The answers to the short questions follow. This document refers to the attached Ipython notebook. Please see the notebook for further detail.

Section 1. Statistical Test

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?

   I used the Mann-Whitney U Test with a one-tail P value.

2. Why is this statistical test applicable to the dataset?

   The data set had a non-normal distribution and the design criteria for the Mann-Whitney U test were reasonably met. Please see notebook for further detail regarding design criteria.

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(Rain): 1105.44637675

   Mean(No Rain) 1090.27878015

   U: 1924409167.0

   p: 0.0193096344138

On average ~ 15 more people/hour ride the subway on a rainy day.

4. What is the significance and interpretation of these results?

   I found there is a difference in the means of ridership as a function of rain (rain vs. no rain).

Section 2. Linear Regression

1. What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

   Gradient descent (as implemented in exercise 3.5)

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

   Is_weekday (weekday or weekend), rain/precipitation (tried both, either one gives the same result in my model), hour, temperature (tried min, max, mean etc. none of them had high predictive value), and Unit as a dummy variable (because Unit is categorical).

3. Why did you select these features in your model? We are looking for specific reasons that lead you to believe that.

My data exploration (please see notebook) suggested that some of these features would be predictive of ridership (Unit, hour, weekend vs. weekday, rain vs. no rain). I tried out the temperature features out of sheer curiosity. I applied each feature from above individually to the model to see its effect on R^2, keeping those that improved the model (raised R^2). I then grouped the winners into my final implementation.

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

   .468

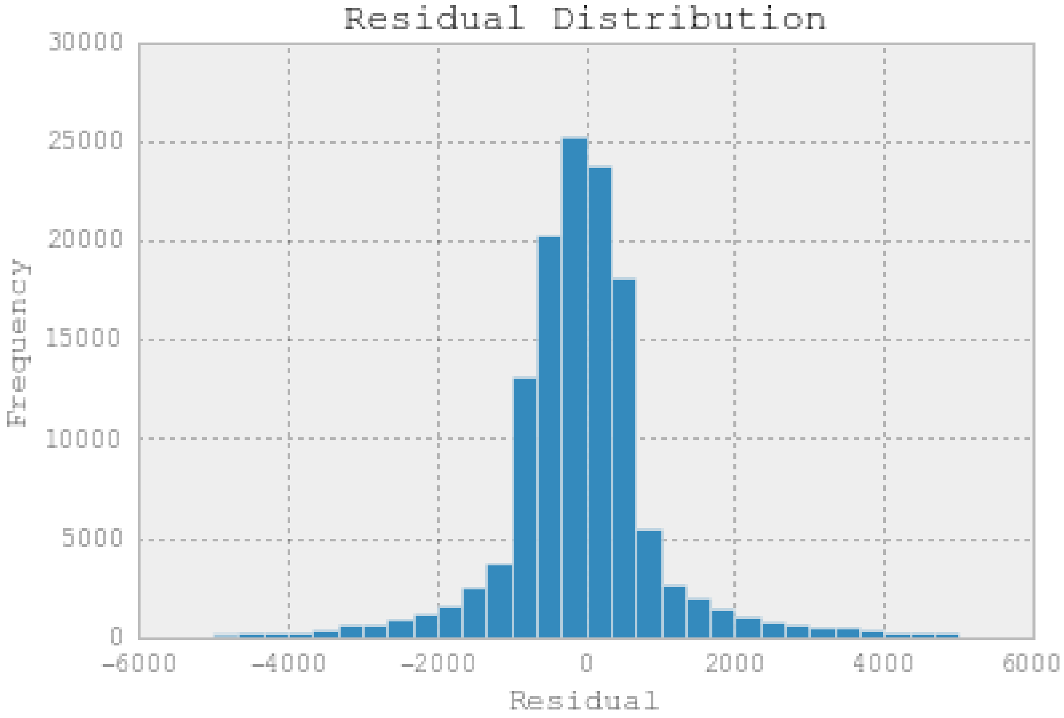
5. What does this R^2 value mean for the goodness of fit for your regression model?

   It means that the model has accounted for roughly 47% of the variability of the data.

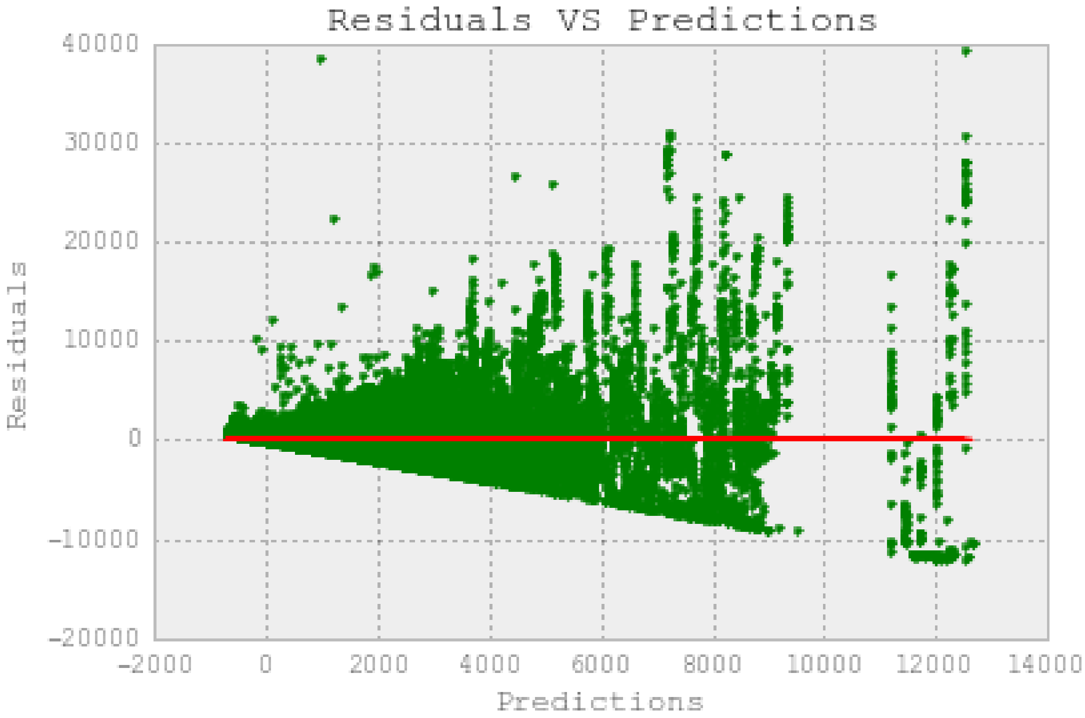
Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

The R^2 values is low, < 50%. Much of the fit was gained by incorporating the rain feature. So, although we have a low R^2 value, the presence of a statistically significant predictor (determined per above) may render our model useful. However, there is still 53% variability we haven't captured. Even more telling in terms of how appropriate our model is, are the residuals.

First their distribution:

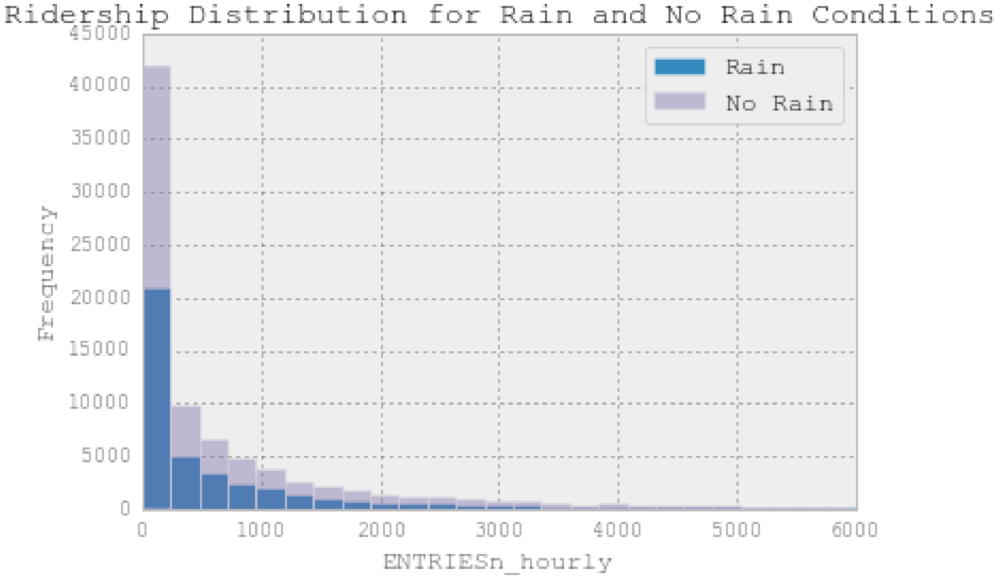


This distribution appears to be close to normal, albeit more values seem to be bunched on the positive side. According to my reading histograms are not the best way to assess the appropriateness of your model. So, I plotted residuals vs. predictions and fit a trend line. It is clear from this visualization that rather than being randomly distributed, the residuals increase as a function of the predicted value. Clearly we have structure not accounted for by our model.

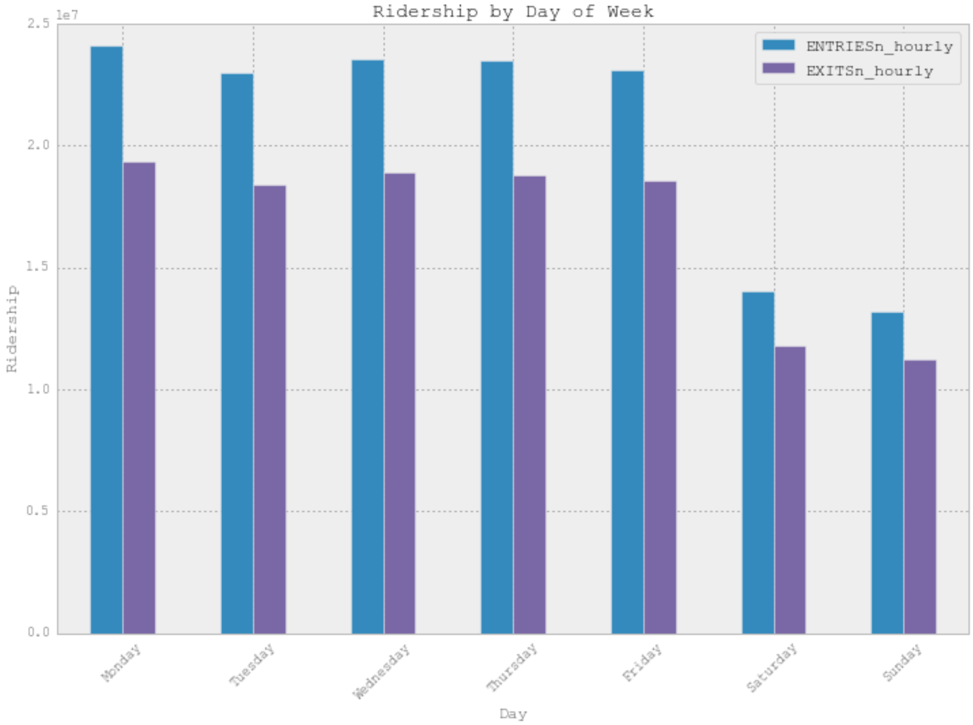


Section 3. Visualization

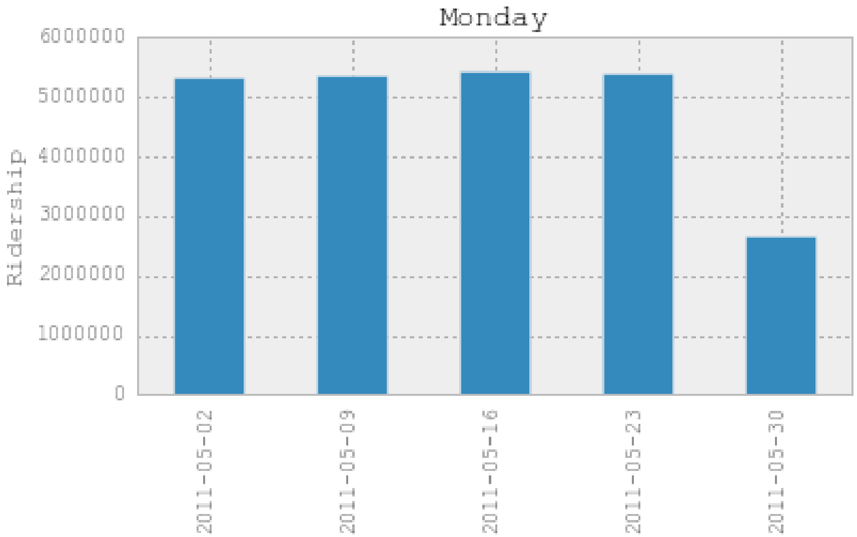
Please see notebook for extended data visualization. Please also note that after updating Anaconda python, ggplot would no longer display the legend.



This histogram shows the distribution of ridership, represented by ENTRIESn_hourly for the NYC subway system in the Month of May 2011. The key insight is that these distributions are non-normal, with similar shape. This informs decisions on statistical model used to interpret the data.I show several visualizations of the data in the attached notebook pdf. One of my favorites was also the simplest, ridership by day of the week, as represented by hourly entries and exits.

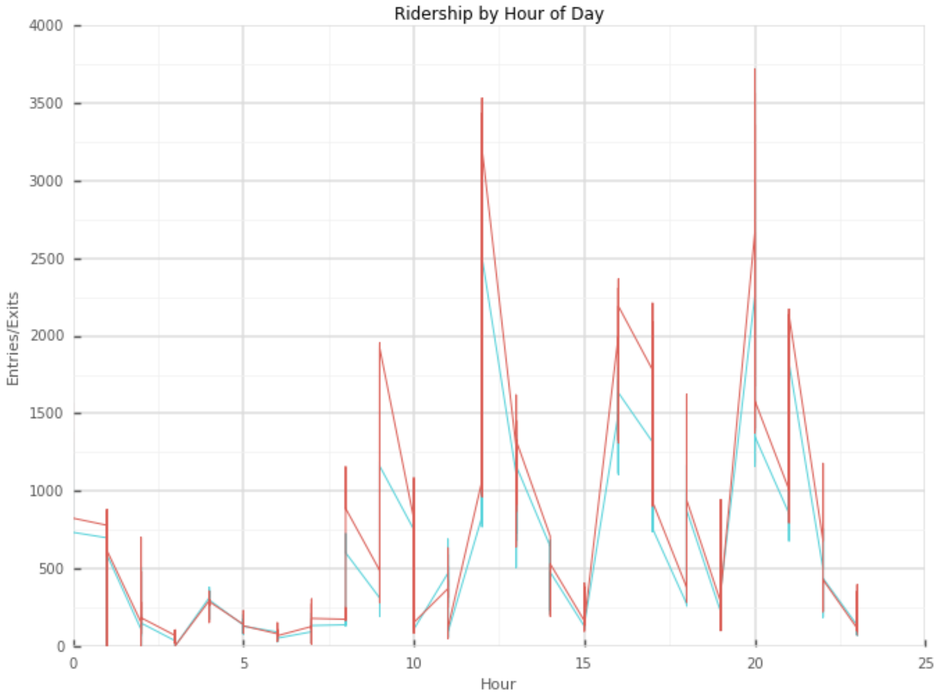


The plot above shows hourly entries and exits summarized by day of the week. This is a simple visualization and the results are intuitive (ridership is down on weekends). It does show a fairly consistent level of ridership on each weekday and on each weekend day. Knowing this, I was able to create a feature for linear regression, splitting the data into two groups, is weekday/is not weekday. Here is what I found most interesting. Often you lose insight when summarizing data. Looking at the data in finder granularity (below) you will see that not all Mondays have equal ridership. What is going on? Memorial Day is going on.

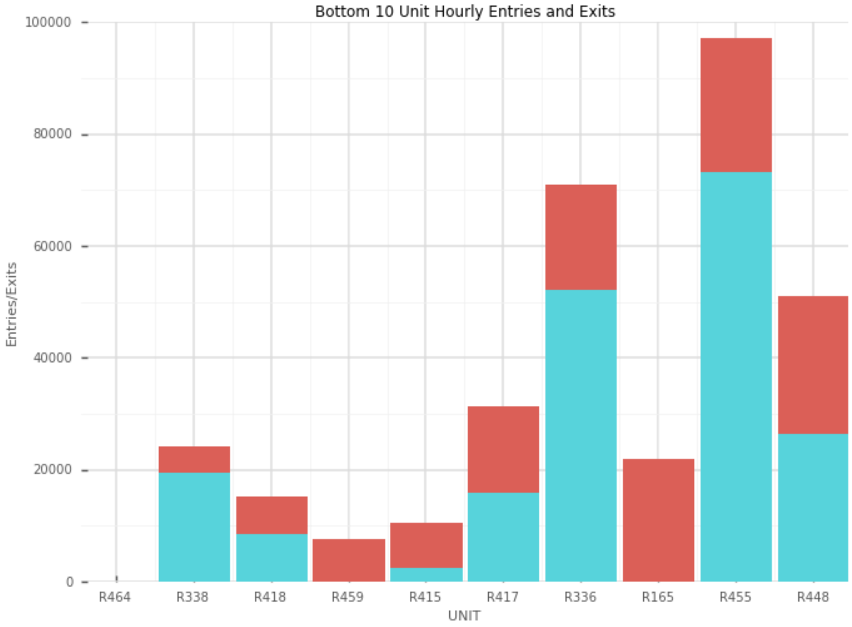


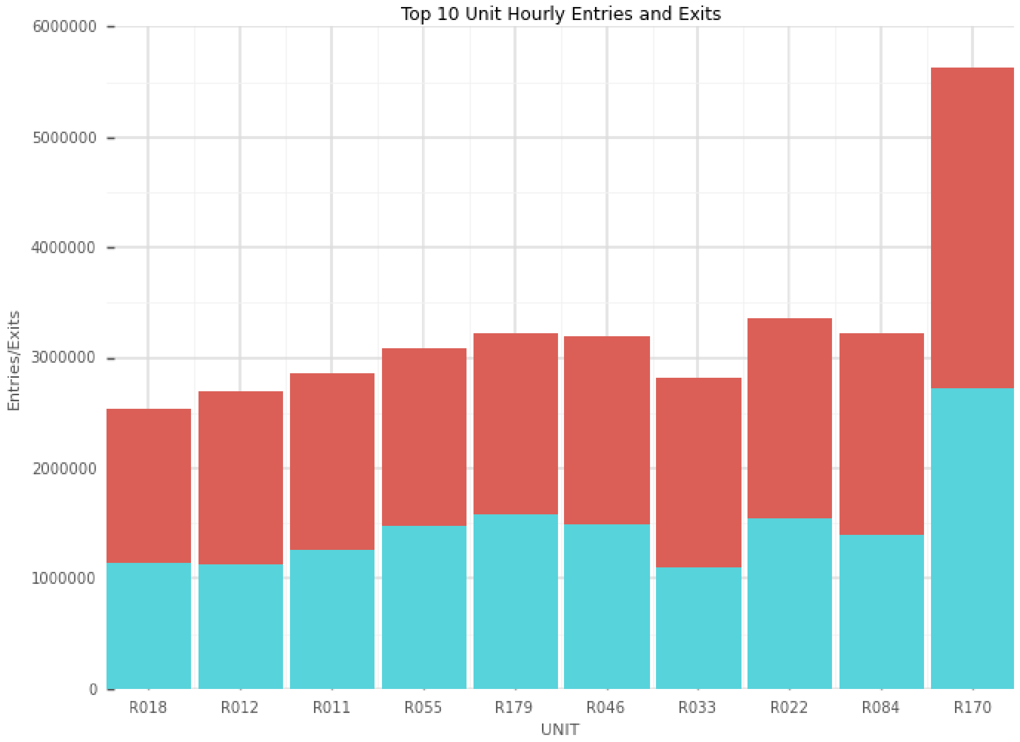
Other Visualizations:

The following visualizations also gave me insight into the data. Please see the attached notebook or exercise solutions for further detail.



I created a datetime index for the dataframe and resampled. Visualizing this with ggplot gave me the insight that ridership varies by hour of the day.





I wanted to know the top and bottom 10 used units. I sliced up the dataframe to get that information and plotted it using ggplot. This was more for fun and to play with PANDAS and ggplot. I was interested in knowing if the proportion of exits at top and bottom N units were proportional to each other for each unit. In the case of the top 10 units, they look reasonable proportional, but less so in the case of the bottom 10.

Section 4. Conclusion

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

1. From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining versus when it is not raining?

For the month of May, 2011, more people rode the subway when it was raining than when it was not raining. Specifically ~15 more people rode the subway per hour when it was raining than when it was not raining.

2. What analyses lead you to this conclusion?

The main analyses that informed my conclusion was the Mann-Whitney U test. The low p-value led me to reject the null hypothesis, that there was no difference in the mean ridership between the rain and no-rain conditions. The mean values provided an estimate as to the magnitude of this difference

(~15/hour). The fact that including rain in the gradient descent model led to a large increase in the value of R^2 further bolstered my confidence in this conclusion.

Section 5. Reflection

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

1. Please discuss potential shortcomings of the data set and the methods of your analysis.

The major shortcoming of my analysis is in the data. In this analysis we are trying to determine how certain features affect ridership. In terms of hour, days, days of the week, we have a decent sample size (1 months' worth), but a larger sample would be better. In terms of tying changes in ridership to features of weather, our data set is woefully inadequate. We only have data from one month at one time of the year (Spring). This means that we don't have a sample of ridership over large fluctuations in weather conditions (temperature, pressure, etc.), it is pretty even. Even for rain vs. not rainy days, we have many more rainy days than non-rainy days.

The methods of analysis are fairly sound, albeit they are limited by the data. The gradient descent model clearly needs some work to produce a better fit to the data and increase predictive value.

Despite the above, I think the current analysis and data set are a great starting point to test computational methods. An improved analysis would incorporate the following:

- A larger and more comprehensive data set:
  - More months and years: The current methods should scale well to additional months years. Additional months would give us a better sample of features, as mentioned above.
  - Extended Features: Additional features which may help predict ridership could be included. One suggestion might be to include sunset and sunrise time (when it gets dark and light) and location of unit or some calculation derived from the location (distance from large population center, etc.) Location is in the extended data set provided for the optional OLS exercise.
  - Improved regression analysis. My regression model is clearly not ideal and is not accounting for all of variability in the data. Incorporating additional features and/or a non-linear model may improve its predictive value.

References:

- GGPlot (http://ggplot.yhathq.com/docs/index.html)
- http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
- https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php
- http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

- Class Page on Piazza
- Python for Data Analysis by Wes McKinney
- Stack Overflow