Final 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.

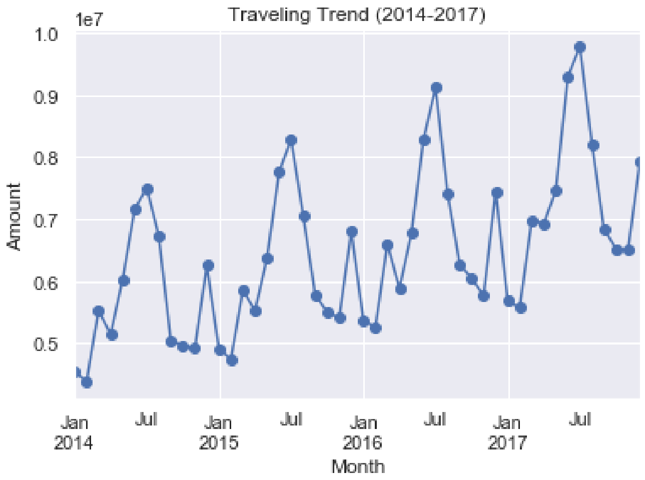
2.3. Summary:

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.

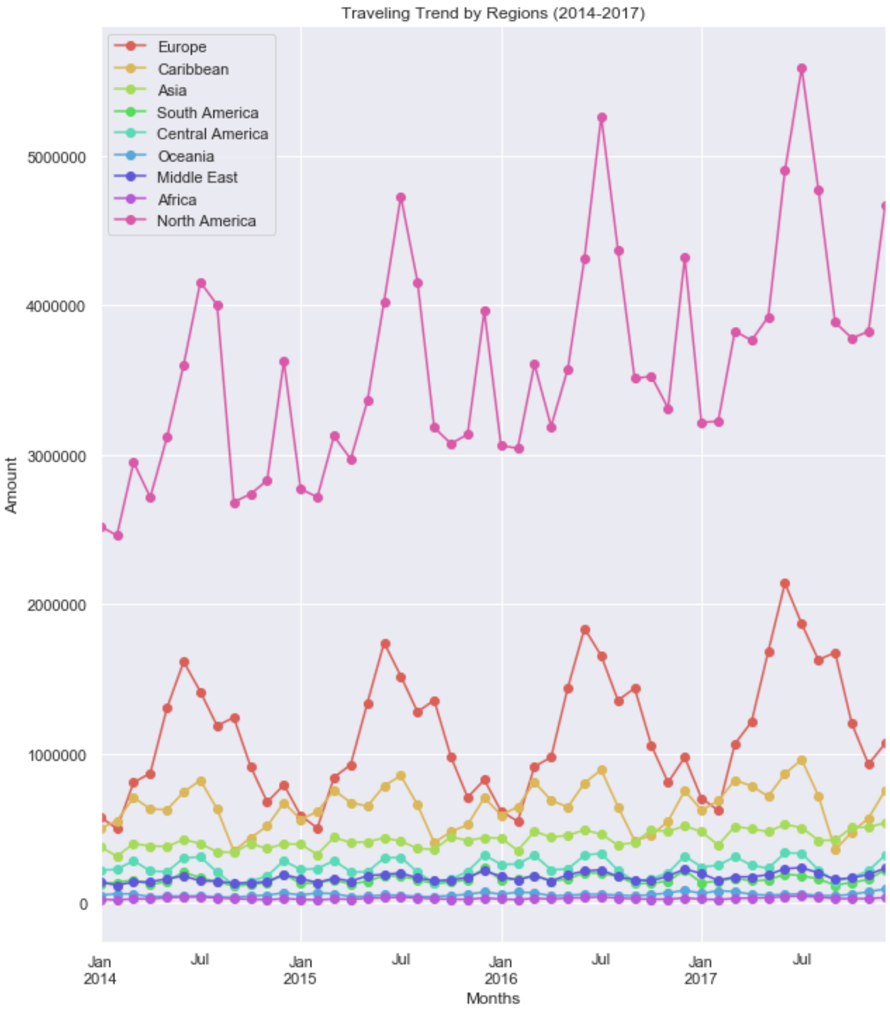
3. Exploratory Data Analysis:

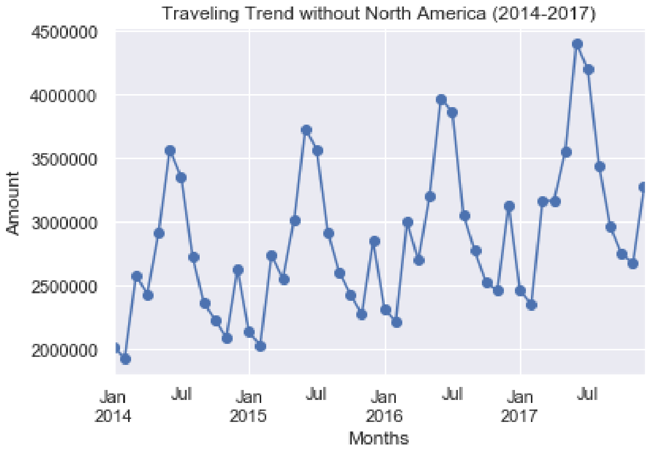
3.1. Initial Exploration:

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.

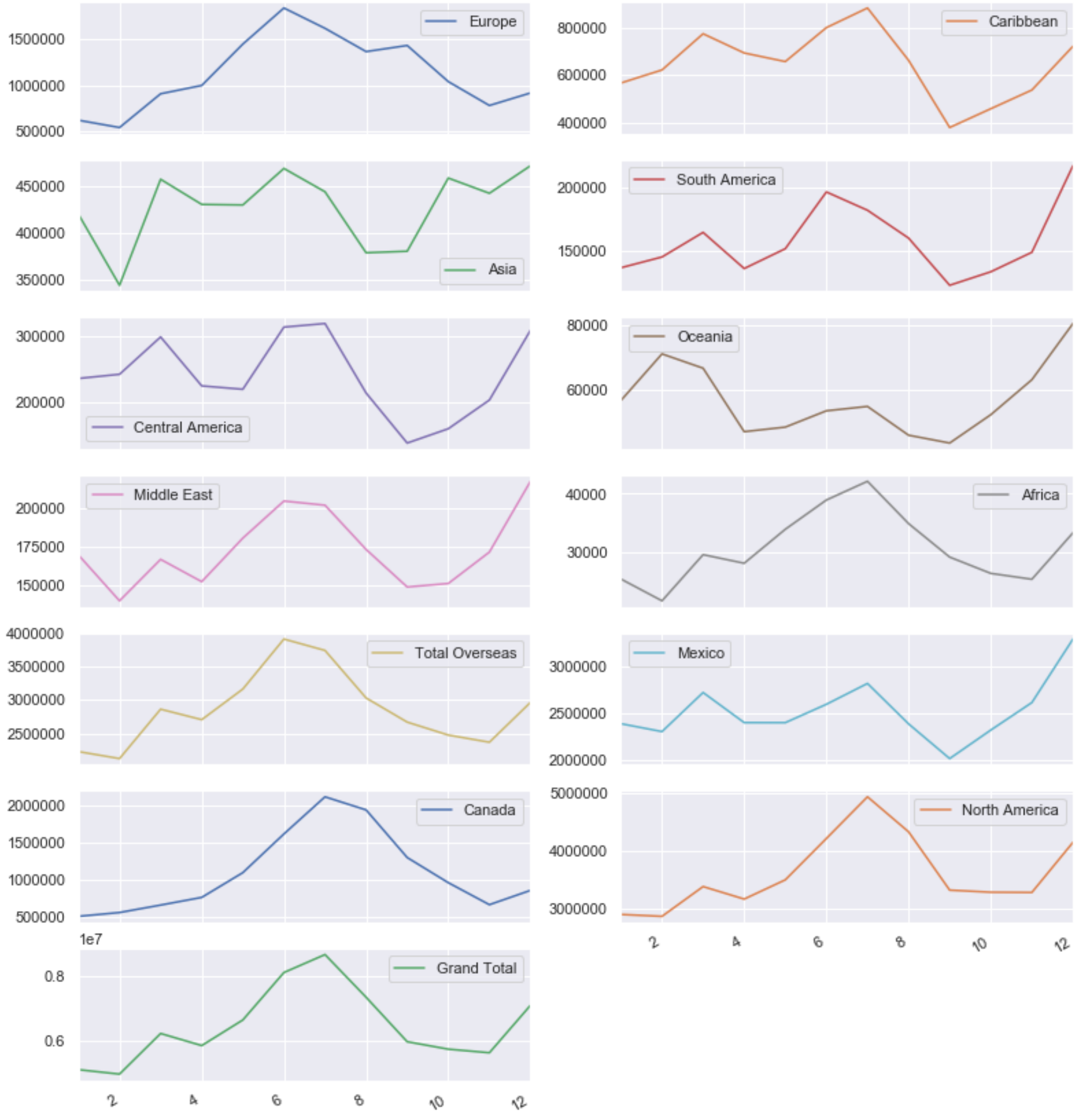


Next, I observed the traveling trend by regions. North America includes Mexico and Canada, which has the highest amount. The trend of North America looks similar to the trend of total amount. Since the amount from North America is significantly higher than other regions, I decided to research if the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



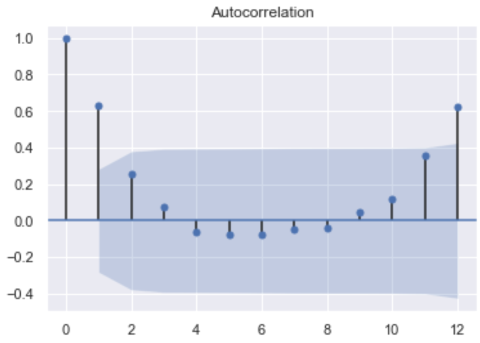


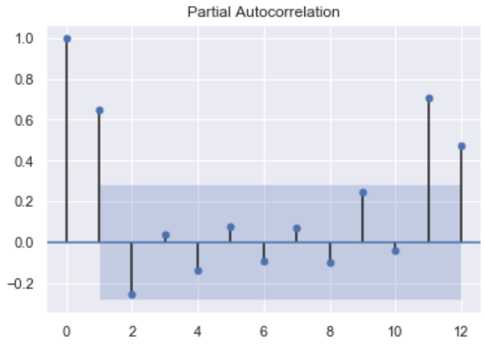
I also observed the average travel amount by month. People travel to Oceania mostly during winter time, which is summer time over there. Generally, there are three travelling peaks during a year. They are March, July and December, with July the most. I found it interesting that Asia has a traveling peak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



3.2. Deeper Exploration:

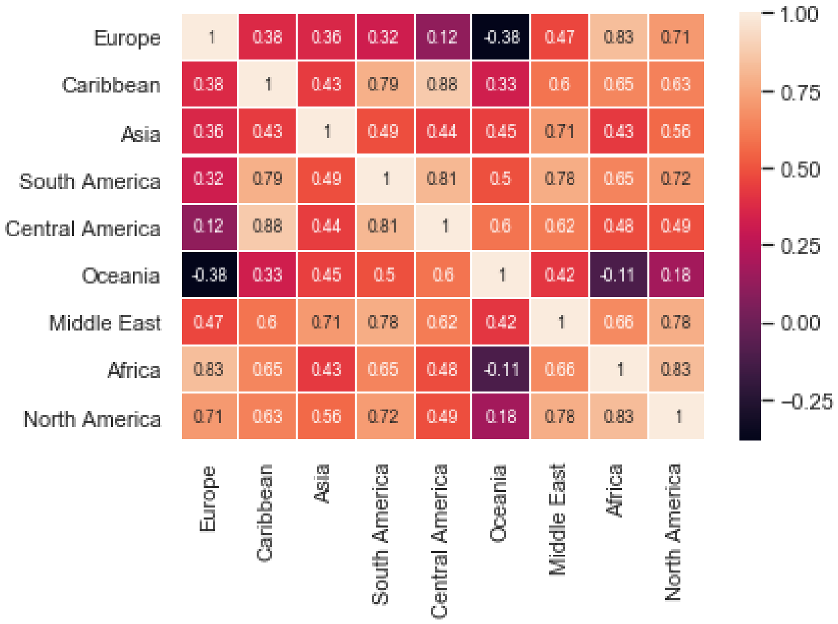
I first check the autocorrelation and partial autocorrelation of *Grand Total* with lags of 12 (a year). The autocorrelation graph shows the observations are negatively correlated from lags = 4 to lags = 8 and positively correlated for other lags. However, only at lags = 0, 1, 12, the observations are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





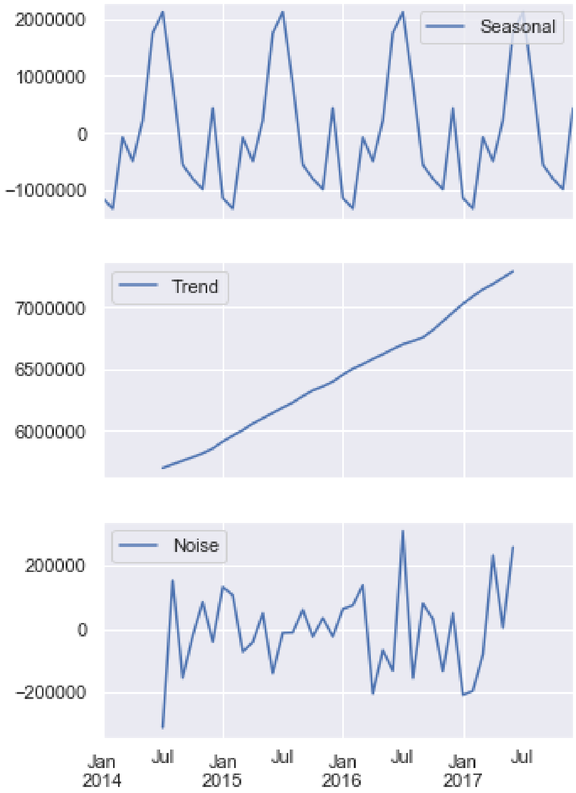
Next, I observed the correlation between regions. I also organized most/least correlated region to each region into a data frame. Oceania and Europe are the least correlated to other region may due to season. Caribbean, Central America and South America are highly correlated with each other may due to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

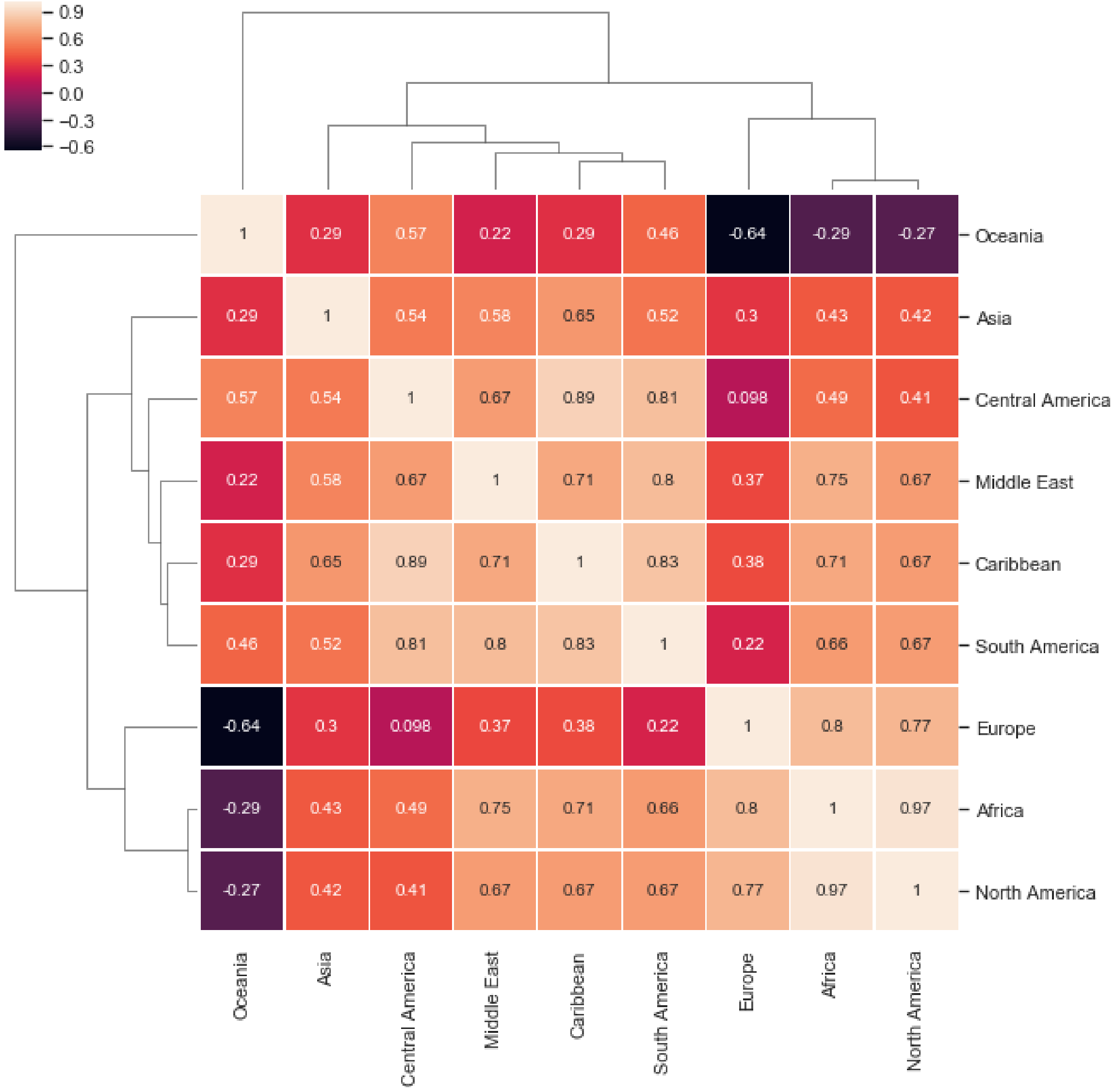
Regions	Most Correlated	Least Correlated
Europe	Africa	Oceania
Caribbean	Central America	Oceania
Asia	Middle East	Europe
South America	Central America	Europe
Central America	Caribbean	Europe
Oceania	Central America	Europe
Middle East	North America	Oceania
Africa	Europe	Oceania
North America	Africa	Oceania



	Europe	Caribbean	Asia	South America	Central America	Oceania	Middle East	Africa	North America
Europe	1	0.38	0.36	0.32	0.12	-0.38	0.47	0.83	0.71
Caribbean	0.38	1	0.43	0.79	0.88	0.33	0.6	0.65	0.63
Asia	0.36	0.43	1	0.49	0.44	0.45	0.71	0.43	0.56
South America	0.32	0.79	0.49	1	0.81	0.5	0.78	0.65	0.72
Central America	0.12	0.88	0.44	0.81	1	0.6	0.62	0.48	0.49
Oceania	-0.38	0.33	0.45	0.5	0.6	1	0.42	-0.11	0.18
Middle East	0.47	0.6	0.71	0.78	0.62	0.42	1	0.66	0.78
Africa	0.83	0.65	0.43	0.65	0.48	-0.11	0.66	1	0.83
North America	0.71	0.63	0.56	0.72	0.49	0.18	0.78	0.83	1

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.





	Oceania	Asia	Central America	Middle East	Caribbean	South America	Europe	Africa	North America
Oceania	1	0.29	0.57	0.22	0.29	0.46	-0.64	-0.29	-0.27
Asia	0.29	1	0.54	0.58	0.65	0.52	0.3	0.43	0.42
Central America	0.57	0.54	1	0.67	0.89	0.81	0.098	0.49	0.41
Middle East	0.22	0.58	0.67	1	0.71	0.8	0.37	0.75	0.67
Caribbean	0.29	0.65	0.89	0.71	1	0.83	0.38	0.71	0.67
South America	0.46	0.52	0.81	0.8	0.83	1	0.22	0.66	0.67
Europe	-0.64	0.3	0.098	0.37	0.38	0.22	1	0.8	0.77
Africa	-0.29	0.43	0.49	0.75	0.71	0.66	0.8	1	0.97
North America	-0.27	0.42	0.41	0.67	0.67	0.67	0.77	0.97	1

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:

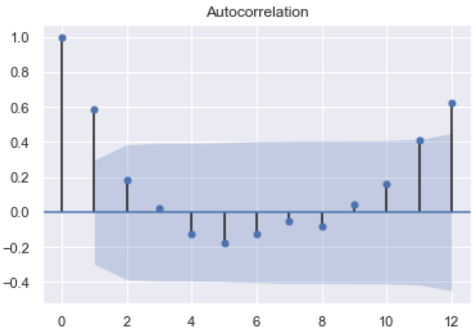
From the observations, we can see there’s an increase trend and a seasonal pattern annually. Overall, travel demands are high in summer and winter holidays. Travel demands to each region are related to the region’s season.

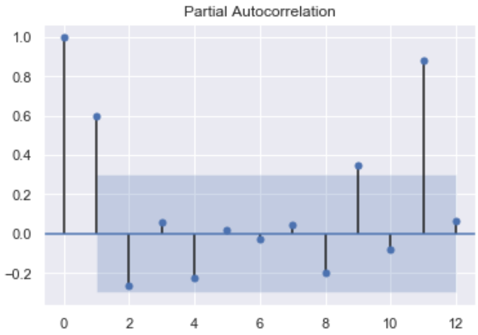
4. Results and In-Depth Analysis using Machine Learning:

4.1. In-Depth Aanlysis:

Prior to data processing and modeling, I created three helper functions. *test_stationary* uses Dicky-Fuller Test and displays the result. *plot_decompose* takes in a time series and returns the decomposition. It applies seasonal decomposition on the input time series, then plot the decomposition. *rmse* calculates the root mean square error, which takes in actual values and predicted values. I splitted the dataset into training and testing sets with ratio of 9:1.

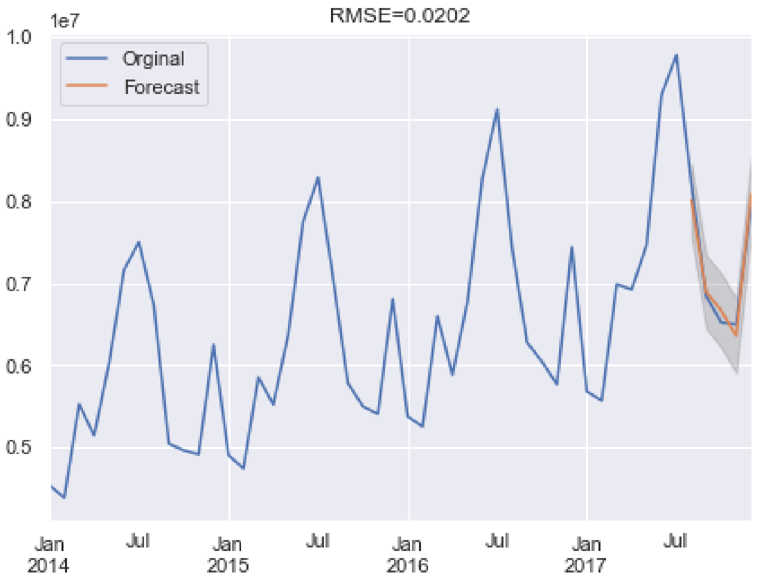
Since my data contains both trend and seasonal pattern. I decided to use SARIMA. I plotted ACF and PACF graph to determine my hyperparameters for the model. First, I looked for the basic (p,d,q) for AR(p) and MA(q) by inspecting the first significant lag in both graphs. PACF has the first significant at lag = 1, which means the parameter p = 1. ACF has the first significant at lag = 1, which means the parameter q = 1. Since there’s a increasing trend, the non-seasonal differencing d = 1. Next, I looked for seasonal (P,D,Q) for seasonal AR (P) and seasonal MA(Q) with S = 12, which is the time span of repeating seasonal pattern. The lag = 12 in ACF is positive, thus P = 1 and Q = 0. The data shows a seasonal pattern, so seasonal differencing D = 1.


Autocorrelation

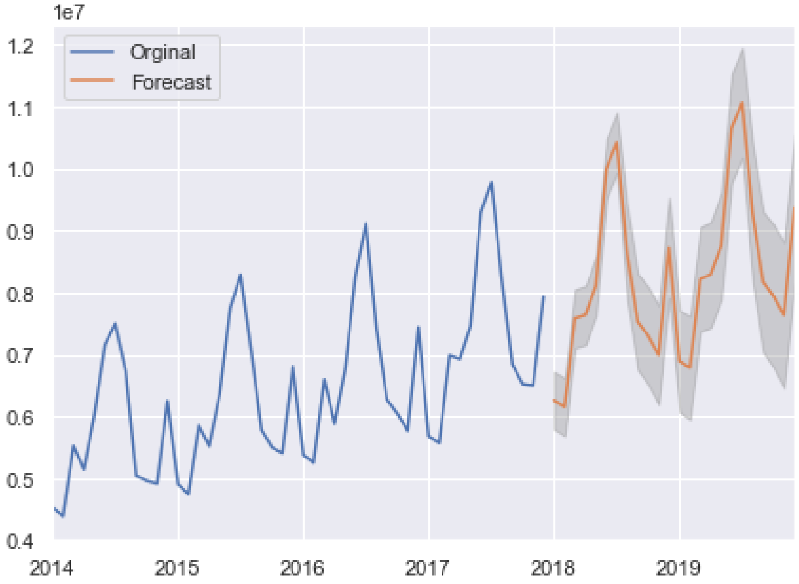

Partial Autocorrelation

To verify (1,1,1) x (1,1,0,12) is the best model, I constructed other hyperparameters combinations and did a grid search. The AIC score showed the best model indeed is (1,1,1) x (1,1,0,12).

I defined a SARIMA model with the best result found in the grid search then fit the model with the total monthly outbounds. I forecasted the value with same date range as the testing set, and compared it with the testing set data to see how the model perform. We can see the predicted values are very close to the original data. The grey area in the plot is the 95% confidence interval of the predicted values. The normalized RMSE is 0.0202.

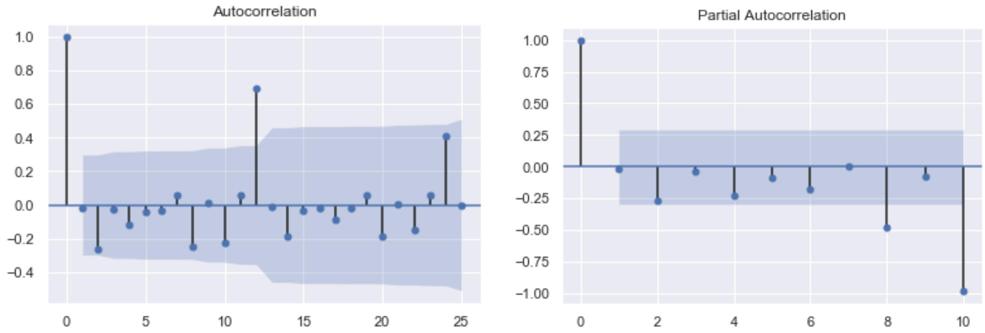


Next, I forecasted the outbounds demand for next 2 years. I plotted the predicted values with the original values. I also plotted the 95% confidence interval as the grey area.

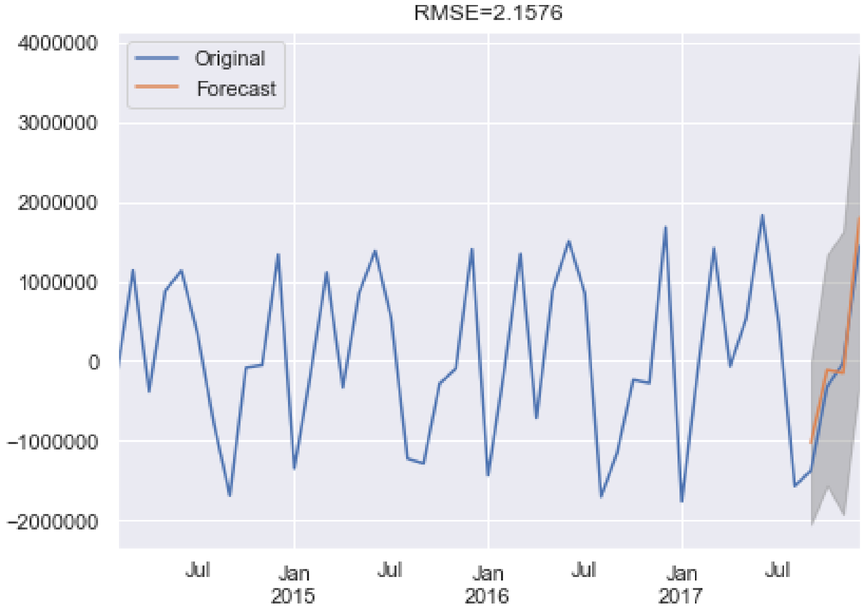


Unsure if making the data stationary would perform better, I decided to make the data stationary and compare the resulted model with the one I just created. I tried logarithm, logarithm with first differencing, first differencing, and seasonal differencing plus first differencing. I used Dicky-Fuller Test on each data to test the stationarity. All data are stationary besides the logarithm one. I chose the first differencing data to train the model considering the size of my dataset and the process required to “convert” the predicted values back later. Before modeling process, I splitted the first differencing data set into training and testing sets with the ratio of 9:1. I used the same process as mentioned before to identify the hyperparameters, which is (0,1,0) x

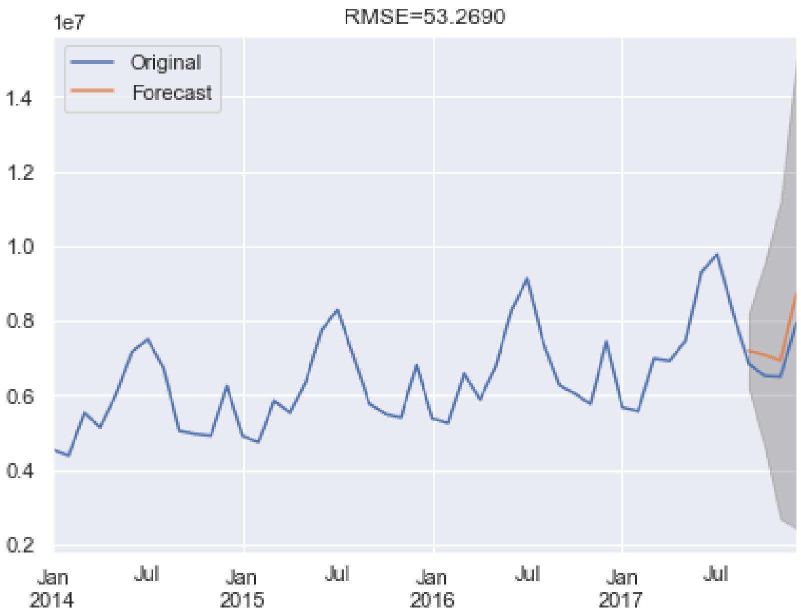
(1,1,0,12). I verified the combination with a grid search. The result shows (0,1,0) x (1,1,0,12) is the best combination.

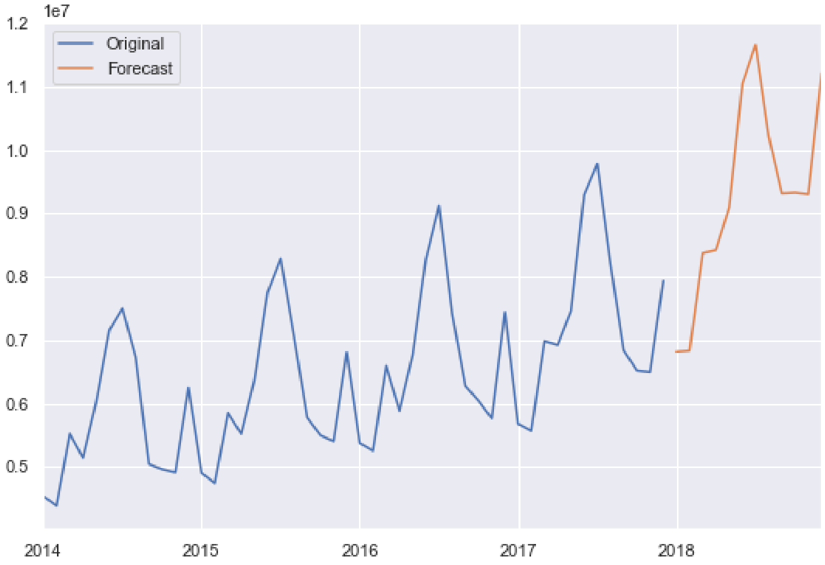


I forecasted the demand with the same data range as in the testing set, and plot the predicted values against the actual values.The normalized RMSE is 2.1576, which indicates this model performs worse than the previous one.



I converted the predicted values back to scale by adding the original value at the last timestamp of the training set, which is ‘2017-08-31’, to the cumulative sum of the predicted values. I plotted both original and predicted values to observe. The result doesn’t seems close to the original data. The normalized RMSE is 53.2690. Despite the fact that the model performs worse than the previous one, I still curious how the future forecast would be like. I forecasted the traveling demand for 2018. The result does show the seasonal pattern and a increasing trend.





4.2. Conclusion:

The SARIMA model with original data performs the best. The reason may be the model itself already taking care of the seasonality and trend with hyperparameter d and D.

I tried 70%, 75%, 80% as training data size before settled with 90%. I found that if training size is below 80%, I don’t have enough data to train. Furthermore, the larger the training set is, the better the model performs. This might due to small size of the data set.

From the predicted values, we can see the seasonal patterns where we have peaks at summer and December. It’s interesting the peak at March isn’t as obvious as previous years, and seems to become milder in 2019 prediction. It may caused by the higher demand in April in 2017. We also see the increasing trend in the prediction.

4.3. Future Work:

Since the data contains regional data, we can further forecast the demand by regions. More travel data will certainly help to improve the model's performance. Economic or weather data may be factors, it would be interesting to see what would be the result if incorporate these data.

Once gather more data and have enough data set, try stationarize the data , then use the ARIMA model instead of SARIMA model for the stationary data. The approach might yield a better result.