Milestone Report

1. Problem Statement:

Nowadays with the convenience of aire plane, people travel internationally for business or vacation. Travel businesses adjust their price based on current demand throughout the whole year. Knowing the demand ahead could help price setting.

2. Description of Dataset:

2.1. General Description:

The datasets are from National Travel & Tourism Office. The three excel files contain monthly international travel counts to different regions from 2014 to 2017.

2.2. Data Wrangling Process:

2.2.1. **Initial Handling**:

Since there are unneeded information, I only read in desired rows from the excel files as DataFrame. I manually checked what rows are needed for the final dataset. The function *create_df* was created to help create dataframe for each year from 2014 to 2017.

2.2.2. **Main Process**:

I intend to have regions as columns and each month as datetime index. I first set *Regions* as index to separate out from other columns. Next, I used stack() and unstack() to swap index and columns. Last, I changed abbreviation of months into number form then convert to Datetime. Additionally, I renamed the columns to remove extra characters. The function *to_time* was created to do the process.

2.2.3. **Final Process**:

All data frames are concatenated as one whole time series dataframe with monthly frequency. The result dataframe *outbounds* was exported as a csv file.

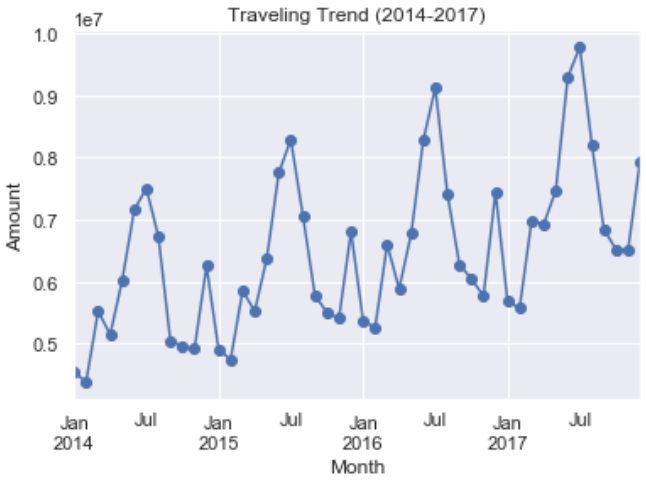
2.3. Summary:

The result dataset *outbounds* contains total of 48 records and 13 columns. The indexes are from 2014-01-31 to 2017-12-31 with monthly frequency.

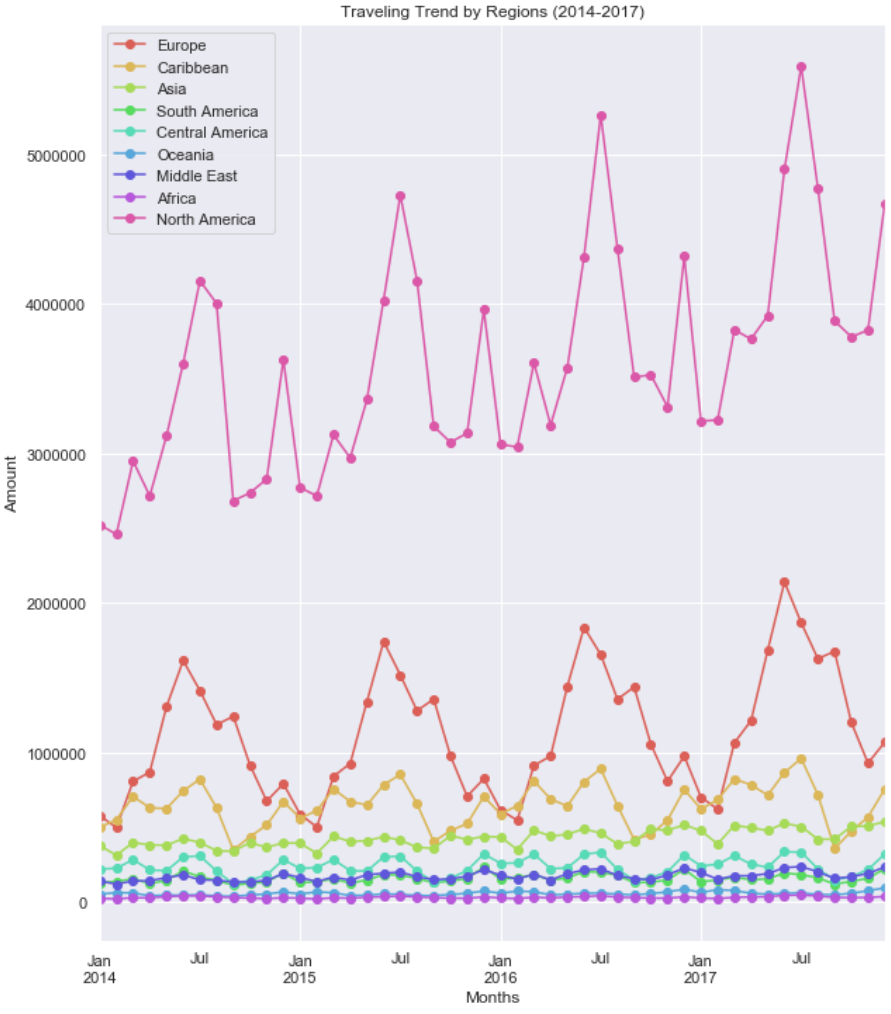
3. Exploratory Data Analysis:

3.1. Initial Exploration:

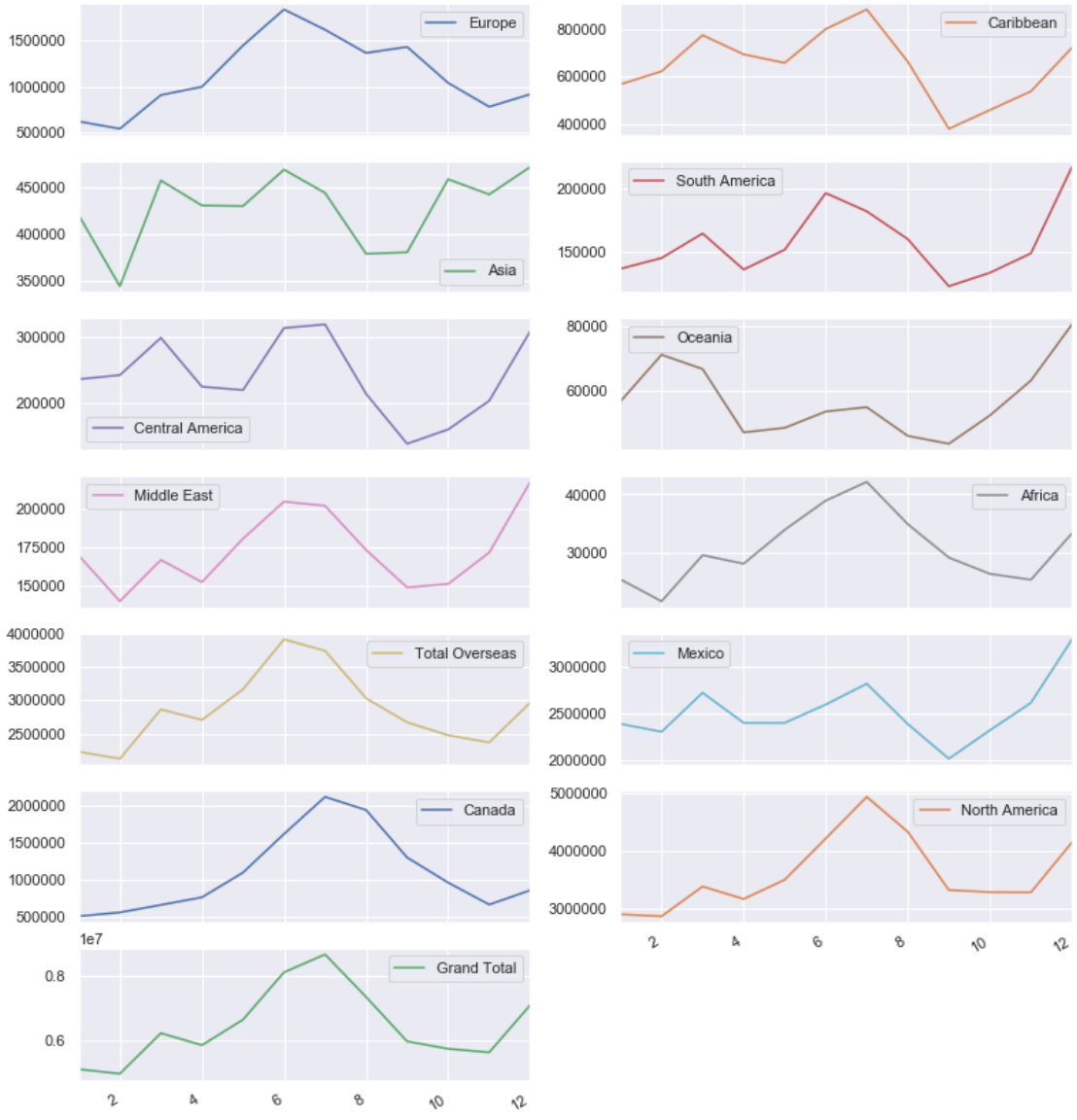
I first observed the general trend over the 4 years then individual year. From the plots, we can see international traveling has been increasing. There are higher demands from May to August. February has the lowest demand. The high demand in December may be resulted from Christmas holidays.



Next, I observed the traveling trend by regions. North America includes Mexico and Canada, which has the highest amount. The trend of North America looks similar to the trend of total amount. Since the amount from North America is significantly higher than other regions, I decided to research if the trend would look different if not including North America. The result is similar to the trend of total amount.

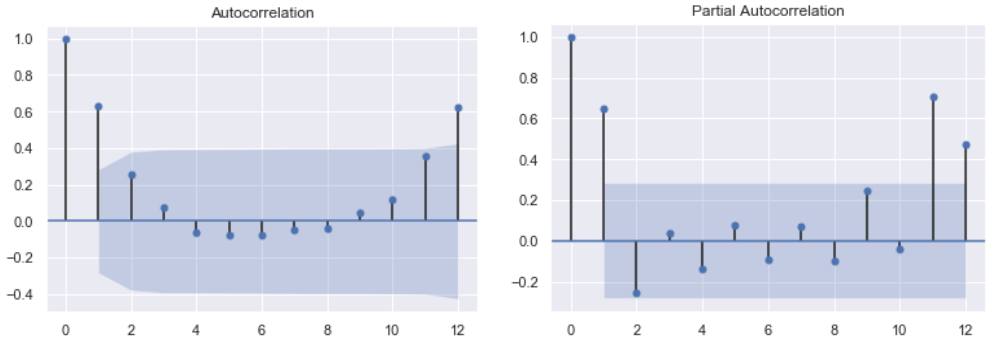


I also observed the average travel amount by month. People travel to Oceania mostly during winter time, which is summer time over there. Generally, there are three travelling peaks during a year. They are March, July and December, with July the most. I found it interesting that Asia has a traveling peak in October besides other three popular months.



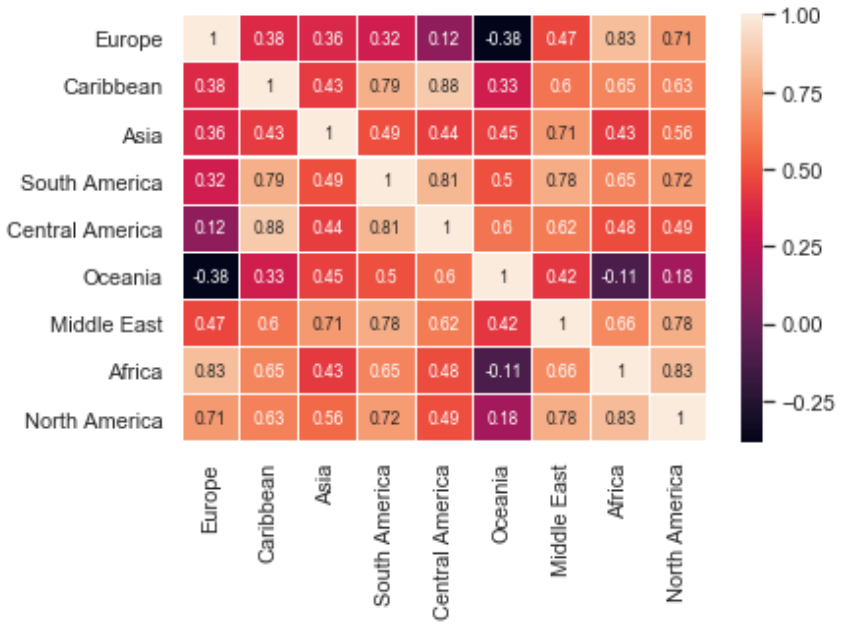
3.2. Deeper Exploration:

I first check the autocorrelation and partial autocorrelation of *Grand Total* with lags of 12 (a year). The autocorrelation graph shows the observations are negatively correlated from lags = 4 to lags = 8 and positively correlated for other lags. However, only at lags = 0, 1, 12, the observations are highly correlated and statistically significant. The partial autocorrelation graph shows the observations are statistically significant at lags = 0, 1, 11 and 12.

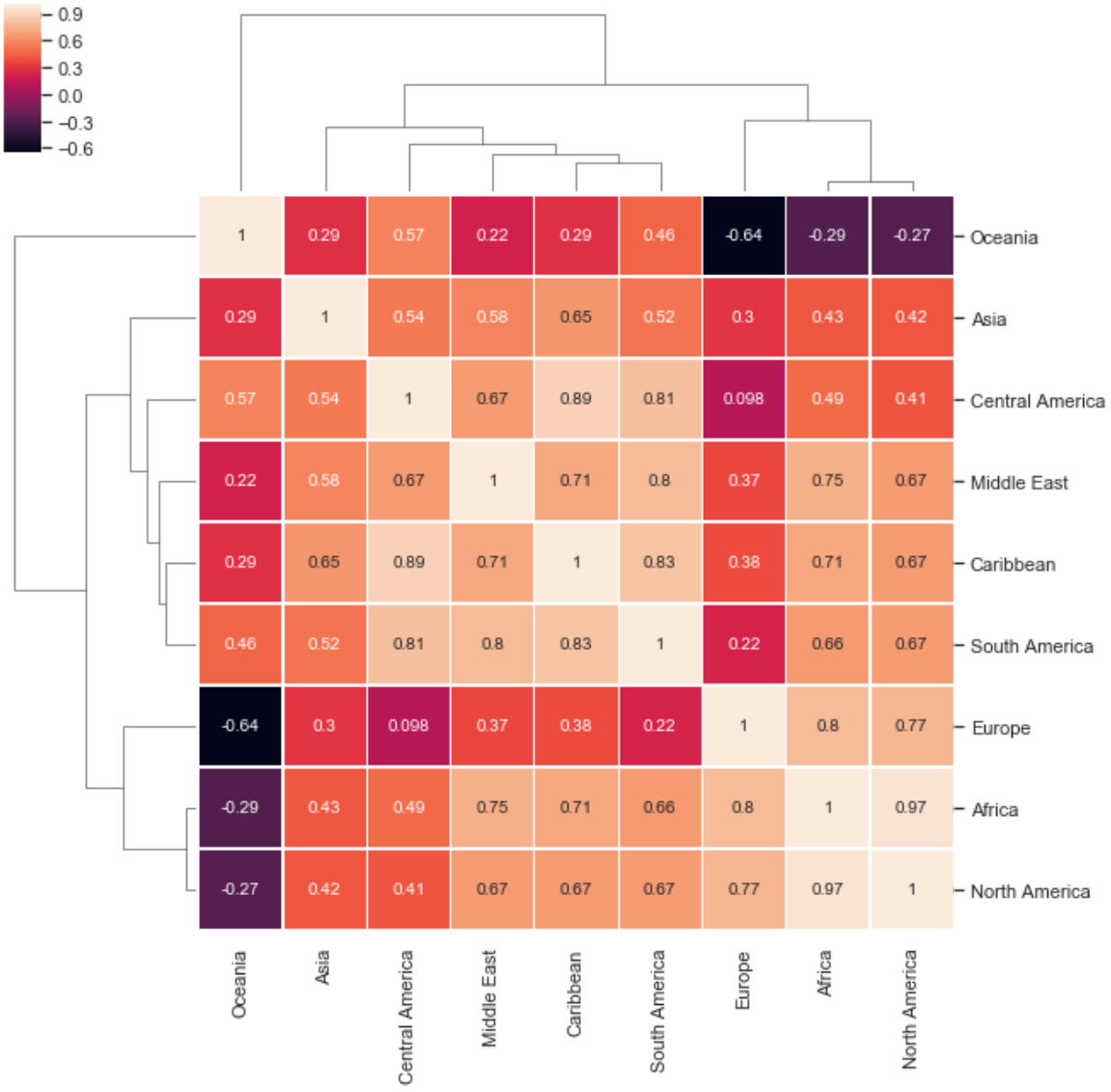


Next, I observed the correlation between regions. I also organized most/least correlated region to each region into a data frame. Oceania and Europe are the least correlated to other region may due to season. Caribbean, Central America and South America are highly correlated with each other may due to geographically they are close to each other.

Regions	Most Correlated	Least Correlated
Europe	Africa	Oceania
Caribbean	Central America	Oceania
Asia	Middle East	Europe
South America	Central America	Europe
Central America	Caribbean	Europe
Oceania	Central America	Europe
Middle East	North America	Oceania
Africa	Europe	Oceania
North America	Africa	Oceania



Next, I applied decomposition to observe the seasonality, trend and noise in the dataset. Grand total and regions are both observed. A cluster heatmap was plotted to see the correlation between each region's seasonality. The correlation result of seasonality is very similar with the previous correlation result. Oceania and Europe still the least with other regions.



Last, I check if the data is stationary with Dicky-Fuller Test. The p-value shows the data is not stationary. Some data transformation will be required to make the data stationary before machine learning modeling.

Test Statistic	-0.124552
p-value	0.946937
#Lags Used	10.000000
Number of Observations Used	37.000000
Critical Value (1%)	-3.620918
Critical Value (5%)	-2.943539
Critical Value (10%)	-2.610400

3.3. Conclusion:

From the observations, we can see there's an increase trend and a seasonal pattern annually. Overall, travel demands are high in summer and winter holidays. Travel demands to each region are related to the region's season.