Section 0. References

1. https://docs.python.org/2/tutorial/floatingpoint.html
2. https://pypi.python.org/pypi/pandas
3. http://www.gregreda.com/2013/10/26/intro-to-pandas-data-structures/
4. http://nbviewer.ipython.org/urls/bitbucket.org/hrojas/learn-pandas/raw/master/lessons/01%20-%20Lesson.ipynb
5. http://pandas.pydata.org/pandas-docs/stable/generated/pandas.io.parsers.read_csv.html
6. https://docs.python.org/2/tutorial/inputoutput.html
7. http://www.numpy.org/
8. http://pandas.pydata.org/pandas-docs/dev/generated/pandas.DataFrame.html
9. https://docs.python.org/2/library/datetime.html
10. http://matplotlib.org/users/pyplot_tutorial.html
11. http://matplotlib.org/api/pyplot_api.html
12. Udacity course “Intro to Descriptive Statistics”
13. Udacity course “Intro to Inferential Statisics”
14. http://en.wikipedia.org/wiki/Welch%27s_t_test
15. http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.html
16. http://ggplot.yhathq.com/
17. https://docs.python.org/2/library/sys.html
18. Bill Lubanovic, “Introducing Python”, First Edition, November 2014. Published by O’Reilly Media Inc.,
19. David M. Beazley, “Python Essential Reference”, Fourth Edition, First Printing June 2009

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?

Mann-Whitney U-statistic test is used to analyze the NYC subway data. It is a two-tail P value and the null hypothesis is the distribution of the number of entries is not statistically significant different between rainy and non-rainy days. My p-critical value is 0.05(5%).

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 distribution of the number of entries is non normal distribution as seen visually via histograms; in this case the non-parametric Mann-Whitney U test is more applicable to the dataset because it does not assume the data is drawn from any particular underlying probability 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.

The mean of the number of entries with rain is 1105.45 and the one without rain is 1090.28. The U_statistics is 1924409167 and the P value is 0.025. Please note this P value that was calculated by scipy.stats.mannwhitneyu function is one-tailed value. In order to get two tailed P value, we'll need to double it resulting 0.05 two-tailed P value.

1.4 What is the significance and interpretation of these results?

The U_statistics and P value are both within critical area, so we should reject null Hypothesis, which means the number of ridership with rain and without rain is statistically significant different.

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 used two approaches:

- Gradient descent (as implemented in exercise 3.5)
- OLS using Statsmodels

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 (input variables) used in my models are 'rain', 'Precipi', 'hour', 'meantempi'. I used 'UNIT' as a dummy variables as part of the features.

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.

I decided to use 'rain' and 'precipi'(Precipitation in inches) because I think that when it's raining there would be more people to ride the subway. This is also based on the statistics test of ridership with rain vs. without rain as seen in Section 1. The ridership would be different on the peak hours and non-peak hours, this is why 'hour' is included as feature. Last one 'meantempi'(daily average of tempi) would also affect the ridership frequency.

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

```
[-11.36156201    20.40035623   409.84363995   -51.11442701  1094.94321152]
```

2.5 What is your model's R2 (coefficients of determination) value?

The R2 (coefficients of determination) value of Gradient descent method is 0.46397, and the R2 value of the OLS method is 0.48340

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?

R2 value, also referred to as coefficient of determination, is a quantitative measure of how good our regression model predicts the data. Here is the equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad where\ data = y_1 \dots y_n, prediction = f_1 \dots f_n, mean\ of\ y = \bar{y}$$

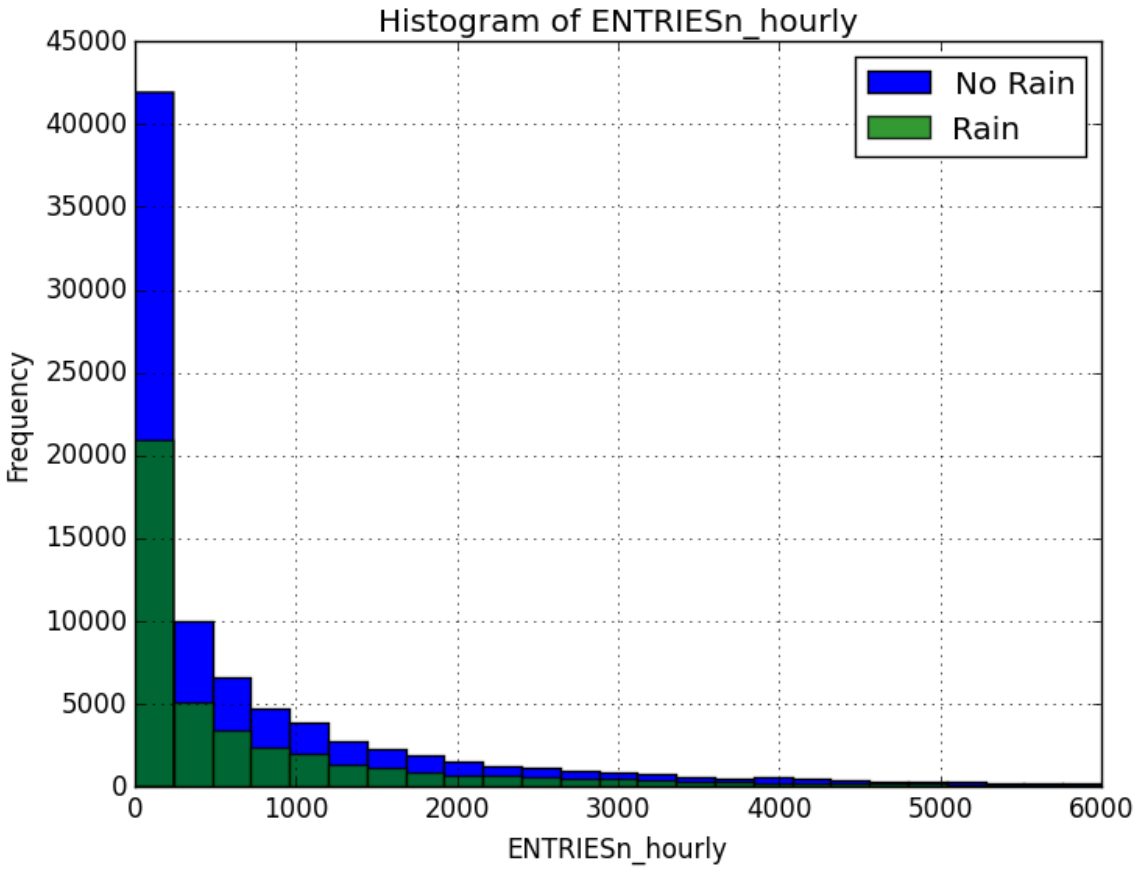
It equals 1 minus the ratio of residual variance and overall variance. When the predictions are close to data R-Square tends to zero; otherwise, when the predictions don't reflect data R-Square tends to 1. Normally R-Square falls in 0 to 1. It explains the percentage of prediction of our model over the original variance.

In this project, the calculated R2 value is about 46% - 48%, which is roughly in the middle of [0 – 1]. It means this model has explained roughly 50% of the original variability. Though further study is preferred, I think to some extend this regression model is appropriate for this dataset.

Section 3. Visualization

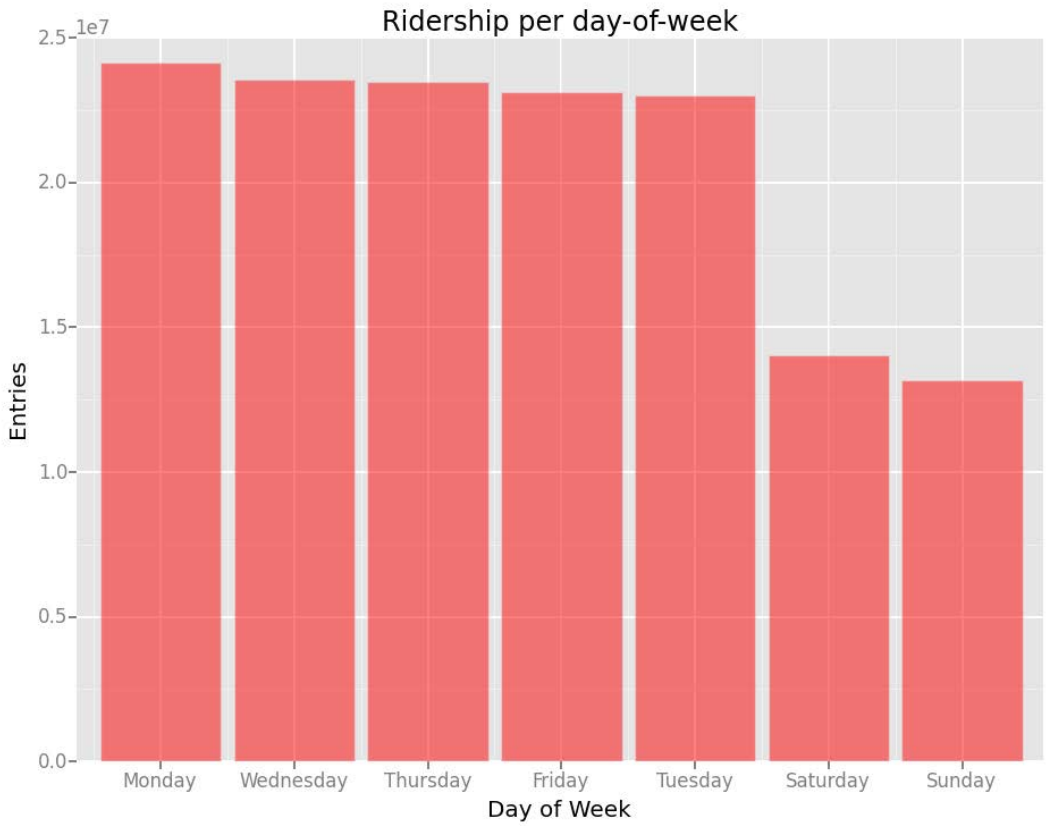
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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



- I combine the two histograms in a single plot.
- In the histograms, I have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis.
- Please note the bins used for “Rain” and “Non-Rain” have been adjusted to identical bin width in order to make the comparison easier to read.

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.



- The second plot shows Ridership by day-of-week. The result is showed as descending order – Monday has the most ridership so Monday is showed as the first one along x-axis and Sunday which has least ridership is the last one in the x-axis.

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?

Based on the analysis and interpretation of the data, my conclusion is more people ride the NYC subway when it is raining than when it is not raining.

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.

The Mann-Whitney U test indicates more ridership when raining vs non-rain. As showed in histogram plot, the ridership (entries_hourly) of NYC subway is not normal distribution (actually it's left skewed). In this case the Mann-Whitney U test is the best fit since Mann-Whitney U test does not assume the data is drawn from any particular underlying probability distribution. The null Hypothesis of Mann-Whitney test is the distribution of the number of entries is not statistically significant different between rainy and non-rainy days. The U_statistic of Mann-Whitney test is 1924409167.0 and two tailed P value is 0.05, both within critical area. So we should reject the null Hypothesis. Since the mean of ridership on rainy day is 1105.45 vs 1090.28 on non-rainy day, we can conclude that the distribution of ridership on rainy day is statistically significant different from distribution of the ridership on non-rainy day – more precisely, more people ride NYC subway when it rains.

Section 5. Reflection

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

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

Regarding the dataset, the turnstile data file only include roughly one month data occurred on May, 2011. So the analysis and conclusion based on this dataset won't reflect the entire year especially winter months. This shortcoming could impact on the conclusion and need to further study/analyze based on wider range of data.

The Mann-Whitney U test used in this project does not take into consideration the turnstile/station factor. As we can see the ridership at each turnstile/station would have different pattern on a rainy/non-rainy day; ridership at some turnstiles/stations could be more sensitive to weather condition than others. Further study of ridership influenced by weather should be done based on each turnstile/station to show potential more accurate result.

Both Gradient-Descent and OLS models used in this project are linear regression model. It's still an assumption and simplification that this analysis has a linear relationship. Hence, a further analysis based on polynomial combinations could have a better or more accurate result.

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

I used ggplot to plot the residual per data point to further analyze the accuracy of the gradient-descent linear regression model used in this project. The result shows as below. Please pay attention to the trend line – it's very clear that the blue trend line of residual is almost proportional to data point (except the area close to original point but that's so small and can be ignored). This relationship may improve the prediction of the data. For example, let's say:

$$Residual = a * Datapoint, a\ is\ known\ from\ the\ trendline$$

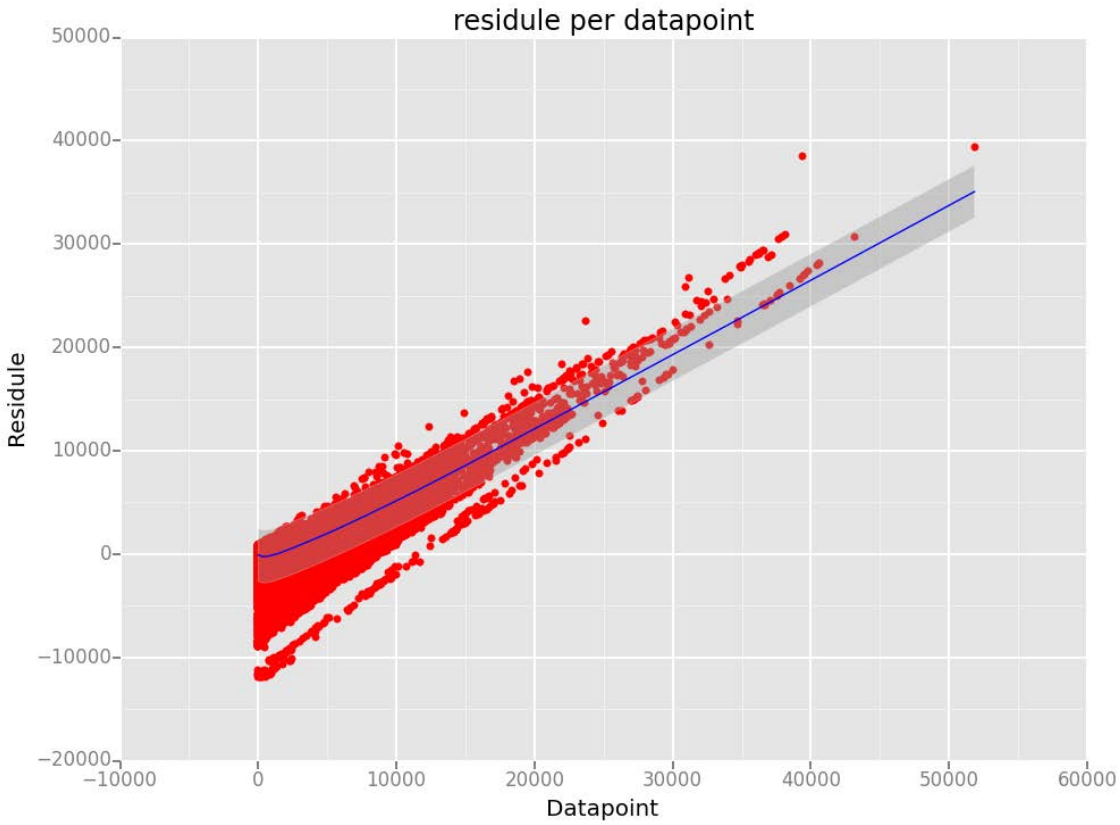
Since:

$$Residual = Datapoint - Prediction$$

We have:

$$Datapoint = \frac{Prediction}{1 - a}$$

We could use above formula to further improve the prediction of our linear regression model.


residule per datapoint
Residule
Datapoint
50000
40000
30000
20000
10000
0
−10000
−20000
−10000
0
10000
20000
30000
40000
50000
60000