Anomaly Detection in Time Series (MTH 442A)

Instructor: Dr. Amit Mitra
Team Members:
Madhur Bansal (210572)
Chirag Garg (210288)
Paritosh Pankaj (210702)
Manav Reddy (210583)
Vijay Soren (211163)

1 Anomaly Detection: Motivation

Anomaly detection is a traditional practice in data science and machine learning domain. It is concerned with identification of data points, observations or events which deviate from the ideal behaviour of data. Time Series Analysis has found it's applications in diverse industrial domains spanning from medical and finance industry to fields like astronomy and weather forecasting.

Looking at the disparate driver applications of time series analysis, detection of anomalies or outliers in time series data becomes an important concern.

2 Procedure

Here are the following steps which we have taken to perform anomaly detection.

- Rearrange the data into daily basis.
- Then we check for trend and and try to remove trend from data.
- After removal of trend, we look for presence of seasonality and try to remove it.
- After this, we perform residual analysis.
- Look for presence of stationarity of data.
- We try to fit ARIMA model into our data by determining adequate values of lags of auto regressive and moving average processes respectively.
- The final step is identification of anomalies based on a threshold value.

3 Experiments and Results

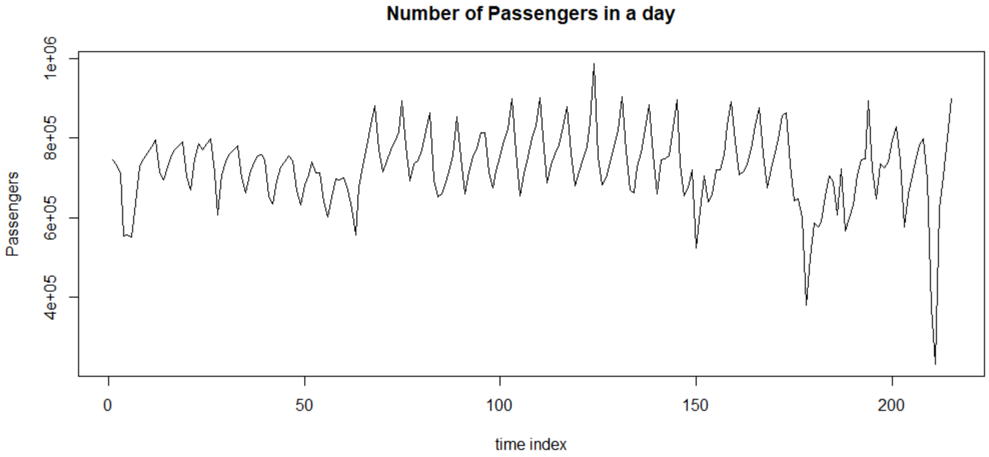
Here we are going to work with the *New York Taxi Dataset* which contains the number of taxis used by public from July 2014 to January 2015. The frequency of taxi's used is calculated every half an hour each day in our dataset.

While working with this dataset, we essentially look to solve two major problems:

- **Problem 1**: Look for the days in the entire time frame when the use of taxis served as an outlier to entire data, that is taxis were used more frequently than normal or when the use was very less
- **Problem2**: Suppose you are a taxi driver and you want to work for certain number of hours on weekends and weekdays, then what should be the appropriate time intervals for operation

For solving the first question, we had to look for daily dataset. Since the data was recorded at every half an hour interval, every data has 48 time points, to obtain the daily data we averaged the original data over 48 time points subsequently.

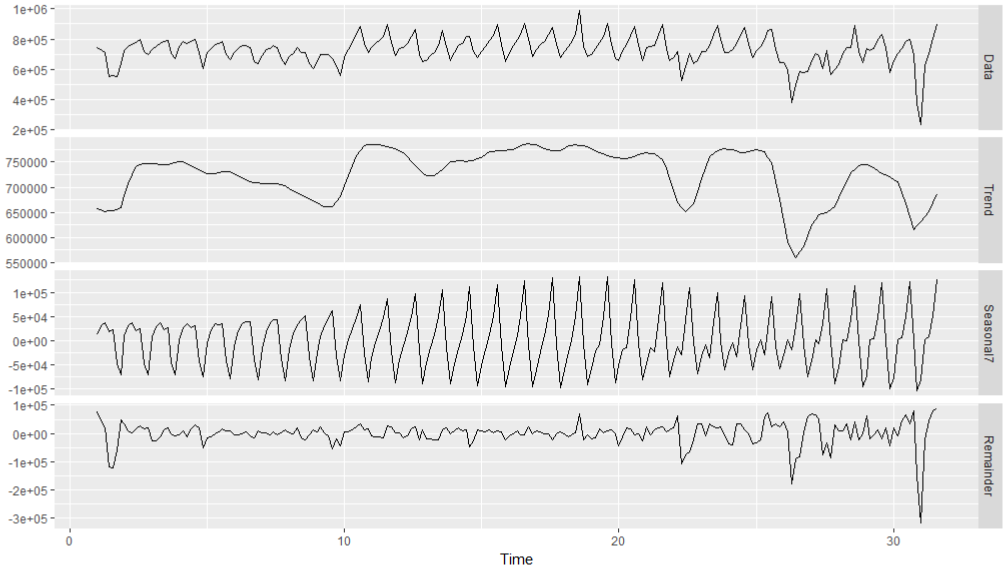
Here is the visualization of the day wise data:



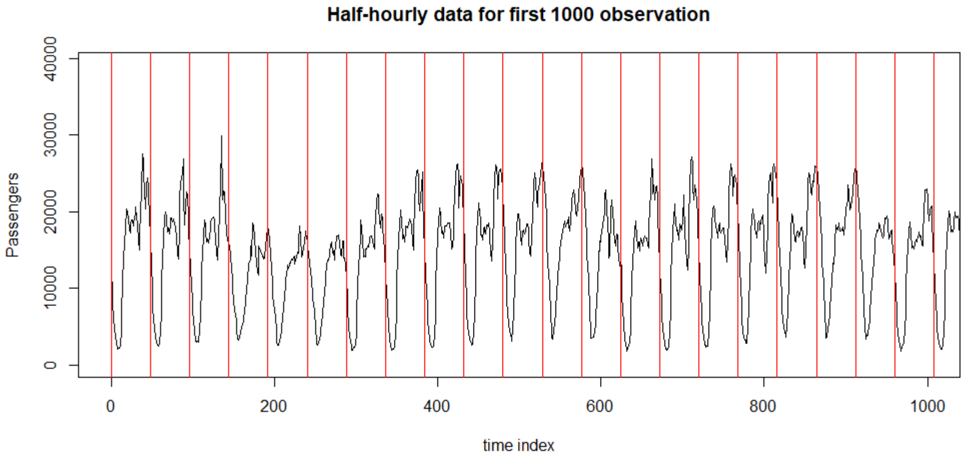
3.1 Trend and Seasonality

For checking presence of trend in the data, we performed "relative ordering test".Although we found that no trend was present in the data. For detection of seasonality, we performed Friedman test.

We also present a visual snapshot of the visualization and trend components of the original data:



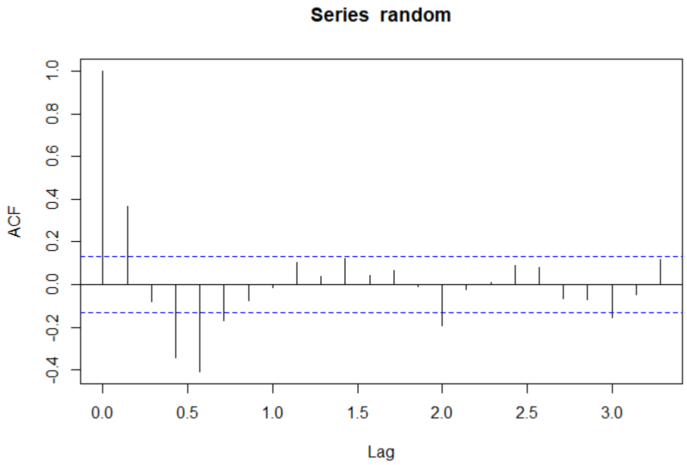
The presence of seasonality in our data can be verified from the following plot which shows first 1000 data points of the original data:



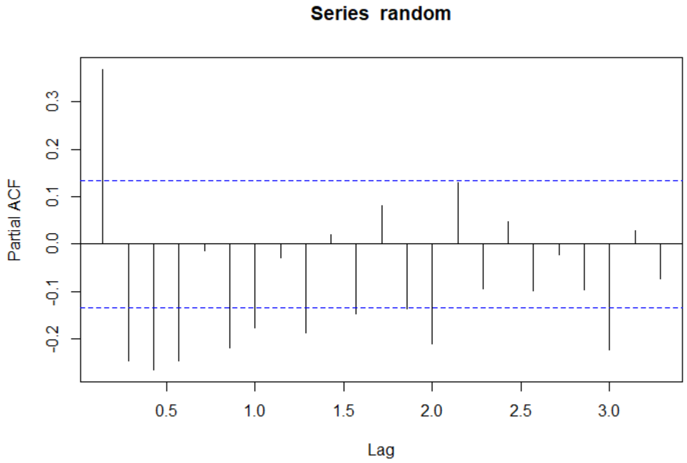
3.2 Fitting ARIMA model

We performed tests for stationarity in the data we performed Augmented Dicky Fuller Test (ADF test), KPSS (Kwiatkowski–Phillips–Schmidt–Shin) test and PP (Phillips–Perron) test. All the three tests suggested that the available data is stationary. After establishing the fact that the data is stationary, we tried to fit an ARIMA(Auto Regressive Integrated Moving Average) model into the data with suitable lag values.Clearly for every method which we are going to implement we know that the value of d parameter is going to be zero as we have already concluded from our previous results that our data is stationary. For finding the lag values, we first tried to plot the ACF (Auto correlation function) and PACF (partial auto correlation function) plots for the data. But we did not get any significant result from the plots. After that we moved to the information theoretic measures. We used AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion).

Here is the plot for ACF (Auto Correlation Function) of the data:



and here goes the plot for partial ACF of the data:

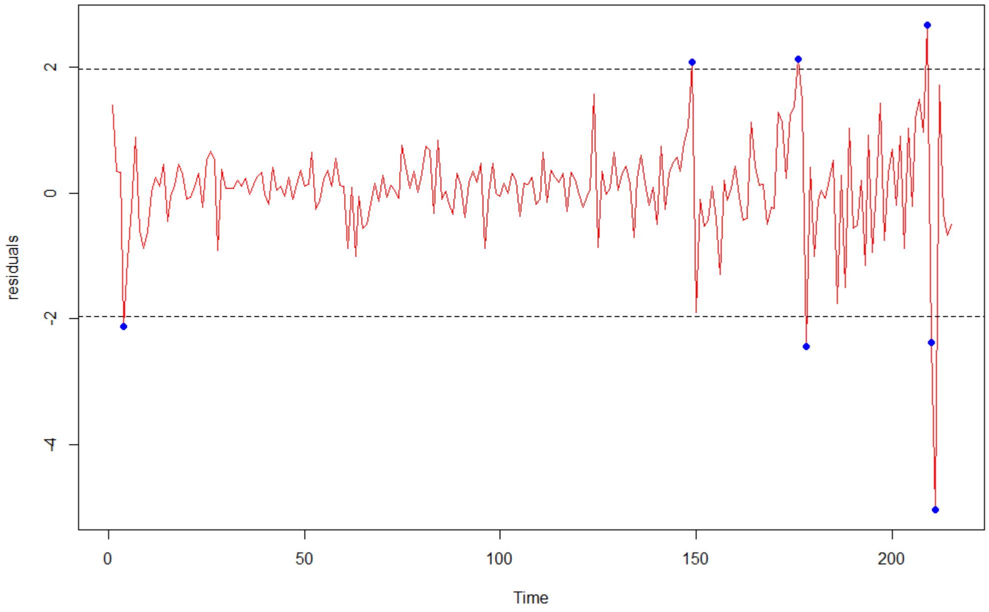


After implementing the methods, we found that the suitable lag values of p (autoregressive lag) is 2 and q (moving average lag) is also 2, So finally we fitted an ARMA(2,2) model into our data.

3.3 Outlier Detection

After fiitng the ARMA(2,2) model to our data, we calculated the errors for every timestamp in our data. We standardized those error values and assumed that the errors follow a standard normal distribution. Using this assumption, we performed hypothesis testing. The threshold for rejection of null hypothesis was α value 0.05. In the end we found out that there were 7 anomalies in total, that is, 7 days which served as outlier to our data. These 7 days were:
4th July 2014
26th November 2014
23rd December 2014
25th December 2014
25th January 2015
26th January 2015, and
27th January 2015.

Here , we present a visualization of the anomalies of our taxi data.



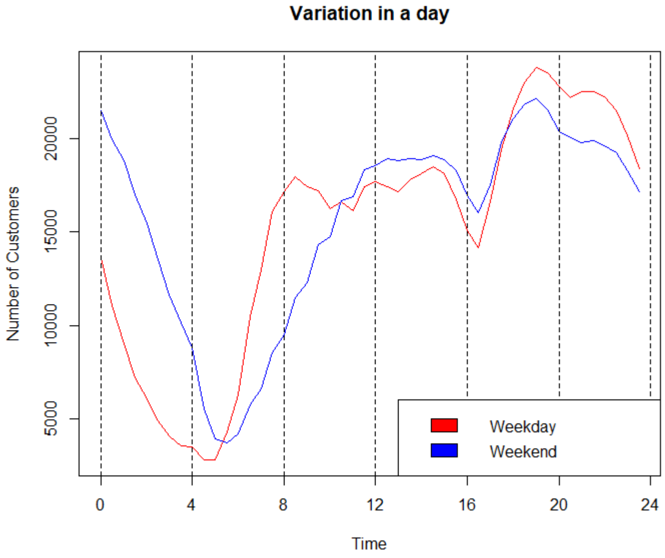
3.4 Result of Problem 2

As discussed earlier, we were trying to find out, from a cab driver's point of view, what time intervals would it be suitable for working on weekdays and on weekends. We have come up with the following results.

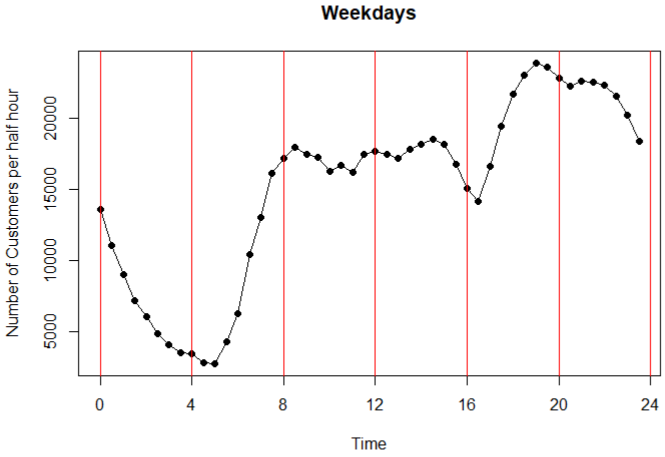
On weekdays, it is profitable to work at following intervals:

- During night, the time interval from 6:00 pm to around 11:00 pm has more passengers.
- During noon, the time interval from 8:00 am to around 3:00 pm has more passengers.

A visual analysis of the combined weekend and weekday result is presented here:



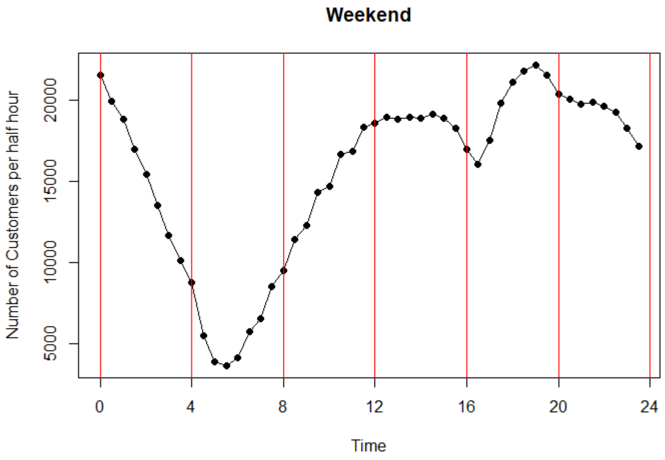
It can be observed that, weekdays have more working hours during the day. While the weekends have more working hours during the night.



In weekends, it is profitable to work on following time intervals:

- During night, the time interval from 6:00 pm to around 1:00 am has more customers.
- During noon, the time interval from 10:00 am to 4:00 pm has more customers

We also present a visualization of the weekdays and weekend data separately to ensure the authenticity of our results.



4 Conclusions

Finally, we are able to find that all the anomalies in our data hold some special importance in USA's and New York's history. 4th July is America's independence day, 26th November is America's thanksgiving day, the final week December serves as Christmas and New Year's Eve while the remaining three days in the anomaly list, that is, 25th, 26th and 27th January 2015 were witness to massive snow storm.

Moreover, when it comes to the analysis of taxi driver's working hours, our results show that during weekdays, the number of passengers is more in morning. This is perhaps because people have to leave for work early in the morning. While on weekends, the number of passengers is more on midnight as most people may like to spend more time outside their household and enjoy their lives.