CSP 571 DPA FINAL PROJECT

TIME SERIES ANALYSIS AND FORECASTING - NYC TAXI FARE

Team:
Yash Pradeep Gupte - A20472798
Amrutham Lakshmi Himaja - A20474105

Research Goal

Objective :

Our objective is to predict the taxi fares of the NYC taxi dataset and understanding the features that impact the taxi fares.

Specific Questions :

- What are the locations where the taxi fare is relatively high and low(demand based on location)?
- What is the time when the fare of the taxi is high as well as low (demand based on time)?
- Does the dataset require additional features to determine the proposed outcome ?
- What features are statistically correlated with each other?
- How does stationarity and seasonality affect our time series analysis?

Findings :

Time series models are better at forecasting future taxi fares when compared to linear models.The taxi fares are highest during early mornings and evenings as people travel to work and airports or famous locations during this time.

Executive Summary

Future Work :

- For the future scope of the project, we would like to explore the correlation between location and timestamps collectively. Considering these features, we would be able to draw conclusions for optimal routes with optimal taxi fares. We would also like to explore the impact on traditional taxis by considering taxi pool services like Uber and Lyft. We would extend the time series forecasting model to predict the exact or approximate taxi fare for a given date and time in the future.

Project Outline

CSP 571 Project Planner

TIME SERIES ANALYSIS AND FORECASTING - NYC TAXI FARE
Yash Pradeep Gupte - A20472798
Amrutham Lakshmi Himaja - A20474105

Tasks and Deliverables	Start Date	End Date	Duration(week)	Assigned to		Complete	Week						
				Yash	Himaja		1	2	3	4	5	6	7
Project - Formation and Ideation - Phase 1													
Project Group & Topic Form	09/18	09/25	1	☑	☑	completed							
Project Proposal & Outline	10/09	10/16	1	☑	☑	completed							
Project Plan & Detail	11/8	11/13	1	☑	☑	completed							
Data Selection - Phase 2													
Features and Sample Selection	10/09	10/16	1	☑	☑	completed							
Prepare Dataset	10/09	10/16	1	☑	☑	completed							
Data Processing - Phase 3													
Data Cleaning	11/8	11/13	1	☑	☑	completed							
Identify Missing values and Imputation	11/8	11/13	1	☑	☑	completed							
Outlier Identification and Elimination (if required)	11/8	11/13	1	☑	☑	completed							
Data Aggregation (if required)	11/8	11/13	1	☑	☑	completed							
Feature Importance	11/8	11/13	1	☑	☑	completed							
Data Transformation - Phase 4													
Define final Time Series and Clustering Data	11/13	11/16	1	☑	☑	completed							
Feature Engineering and Lags	11/13	11/16	1	☑	☑	completed							
Peform Distribution for Each Features	11/13	11/16	1	☑	☑	completed							
Transform Features (if required)	11/13	11/16	1	☑	☑	completed							
Data Analysis - Phase 5													
Time Series - Stationarity and Seasonality Check	11/16	11/20	1	☑	☑	completed							
Transform Time Series - Differencing (if required)	11/16	11/20	1	☑	☑	completed							
Auto correlation Plots - ACF, PACF	11/16	11/20	1	☑	☑	completed							
Identify optimal Lags	11/16	11/20	1	☑	☑	completed							
Split Data for Time Series - Train / Test	11/16	11/20	1	☑	☑	completed							
Clustering Analysis on entire Dataset - Elbow Plot	11/16	11/20	1	☑	☑	completed							
Identify optimal K value for clustering(SH score)	11/16	11/20	1	☑	☑	Incomplete							
PCA For Feature And Dimensionality Reduction	11/16	11/20	1	☑	☑	completed							
Training/Testing Set Split	11/16	11/20	1	☑	☑	completed							
Data Modeling and Inference - Phase 6													
Regression or XGBoost	11/21	12/3	2	☑	☑	completed							
Clustering Model - K Means	11/21	12/3	2	☑	☑	completed							
Time Series Modeling - ARIMA and SARIMA	11/21	12/3	2	☑	☑	completed							
Model Conclusions and insights - Evaluation	11/21	12/3	2	☑	☑	completed							
Model Comparision	11/21	12/3	2	☑	☑	completed							
Model Inference	11/21	12/3	2	☑	☑	completed							
Critique	11/21	12/3	2	☑	☑	completed							
Final Report	11/21	12/3	2	☑	☑	completed							
Project Presentation	11/21	12/3	2	☑	☑	completed							

Dataset

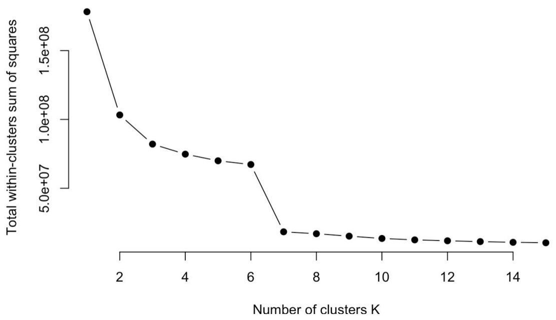
- The data is about 5.5gb size, to process this huge chunk of data, we consider only a subset of this dataset - about 2M observations from the train set and proceed for further preprocessing and analysis.

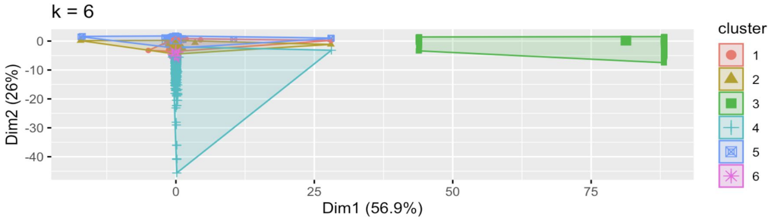
	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
1	2009-06-15 17:26:21	4.50	2009-06-15 17:26:21	-73.84431	40.72132	-73.84161	40.71228	1
2	2010-01-05 16:52:16	16.90	2010-01-05 16:52:16	-74.01605	40.71130	-73.97927	40.78200	1
3	2011-08-18 00:35:00	5.70	2011-08-18 00:35:00	-73.98274	40.76127	-73.99124	40.75056	2
4	2012-04-21 04:30:42	7.70	2012-04-21 04:30:42	-73.98713	40.73314	-73.99157	40.75809	1
5	2010-03-09 07:51:00	5.30	2010-03-09 07:51:00	-73.96810	40.76801	-73.95665	40.78376	1

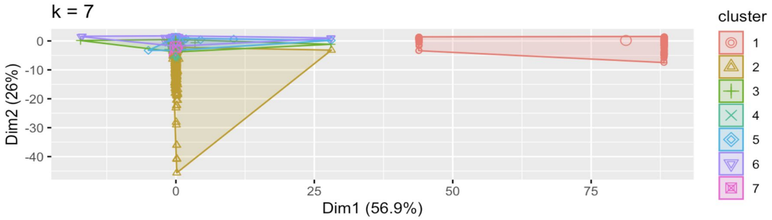
- Data Issues and Data Cleaning
    - Removing NaN values
    - passenger count between 1 to 6
    - pick up and drop off latitude and longitude
    - fare amount between 1 to 500.

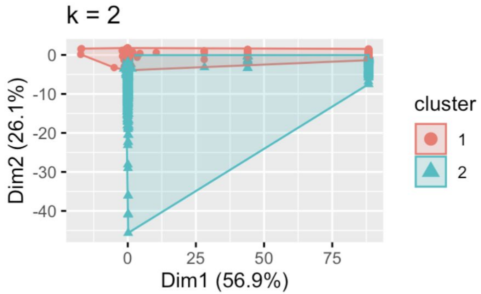
Clustering

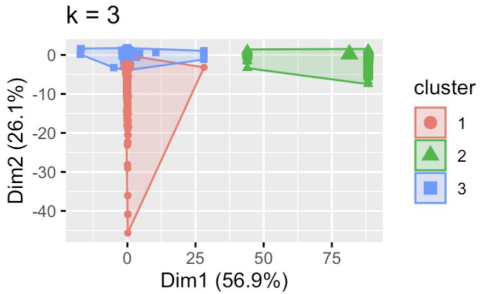
- Elbow method and forming clustering based on the k values

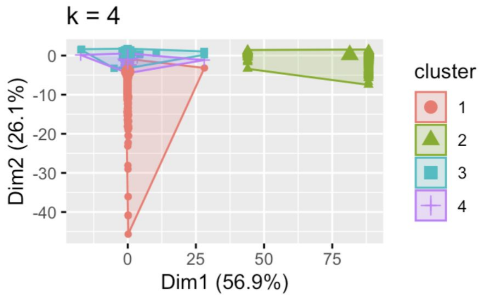


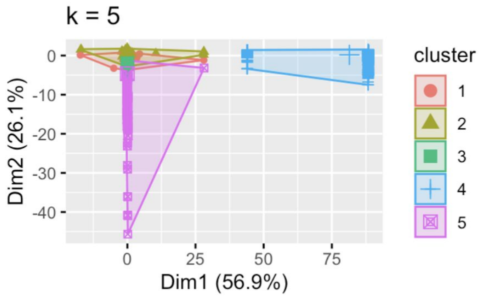






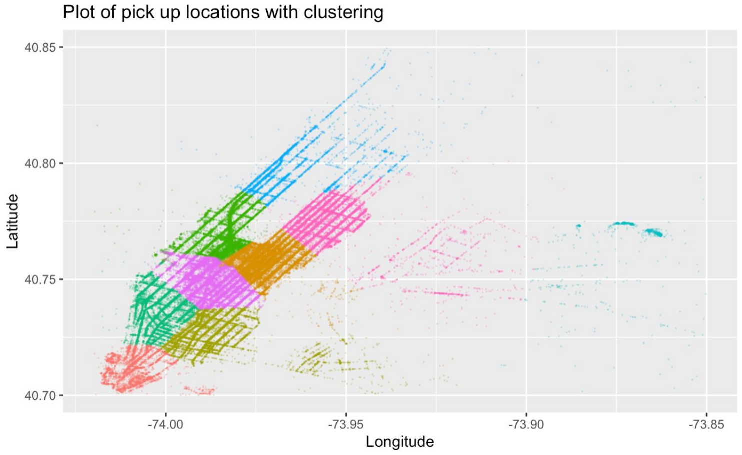


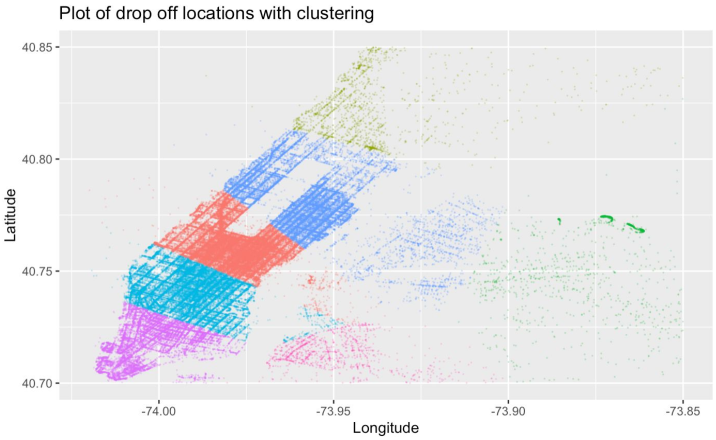




K means Clustering

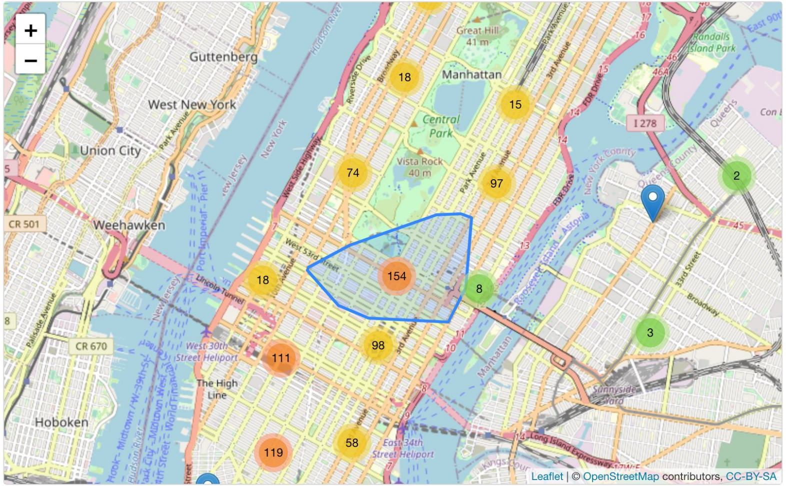
- Clustering based on the pickup and drop off locations of combined test and train dataset.

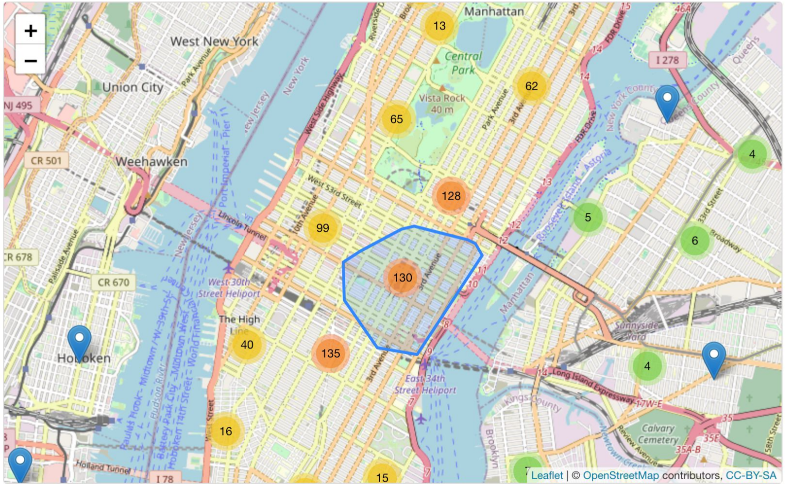




Leaflet Clustering Algorithm

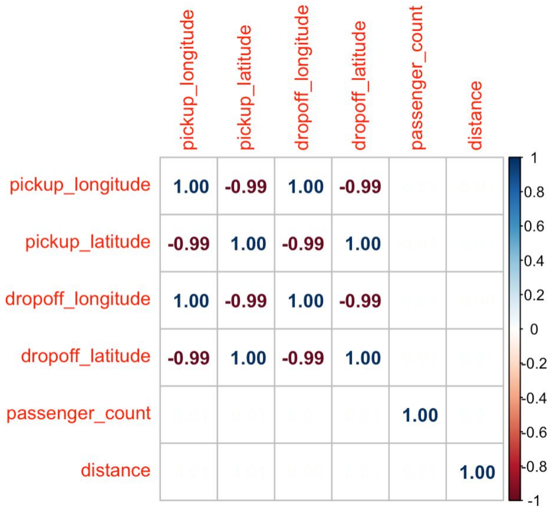
- Using the leaflet clustering algorithm, we can identify the prime pickup and dropoff locations in New York city where the demand for taxis is higher.

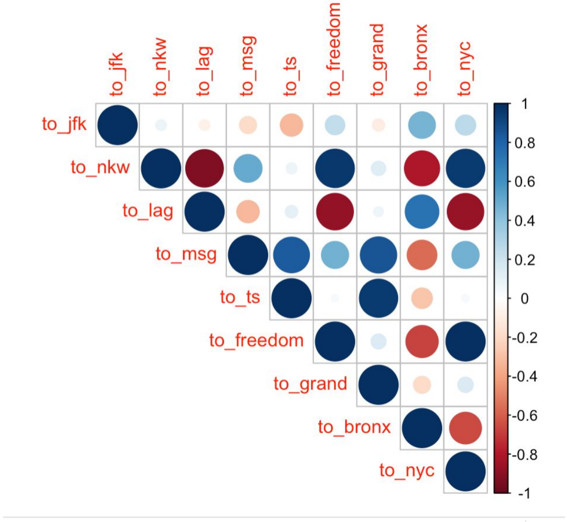




Correlation

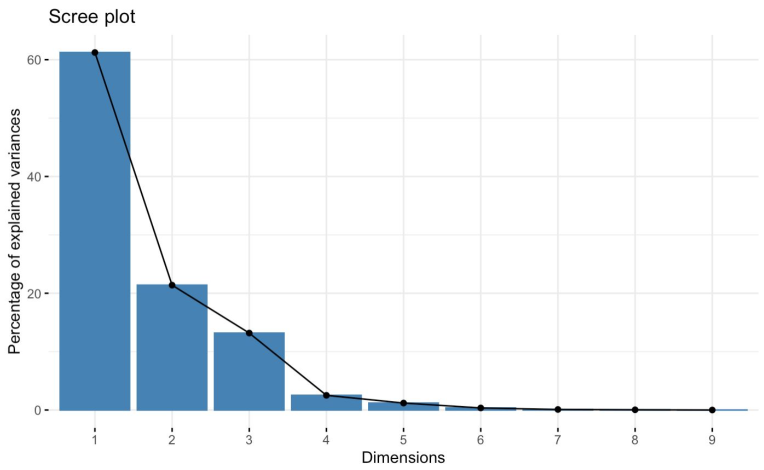
- Using Spearman's correlation to look up the correlation between the different variables on the entire dataset and also for the new features to the dataset.
- The correlation plot for the entire dataset have either positive or negative correlation between the variables. Distance to Laguardia is the airport factor with the lowest correlation. Times Square and Grand Central Station seem to have a direct correlation.





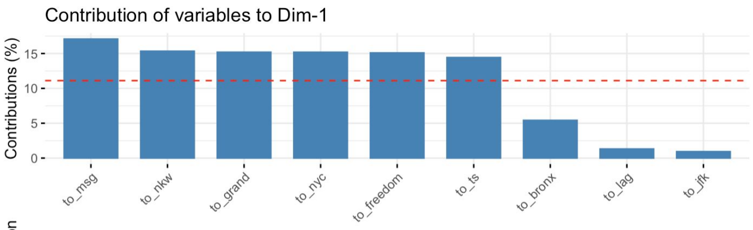
PCA

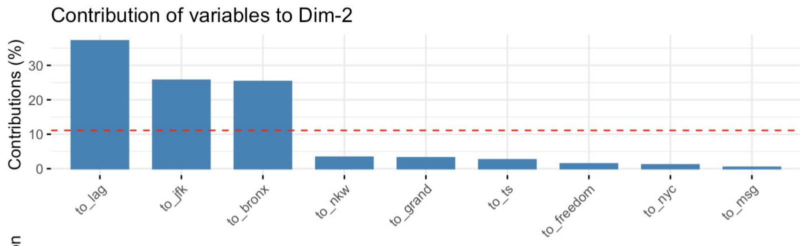
- Performing Principal Component Analysis understanding the variability. Employing the package factoextra to visualize the PCA. Using the plot we examine how much variance is accounted for by the primary components. The first main component appears to explain more than 60% of the variance! By the second and third, it had dropped to almost 15%

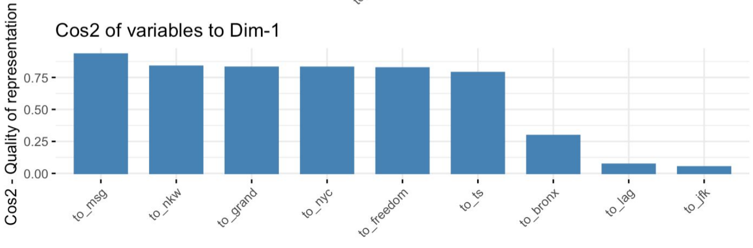


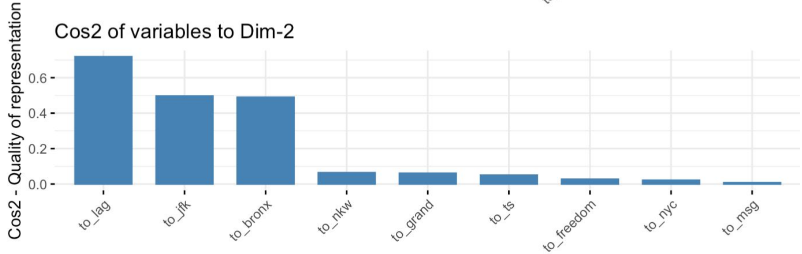
PCA 1 & 2

- Considering the first principal component. It is primarily made up of Madison Square Garden, Newark Airport,Grand Central Station,Center of New York City, Freedom Tower and Time Square locations which are above the red line in the below graph. From the analysis we can see that more than 80% of the Madison Square Garden location is represented by the first principal component.
- Considering the second principal component. It is primarily made up of Laguardia, JFK Airport and Bronx locations which are above the red line in the below graph which did not contribute in the first principal component. The Laguardia has the highest quality of representation of about 70% information retained.



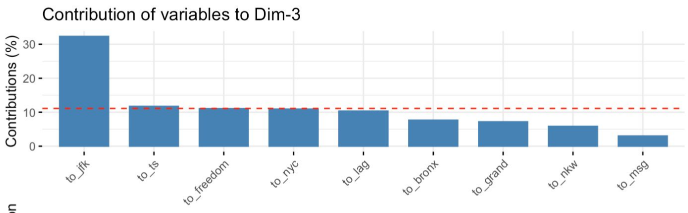


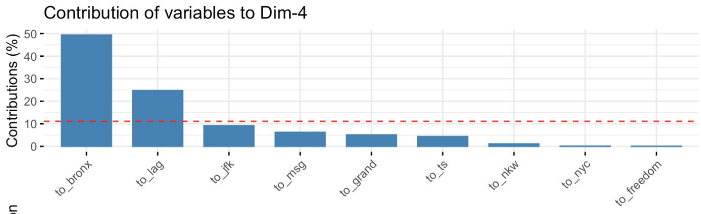


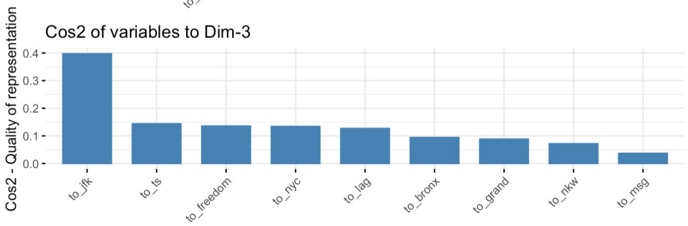


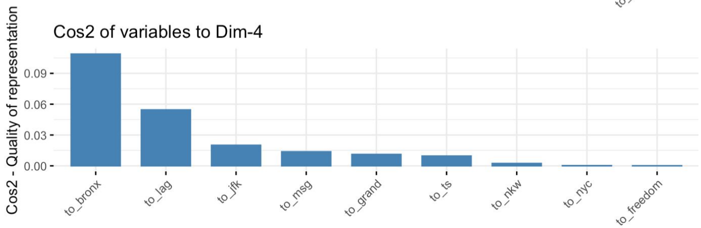
PCA 3 & 4

- Taking into account the third principal component.The first two are what make it up. Given how much variety the first two described, this seems logical. Using the first three principal components looks like a wise choice if we were interested in dimension reduction. We lose about 40% of the information from JFK and 15% from Time Square if we simply use the first two.
- Considering the fourth principal component, It appears about 12% of information from the Bronx would be lost and 5% for Laguardia.



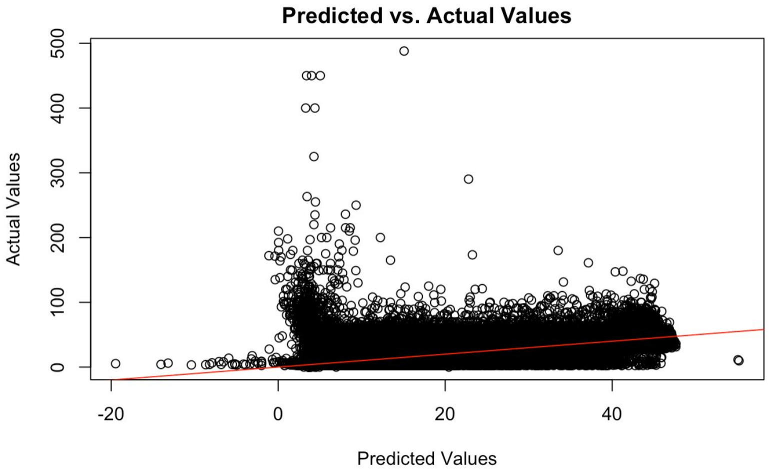






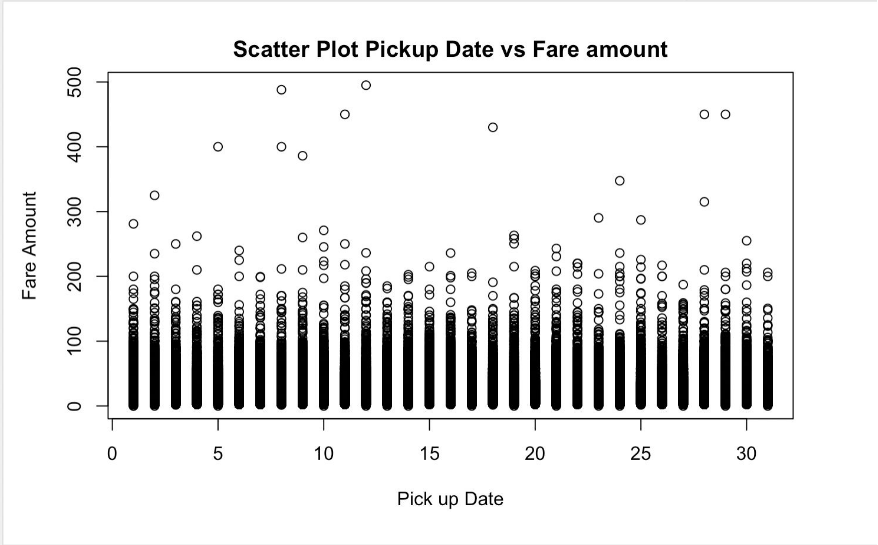
Linear Regression

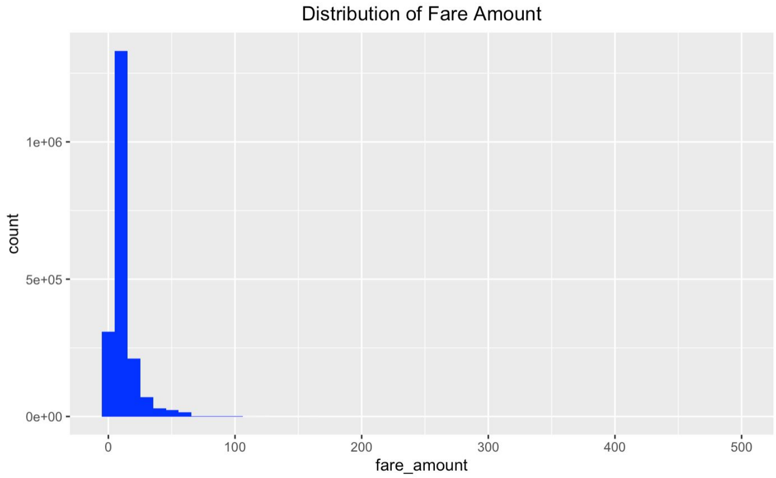
- Building the linear model on the entire dataset, and predicting the fare amount based on the train dataset. The RMSE score for the linear model is 4.2, MSE score is 17.8 and R-square is 0.68



Time Series

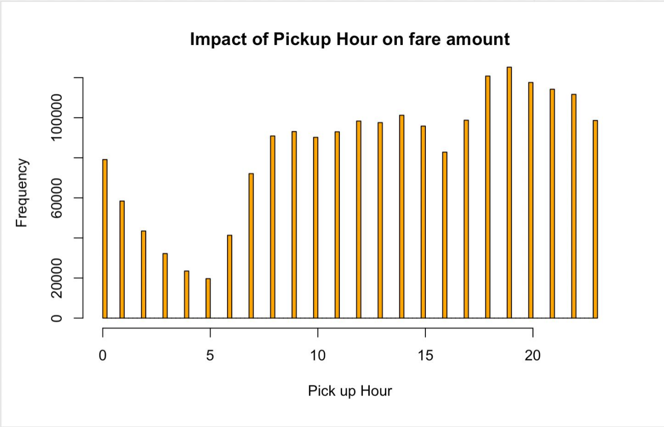
- Impact of the Pick up date on Fare Amount

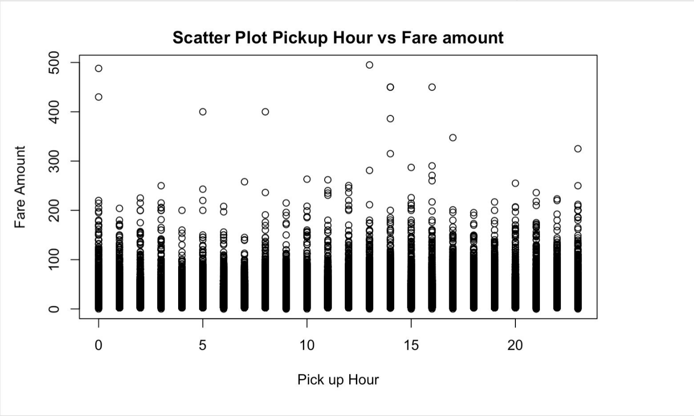




Time Series

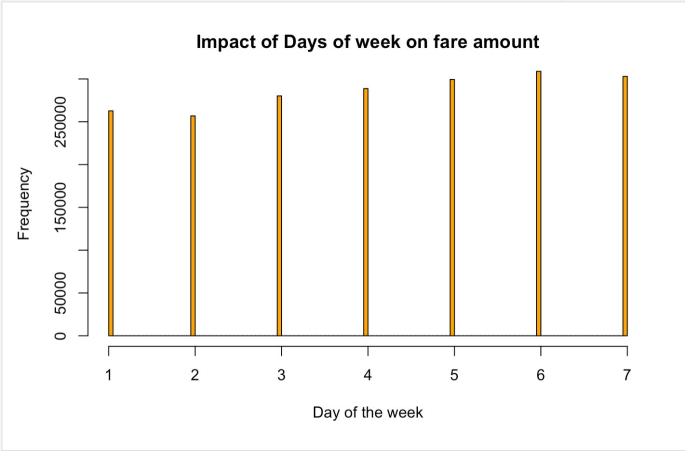
- Impact of Pick up hour on Fare Amount

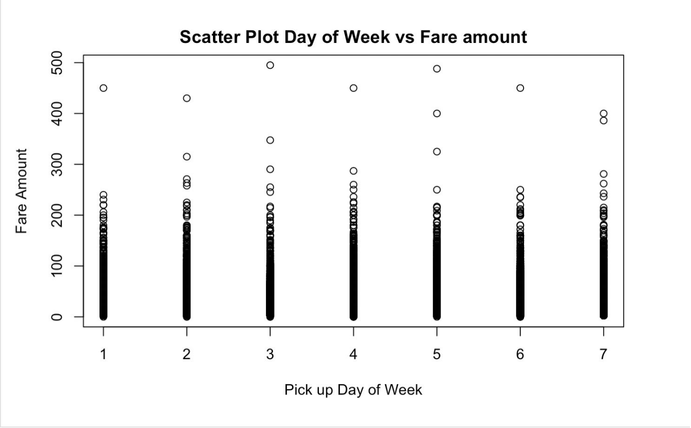




Time Series

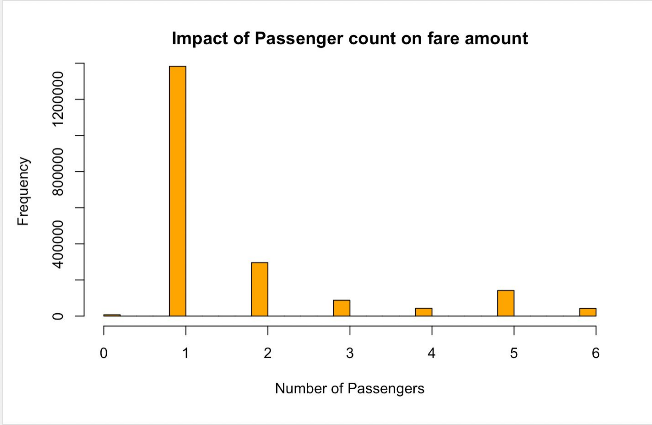
- Impact of Days of week on fare amount

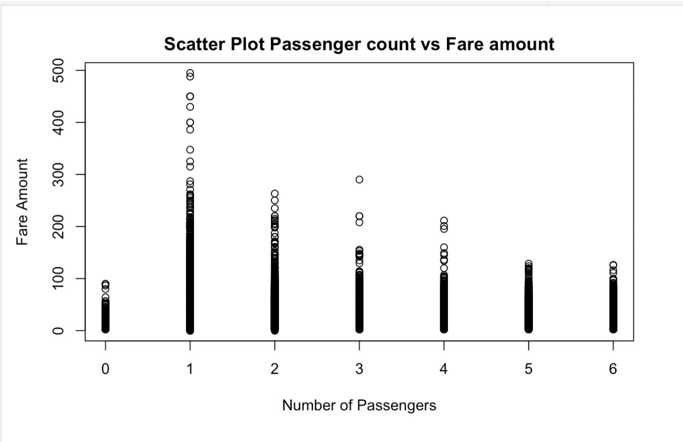




Time Series

- Impact of Passenger count on Fare Amount





Feature Engineering for Time Series Analysis

- Two versions based on features for time series analysis

	date_ym	fare_amount
1	2009-01-01	242303.5
2	2009-02-01	242141.0
3	2009-03-01	259715.2
4	2009-04-01	254463.8
5	2009-05-01	267341.0

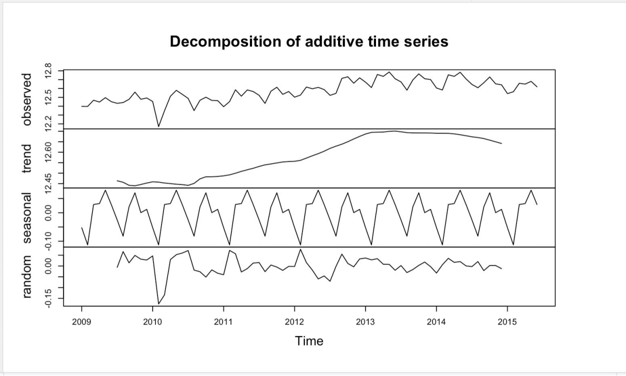
Fig. year_month dataframe

	date_hr	fare_amount
1	2009-01-01 00:00:00	266.30
2	2009-01-01 01:00:00	216.30
3	2009-01-01 02:00:00	176.30
4	2009-01-01 03:00:00	182.20
5	2009-01-01 04:00:00	186.90

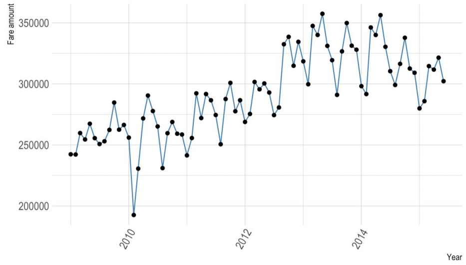
Fig. hourly dataframe

Feature Engineering for Time Series Analysis

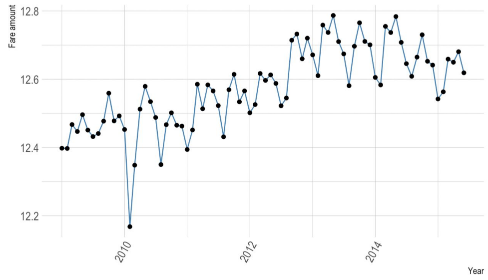
- Analyze the year_month data.There's an upward trend and we can also observe a seasonal pattern which is repeating every year



Monthly time series

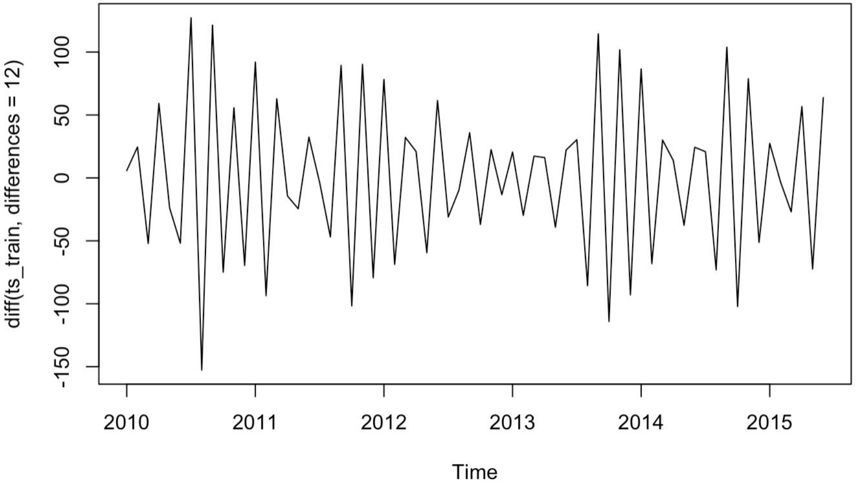


Log transformed Monthly time series



Feature Engineering for Time Series Analysis

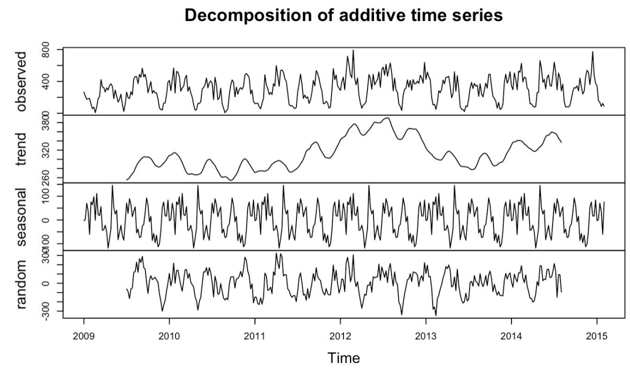
- Stationary year_month time series



Augmented Dickey Fuller
Test - ADF
P - value = 0.01

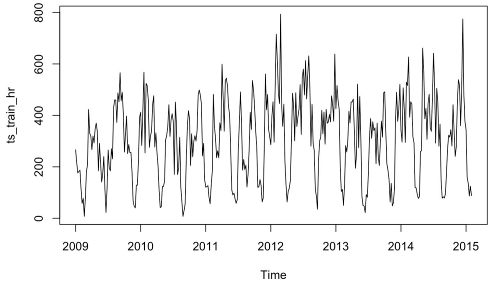
Feature Engineering for Time Series Analysis

- Analyze the hourly time series.There's an upward trend and we can also observe a seasonal pattern which is repeating every year

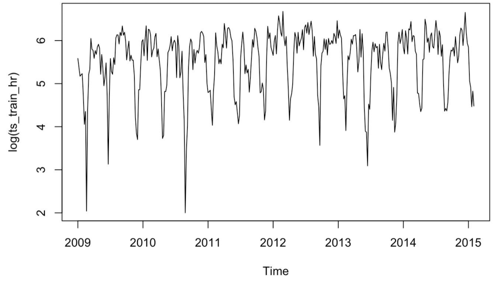


Augmented Dickey Fuller Test - ADF
P - value = 0.01

Hourly time series

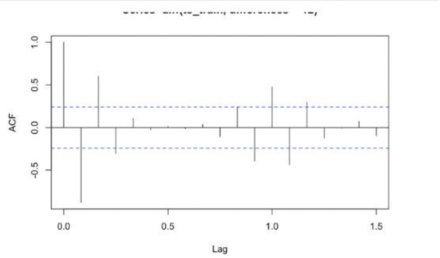


Log transformed Hourly time series

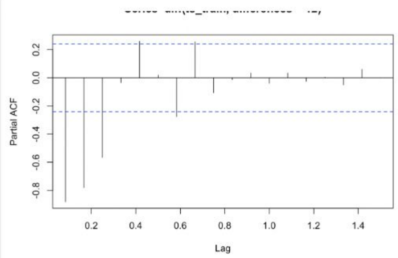


Model selection for Time Series Analysis

- Year Month Time series - ACF and PAC

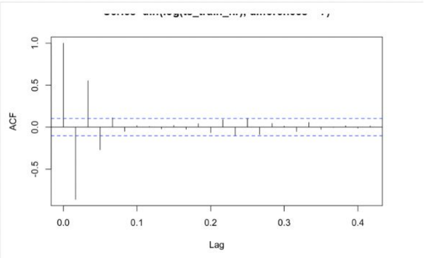


ACF plot for Monthly TS

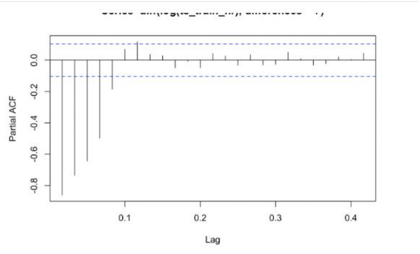


PACF plot for Monthly TS

- Hourly Time Series - ACF and PACF



ACF plot for Hourly TS



PACF plot for Hourly TS

Model Selection for Time Series Analysis

ARIMA

AR - Auto Regression models consider lags , meaning we are trying to predict something for today based on its value on previous days. AR Models capture a pattern and predict the future values.

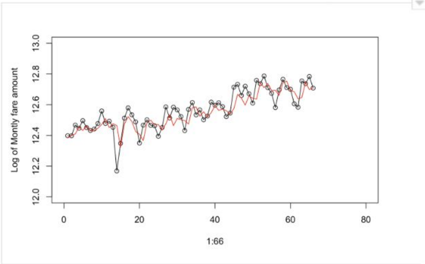
I - Differencing can help stabilize the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality.

MA - Moving Average or Rolling Mean model considers time period t impacted by unexpected external factors in previous time slots. These impacts are called as Errors or residuals and the MA model predicts the future values by considering these residuals from the past data.

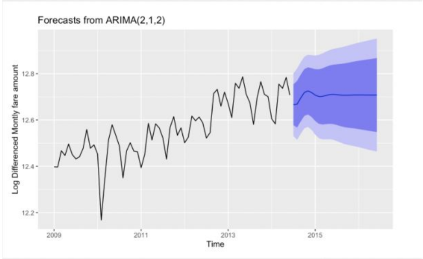
Model Validation for Time Series Analysis

MONTHLY FARE AMOUNT FORECASTING RESULTS

1] Model 1 : (p=2,d=1,q=2)

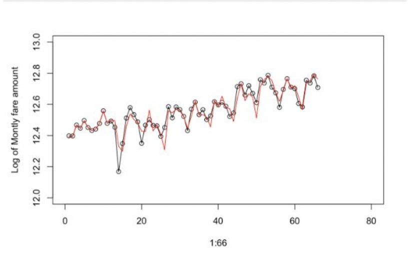


Train forecasts (red) for Monthly Fare Amount - Model 1

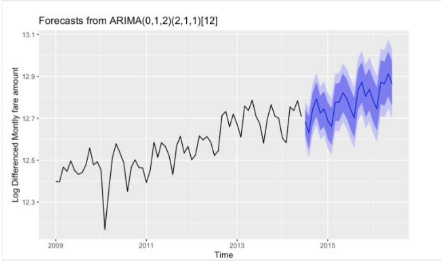


Test forecasts(blue) for Monthly fare amount - Model 1

2] Model 2: order of (0,1,2) and seasonal order of (2,1,1)



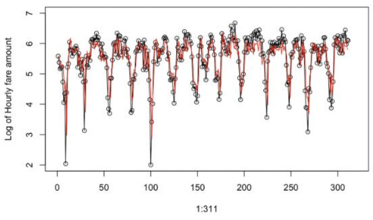
Train forecast(red) for Monthly fares - Model 2



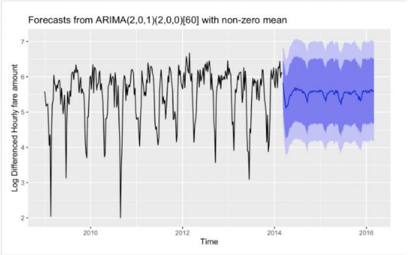
Test forecast for Monthly fares - Model 2

HOURLY FARE AMOUNT FORECASTING RESULTS

3] Model 3 : order(2,0,1) and seasonal order (2,0,0)



Train forecast(red) for Hourly fares - Model 3



Test forecast(blue) for Hourly fares - Model 3

Model Evaluation for Time Series Analysis

Models	MLE	AIC	RMSE		MAPE	
			Train	Test	Train	Test
Model 1 - monthly fare	99.1	-188.2	0.068	0.80	0.0039	0.0052
Model 2 - monthly fare	97.45	-182.89	0.042	0.108	0.0022	0.0078
Model 3 - hourly fare	-240.72	494.44	0.480	0.5976	0.0729	0.0947

Conclusion

- As we performed predictive analysis, we realized that time series models are better at forecasting future taxi fares when compared to linear models. There's definitely an upward trend present in the dataset which comes from the fact that taxi prices are shooting up on a yearly basis.
- There's a high amount of taxi fare ratio for single passengers. The taxi fares are highest during early mornings and evenings as people travel to work and airports or famous locations during this time. We observed crucial factors impacting taxi fares during month of year where people tend to travel more during vacations and holiday seasons and the demand for taxis increases.
- The time series model we built was able to predict monthly fare amounts accurately. The pickup and drop off location features added for cluster analysis, such as famous locations in NYC - airports, parks and tourist attractions, also contribute towards the hike in fare amounts. Outliers related to distance features were identified and removed for further analysis.