Analyzing the NYC Subway Dataset

Section 0 References

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.mannwhitneyu.html

https://storage.googleapis.com/supplemental_media/udacityu/4332539257/MannWhitneyUTest.pdf?GoogleAccessId=1069728276824-fdhtlb98k1m9qrmdgj4jgc7gjp2l1lsm@developer.gserviceaccount.com&Expires=1433275960&Signature=fPRzKzw1qMBzQEOyFbv4MBlyHIkVimzzY9%2B9JrIXDmfBsRjoIk034Lta/chJQyji01J4Bm8MUuhRP9p5oWPNi/x8OLy%2Btnhecw0cMNQU/Y2Z7QAjSxONeOClkuh0ZQbERpdd0G/2k7FMiQsXHuW4lKL1I2LZnfT8MI9h%2B/ARe4A%3D

http://www.statsdirect.com/help/default.htm#nonparametric_methods/mann_whitney.htm

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

http://people.duke.edu/~rnau/rsquared.htm#punchline

http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/

http://stackoverflow.com/questions/20181456/sum-up-column-values-in-pandas-dataframe

http://docs.ggplot2.org/0.9.3.1/geom_bar.html

http://stackoverflow.com/questions/30406564/python-ggplot-how-do-i-layer-histograms

Section 1 Statistical Test

1.1 Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

For the NYC Subway data, we used the Mann Whitney U test which is also known as Wilcoxon rank sum test. This is a non-parametric test that can be used to test two populations with unknown distributions. For our example, we used the one-tailed P value, with a p-critical value of 0.05. The null hypothesis for this tests whether or not two populations are the same.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

The Mann-Whitney U test is applicable for the NYC subway dataset because it is not normally distributed. In our problem sets, we plotted a histogram of the two distributions and from the graph it was evident that the dataset is not normally distributed. With a non-parametric test, we do not assume that the data is drawn from any particular probability distribution.

1.3 What results did you get from this statistical test? This should include the following values: p-values, as well as the means for each of the two samples under test.

From the analysis of turnstile weather data comparing entries between ridership on rainy and non-rainy days. I got the following results:

With rain mean: 1105.4463767458733

Without rain mean: 1090.278780151855

U value: 1924409167

p value: 0.024999912793489721

1.4 What is the significance and interpretation of these results?

Given that the p value is the probability of obtaining a test statistic at least as extreme as ours if the null hypothesis (whether or not they are from the same population) is true is small and the fact that the p-value is smaller than our p-critical value, it means that we reject the null hypothesis and therefore we can say that the distribution of the entries in our turnstile data is statistically different between rainy and non-rainy days. The large U value, observed as the number of times one sample can be observed as being greater than the second in terms of ranking further proves that the two distributions are statistically different, therefore we can infer that there is a difference between ridership on the NYC subway during rainy and non-rainy days.

Section 2 Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model?

I did not use OLS or gradient descent using the Scikit Learn, instead implemented gradient descent using numpy and pandas.

```
def compute_cost(features, values, theta):

   m = len(values)

   sum_of_square_errors = numpy.square(np.dot(features, theta) - values).sum()

   cost = sum_of_square_errors / (2*m)

   return cost

def gradient_descent(features, values, theta, alpha, num_iterations):

   m = len(values)

   cost_history = []

   for i in range(num_iterations):

      difference = (values - np.dot(features,theta))

      fTrans = features.transpose()
```

```
        theta = theta + (alpha/m)*(np.dot(fTrans,difference))

    return theta, pandas.Series(cost_history)
```

Ordinary Least Squares using Statsmodel was used later on in problem set 3

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

The features I used in the model are:

- Rain (Yes/No)
- Fog (Yes/No)
- Precipitation in inches
- Hour of travel
- Daily average temperature in Fahrenheit
- Unit

I did have to use dummy variables as part of the features as Unit is not a value that can be converted into a numerical value. Hence dummy variables needs to be created (one for each unique Unit entry in that column) and then populated with a 1s or 0s.

2.3 Why did you select these features in your model?

- Rain (Yes/No) – selected this as in the previous Mann-Whitney U test, we were able to show that there is a statistical difference between ridership and rainy versus non-rainy days.
- Fog (Yes/No) – this was selected mainly based on intuition, I made the assumption that more people would ride the subway if it was foggy.
- Precipitation in inches – I assumed that as rain was a good indicator, we can further deduce that the volume of rain could have a factor, this was proved by an increase in the R2 value calculated
- Hour of travel – Again, initially assumed that more people would use the subway based on time of day (late at night, or early in the morning) this was confirmed with an increase in R2 value.
- Daily average temperature in Fahrenheit – Added as there seems to be a trend in the raw data that the largest number of entries seems to correspond to colder temperatures
- Unit – added to allow us to distinguish between ridership at different stations/exits. So removes the assumption that all stations have the same volume

2.4 What are the coefficients of the non-dummy features in your linear regression model?

For my 3 non-dummy features, coefficients are as follows:

- Rain = -14.26727847
- Fog = 65.82026985
- Precipitation = -9.622894
- Hour = 468.39830069
- Daily Average temperatures in Fahrenheit = -74.59692835

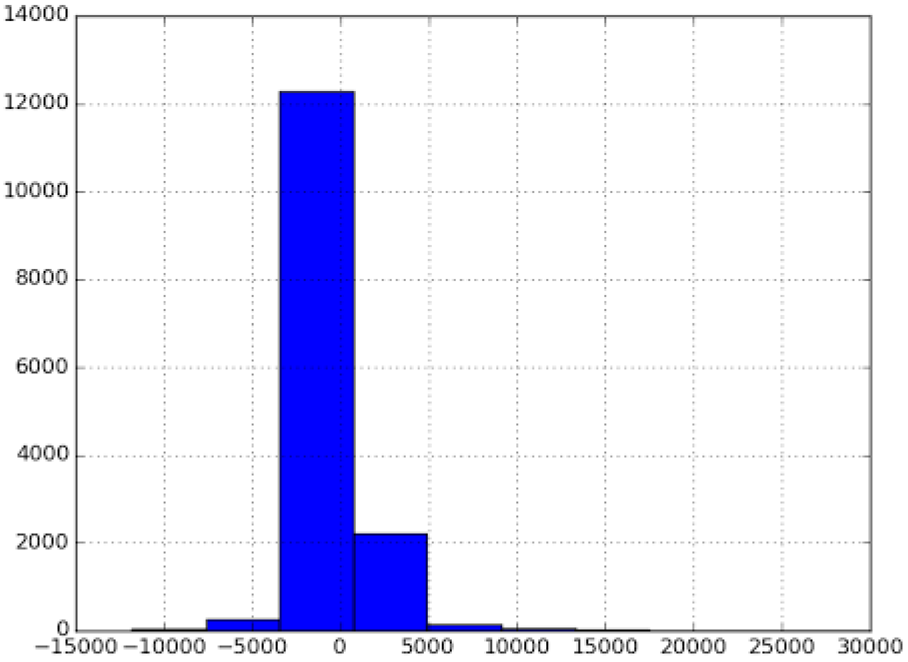
2.5 What is your model's R2 value?

My value for R2 for the above model is 0.464492936908.

2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset given this R2 value?

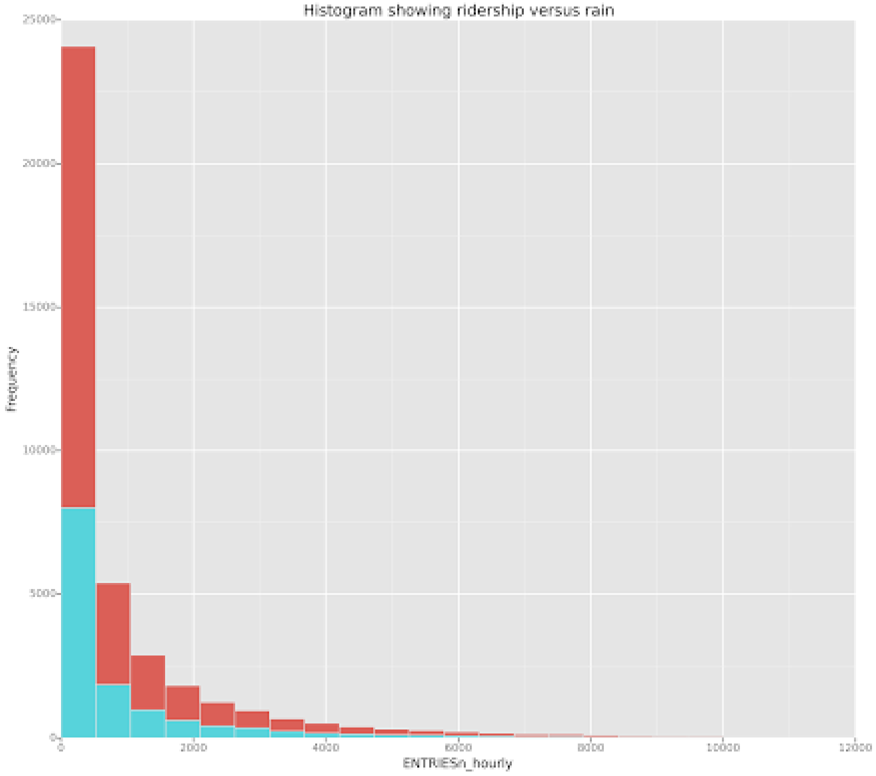
The R squared value tells us how close the data matches to our linear regression line, the closer the R2 value is to 1, the better the fit. During my first interpretation, I assumed that it was a poor fit as it was less than 50% accurate, however on further reading, I don't believe the R2 value alone is enough to predict whether or not our model is sufficient to predict ridership on the NYC subway. A low R2 result also does not automatically mean that the data is a poor fit, it can be due to many other factors such as the fact that human behaviour is difficult to predict. Implementing a histogram showing the residual values we can see that the majority of the values are clustered around 0, although it shows more negative values indicating that our predictions were too high.

In conclusion, I believe that this linear model, along with the residual plot and Mann-Whitney U test suggests that there are strong relationships between ridership on the NYC subway and weather.



Section 3 Visualization

3.1 One visualisation should contain two histograms: one of ENTRIESn_hourly for rainy days and one for ENTRIESn_hourly for non-rainy days.

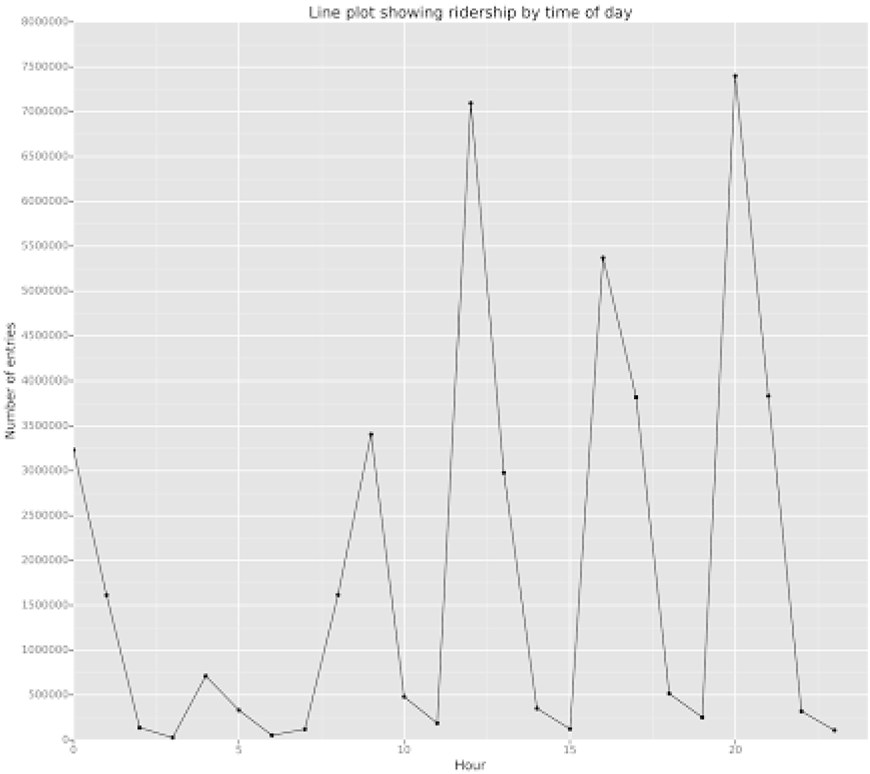


```
df = turnstile_weather[turnstile_weather['ENTRIESn_hourly']<10000

plot = ggplot(df, aes(x=df.ENTRIESn_hourly, fill = 'rain')) + geom_histogram(binwidth = 500) + ggtitle('Histogram showing ridership versus rain') + xlab('ENTRIESn_hourly') + ylab('Frequency')
```

The above graph shows the ridership data between rainy and non-rainy days. This histogram however, does not depict that more people ride the subway on rainy days. Instead, it shows the distribution of the data. There are fewer samples for rainy days compared to non-rainy days.

3.2 Freeform visualization



```
df = turnstile_weather

df = df.groupby(['Hour'], as_index=False).sum()

plot = ggplot(df, aes(x=df.Hour, y = df.ENTRIESn_hourly)) + geom_line() + geom_point() + ggtitle('Line plot showing ridership by time of day') + xlim(0,24) + ylim(0,8000000)+ xlab('Hour') + ylab('Number of entries')
```

This line plot depicts the ridership on the NYC subway system at different hours in the day. Fewer people ride the subway during the early hours in the morning and peak in ridership is during lunch and evenings.

Section 4 Conclusion

Overall, from the various analysis and interpretation of ridership on the NYC subway and weather data, I can say that there is a strong relationship between ridership versus rainy and non-rainy days, however I don't believe it is possible to determine if more people ride the subway when it is raining.

The assumption was first made through intuition where more people will decide to use the subway more often if it is raining outside. Using basic histogram plots, it was also evident that there was a entries per hour for rainy and non-rainy days did not follow a normal distribution and that there are more samples for non-rainy days compared to rainy days. The Mann-Whitney U statistical test then showed that there is a statistical difference between the two distributions (rainy versus non-rainy days) via the rejection of the null hypothesis. Implementing linear regression with the rain variable did improve the R2 value, but it was not the variable that provided the most significant increase in value out of my parameters, that value was 'Hour'.

Section 5 Reflection

From this project, I learnt that there are many different tools and techniques that can be applied to data sets to make interpretations. However, no one tool or technique is able to provide enough proof to make a strong case of future outcomes. Instead, we have to focus on a combination of techniques to build an overall picture of the data. I also found that it is important to have an overview and understanding of the data and the problem we are trying to solve as this allows us to make initial assumptions and use our intuition as a starting point in finding patterns in the data and decide on the most appropriate tools or parameters to focus on.

The dataset although large, needed some cleansing and transformation before we were able to leverage them in our models and there is not always the same proportion of sample size between the different parameters we are trying to measure. Linear regression, as the name suggest also assumes a linear relationship between the variables and is also sensitive to outliers and can skew the results.

Regarding the dataset, I noticed that the thunder column did not vary at all. There was no data samples for ridership when there was thunder. I think it would also be interesting if we could map the UNIT to location of the unit in New York, it would be interesting to see if there is a relationship between ridership and location in the city. The data only included the month of May in 2011 as well, so would be interesting to see how ridership differs across different months that year and perhaps even compare ridership for the month of May for different years.