# 4.1 Implementation

The implementation of this study, covers deriving an insight from the dataset used and creating a dashboard using power bi. The process starts by launching anaconda, which has Jupyter notebook installed in it. Jupyter note book is a text editor design to write python codes.

To begin the experiments, it starts by importing the library needed. The first library imported is pandas, pandas is used for data manipulation. The pandas were used in importing the dataset. Numpy was used for arithmetic computation. The seaborn and matplotlib works together to plot the charts needed. After the library has been imported, the next step is to use pandas to import the dataset and assign the dataset to a data frame. The dataframe is a variable name assigned to the dataset. The dataframe can be previewed and cleaned. Data cleaning involves checking if there are any bias values in the dataset. This can be checked using a python command and can be seen in the appendix section.

**4.2 Data Cleaning**

From the raw dataset. It shows that in column sex, visits, nights and sample contains null values which cannot be empty cells. Because we have a large number of dataset samples, deleting the row with any null values would not have much impact on the result of the analysis. The entire dataset has 425,024 samples with 14 columns before data cleaning. After removing the null values the number of samples decreased to 410,573 samples. A thorough checking of the values of each columns showed that columns like sex, age and duration of visits contains some unwanted inputs. Sex contains “Don’t know and ‘ ‘ ,an empty quote” , age contains “D/K” as values while duration of visits contains “Stay not Known”.

After the unwanted values were removed the number of samples decreased to 404,775.

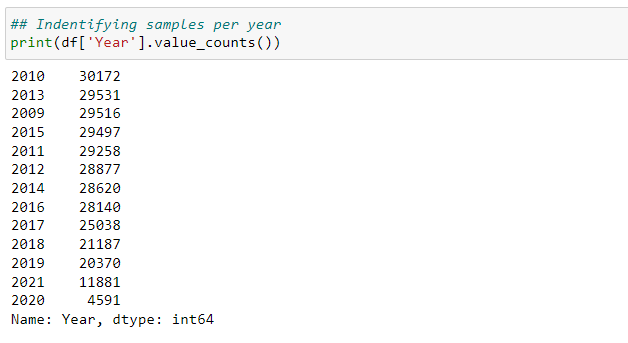
The cleaned dataset was exported to a new data sample and then re-imported and begin the process of removing outliers.

**4.3 Removing outliers**

Outliers, or extreme values that are significantly different from the majority of data points, can have a significant impact on the analysis and interpretation of a dataset. They can skew statistical measures such as the mean and standard deviation, and can also affect the results of statistical tests and models. Outliers can also lead to errors in data visualization, such as distorting the scales on a graph. It is important to identify and handle outliers appropriately in order to accurately understand and interpret the data. A box plot of some columns showed that there are outliers in them. The outliers were removed and this reduced the dataset to 316,678 samples.

**4.4 Identifying samples per year**

From the entire dataset. 2010 has the highest number of data samples, while 2020 has the least as shown in fig 1 below



*Figure 1:Dataset samples per year*

# 4.2 Analysis Discussion

The insight derived is as a response to some research questions, such as;

1. What are the insights related to the travelers as regards their country, night spent, mode of travel and package of travel?
2. What necessitates travelling?
3. Does age groups affects travelling?

In other to make the analysis easier , and for proper insight generation. I concluded to segment the data. Segmentation helps me to break down all the data into smaller chunks to simplify the complexity of the analysis. The first segmentation done is breaking the dataset into two categories, using the ukos (where\_contact\_lives) variable which identifies where contacts lives. With this, I now have dataset of travelers who lived in the UK and the travelers who lived in the overseas.

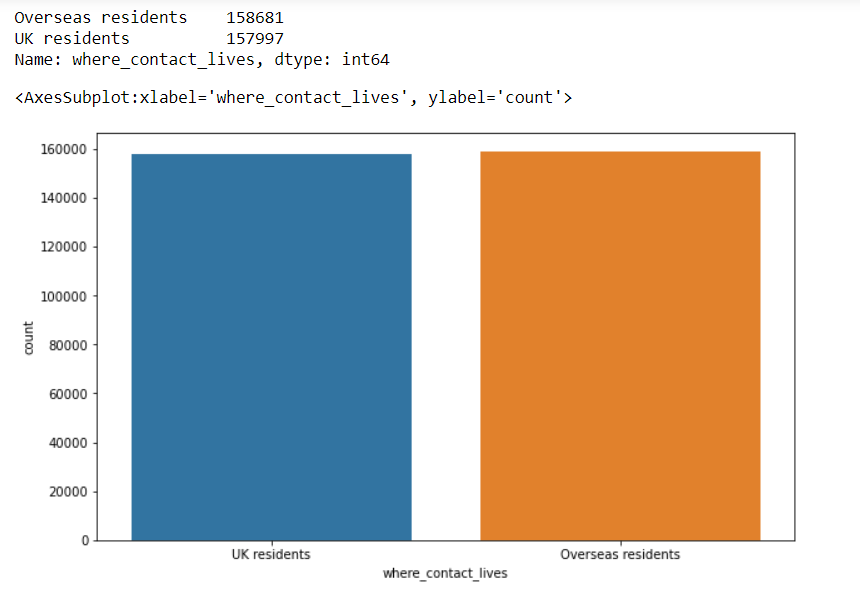


Figure 2: Where contact lives

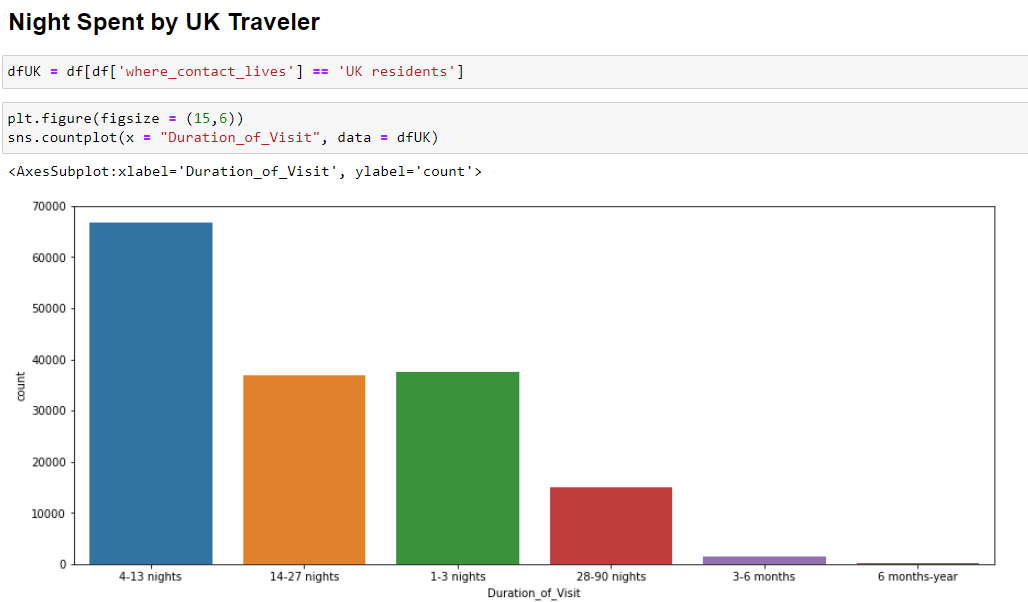
Fig 2 shows that from the overall dataset we have more oversea travelers than the UK travelers. Going further with the classification to identify what necessitates travelling, I considered identifying how many nights spend between both travelers. This process involves comparing the result from the UK and the Overseas travelers to identify the similarities and differences and detect the insights. 

Figure 3: Night spent by UK Traveler

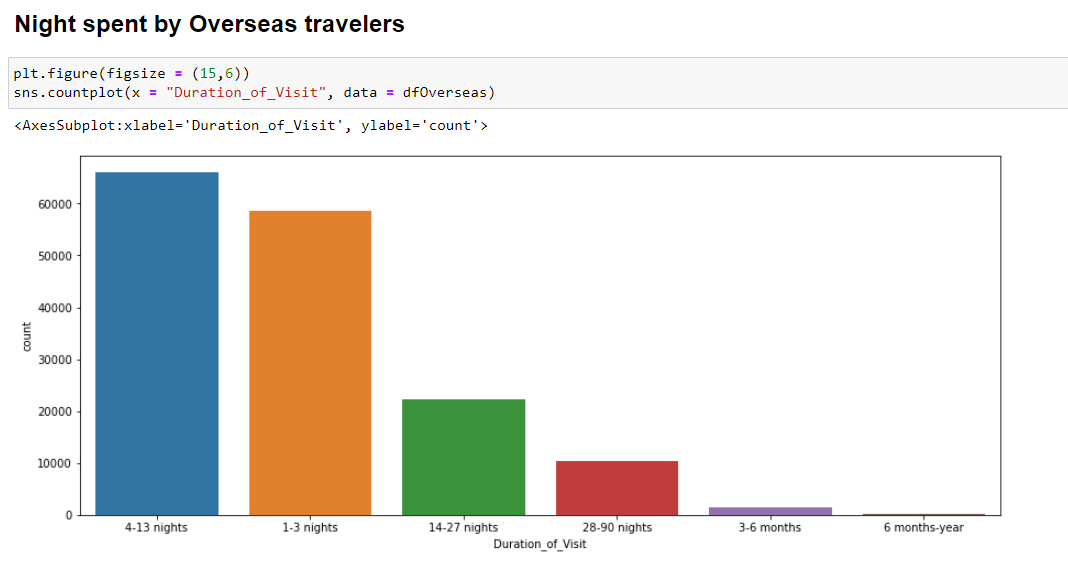


Figure 4: Night spent by Overseas travelers

The bar chart above shows of all the night available, travelers who lives in the UK and overseas travelers spent 4 to 13 nights more ,the number of UK travelers who spent 14-27 nights and 1-3 nights hovers around the same mean, while the margin between oversea travelers who 1-3 nights more and 14-27 nights is so wide with 1-3 nights more than 14-27 nights. I focused more on these because of its larger population, to identify what necessitate their travelling.

More segmentation was done, where the traveler that spent between 4 to 13 nights was further classified to identify the purpose of travelling.

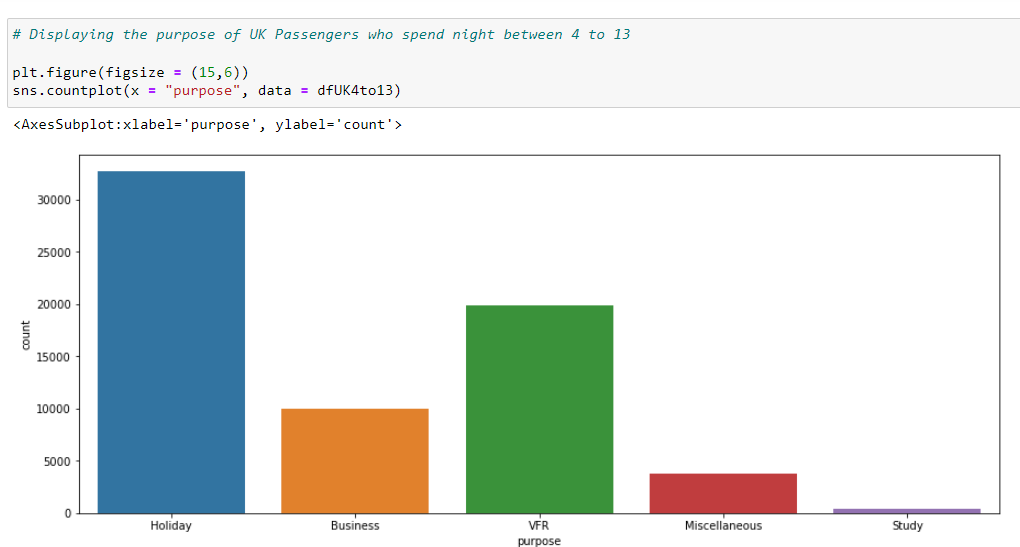


Figure 5: Purpose of passengers who spent 4 – 13 night that stays in the UK

As seen in fig 5, for travelers that stay in the UK, the major reason for travelling is for holiday, the second major reason is visitation.

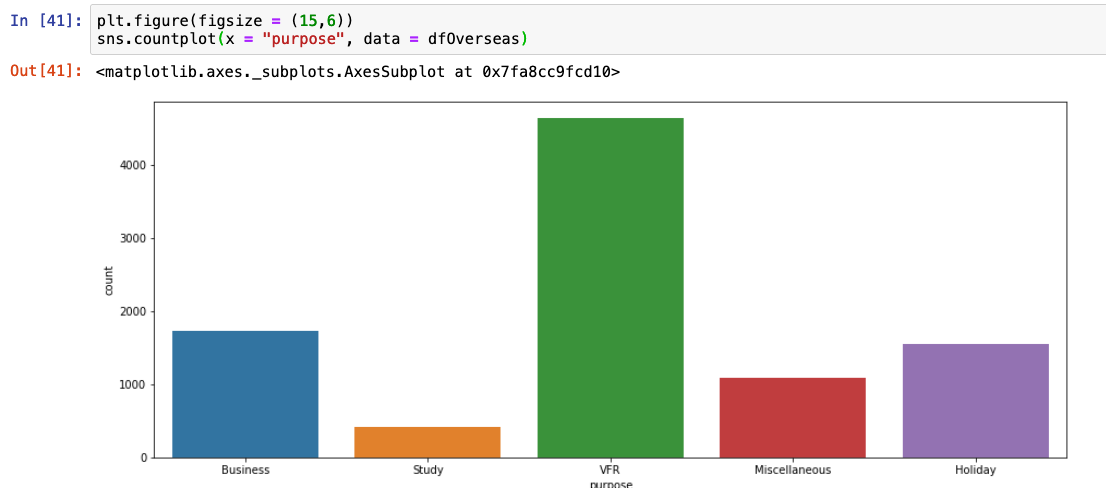


Figure 6: Purpose of passengers who spent 4 – 13 night that stays in the Overseas

As seen in fig 6, it shows that the major purpose of travelling by passengers that spend between 4 to 13 night in the overseas are traveling for visitation purposes while the second major reason is for business. With this information we can conclude that the visitation is one of the major purpose of the travelling either those that lived in the UK or overseas .

Going further, the dataset was grouped by age. This is done to see the kinds of people who travelled for visitation. From the segmentation its shows that traveler between the age 25 – 65 mostly go for visitation. This is very relatable as the age categories covers university graduates who might want to visit home when school are on break or adult who are travelling back home for visitation.

Figure 7: Age groups of passengers who spend 4 to 13 night in the overseas

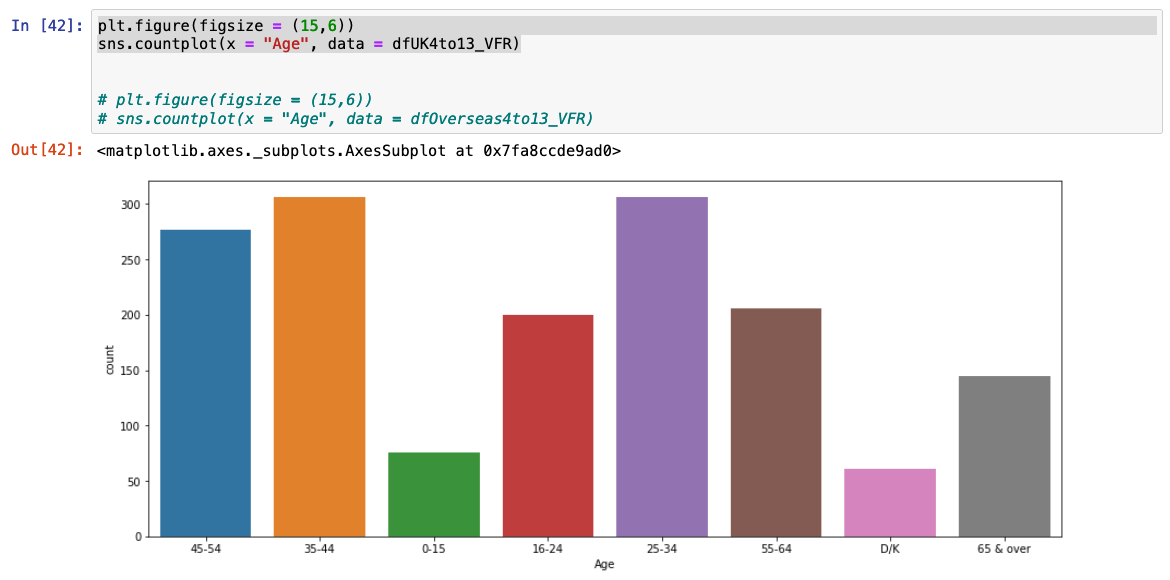


Figure 8: Age groups of travelers who spend 4 to 13 night in the UK.

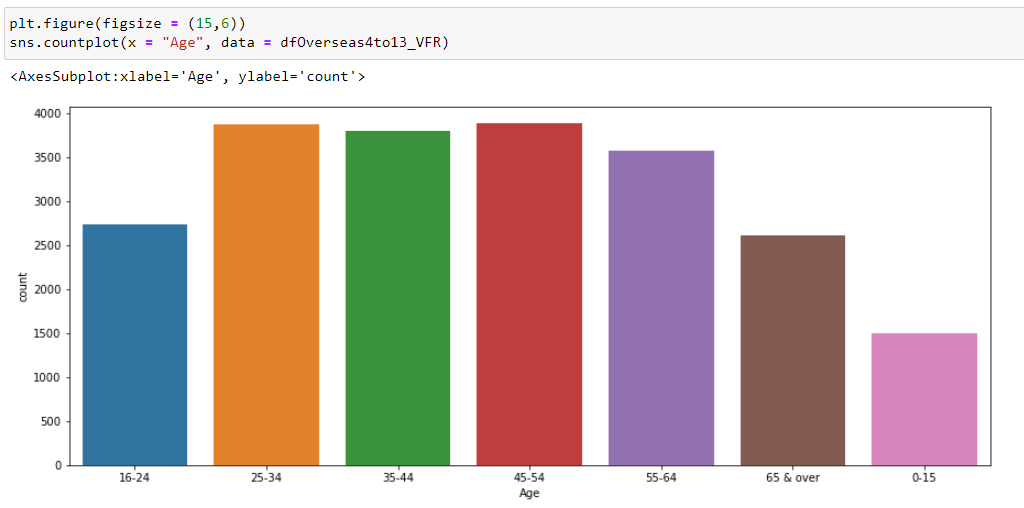


Figure 9: Age groups of travelers who spend 4 to 13 night overseas

From figure 8 and 9 above, there are similarities in the age group of the passengers travelling. From age 25 – 55, we have larger number of passengers in that age group, considering that we have more passengers who travel for visitation and holidays, we can conclude that most passengers between the age of 25 – 55 are travelling for holidays and visitation purpose, more reason passengers are spending 1 to 13 nights during their visits.

Categorizing the dataset into years and identifying the number of visits yearly, as seen in fig below, on a yearly bases, starting from 2009 till 2019, we can see an increase in the number of passengers and a huge drop in the year 2020, this is basically caused by COVID-19 lock-down and in the year 2021, it started increasing gradually.

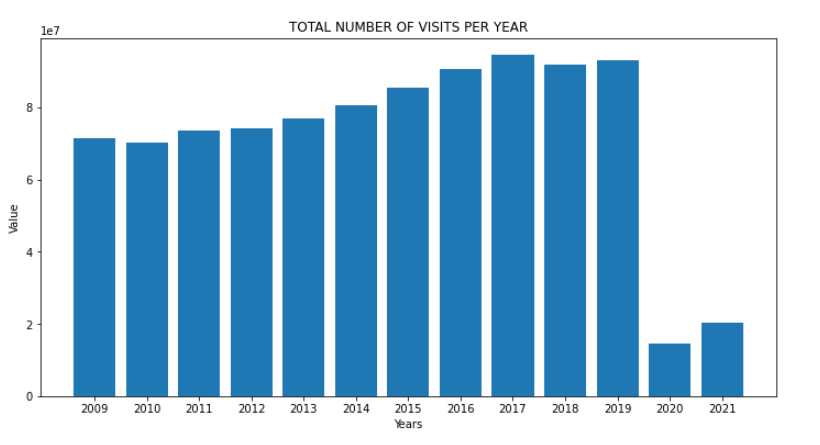


Figure 10: Total number of visits per year

Same analysis is applicable to the number of nights spent as explained above.

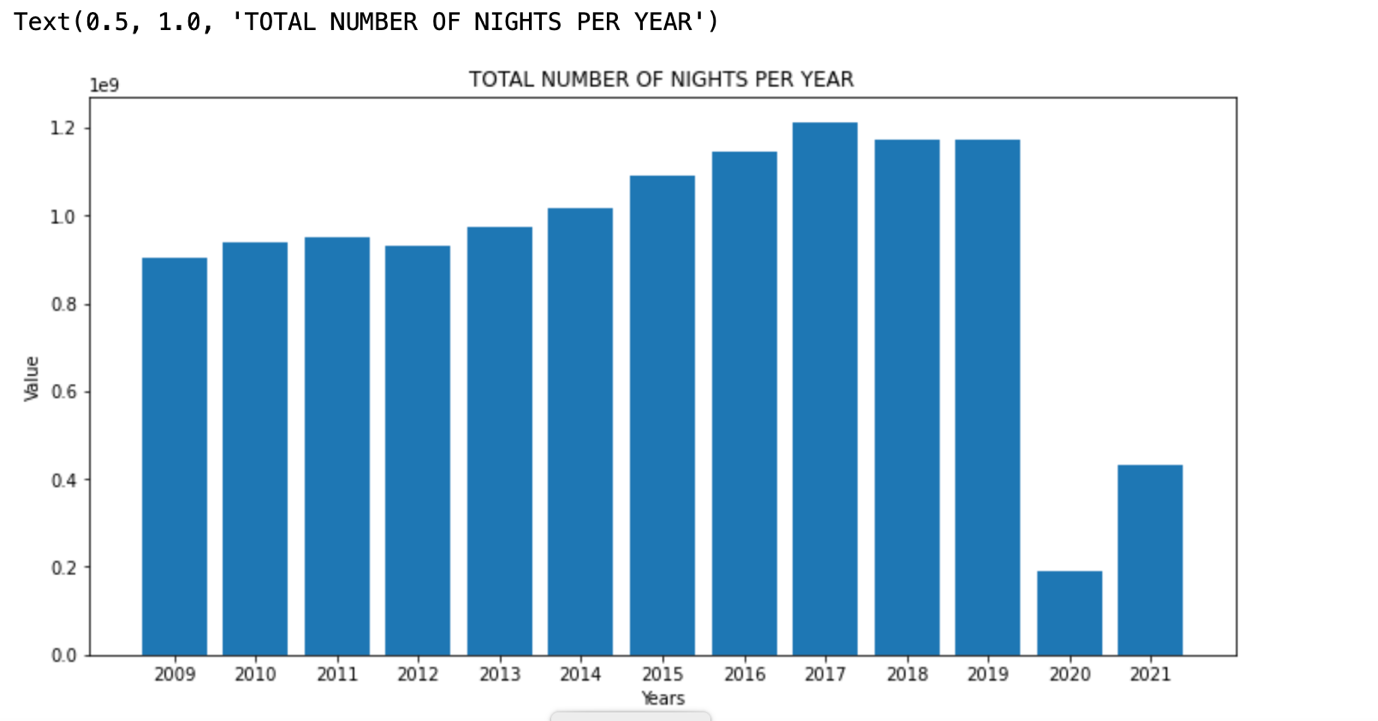
****

Figure 11: Total number of nights per year

From the entire dataset ranging from year 2009 to 2021, It shows that place of residence for overseas residents or of visit for UK residents with the highest number are from France, with 35,021 passengers while Cyprus which is the list has 690 passengers.



Figure 12: Country with the highest number of travelers and Least number of travelers

# 4.3 Time series Forecasting Discussion

A time series is a collection of data points that are ordered. It includes techniques for deriving useful statistics and other aspects of time series data through analysis. This is the foundation for predicting future events by evaluating historical data. An ARIMA model was employed as the forecasting algorithm. A statistical analysis model called an autoregressive integrated moving average, or ARIMA, uses time series data to either better comprehend the data set or forecast future trends. (Mingda, 2018).Time series analysis is being implemented in different field or organization to identify why some specific variables changes over some period of time. It is efficient for stationary dataset i.e., data that the values are prone to changes.Time series analysis is frequently used in the financial, economic, and retail sectors since sales and currencies are continually changing.It is one of the methods used in predictive analytics that shows potential changes in the data, such as cyclical or seasonal activity, which provides a clear understanding of the data factors and aids in improved forecasting.

For this work, time series analysis would be implemented on different variables which are the amount of expenditures of the travelers, the number of nights spent by the travelers and the number of visits of the passengers for the next 5 years. The first experiment done was forecasting the amount of expenditure for the next 5 years.

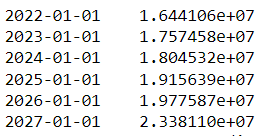


Figure 13:Forecast of Total Amount spent

After the dataset has been imported, it needs to be checked if its stationary or not. Time series analysis relies heavily on the concept of stationarity, which has a major impact on how the data is interpreted and forecasted. Most time series models make the assumption that every point is independent of every other point for forecasting or predicting the future. The dataset of historical instances being stationary is the strongest indicator of this. The statistical characteristics of a system must remain constant across time for data to be considered stationary. This does not imply that the values for each data point must be identical, but rather that the general pattern of the data should not change.

This can be identified using the P-Values of ADF test, which needs to be small as possible, a. higher P-Values shows the dataset is not stationary while a lower P-Values shows the dataset is stationary.

**Augmented Dickey–Fuller Test**

The Augmented Dickey-Fuller Test is used to determine if time-series data is stationary or not. Similar to a t-test, we set a significance level before the test and make conclusions on the hypothesis based on the resulting p-value.

**Null hypothesis and alternative hypothesis for ADF test**

Null Hypothesis: The data is not stationary.

Alternative Hypothesis: The data is stationary.

For the data to be stationary (ie. reject the null hypothesis), the ADF test should have:

p-value <= significance level (0.05).

If the p-value is greater than the significance level then we can say that it is likely that the data is not stationary.

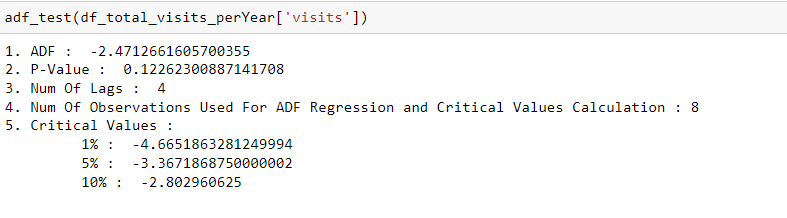


Figure 14: P - Values

From the result above, the P-Value is greater than 0.05. So the null hypothesis accepted, this means that the dataset is not stationary.

A best fit model search was done using auto arima stepwise algorithm, ARIMA(0,2,1) was suggested.

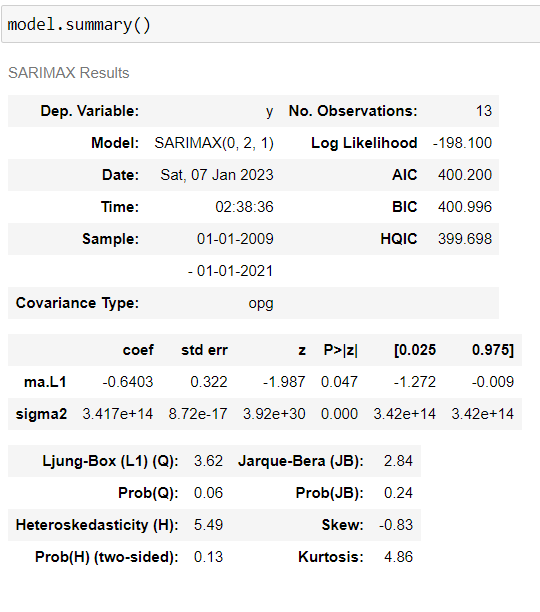


Figure 14 Suggested model

Prediction made using the arima model above generated some negative values. This suggests that the model is not good enough to use as a predictor. A sarimax model of order(0,1,1) was later fitted. This showed to be a promising model as the plot of the predictions and the real data look similar.

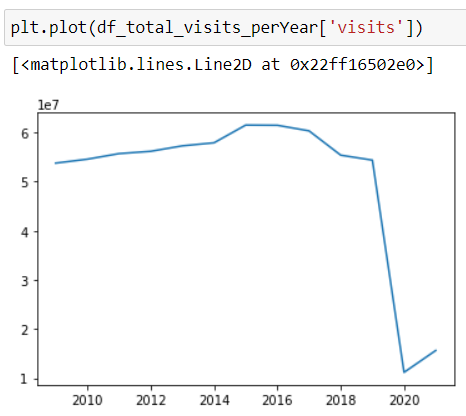


Fig 15:Plot of total visits per year

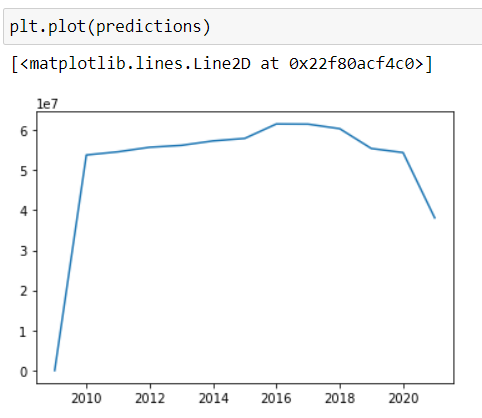


Fig 16:Plot of predicted values of total visits per year

**Forecasting**

Forecasting results for expenditure for the next 5 years (Year 2022 to 2027), it is observed that year 2023 will have the minimum expenditure amount of over 13 billion, anything less is not seen while 2027 is expecting over 16 billion

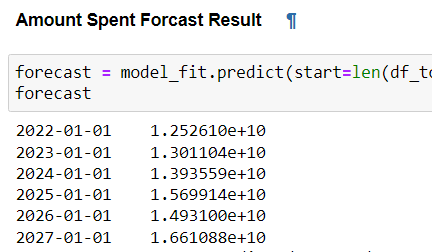


Figure 17: Expenditure Forecast Results

The same approach is being implemented in the total of visits per year, to forecast the number of visits for the next 5 years. There will be an increment after the year 2021. From the analysis being implemented, a static value was derived after the year 2020, which shows that the population of travelers would increase after the year 2022 and that would affect the total number of nights spent of the travelers to increase.

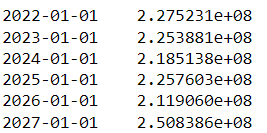


Figure 18: Forecast for Visits

Analyzing the number of nights spent using the SARIMAX model to forecast the number of night spent in the next 5 years. It shows that there is an increment after the year 2021. From the analysis being implemented, a static value was derived after the year 2020, which shows that the population of travelers would increase after the year 2022 , have ups and downs through the year 2023 to 2027, with 2027 estimate of over 250 million trevellers

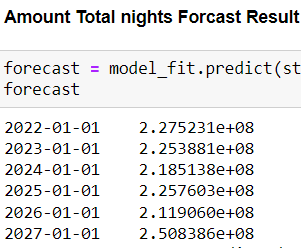


Figure 19: Forecast for Nights spent

# 4.4 Machine Learning Model Fitting

Machine learning is a method of teaching computers to learn and make decisions based on data, without explicitly programming them. It involves feeding a computer system a large amount of data, and allowing the system to use statistical analysis to identify patterns and make decisions. There are many different techniques and approaches to machine learning, including decision trees, support vector machines, and neural networks. These techniques can be used for a wide range of applications, such as image and speech recognition, natural language processing, and predictive modeling.

There are many different types of machine learning algorithms, but they can generally be grouped into one of four categories:

Supervised learning: This type of algorithm is trained on labeled data, where the correct output is provided for each example in the training set. The goal is to make predictions about unseen data based on the patterns that it learned from the training data. Examples of supervised learning algorithms include linear regression, support vector machines, and decision trees.

Unsupervised learning: This type of algorithm is not given any labeled training data and instead must find patterns and relationships in the data on its own. One common application of unsupervised learning is clustering, where the goal is to group similar examples together. Examples of unsupervised learning algorithms include k-means and hierarchical clustering.

Semi-supervised learning: This type of algorithm is trained on a dataset that is partially labeled and partially unlabeled. The goal is to use the labeled data to make predictions about the unlabeled data.

Reinforcement learning: This type of algorithm learns through trial and error, by taking actions in an environment and receiving rewards or penalties for those actions. The goal is to maximize the cumulative reward over time.

There are many variations and nuances within these categories, and different algorithms are better suited for different types of tasks and data. In this report, we will use the supervised machine learning to fit a model. The process of supervised machine learning generally involves the following steps:

Collect and preprocess the data: This includes gathering the relevant data, cleaning it to remove any errors or inconsistencies, and formatting it in a way that the algorithm can understand.

Split the data into a training set and a test set: The training set is used to train the model, while the test set is used to evaluate the model's performance.

Choose an appropriate model and training algorithm: There are many different models and algorithms to choose from, and the appropriate choice depends on the nature of the data and the task at hand.

Train the model: This involves feeding the training data to the model and using an optimization algorithm to adjust the model's parameters so that it can make predictions that are as accurate as possible.

Evaluate the model: The model's performance is evaluated using the test set. Common evaluation metrics include accuracy, precision, and recall.

Fine-tune the model: If the model's performance is not satisfactory, various techniques such as hyperparameter optimization can be used to improve the model.

Make predictions on new data: Once the model is trained and fine-tuned, it can be used to make predictions on new, unseen data.

For this work a supervised learning would be used.

**4.41 Division of data**

After thorough cleaning of the original dataset, the outliers were sieved out. The dataset is then divided into two subset, the Holiday package is the Y-data while the remaining data such as year, quarters, ages, number of nights, purpose are our X-data. Each of the X-data and Y-data is then split into Y-train, Y-test, X-train and X-test. The X data contains both categorical data and continuous data. The categorical data needs to be transformed into numeric format, dummies , that the computer can understand.

# 4.42 Model Fitting for Holiday Package

An initial accuracy test was for potential model was conducted using K-fold evaluation. Logistic regression and random forest classification have the highest initial accuracy ,as random forest have the highest accuracy, it was chosen. After fitting the a model with it, the model performed well with a model score of over 85% .

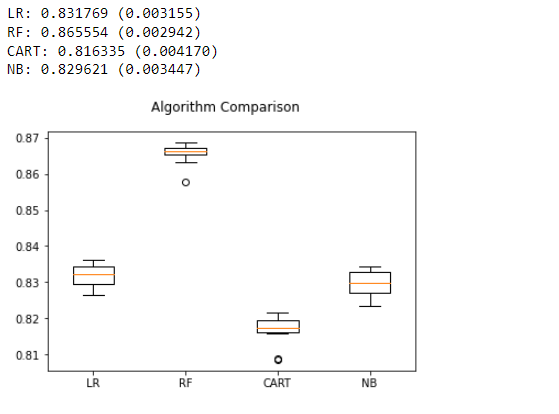


Fig 20: Initial accuracy test to choose a good model

4.43 Model accuracy of Holiday package.

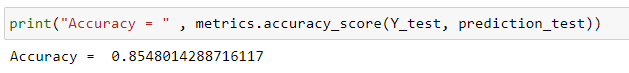


Fig 21: Model accuracy for Holiday Package

4.44 Feature importances of the Holiday package model

sklearn has a method called feature\_importances. Feature importances show the respective contribution of each of the feature and sub-feature to the accuracy of the model.

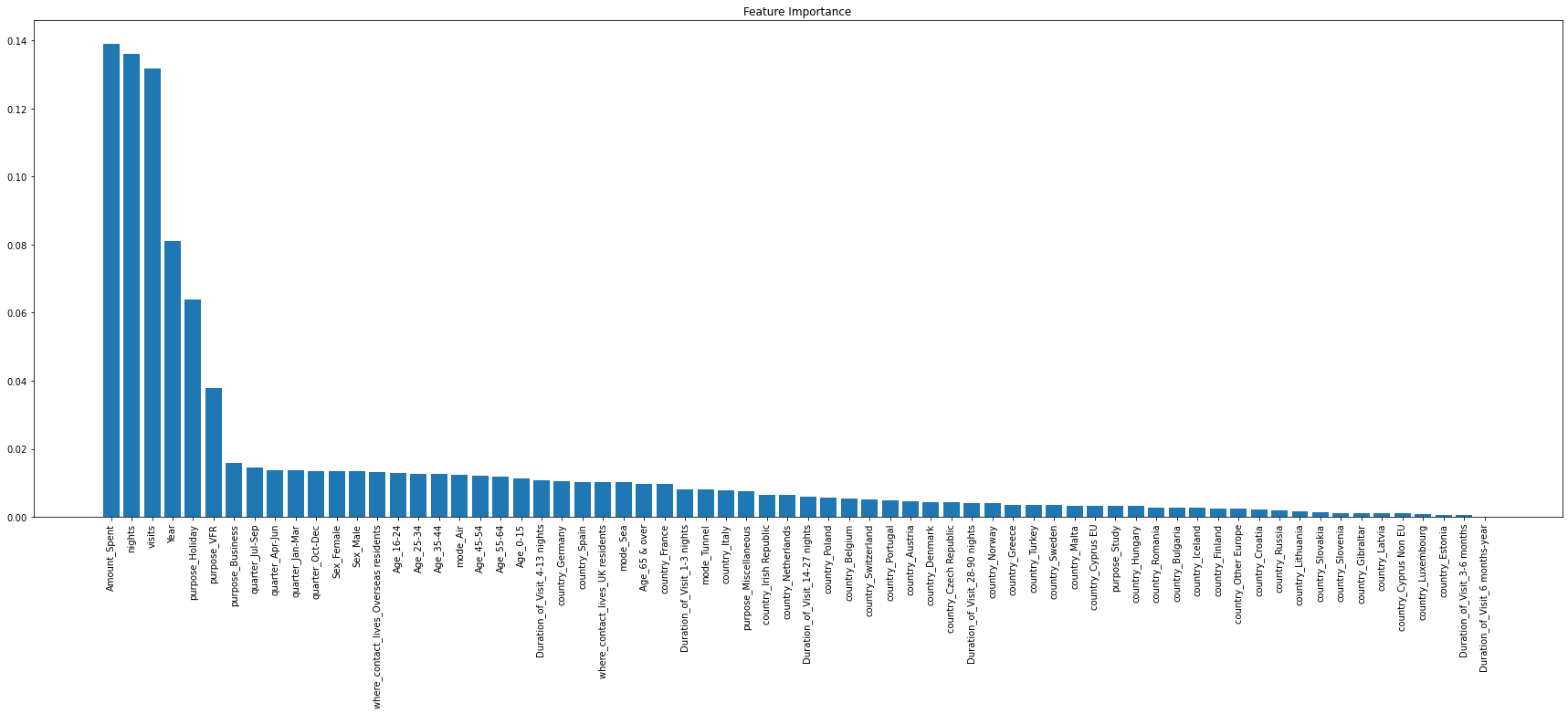


Fig 22: A barchart showing the features importances of each dependent variables to Holiday package in descending order

The best five features that contribute to the accuracy of the model are shown below.

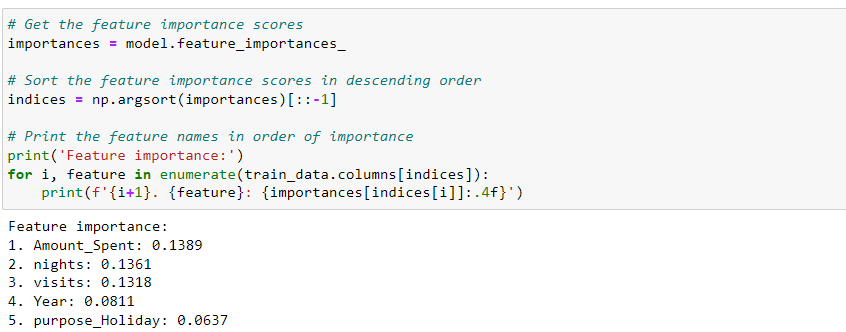


Fig 23: The best five features that contribute to the prediction of Holiday package

This showed that the amount spent contribute 13% to predicting Holiday package, nights 13%, visits 13%, year 8% and purpose\_Holiday, which is a sub-feature of holiday 6%, are the best predictor of the type of holiday package a traveler will have.

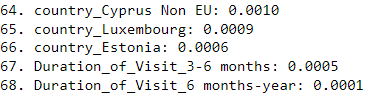


Fig 24: the least five features that contribute to prediction of Holiday package

4.45 Feature engineering

In an attempt to increase the accuracy of the machine learning model, two other models were fitted. A second model was fitted using all the features except duration of visit, the model have an accuracy of 85.1 % compared to the first mode with an accuracy of approximately 85.5% , 0.4% difference. This means that duration of visits(number of nights) is almost insignificant in predicting holiday package. The third model was fit omitting duration of visit and mode of transportation. It eventually have an accuracy of 84.1%, a 1.4% difference with the first model. Since both the second and third model didn’t increase the accuracy, we will stick to the first model.

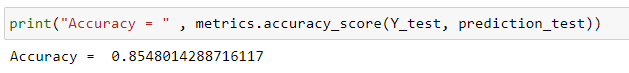


Fig 25: Model accuracy for first model using all features

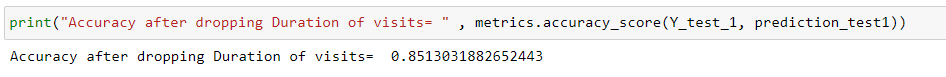


Fig 26: Model accuracy of the second model after dropping duration of visit

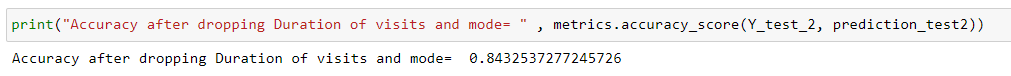


Fig 27: Model accuracy of the third model after dropping duration of visit and mode of transportation

4.5 Model fitting for Amount Spent(expend)

After fitting a model and making prediction with a time series ARIMA. We fitted a machine learning model considering the comparative advantages below.

Machine learning models and ARIMA (AutoRegressive Integrated Moving Average) models are both techniques used for time series forecasting. They both have their own strengths and limitations, and the appropriate one to use depends on the specific problem at hand.

One advantage of machine learning models is their ability to automatically learn complex patterns in data and make predictions based on those patterns. They can handle a large number of input features and can often make more accurate predictions than simpler models like ARIMA.

On the other hand, ARIMA models are simpler and easier to implement, especially for users who are not familiar with machine learning. They are also well-suited for forecasting when there is a clear trend or seasonality in the data.

In general, machine learning models may be a better choice when there is a large amount of data available and there is a need for high prediction accuracy. ARIMA models may be a good choice when the focus is on understanding the underlying patterns in the data rather than making the most accurate predictions possible.





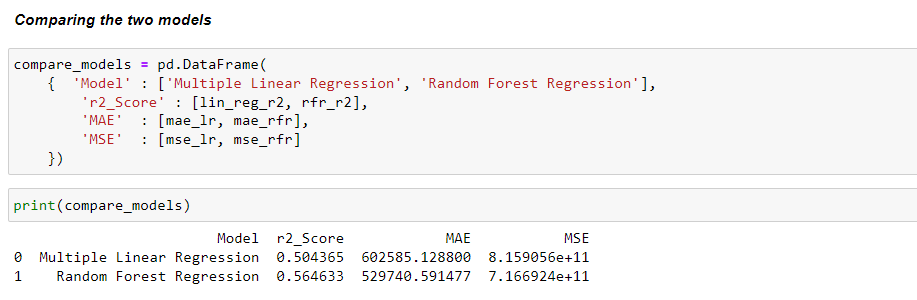


Fig 28: comparing model score

We can now see the score and error of Linear regression and Random forest classifier and compare them. Score of Random forest Regression is greater than Linear regression and error is also less. Thus, Random forest Regression will be the right choice for our model.

The r squared score of our machine learning model is 0.56. This means that only 56 percent variations in amount spent(expend) by travelers can be explained by all the features. This model is not good enough for prediction. In an attempt to make the model better , an effort was made to determine the relationship between some independent variables and amount spent as seen below. Pearson correlation coefficient was used for calculating the correlation coefficient between amount spent and other continuous variables like nights, visits and year.

While covariance was used to find extent of relationship between amount spent and other categorical variables like age and purpose as shown below.

4.51 Feature selection for Amount spent(expend)

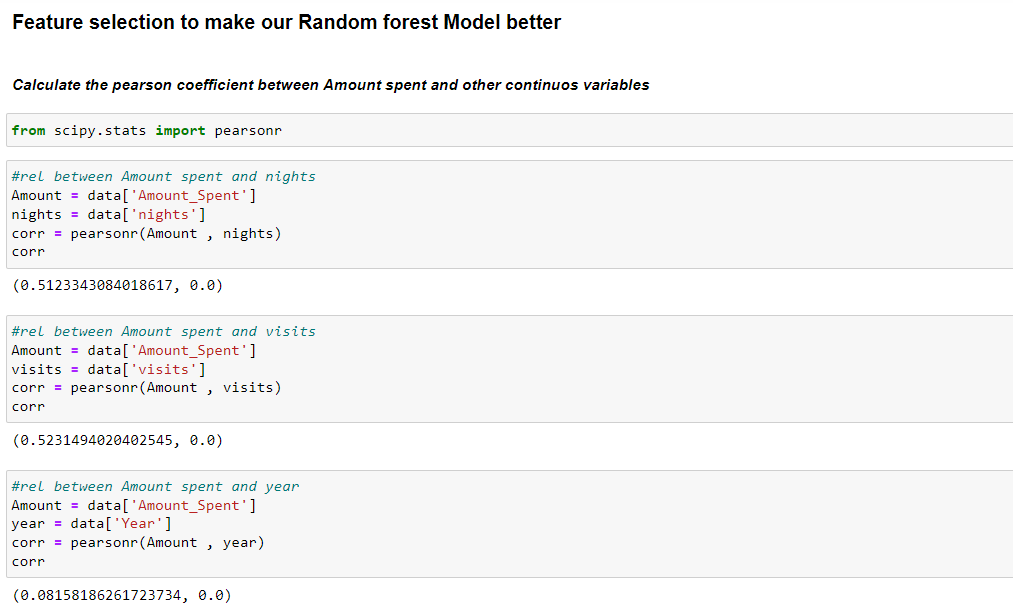


Fig 29: pearson correlation coefficient showing correlation between Amount spent and nights,visits and year

Result: Pearson correlation coefficient shows a high correlation between Amount spent and nights , visits and a weak correlation between amount and year.

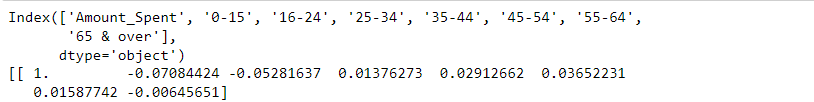


Fig 30: correlation matrix showing extent of relationship between Amount spent and Age

Result: The amount spent has a weak negative correlation with ages 0-15 , 16-24 and 65&over , While it has a weak positive correlation with the remaining ages, i.e ages 25-34 , 35-44 , 45-54 and 55-64. In other words , Amount spent is slightly increased if the individual is a working class , while it reduces with more non working class

Conclusion: Amount spent has a weak correlation with ages



Fig 31: Showing the covariance result between Amount spent(expend) and purpose of travelling.

Result: A weak positive relationship exist between Amount spent and Business , holiday and Study, while a weak negative relationship exist between amount spent and those who travel for visitation and Miscellaneous

Conclusion: An increase in the number of those who travel for Business, Holiday and Study translates to an increase in amount spent. An increase in the number of those who travel for visitation and Miscellaneous translates to a decrease in Amount spent

4.52 Comparing Model score after removing some features

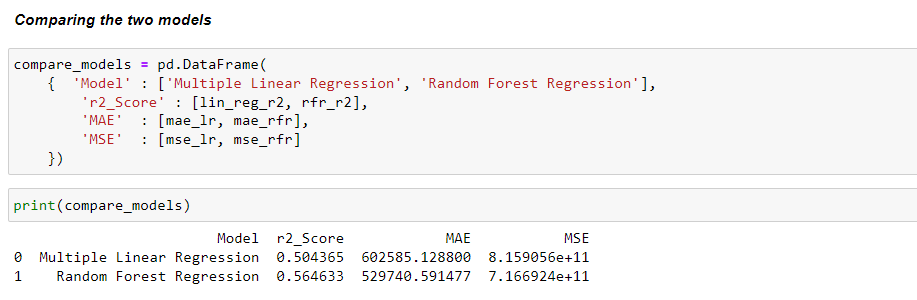


Fig 32: Model score of the first model before removing any feature

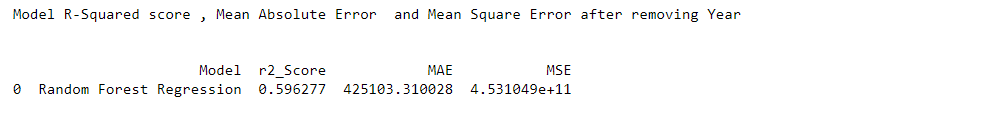


Fig 33: Model score of the second model after removing year

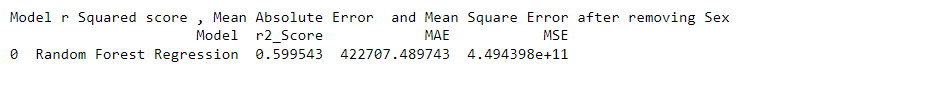


Fig 34: Model score of the third model after removing Sex

Conclusion: Since our model score is not becoming better, the best score so far is 60% accuracy. The result from its prediction cannot be accurate. For that, its better to stick with the prediction we got from our time series model.

# 5.1 Conclusions

Travelling into the UK can be easy for a passenger but as a travelling agency, it is their duty to ensure the transportation is not vulnerable to any attacks. With the survey and reviews conducted, its shows that there is a limited platform to gain access to information about travelling passengers in and out of the UK or perhaps let’s say it isn’t for the public to benefit.

Such information can be useful to researchers or to smaller travelling agencies that do not have the capacity to get the information themselves. More so, lack of information about passengers can make the travelling agencies vulnerable to attacks, having a better understanding about the passengers and the travelling can guide against any negative effects.

These challenges lead to the development of this study. To better solve the problems, we asked some research questions, such as.

i. What are the insights related to the travelers as regards their country, night spent, mode of travel and package of travel?

ii. What necessitates travelling?

iii. Does age groups affects travelling?

Two approaches were made. The first one involved using python programming libraries and analyses to fit time series model and machine learning models for some variables. Timeseries models forecasted the expected yearly outcomes of some of our variables for five years. Secondly,a dashboard was created using Power Bi using a dataset collected. Insights can then be viewed from the dashboard.

The summary of insights generated are as follows:

2010 has the highest number of data samples, while 2020 has the least

There are more oversea travelers than the UK travelers

Travelers who lives in the UK and overseas travelers mostly spent between 4 to 13 nights.

Further segmentaion of travellers who spent 4-13 nights revealed that for travelers that stay in the UK, the major reason for travelling is for holiday, the second major reason is visitation. Most overseas travellers are traveling for visitation purposes while the second major reason is for business.

To see the kinds of people who travel for visitation. The segmentation shows that travellers between the ages of 25 and 65 mostly go for visitation. This is very relatable as the age categories cover university graduates who might want to visit home when school are on break or adults who are travelling back home for visitation.

From age 25 – 55, we have larger number of passengers in that age group, considering that we have more passengers who travel for visitation and holidays, we can conclude that most passengers between the age of 25 – 55 are travelling for holidays and visitation purpose, more reason passengers are spending 1 to 13 nights during their visits

Categorizing the dataset into years and identifying the number of visits yearly, starting from 2009 till 2019, There is an increasing trend in the number of passengers but a huge drop in the year 2020, this is basically caused by COVID-19 lock-down and in the year 2021, it started increasing gradually

The place of residence for most overseas residents or of visit for most UK residents is France, with 35,021 passengers while Cyprus, which is on the least has 690 passengers.

The best four attributes that contribute to the prediction of the type Holiday package a traveler will have are: amount spent(expend), number of nights spent, number of visits and year.

Forecasting results for expenditure for the next 5 years (Year 2022 to 2027), it is observed that year 2023 will have the minimum expenditure amount of over 13 billion, while 2027 is expecting over 16 billion.

The model score machine learning model is 0.56. This means that only 56 percent variations in amount spent(expend) by travelers can be explained by all the features. This model is not good enough for prediction, this suggests that to predict the amount a certain traveler with some travel qualities correctly, we need more attributes.

There is strong relationship between Amount spent and nights , visits while we have weak relationship between amount and year.

The amount spent has a weak negative correlation with ages 0-15 , 16-24 and 65&over , Amount spent is slightly increased if the individual is a working class , while it reduces with more non working class.

An increase in the number of those who travel for Business, Holiday and Study translates to an increase in amount spent. An increase in the number of those who travel for visitation and Miscellaneous translates to a decrease in Amount spent

Several other insights were derived and were used in creating the dash board.

One key thing I learnt during the development of this project is consistency, initially my writing and programming skills was so amateur. But while progressing deeper into the report, I was able to revert back to what I have written before to make correction to what I have newly learnt about dissertation writing. Also, my python knowledge was minimal, I was able to go through series of programming tutorial to get some things done. Numerous code samples were reviewed, which helped in the analysis of this experiment. Also, for the development of the dashboard, this would be the first time I would be working with Power Bi, its an interesting tool to work with. This project introduced me to different tools and techniques as well.

Several problems were encountered, one of the big problems faced was the process of setting up Power Bi. The software is owned by Microsoft and can only be installed on a windows system. I am using a mac book air which requires I make a separate partition to install windows OS in other to install Power Bi. Another problem are the skills needed to complete the analysis. Several programming tutorials were downloaded and watched repeatedly to understand the concept of how python programming languages is being used and its application for data analysis.

During the development of the project, several mistakes were encountered such as programming error due to mistakes from wrong syntax. Also, in the dataset to be used. Initially, the dataset downloaded was just a year, at the point of completing the report, it shows that just a single year dataset can’t be used. All process was repeated over again, this time the dataset started from 2009 to 2021. This process reflected on the project, more insight was derived from this act which leads to yearly comparison of results.

In the future, this result in this analysis and the process can be converted into a web based application where information can be access by anyone. The dataset for the work is only for UK passengers, this can be extended to cover more regions.