Analyzing the NYC Subway Dataset

Questions

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.

This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

Section 0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

U-test
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu
Drop non-numeric columns: http://stackoverflow.com/questions/26771471/remove-rows-where-column-value-type-is-string-pandas
Feature selection:
http://scikit-learn.org/stable/modules/feature_selection.html
R2:
http://en.wikipedia.org/wiki/Coefficient_of_determination
Numpy tutorial:
http://cs231n.github.io/python-numpy-tutorial/
Decision tree:
http://scikit-learn.org/stable/modules/tree.html
OLS tutorial:
http://www.datarobot.com/blog/ordinary-least-squares-in-python/
How to interpret R-squared: http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

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?
The statistical test used was Mann Whitney U-test.
1-sided p-value.
Null hypothesis is that the two populations have the same distribution.
P-critical value: 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.
Because the two samples are not normally distributed, the statistical tests such as Welch's T test can not be used.
Mann Whitney U-test being a non-parametric test can be used.

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.

```
Mean_with_rain: 1105.4463767458733

Mean_without_rain: 1090.278780151855

U: 1924409167.0

p-value: 0.024999912793489721
```

1.4 What is the significance and interpretation of these results?
Since p-value is less than the p-critical value we can reject the null hypothesis and conclude the raining does make a difference in subway ridership.

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:

a. Gradient descent (as implemented in exercise 3.5)
b. OLS using Statsmodels
c. Or something different?

Gradient descent was used

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
Features used: 'minpressurei', 'precipi', 'Hour', 'meantempi', 'UNIT'
'UNIT' was converted to dummy variables

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."

From visualizing subway data (I actually did the problem set 4 first and then came back to PS3) I find there are certain hours with much higher ridership than others, which is also common sense because people ride to work 8~9am and ride home 4~5pm.
From the previous section it was found rain has an effect on ridership so I decided to use 'precipi' as a predictor – and by the same vein I also throw in 'minpressurei' and 'meantempi' just by trial and error (ie, R^2 improved). 'rain' and 'fog' are not used because those are correlated with other weather measures already. Last but not least, 'UNIT' was also used as a predictor because the popularity of stations vary greatly, we want to allow gradient descent to assign weights based on different stations.

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

Table of features and their weights:

'minpressurei'	'precipi'	'Hour'	'meantempi'
-6.45933787e+01	8.37120079e+00	4.67969047e+02	-6.79979343e+01

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

```
0.464682234412
```

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?

R^2 is the ratio between explained variation and total variation.
R^2 of 0.46 means the regression model accounts for 46% of the variation in the data. If the regression model is perfect R^2 would be 1.
According to [1], "any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%," so I guess 46% is okay, however the same source cautions us to look at the residual plot to find out if the model is biased, that may actually be more important than R^2.

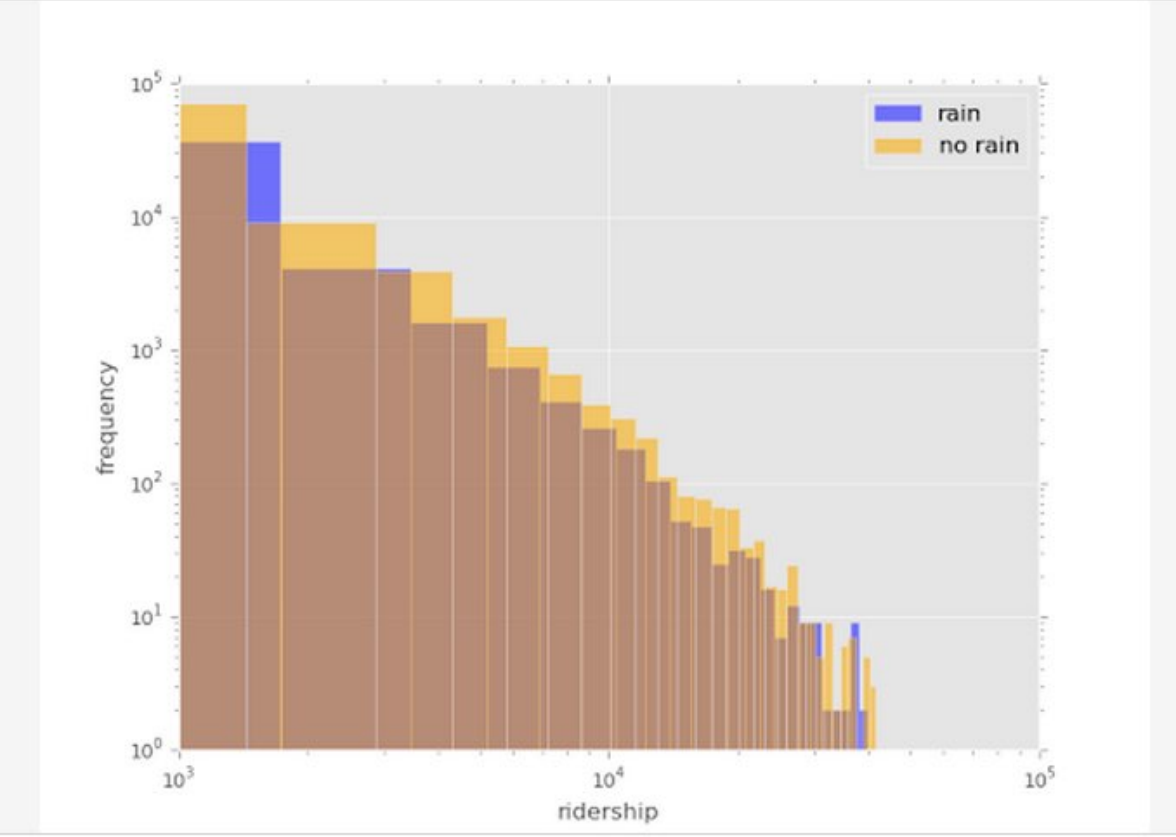
[1]: http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

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.

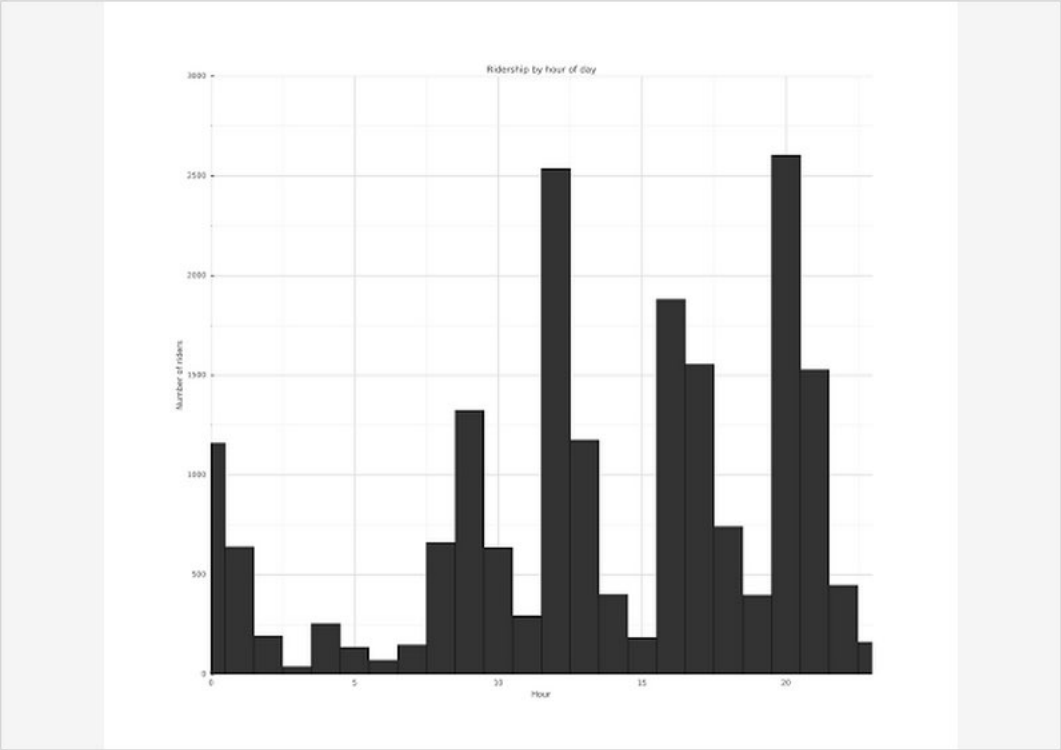
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, 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

```
The image produced by your code is shown below
```



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?

More people ride subway when it's 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 subway data was partitioned into 2 groups, group 1 pertains the raining condition and group 2 pertains the not raining condition, the group 1's average ridership is higher than group 2. To assess statistical significance of this result, Mann Whitney U-test was used and yielded 1-sided p-value of 0.024 which is less than p-critical value of 0.05, thus we can reject the null hypothesis and conclude the two

groups are not from same population. The OLS regression model also supports the view that ridership increases when it rains. One of the independent variables used was precipitation and its corresponding coefficient is positive, suggesting the model expects ridership to increase from each unit-increase in precipitation.

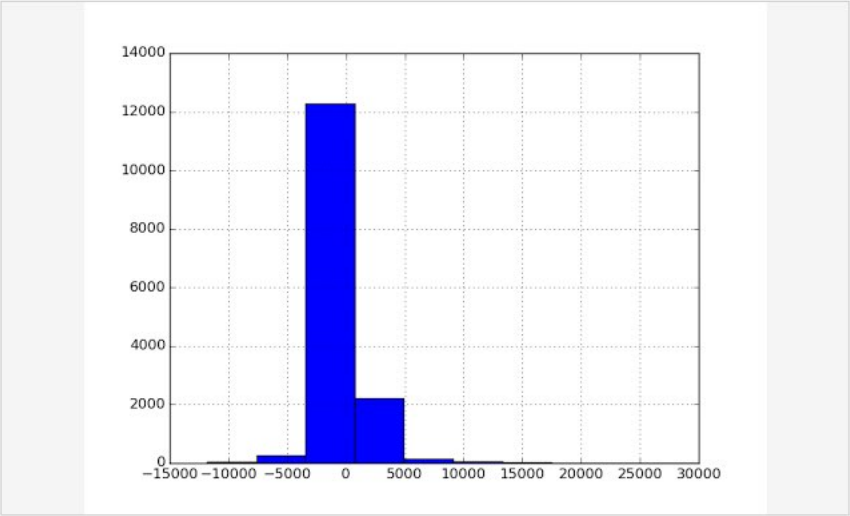
Section 5. Reflection

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:

1. Dataset,
   The most obvious thing that can improve the result is to have understanding of the geography. If we know the connectivity between the stations and the distance between each pair that would help prediction. If we have this data we can identify major sources and sinks of traffic, coupled with time of day, we can model morning rush and evening rush behaviors.

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



   From plotting the residual I find my regression model is biased, it tends to underpredict. However, rather than trying to beat a dead horse, I think a better model can be made using decision tree. (Note, I was taking udacity machine learning class before signing up for nanodegree). With decision tree, you naturally find cuts in the data that separate them into classes that minimize error. I think that's better suited to the type of data we have.

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