Section 0: Refrences


http://fch808.github.io/Intro-to-DS-Exercises.html

http://www.statisticssolutions.com/mann-whitney-u-test-2/

http://nbviewer.ipython.org/github/modqhx/DataScienceDegree/blob/master/IntroDataScience/Lesson4_DataVisualization.ipynb

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

http://fch808.github.io/Intro-to-DS-Exercises.html

https://pypi.python.org/pypi/ggplot/

http://stackoverflow.com/questions/25061341/cant-import-ggplot-into-my-python-code

https://www.jetbrains.com/pycharm/help/viewing-diagram.html

https://www.jetbrains.com/pycharm/help/viewing-model-dependency-diagram.html

https://statbandit.wordpress.com/2011/07/29/a-ggplot-trick-to-plot-different-plot-types-in-facets/

http://matplotlib.org/api/pyplot_api.html

http://forum.jetbrains.com/thread/PyCharm-1148

http://stackoverflow.com/questions/14365542/read-csv-file-and-return-data-frame-in-python

http://www.codeitive.com/0SSeUkgeVU/getting-legend-coloring-in-an-empty-ggplot.html

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

## 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 the subway data because the distribution of the data we want to analyze is not normally distributed which means we cannot use Welch's t-Test.

When I use the Mann-Whitney U-test there was a one-tail test.

For the Mann-Whitney we the null hypothesis is that two samples come from the same population.

P-value = 0.024999912793489721

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.

Because the sample data in dataset is large enough to represent the population. And the two populations should be comparable.

For this case, all the observations from both groups are independent of each other because one is about the entries hourly in rainy days and the other is on non-rainy days. The responses are ordinal based on the date and time. And the distribution of both groups are equal under the null hypothesis.

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.

Since the p-value= 0.024,and the mean of entries hourly in rainy day is the m 1105.4463767458733,and the mean for non-rainy day is 1090.278780151855 so we can see the result is significant and we can reject the null hypothesis that those two populations are the same.

1.4 What is the significance and interpretation of these results?

The significance of the test is assumed that with each sample at least n>30 of U-value from the sample approximate normal distribution. So it is significant that those two populations are not parametric which means they are not drawn from any underlying distribution.
Since the P value is small, we can reject the null hypothesis and conclude that the difference between the ridership on  rainy and non-rainy days is not only due to random sampling and instead the population is distinct.

## 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:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

I used the first two ways to compute the coefficients theta.

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

I used the rain, precipi, hour, and meantempi as the independent variables to run the regression with the entries_hourly in the first model and also I used the dummy variable which is unit.

In the second model, I used ['Hour','fog', 'precipi','rain','EXITSn_hourly'] to run the regression with the dependent variable ['ENTRIESn_hourly'].

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."

The reason for me to choose those independent variables is mostly based on my intuition first since I thought when it's more likely that people will use the subway when it is rainy, foggy, or has precipitation and when the temperature is low or too high.

But in the second model when I tried to use those variables I found out that the R^2 is too small so I just drop the certain irrelevant variables based on the outcomes of their t-stats which is less than the critical ones and added the exitsn_hourly in the model and the R^2 is significantly increase afterwards.

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

The coefficients of hour, fog, precipi, rain, EXITSn_hourly are 19.4962,111.0458, -63.1389, 27.1519, 0.7669 respectively and the constant coefficient is 152.3323.

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

The R^2 of the first model is 0.4636.
And for the second one R^2 is 0.554.

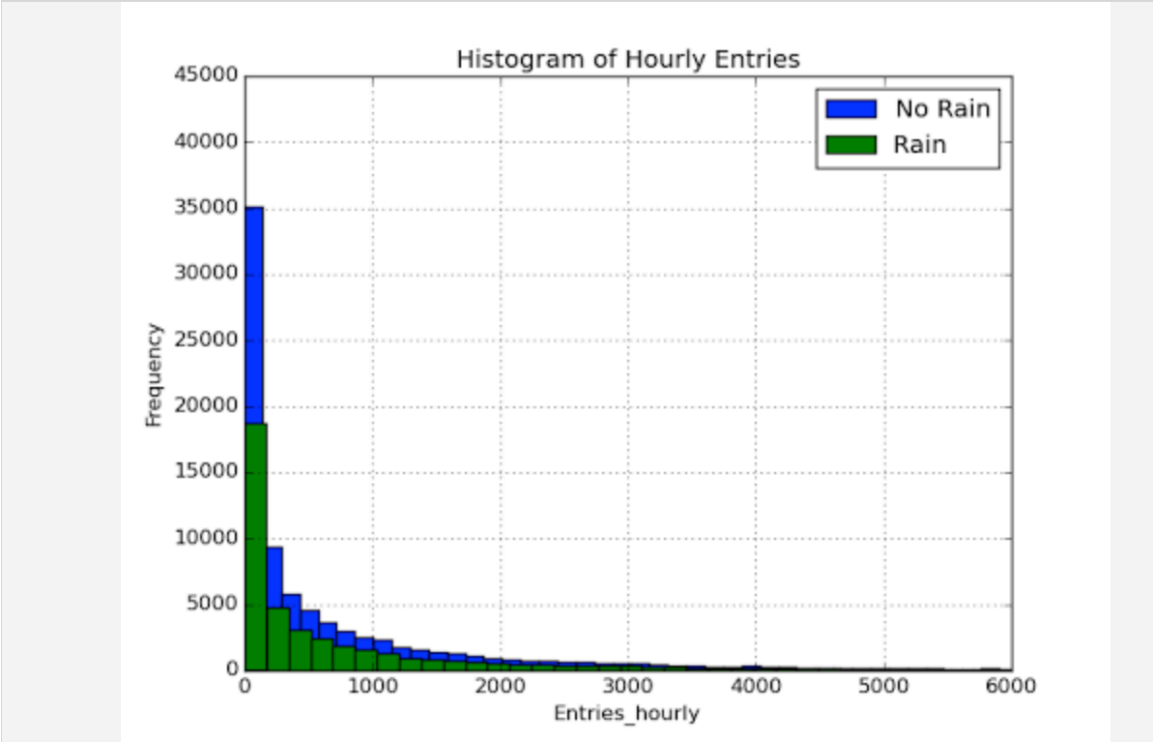
2.6 What does this $R^2$ value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this $R^2$ value?

If the R^2 is large enough which means the goodness of fit for my model is good.

Since the R^2 of the two model are approximately around 0.5 which means the goodness of fit of the model is fair and it can be used to predict the future trend but may not be accurate overall. If the R^2 can be as large as 0.8 or bigger which means the goodness of fit for model is much better.

Section 3: Visualization

3.1




3.2

Plot on Pycharm

Section 4: Conclusion

*Please address the following questions in detail. Your answers should be 1-2 paragraphs long.*

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?

From the data outcome and the graph outcome we can find that there are more people using the subway when the weather is non-rainy although it's partially because the overall non-rainy days is larger than rainy days.
And from the outcome of Mann_whitney U-test we get the outcome of the mean ridership on different weather type: (1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721). From this outcome we can find that the mean ridership on rainy day is 1105.446 which is larger than the mean ridership on non-rainy day which is 1090.279 and the p-value proves that this two populations are significant different from each other.

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.

From the graph we plotted before and the linear regression I did, as well as the statistic test I did.

Firstly, from the outcome of the Mann-Whitney U-test we can find that the two samples are different besides the reason of random selecting. And then we do the linear regression on variable 'rain' and variable 'ENTRIESNs_ hourly' and find out that whether it is raining or not does influence the hourly entries and the mean entries of the rainy day is larger than the mean of non-rainy days.

Although the linear regression outcome shows that the if it is raining or not may not have huge impact on the ridership to some extent.

## 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.

1.The date included in the dataset is May and the data may be biased because the season changed and the weather conditions, temperature and human's behavior may change as well.

And the information provided such as meanpressurei, meantempi, maxtempi, or mindewpti is useless especially when we want to run the regression and find the relationship between independent variables and dependent variables.

2. The linear regression we run in the project shows that there is no strong correlation between chosen independent variables and independent variables since the R^2 is around 0.5 and the t-test of each independent variable is not significant sometimes and there is not many other choices left to replace the unqualified ones.

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