Analyzing the NYC Subway Dataset

Section 0. References

PS 2-1: https://dev.mysql.com/doc/refman/5.1/en/counting-rows.html
PS 2-5: http://goo.gl/HBbvyy
http://docs.python.org/2/library/csv.html#examples
http://www.learnpython.org/en/Loops
PS 2-8: http://stackoverflow.com/questions/10982089/how-to-shift-a-column-in-pandas-dataframe
PS 2-9: http://stackoverflow.com/questions/3098248/time-to-decimal-time-in-python
PS 2-11: http://docs.python.org/2/library/datetime.html#datetime.datetime.strptime

PS 3-1: http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms
https://discussions.udacity.com/t/matplotlib-histogram-for-days/24272/13
PS 3-3: http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
PS 3-5: http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
PS 3-6: http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm
PS 3-8: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html
https://discussions.udacity.com/t/problem-set-3-8-recovering-parameters/21855/2

PS 4-1,2: https://discussions.udacity.com/t/prob-set-4-2-error-pivot-table-got-an-unexpected-keyword-argument-rows/30187
https://discussions.udacity.com/t/lesson-4-lineplot-compare-legend-not-generated/19333

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 the Mann Whitney U-test to analyze the NYC subway ridership (i.e. value of ENTRIESn_hourly). Two-tailed p-value is used in the analysis. The null hypothesis is that the number of entries of the NYC subway in rainy days is not different from the non-rainy days. The p-critical value is set to 0.05.

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.

In the two samples, the riderships are not normally distributed. Therefore, it violates the normal distribution assumption in Welch's T-test. Mann Whitney U-test is the equivalent test for the non-normally distributed data.

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 number of entries in rainy days is 1105.45; the mean number of entries in non-rainy days is 1090.28. In the Mann-Whitney U-test, U = 192, p = .025. Therefore, at 95% confidence level, the p-value for two-tailed test is doubled to p=.05.

1.4 What is the significance and interpretation of these results?

The number of entries in rainy days is marginally different from non-rainy days (p=.05 two-tailed). It also suggests that people use the subway more often in rainy day, when compared to non-rainy days (mean_rainy > mean_non-rainy, p=.025 one-tailed).

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 OLS in Statsmodels in PS 3-5, and tried Gradient descent in sklearn.linear_model in PS 3-8.

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

The following features are used in the model, including rain, precipitation, fog, the time of entries (i.e. 'Hour'), the mean of temperature, the mean of wind speed and a dummy variable 'UNIT'.

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, precipitation, fog, the mean of temperature and the mean of wind speed is because I believed that worse weather condition might drive people to use the subway more often.

I decided to use the time of entries (i.e. 'Hour') is because I guessed some specific time, such as the rush hours, might drive people to use the subway more often.

I used the dummy variable 'UNIT' because the R^2 value improved after including it to the model.

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

Using the OLS approach, the parameters of the non-dummy features are as followed:

```
rain              0.040560
precipi         -73.976935
Hour             65.364525
meantempi        -9.491420
fog             214.086883
meanwindspdi     32.568316
```

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

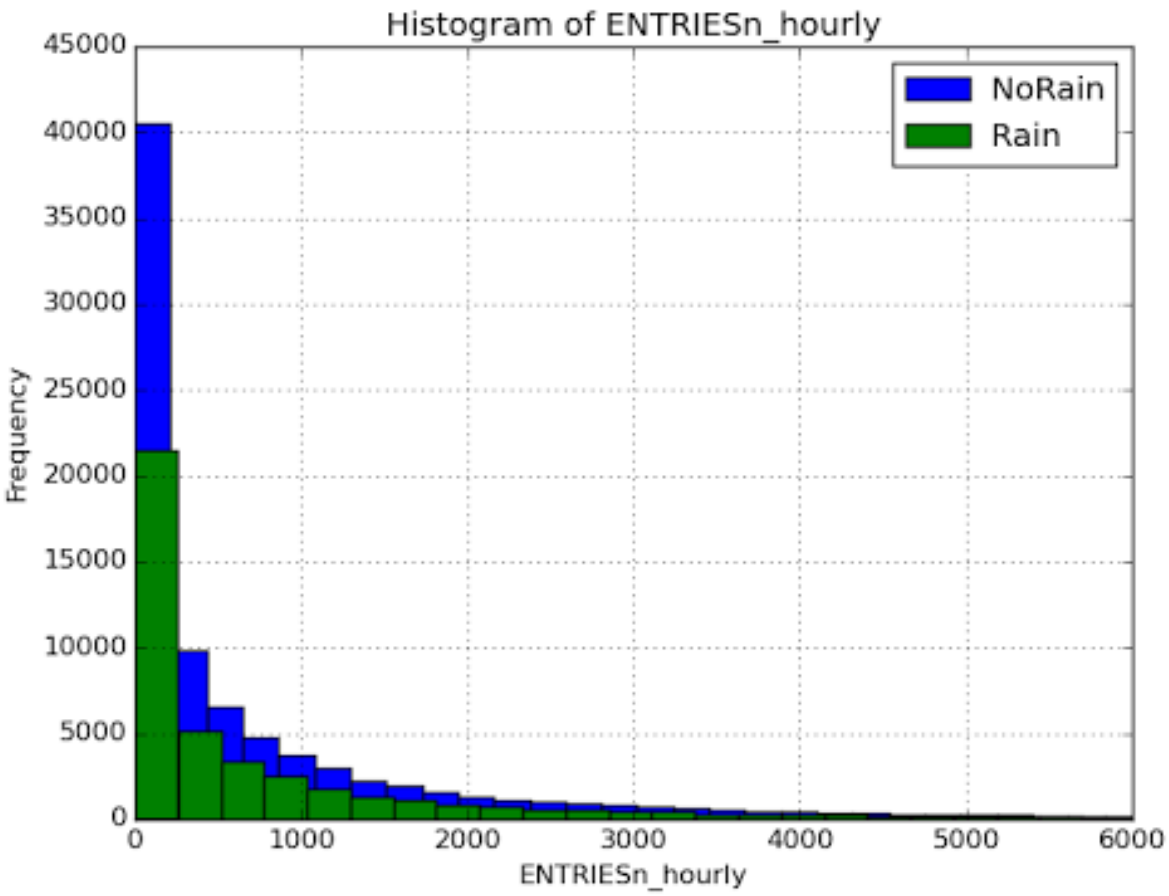
Using the OLS approach, the model's R^2 value is 0.480.

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?

The model's R^2 value indicates that the model explains 48% of the variability of the ridership around its mean. Given this R^2 value and the model's residual plot, I think the model to predict ridership is appropriate for this dataset. However, the histogram of the residual plot showed long tails at both ends, indicating it may not be normally distributed and there are some very large residuals generated by this model, which can be a reason to question this linear regression model.

Section 3. Visualization

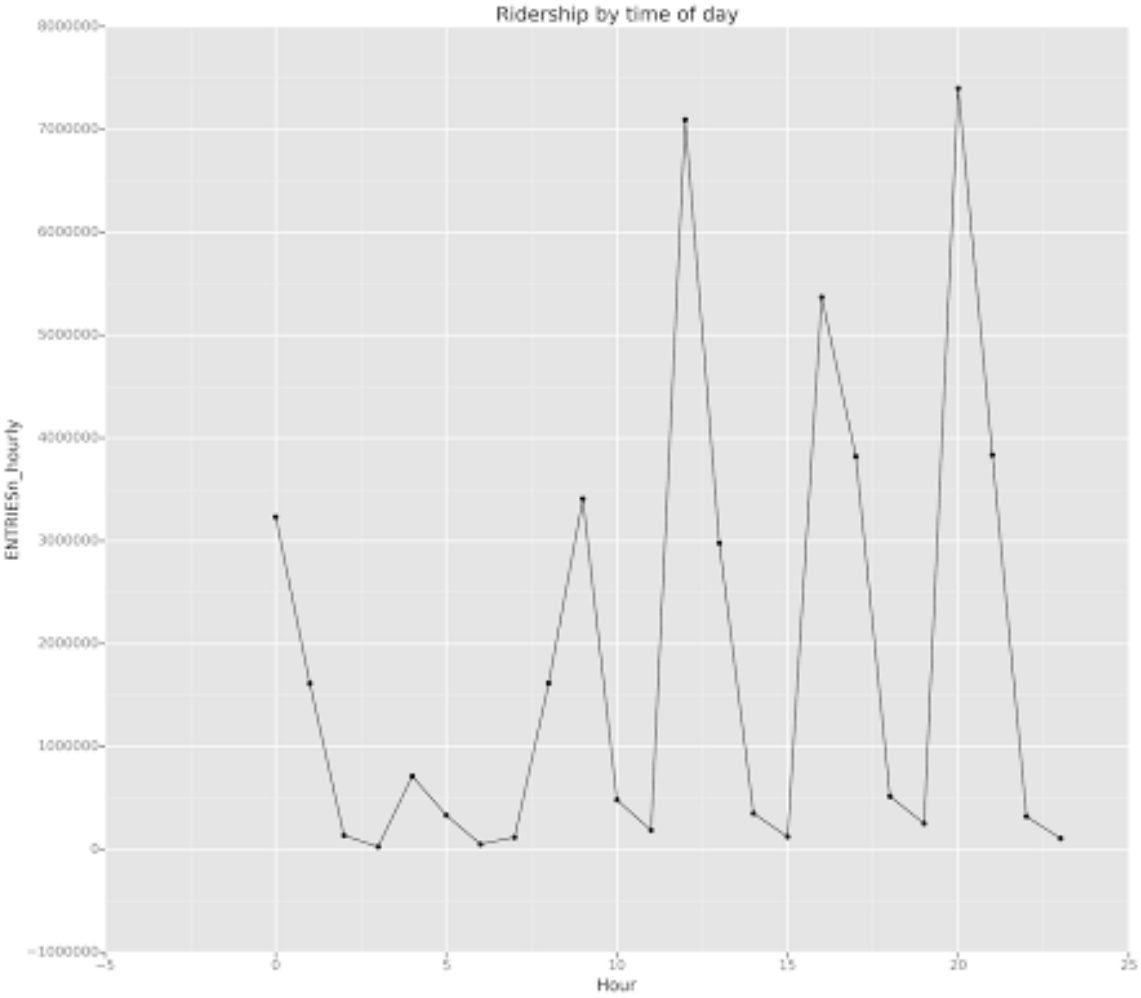
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



The histogram is generated using matplotlib.pyplot. The bin size is 200.
The x-axis in this histogram has been truncated at 6,000 cutting off outliers in the long tail which extends beyond 50,000.

From the histogram, I noticed that the sample of non-rainy days is larger than the sample of rainy days. In addition, both samples are positively skewed, with a long tail extending beyond 50,000. Therefore, both samples are non-normally distributed.

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.



From the line plot of the ridership by time-of-day, I noticed that there are several peaks of entries at the time of midnight, 9AM, 12PM, 16PM and 20PM. It suggests that besides the rush hours (i.e. 9AM, 12PM and 16PM), people tend to ride more often in the evening (i.e. the highest peak at 20PM) and at the midnight (i.e. 0AM). In contrast, the entries at 2-3AM, 6-7AM are the lowest.

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?

The analysis and interpretation of the data suggest that more people ride the NYC subway in the rainy days when compared to the non-rainy days.

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.

First, using the Mann Whitney U-test, group comparison between the rainy days and the non-rainy days showed that the mean entries of NYC subway in rainy days (mean_rainy = 1105.45) is significantly greater than the mean entries in non-rainy days (mean_nonrainy = 1090.28, U=192, p=.025, two-tailed) at 95% confidence level.

Second, in the linear regression the coefficient of rain is 0.041, suggesting that compared to the non-rainy days, the number of hourly entries in rainy days increase 0.041. Given that the model accounts for 48% of the variability (R^2 =.48), it suggests the model to predict ridership is appropriate for this dataset.

Taken together, both statistical test and linear regression can lead to the above conclusion.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test.

One of the shortcomings can be the different sample sizes of rainy and non-rainy days. From the histograms, I noticed that the sample of rainy days is smaller than the non-rainy days. This imbalance sample size might lead to potential type I error in the statistical test. In addition, the time spam in this dataset seems not long enough to capture the ridership.

In the regression model, the residual plot showed long tails, indicating there are some very large residuals and the distribution of residual may not be normal. Therefore, the influence of potential outliers should be accessed in the linear regression model. Given that the outliers can significantly influence the regression result, screening outliers are necessary. In addition, some of the features (i.e. minimum, mean and maximum

temperature) in this dataset may be inter-correlated, which increases the risk of problems with collinearity.

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

From the linear regression, I also noticed that people tend to ride the subway more often when it is foggy (fog coefficient = 214.09), or when the mean temperature drops (meantempi coefficient = -9.49). They also tend to ride the subway more often when it's more windy (meanwindspdi coefficient = 32.57), or less precipitated (precipi coefficient =-73.98).