001$). The alternative hypothesis (H_{A4}) states that there will be a significant relationship between Public Expenditure on Education as a % of Government Expenditure and the passenger-to-population ratio of a country.

Research Question 2 asked, "Can a predictive model using various socio-economic variables predict the passenger-to-population ratio of a country?". Based on the results of the statistical analyses, the null hypothesis (H_{05}) states that the socio-economic variables of a country will not create a predictive model of the passenger-to-population ratio of a country and can be rejected, and the alternative hypothesis is accepted. The alternative hypothesis (H_{A5}) states that the socio-economic variables of a country will create a predictive model of the passenger-to-population ratio of a country.

Summary of Hypothesis Testing

The purpose of this study was to observe and identify any relationships that exist, either individually or as groups, between select socio-economic variables and the passenger-to-population ratio of a country. Overall, two research questions and five hypotheses were tested, and all five of the hypotheses were rejected successfully. The research rejected null hypothesis one and thus suggests that there is a relationship between socio-economic variables and the passenger-to-population ratio of a country. The research

rejected null hypothesis two and thus suggests that there is a relationship between Labour Force Participation and the passenger-to-population ratio of a country. The research rejected null hypothesis three and thus suggests that there is a relationship between GDP per capita and the passenger-to-population ratio of a country. The research rejected null hypothesis four and thus suggests that there is a relationship between Public Expenditure on Education as a % of Government Expenditure and the passenger-to-population ratio of a country. The research rejected null hypothesis five as the results of the study show that the best model, even with the inclusion of cross-fold validation, shows a solid R^2 score of 0.75 that suggests a predictive model of good accuracy could be developed.

Summary

The analyses conducted during this study provide support to the existence of a relationship between socio-economic variables and the passenger-to-population of a developed country. Naturally, there was variation in each socio-economic variable when it came to their relationship with the passenger-to-population ratio. Individually, the independent variable Public Expenditure on Education as a % of Government Expenditure had the most supporting evidence when it came to a relationship with the dependent variable passenger-to-population ratio. Using the independent variables to build predictive models yielded great success; actually, it was too successful, and the previously discussed fear of overfitting became apparent, which is why five-fold cross-validation with feature selection was used to display all of the models' capabilities for prediction rather than just fitting a line. Overall, the results from the analyses provide us with the insight that the strongest variables individually also appeared in the best models consistently and that

those variables tended to be more financial-centric. This suggests that in this study, we can suggest that economic variables might hold more say when it comes to the passenger-to-population ratio of a developed country, which is supported by plenty of previous studies as captured in the literature review of this study.

Chapter 5

Conclusion

Research Summary

This study was designed to measure the relationships between existing data, from the United Nations (2019) database, from both social and economic areas to find interesting patterns with the passenger to population ratio and see if there could be a predictive model that could be useful. The literature review conducted beforehand highlighted specific economic variables that would likely have major impact on the DV. These specific variables were commonly used in other predictive models for passenger count prediction between airport pairs and on a national level. Variables, such as GDP, were commonly used in an algorithm for predicting passenger count and thus would likely yield some valuable results when also applied to the passenger to population ratio.

The study proposed two RQs to be answered by the study results:

1. What is the relationship between various socio-economic variables and the passenger-to-population ratio of a country?
2. Can a predictive model using various socio-economic variables predict the passenger-to-population ratio of a country?

Alongside two RQs, there were five hypotheses to test; the hypotheses are as follows:

- There will not be a significant relationship between socio-economic variables and the passenger to population ratio of a country.
- The socio-economic variables will not create a predictive model of the passenger to population ratio of a country.

- There will not be a significant relationship between Labour Force Participation and the passenger to population ratio of a country.
- There will not be a significant relationship between GDP per capita and a country's passenger to population ratio.
- There will not be a significant relationship between Public Expenditure on Education as a % of Government Expenditure and the passenger to population ratio of a country.

There was no specific instrument for data collection purposes because the data owner had already collected the data. This research study used Microsoft Excel for the statistical analyses along with the extension Data Toolpak to allow for the appropriate statistical analyses.

The study's target population consisted of highly developed countries as defined by the United Nations. The sample used convenience sampling to gather data about each variable from sample countries held by the United Nations. The study used a correlational research design, as is appropriate, as there is only one group and analysis between that one group and multiple other IVs is the primary goal of this study. Quantitative data were analyzed using two methods: linear regression and multiple regression. The analysis used linear regression, where the analysis consisted of one IV and the DV. The analysis used multiple regression, where the analysis consisted of multiple IVs to form a model to predict the DV. Regression analyses yielded R^2 values that provided a baseline value for the quality of each variable; more specifically, providing some evidence as to whether the IV/s could have a relationship with the DV.

After R^2 values were collected from each IV, further analysis had to be completed to ensure that overfitting was not obscuring the quality of the IV(s). Thus, this research study employed five-fold cross-validation to remove the potential obscurity of overfitting. The study results were statistically significant, with some individual variables and all variations of IVs combinations scoring high R^2 values and including five-fold cross-validation testing.

Discussion and Interpretation of Findings

General Discussion

The extreme multivariate substrate of each country makes it very difficult to solidly claim any causal relationship between the IVs and DVs. From individual social and economic policies in a country to the geographical appropriateness of aviation for each country, there are infinite possibilities in the microvariability at a country level. For example, even though the sample countries are within the same highly developed bracket, the countries might have arrived to that stage of development through different methods. Each country's growth may have also happened for other reasons and through varying different cultures. If you look at Hong Kong as an example, Hong Kong has benefitted from being under the reign of both China and the United Kingdom. These two countries are likely to have two very different approaches to development as one had a communist government and the other a constitutional monarchy. The study results provide evidence towards the idea that the selected socio-economic variables do have a relationship with the passenger-to-population ratio of a country. Some variables offer very little evidence of a relationship between them and the passenger-to-population ratio of a nation.

Conversely, some variables, even as individual IVs, provide good evidence of a relationship with a country's passenger-to-population ratio. The best IV that provided evidence of a relationship with the DV is Public Expenditure on Education as a % of Government Expenditure that holds an R^2 value of 0.80 and a p-value of $p = <0.00001$. This result suggests that the analysis of the variable Public Expenditure on Education as a % of Government Expenditure provides robust evidence as to the possibility of a relationship between one of our IVs and the DV, which alone tells us only that as this variable changes so does the DV. The worst IV that didn't prove a relationship with the DV is Balance of Payments: Current Account, with an R^2 value of 0.002 and a p value of $p = 0.88$. This result suggests that the analysis of the variable Balance of Payments: Current Account provides extremely weak evidence as to the possibility of a relationship between this IV and the DV.

Concerning the first research question, "What is the relationship between various socio-economic variables and the passenger-to-population ratio of a country?" we can conclude that there is most certainly some relationship between some of the socio-economic variables and the DV. As previously stated, the best individual IV has an R^2 value of 0.80 that alone provides strong evidence of a relationship between the IV and DV. The best variable's p-value was $p = <0.00001$, which is exceptionally high by any measure of strength and satisfies the requirements for this study. The worst variable scored an R^2 value of 0.002 and a p-value of $p = 0.88$, which provides almost zero evidence that there is any relationship between this IV and the DV. Low scores don't necessarily mean that the worst IV will not be helpful when we consider multiple variables to create a predictive

model. The best variable's results could be sufficient by themselves because the best variable's data points almost match the DV's data points in terms of their change over time. For example, when the DV increases, so does the best IV. Simply fitting an existing pattern is why adding the five-fold filter-based cross-validation was necessary to void any results that occurred due to overfitting.

When discussing the second research question, which asks whether this study can create a useful predictive model, this research concludes that a predictive model can be successfully produced from the IVs. As we already have one variable with an R^2 score of 0.80, this variable alone should make a good prediction of the DV. Introducing other IVs certainly should increase the strength of the created model's ability to predict the DV, and in practice, that is exactly what happened. In fact, the predictive models were very successful and had an R^2 score of 0.99 from seven combined IVs upwards. From nine combined IVs upwards, the predictive models had R^2 scores of 0.999. These very high R^2 numbers provide very strong predictive power, and the next step would be to test these models against larger or different development stage datasets. The R^2 numbers are very high when combining multiple IVs, which is likely because the dataset is relatively small. Thus, the chances of enough combinations of IVs simply fitting the DV dataset are high. Again, this is why the inclusion of five-fold filter-based cross-validation was introduced to increase the accuracy of the predictive models and avoid the concept of overfitting.

Individual Variables

This study used linear regression to analyze each of the IVs to try and establish some evidence for the presence of a relationship between that IV and the DV. When each IV was analyzed, there was good variance in evidence provided about the strength of each variable and its relationship with the DV. Provided below is table 5.1 with results from the analysis of each of the IVs.

Table 5.1

Individual IV Linear Regression Results

Variable Name	R^2 Score	*p* Value
Tourist/Visitor Arrivals	(R^2 = 0.22)	(p = <0.09)
Tourism Expenditure	(R^2 = 0.11)	(p = <0.24)
Public expenditure on education: as a % of Government expenditure	(R^2 = 0.80)	(p = <0.00001)
Public expenditure on education: as a % of GDP	(R^2 = 0.04)	(p = 0.52)
GDP per capita	(R^2 = 0.31)	(p = 0.04)
Labour Force Participation	(R^2 = 0.29)	(p = 0.04)
Total Population of Concern to UNHCR	(R^2 = 0.22)	(p = <0.09)
Balance Imports/Exports	(R^2= 0.03)	(p = 0.58)
Balance of Payments: Current Account	(R^2= 0.00)	(p = 0.88)
Unemployment Rate	(R^2 = 0.20)	(p = 0.11)
Consumer Price Index	(R^2 = 0.14)	(p = 0.18)
Population	(R^2 = 0.22)	(p = <0.09)
Patents In Force	(R^2 = 0.23)	(p = <0.09)
Gross Domestic Expenditure on R & D as a % of GDP	(R^2 = 0.23)	(p = <0.28)

Among the individual testing of each IV came some unexpected results, with some variables that the literature review indirectly suggested having a relationship with passenger count not being as significant in these analyses. The population variable is likely to have some relationship with the DV because the DV is the population modified by the number of passengers in a country. The fact that population has an almost significant p-value supports the argument that the variance in how many passengers per population member is not very high between the selected highly developed countries. This trend could be tested by further research with other countries at different stages of development. Intuitively, we could assume that with the variance in the DV growing as the divergence in development does, then population as a variable may be less significant.

Some of the more logically correlated variables also hold relatively strong when analyzed against the DV. Labor force participation could be logically correlated with the DV as the more people working, the more people who may have to fly somewhere for work. Likely, there is a strong correlation between working and the ability to fly for leisure or business. This may be highlighted in the results of this study by the GDP per capita variable. GDP per capita is a well-known and commonly used for estimating economic well-being that tells us how much productive value each citizen brings to the country. Assuming that in some way, the benefits of providing more value to a country do trickle down to each citizen. That trickle-down benefit could be an excess of capital for said citizen that could be used for national or international travel by aircraft. GDP per capita could also have a good relationship with the DV because the more economic value a country has, the better the infrastructure will be. Services such as aviation travel, which we

could consider a first-world luxury or convenience, probably develop faster when the economic value is more significant.

Tourist and visitor arrivals seem like a solid choice for a variable that logically may be related to the DV. In the early stages of this research, the author was almost assured that analysis of the tourism-related variables would provide strong evidence of a relationship between them and the DV. As a standalone variable, tourist and visitor arrivals prove statistically insignificant, with a p-value of 0.09 and an R^2 score of 0.22. Still, they are significant in the best cross-validated predictive models with a very low p-value.

The CPI variable provides weak evidence of relationships with the DV with a p-value of 0.18 and an R^2 score of 0.14. This may well be because the consumer price index is influenced by many other factors than just the development of a country. Some countries may have a higher consumer price index because they are island nations, and having to import certain goods causes the prices of those goods to increase. Some countries may be geographically located in places that may not allow certain types of goods to grow, and this may cause goods to go up in cost. Ultimately, as with all other IVs, there are so many extraneous factors at play that it would be challenging to link an IV alone causally.

Public expenditure on education: as a % of Government expenditure is the standout variable of this research study with an outstanding R^2 score of 0.80 and a *p* value of <0.00001. The results of this aspect of the study provide strong evidence of a relationship between the DV and this outstanding IV. Though the results suggest that the relationship is correlational, it should be encouraged to check for a similar relationship with the DV with countries at other stages of development. The best IV is used when a predictive model is

created and likely acts as the equation's backbone for predicting the DV. The reasoning behind why this variable has such a strong relationship with the DV is unknown. However, one explanation could be that a country that spends more on education should have a higher level of education across the population. With a higher level of education across the population comes greater technology and faster technological advances. Aviation being a part of the specialized travel sector is likely boosted by an educated population from production, design, and service standpoints. It could also be that a developed country puts more focus on higher education levels, and only an excess of funds would allow a government to focus on the higher education levels. When discussing sectors of industry that a country primarily uses, a country does not necessarily need a highly educated population to work agriculture jobs for we might say that if you are majority primary industry, you. However, suppose the country you are in is providing a lot of services from the tertiary industry. In that case, you may need education in a country to be higher to create employees for that level of work. This all lends itself to possible reasons why education expenditure as a percentage of government expenditure strongly relates to the DV, yet it could just be a coincidence.

Variables Combined into a Model

RQ2 refers to whether the IVs can be combined to form a predictive model. As previously discussed in Chapter 4, the predictive models came out with great success by the numbers. We must partly attribute this to the relatively small sample that makes the models look better than they probably are when applied to a larger dataset. Table 5.2 displays the variables and p-values that went into combining the best predictive model

created by some combination of the IVs. This study ended up with a model containing nine variables as the most accurate predictive model. After five-fold cross-validation, the nine variable models had an R^2 score of 0.75, representing a reasonably good predictive strength for the DV. Future research studies could use this predictive model to see if the predictions are as strong for countries at different stages of development. The assumption would be that because the predictive model has been created to predict countries at higher stages of development, the model would likely not be as successful at predicting the DV across a range of countries at different stages of development.

Table 5.2

Best Predictive Model Contained Variables

Variable Name	*p* Value
Patents in force (number)	0.02
Gross domestic expenditure on R & D: as a percentage of GDP (%)	0.00
Tourist/visitor arrivals (thousands)	0.00
Tourism expenditure (millions of US dollars)	0.00
Public expenditure on education (% of government expenditure)	0.00
Public expenditure on education (% of GDP)	0.00
GDP per capita (US dollars)	0.00
Labor force participation - Total	0.02
Population	0.00

As displayed in Table 5.2, the *p* values from the variables of the best predictive model are all considered significant at less than 0.05. As expected, the best individual variable features in the best predictive model have a p-value of 0.000003090 compared to 0.0000159348 as an individual IV. GDP per capita was a significant variable in the current study which was in alignment with the literature review suggesting that it would also be

significant. GDP per capita was used or discussed in every reviewed study related to passenger number prediction in the air and on the ground.

As previously discussed, it would be easy to create a predictive model that replicates the DV's slope. Further measures needed to ensure that the model wasn't simply overfitting and thus unable to predict new data in the future. Five-fold filter-based cross-validation was used to avoid any overfitting complications. This process splits the dataset into groups, five groups hence five-fold, and tests the model against each data set.

Significance of Study

The significance of the study is almost yet to be determined, as the usability of the predictive model will have to be tested against more extensive and more varied datasets to be generally applicable on a worldwide scale. In the recommendations for further research section, multiple ideas for other predictive model uses arose from this research study.

The study results suggest that the predictive model with an R^2 score of 0.75 has strong predictive power of the countries considered highly developed by the United Nations. This model could now predict future passengers per population figures for each country and allow the numerous entities whose business revolves around the aviation industry to make simple estimations of their business demands.

The results of this study on an individual IV level are also significant when considering that the best individual IV could well be a good indicator of the DV. This IV could be investigated by itself to see if it retains its relationship with the DV at countries at other stages of development and to see why it has the relationship that it does.

Recommendations for Future Research

The population count of countries generally trends up, assuming the country's population rate is positive. Simply put, there are more births than deaths over a given period, excluding periods inclusive of historical effects like Covid-19. The passenger count of countries shows a similar trend that follows logical assumption: the more people there are, the more that will need transporting. This section provides a discussion of some ideas that could improve upon this study going forward.

Based on the findings of this study, it would be exciting to see if repeated research with a more extensive data set would yield the same or similar results. Future research could compile a larger dataset by utilizing other countries and not just using the countries considered highly developed by the United Nations. Whether to study the individual variables or look at a combination of variables, which could create a predictive model, both paths are likely worth investigating. This further research could reveal new relationships that this study did not disclose, perhaps with the inclusion of countries at other stages of development. The highest-scoring individual variable tested in this study was public expenditure on education as a percentage of government expenditure that scored $R^2 = 0.80$, $p = <0.00001$, and featured in the best predictive model.

Further research could test this variable against other countries at different stages of development and see if there is indeed a pattern with this variable alone. Theoretical support for this future study idea comes from the well-known Maslow's Hierarchy of Needs (1943). This theory states that people need different things depending on their state of motivation and at the bottom are physiological needs and atop the hierarchy is self-

actualization, which could indicate some differences between countries at different levels of development. If that were indeed the case, it would be insightful to delve deeper into why this variable is so powerful when looking at a country's passenger to population ratio, potentially encouraging governments to spend more of their money on education. The logic of the better the education of a population, the more flights per population holds some ground intrinsically but seeing this relationship amongst other countries at different stages of development would provide additional evidence on a correlational level.

A second future research idea would be to complete a similar function on a more enormous data set by using more years of data rather than differing the countries by development status according to the UN. By expanding the data set by years, the data set becomes more extensive, making any results that arise from the future research more valid as the size of the data set is the most significant limitation of this research. The expansion of the data set by years is not as simple as it seems, however, the growth of the dataset will allow more potentially adverse effects to occur. Historical effects occur at a certain point that affects everything in a particular geographical region or the entire world. One example is passenger numbers have dropped since the pre to mid-Covid-19 pandemic. Increasing the date range of the dataset would give us a better idea of how strong the relationships that have appeared in this research study would be but would also be susceptible to further errors that any future researcher should be aware of.

The third future research recommendation would be to combine the previously mentioned future research recommendations. Growing the data set by expanding the year range of the data that has been used and by using a greater variety of countries would be

the most robust test for the results of this research. There are going to be further limitations with the expansion of the dataset. This study only used a dataset from a couple of years because the data was not consistently available at the UN source for all of the countries. Finding the missing data, or perhaps using some statistically appropriate method of extrapolation, to fill in the blanks available at the data source would have to be considered to ensure a clean and usable dataset. This dataset would be by far the largest of all research recommendations and be the most susceptible to errors of all kinds due to its scale. Once achieved, the research results would hold the most significant strength as the dataset is likely the most representative dataset available at this given moment.

The fourth recommendation for future research consists of conducting a similar study as mentioned in the third recommendation but on each category of developed or undeveloped countries as defined by the UN. This study only used highly developed countries, as defined by the UN. With that small dataset, the research results become difficult to argue for in terms of representativeness and expandability to other countries in different stages of development. The way to solve this would be to run a similar study for each category as individual categories. The results of this study would help narrow down the apparent relationships found in this research. For example, if we found that a variable that had been seen to have strong evidence from this study also had a similar effect with developing countries, the variable's potential strength becomes greater. On the other hand, if there is one variable with solid evidence in developing countries but not in highly developed countries, then an investigation into why that variable works differently across development stages should begin.

The fifth recommendation for future research would be to conduct a research study on every combination of the countries within each development stage. For example, highly developed, developing, and least developed countries would be the three groups, and each variation of these three groups could have a similar research study as this research conducted on them. This would allow a more detailed investigation into which variables most likely affect a country's passenger to population ratio within the representative development stage. Per the first future research recommendation, where the three groups were analyzed as one group, this could yield useful and representative results for the entire population. However, analyzing each possible combination of the three groups could deliver more specific results as it may provide more insight into the relationships between the developmental stages. If the least developed and developing countries' analysis suggests that economic-related variables primarily control the passenger to population ratio, the results could infer those economic variables are more important at earlier stages of development.

The sixth future research recommendation could be to complete a similar research study but with a different set of independent variables. The United Nations, and other similar organizations, likely have many different datasets encompassing every aspect of life that could be analyzed to see if any other relationships were to be discovered. This would also enable the set of countries the dependent variable is based on to be larger or more varied and, as a result, add external validity.

Conclusion

The purpose of this study was twofold, to investigate any potential relationships between socio-economic variables and passenger to population ratio, and to see if the investigated variables could combine to form a predictive model of passenger to population ratio. The findings suggest that there are strong relationships between a number of the selected socio-economic variables at least at a correlational level. In terms of creation of a predictive model, the findings show that a predictive model with an R^2 score of 0.75 was created out of nine of the IVs. The inclusion of feature selection and five-fold cross-validation was used to ensure that the model was not overfitting the DV. The use of the knowledge gained in the analyses performed in this study should prove useful when expansion of the data set occurs for any future research, whether by increasing the date range of the data or including additional subsets of countries.

References

Agras, J. & Chapman, D. (1999). A dynamic approach to the Environmental Kuznets Curve hypothesis. *Ecological Economics*, 28 (2): 267-277.

Akgüngör, A. P., & Doğan, E. (2009). An artificial intelligent approach to traffic accident estimation: model development and application. *Transport,* 24(2):135-142. doi: 10.3846/1648-4142.2009.24.135-142

Azam, M. (2016). "Does environmental degradation shackle economic growth? A Panel Data Investigation on 11 Asian Countries", Renewable and Sustainable Energy Reviews, 65, p. 175-182.

Bartlett, H.C. (1965). *The demand for passenger air transportation*, 1947-1962. [Doctoral Dissertation, The University of Michigan.] https://books.google.com/books?id=aOvStwEACAAJ

Boyle, G. E., & McCarthy, T. G. (1999) Simple measures of convergence in per capita GDP: a note on some further international evidence, Applied Economics Letters, 6:6, 343-347, DOI: 10.1080/135048599353041

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. European Journal of Operational Research, 184(3), 1140-1154.

Doganis R (2009). Flying off course: airline economics and marketing. 4th edition. Abingdon: Routledge.

Güvercin, D. (2019). The impact of GDP per capita and crime rate on the CO^2 Emission. *Selcuk University Social Sciences Institute Journal*, 25–33.

Grossman, G. M., & Krueger, A. B. (1993). Environmental impacts of a North American Free Trade Agreement, Garber P.(éd.), The US-Mexico Free Trade Agreement, MIT Press, Cambridge, MA, 1655177

Grossman, G. M., & Krueger, A. B. (1995). "Economic growth and the environment", *The Quarterly Journal of Economics*, 110 (2): 353-377

Harvey, D. (1951). "Airline passenger traffic pattern within the United States". *Journal of Air Law & Commerce*. 18, 157 https://scholar.smu.edu/jalc/vol18/iss2/2

Holtz-Eakin, D., & Selden, T. M. (1995). "Stoking the fires? CO2 emissions and economic growth", *Journal of Public Economics*, 57 (1): 85-101.

International Air Transport Association. (2019). World Air Transport Statistics 2020. Retrieved November 2, 2020 from https://www.iata.org/contentassets/a686ff624550453e8bf0c9b3f7f0ab26/wats-2020-mediakit.pdf

Jacobson, M. Z. (2008). "On the causal link between carbon dioxide and air pollution mortality", Geophysical Research Letters, 35 (3): 1-45

Karlaftis, M. G., Zografos, K. G., Papastavrou, J. D., & Charnes, J. M. (1996). Methodological framework for air-travel demand forecasting. *Journal of Transportation Engineering*, 122(2),96. https://doi-org.portal.lib.fit.edu/10.1061/(ASCE)0733-947X(1996)122:2(96)

Laik, M. N., Choy, M., & Sen, P. (2014). "Predicting airline passenger load: A case study," 2014 IEEE 16th Conference on Business Informatics, Geneva, 2014, pp. 33-38, doi:10.1109/CBI.2014.39.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review, 50*(4), 370–396. https://doi.org/10.1037/h0054346

Mayhill, G. R. (1953). "A critique of CAA studies on air traffic generation in the United States". *Journal of Air Law & Commerce*, 20, 158-177 https://scholar.smu.edu/jalc/vol20/iss2/3

Richmond, A. K. & Kaufmann, R., (2006). "Is there a turning point in the relationship between income and energy use and/or carbon emissions?", *Ecological Economics*, 56, issue 2, p. 176-189, https://EconPapers.repec.org/RePEc:eee:ecolec:v:56:y:2006:i:2:p:176-189.

Sala-i-Martin, X., (1996). The classical approach to convergence analysis, *Economic Journal*, 106, issue 437, p.1019-36, https://EconPapers.repec.org/RePEc:ecj:econjl:v:106:y:1996:i:437:p:1019-36.

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 130–135. https://doi.org/10.11919/j.issn.10020829.215044

Srisaeng, P., Baxter, G., Richardson, S., & Wild, G. (2015). A forecasting tool for predicting Australia's domestic airline passenger demand using a genetic algorithm. *Journal of Aerospace Technology and Management*, 7(4), 476-489. https://doi.org/10.5028/jatm.v7i4/475

United Nations Development Programme. (2019). Human Development Reports. Retrieved September 10, 2020, from http://hdr.undp.org/en/content/human-development-index-hdi

United Nations. (2019). Gross domestic product and gross domestic product per capita. United Nations Data. Retrieved November 10, 2020, from http://data.un.org/_Docs/SYB/PDFs/SYB63_230_202009_GDP%20and%20GDP%20Per%20Capita.pdf

Watson, R. T., Patz, J., Gubler, D. J., Parson, E. A., & Vincent, J. H. (2005). "Environmental health implications of global climate change". *Journal of Environmental Meteorology*, 7 (9):834-843.

Wensveen, J., G., (2011). Air transportation: a management perspective. 7th edition. Farnham: Ashgate Publishing

Young, A. T., Higgins, M. J., & Levy, D. (2007) Sigma convergence versus beta convergence: Evidence from U.S. county-level data. *Emory Law and Economics Research Paper* No. 07-4, http://dx.doi.org/10.2139/ssrn.441460