London Tourism Analytics

by Ziyi Zhao, Hongbo Liu, Yuxuan (Leo) Liu, Jingze Zhang, and Erya Ouyang

Masters of Science in Business Analytics, December 2018,
George Washington University School of Business

A Thesis submitted to

The Faculty of
The School of Business
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science
in Business Analytics

December 2018

Thesis directed by

Dr. Shivraj Kanungo

Chair of the Department of Decision Sciences
Faculty Director of Decision Sciences
Associate Professor of Decision Sciences & of Info Systems & Tech Management

Acknowledgement

We would like to thank Kyle Harbacek, Jason Houghton, and Lauren Mandell of the Deloitte for providing the topic and for mentoring our team throughout this project. We would also like to thank Dr. Shivraj Kanungo of the George Washington University School of Business for guiding our team on the data analytics techniques used in this project.

Abstract

London, the capital of England and the United Kingdom, is a world's leading tourism destinations.

According to Wikipedia, the city attracted nearly 20 million international visitors in 2016, which making it one of the world's most visited in terms of international tourists. Research has been showing that tourists that enter the UK under for significant durations can have a higher probability of staying beyond the legally determined length of visits. International tourists who remain in the Greater London Area (GLA) can influence the city's KPI levels aligned to job performance and socioeconomic indicators. The current analysis was carried out to construct foundational classification capabilities, centered on detecting patterns regarding individual sovereign nation's tourism patterns, given their respective socio-economic prosperity indicators to the GLA have the highest probability levels of having citizens stay 15+ days. The success of the classification model and visualization platform will provide a basis for next generation predictive capabilities in detecting international prosperity measures and London tourism patterns.

Executive Summary

Project Objective

This project will be delivered to The City of London's Local Ministry and National Security officials in order to serve as a long-term analytics maturity model to advise existing security practices. Among all EU member nations, London has been selected to be the pilot city in this project to pioneering this predictive capability. A successful development of the first London Tourism Analytics prototype serve as a foundation solution can translate a multi-year contract to maintain and expand this capability to other cities.

The various elements of the project are summarized as follows:

a. To describe the background of the project
b. To describe the process used to conduct the analytics
c. To outline findings and results
d. To provide some suggestions about what to do next

Background

This analytical study is a first phase analysis conducted by us. It was done in partnership with Deloitte Touche Tohmatsu Limited, commonly refer to as Deloitte, a multinational professional services network, and George Washington University (GWU), a private research university in Washington, D.C.

Process

This project last for 4 months and was based on two datasets. One is the International Tourism Data from 2002 to 2017 provided by office for National Statistics (ONS) and International Passenger Survey. The other is the Global Prosperity Population Data from 2007 to 2017 provided by the Legatum Institute. In order to effectively processing the analysis, we merged two datasets into one and apply both supervised and unsupervised methods, including Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors (KNN), Naïve Bayes, and Kernel Support Vector Machine (SVM).

Finding and Conclusion

Based on all the scores and validations (figure 1 below), we found Logistic Regression and Random Forest Classification are two best models to predict whether a new visitor will stay 15+ days in Great London Area (GLA).

Model	Accuracy Mean	Accuracy Standard Deviation
Logistic Regression	0.9209	0.00334
Decision Tree Classification	0.8652	0.00518
Random Forest Classification	0.9067	0.00431
K-Nearest Neighbors (KNN)	0.8655	0.00675
Naïve Bayes	0.8480	0.00530
Kernel SVM	0.8655	0.00675

Figure 1. Models Accuracy and Variance

Recommendation for actions

By choosing Logistic Regression as our final model, we decide to further improve it, so that the City of London's local ministry and National Security officials able to use it to ensure citizen safety and then to make policy-based decisions regarding economic prosperity and development to next generations.

Table of Contents

Glossary of Terms

Package

- **Pandas**: In computer programming, panda is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.
- **Numpy**: a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- **Python Dashboard**: Dash is Python framework for building web applications. It built on top of Flask, Plotly.js, React and React Js. It enables you to build dashboards using pure Python.

Functions

As you might know by now, we can't have text in our data if we're going to run any kind of model on it. So, before we can run a model, we need to make this data ready for the model.

- **Encoding**: The process of applying a specific code, such as letters, symbols and numbers, to data for conversion into an equivalent cipher.
- **Label Encoding**: An approach to encoding categorical values in a column to a number.
    - What label encoding does is, if you have three countries in the first column have been replaced by the numbers 0, 1, and 2.

- **One-Hot Encoding**: A process by which categorical variables are converted into a form that could be provided to machine learning algorithms to do a better job in prediction.
    - What one hot encoding does is, it takes a column which has categorical data, which has been label encoded, and then splits the column into multiple columns. The numbers are replaced by 1s and 0s, depending on which column has what value.
- **Target encoding**: the process of replacing a categorical value with the mean of the target variable.
    - Unlike label encoding, which gets the work done efficiently but in a random way, mean encoding tries to approach the problem more logically. In a nutshell, it uses the target variable as the basis to generate the new encoded feature.
- **Feature Engineering**: The process of using domain knowledge of the data to create features that make machine learning algorithms work. Feature engineering is fundamental to the application of machine learning and is both difficult and expensive.
- **Principle Component Analysis**: a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.

List of Figures

Chapter 1: Introduction

This analysis was partnership with Deloitte Touche Tohmatsu Limited and George Washington University and will be delivered to The City of London's Local Ministry and National Security officials in order to serve as a long-term analytics maturity model to advise existing security practices. Among all EU member nations, London has been selected to be the pilot city in this project to pioneering this predictive capability. The two institutions developed this project as a way to provide a foundation predictively solution, and then translate a multi-year contract to maintain and expand this capability to other cities.

The project was framed around two main study questions:

1. ***Whether the potential tourist staying 15+ days in Great London Area have higher risk of immigration?***
2. ***Whether visitors coming from certain countries have higher probability to stay 15+ days in Great London Area?***

Besides above two main questions, we also interested in to find out the following topics/questions:

1. ***What are the main factors that affect visitors spend 15+ days in Great London Area?***

   By using the macro and micro indicators, as well as other inputs that provided in the two data sets to answer this question, so that The City of London's Local

Ministry and National Security officials could know which factors they need to pay more attention to.

2. ***How the international visitors spending pattern affects the different fields of prosperity?***

   By analysis of countries' spending and their potential purchasing power, the City of London's Local Ministry and National Security officials could attract more visitors (who spend more money in London) while limit visitors (who are likely to illegal immigrate to London).

3. ***What are the changes of international tourism trend?***

   By using the macro and micro indicators, as well as other inputs that provided in the two data sets to answer this question, so that The City of London's Local Ministry and National Security officials could better understand the international tourism trend, performance, and thus adjusting policies to make improvement.

Chapter 2: Background & Data Source

London Tourist Board

The London Tourist Board, the official regional tourist board for London since 1969, is responsible for marketing and promoting the capital, providing tourist information services and recommending improvements to the infrastructure and facilities for the growth of tourism.

With the establishment of Visit London in 2003, the London Tourist Board was divided into two parts, Visit London and the London Development Agency. Visit London took over the marketing and promotion of London, while the London Development Agency was responsible for planning, research and development. Visit London was subsequently put into administration by its main funder Greater London Authority and replaced with a new organization London & Partners.

Since tourism is now one of the London's three most important industries, it brings streams of economic flow for London, but also alarms for London's safety. The London local ministry, under the increased visibility from MI5 and Prime minister, hopes to ensure citizen safety with the help of TTW analytics team by predicting the probability of international visitors staying beyond the legally determined length of visit based on the dataset International Visitors in London provided by the London Tourist Board.

In addition to citizen safety, the success of this analysis is expected to be used in detecting international prosperity measures and London tourism patterns.

Legatum Institute

The Legatum Institute is a London based independent educational charity with the global mission of "to see all people lifted out of poverty". As it is established, its objective is to understand, measure and explain the journey from poverty to prosperity for individuals, communities and nations.

Their events, reports and publications aim to advance the research in the fields of economic, political and social policy. All reports and data from Legatum Institute, in connection with programs regarding economics, government, migration, culture, global trade, are designed to educate the public and further their charitable objectives.

Availability of Datasets

The first dataset London Tourism Data, available for Visit London website, shows the London totals for nights, visits and spend with break-down by purpose, duration, mode and country. It provides empirical data from 2002 to 2017 with quarterly update. The detailed data-taken process is conducted by International Passenger Survey (IPS).

The three most important variables in this project are

1) Nights

Provides the information on how many nights foreign residents from same country with same predicted duration of stay in a quarter of a certain year stayed in the London

2) Spends

Shows the amount of UK money spent by foreign residents from same country with same predicted duration of stay in a quarter of a certain year

3) Visits

Indicates the times of visits foreign residents from same country visited London during one quarter in a year with different transportation mode and purposes of visits.

The second dataset Legatum Institute Prosperity Indicators shows nines pillars of prosperity measuring both wealth and wellbeing with break-down variables in each pillar. The most significant difference from other prosperity index is that they measure beyond GDP. Governance, Personal Freedom, Social Capital and Safety & Security as four institutional foundations of prosperity index provide the basis for trust and constructive social and economic relationships among people, business and government.

1) Governance

Governance examines the institutions that empower and constrain government action. Since the rule of law, strong institutions and regulatory quality contribute significantly to economic growth and have significant impact on prosperity. In addition, effective, fair and accountable governments increase public confidence, and, ultimately, result in higher levels of life satisfaction among citizens. The variables under Governance are Rule of Law, Government Integrity, Government Performance and Political Participation. *Rule of Law* refers to the law applied equally to all and to the quality of contract enforcement, property rights, polices and courts. *Government Integrity* assesses how well the government is operated in a transparent manner with minimum corruption. Citizens' awareness of political participation strengthens the government accountability to a large extent. *Government Performance* measures the performance of civil services and the quality of regulation and policy incentivize business. *Political Participation* assess how strong their willingness to engagement in political process.

2) Personal Freedom

A country will benefit from its high-level personal freedom since protected personal liberties will enhance the citizens' satisfaction. *Basic legal rights* assess the degree of personal autonomy. High-level personal autonomy indicates the right to one's own person and freedom around what information they can publish and consume. *Individual Liberties* measure the capability of citizens to choose what life to live with, including the free choice of religion and belief and the right to own their property without the intervene

of government. *Social Tolerance* assesses how well the government, community and society respect diversity of people's religious beliefs, ethnicities, origins and sexual orientations. This embedded diversity indicating new markets and fresh ideas, create innovation for the society. Simultaneously, the society will benefit from the cultural interactions among people from different backgrounds.

3) Social Capital

Social Capital describes the social networks among people, assessing the degree of trust among people and support from their friends or families. The "capital" stresses the contribution of social networks as an asset that benefits the whole society with increase economic returns and people with enhanced belongingness. *Personal and Social Relationships* measures the support and bonding of social capital among families, personal relationship and genuine community. *Social Norms* describes the trusts in institutions and the level of respect people receive from each other in a community. *Civil Participation* measures the degree of civic and political participation in the form of volunteering, donating and some political processes.

4) Safety and Security

Since it is necessary for a nation to attract investments and keep economic growth in a safe and secure environment, safety and security are integral to securing prosperity. *National Security* assesses how well people within the country are kept safe from conflict and violence, including coups, state-sanctioned killings, torture, disappearances and political imprisonment. *Personal Safety* measures how well the property and person are

protected and respected, including thefts, homicides and safety while walking alone. *Security of Living Conditions* examines the degree of safety and security of the environment where people live in. This includes the secure housing, food supply, the safety of the living environment and infrastructure.

5) Education

Human capital is an asset for a society as better educated people contribute more to the society's economic development and prosperity. In addition, more exposure to education allows them to better fulfill people themselves. *Access to Education* examines how wide access to education for all ages, genders, ethnicities people. *Quality of Education* assesses the quality of education attained by the people within the nation by measuring the number of people achieving primary and secondary education, the number of high-educated people, the degree of international conference participation of universities and the international reputation of universities. *Human Capital of the Workforce* measures the skills of research and development within institutions. These are measure at secondary and tertiary level.

6) Economic Quality

Economic quality is usually measured by the level of people's satisfaction with their standard of living. *Standard of Living* is one of the variables determining how well the economic quality is. It measures whether people have access to affordable goods and services. *Economic Inclusiveness* examines the degree of people's participation in the economic activities with available resources and opportunities. *Anti-Monopoly Policy*

assesses who well merchants can compete freely in the market and consumers have their rights to choose what they buy and buy from which merchants. *Labor Force Participation* measures the extent people participate in the workforce.

7) Business Environment

A strong business environment, comprised of an entrepreneurial climate, openness to new ideas and opportunities, leads to more wealth and greater social wellbeing. *Entrepreneurial Environment* assess how easy to start and run a company. *Business Infrastructure* examines the infrastructure that enables market access for individuals and firms, which includes transport/logistics, utilities and communications. *Access to Credit*, *Investor Protections and Labor Market Flexibility* are indicators to measure how well the business environment is for an entrepreneur to start businesses.

8) Health

An effective health infrastructure is critical for people living in the nation, since those who enjoy good physical and mental health report high levels of wellbeing. *Health Outcomes*, including mental and emotional wellbeing, as well as basic measure of mortality and life expectancy, are the basic outcomes expected by members of a prosperous society. *Health Systems Quality* looks at the adequacy of health infrastructure, service quality and preventative care (including sanitation, immunization and broader public health). This is reflected in people's satisfaction with their healthcare.

9) Nature Environment

A high-quality natural environment conveys a sense of wellbeing and satisfaction to a country's population through characteristics that may be physical (such as air quality), social (such as green areas in which to meet) or symbolic (such as national parks).

Chapter 3: Description of Work Undertaken

Part 1: Data Visualization

Because this analytical study is the initial prototype conducted by us, in order to find any interested patterns in the data sets, we have to first look at the data. One of the best ways to look at data is data visualization. So, we put our main data set, the International Tourism Data 2002-2017 in Tableau, an application allows for instantaneous insight by transforming data into visually appealing and interactive visualizations.

We first look at the International Tourism Data 2002-2017. There are total 62 countries in the International Tourism Data 2002-2017. By looking at *Figure 2: Countries Distribution & Counts*, the darker the color, the more data we have in hands. For example, most of the visitors in our data are from France while least of the visitors in our data are from Qatar.

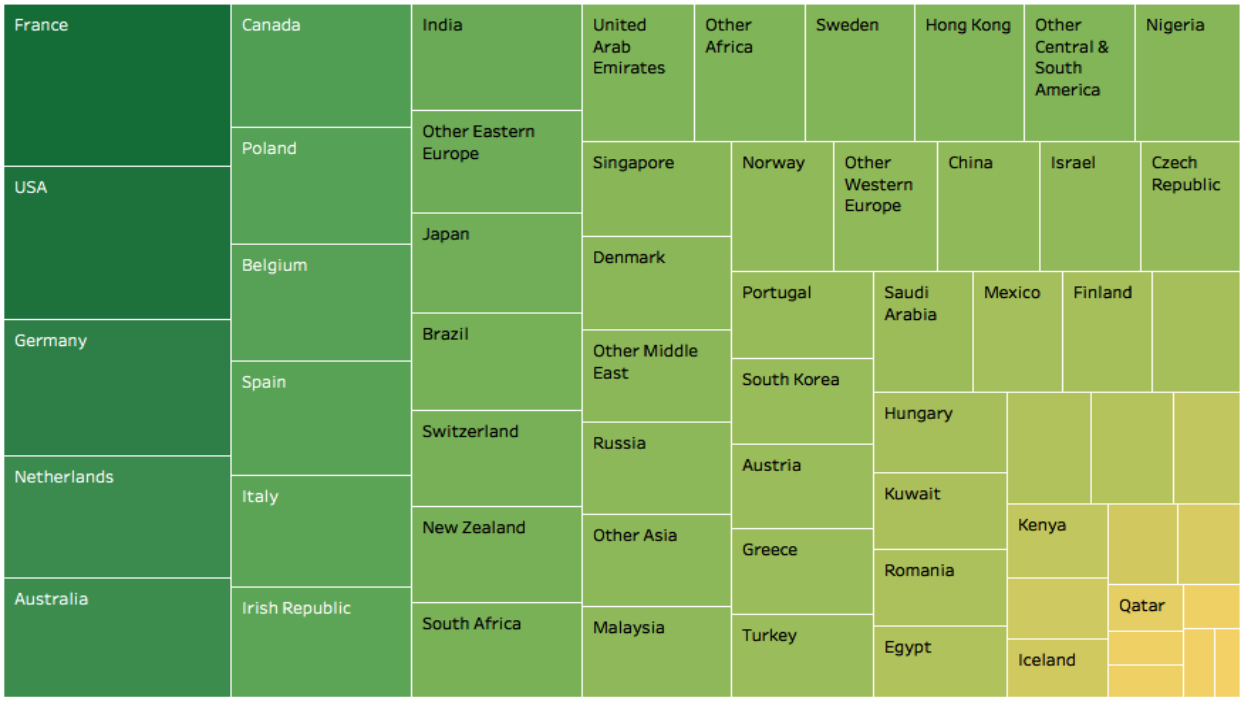

Figure 2: Countries Distribution & Counts

There are total 5 purposes that visitors travel to Great London Area, including

1) Visit Friends and Relatives = VFR (green)
2) Study (light blue)
3) Miscellaneous (red)
4) Holiday (orange), and
5) Business (dark blue)

By looking at *Figure 3: Purpose by Countries & Counts*, the higher the bar chart of each purpose stands for the more visitors travel to London because of that purpose. For example, visitors come to London to study has the least number while visitors travel to London for Holiday and VFR have a comparably large number in those 62 countries.

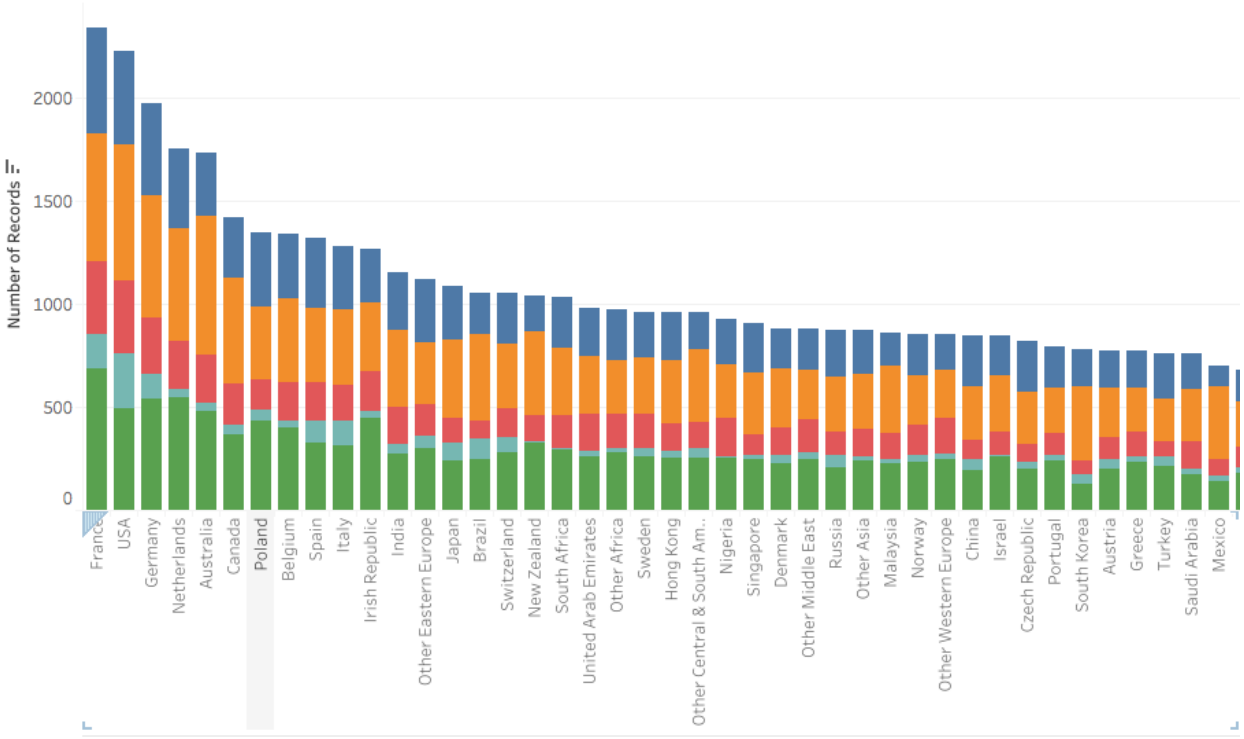

Figure3: Purpose by Countries & Counts

By looking at *Figure 4: Purpose, Nights, Spends by Year*, it can be seen that business people and holiday travelers have an overall increasing trend, whereas for other purposes, the trend lines are flat. The situation is the same for future prediction.

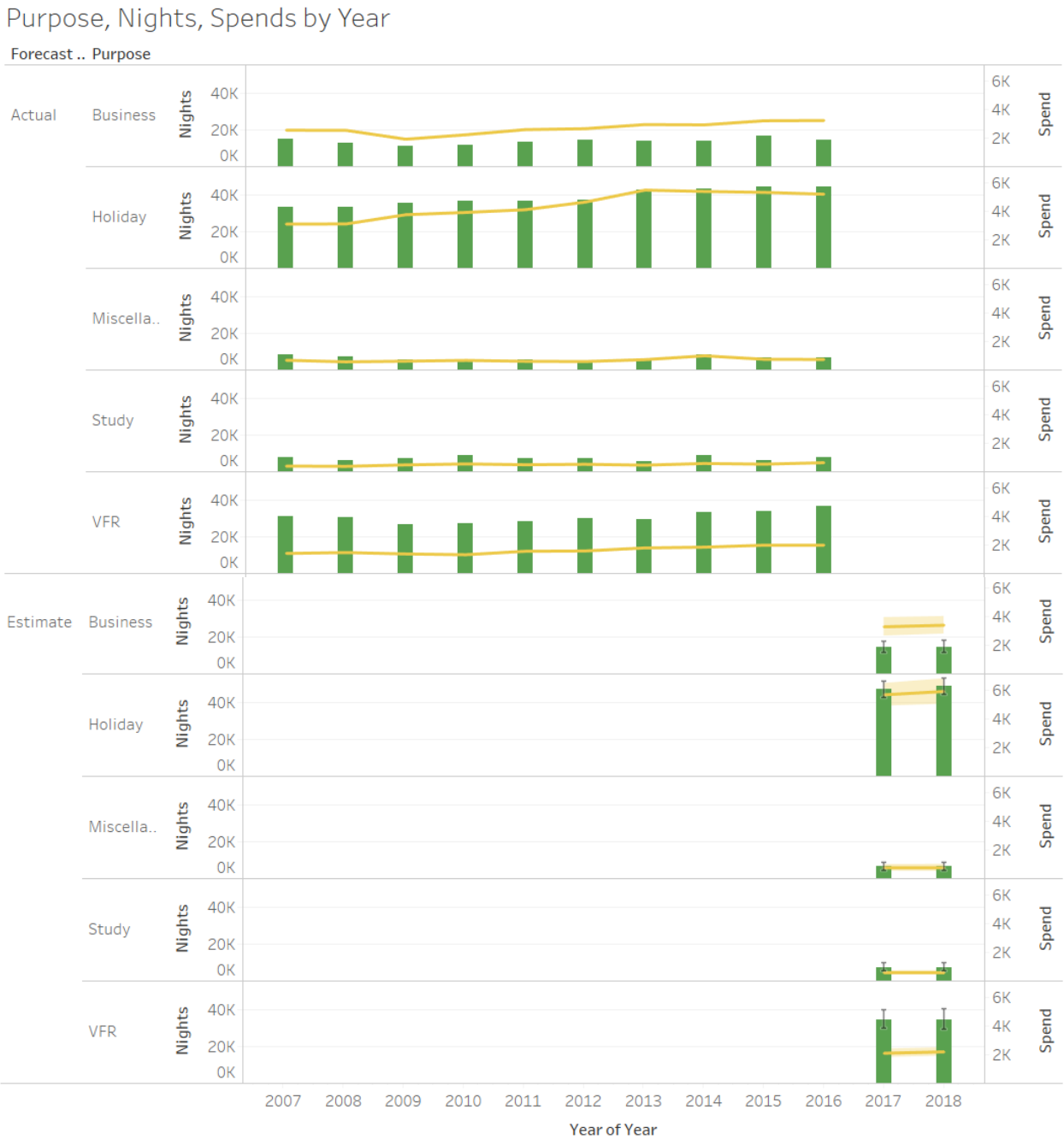

Figure 4: Purpose, Nights, Spends by Year

The International Tourism Data 2002-2017 divided people duration of stay in Great London Area into 4 categories, they are visitors stay

1) 1-3 nights
2) 4-7 nights
3) 8-14 nights, and
4) 15+ nights

By looking at *Figure 5: Duration of Stay & Counts*, please pay attention to the last categories, visitors that spend 15+ night in Great London Area. This collection of people is the target group in this analytical project.

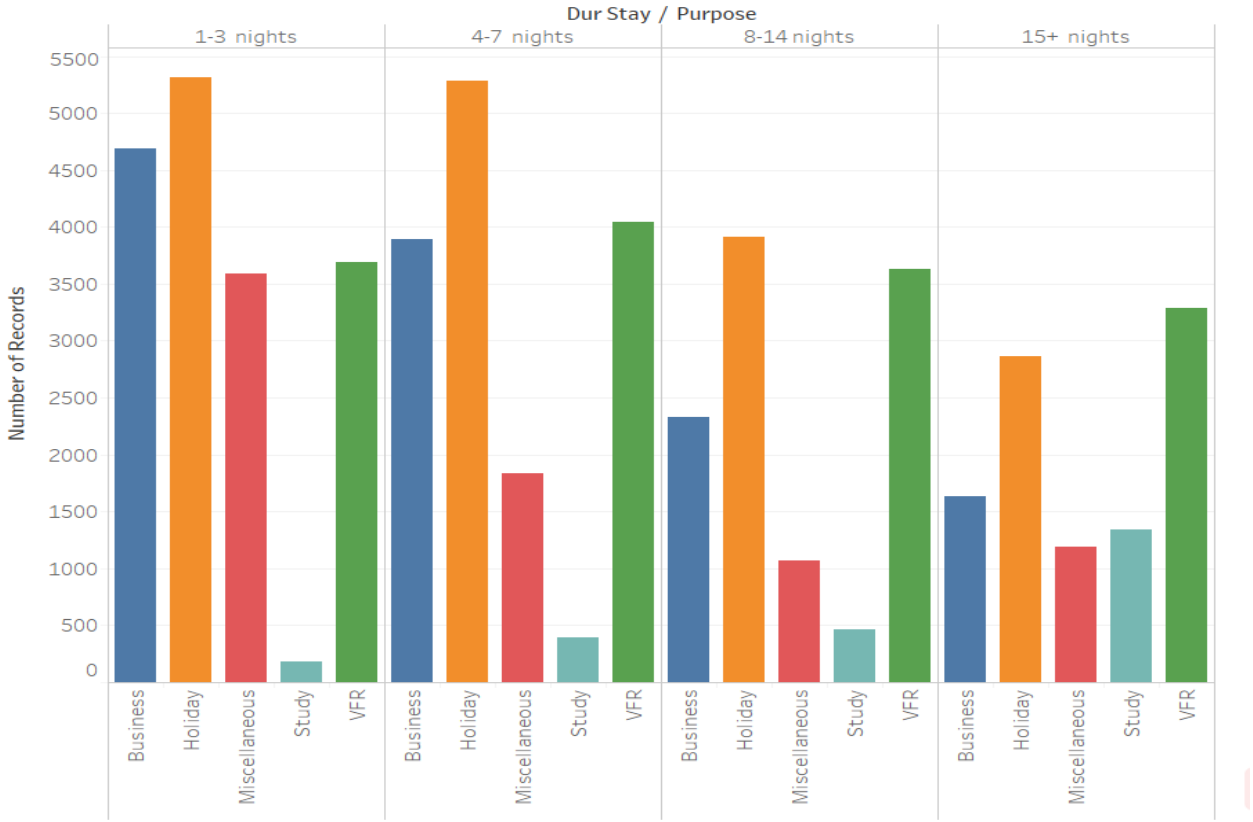

Figure 5: Duration of Stay & Purpose Counts

We then look at our second data, the Global Prosperity Population Data from 2007 to 2017. Our final purpose and main concern of this project is the security issue. Therefore, we put the Safe Column of the Global Prosperity Population Data in the Tableau to further analysis it.

By looking at *Figure 6: Country & Safety*, the darker the color, the safer the countries are. Over the 149 countries in our second data set, Singapore, Iceland and Japan are the top 3 safest countries while Democratic Republic of Congo, Iraq and Central African Republic are the top 3 most dangerous countries.

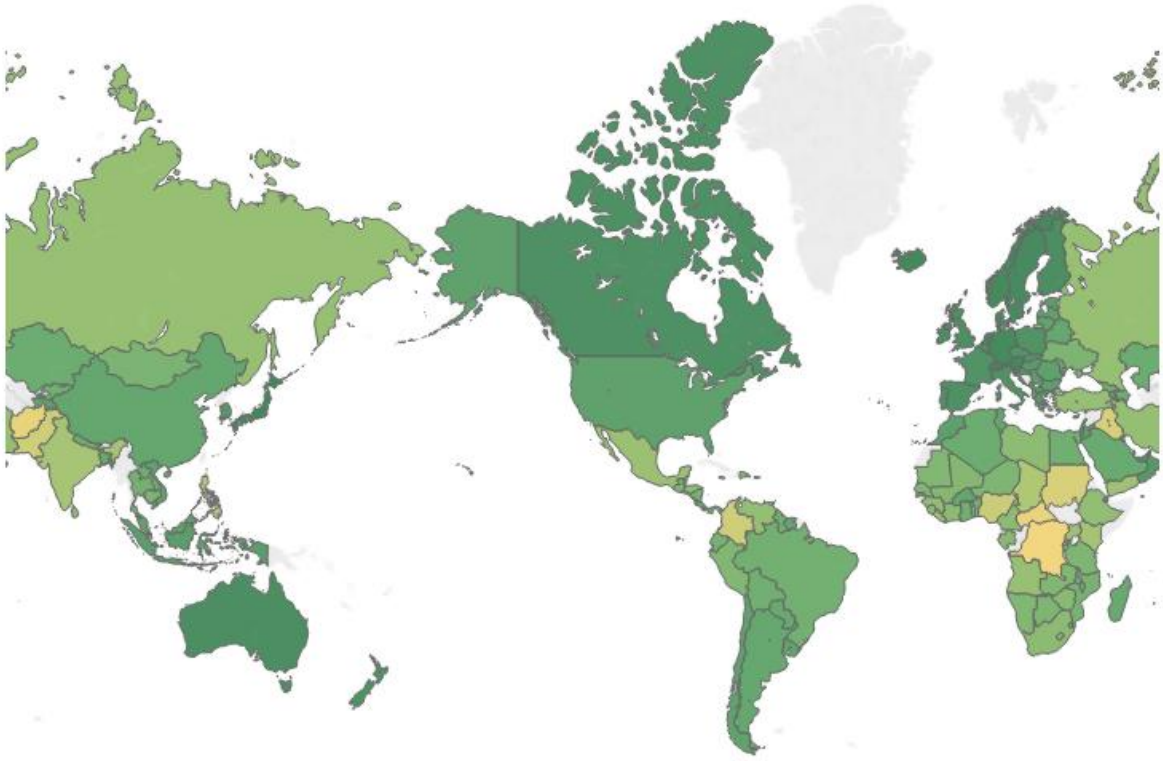

Figure 6: Country & Safety

By looking at *Figure 7: Spend, Visits and Average safe index*, it shows that USA is the one with a medial safe index, spending and visiting the most in London. Besides, the majority of European

countries, for instance, France, Germany, Spain, Belgium and Netherlands, have a relatively higher average spend and average visit days.

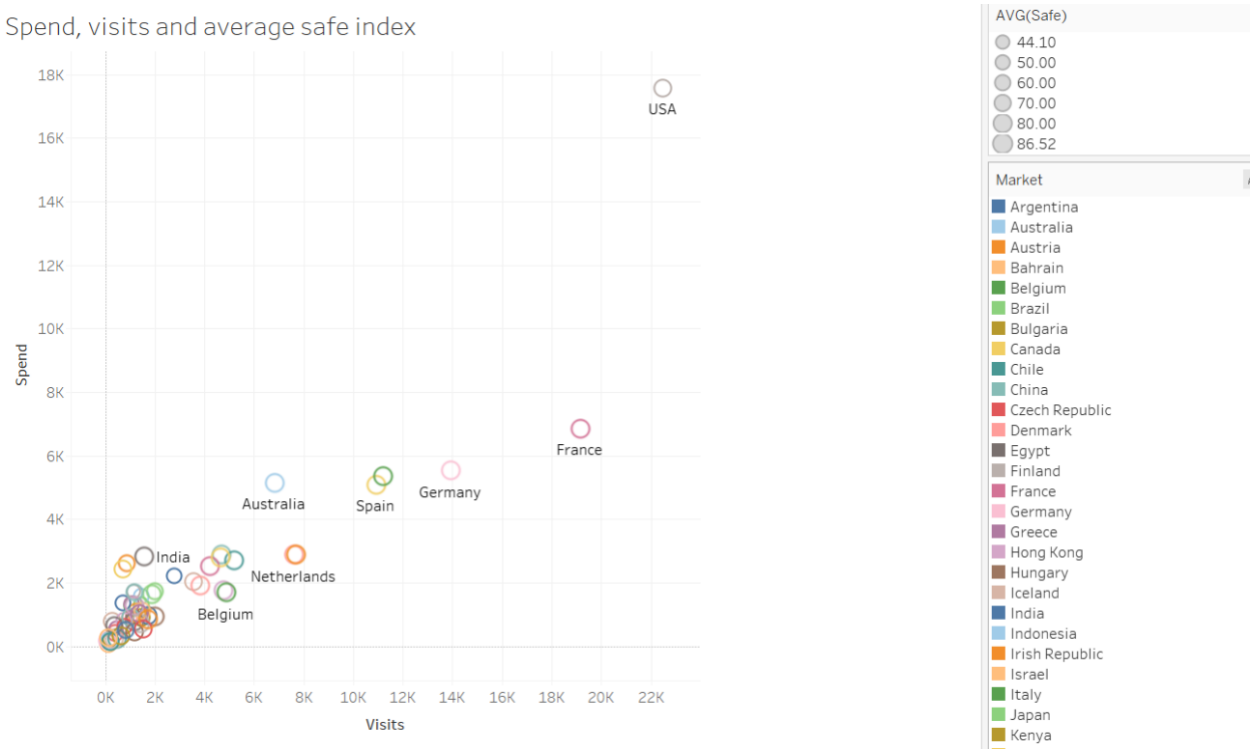

Figure 7: Spend, Visits and Average safe index

By looking at *Figure 8: Bottom 10 countries by Average safe index*, it is clearly that countries with a lower safe index spend less and visit less. However, we should pay more attention on the countries staying in the bottom right corner. It seems like they tend to spend little but stay longer in London, which is not good for London's safety.

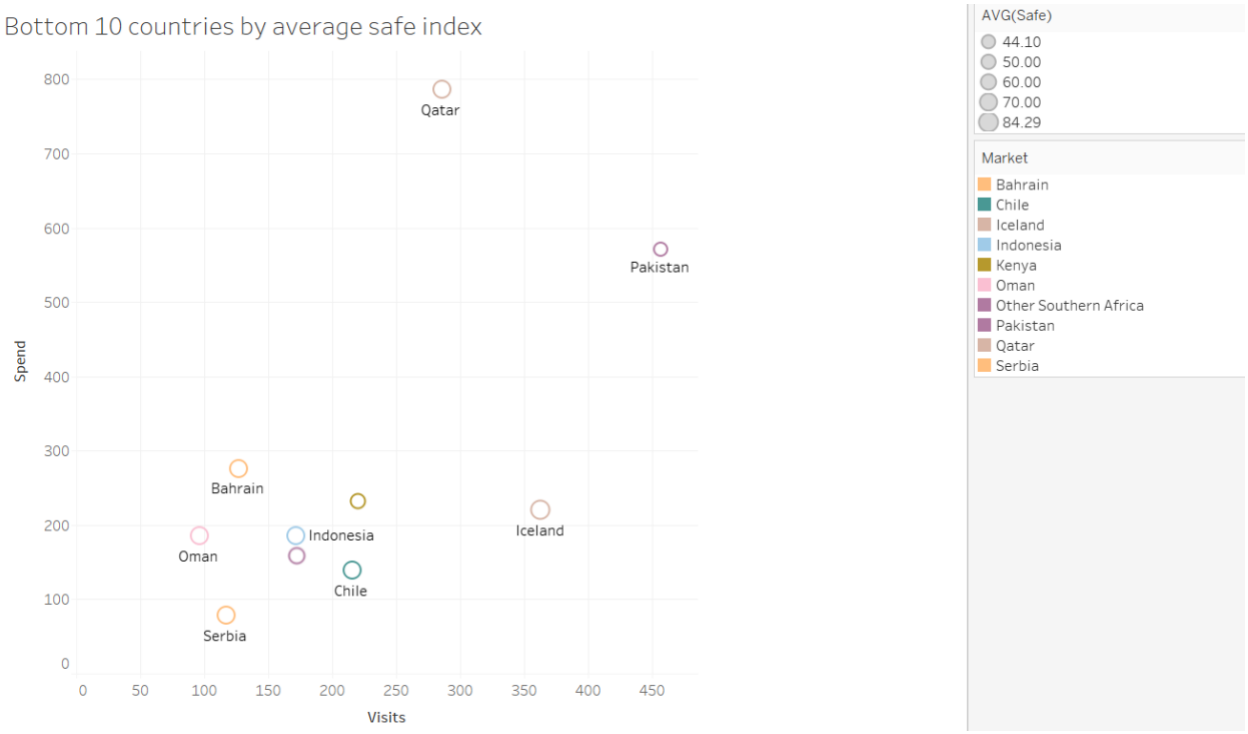

Figure 8: Bottom 10 countries by Average safe index

Part 2: Data Preparation & Exploration

After picturing some important data in Tableau, we have a basic understanding of the project and gained some preliminary ideas about what we are going to analyze and explore. However, we found there are three main problems in the data sets, which we must solve them first before we proceed to the next step of analysis.

The three problems are

1) **Two data sets do not match**

   a) *Missing 5 years' records in the International Tourism Data*

   The Global prosperity Population Data contain records from 2002 to 2017. However, the International Tourism Data comprise of records from 2007 to 2017. Therefore, we

have to choose the common year to analyze, which means we need to drop the records in 2002, 2003, 2004, 2005, and 2006 from the Global prosperity Population Data.

b) *Countries' names do not match*

For examples, "United States" and "Ireland" are the name used in the Global prosperity Population Data. But "USA" and "Irish Republic" are the name used in the International Tourism Data. We need to make them consistent because we know USA is United States and Ireland is Irish Republic. So, we replace the name in Global prosperity Population Data by using the countries name in the International Tourism Data.

c) *Number of countries are different in two data sets*

Our first data, the International Tourism Data, has 62 countries in total. However, our second data, the Global Prosperity Population Data has 149 countries in total. After carefully examination the two data sets, we drop "Taiwan" not only because of its sensitive political issues, but also because it only appears in one data set. We categorize countries into different regions as the table shown below.

Asia	India	South Asia
	Pakistan	West Asia (Middle East or Near East)
	Bahrain	
	Israel	
	Kuwait	

	Oman	
	Qatar	
	Saudi Arabia	
	United Arab Emirates	
	Turkey	
	Other Middle East	
	China	East Asia
	Hong Kong	
	Japan	
	South Korea	
	Indonesia	Southeast Asia
	Malaysia	
	Singapore	
	Thailand	
	Other Asia	Central Asia
Europe	Denmark	Northern Europe
	Finland	
	Iceland	
	Norway	
	Sweden	
	Greece	Southern Europe
	Italy	
	Portugal	Southwest Europe
	Spain	
	Bulgaria	Southeast Europe
	Romania	
	Serbia	
	Belgium	Western Europe
	France	
	Irish Republic	
	Luxembourg	
	Netherlands	
	Other Western Europe	
	Austria	Central Europe
	Czech Republic	
	Germany	
	Hungary	
	Poland	

	Switzerland	
	Russia	Eastern Europe
	Other Eastern Europe	
Africa	Egypt	North Africa
	South Africa	South Africa
	Other Southern Africa	
	Nigeria	West Africa
	Kenya	Eastern Africa
	Other Africa	Central Africa
America	Canada	North America
	Mexico	
	USA	
	Argentina	South America
	Chile	
	Brazil	
	Other Central & South America	Central America
Oceania	Australia	Oceania
	New Zealand	

2) Global Prosperity Population Data

There is too many detailed information regarding how Legatum Institute calculated governance, personal freedom, social capital, safety and security, education, economic quality, business environment, health, and nature environment variables. Since the purpose of this project is not to understand how this data is being prepared for us. We do not need to pay attention to how Legatum Institute calculate those variables. Therefore, we only use the "pillar" sheet, which contains the information summary of all PI index, in the Global prosperity Population Data to proceed our predictive analysis.

3) International Tourism Data

a) *Dropped column "Sample"*

The reason we dropped column "sample" is because how the office of the National Statistics (ONS) and International Passenger Survey got the sample and the method they calculated it are too complex to analyze. The column "sample" is not helpful to the final prediction process.

b) *Dropped column "Area"*

The reason we dropped column "area" is because we know we are focusing on the Great London Area, and all the data we have in hands is from London. Thus, it is not necessary to have this redundant column.

c) *Dropped column "Quarter"*

The reason we dropped column "Quarter" is because there is no relevant columns in the Global Prosperity Population Data. Since we are not going analyze London tourism pattern by quarters, we simply eliminate this column to avoid complexity.

Part 3: Data Combination

After data exploration and solved the problems in the data prepare process, our next step is to combine two datasets into one final dataset. The final dataset has 19 columns, including year, market (country), dur_stay (duration of stay), mode, purpose, visits, spend, nights, country's general PI score, and scores in 9 different measures of the prosperity described above.

As for the detailed datasets merging process, it will be explained in three steps as followed.

1) **Handling Prosperity Index for binned countries**

   Due to datasets inconsistency, some countries are binned. Because the prosperity index of these binned countries is missing, we need to integrate them. Since countries in the same binned group have similar prosperity index in each year, we decide to average these countries' index as the final binned group's index of each year.

2) **Merge International Tourism Dataset and Prosperity Index**

   Although our analysis focus on the 'country' level, we also want to make 'year' as the prediction feature. Therefore, two datasets will be merged based on different countries and years. The completed dataset will have 61 countries in 2007-2017, including 7 binned groups.

3) **Drop repeat columns**

   After datasets merging process, we dropped the repeated columns "country" and "year". Besides the repeated columns, we also drop all 9 measures of prosperity's ranking score,

which are originally in the Global Prosperity Population Data. Because country's ranking columns in Global Prosperity Population Data shares the same information with detailed index score in the same data, so the index score columns are non-relative. This drop process can make the data as a better input for future prediction model.

Part 4: Exploration of Final Dataset

Before moving forward to the modeling stage, we would like to do an exploratory analysis to the new merged dataset. By this exploration, we can know whether we need to do further data manipulation. If so, what should we do. By conducting an exploration, we have three noticeable points that we like to introduce.

1) Skewness of the three main features

By reviewing the histogram of three main features "nights", "spend", and "visits" as *Figure 9: Frequency of Visits/Spend/Nights*, we saw a strong right skewness in these three main features. Because properly handle skewness issue is a necessary step before fitting them into models, we expected to explore the correlation between these three variables with the stay duration respectively.

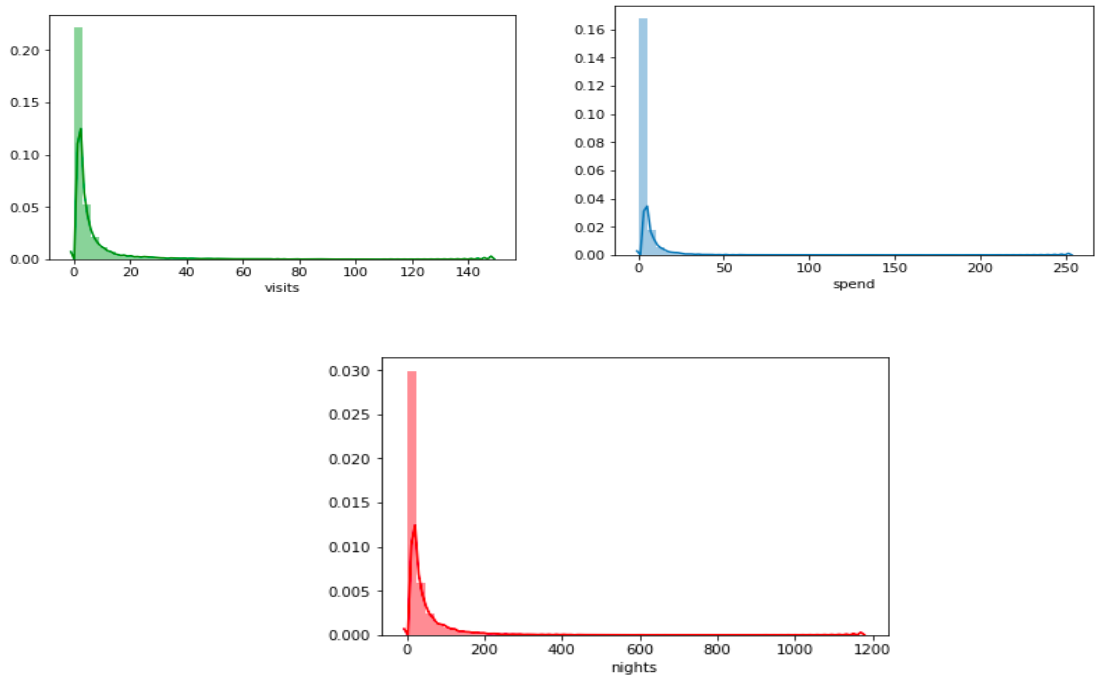

Figure 9: Frequency of Visits/ Spend/ Nights

2) Outliers (Extreme records)

Given the histogram of three main features in the final dataset, we decided to analyze the influence of some extreme records. Therefore, we found records that has the maximum value in the column “visits”, “spend”, and “nights” respectively.

a. *Visits*

year	2010
quarter	Q4
market	France
dur_stay	<15 nights
mode	Tunnel
purpose	Holiday

For this group, it has the maximum visits value, which is 148.193.

b. *Spend*

year	2017
quarter	Q3
market	Saudi Arabia
dur_stay	<15 nights
mode	Air
purpose	Holiday

For this group, it has the maximum spend value, which is 252.552.

c. *Nights*

year	2014
quarter	Q2
market	USA
dur_stay	15+ nights
mode	Air
purpose	Study

For this group, it has the maximum nights value, which is 1172.

3) Two new columns for better exploration

As showed above, the difference in value between outlier and most records are huge. In order to better explore the final dataset, having a quick look for prediction expectation, we create two more columns, "*nights per visit*" and "*spend per visit*".

Reviewing the nights and corresponding spend for each visit can provide us a hint that which combination in year, country, mode and purpose would show a negative or target pattern that stay more but spend less.

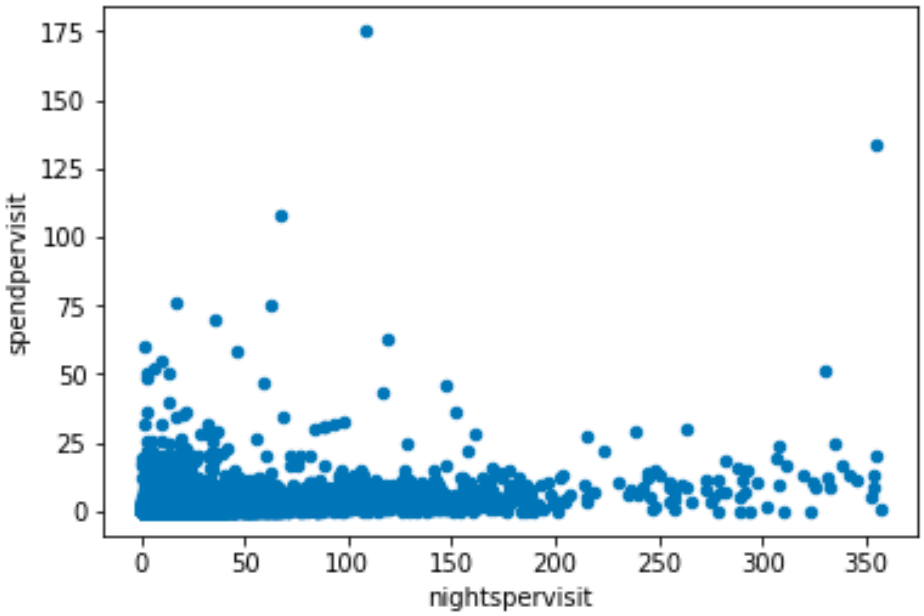

Figure 10: Scatter Plot between Nights per Visit & Spend per Visit

The *Figure 10: Scatter Plot between Nights per Visit & Spend per Visit* showed above is the scatter plot between nights per visit and spend per visit for all records in the final dataset. From this view, we can see that there are some records show negative pattern, which is low spend per

visit but high night per visit. Therefore, we found those records whose spend per visit is 5% lowest, and nights per visit is 5% highest. Among all records, only 5 records can fulfill the conditions above, and they are showed in *Figure 11: Countries with high dur_stay but low spend*.

.

year	quarter	market	dur_stay	mode	purpose
2007	Q1	Spain	15+ nights	Air	Business
2007	Q2	Other Asia	15+ nights	Air	Holiday
2007	Q2	United Arab Emirates	15+ nights	Air	Miscellaneous
2007	Q4	Switzerland	15+ nights	Air	VFR
2007	Q4	Other Eastern Europe	<15 nights	Tunnel	VFR

Figure 11: Countries with high dur_stay but low spend

Part 5: Data Manipulation for Modeling

Based on a glance at the final dataset above, we found an issue of highly skewed data in three main columns. Then, the expectation of prediction is also showed. However, all current results are based on a relatively easier analysis of the final raw data. In order to perform an accurate and quantified analysis of possibility in staying more than 15 days, statistical models are needed.

To achieve a relatively accurate prediction result, our next step is data manipulation. There are three processes in this step: standardization on numeric variables, target encoding on categorical variables, and target variable transformation.

1) **Standardization**

   We use average-by-level approach to do the standardization process. In this process, for each numeric column, we replace the original data with a quotient that is the difference between each record and the mean of the column divided by the standard deviation of the column. After this standardization process, both skewness in three main features that we discussed above, and the different scales in two datasets are handled appropriate without further drop any record.

2) **Target encoding on Categorical variables**

   Not only numeric variables, but also categorical variables are needed to be encoded. For some models, such as regression model, the existence of categorical variable may lead a

less accurate model training, and thus a target encoding on categorical variables are also implemented.

We have 5 categorical variables in this final dataset, "year", "market" (country), "mode", and "purpose". For the target encoding on categorical variables process, we use rate-by-level approach. In this approach, we replace each unique level in certain categorical column with this level's event rate in our target variable (dur_stay). Then, apply this transform to the same level among all records. As a result, all categorical variables become numeric variables, which are also highly relate to the target variable.

3) Target Variable Transform

Last but the least step in data manipulation is to transform the target variable. Our target variable is dur_stay (duration of stay). The original dataset provides different levels in this column, such as 1-3 days, 4-7 days, 15+ days. However, we believe two levels can be better interpreted our main goal, whether visitors will of stay will 15+ day. So, we combined all groups that less than or equal to 15 days as one group and leave the other group as 15+ days as original. Then, we replace all records' duration stay in the <15 days group with 0 and replace another group's duration stay as 1. After this final step, our dataset thus includes all numeric variables that are convenient to be trained for different types of models.

Part 6: Feature Engineering (Principal Component Analysis)

Principal component analysis (PCA) is a technique used for identification of a smaller number of uncorrelated variables known as principal components from a larger set of data. The technique is widely used to emphasize variation and capture strong patterns in a data set. Principal component analysis is focused on the maximum variance amount with the fewest number of principal components. Especially makes use of principal component analysis to eliminate the number of variables or when there are too many predictors compared to number of observations or to avoid multicollinearity.

It is true that the less components are chosen, the more information will be lose. In order to keep partially variance and have a relatively higher accuracy, we kept adding and subtracting the number of principal components from 1-17. Finally, we decided to use 12 principle components. The highest two explained variances are 0.50152429, 0.13241758.

Part 7: Models Building

Since the dependent variable is dur_stay, which is a binary variable after data manipulation. As a result, we build 6 classification models and try to find the best fitting model.

1) Logistic Regression

In regression analysis, logistic regression is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, these are represented by an indicator variable, where the two values are labeled "0" and "1".

- Parameters:

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=0, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

Figure 12: Logistic Regression Parameters

- Confusion Matrix:

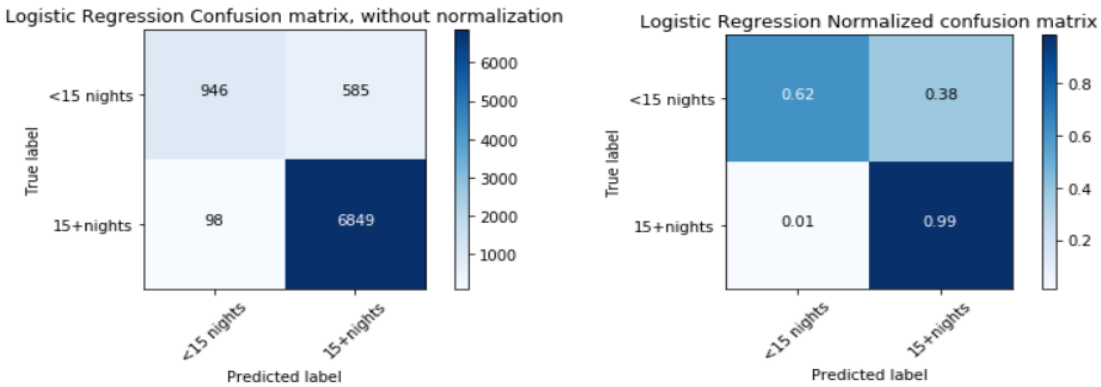

Figure 13: Logistic Regression Confusion Matrix

2) Decision Tree Classification

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

- Parameters:

```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best')
```

Figure 14: Decision Tree Parameters

- Confusion Matrix:

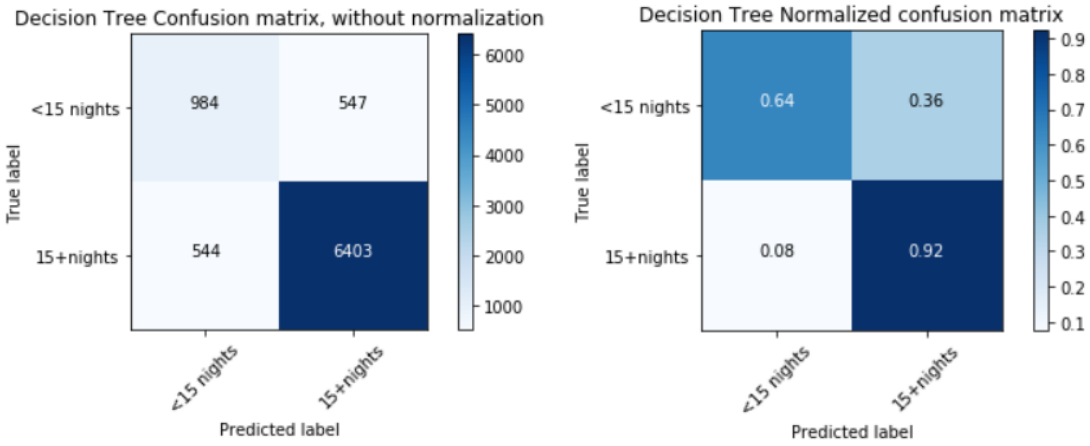

Figure 15: Decision Tree Confusion Matrix

3) Random Forest Classification

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn with replacement if bootstrap=True (default).

- Parameters:

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='entropy',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=500, n_jobs=1,
            oob_score=False, random_state=0, verbose=0, warm_start=False)
```

Figure 16: Random Forest Parameters

- Confusion Matrix:

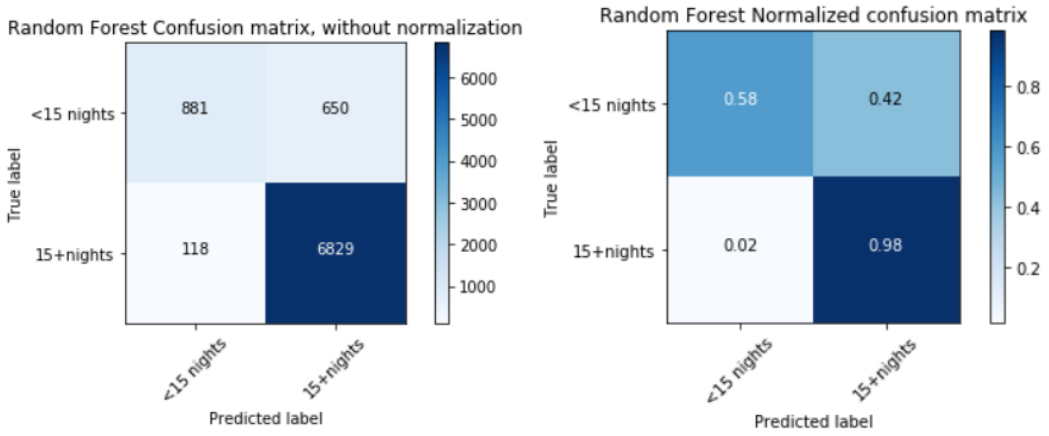

Figure 17: Random Forest Confusion Matrix

4) K-Nearest Neighbors (KNN)

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

- Parameters:

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=1, n_neighbors=4, p=2,
           weights='uniform')
```

Figure 18: K-Nearest Neighbors Parameters

- Confusion Matrix:

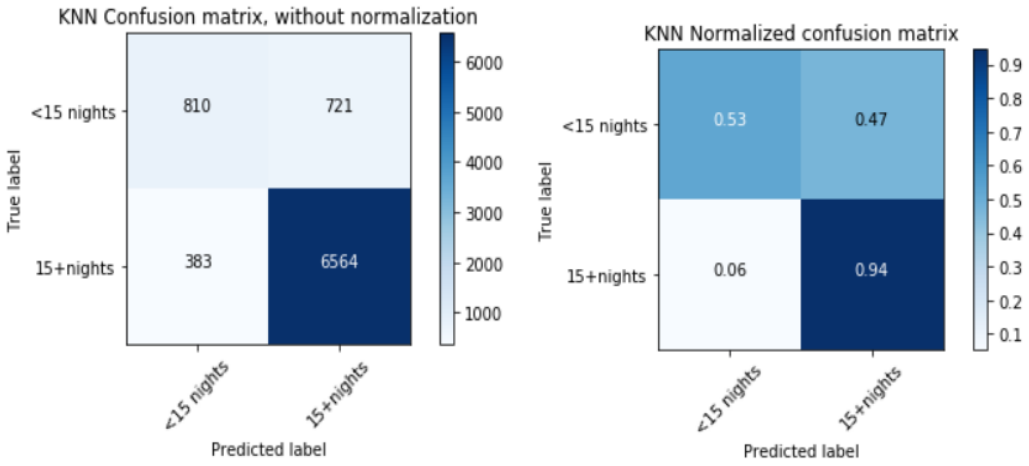

Figure 19: K-Nearest Neighbors Confusion Matrix

5) Naïve Bayes

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

- Parameters:

```
GaussianNB(priors=None)
```

Figure 20: Naïve Bayes Parameters

- Confusion Matrix:

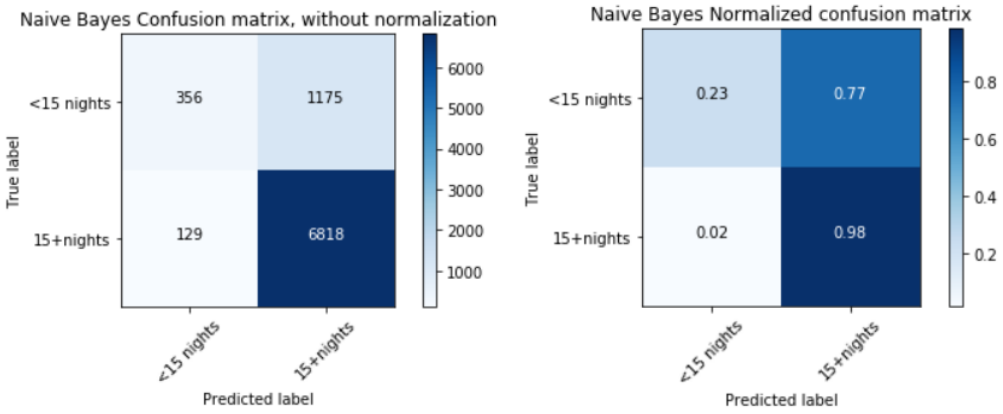

Figure 21: Naïve Bayes Confusion Matrix

6) Kernel Support Vector Machine (SVM)

Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary kernel classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

- Parameters:

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma='auto', kernel='rbf',
  max_iter=-1, probability=False, random_state=0, shrinking=True,
  tol=0.001, verbose=False)
```

Figure 22: Kernel Support Vector Machine Parameters

- Confusion Matrix:

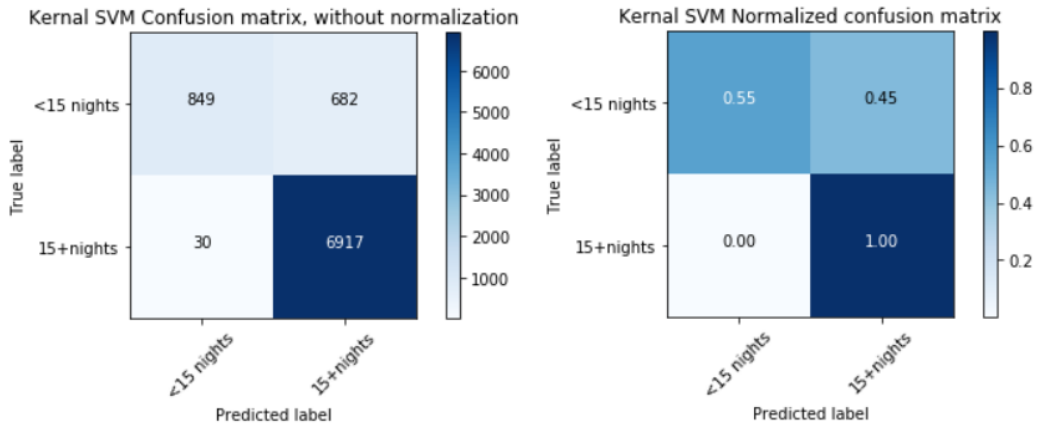

Figure 23: Kernel Support Vector Machine Confusion Matrix

Part 8: Dashboard

To achieve our goal of embedding the model into reality manipulation, we decided to design a deliverable dashboard to reveal and conclude our results. The dashboard is designed by using Python Dash package. In addition, the dashboard could be improved with better data construction and external_css frame for further use. By using this dashboard, our client can easily to get the information they are seeking for on country level in different years. Moreover, the clients can develop and change their policies associated with new input tourism data into our model.

Construction of the dashboard: The first thing to apply our model and results to our dashboard is to add longitude and latitude to our predicted datasets. By doing this, the countries' results could transform to map data on our first dashboard graph. As is shown in the figure below, by hovering the different data points on the map, the dashboard can reveal the details of the original value of the dur_stay, predicted value of dur_stay and the probability that certain group of people would stay more than 15 days. The dashboard also gives descriptive graph on country level, based on the prosperity indexes and original country data information. For instance, France is set as an example to show that there is over 99% probability for those who come from France may stay more than 15 days illegally in London.

Figure 24: Dashboard Overview

Chapter 4: Analysis and Results

Study Question 1

The first question addressed is that whether the potential tourist staying 15+ days in Great London Area have higher risk of immigration. We explored this question by applying both supervised and unsupervised methods, including Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors (KNN), Naïve Bayes, and Kernel Support Vector Machine (SVM). After comparing the performances among different models on this dataset by looking at Confusion Matrix, we select Logistic Regression model predicting the probability of immigration for those who stayed more than 15 days in London. As is the confusion matrix shown, below, there is 95% probability that we predict accurately for those who spent 15+ days in London actually staying 15+ days in London. For those who planned to stay less than 15 days actually stay less than 15 days, we have 54% probability to predict correctly. It is evident that the precision in general is 92% as shown in the table which is attained by applying sklearn.metrics package. All the figure mentioned above indicate that the potential tourist staying more than 15 days in Great London Area have higher risk of immigration. After applying confusion matrix, the 0.95 ROC curve area shown below has a clearer vision on how accurate the prediction on the probability of staying more than 15 days is.

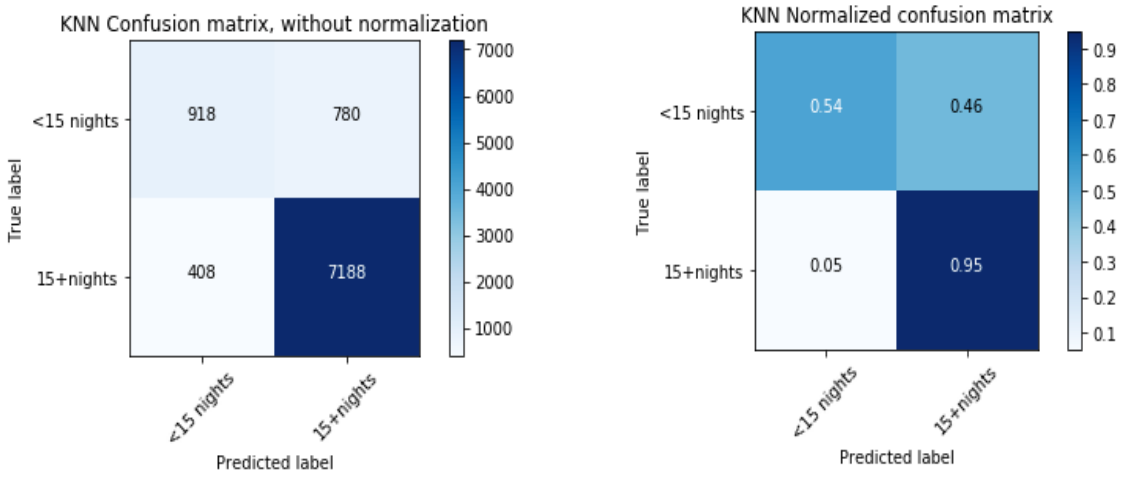

Figure 13: Logistic Regression Confusion Matrix

	Precision	Recall	F1-score	Support
< 15 nights	0.90	0.64	0.75	1698
≥ 15 nights	0.92	0.98	0.95	7596
Avg /Total	0.92	0.92	0.92	9294

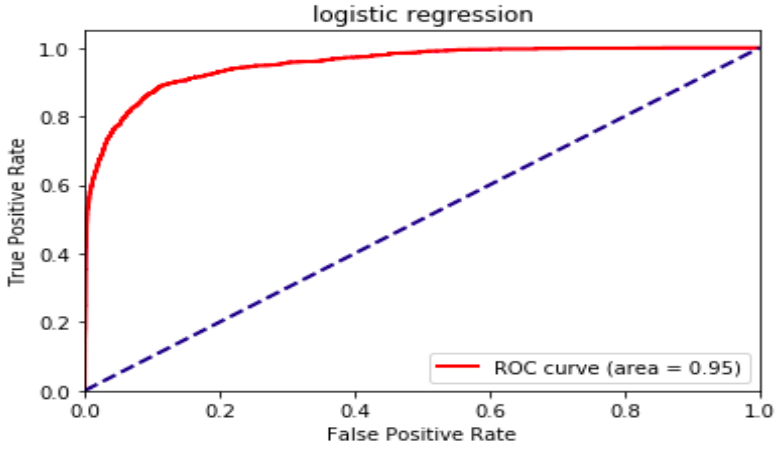

Figure 25: Logistic Regression ROC Curve

Study Question 2

The second question addressed is that Whether visitors coming from certain countries have higher probability to stay 15+ days in Great London Area.

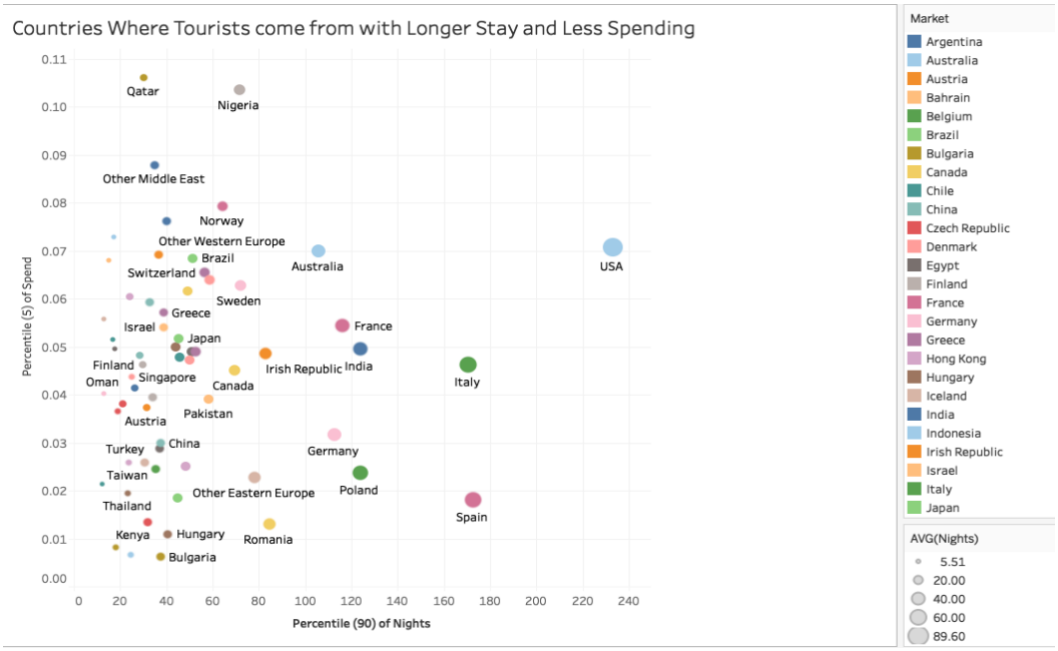

Figure 26: Scatter Plot of Average Nights Versus Average Spending on Different Countries

Besides predicting the probabilities of illegal staying for foreign tourists in model, this question can be explored in a relatively direct method. We plot the scatter figure with the size of circle representing the average nights they stay in London and colors representing different markets. The bottom of the figure is the area we are interested in. The bottom area represents the countries with the highest probability of illegal staying in London, such as Kenya, Spain, Romania, etc. In addition to those countries above, more detailed answers of countries whose visitors have higher probability to stay 15 more days, can also be found in our dashboard, that when certain country has higher probability, its color in dashboard will be more deep red.

Additional Questions

Besides above two main questions, we also interested in to find out the following topics/questions:

1. *What are the main factors that affect visitors spend 15+ days in Great London Area?*

   By using the macro and micro indicators, as well as other inputs that provided in the two data sets to answer this question, so that The City of London's Local Ministry and National Security officials could know which factors they need to pay more attention to.

   As the figure shown below, the economic development, business environment and PI index of the country tourists come from, tourists' spending in London and the origin country these tourists come from are the utmost important factor impacting whether these tourists will stay in London illegally or not.

```
[(0.00355, 'econ_std'),
 (0.00355, 'busi_std'),
 (0.0029, 'spend_std'),
 (0.0027, 'market'),
 (0.00267, 'PI_std'),
 (0.00266, 'visits_std'),
 (0.00264, 'purpose'),
 (0.00264, 'nights_std'),
 (0.00255, 'gove_std'),
 (0.00253, 'year'),
 (0.0025, 'mode'),
 (0.0016, 'heal_std'),
 (0.00081, 'soci_std'),
 (0.00057, 'safe_std'),
 (0.00049, 'envi_std'),
 (0.00041, 'educ_std'),
 (0.00031, 'pers_std')]
```

Figure 27: Feature Selection on the Importance of Variable

2. *How the international visitors spending pattern affects the different fields of prosperity?*

By analysis of countries' spending and their potential purchasing power, the City of London's Local Ministry and National Security officials could attract more visitors (who spend more money in London) while limit visitors (who are likely to illegal immigrate to London).

Before we explore the spending pattern of tourists, we first plot a spending map of different countries. By looking at *Figure 28: Average Spend by Country*, the darker red the area is, the more tourists coming from this country spend in London. Obviously, the United States and most of the European countries play an important role that enhancing London tourism.

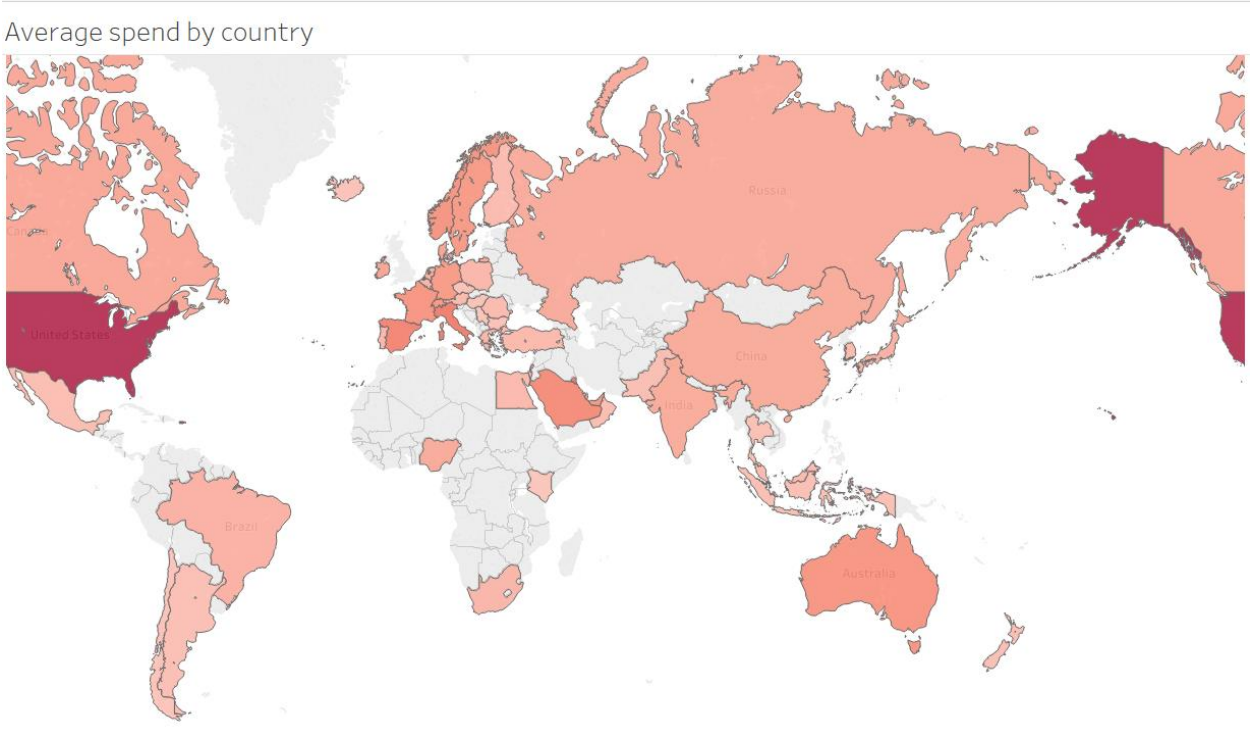

Figure 28: Average Spend by Country

We then plot the scatter picture of night_per_visit (X axis) Versus spend_per_visit (Y axis). Most of spots are centered in the bottom of the plot. We are interest in those outliers representing those tourists who stayed longer in London while spending extremely less during their stays.

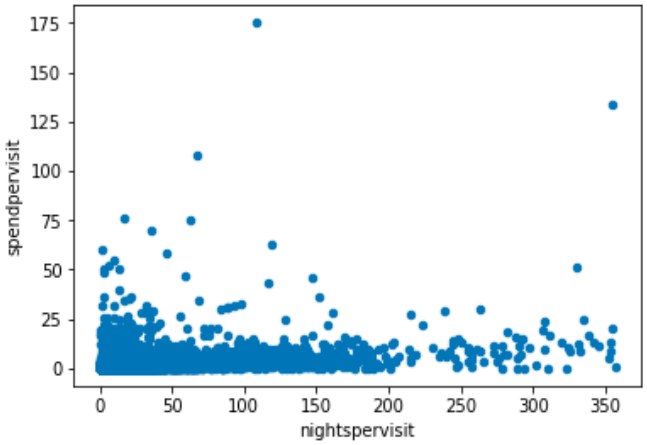

Figure 10: Scatter Plot between Nights per Visit & Spend per Visit

Tourists from Spain, United Arab Emirates and some Eastern European countries have the higher risk of staying illegally in London than those tourists from other countries.

3. *What are the changes of international tourism trend?*

By using the macro and micro indicators, as well as other inputs that provided in the two data sets to answer this question, so that The City of London's Local Ministry and National Security officials could better understand the international tourism trend, performance, and thus adjusting policies to make improvement. Figure showed followed is the international tourism trend between 2007 to 2017: the length of each bar means the total worldwide visits to London of each year, and the color means the average spend per

night of each year. As we can see, both visits and spend are slightly decreased from 2007 to 2009, and then they are all increased from 2009 to 2017.

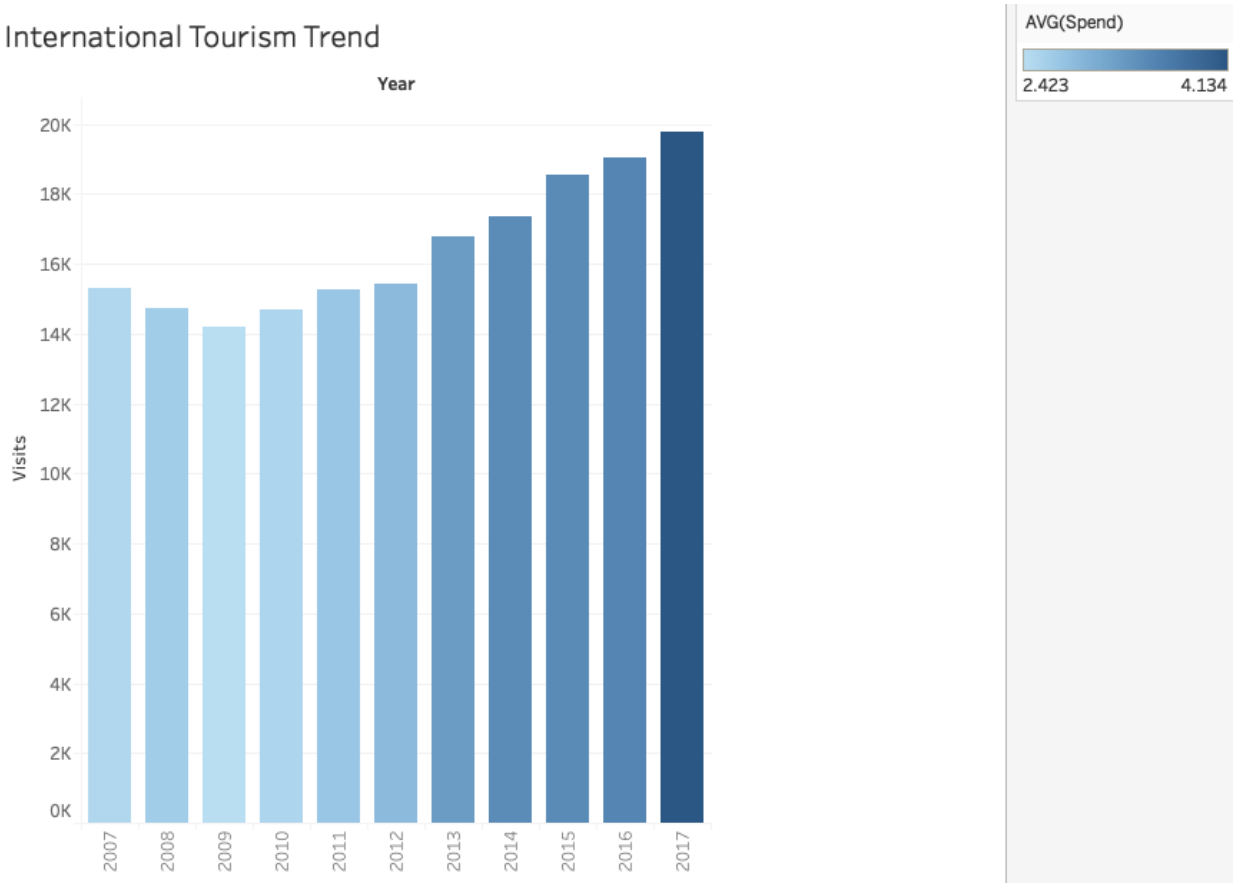

Figure 29: International Tourism Visits & Spend Trend

Chapter 5: Conclusion

The purpose of this paper was to answer the first two study questions and we will continue our analysis to answer the flowing three study questions.

1. ***Whether the potential tourist staying 15+ days in Great London Area have higher risk of immigration?***
2. ***Whether visitors coming from certain countries have higher probability to stay 15+ days in Great London Area?***
3. *What are the main factors that affect visitors spend 15+ days in Great London Area?*
4. *How the international visitors spending pattern affects the different fields of prosperity?*
5. *What are the changes of international tourism trend?*

With extensive data cleaning, wrangling, and employing various tools such as Tableau, Python, and Dashboard, valuable insights were discovered, interested observations were formed, and supervised and unsupervised predictivity models were built.

Among the six models, Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors (KNN), Naïve Bayes, and Kernel Support Vector Machine (SVM), Logistic Regression is our best model to predict whether a new visitor will stay 15+ days in Great London Area (GLA) based on the accuracy and validations scores.

Therefore, our predictive model would be delivered to The City of London's Local Ministry and National Security officials to serve as a long-term analytics maturity model to advise existing security practices and the two institutions could use this project as a way to provide a foundation predictively solution, and then translate a multi-year contract to maintain and expand this capability to other cities.

Chapter 6: Future Work

As the main problem of this practicum, safety issue of London standing in the aspect of tourism is the most significant factor. However, not the only thing we should consider. During our exploration, we found some countries stay long but spend less than the average, such as Spain. In the future work, figuring out what reason hinder their consumption and stimulating London economics development especially on international tourist's consumption by tracking foreign tourists spending pattern can be implemented.

Besides a deeper understanding of current data, it is also important to test the accuracy of our model with the new data coming in. Therefore, the optimal model we select now to predict the probability may change. As a result, exploration more on this direction in our future work is also expected.

Last but not least, our dashboard can be improved and customized to show more visualization, directly fulfilling expectation of all audience. The improved dashboard visualization can become a useful tool for our client to perform a detailed analysis on international tourism analysis, especially in the point of immigration safety, and for different government or organization to analyze different city's tourism pattern.

References

https://en.wikipedia.org/wiki/Pandas_(software)

https://en.wikipedia.org/wiki/NumPy

https://www.datacamp.com/community/tutorials/learn-build-dash-python

https://www.techopedia.com/definition/948/encoding

https://medium.com/@contactsunny/label-encoder-vs-one-hot-encoder-in-machine-learning-3fc273365621

https://towardsdatascience.com/why-you-should-try-mean-encoding-17057262cd0

https://en.wikipedia.org/wiki/Feature_engineering

https://en.wikipedia.org/wiki/Principal_component_analysis

https://www.londonandpartners.com/media-centre/insights-and-statistics

https://data.london.gov.uk/dataset/number-international-visitors-london

https://prosperitysite.s3accelerate.amazonaws.com/3515/1187/1128/Legatum_Prosperity_Index_2017.pdf

https://www.prosperity.com/

https://en.wikipedia.org/wiki/London_Tourist_Board

https://en.wikipedia.org/wiki/Legatum_Institute