DATA ANALYSIS OF BIKE SHARE DEMAND

Presented By:
Arundathi Sandikar
Pallavi Madhuranath
Sindhuja Chinnathambi

BIKE SHARE DEMAND

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

DATA ACQUIRING

Data acquired from - https://www.kaggle.com/c/bike-sharing-demand
We are analysing the usage patterns with weather data as a factor in order to forecast bike rental demand in the Capital BikeShare program in Washington, D.C.

HYPOTHESIS

1. Bike usage has more demand in the peak office hours. Morning from 7am -10 am and Evening 5pm - 7 pm.
2. Bike usage is high when the weather is clear. One raining day it has least usage.
3. Bike usage is low on a weekend compared to that of a weekday.

UNDERSTANDING THE DATA SETS

The dataset shows hourly rental data of 20 days per month spanning for two years (2011 and 2012).Dataset has 10886 observations spread across 12 variables . We split the data set into training and test data sets. Training consists of first 15 days and test data has 16 - 19 days from the original dataset. Training data has 8600 observations and test has 2264 observations across 12 variables.

Variable data types:

```
> str(bike_data)
'data.frame':   10886 obs. of  12 variables:
 $ datetime  : Factor w/ 10886 levels "1/1/11 0:00",..: 1 2 13 18 19 20 21 22 23 24 ...
 $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
 $ weather   : int  1 1 1 1 1 2 1 1 1 1 ...
 $ temp      : num  9.84 9.02 9.02 9.84 9.84 ...
 $ atemp     : num  14.4 13.6 13.6 14.4 14.4 ...
 $ humidity  : int  81 80 80 75 75 75 80 86 75 76 ...
 $ windspeed : num  0 0 0 0 0 ...
 $ casual    : int  3 8 5 3 0 0 2 1 1 8 ...
 $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
 $ count     : int  16 40 32 13 1 1 2 3 8 14 ...
>
```

Independent variable : datetime, season, holiday, workingday, weather, temp, atemp, humidity, windspeed, day, hour.

Dependent variable : count, registered,casual.

DATA CLEANING

```
> summary(is.na(bike_data))
  datetime        season         holiday        workingday       weather          temp           atemp
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:10886     FALSE:10886     FALSE:10886     FALSE:10886     FALSE:10886     FALSE:10886     FALSE:10886
 NA's :0         NA's :0         NA's :0         NA's :0         NA's :0         NA's :0         NA's :0
  humidity       windspeed        casual         registered       count
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:10886     FALSE:10886     FALSE:10886     FALSE:10886     FALSE:10886
 NA's :0         NA's :0         NA's :0         NA's :0         NA's :0
```

1.Missing Values

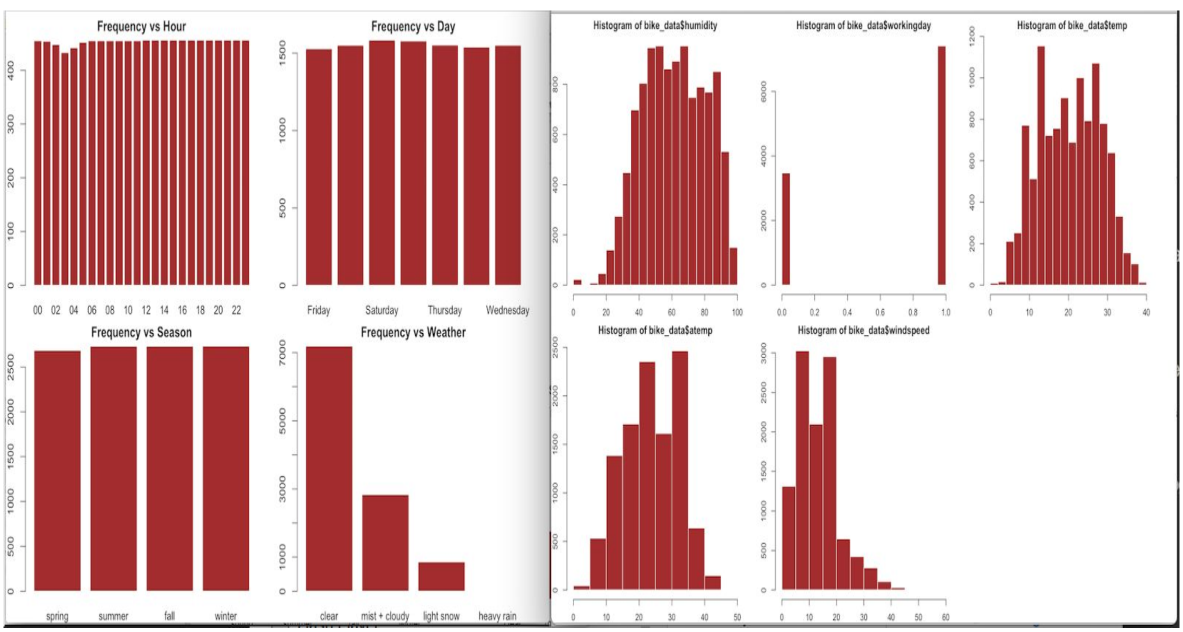
We ran a summary on the data set and found no missing value.

```
> summary(bike_data)
         datetime          season          holiday           workingday        weather           temp
 1/1/11 0:00 :     1   Min.   :1.000   Min.   :0.00000   Min.   :0.0000   Min.   :1.000   Min.   : 0.82
 1/1/11 1:00 :     1   1st Qu.:2.000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:13.94
 1/1/11 10:00:     1   Median :3.000   Median :0.00000   Median :1.0000   Median :1.000   Median :20.50
 1/1/11 11:00:     1   Mean   :2.507   Mean   :0.02857   Mean   :0.6809   Mean   :1.418   Mean   :20.23
 1/1/11 12:00:     1   3rd Qu.:4.000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:26.24
 1/1/11 13:00:     1   Max.   :4.000   Max.   :1.00000   Max.   :1.0000   Max.   :4.000   Max.   :41.00
 (Other)     :10880
     atemp          humidity         windspeed          casual          registered        count
 Min.   : 0.76   Min.   :  0.00   Min.   : 0.000   Min.   :  0.00   Min.   :  0.0   Min.   :  1.0
 1st Qu.:16.66   1st Qu.: 47.00   1st Qu.: 7.002   1st Qu.:  4.00   1st Qu.: 36.0   1st Qu.: 42.0
 Median :24.24   Median : 62.00   Median :12.998   Median : 17.00   Median :118.0   Median :145.0
 Mean   :23.66   Mean   : 61.89   Mean   :12.799   Mean   : 36.02   Mean   :155.6   Mean   :191.6
 3rd Qu.:31.06   3rd Qu.: 77.00   3rd Qu.:16.998   3rd Qu.: 49.00   3rd Qu.:222.0   3rd Qu.:284.0
 Max.   :45.45   Max.   :100.00   Max.   :56.997   Max.   :367.00   Max.   :886.0   Max.   :977.0
```

2. Check for outliers in independent variables

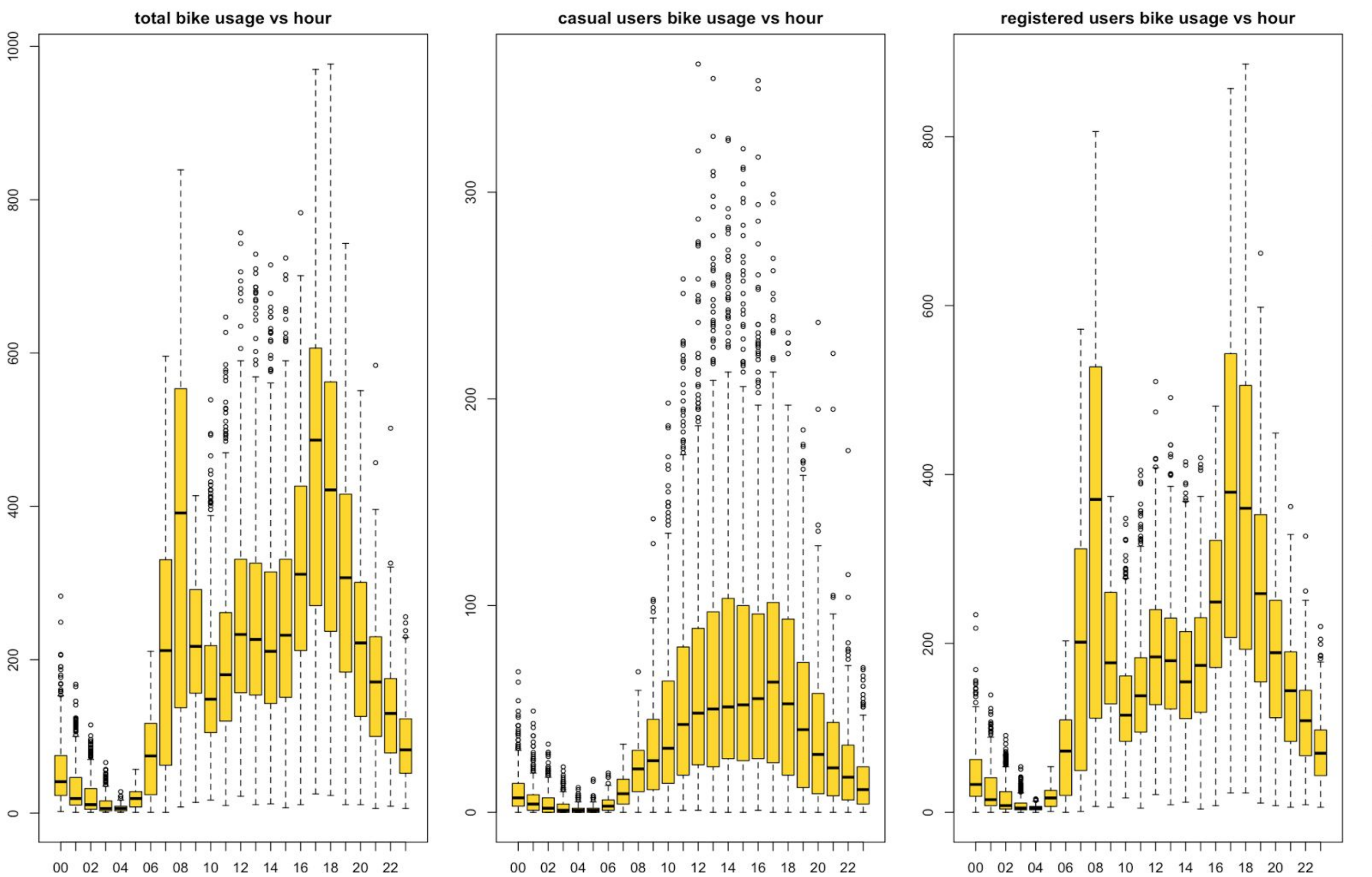
In the summary, we find mean and median of every independent variables are almost same. There is no huge difference between mean and median values per independent variables. Thus, we can say variables are normally distributed.

This can be further proved by plotting a histogram of all variables.

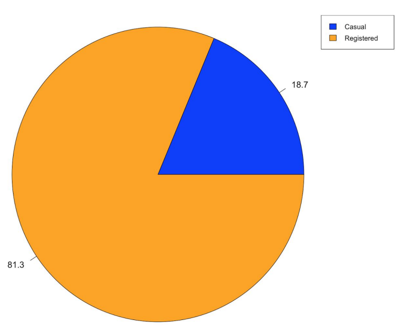


ANALYSIS FROM CLEANED TRAINING DATA

Plotting a boxplot to see the hourly trend of total bike users, casual and registered users over hours

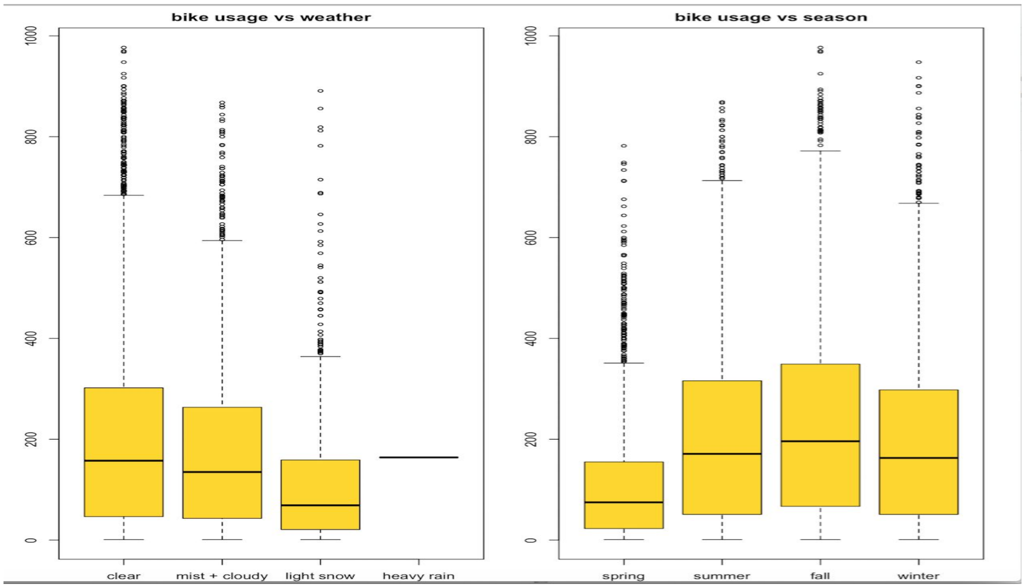


From the graph of total bike usage vs hour, we can say that bike usage is high 7am - 9am and 4pm - 7pm. Which is almost same as our hypothesis.We also obtained a similar pattern in registered usage vs hour. So, we can conclude that registered data is more significant compared to casual user's data.



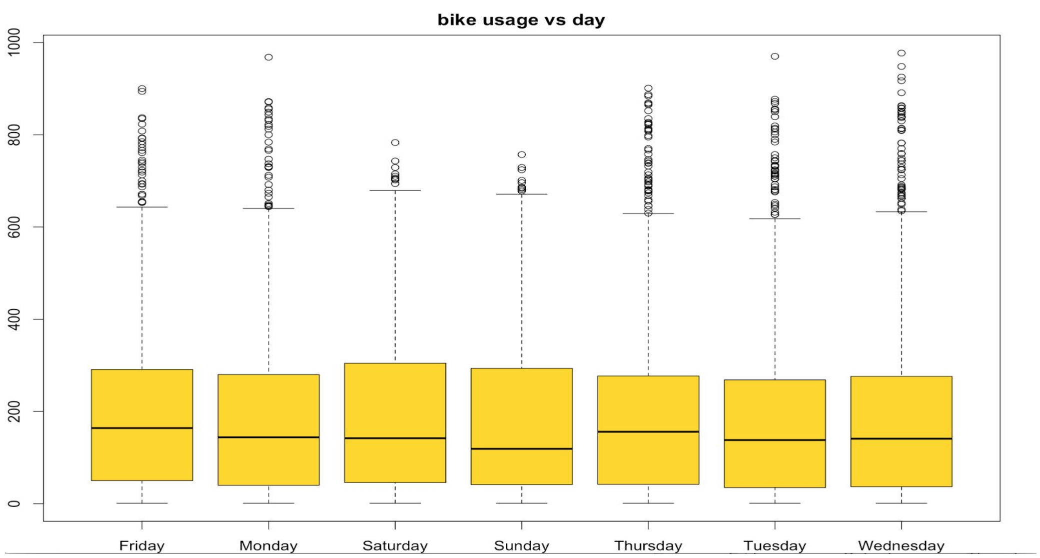
Registered bike users account for 81.3% of total usage, And rest 18.7% are casual users.

Plotting Boxplot to see usage trend over weather, season and day



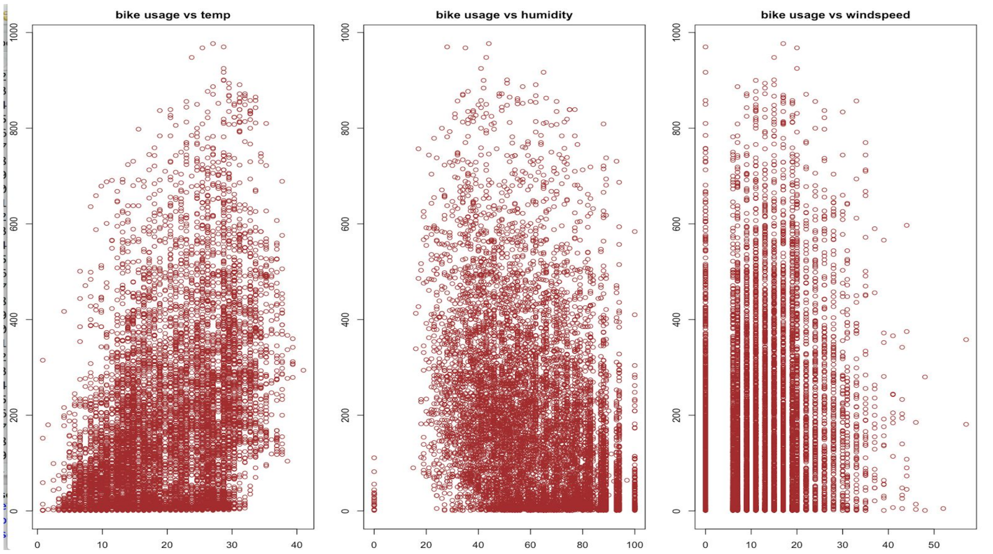
From analysing the graph bike usage vs weather we can say that usage is high when the weather is clear and when there is a heavy rain, usage is zero.

There is relatively high bike usage in fall and less in spring.



When analysing the boxplot for bike usage and day, we can see that on sunday usage is less compared to weekdays.

Scatter plot to see relationship between bike usage vs temperature, humidity and windspeed



Analysis from the graph,

1. **bike usage vs temp -** there is some positive relationship between the variables, but data is mostly scattered.
2. **bike usage vs humidity -** we see no relationship between the variables.
3. **bike usage vs windspeed** - there is negative relationship between the variables.

Below is the table that holds the correlation values between count (dependent variable) and other independent variables.

	bike_data_training_cor.count
bike_data_training_cor.temp	0.3977556
bike_data_training_cor.humidi	-0.3209817
bike_data_training_cor.atemp	0.3976934
bike_data_training_cor.windsp	0.1165921
bike_data_training_cor.season	0.1784616
bike_data_training_cor.weathe	-0.125428831
bike_data_training_cor.hour	0.400160384
bike_data_training_cor.workin	0.009301636

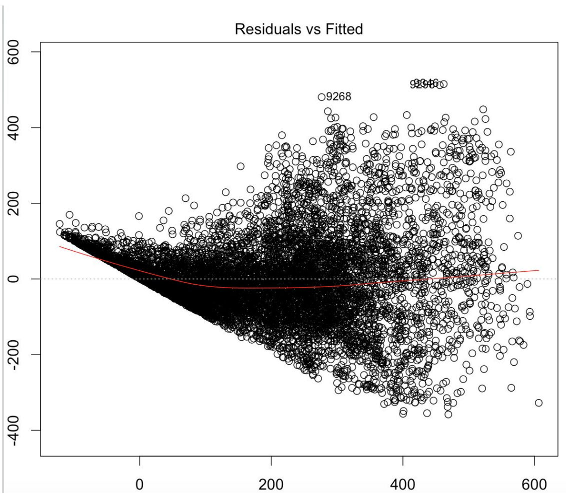
From the table we draw an inference that hour and temp are positively correlated to count. And humidity is negatively correlated.

LINEAR MODEL

Built a linear model from the training data with count as a dependent variable and hour, day , temperature,humidity and season as independent variables.

Model 1 :
bike_data_training_lm <- lm(count ~ temp+humidity+hour+day+season, data=bike_data_training)

```
Residual standard error: 110.7 on 8565 degrees of freedom
Multiple R-squared:  0.6261,    Adjusted R-squared:  0.6246
F-statistic: 421.9 on 34 and 8565 DF,  p-value: < 2.2e-16
```

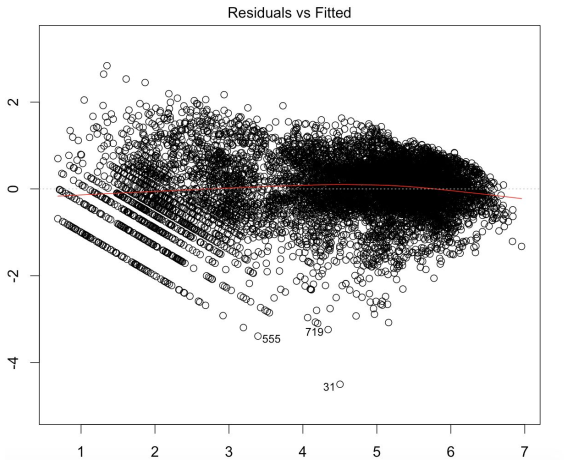


From the plot we found that the variance is not constant and the heteroscedasticity exists. So we used log of dependent variable to redesign the model.

Model 2 :
Bike_data_training_lm <- lm(log(count) ~ temp+humidity+hour+day+season, data= bike_data_training)

Summary of linear model:

```
Residual standard error: 0.6851 on 8565 degrees of freedom
Multiple R-squared:  0.7912,    Adjusted R-squared:  0.7904
F-statistic: 954.7 on 34 and 8565 DF,  p-value: < 2.2e-16
```



This model explains 79.04% of the variance of count and rest 21% is unexplained.
P-value is less zero so there is a significant relationship between independent variables and dependent variables.
From the plot we can see the pattern in the residuals vs fitted and also we got the higher R squared value.

Using the linear model 2, we predict bike usage for the test data.

Predict_count <- predict(Bike_data_training_lm, newdata = bike_data_test)

RMSE CALCULATION

exp(rmse(log((bike_data_test$count),bike_data_test$predictedcountlog))

RMSE Value : 2.06288

Sample of *bike_data_test_predicted.csv tables :*

DateTime	Actual_count	Predicted_count
5/17/2011 23:00:00	56	55
5/18/2011 0:00:00	23	55
5/18/2011 1:00:00	12	20
5/18/2011 2:00:00	6	7
5/18/2011 3:00:00	9	7
5/18/2011 4:00:00	3	7
5/18/2011 5:00:00	9	20
5/18/2011 6:00:00	101	55
5/18/2011 7:00:00	274	148
5/18/2011 8:00:00	453	403
5/18/2011 9:00:00	202	148
5/18/2011 10:00:00	106	148

The predicted count values in comparison with the actual count values is in the csv file *bike_data_test_predicted.csv.*

CONCLUSION

From the analysis we can confirm our proposed hypothesis that Bike usage depends on the time of the day, weather conditions and weekday or weekend. In our linear model 1 we found residual plots scattered and also the model was not Homoscedastic. Hence in our model two we applied a logarithmic transform on the dependent variable and tried to derive a Homoscedastic model. We used model 2 to predict bike usage in the test data.

REFERENCES

https://www.statisticssolutions.com/homoscedasticity/
http://www.statmethods.net/stats/regression.html
https://www.kaggle.com/c/bike-sharing-demand
http://www.statsmakemecry.com/smmctheblog/confusing-stats-terms-explained-heteroscedasticity-heteroske.html