DIVVY BIKE SHARE ANALYSIS

AKHIL, ASHWIN, POOJITHA, SENTHIL

CSP 571 DATA PREPARATION AND ANALYSIS

PROBLEM DEFINITION

- **Description**

Divvy is a bicycle sharing system in the City of Chicago operated by 'Motivate' for the Chicago Department of Transportation. It operates 5800 bicycles at 580 stations. Divvy is a fun and affordable way to get around Chicago with a big customer base.

- **Problem Statement**

To identify patterns in Divvy bicycle usage to benefit Divvy's business operation

- **Goal**

- To build a time series regression model using Auto Regressive Integrated Moving Average method(ARIMA)
- Forecasting and plotting of the obtained model for the upcoming year(2018)

DATA PREPATATION

Each row in the dataset is recorded from an individual's usage with the below columns

Variable Name	Description
trip_id	ID attached to each trip taken
starttime	day and time trip started
stoptime	day and time trip ended
bikeid	Id attached to each bike
tripduration	time of trip in seconds
from_station_id	ID of station where trip originated
from_station_name	name of station where trip originated
to_station_id	ID of station where trip terminated
to_station_name	name of station where trip terminated
usertype	"Customer" is a rider who purchased a 24-hour pass, "Subscriber" is a rider who purchased an annual membership
gender	gender of rider
birthyear	birth year of rider.

DATA PREPROCESSING

- **Filling in missing values**
  - Variables such as gender and birth year have the missing values.
  - Fill in with ‘Unknown’ for the missing values in variable gender.
  - For variable birth year, fill in with the mean value for it.
- **Handling date and time**

  For handling date and time we use POSIXt function using the following format ‘%m%d%y %H%M’
- **Dropping off unused levels**

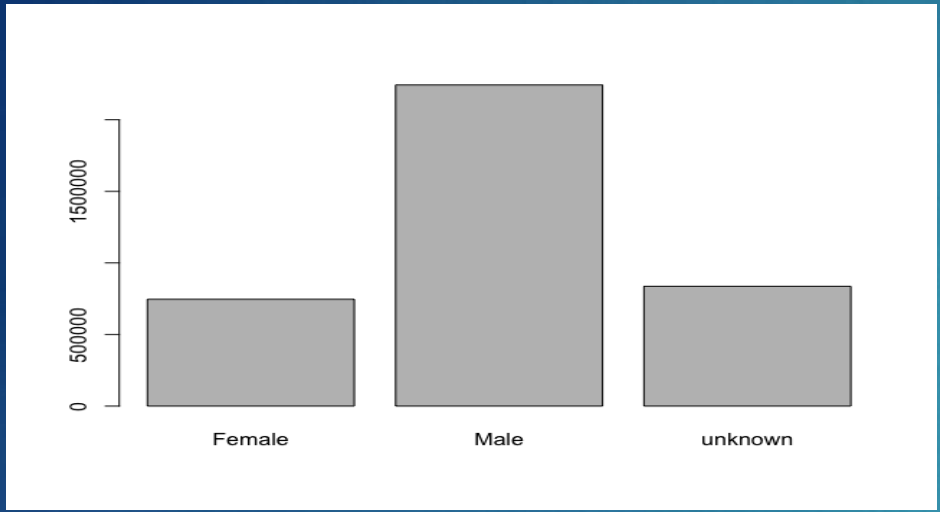
  Among the subscriber, customer and dependent we are dropping levels of dependent as it has very less instance values from the variable user type.

ANALYSIS OF 2017 DATA

Following are some of the analysis done on the 2017 data before performing time series analysis:

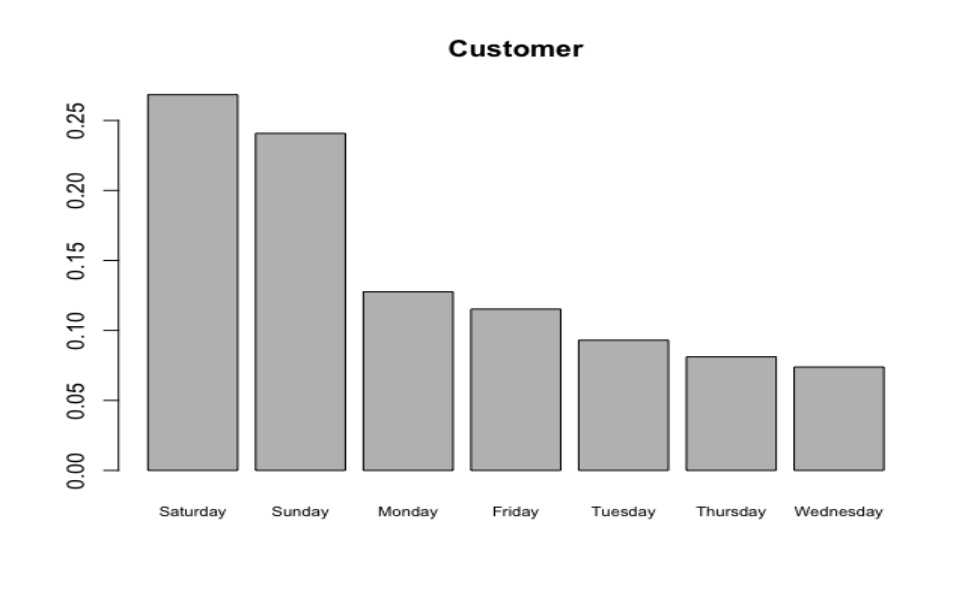
- Analysis on user type and gender variables
- Plotting of top 10 best used stations.
- Plotting of top 10 least used stations.
- Visualization of the top 10 best and least used stations in a map.
- Finding and plotting the frequency of rides for all the days in a week in the year 2017.
- Finding the frequency of rides throughout the year.

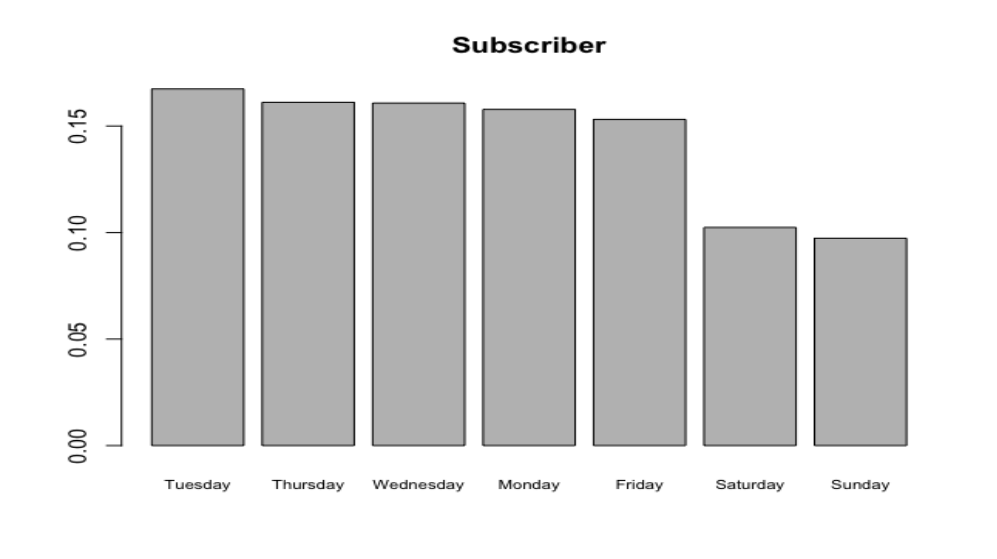
Analysis on user type and gender variables



- The above plot shows the total number of user types based on the gender.
- The below plot shows the number of customers and subscribers after dropping level ‘dependent’.

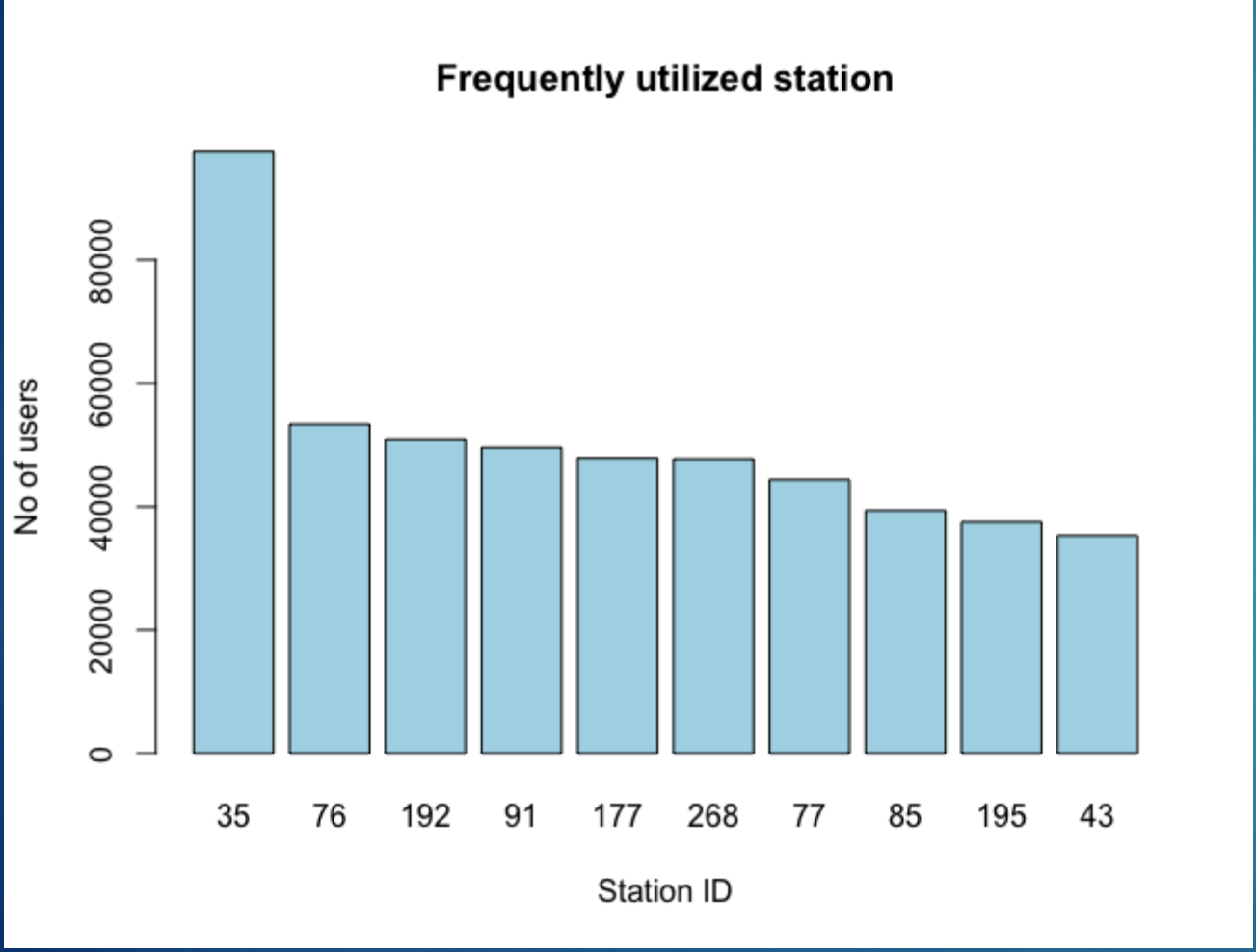
Analysis on user type and time variables





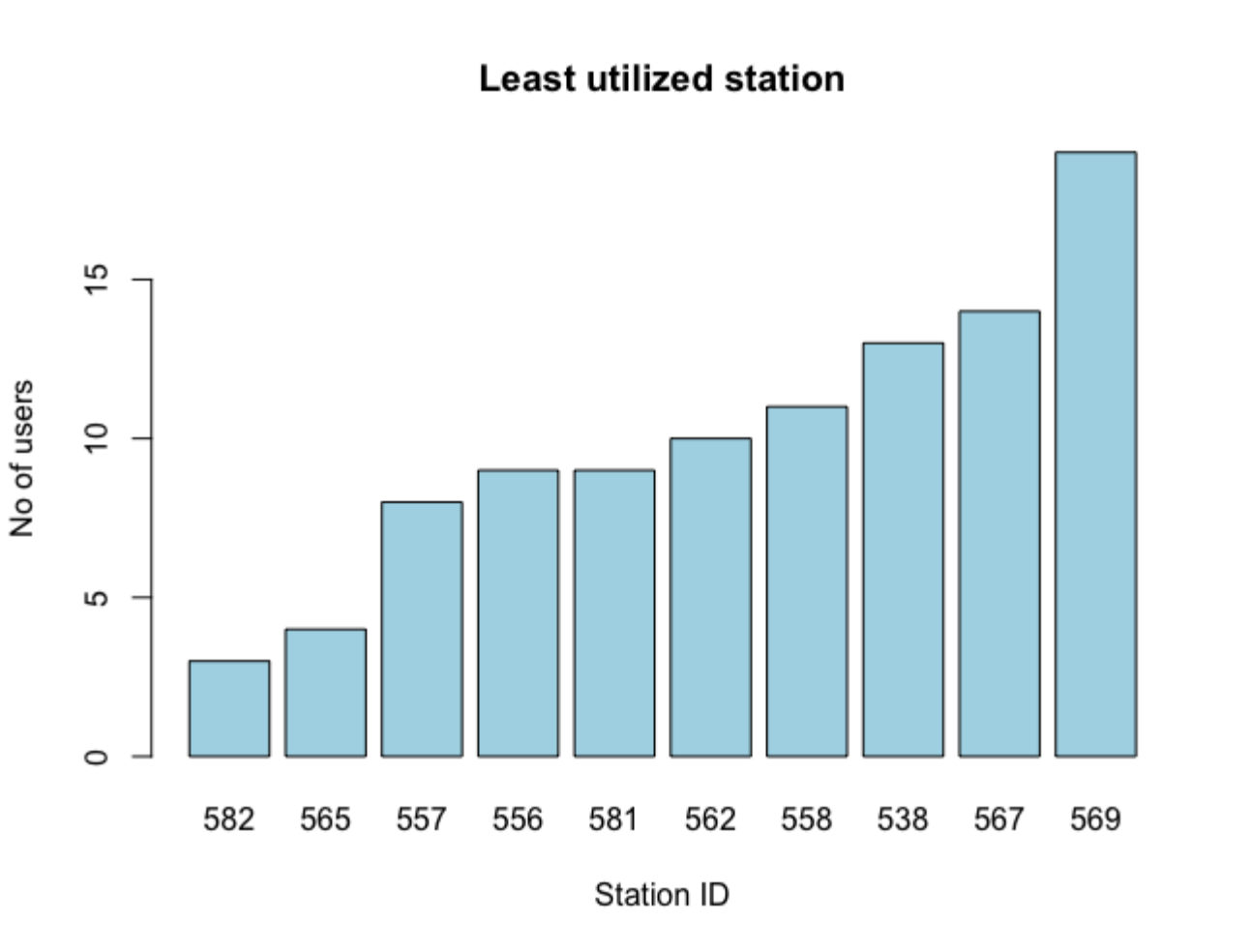
- The graphs show the number of rides done by customers and subscribers on a weekly basis in the increasing order.
- It can be seen that Customers use bikes on Weekends mostly and Subscribers on Weekdays than on the weekends.

Plotting of top 10 best used stations.



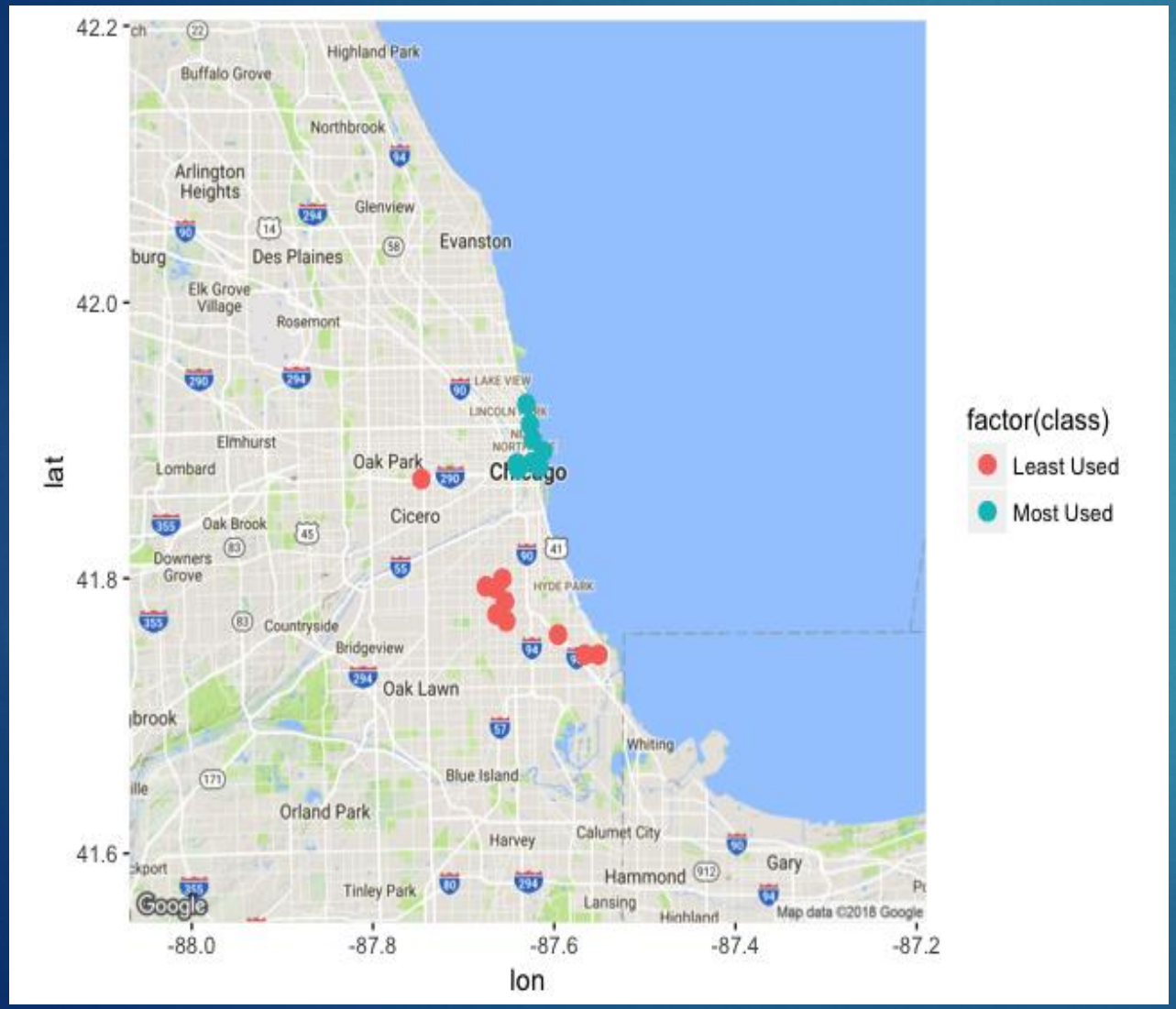
- The bar plot shows the frequently used station with No. of users vs Station ID.
- Among the station IDs, station 35 i.e. ‘Streeter and Dr Ave’ is the best used station.

Plotting of top 10 least used stations.



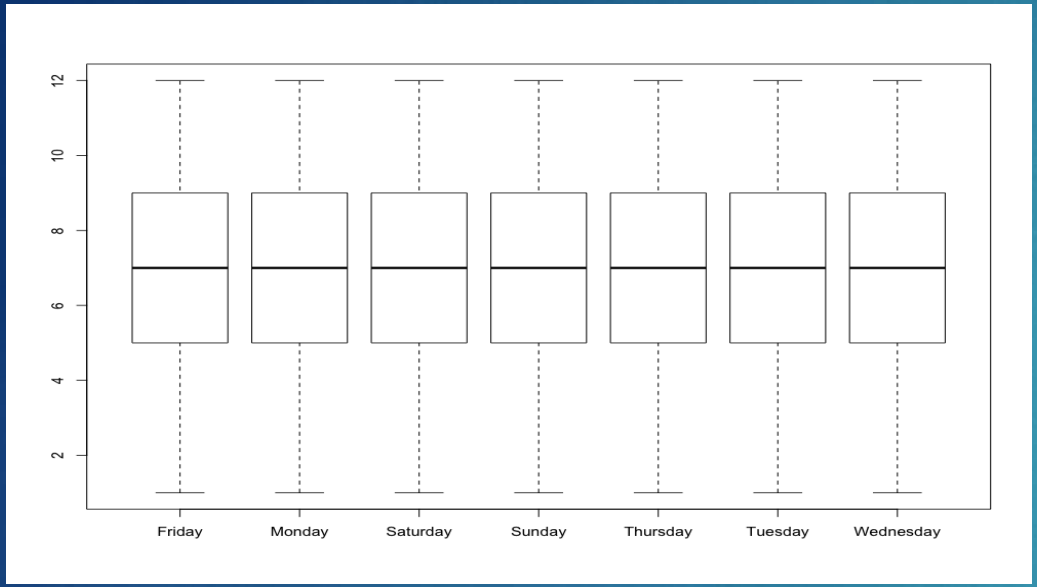
- The bar plot shows the least used station with No. of users vs Station ID.
- Among the station IDs, station 582 i.e. 'Philip Ave and 82nd Street' is the least used station.

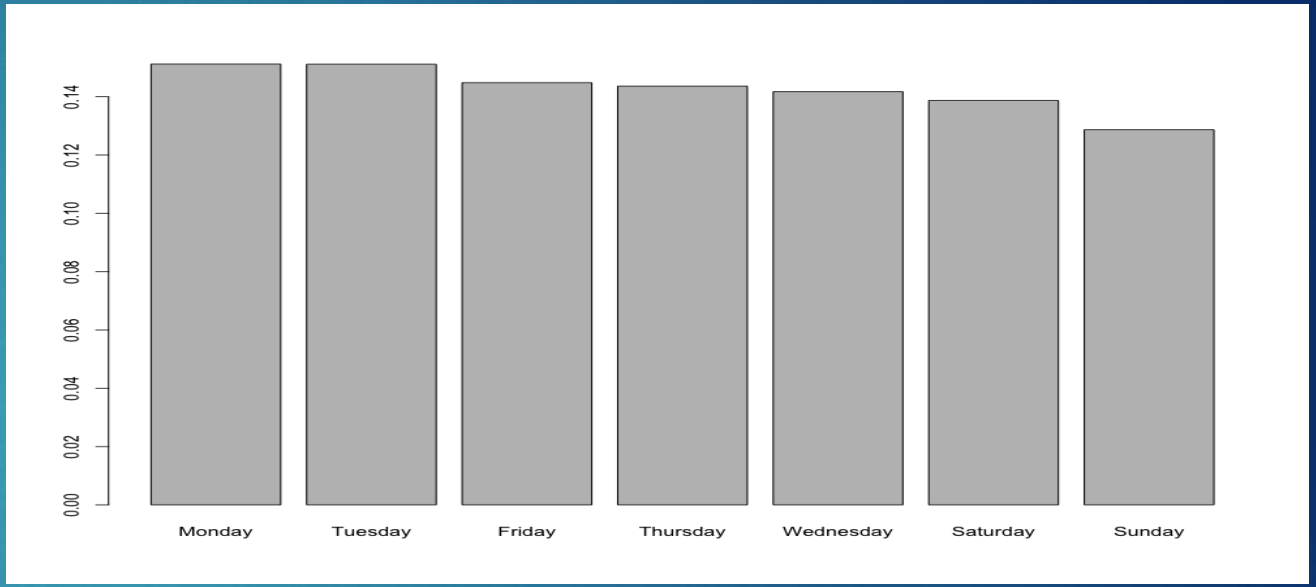
Visualization of the top 10 best and least used stations in a map.



- Bind Divvy trip's data to the Divvy station's data and generate 2 points which matches the top 10 best and the least frequently used stations.
- Bind these 2 points using rbind function and using ggmap function plot them in a map using getmap.

Finding and plotting the frequency of rides for all the days in a week in the year 2017.





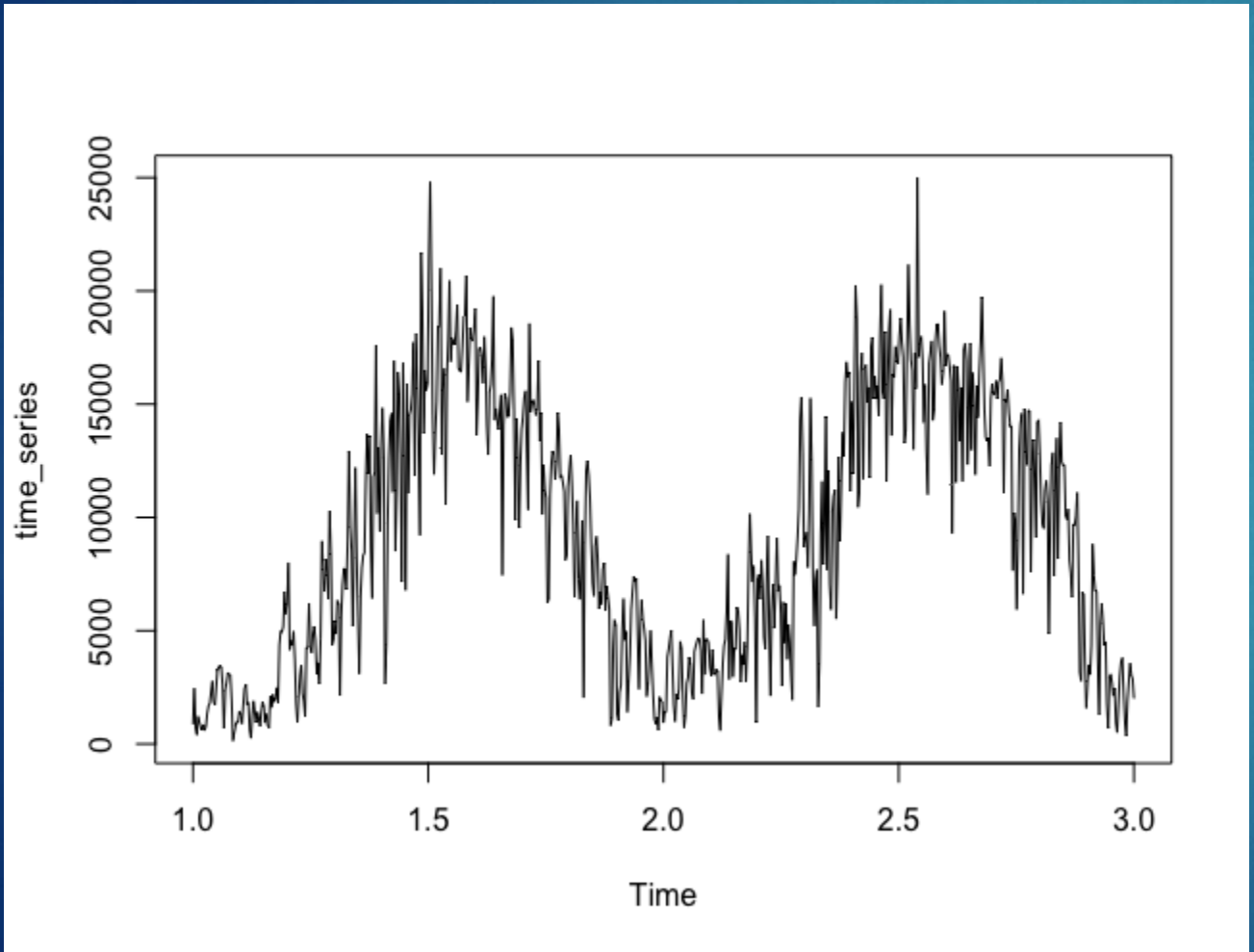
- The bar plot on the right shows the average no. of rides for each day in a week in the increasing order.
- The plot on the left is obtained by comparing the previous output with the frequency of the starting day of the month in that year in an increasing order. Therefore, bikes used on Fridays are comparatively higher.

TIME SERIES ANALYSIS

Following are the steps done for the time series analysis:

- Generating a time series using ts function and plotting the same.
- Decomposing time series.
- Plotting of the decomposed time series.
- Computing acf and pacf for the time series.
- Generating ARIMA model for the obtained time series.
- Forecasting and plotting the obtained model.

Generating a time series using ts function and plotting it



- The function ts is used to create time series object. On applying this function we generate time series and plot it.
- Ts function just predicts the time series for the objects present and does not forecast time series object.

Decomposing time series.

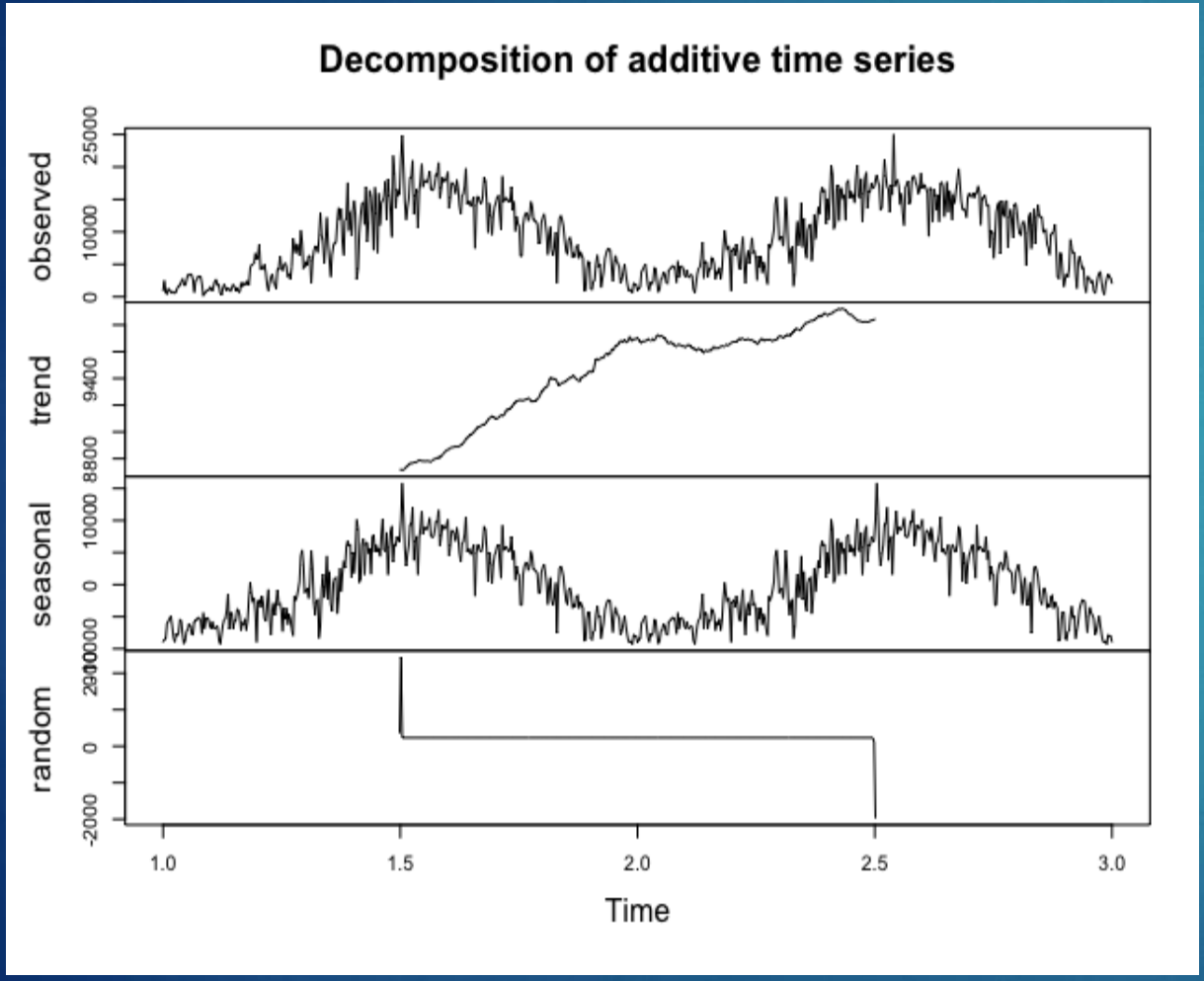
The decomposition of time series is a statistical task that deconstructs a time series into several components, each representing one of the underlying categories of patterns.

- Decomposition based on rates of change:

This is the technique which we have used because it does seasonal adjustments. In this method the time series are decomposed into:

- T_{t}, the trend component at time t, which reflects the long-term progression of the series . A trend exists when there is a persistent increasing or decreasing direction in the data.
- S_{t}, the seasonal component at time t, reflecting seasonality. A seasonal pattern exists when a time series is influenced by seasonal factors. Seasonality occurs over a fixed and known period.

Plotting of the decomposed time series.

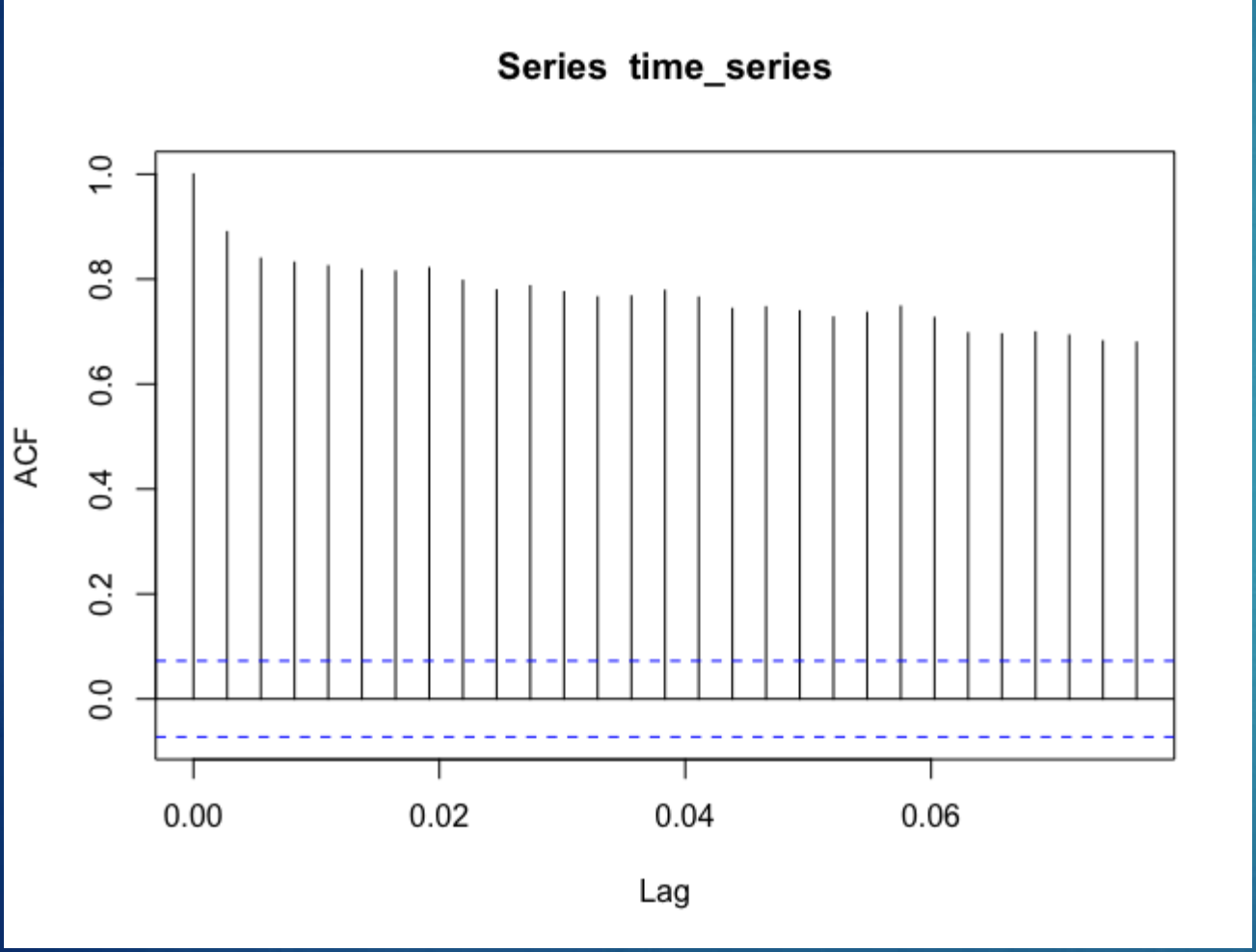


The components in the graph are defined as follows:

- Trend – An increasing or decreasing value in the series.
- Seasonal – The repeating short term cycle in the series.
- Noise – The random variation in the series.

It can be seen from the graph that observed component and seasonal component are pretty much the same with the trend component increasing which makes obvious that the time series prediction depends on the seasonal component.

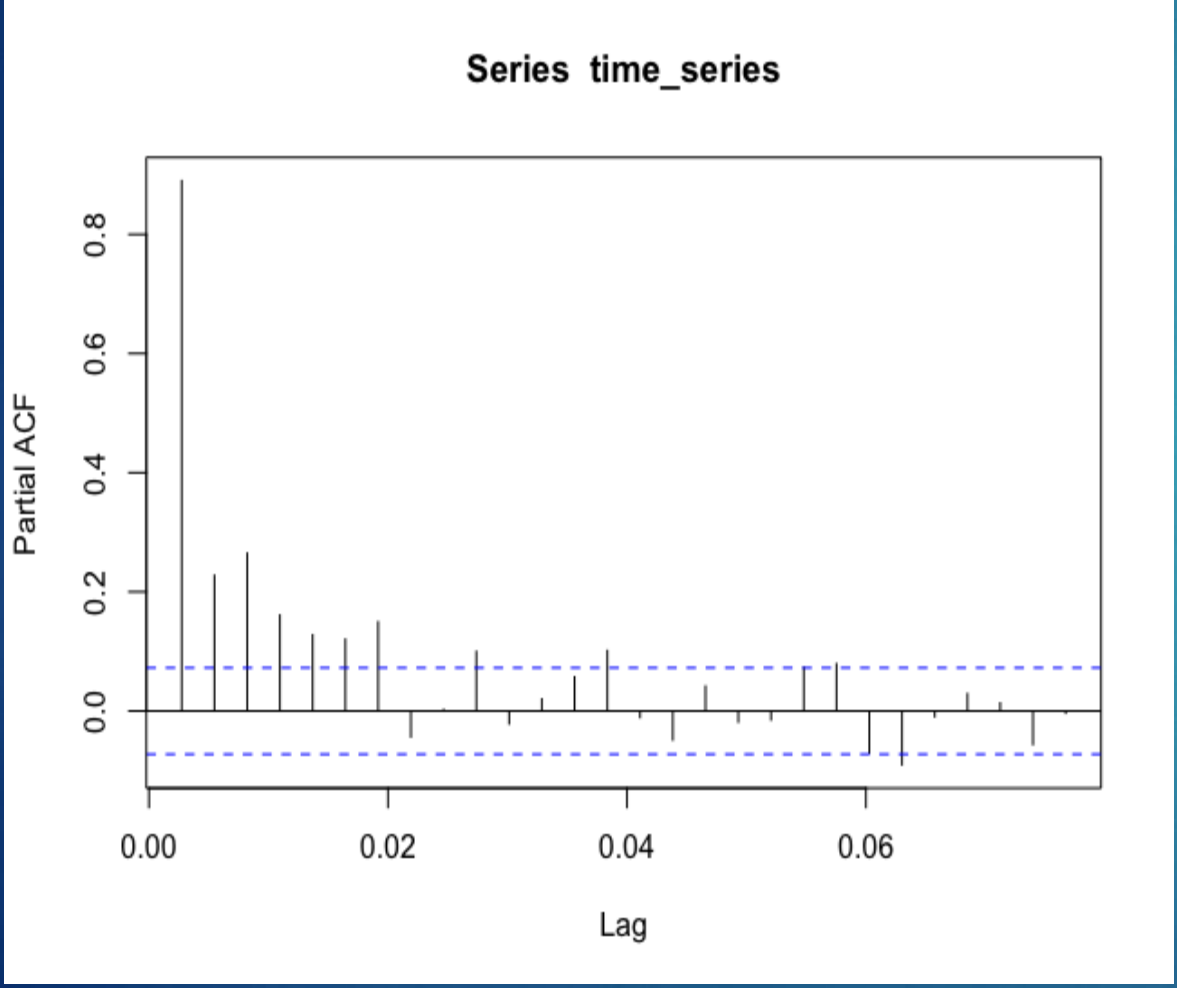
Computing acf for the time series.



The ACF property defines a distinct pattern for the autocorrelations. For a positive value of φ1, the ACF exponentially decreases to 0 as the lag h increases. For negative φ1, the ACF also exponentially decays to 0 as the lag increases

- From the obtained acf graph we can see that the time series is non stationary.
- Hence acf is decaying or decreasing very slowly well above the significance range(dotted blue lines).
- This proves that the series is non stationary.

Computing pacf for the time series.



In time series analysis, the partial autocorrelation function (PACF) gives the partial correlation of a time series with its own lagged values, controlling for the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags.

- From the obtained pacf graph we can see that there are some lags contracting the acf graph.
- Hence counting the lags in pacf to determine the order of the moving time series model.

Generating ARIMA model for the obtained time series.

- What is ARIMA ?

  ARIMA stands for Autoregressive Integrated Moving Average models.

  Univariate (single vector) ARIMA is a forecasting technique that projects the future values of a series based entirely on its own inertia.

  Its main application is in the area of short term forecasting requiring at least 40 historical data points.

  It works best when your data exhibits a stable or consistent pattern over time with a minimum amount of outliers.

- Determining the parameters for the model(p,d,q)

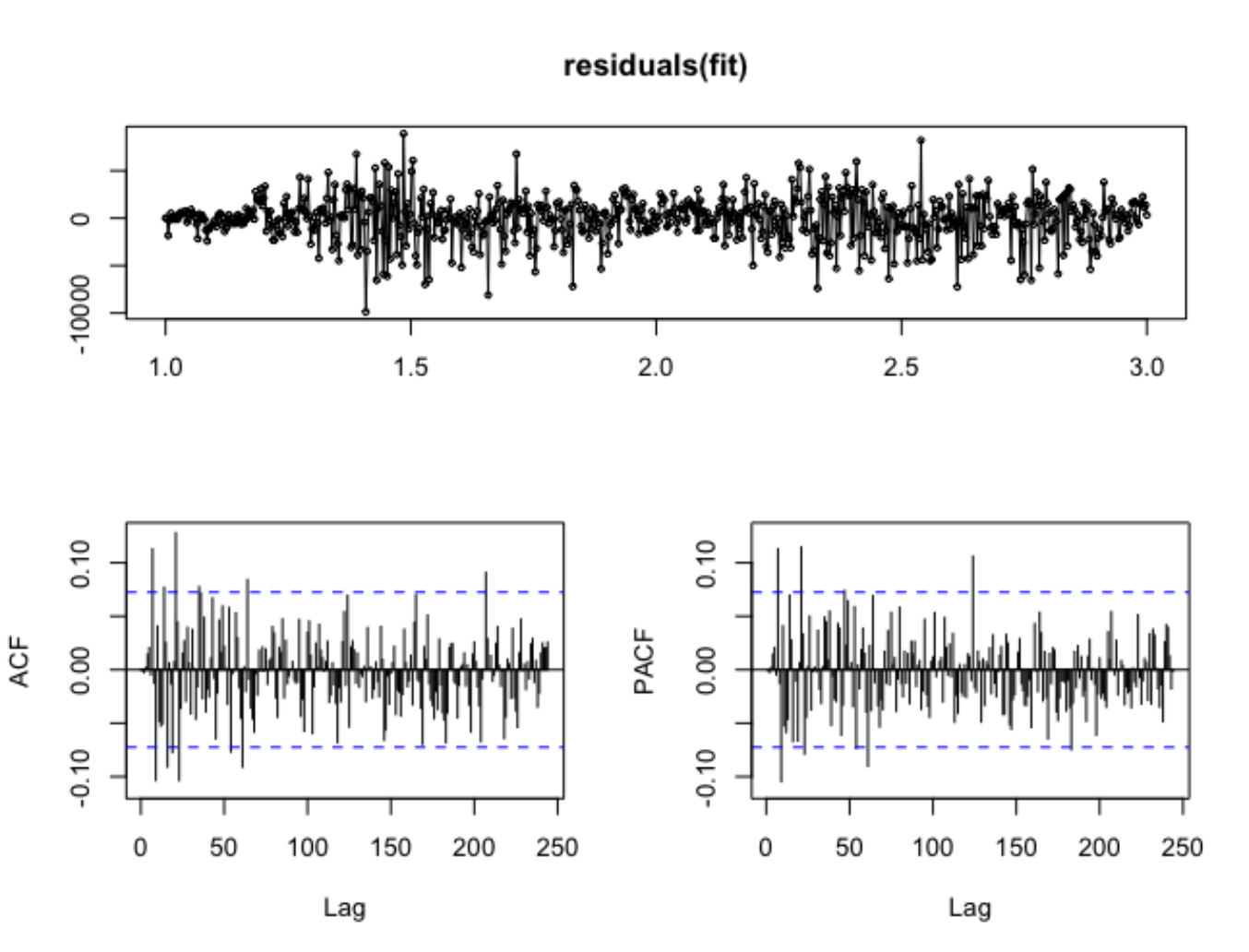
  p refers to the number of autoregressive terms.

  d refers to the degree of differencing.

  q refers to the number of lagged forecast errors.

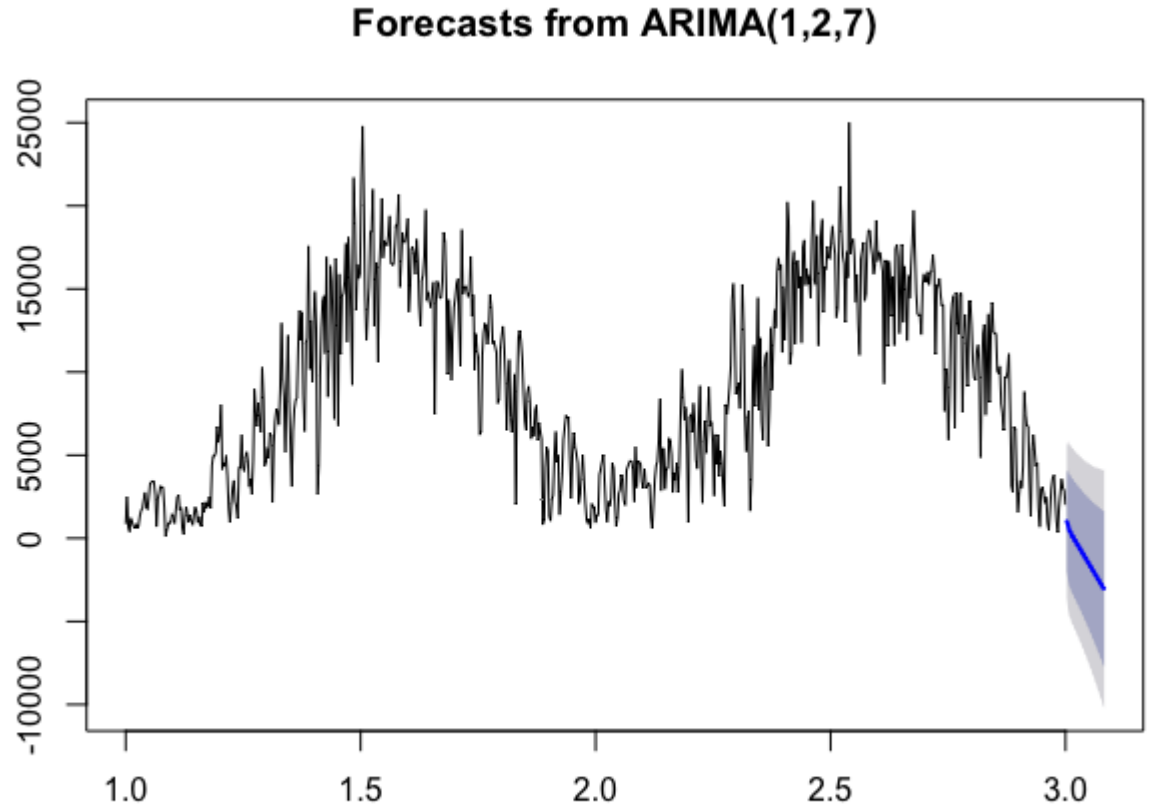
- The model is valuated on the obtained aic or bic value

PLOTTING ARIMA(1,2,7) MODEL



- From the lags in acf and pacf obtained in the graph we can see that there is a lag at 7.
- Hence we can see that the ARIMA model fits better when q=7 and there are no significant lags after q=7.
- The residuals obtained also shows small error range when compared to other parameter values.

Forecasting and plotting the obtained model



- Using forecast function and the obtained ARIMA (1,2,7) model we can predict how the model performs in the future
- From the obtained graph we can see that the forecast is decreasing.
- We conclude that the plotted predictions are based on the assumptions that there will be no other seasonal fluctuations in the data and the change in number of bicycles from one day to another is more or less constant in mean and variance.

THANK YOU

Any Questions?