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.

To prepare for the “Intro to Data Science” course, I first brushed up on statistics by watching all of the Khan Academy videos on Descriptive and Inferential Statistics:
https://www.khanacademy.org/math/probability/descriptive-statistics
https://www.khanacademy.org/math/probability/statistics-inferential

I also reviewed my college statistics textbook:
Kolstoe, Ralph H. *Introduction to Statistics for the Behavioral Sciences*. Revised Edition. Homewood, IL: The Dorsey Press, 1973. Print.

I went to Google to search for many Python-related things. For example, I used the following site to review Python Dictionary methods:
http://www.tutorialspoint.com/python/python_dictionary.htm

On the Udacity site, I participated in the following discussion forum to learn more about the dummy variables (UNIT values used as category data). I am thomas_219260 in this discussion thread:
http://discussions.udacity.com/t/dummy-variables-in-linear-regression/14451/2

Also on the Udacity site, I looked over the following discussion about one-sided and two-sided p values for the Mann-Whitney U Test.
http://forums.udacity.com/questions/100153716/if-the-mann-whitney-u-test-returns-a-one-sided-p-value-what-is-the-null-hypothesis

To learn more about the Mann-Whitney U Test and how to interpret its results, I looked here:
http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://www.isixsigma.com/tools-templates/hypothesis-testing/making-sense-mann-whitney-test-median-comparison/

The following sites helped me to learn more about analyzing the results of the linear regression model:
http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html

Finally, the following pages provided an explanation of how to interpret R^2 values in regression analysis:
http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values

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?

To analyze the data, I used the Mann-Whitney U Test. While the Python implementation of this test yields a one-tailed p-value, I used the two-tailed p-value (arrived at by doubling the one-tailed P value). The null hypothesis assumes that both samples came from the same population. My p-critical value was 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.

The Mann-Whitney U Test is an example of a non-parametric test, which does not assume that the data is drawn from any particular underlying probability distribution. In particular, it does not require the underlying data to be normally distributed.

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 Python implementation of the test returned a U-value of 1,924,409,167.0 and a one-sided p-value of 0.02499991 (doubling of which yields the two-sided p-value of 0.04999982). The mean of the “subway ridership when raining” sample was 1105.446 and the mean of the “subway ridership when not raining” sample was 1090.279.

1.4 What is the significance and interpretation of these results?

The two-sided p-value of 0.04999982 is less than critical p-value of 0.05, so I reject the null hypothesis that the two samples came from the same population. In other words, I observed a statistically significant difference between the sample of subway ridership data taken while it was raining, and the sample of ridership data taken while it was not raining. Put simply, subway ridership is higher overall when it rains. However, as depicted in the visualization in Section 3.2 below, subway ridership actually is lower on weekends when it rains, when perhaps people would be more inclined to stay home.

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?

I used the Gradient Descent approach in exercise 3.5 and the OLS using Statsmodels approach in optional exercise 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?

My features included rain, precipi, Hour, and meantempi. One set of dummy variables were included, which provided category values based on each distinct UNIT value present in the input data.

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."

To be honest, I stuck with the default set of feature variables (i.e., rain, precipi, Hour, and meantempi) included in the code provided with the exercise. Intuitively, it seems reasonable to expect subway ridership to be higher when it's raining (and the application of the Mann-Whitney U Test in the earlier exercise provided statistical evidence to justify this intuition). Hence, the rain-related variables should help in predicting ridership. Similary, temperature (meantempi) should be a decent predictor, as it seems reasonable to expect ridership to be higher during cold or hot weather, rather than during mild weather (in which people might be more willing to walk to their destination). Finally, time of day (Hour) should be a good predictor, with higher ridership expected during rush hour commutes.

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

In my final run, my theta coefficient for the rain variable was 2.924, for the precipi variable was 14.653, for the Hour variable was 467.709, and for the meantempi variable was -62.218.

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

In my final run, the model's calculated R^2 value was 0.464.

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 linear model's calculated R^2 value of 0.464 might be considered midling, but it confers that the model does explain 46.4% of the variation of the predicted variable's (i.e., ENTRIESn_hourly) values from its actual values. In other words, the model does expose a trend indicating that the feature variables do provide useful information about the predicted variable.

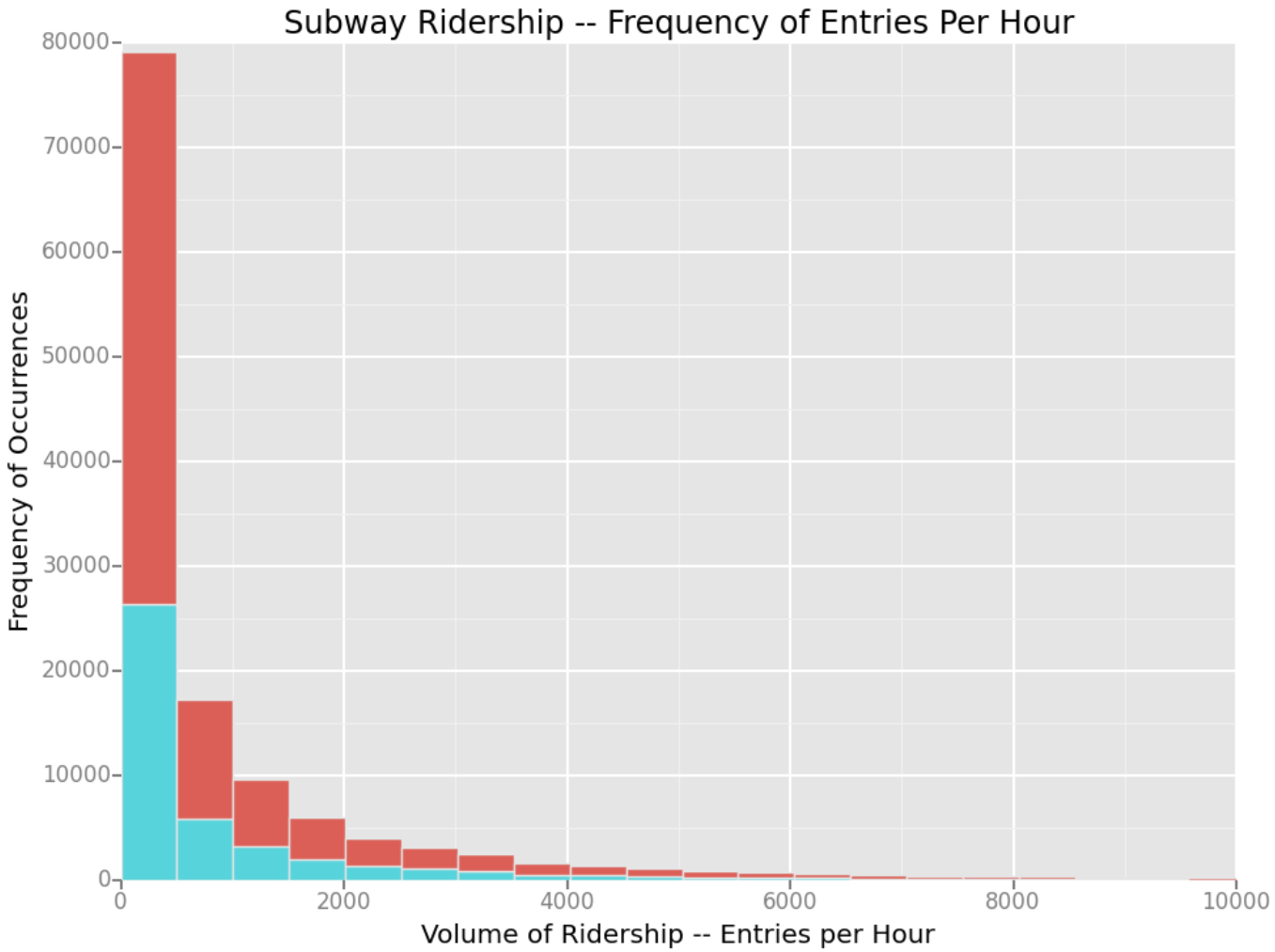
Seeing a coefficient of determination value closer to 1, or even greater than 0.5, would give me more confidence in this model's ability to predict subway ridership accurately. Riding the subway is however a human behavior, which will limit the exactness of any prediction model. I did try improving upon the model by processing the "improved dataset" provided in the instructions for this project (turnstile_weather_v2.csv) on my local machine. With the same set of feature variables (rain, precipi, Hour, and meantempi) produced an R^2 value = 0.445. Including the weekday variable improved the fit somewhat, producing an R^2 value = 0.467. That is, drawing a distinction between weekdays and weekends yielded a slight improvement in the coefficient of determination.

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.



The above stacked histogram depicts the distribution of NYC subway ridership volume frequencies falling within intervals of five hundreds, comparing “entries per hour” data from when it is raining (blue bars) to when it is not raining (red bars). For example, the leftmost bar shows that there were 26305 occurrences (i.e., rows of data in the collection) in which the ridership volume was less than or equal to 500 when it was raining,

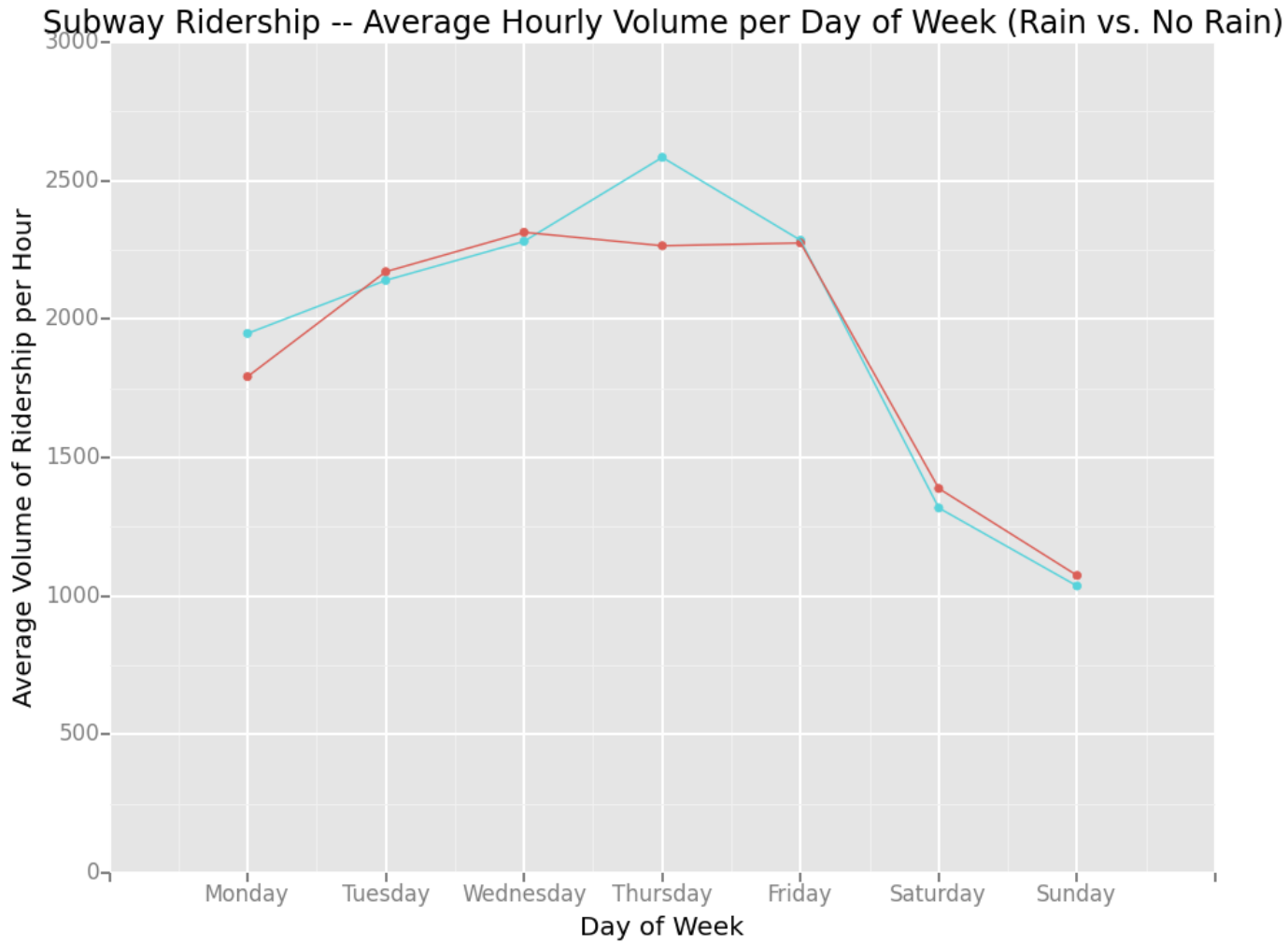
and 52588 occurrences of ridership volume in the same interval when it was not raining. Outlier data (x > 10000) is not shown in the graph to improve readability. Overall in each interval, the ridership volume frequency is greater when it is not raining. This result is not too surprising, as the overall data collection contains approximately twice as many records corresponding to when it is not raining compared to when it is raining. Generally however, the distributions appear to have the same overall (non-normal) shape.

The histogram was generated by parsing the turnstile_data_master_with_weather.csv file on my local machine. Here is the complete python program used to generate the graph:

```
from pandas import *
from ggplot import *

def plot_weather_data(dataframe):
      plot = ggplot(dataframe, aes('ENTRIESn_hourly', fill='rain')) + geom_histogram(binwidth=500) +
xlim(0,10000) + ggtitle('Subway Ridership -- Frequency of Entries Per Hour') + xlab('Volume of Ridership --
Entries per Hour') + ylab('Frequency of Occurrences') +
scale_x_discrete(labels=['0','2000','4000','6000','8000','10000'])

      return plot

if __name__ == "__main__":
      filename = 'turnstile_data_master_with_weather.csv'
      dataframe = pandas.read_csv(filename)
      plot = plot_weather_data(dataframe)

      print plot
```

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 above line plot depicts average volumes of NYC subway ridership, comparing average “entries per hour” data from when it is raining (blue line) to when it is not raining (red line). While overall ridership hourly volume averages are higher when it is raining, the hourly volume averages are lower on rainy weekend days (curiously, within this data collection the hourly volume averages were slightly lower on rainy Tuesdays and Wednesdays as

well, and quite a bit higher on rainy Thursdays). As mentioned in Section 1.4, observing lower ridership hourly volume averages on rainy weekend days is not too surprising, when perhaps people would be more inclined to stay home.

The line plot was generated by parsing the "improved" turnstile_weather_v2.csv file on my local machine. Here is the complete python program used to generate the graph:

```
from pandas import *
from ggplot import *
import pandasql

def plot_weather_data(dataframe):
      dataframe.rename(columns = lambda x: x.replace(' ', '_').lower(), inplace=True)
      q = 'select day_week, rain, avg(entriesn_hourly) as avg_entriesn_hourly from dataframe group by
day_week, rain'
      results = pandasql.sqldf(q.lower(), locals())

      plot = ggplot(results, aes('day_week', 'avg_entriesn_hourly', color='rain')) + geom_point() +
geom_line() + ggtitle('Subway Ridership -- Average Hourly Volume per Day of Week (Rain vs. No Rain)') +
xlab('Day of Week') + ylab('Average Volume of Ridership per Hour') + xlim(-1,7) +
scale_x_discrete(labels=['','Monday','Tuesday','Wednesday','Thursday','Friday','Saturday','Sunday','']) +
ylim(0,3000)

      return plot

if __name__ == "__main__":
      filename = 'turnstile_weather_v2.csv'
      dataframe = pandas.read_csv(filename)
      plot = plot_weather_data(dataframe)

      print plot
```

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?

Overall, NYC subway ridership volumes are higher when it is raining. However, an additional insight was gleaned from studying the "improved" turnstile_weather_v2.csv dataset: hourly volume averages actually are lower on rainy weekend days, but not enough to invalidate the overall conclusion. Intuitively, higher ridership on rainy weekdays makes apparent sense, as commuters still are compelled to go to work. Lower ridership on rainy weekends might be explained by people simply opting to stay home.

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.

Given the two samples of subway ridership volumes (with and without rain), the Mann-Whitney U Test results allowed me to reject the null hypothesis that the two samples came from the same population. In other words, I observed a statistically significant difference between the sample of subway ridership data taken while it was raining, and the sample of ridership data taken while it was not raining. Additionally, the mean of the "with rain" sample was 1105.446, while the mean of the "without rain" sample was 1090.279. Finally, the graph depicted in Section 3.2 above provided some visual evidence of the overall conclusion.

I applied the gradient descent algorithm using feature variables rain, precipi, Hour, and meantempi (and one set of dummy variables providing category values based on each distinct UNIT) to predict hourly ridership volumes. In my final run, my theta coefficient (i.e., weight) for the rain variable was 2.924, for the precipi variable was 14.653, for the Hour variable was 467.709, and for the meantempi variable was -62.218. The model's calculated R^2 value was 0.464. A consideration of the resulting weights shows that knowing whether or not it is raining does help to predict the hourly ridership volumes, but not nearly as much as knowing the hour helps in making the prediction.

Section 5. Reflection

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

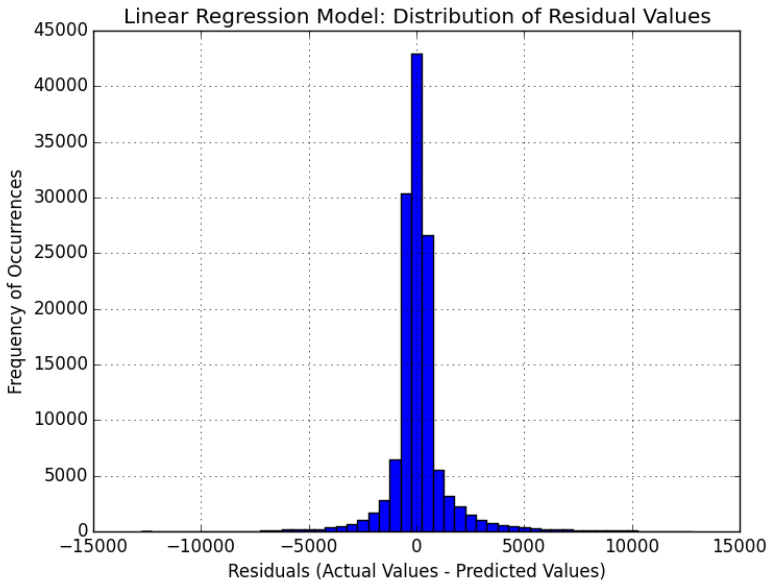
5.1 Please discuss potential shortcomings of the methods of your analysis, including:

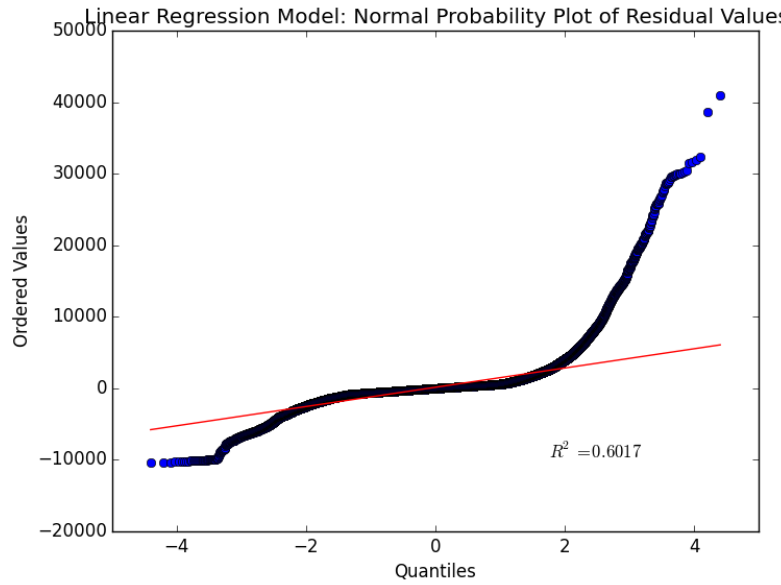
1. Dataset,
2. Analysis, such as the linear regression model or statistical test.

My conclusion from applying the Mann-Whitney U Test was that there existed a statistically significant difference between subway ridership hourly volumes when it was raining compared to hourly volumes when it was not raining. Yet in applying the Gradient Descent linear regression algorithm and comparing the resulting theta coefficients of my chosen feature variables, I concluded that rain was a much weaker predictor of hourly ridership volumes than was the hour of the day. Using the "improved" dataset allowed me to compare average hourly ridership volumes per day of the week. In doing so, I found that overall volumes were higher when it was raining, but that the opposite was true on rainy weekend days. While this weekend result is supported by intuition, I also noted the curious (and unexplained) finding that ridership volumes were slightly lower on rainy Tuesdays and Wednesdays, and quite a bit higher on rainy Thursdays. If pressed to venture a guess about rainy Thursdays, I might posit that the higher ridership volumes could be due to workers going out together at the end of the workday and being more inclined to stay dry by taking the subway. It would be interesting to see if these secondary findings hold up when evaluating a dataset that encompassed a longer timeframe.

With all of my considered combinations of feature variables, my best calculated R^2 value was around 0.467. Most of the "bang for the buck" here was made possible by including the UNIT category values as a dummy feature variable. I found these results to be somewhat disappointing, and would have liked to have achieved a calculated R^2 value closer to 1, or simply greater than 0.5. One possible shortcoming of the dataset might be the existence of cross-correlation among some of the variables. For example, it may not add value to consider both rain and barometric pressure at the same time as predictors, as barometric pressure readings are usually low when it is raining. One factor that might improve correlations would be to consider historical data. I would hypothesize that the best predictor of future ridership for a given hour at a given station would be past ridership during the same hour (and same day of the week) at the same location.

To examine the shortcomings of the linear regression model, let's consider the frequency distribution of the residuals (i.e., the error values derived by subtracting the predicted values from the actual values), as well as the normal probability plot of the residuals. The two graphs are shown below:





The frequency distribution of the residual values does appear to be reasonably normal about zero, but with some degree of overestimation in some of the predicted values. One example of this overestimation can be seen by comparing the bars immediately to the left and right of the center bar. Here, the frequency of residuals in the range (-750, -250) is a bit over 30000, while the frequency of residuals in the range (250, 750) is approximately 27000. The normal probability plot of the residual values produced a line of best fit having an R^2 value of 0.6017. If the residuals were distributed more normally, I would expect this line of best fit to have an R^2 value closer to 1. A more normal distribution of the residuals would imply greater suitability of the model in predicting subway ridership.

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