ALY 6050: Data Mining Applications

Week 1 Assignment: Finding Patterns in Data and EDA

Student Name: Snehal Dahiphale

Professor Name: Prof. Marcus Ellis

Spring 2018

Date: 2nd June 2018

Bike Sharing Systems have revolutionized the traditional bike rental by automating the process of membership, rental and return. Bike sharing systems allow individuals to borrow bikes for a short period of time from one position and return it to another. Bike sharing systems have advantages like flexible mobility, emission reductions, health benefits, reduced traffic congestion and fuel use. Today, there are over 500 bike-sharing programs with over 500 thousand bikes available on a sharing basis for a nominal fee with information about nearby pick up and drop off spots in the app. The data obtained from bike sharing systems such as trip time, departure time, arrival time, pick up point, drop off point opposed to the bus and train transport help us sense and analyze the mobility throughout the city. It will also help us make conclusions about where are the hotspots in the city that need better public transportThis dataset consists of daily count of rental bikes from 2011 to 2012 in Washington DC with the following fields described below:

S no.	Field Name	Field Description
1	Instant	Record index
2	Dte	Date (Year/Month/Day format)
3	Season	Seasons: 1 to 4 (1-Spring, 2-Summer, 3-Fall, 4-Winter)
4	Yr	Year 0 or 1 (0-2011, 1-2012)
5	Month	Month: 1 to 12 (1-jan, 2-Feb, 3-March,..,12-December)
6	Hr	Hour: 0 to 23
7	Holiday	Whether the day is a holiday or not
8	Weekday	Day of the week: 0 to 6
9	Workingday	Working days= 1 and Holidays = 0
10	Weathersit	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

11	Temp	Temp
12	Atemp	Feels-Like Temp
13	Hum	Humidity
14	Windspeed	Windspeed
15	Casual	Count of Casual Users
16	registered	Count of Registered Users
17	cnt	Count of total rental bikes (casual + registered)

We load the dataset using read.csv() and perform ncol, nrow and summary to get the total rows, total columns and summary of the dataset with information of minimum, maximum, mean, 1st quartile and 3rd quartile value of each field. Applying matrix function on the data gives us an idea about the correlation between all fields to help us understand the highly correlated fields before beginning our analysis. By looking at the summary we can see that, temp, atemp, hum, windspeed have similar mean values and may be grouped in one cluster as temperature fields and similarly, cnt and registered fields can be clubbed together as they represent the total no. of bikes. The matrix graph verifies that the field pairs such as registered, cnt and temp, atemp have a good correlation between them respectively and yr, mnth, holiday, weekday do not.

```
> nrow(bikedata)
[1] 731
> #Total number of columns in day.csv using ncol()
> ncol(bikedata)
[1] 16
> #Dimension of day.csv using dim()
> dim(bikedata)
[1] 731  16
> #Summary of bikedata using s
> summary(bikedata)
    instant          dteday        season          yr              mnth
 Min.   :  1.0   2011-01-01:  1   Min.   :1.000   Min.   :0.0000   Min.   : 1.00
 1st Qu.:183.5   2011-01-02:  1   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.: 4.00
 Median :366.0   2011-01-03:  1   Median :3.000   Median :1.0000   Median : 7.00
 Mean   :366.0   2011-01-04:  1   Mean   :2.497   Mean   :0.5007   Mean   : 6.52
 3rd Qu.:548.5   2011-01-05:  1   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:10.00
 Max.   :731.0   2011-01-06:  1   Max.   :4.000   Max.   :1.0000   Max.   :12.00
                 (Other)   :725
    holiday           weekday        workingday      weathersit         temp
 Min.   :0.00000   Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :0.05913
 1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.33708
 Median :0.00000   Median :3.000   Median :1.000   Median :1.000   Median :0.49833
 Mean   :0.02873   Mean   :2.997   Mean   :0.684   Mean   :1.395   Mean   :0.49538
 3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:0.65542
 Max.   :1.00000   Max.   :6.000   Max.   :1.000   Max.   :3.000   Max.   :0.86167

     atemp              hum           windspeed          casual         registered
 Min.   :0.07907   Min.   :0.0000   Min.   :0.02239   Min.   :   2.0   Min.   :  20
 1st Qu.:0.33784   1st Qu.:0.5200   1st Qu.:0.13495   1st Qu.: 315.5   1st Qu.:2497
 Median :0.48673   Median :0.6267   Median :0.18097   Median : 713.0   Median :3662
 Mean   :0.47435   Mean   :0.6279   Mean   :0.19049   Mean   : 848.2   Mean   :3656
 3rd Qu.:0.60860   3rd Qu.:0.7302   3rd Qu.:0.23321   3rd Qu.:1096.0   3rd Qu.:4776
 Max.   :0.84090   Max.   :0.9725   Max.   :0.50746   Max.   :3410.0   Max.   :6946

      cnt
 Min.   :  22
 1st Qu.:3152
 Median :4548
 Mean   :4504
 3rd Qu.:5956
```

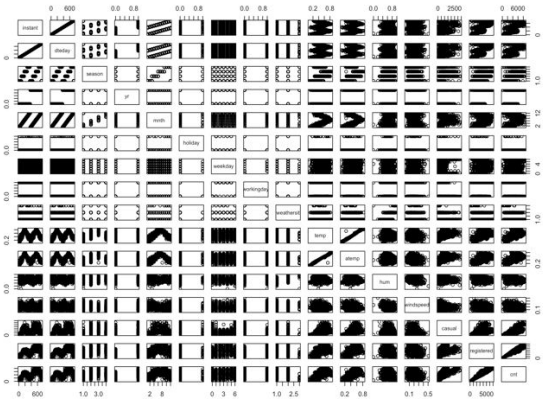

Fig: Summary of bike sharing data

Fig: Matrix of correlations

Next, we obtain the table of distribution of number of rental bikes for every season, workday, holiday and year by applying the factor and table function on our dataset by plugging in as fall, winter, spring, summer for season for season levels 0,1,2,3 and years as 2011,2012 for level 0,1.

```
> table(bikedata$season)

Spring Summer   Fall Winter
   181    184    188    178
```

```
> table(bikedata$yr)

2011 2012
 365  366
```

```
> table(bikedata$holiday)

Working Day     Holiday
        710          21
```

```
> table(bikedata$weathersit)

     Good:Clear/Sunny      Moderate:Cloudy/Mist          Bad: Rain/Snow/Fog
                  463                       247                        21
Worse: Heavy Rain/Snow/Fog
                    0
```

Fig: Bike rental numbers during different seasons, years, days and weather

The no. of rental bikes during fall season is the most being 188 and least during winter season being 178. The total rental bikes used in 2011 and 2012 is almost the same with values 365 and 366 respectively. Bikes were used most on working days and least on holidays. People preferred to use bikes most on good: clear/sunny day 463 timesand not even once during worse: heavy rain/ snow days but however, some people used bikes on bad: rain/snow/fog days 21 times.

After this, we create new attributes to denormalise actual values because, the normalised values were low and factorised the categorical attributes.

```
bikedata$actual_temp <- bikedata$temp*41
bikedata$actual_feel_temp <- bikedata$atemp*50
bikedata$actual_windspeed <- bikedata$windspeed*67
bikedata$actual_humidity <- bikedata$hum*100
bikedata$mean_acttemp_feeltemp <- (bikedata$actual_temp+bikedata$actual_feel_temp)/2
```

Fig: create normalised attributes for actual, feel, mean temperatures and humidity

Visualizations

Boxplot (box and whisker diagram) can be defined as a standardized way of showing data distributions on the basis of: minimum, maximum, median, first quartile and thir quartile values.

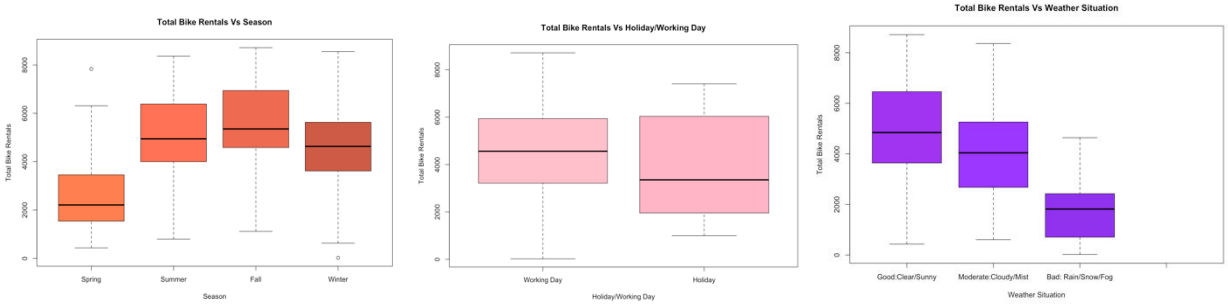

Fig: Bike rentals vs Season, Working day/Holiday, Weather Situation

Average bike rentals are highest during fall and summer season and higher on working days compared to holidays. A decreasing trend can be observed for rental bikes as the weather goes from good, moderate, bad to zero at worse.

Histograms are similar to bar charts but, they represent numerical data by plotting against frequency. In R, we can use the hist() function to make histograms and here we made it for total and registered bike rental count against their frequency of use.

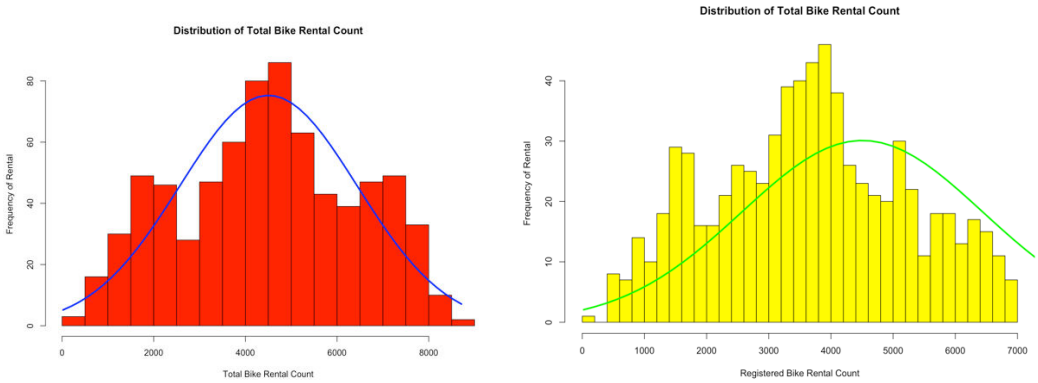

Fig: Histograms for total bike rental count and registered bike rental count

We obtain the mean for graph 1 at 4500 and for graph 2 at 4800. Both distributions are normal. The mean and variance are almost at the same levels for graph 1 however for graph 2, the variance is more comparatively because only registered users were taken into account and casual users were not counted. In the graph below, over the span of two years use of bike rentals has increased significantly and we can see the variation during different seasons represented through the ups and downs in the bike rental data.

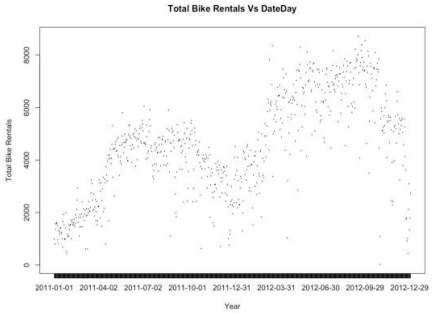

Fig: Total bike rental vs year

Scatterplots are two variables plotted with correlation points scattered and in r, we use lm() function to plot the line of fit for analyzing the effect of temperature, windspeed and humidity on total bike rentals for different seasons fall, spring, summer and winter marked in 4 colors.

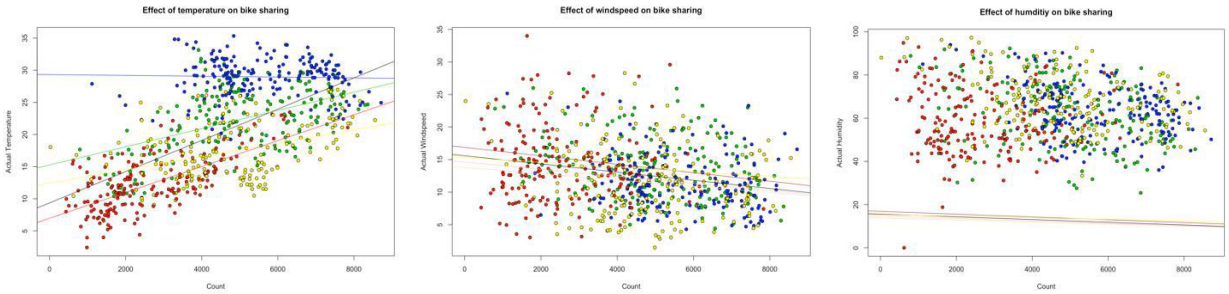

Fig: Effect of temperature, windspeed and humidity on total no. of bike rentals

In the first graph, the black line shows the overall good correlation between bike rental count and increasing temperatures. The blue line representing fall season remains almost stable however, for summer, spring and winter seasons as the temperatures increase, the total no. of bike rentals kept increasing. In graph 2, we study the effect of windspeed on bike rentals and observe that correlations for all seasons are negative because as the windspeed kept increasing, the total no. of bike rentals kept going down. In graph 3, we analyze the effect of humidity on the total no. of bike rentals for different seasons but we notice that, humidity more or less has barely any effect. But, we can see a slight negative correlation which also indicates that with growing humidity the bike rentals decrease a little bit.

Regression

Regression is a commonly used modelling technique used to predict the relationship between dependent and independent variables. We test a linear model using the lm() function to study the effect or dependence of actual temperature i.e actual_temp variable on total no. of bike rentals i.e cnt variable. The summary of test below shows us that the minimum and maximum value of residuals is -4615.3 and 3737.8 respectively. The p-value of the test is smaller than the

significance value and the value of f-statistic is high therefore, the test can be considered to be good which means that the actual_temp highly affects the total cnt.

```
Call:
lm(formula = bikedata$cnt ~ bikedata$actual_temp)

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)         1214.642    161.164   7.537 1.43e-13 ***
bikedata$actual_temp  161.969      7.444  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
```

Fig: Summary of lm_test model

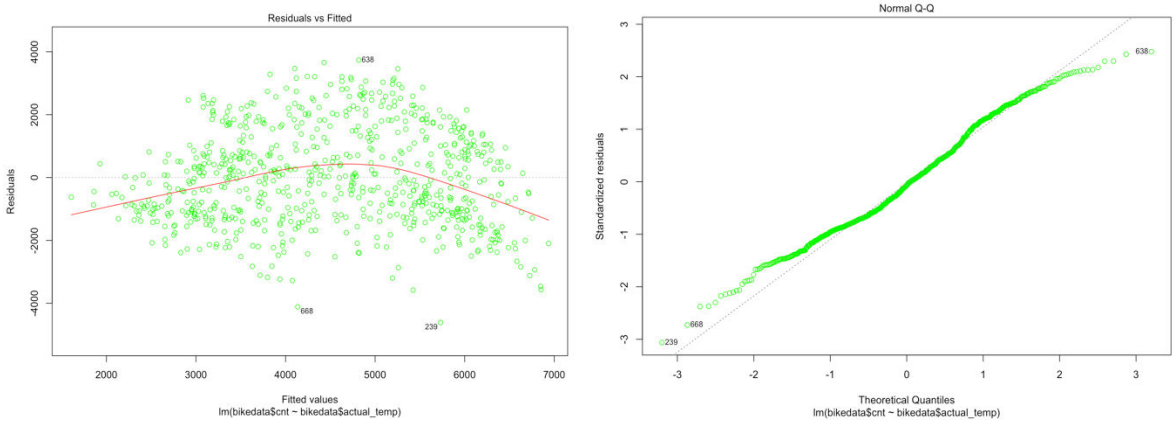

Fig: Residuals vs Fitted and Normal Q-Q plot

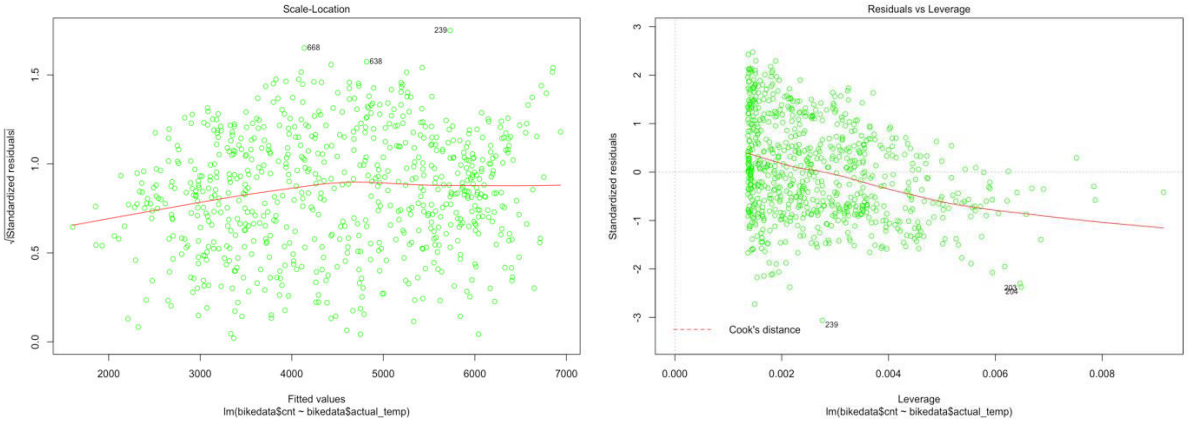

Fig: Standardized Residuals and Residuals vs Leverage

The graphs of the linear model for residuals show us that the residuals are equally distributed below and above the mean red line and the q-q plot shows the line which is almost linearly formed between theoretical and standardized residual values. Both these factors are an indication of test being accurate but however, in the last graph there are more points above the red line than below which also means that there may be a test which could outperform this one.

References:

1. https://www.kaggle.com/marklvl/bike-sharing-dataset (Data Source)

2. https://www.r-bloggers.com/simple-linear-regression-2/

3. Eckerson, W. W. (2012). *Secrets of analytical leaders*: Westfield, NJ: Technics Publication