The answers to the short questions follow. This document refers to the attached Ipython notebook.

Section 1. Statistical Test

1. Statistical Test Used: Mann-Whitney U Test converted to a two tail p-value.
   Null Hypothesis: There is no difference between the means of ridership on rainy vs. non-rain days.
   Desired Statistical Level: 95%

2. Type of test and p-value: The data set had a non-normal distribution and design criteria for the Mann-Whitney U test were reasonably met. A one-sided t-test assumes the direction of the change, prior to data acquisition, a two tailed test does not. I multiplied the p-value by 2, to obtain the two sided p-value.

3. Results of Test: Mean(Rain): 1105.44637675, Mean(No Rain) 1090.27878015, U: 1924409167.0
   One-tailed p: 0.025 (Udacity IDE value, differs from ipython notebook value, ticket open in scipy)
   Two tailed p = .05 ,100 – two tailed p = 95%

Within a 95% confidence interval, an average ~ 15 more people/unit ride the subway on a rainy day.

4. Significance of Interpretation: Mean ridership differs as a function of rain.

Section 2. Linear Regression

1. Approach to compute the coefficients theta and produce prediction for ENTRIESn_hourly in regression model: Gradient descent

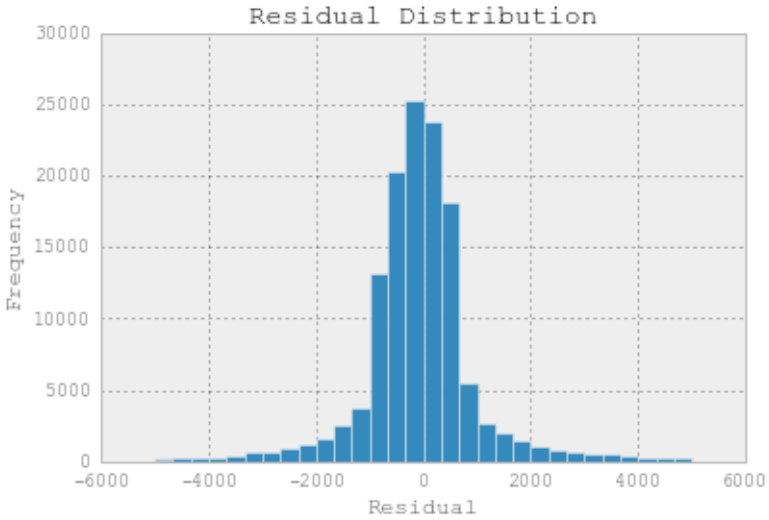
2. Model Features: ['rain','Hour','isWeekday','maxtempi']] and Unit as a dummy variable

3. Rationale for feature selection: Data exploration revealed these features to be predictive of ridership. Each feature was applied individually to determine its effect on fit. I grouped those that improved fit (increased R^2) into my final implementation.

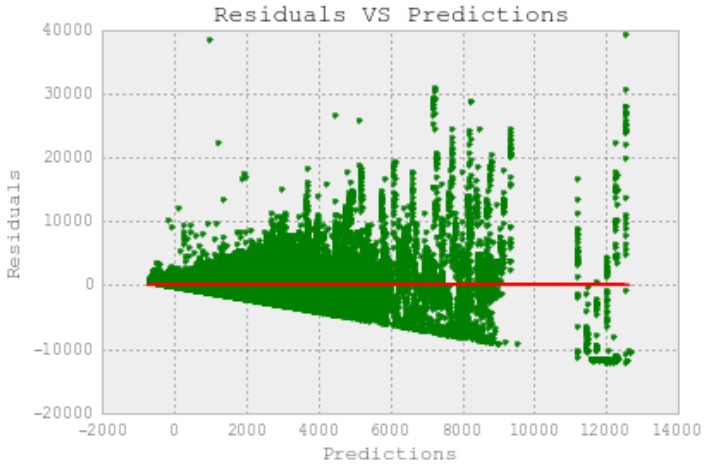
4. Model R^2: .468

5. Meaning of resultant R^2: The model accounts for roughly 47% of the variability of the data.

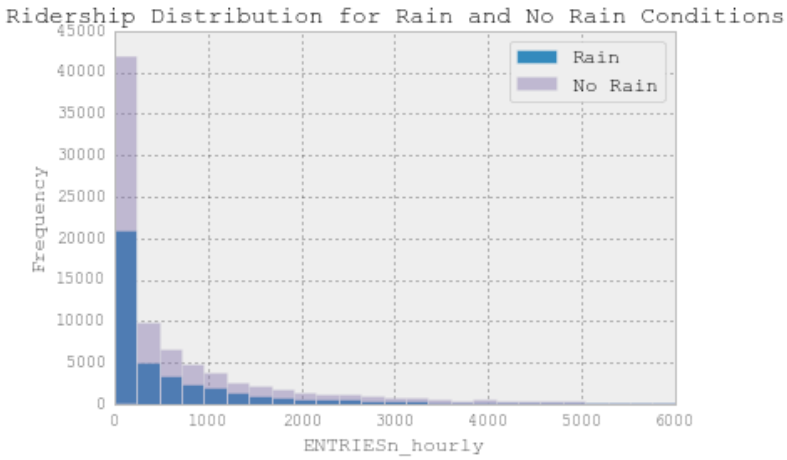
6. Predictive value of the model: The R^2 values is low, < 50%. The rain feature accounts for most of the fit. Despite the low R^2 value, the presence of a statistically significant predictor (rain) adds value to the model. However, we haven't captured 53% of the variability. The residuals tell a more detailed story.



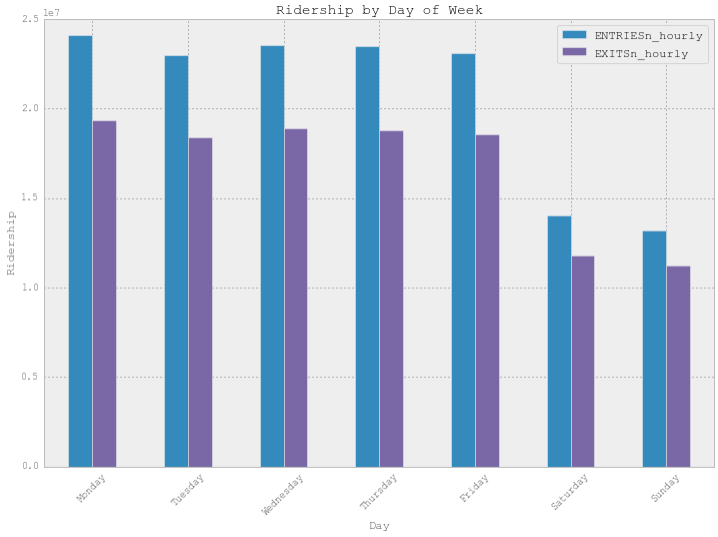
This distribution is ~ normal, histograms have limitations. A plot residuals vs. predictions, fit to a trend line, reveals that residuals increase as a function of predicted value. Clearly structure exists that is not accounted for by our model.



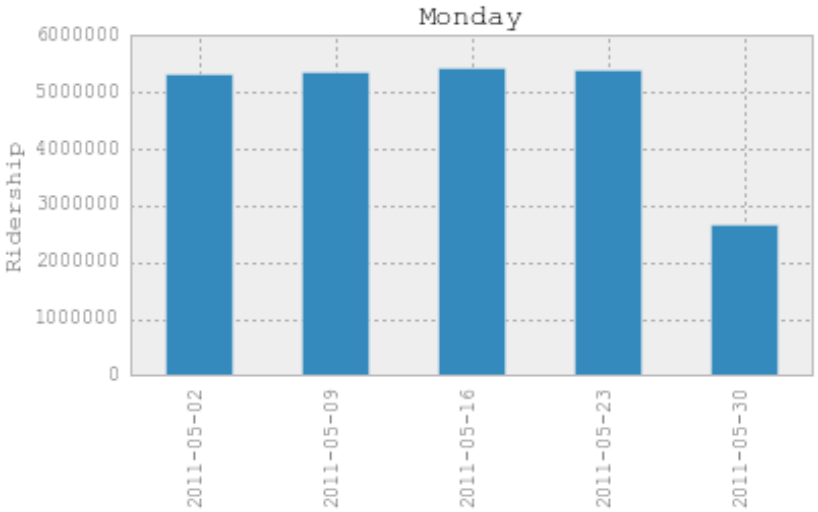
Section 3. Visualization



Plot: Distribution of ridership for the NYC subway system in the Month of May 2011. The
Key insight: Distributions are non-normal, with similar shape. This informs decisions on statistical model used to interpret the data.

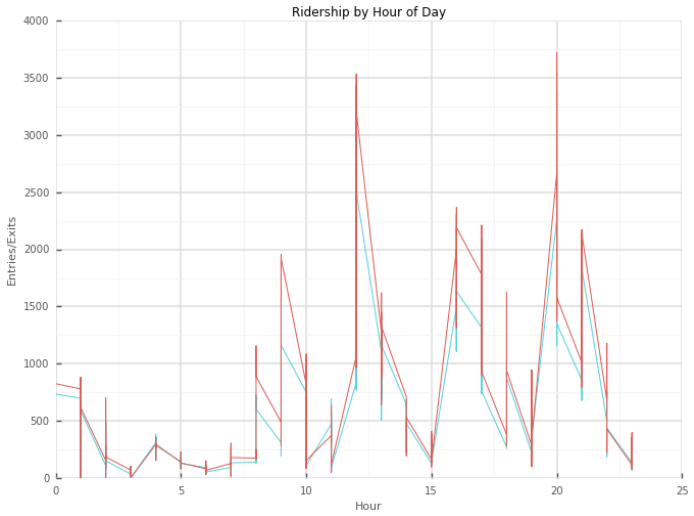


Plot: Entries and exits summarized by day of the week.
Key Insights: Ridership is down on weekends.

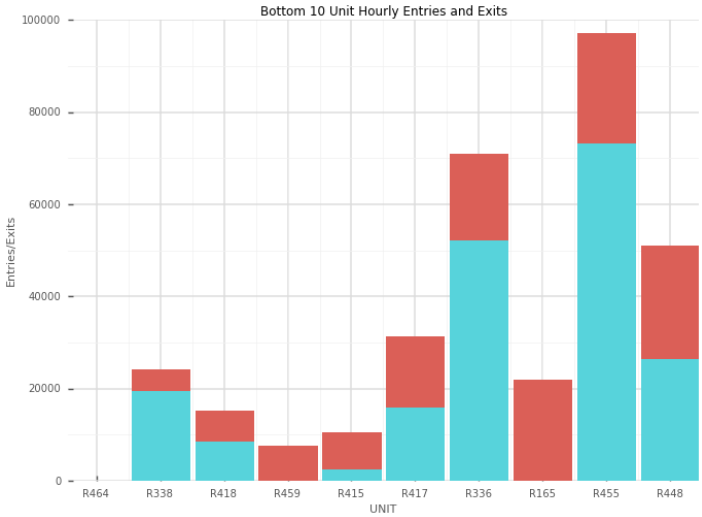


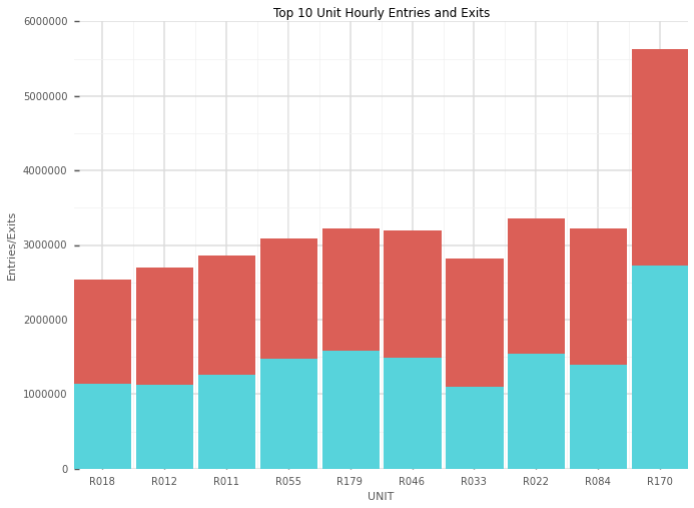
Plot: Ridership on Monday for May 2011
Key Insight: Ridership is down on Memorial Day

Other Visualizations:



Plot: Ridership by hour, via resampling.
Key Insight: Ridership varies by hour of the day.





Plot(s): Top and bottom 10 units.
Key Insight: Top units entries and exists appear proportional, bottom, less so.

Section 4. Conclusion

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

1. Do more people ride the NYC subway when it is raining versus when it is not raining?

Specific to the month of May, 2011, more people rode the subway when it was raining than when it was not raining at ~ 15 people per unit.

2. Basis of conclusion: The Mann-Whitney U test was the primary informer. The low p-value led to rejection of the null hypothesis. Mean values provided an estimate of the magnitude of this difference (~15/unit). 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. Shortcomings:

Data: In terms of hour, days, days of the week, we have a decent sample size, but a larger sample would be better. In terms of tying changes in ridership to features of weather, our data set, one month, is woefully inadequate. We lack a sample of ridership over large fluctuations in weather conditions.

Methods: The gradient descent model requires refinement to increase predictive value.
The current analysis and data set are great starting points to test computational methods. 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.
    - Extended Features: Additional features may help predict ridership. These may include sunset and sunrise time or location of unit.
    - Improved regression analysis. A non-linear model may improve predictive value.

References: See notebook