Analyzing the NYC Subway Dataset

Section 0. References

As per my knowledge to strengthen my statistics and python knowledge, I used the following references:

- http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
- http://en.wikipedia.org/wiki/Student%27s_t-test
- http://stackoverflow.com/questions/14941366/pandas-sort-by-group-aggregate-and-column
- https://docs.python.org/2/library/time.html
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu
- https://www.udacity.com/course/ud827
- https://www.udacity.com/course/ud201
- http://www.codecademy.com/en/tracks/python
- http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.hist.html
- http://strftime.org/

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze 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?

I used Mann-Whitney U Test to analyze subway data. Reason to use this test was appropriate, since the dataset was non-parametric and non-normal distribution.

I used two tail P value as we want to assess if there is any difference in the distribution of number of entries for rainy and non-rainy days.

Null hypothesis was distribution of number of entries between rainy and non-rainy day is same.

P-critical value = 0.05, so p value less than p-critical value will reject null hypothesis.

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.

Reason to use this test was appropriate, since the dataset was non-parametric and non-normal distribution

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

Rain mean = 1105.44
Without Rain = 1090.28
P Value = 0.0249
U statistics = 1924409167.0

1.4 What is the significance and interpretation of these results?

The mean of subway entries with rain is 1.38% higher than mean of subway entries without rain.

U statistics is half of count (Entries with Rain) * count (Entries without rain).

P value is less than 0.05. So we can say with 95% confidence that the distribution of subway riders on rainy day is different than on a non-rainy day.

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?

- OLS using Statsmodels or Scikit Learn
- Gradient descent using Scikit Learn
- Or something different?

OLS using Statsmodels or Scikit Learn

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

- Rain
- Precipitation (precipi)
- Hour of the day (hour)
- Temperature mean (meantempi)
- Mean wind speed (meanwindspdi)
- fog
- Dummy variables for subway stations (UNIT)

Yes, I used it for unit, since unit is a category and cannot be used for linear regression model. I used pandas.dummies.

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

- Your reasons might be based on intuition. For example, response for fog might be: "I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often."
- Your reasons might also be based on data exploration and experimentation, for example: "I used feature X because as soon as I included it in my model, it drastically improved my R^2 value."
- I included rain in my features list because the results from Mann-Whitney U test showed that there was change in number of riders whether it rained that day or not.
- I included hour because it made sense that number of riders would vary based on the time of the day.
- I included precipi because if it is snowing then people would prefer to use the subway than driving in the city themselves. On similar notion, I included meantempi and fog.
- I included mean wind speed because once I included it, it drastically improved my R^2 value.
- UNIT was one of the important features to be considered because there would be few prime subway stations which would be used more often than a few low profile subway stations.

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

- Rain – 97.37
- Precipitation (precipi) – 635.38
- Hour of the day (hour) – 2835.01
- Temperature mean (meantempi) – 319.09
- Mean wind speed (meanwindspdi) – 828.76
- Fog – 1460.41

2.5 What is your model's R^2 (coefficients of determination) value?

For OLS using Statsmodels implementation R^2 was 0.480456675828

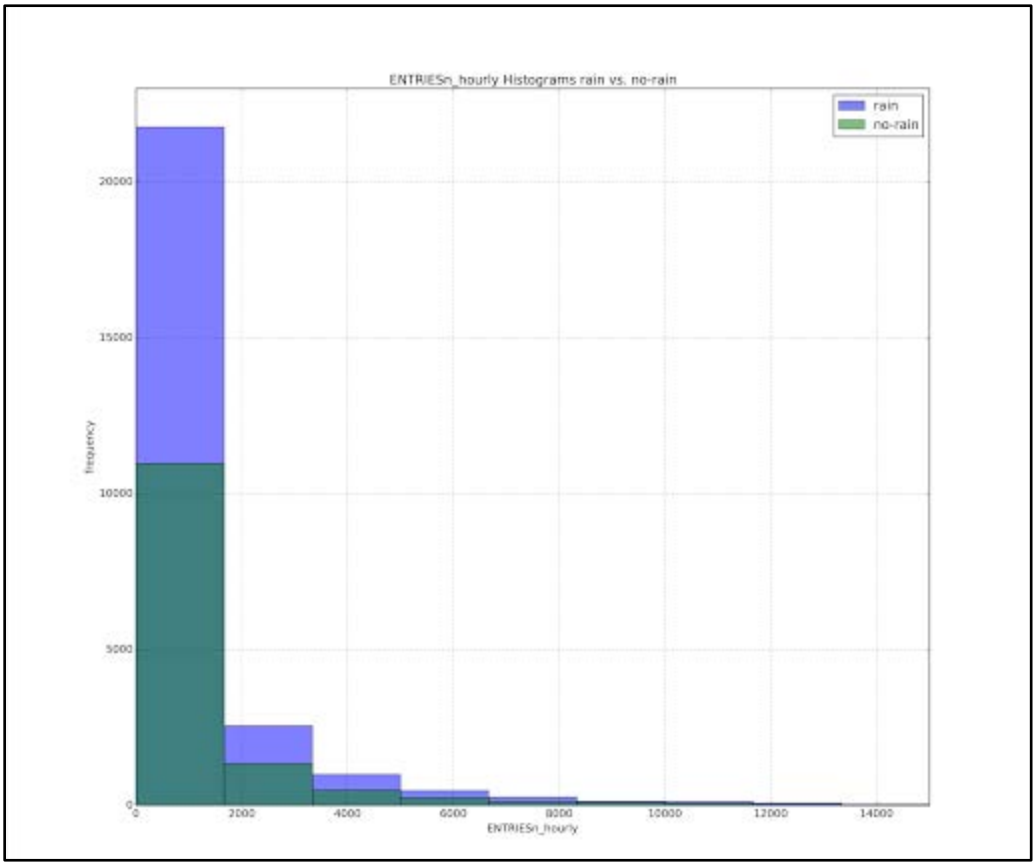
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

This means that 48.04% of the proportion of total variance can be explained by this model. I don't think this model is appropriate to predict the ridership, since it is less than 50 %. We may need to remove the outliners to get a better understanding.

Section 3. Visualization

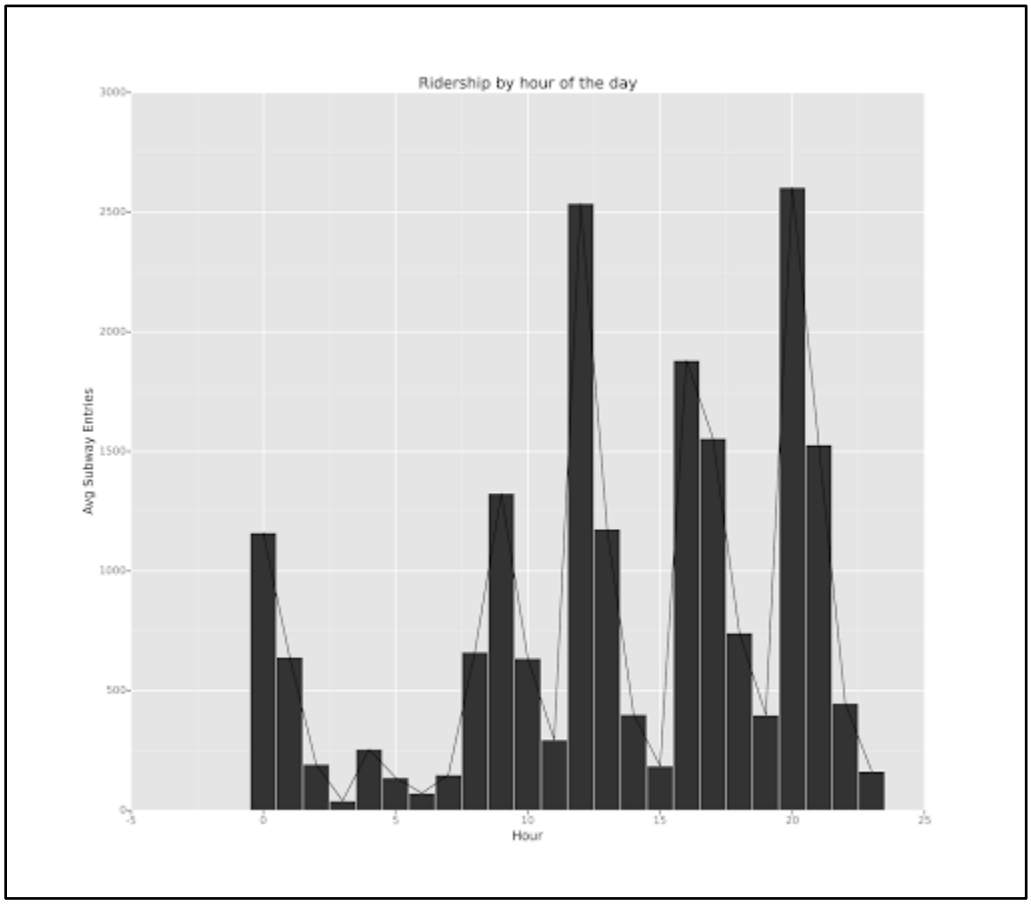
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

- Ridership by time-of-day
- Ridership by day-of-week



Ridership by the hour of the day depicts the average number of NYC riders from all stations riding by the hour of the day

Section 4. Conclusion

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

Even without this analysis, people will ride the NYC subway when it is raining. In addition, to the above conclusion, Mann Whitney U test also leads to that conclusion.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

If we want to assess the effect of rain on riders, we need to only consider rain and units only. So, if we create a linear model with rain and dummy values of UNITs, then coefficient of rain comes out to be positive.

This indicates, if keeping all UNITs constant, rain will have a positive effect on the ridership.

UNIT was one of the important features to be considered because there would be few prime subway stations which would be used more often than a few low profile subway stations. Different subway stations have different geographic locations, connectivity to other subway stations and other factors which affect the number of riders.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset,
2. Analysis, such as the linear regression model or statistical test.

- The first limitation is the dataset is, it is provided only for month of May, this is a huge limitation since data for May, may have similar season for the entire month. This may lead to a conclusion that data is biased. To overcome this limitation data should have been provided for the entire year.
- Second thing I would like to add here is that using dummy variables we included all the subway stations in our linear model. Instead of this, we should have focused on one subway station at a time and then create a different model for each subway station. Because, different subway stations have different geographic locations, connectivity to other subway stations and other factors which affect the number of riders. Further, we should also focus on hour of the day since hour of the day also affects the number of riders. So, by selecting one subway station and one hour of the day we will truly be focusing on effects of weather on number of subway riders.
- Thirdly, we have considered implementing linear model, but we should consider other regression models like polynomial regression models as well because it is not necessary that linear model will fit in for all kinds of dataset. A visual inspection of residuals can further strengthen the conclusion.
- Non –environmental factors like any social events or holidays might see a sudden hike or drop in number of riders which can serve as potential outliers.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?