Aishwary Gawande
196009219

Bike Sharing Analysis

Introduction:

This dataset given contains the parameters such as weather, days, humidity, hourly and daily count of Capital bikeshare system rental bikes between years 2011 and 2015 in in Washington, DC with the corresponding weather and seasonal information. With the help of the given information, it will help us to analyze and track demand of bike sharing and the results will help us to optimize the business strategies. This data was taken from Capital Bike Share website.

Dataset Link: https://www.capitalbikeshare.com/system-data

About the Data:

Data consists of Various Parameters:

```
> summary(day)
    instant          dteday               season            yr              mnth          holiday            weekday          workingday
 Min.   :  1.0   Min.   :2011-01-01   Min.   :1.000   Min.   :0.0000   Min.   : 1.00   Min.   :0.00000   Min.   :0.000   Min.   :0.000
 1st Qu.:183.5   1st Qu.:2011-07-02   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.: 4.00   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.000
 Median :366.0   Median :2012-01-01   Median :3.000   Median :1.0000   Median : 7.00   Median :0.00000   Median :3.000   Median :1.000
 Mean   :366.0   Mean   :2012-01-01   Mean   :2.497   Mean   :0.5007   Mean   : 6.52   Mean   :0.02873   Mean   :2.997   Mean   :0.684
 3rd Qu.:548.5   3rd Qu.:2012-07-01   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:10.00   3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:1.000
 Max.   :731.0   Max.   :2012-12-31   Max.   :4.000   Max.   :1.0000   Max.   :12.00   Max.   :1.00000   Max.   :6.000   Max.   :1.000
   weathersit         temp               atemp              hum            windspeed            casual           registered        cnt
 Min.   :1.000   Min.   :0.05913   Min.   :0.07907   Min.   :0.0000   Min.   :0.02239   Min.   :   2.0   Min.   :  20   Min.   :  22
 1st Qu.:1.000   1st Qu.:0.33708   1st Qu.:0.33784   1st Qu.:0.5200   1st Qu.:0.13495   1st Qu.: 315.5   1st Qu.:2497   1st Qu.:3152
 Median :1.000   Median :0.49833   Median :0.48673   Median :0.6267   Median :0.18097   Median : 713.0   Median :3662   Median :4548
 Mean   :1.395   Mean   :0.49538   Mean   :0.47435   Mean   :0.6279   Mean   :0.19049   Mean   : 848.2   Mean   :3656   Mean   :4504
 3rd Qu.:2.000   3rd Qu.:0.65542   3rd Qu.:0.60860   3rd Qu.:0.7302   3rd Qu.:0.23321   3rd Qu.:1096.0   3rd Qu.:4776   3rd Qu.:5956
 Max.   :3.000   Max.   :0.86167   Max.   :0.84090   Max.   :0.9725   Max.   :0.50746   Max.   :3410.0   Max.   :6946   Max.   :8714
>
```

Instant vary from 1 to 731 with a median and mean of 366.Temperature has a range of 0.05913 to 0.086167 with median of 0.48673.

Humidity and windspeed having median 0.6267 and 0.18097 with max range going upto 0.9725 and 0.50746.

Total no of casual and registered user have a range upto 8714 with mean of 4504.

Aishwary Gawande
196009219

Name of the Parameters

```
> # Various Parameters about the data
> names(day)
 [1] "instant"    "dteday"     "season"     "yr"         "mnth"       "holiday"
 [7] "weekday"    "workingday" "weathersit" "temp"       "atemp"      "hum"
[13] "windspeed"  "casual"     "registered" "cnt"        "date"       "year"
[19] "month"      "day"
```

Data cleaning

```
> print(paste("The total number of missing data are",sum(is.na(day))))
[1] "The total number of missing data are 0"
>
>
```

There are no missing values and data need no cleaning.

Purpose of Analysis

- The purpose of this analysis is to create a predictive model and forecast the future ridership.
- How different parameters such as weather, days, humidity etc affects the overall ridership of registered as well as casual users.
- Determine which model would be best fit to accurately predict the results.

  - Simple Linear Regression

  -Multiple Linear Regression

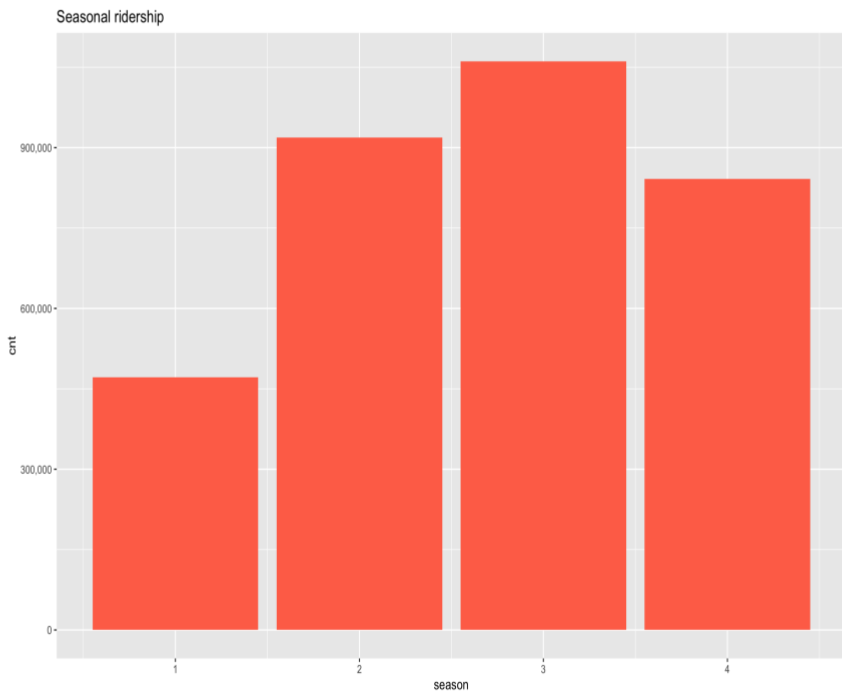
  - ARIMA Forecasting

Aishwary Gawande
196009219

Exploratory Data Analysis:

1. **Seasonal Ridership**

```
season_count<- day %>%select(season,cnt)
point <- format_format(big.mark = ",", scientific = FALSE)
ggplot(season_count, aes(season, cnt))+ geom_bar(stat = "identity", fill="coral1") +
labs(title="Seasonal ridership")+
scale_y_continuous(labels = point)
```



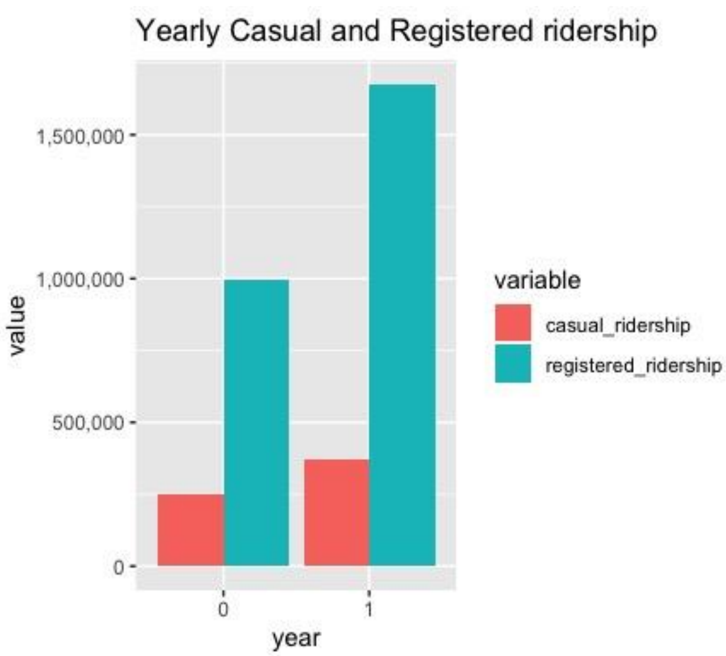
Bikes are least rented in the first quarter(winter) and increases in Spring.

After Spring, there is sudden increase in bike renting during the summer season.

Aishwary Gawande
196009219

2. Yearly Casual and Registered Ridership

```
> day <- Year_count%>%
+   group_by(year) %>%
+   summarise(casual_ridership=sum(casual),
+             registered_ridership = sum(registered))
> day <-as.data.frame(day)
> day$year <- as.character(day$year)
>
> dfm <- melt(day[,c('year','casual_ridership','registered_ridership')],id.vars = 1)
> point <- format_format(big.mark = ",", scientific = FALSE)
> ggplot(dfm,aes(x = year,y = value)) + labs(title="Yearly casual and registered ridership
 ")+
+   geom_bar(aes(fill = variable),stat = "identity",position = "dodge") + scale_y_continuo
us(labels = point)
>
```

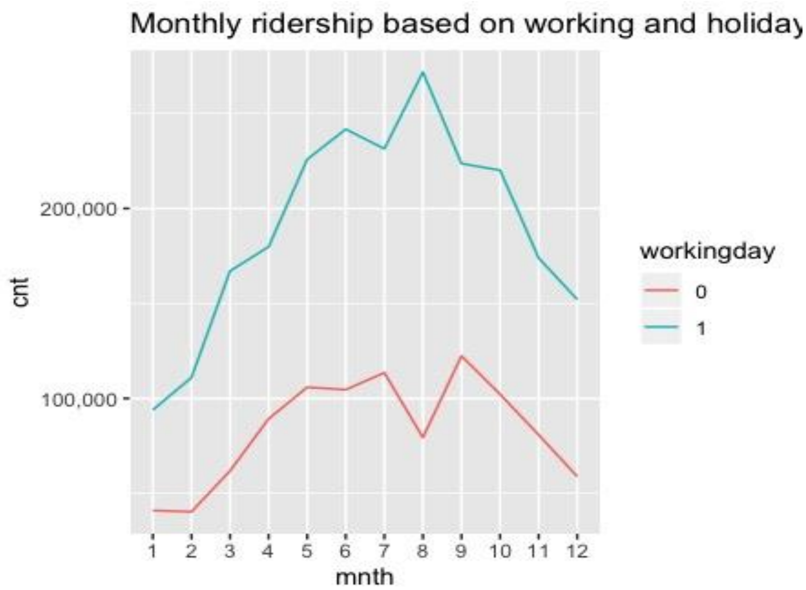


Registered ridership is high in comaprison to the casual ridership in both years.

Aishwary Gawande
196009219

3. Monthly Ridership based on Working and Holiday

```
> day <- mnth_count%>%
+   group_by(mnth,workingday) %>%
+   summarise(cnt = sum(cnt))
> day$mnth <- as.factor(day$mnth)
> day$workingday <- as.character(day$workingday)
> point <- format_format(big.mark = ",", scientific = FALSE)
> ggplot(day, aes(mnth,cnt)) + labs(title="Monthly ridership based on working and holida
y")+
+   geom_line(aes(color=workingday, group=workingday))+ scale_y_continuous(labels = point)
> 
```

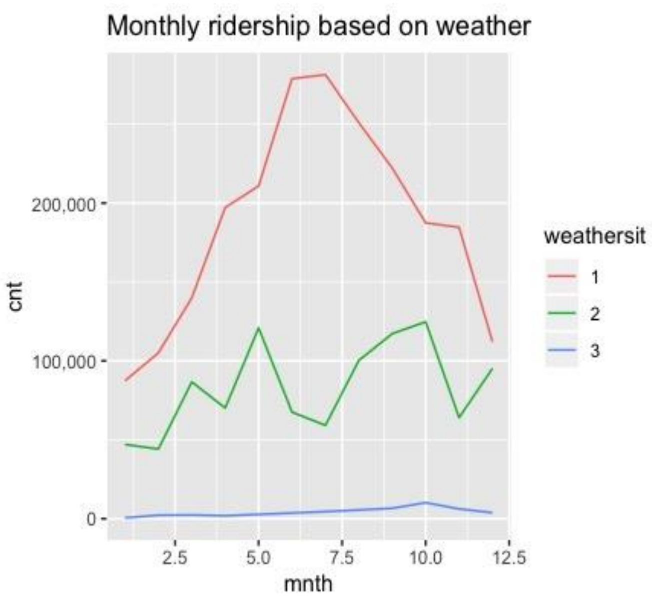


Ridership is maximum in the 8th month i.e. August in working days.

Aishwary Gawande
196009219

4. Monthly Ridership based on Weather

```
weather_count<- day %>%
select(mnth,weathersit,cnt)
weather_df <- weather_count%>%
group_by(mnth,weathersit) %>%
summarise(cnt = sum(cnt))
weather_df$month <- as.factor(weather_df$month)
weather_df$weathersit <- as.character(weather_df$weathersit)
point <- format_format(big.mark = ",", scientific = FALSE)
ggplot(weather_df, aes(mnth,cnt)) + labs(title="Monthly ridership based on weather")+
  geom_line(aes(color=weathersit, group=weathersit))+scale_y_continuous(labels = point)
```

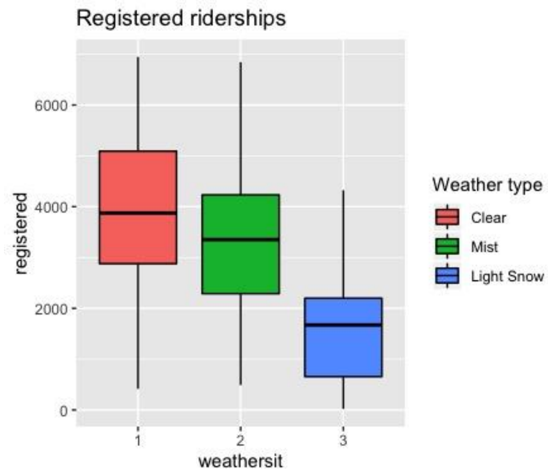


- Weather is clear during the between summer and fall quarter and number of ridership is also high during these month.

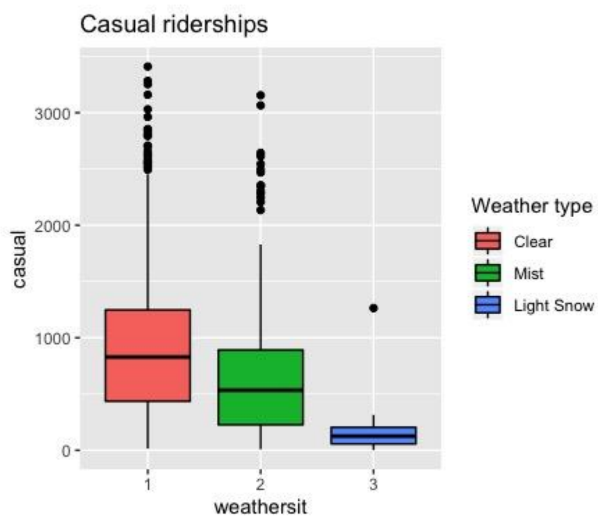
Aishwary Gawande
196009219

5. Registered and Casual user

```
> day%>%
+   mutate(weathersit= factor(weathersit))%>%
+   ggplot(aes(y=registered , x=weathersit, fill=weathersit))+
+   geom_boxplot(colour="black")+labs(title="Registered riderships")+ scale_fill_discrete
(name="Weather type",
+   labels=c("Clear", "Mist", "Light Snow", "Heavy Rain"))
```




```
> day%>%
+   mutate(weathersit= factor(weathersit))%>%
+   ggplot(aes(y=casual , x=weathersit, fill=weathersit))+
+   geom_boxplot(colour="black")+labs(title="Casual ridership")+scale_fill_discrete(name
="Weather type",
+   labels=c("Clear","Mist", "Light Snow", "Heavy Rain"))
```



Ridership is based on weather as well, when weather is clear no of users are also high.

Aishwary Gawande
196009219

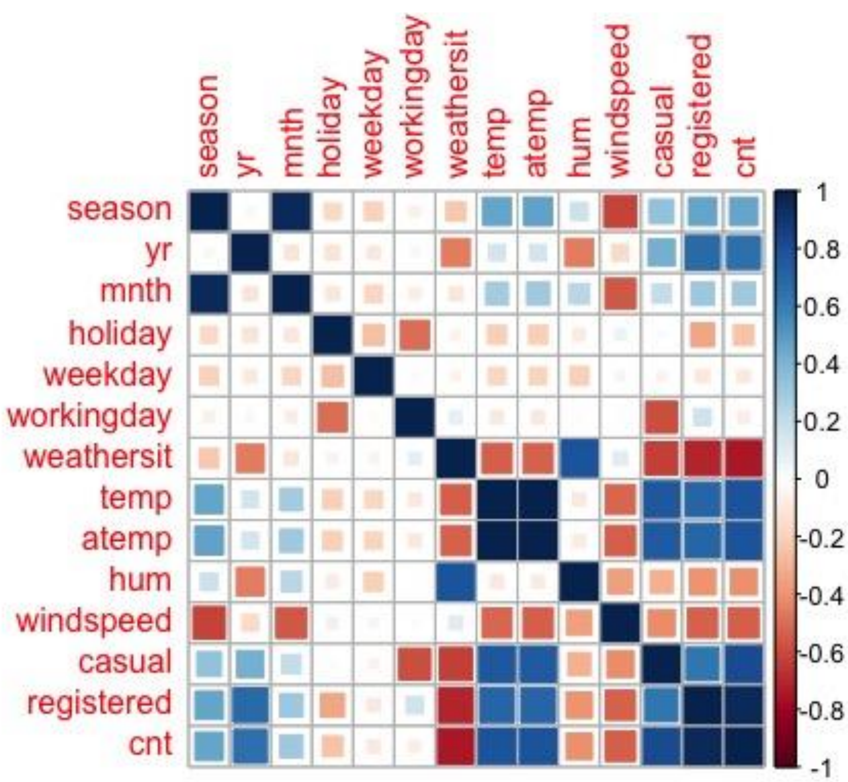
Data Analyzation:

Correlation using Scatterplot:

Correlation:

It is a technique of finding the relationship between two quantitative variables.

```
> #CORELATION MATRIX
> correlation <- mutate_all(day, function(x) as.numeric(as.character(x)))
Warning messages:
1: In (function (x)  : NAs introduced by coercion
2: In (function (x)  : NAs introduced by coercion
> df<- cor(day[,3:16])
> corrplot(cor(df), method = 'square')
```



The above plot shows that there is a higher relation between the casual & registered users and temperature, humidity, month, year and season.

Aishwary Gawande
196009219

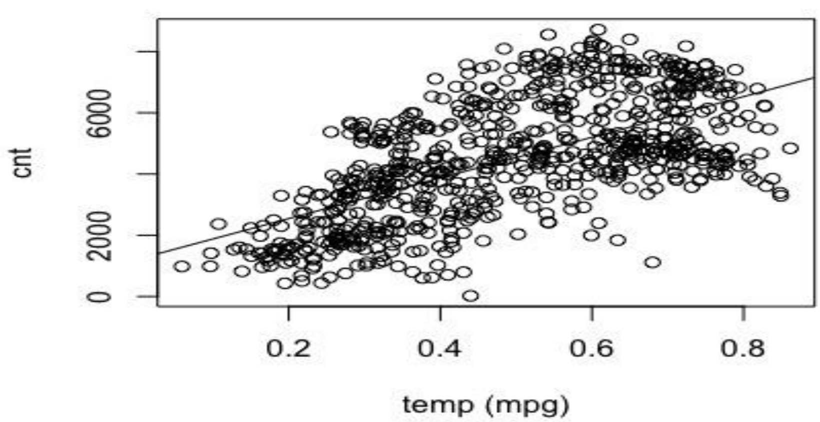
Techniques Used:

1] Simple Linear Regression:

Regression analysis is a statistical process which helps in finding relationship between a dependent(unknown) and independent variable (known). Forecasting or predictive analysis is one of many applications of regression analysis that is used in business applications.

It helps us to establish relationship between two variables and helps us get to know how they are interconnected and affect the overall outcome. of impact is determined by the relationship of the independent variables and the known variables.

```
> plot(jitter(cnt) ~ jitter(temp),
+ xlab="temp (mpg)",ylab="cnt",data=day)
> fit= lm (cnt~temp, data= day)
> abline(fit)
```



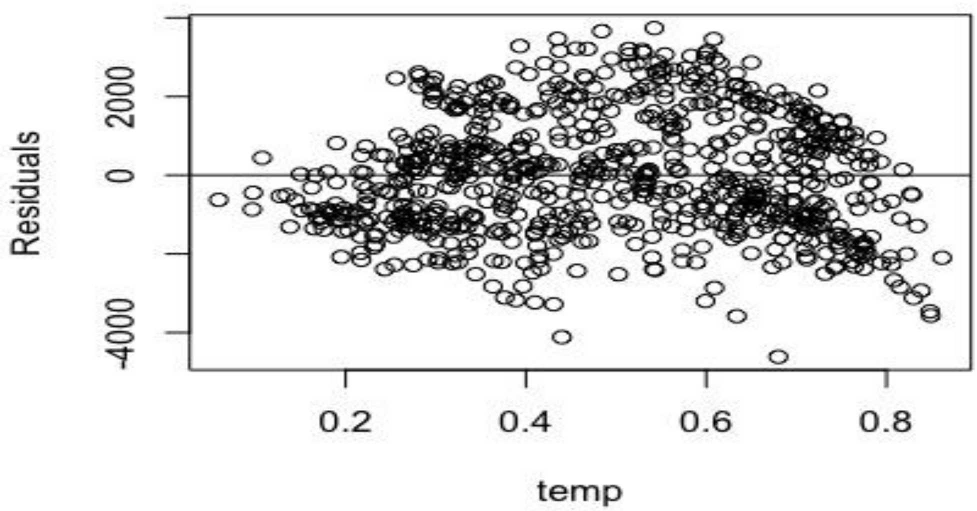
Linear regression considering temperature as a independent variable and registered users as dependent variable.

Aishwary Gawande
196009219

Residual Plot:

Residuals in regression analysis gives you more insight about the model and helps us to discover the parameters/data which otherwise would not have been found. It also helps in finding the credibility of the model and the fluctuations between the actual and the discovered value. And helps us to find different underlying patterns and analogies within the model.

```
> res <- residuals(fit)
> plot(jitter(res)~jitter(temp),
+ ylab="Residuals",xlab="temp",data=day)
> abline(0,0)
```



The residuals are uncorrelated and they follow a random pattern. They are randomly scattered over the scatter plot. If there are correlations between residuals then there is information left in the residuals which should be used in computing forecasts The residuals have zero mean.

Above scatter plot shows that reisduals are not randomly scatter and it is not a good model.

Aishwary Gawande
196009219

Goodness of Fit (R square):

```
> summary(fit)

Call:
lm(formula = cnt ~ temp, data = day)

Residuals:
    Min      1Q  Median      3Q     Max 
-4615.3 -1134.9  -104.4  1044.3  3737.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1214.6      161.2   7.537 1.43e-13 ***
temp          6640.7      305.2  21.759  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1509 on 729 degrees of freedom
Multiple R-squared:  0.3937,	Adjusted R-squared:  0.3929 
F-statistic: 473.5 on 1 and 729 DF,  p-value: < 2.2e-16
```

R square give the goodness of fit and how close the value fits the regression line. R square is very less which shows that model is not good for prediction.

Aishwary Gawande
196009219

2] Multiple Linear Regression:

Multiple Linear Regression analysis is a statistical process which helps in finding relationship between a dependent(unknown) and multiple independent variables (known).

```
> A0=lm(cnt~ temp+ weathersit + yr + mnth ,data=day)
> stepAIC(A0 , direction = "backward")
Start:  AIC=10059.82
cnt ~ temp + weathersit + yr + mnth

             Df Sum of Sq        RSS   AIC
<none>                     683307892 10060
- mnth        1  76385251  759693143 10135
- weathersit  1 122803132  806111024 10179
- temp        1 741326018 1424633910 10595
- yr          1 769024805 1452332697 10609

Call:
lm(formula = cnt ~ temp + weathersit + yr + mnth, data = day)

Coefficients:
(Intercept)         temp   weathersit           yr         mnth
    1085.11      5700.06      -760.93      2055.71        96.32
```

Demand is influenced by factors such as the Season, Month and Weather Conditions.

Model shows the prediction of the bike users according to the factors like temperature, season, workingday, humidity and month to predict the count of the number of the bike users.

Aishwary Gawande
196009219

```
> summary(mod2)

Call:
lm(formula = cnt ~ temp + weathersit + yr + mnth, data = day)

Residuals:
    Min      1Q  Median      3Q     Max 
-4307.3  -490.6    60.8   630.7  3024.5 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1085.11     156.73   6.924  9.7e-12 ***
temp         5700.06     203.10  28.065  < 2e-16 ***
weathersit   -760.93      66.62 -11.423  < 2e-16 ***
yr           2055.71      71.92  28.584  < 2e-16 ***
mnth           96.32      10.69   9.009  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 970.2 on 726 degrees of freedom
Multiple R-squared:  0.7506,	Adjusted R-squared:  0.7492 
F-statistic: 546.2 on 4 and 726 DF,  p-value: < 2.2e-16
```

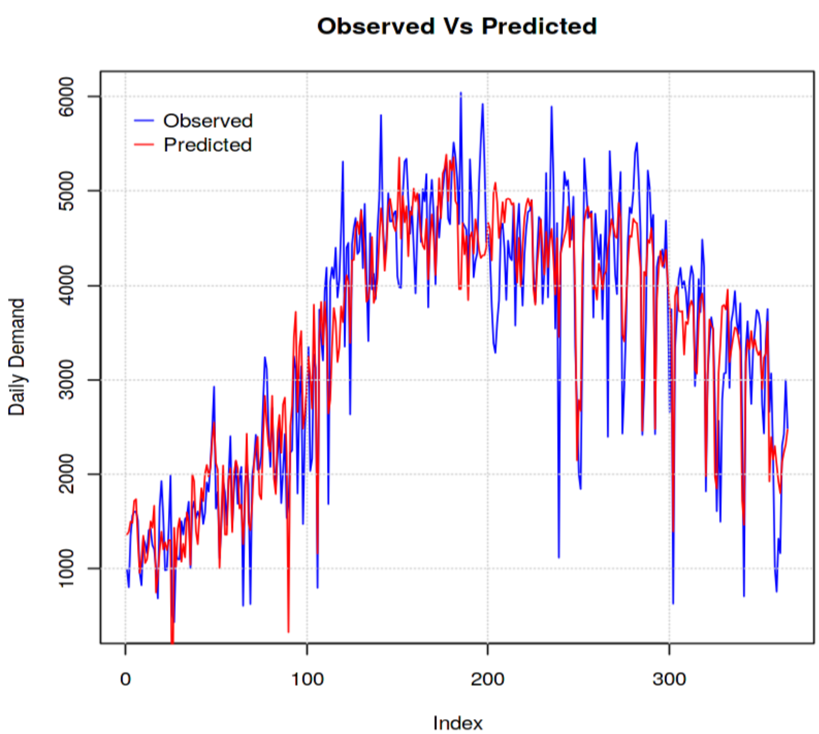
R square is 0.75 which shows that points are closely fit to the regression line than before.

Aishwary Gawande
196009219

Testing the Model

Values are predicted on the based of the previous vales and are fitted to check the fluctuations between the observed and the predicted value.

```
> newdata <- day %>% select_("mnth","weathersit","temp","yr")
> predictions <- predict(mod2,newdata = newdata)
> head(predictions)
       1        2        3        4        5        6
1621.341 1731.415 1539.789 1560.514 1714.171 1585.298
```



It can be noticed how the red line(predicted) tries to mimic the blue line(observed).

Aishwary Gawande
196009219

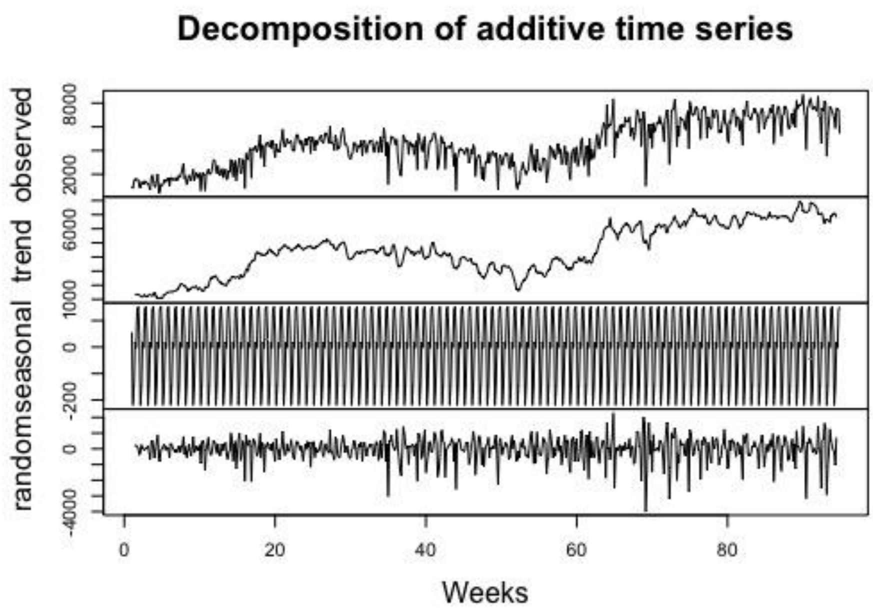
3] ARIMA Modelling

- ARIMA stands for Auto Regressive Integrated Moving Average.
- It is combination of both Auto regressive (AR) model and Moving Average (MA) model.
- In AR model output is predicted on the basis of past values and in MA model output is predicted on the basis of previous errors.
- ARIMA modelling is applied on time series data for analyzing and forecasting.

First step in ARIMA modelling is to check wether the the data is stationery or not. After decomposing the time series data we have to check wether it is stationery or not. This is done by KPSS Unit Root test.

```
> plot(decompose(train_ts, type='add'), xlab="Weeks")
>
> #Test for stationery
> adf_test <- adf.test(train_ts, alternative='stationary')
> library(urca)
> Test=ur.kpss(train_ts)
> summary(Test)

#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 6.2145

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Aishwary Gawande
196009219



We can see that in the first row it is time series data which is decomposed into trend, seasonality and random error.

We then applied KPSS test which gives us t-statistics 6.2145 which is very higher than 1. Test shows thst it's not stationery as test stat is higher than 1 which shows that data is not stationery and we have to apply differencing in order to make it stationery.

Seasonality is removed and differencing is applied to make the data stationery.

We call this an ARIMA(*p*,*d*,*q*) model, where

p = order of the autoregressive part.
d= degree of first differencing involved.
q= order of the moving average part.

Aishwary Gawande
196009219

```
> fit1 <- Arima(train_ts, order=c(7,1,0),seasonal=c(6,1,0),
+               method = "CSS", optim.method = "BFGS")
> fit1
Series: train_ts
ARIMA(7,1,0)(6,1,0)[7]

Coefficients:
          ar1      ar2      ar3      ar4      ar5      ar6     ar7     sar1
      -0.6320  -0.5904  -0.5045  -0.4744  -0.4031  -0.2578  0.0258  -1.1676
s.e.   0.0403   0.0479   0.0501   0.0526   0.0533   0.0575  0.0538   0.0784
         sar2     sar3     sar4     sar5     sar6
      -1.0528  -0.8372  -0.6047  -0.3541  -0.2011
s.e.   0.1103   0.1080   0.0910   0.0673   0.0411

sigma^2 estimated as 816273:  part log likelihood=-5365.28
> accuracy(forecast_ts, test_ts)
                       ME     RMSE       MAE         MPE      MAPE      MASE
Training set    -7.577805  888.944  615.5461   -6.527716  19.03185 0.7127392
Test set     -1113.720379 2017.768 1517.1759 -442.716130 448.68648 1.7567340
                      ACF1 Theil's U
Training set -0.0002364113        NA
Test set      0.6449960509  4.432097
```

The RMSE for your training and your test sets should be very similar if you have built a good model. If the RMSE for the test set is much higher than that of the training set, it is likely that it is badly over fit the data, i.e. we have created a model that tests well in sample, but has little predictive value when tested out of sample.

Aishwary Gawande
196009219

Evaluating the performance of the models:

Error metrics used to perform evaluation:

Mean absolute error (MAE) : It's the average of the absolute differences between prediction and actual observations.

Mean square error (MSE): Similar to the MAE but squares the difference before summing them all instead of using the absolute value.

Root means square error (RMSE): It's the square root of the average of squared differences between prediction and actual observation.

Mean Absolute Percentage Error (MAPE): is the percentage equivalent of MAE. The equation looks just like that of MAE, but with adjustments to convert everything into percentages.

R Square (Goodness of Fit) : It tells us the how closely the points fit the reression line. R square is between 0 and 1. More closely it is to 1 the more accurate is your model.

Aishwary Gawande
196009219

SIMPLE LINEAR REGRESSION:

- R square is 0.379 very less which shows that model is not good for prediction.

- Residual plot is not Random.

MULTIPLE LINEAR REGRESSION:

- R square is 0.75 which shows that points are closely fit to the regression line than before and it can be used for predicting.

- Good Model for Prediction.

ARIMA MODELLING:

-RMSE is 888.944 for Testing Data and 2017.768 for Training data, the difference is very High.

-For a good model, Training and Test data should have very similar RMSE values.

-This is not a good model for prediction.

Aishwary Gawande
196009219

Conclusion:

- Linear regression model is descent enough for prediction as well, although there are lot of outliers in prediction.
- The Multiple Regression Model with the features from corr-plot increased the accuracy of the prediction when considered other independent variables.
- It should be noted that the prediction is based on temperature, humidity and other weather conditions. Overall model accuracy will be dependent of accuracy of weather predictions.
- The RMSE for the test set is much higher than that of the training set, the model has badly over fit the data, i.e. the model tests well in sample, but has little predictive value when tested out of sample.
- For the time series analysis, only time series is not enough for prediction. Other features should also be considered.

Aishwary Gawande
196009219

References:

- https://learn-us-east-1-prod-fleet01-xythos.s3.us-east-1.amazonaws.com/5af494bf200ea/3829000?response-content-disposition=inline%3B%20filename%2A%3DUTF-8%27%27Week6.2%2520Multiple%2520Linear%2520Regression.pdf&response-content-type=application%2Fpdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20191202T225804Z&X-Amz-SignedHeaders=host&X-Amz-Expires=21600&X-Amz-Credential=AKIAIBGJ7RCS23L3LEJQ%2F20191202%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=683349d63869c84388ff1590b24dd869de99565e8ba75215e50b7483692d95af

- https://learn-us-east-1-prod-fleet01-xythos.s3.us-east-1.amazonaws.com/5af494bf200ea/3732122?response-content-disposition=inline%3B%20filename%2A%3DUTF-8%27%27BF_Week4_1_Simple%2520Linear%2520Regression.pdf&response-content-type=application%2Fpdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20191202T222635Z&X-Amz-SignedHeaders=host&X-Amz-Expires=21599&X-Amz-Credential=AKIAIBGJ7RCS23L3LEJQ%2F20191202%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=005731e0b5001ab5c3eca2a054a158d8a7cc6b6d3b6f9a3c3f64e9acb28dca8a

- https://stats.stackexchange.com/questions/56302/what-are-good-rmse-values

- https://www.japantimes.co.jp/life/2017/10/21/lifestyle/pedal-power-bike-sharing-services-expand-in-japan/#.XeLczi2ZNQI

- http://capitalbikeshare.com/system-data