Project 02-Analyzing the NYC Subway Dataset

Section0: References

http://www.slideshare.net/mhsgeography/mann-whitney-u-test-2880296

http://www.statstutor.ac.uk/resources/uploaded/mannwhitney.pdf

http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

Section1: Statistical Data

1.1 Since the data used were not normally distributed we used the Mann-Whitney U test. The null hypothesis for this test was that the two data i.e. entries during rainy and non-rainy days were from the same population or that they are not significantly different. As such, a two sided test was used. A p-critical value of 0.05 was selected.

1.2 As mentioned in 1.1, this test is applicable because the data were not normally distributed. The test assumes no information on how the data is distributed, making it more relevant.

1.3 The following results were obtained (with_rain_mean, without_rain_mean, U value, P-value):

```
1105.4463767458733, 1090.278780151855, 1924409167.0, 0.0499*

* The P-value obtained from the test is 0.024999912793489721. But a two tailed test we multiplied that by 2.
```

1.4 The p-value obtained from the test is smaller than the p-critical indicating that the data compared are significantly different or that they are not from the same population. In our case, the result shows that rain had a significant impact on the number of riders.

Section2: Linear Regression

2.1 OLS is used for prediction.

2.2 The features used for regression were: ***'Hour','fog','rain','meantempi','precipi','meanwindspdi'***. ***UNIT*** is used as a dummy variable.

2.3 These features were selected for the following reasons:

Hour- It is expected that the number of riders is going to vary hourly. You would expect more riders during the morning rush hour, lunch break, and evening. This information is also visualized using a scatter plot of the data. As such both intuitively and from the plot, I expected a certain degree of correlation between hour and ridership.

Fog- During fogy days, from my experience I wouldn't be walking or even driving. So, this was an Intuitive selection in that more riders are expected during fogy days.

Rain- for similar reasons as fog, I expected more riders during rainy days.

Meantempi- if the temperature is cold for instance during winter or so, I expect more people to ditch walking for public transport/personal cars.

Percipi- This was an intuitive selction.

Meanwindspdi- During a very windy day, I would think less people would be walking in the streets of NYC. In addition, this feature was introduced as a result of experiment and it improved the R^2 value.

2.4 The weights for the features are:

```
Hour              65.364525
fog              214.086883
rain               0.040560
meantempi         -9.491420
precipi          -73.976935
meanwindspdi      32.568316


P.S after several iterations, it become clear that the features Fog, rain and Precipi
are highly correlated causing the coefficients and so the model to be very unstable.
This is explained by multicollinarity.
```

2.5

```
Your r^2 value is 0.480456675828
```

2.6 The resulted R^2 does meet the minimum required R^2 for the exercise. R^2 indicates how well the linear model obtained relates to the response variable. In our case R^2 is 0.48. This is indicating that our linear model developed can only explain or account to 48% of the variability in the data. The rest 52% if the variability cannot be explained or accounted by the model making it less accurate or dependable.

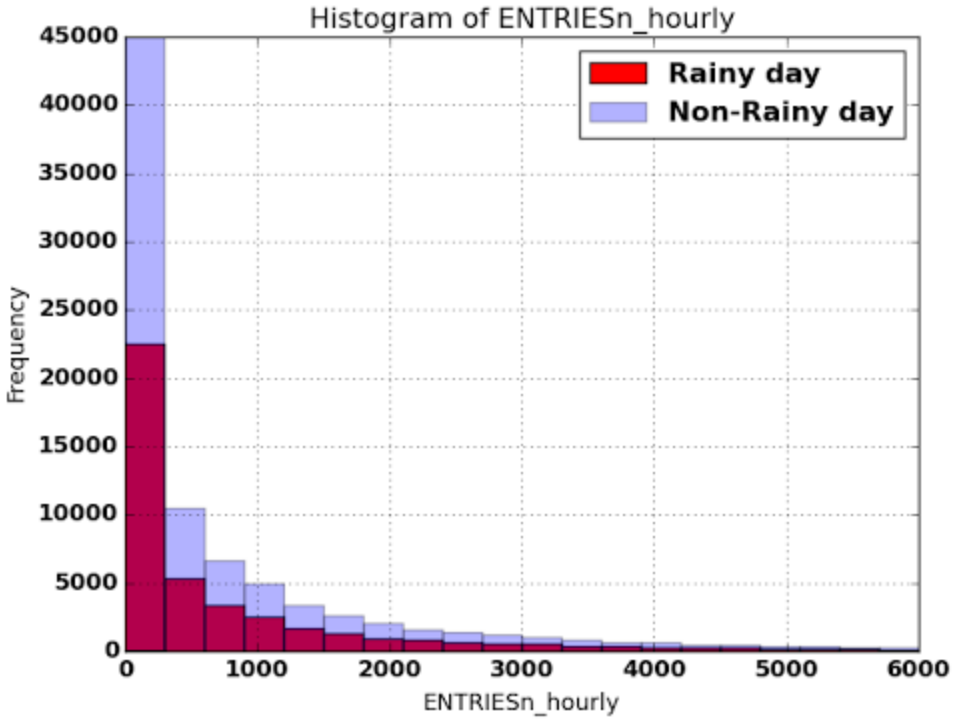
The correlation between our feature and the value (ENTRIESn_hourly) can be explained by the square root of R^2, which is going to be ~0.07 or 7%. This is a very low correlation.

Section 3- Visualization

3.1

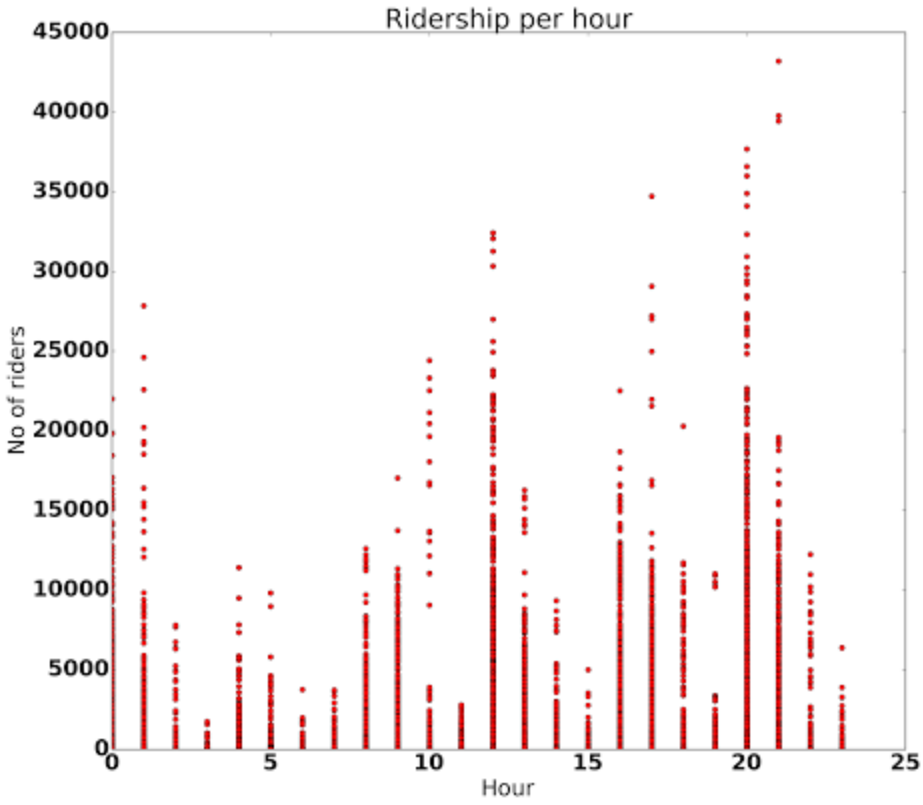
Here is the histogram plot of hourly riders during rainy and non-rainy days. The histogram shows that the data in both rainy and non-rainy days, it is positively skewed with outliers. Because of the non-uniform distribution of the data, we should be careful when choosing statistical tests. It is also apparent we have fewer data for rainy days than non-rain days.

For clarity purposes, the histogram plot is truncated at 6000 for x-axis as most of the values are within that range.



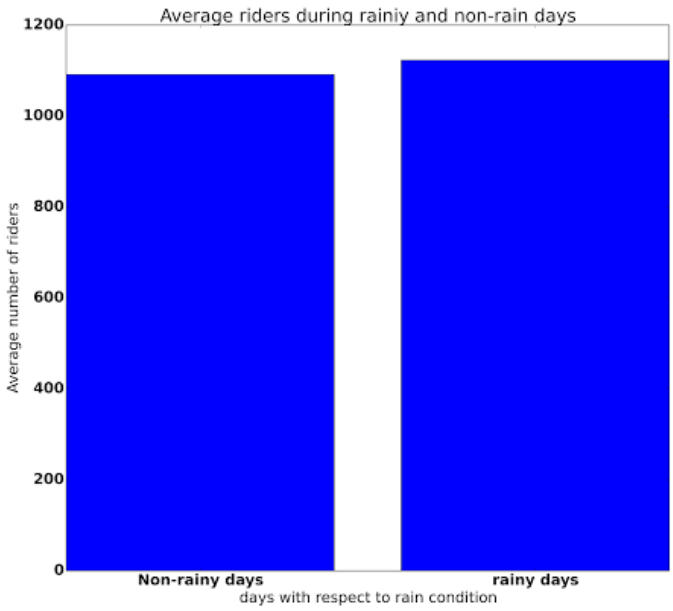
3.2

Here we have a point plot of the number of riders on each hour. As can be inferred from the graph, we have more riders around 1:00 am, noon and 8:00 pm. We also have large number of riders during the morning hours around 9:00 am and 5:00pm. This could be related to folks going and coming from work respectively.



Section 4- Conclusion

4.1 The conclusion is more people ride the NYC subway during rainy days than non-rainy days. Below is a histogram showing the average number of riders during rain and non-rainy days.



4.2

From the statistical analysis it is clear that the number of riders during the rainy day is significantly different from non-rainy days. We also see from the linear regression that the coefficient or weight for rain is positive 0.04. This is indicating there is a positive relation or correlation between rain and number of hourly riders.

Section 5- Reflection

5.1 With regard to the data set, we have significantly much fewer data sets for rainy days. It would have been more robust to have as many dataset for rainy days as non-rainy days. We also had some significant outliers in the data set that could affect our analysis.

We used linear regression technique with R^2 <50%. The linear regression model has a mean error of about 830(from the residual equation). This is significant when compared to the fact that the average number of riders is about 1100. This is also indicated by the low R^2.

A discussion on multicollinearity is given in section 2.4.