Analyzing the NYC Subway Dataset

Section 0. References

https://www.moresteam.com/whitepapers/download/dummy-variables.pdf

https://github.com/alfredessa/pandascookbook

http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.hist

http://pandas.pydata.org/pandas-docs/stable/visualization.html#visualization-hist

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

https://pypi.python.org/pypi/pandasql

http://blog.yhathq.com/posts/pandasql-sql-for-pandas-dataframes.html

Udacity discussion forums

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

http://www.tripadvisor.com/Travel-g60763-s208/New-York-City:NewYork:Weather.And.When.To.Go.html

Section 1. Statistical Test

1.1 The Mann-Whitney U test was used to analyze the NYC subway data. A two-tail P value was used.

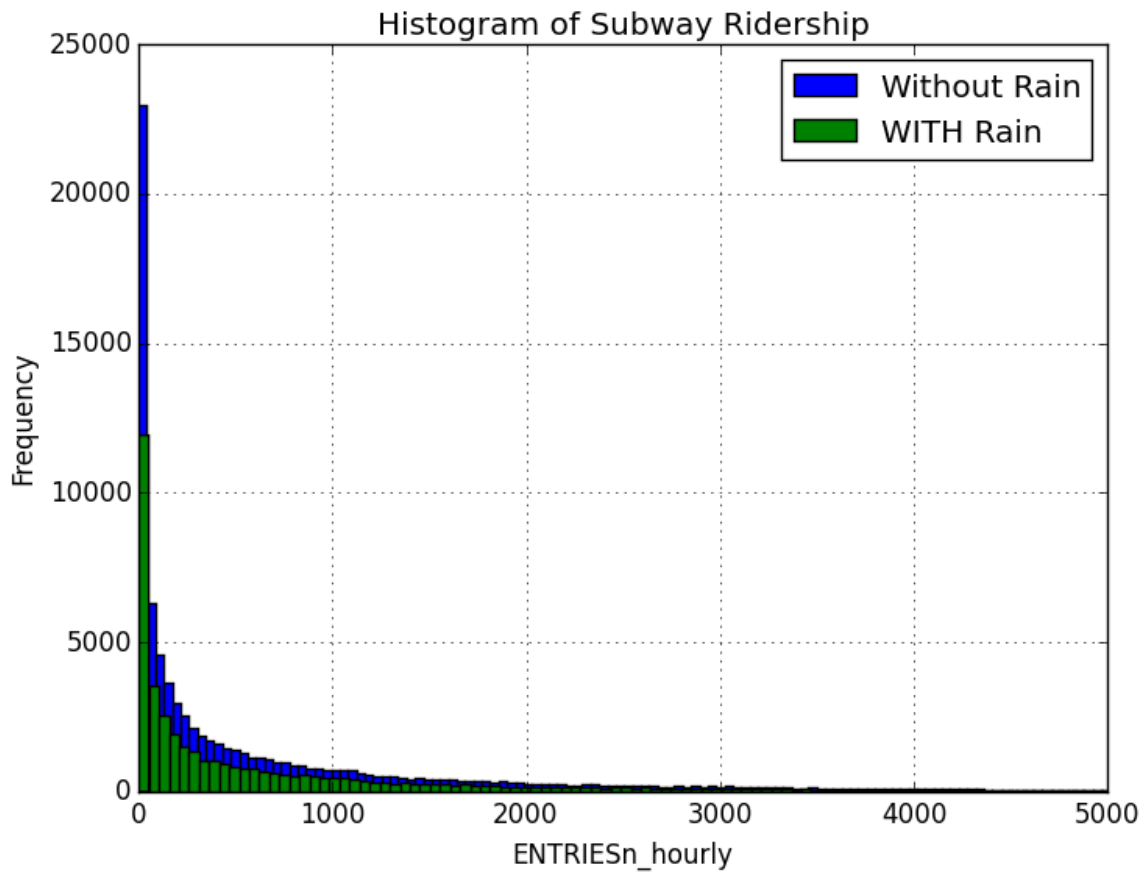
The Null hypothesis H_o : The mean of the subway ridership with Rain = The mean of the subway ridership *without* Rain

The Alternative hypothesis H_a : The mean of subway ridership with Rain is NOT equal to The mean of subway ridership *without* Rain

The P- critical value of 0.05 is used for the U test.

1.2 The Mann-Whitney U test is applicable since it is a non-parametric test. The U test doesn't assume that the data is drawn from any particular probability distribution. A T-test was also considered for the analysis of the data but the T-test requires that the underlying data be normally distributed. Creating histograms for the subway ridership showed that the data was positively skewed and non-normal (see figure 1). Thus the Mann-Whitney U test is more suitable for the comparison of subway ridership data on rainy Vs non-rainy days.

Figure 1 Subway Ridership data



1.3 The results of the Mann-Whitney U Test are summarized below:

P-Value = 0.02499 *2 (for two-tailed test) = 0.4998 ~ 0.5

U-statistic = 1924409167.0

Mean when raining = 1105.45

Mean when NOT raining = 1090.28

1.4 The P-value of ~0.5 means that there is a significant difference between the two data sets – Ridership when raining Vs Ridership with No rain. With a P- critical = 0.05, the P-value is in the P-critical region. Which means that we can reject the Null hypothesis. Furthermore, the

Mean ridership when raining > the Mean ridership when not raining, bolstering the claim that subway ridership increases when raining.

2 Section 2. Linear Regression

2.1 The Gradient Decent approach was taken to compute the coefficients of theta and produce prediction for ENTRIESn_hourly.

2.2 The input variables are as follows:

'rain' , 'precipi',' maxtempi' and 'mintempi'

'rain' is the dummy variable since it has binary values only.

2.3 I picked 'rain' , 'precipi',' maxtempi' and 'mintempi' since the amount of rain probably had an effect on subway ridership, the higher the rain amount the more likely people are take the subway Vs driving. Temperature too may play a role in determining whether or not to take the subway.

2.4 The weights of the non-dummy features in the linear regression are in the table below:

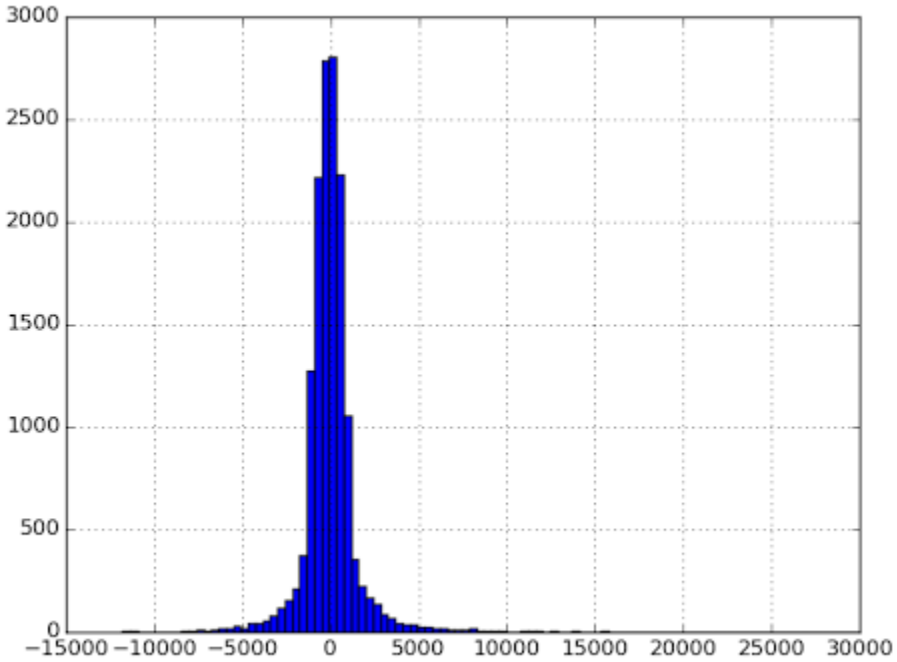
Table 1 Variables and associated Weights

Variable	Weight
precipi	1.79702867e+01
maxtempi	2.54623842e+01
mintempi	-9.01864627e+01

2.5 R2 (coefficients of determination) value = 0.43

2.6 The R2 value means that 43% of the variation in the subway ridership is related to or explained by the variables in table 1. A better linear regression model may be needed to make meaningful conclusions about the relationship between subway ridership and rain. Figure 2 shows a histogram of the residuals plot. Given that the histogram has large tails indicating that there are some very large residuals, the linear regression model is not a very good fit.

Figure 2 Histogram of Residuals



Section 3. Visualization

3.1 Histogram of Subway Ridership:

Figure 3 Subway Rider Ship Histogram

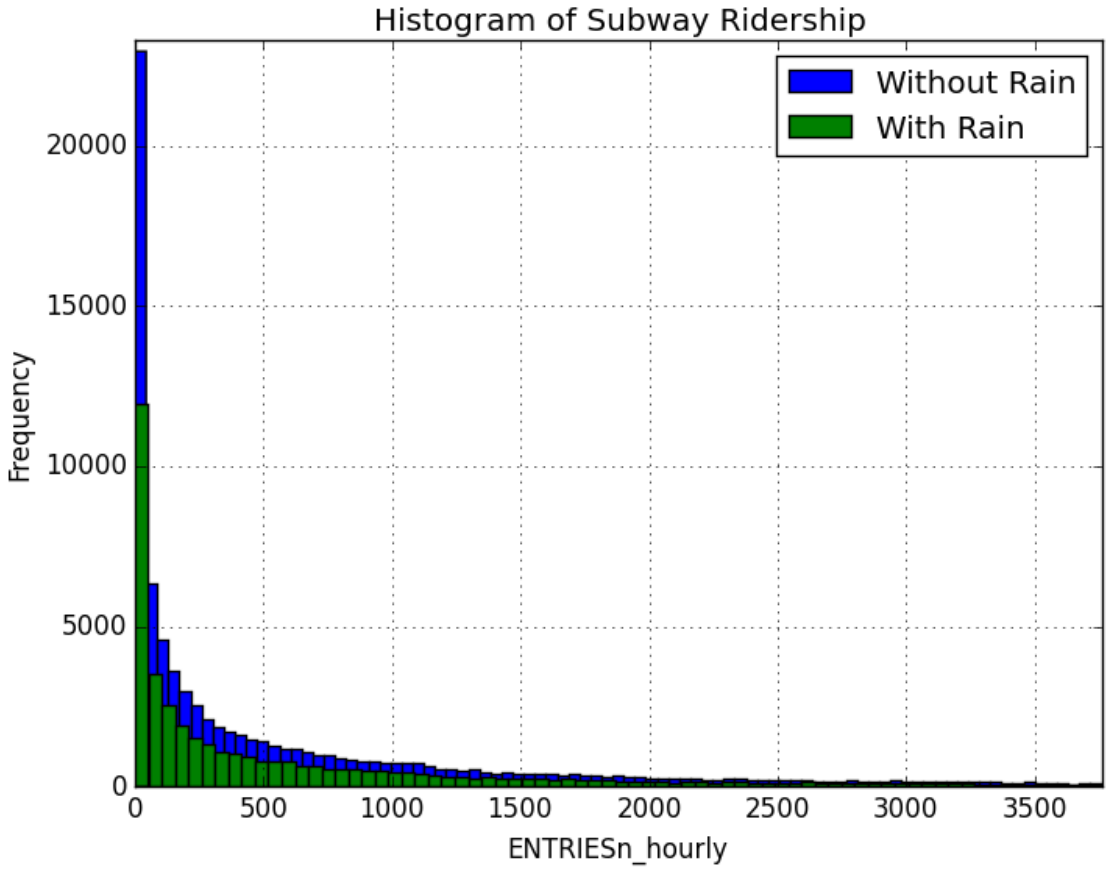


Figure 2 shows two histograms:

- The blue histogram shows subway ridership when there is no rain
- The green histogram shows subway ridership when it is raining

The two histograms illustrate the type of distribution for each data set. Both are positively skewed and non-normal.

3.2 Plot showing the variation in subway ridership by day of the week

Figure 4 Plot of Average Subway Ridership by day of the week

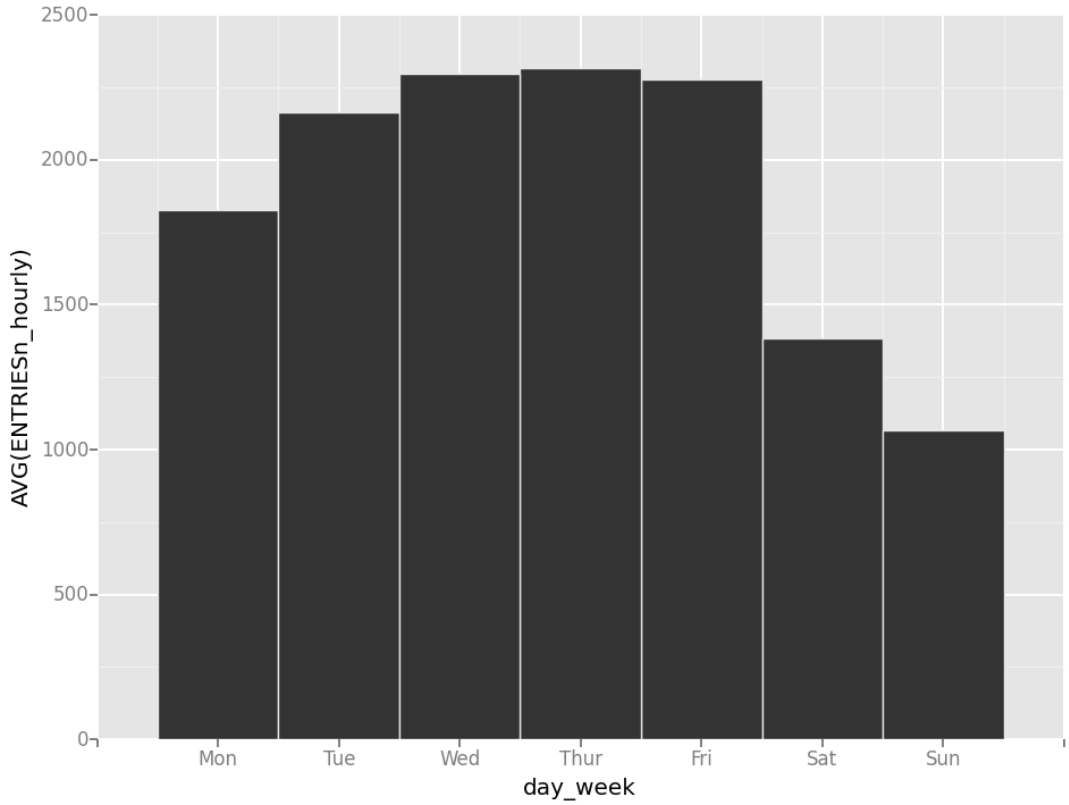


Figure 3 shows the average variation in subway ridership based on the day of the week. The weekends have less traffic which is consistent with fewer travelers commuting to work. The average ridership goes up until Thursday and then starts to drop off. Monday is the best work day to pick subway travel for commuting.

Section 4. Conclusion

4.1 Based on the data analysis weather does have an effect on the subway ridership. In particular the ridership increases when it is raining. The Mann-Whitney U test P- value 0.4998 ~ 0.5 and Mean when raining = 1105.45, which is greater than the Mean when not raining =1090.28, point to an increase in subway ridership when it is raining. Table 2 shows the mean ridership values as a well as the median ridership values. Since the data is skewed (see figure 1) the median is a better measure of assessing ridership. The higher median value for ridership when it is raining points to an increased ridership during rain.

Table 2 Mean and Median comparison when raining Vs not raining

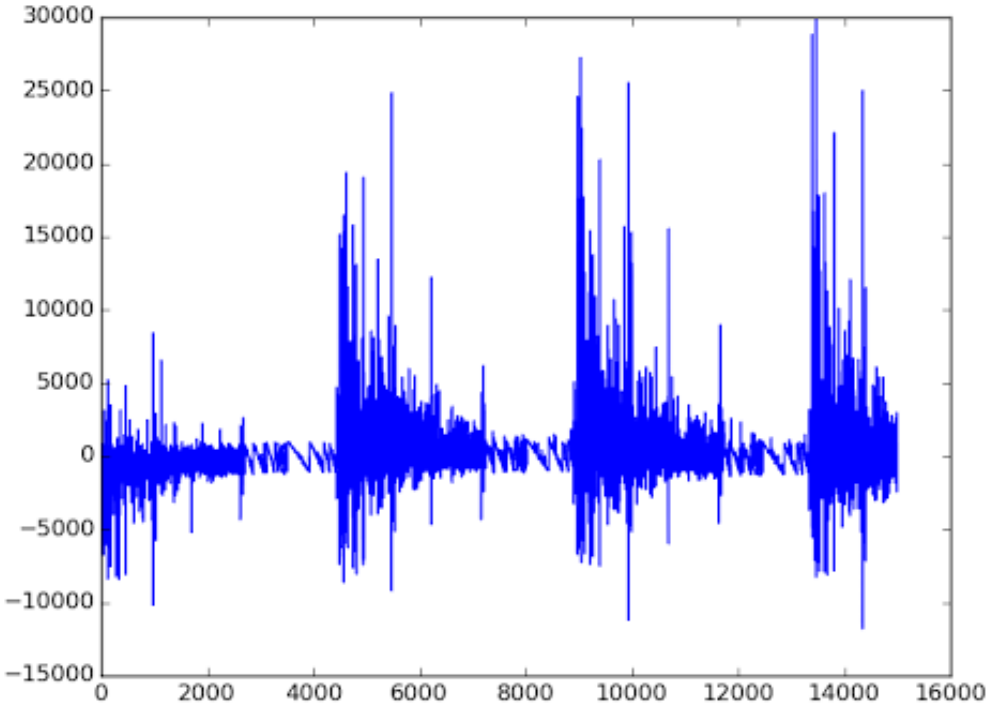
Mean Ridership when Raining	Mean Ridership when NOT raining	Median Ridership when Raining	Median Ridership when NOT raining
1105.44	1090.28	**282**	278

The Linear regression analysis also supports the notion that weather has an effect on subway ridership. Using the variables – 'rain' , 'precipi',' maxtempi' and 'mintempi' yielded R2 value = .43. Which implies that 43% of the variation in the subway ridership is related to or explained by the variables. Although the R2 value is lower than preferred, any prediction regarding human behavior is difficult to model. Taking the R2 value along with the statistical tests it is safe to assume that the subway ridership increases when it is raining. This result also makes sense on a logical level where more people would likely opt to stay off the roads and avoid the stress of driving in rain and low visibility and instead opt for the peace and calm of the subway ride.

Section 5. Reflection

5.1 By plotting the difference between the data points and the predicted values produces a cyclical plot (see figure 5). This indicates that the underlying data is non-linear and therefore a linear regression model is not suitable. Instead utilizing a non-linear model would be a better fit.

Figure 5 Residuals Pot



The data collection period is only through the month of May. The weather changes in NYC quite a bit based on the month. For example July and August weather is described as hot, humid, and sticky and the temperature in the subway can reach ~100 degrees Fahrenheit. In such cases the subway ridership maybe less dependent on rain and more on the temperatures. Just using the month of May doesn't do a good job of capturing the variation in weather. A model for subway ridership derived from only May data is insufficient for making predictions which are valid year round.