Section 0. References

Scipy. “scipy.stats.pearsonr”.
http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.pearsonr.html

Statwing. “Statwing t tests”. http://help.statwing.com/knowledge_base/topics/statwing-t-tests

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?

A Mann-Whitney U-Test was used to compare ridership on rainy vs. non-rainy days.

A two-tail P value was used. The null hypothesis is that there is no significant difference between the mean ridership on rainy days and the mean ridership on non-rainy days. The two-tailed P value is .05, and is significant at a p-critical value of .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.

Unlike other t-tests, the Mann-Whitney U-Test does not require the distribution to be normal. “ENTRIESn_hourly” is the response variable, and since it is positively skewed and not normally distributed, a t-test such as Welch’s t-test that assumes a normal distribution cannot be used (Statwing).

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 group means were 1105 for rainy day hourly ridership and 1090 for non-rainy day hourly ridership. The test results were a U statistic of 1924409167 and a p value of .05.

1.4 What is the significance and interpretation of these results?

There is a statistically significant difference in mean hourly ridership on rainy vs. non-rainy days at an alpha level of .05. People are more likely to ride the subway on rainy days. That could be because people avoid walking on rainy days. More context would be needed to determine the real world significance of an extra 15 people riding the subway at each station each hour, such as the total number of subway stations in NYC and the mean price of a ticket.

To provide further context for these results, ideal future analyses would examine other transportation patterns on rainy vs. non-rainy days (cars, buses), as well the transportation patterns of subway systems in other cities on rainy vs. non-rainy days.

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:

1. Gradient descent (as implemented in exercise 3.5)

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:

- 'UNIT', which is a proxy for subway station, was made into dummy variables.
- 'Traffic_bucket', which bucketed 'Hour' into 3 categories based on mean ridership during that hour: high traffic, mid traffic, and low traffic.
- 'Hour'.

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.

UNIT was made into dummy variables, because those dummy variables will capture differences between stations that are not captured by the other features, such as baseline ridership. The UNIT dummy variables explained 40% of ridership variation.

The local dataset was divided into a training and holdout set, and R^2 were computed for various models.

Including Hour as a feature increased R^2 from 40.3% to 43.8%, and then including Traffic_bucket increased R^2 from 43.8% to 45.5%. Hour was first selected as a feature because the Figure 2 chart of ridership by hour showed peaks in ridership at certain hours. Traffic_bucket was engineered as a feature in order to more heavily weigh whether or not an hour was a peak hour or not, because considering whether or not an hour is a peak hour could be more important than the hour value itself when predicting ridership.

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

Traffic_bucket - 639
Hour - 130

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

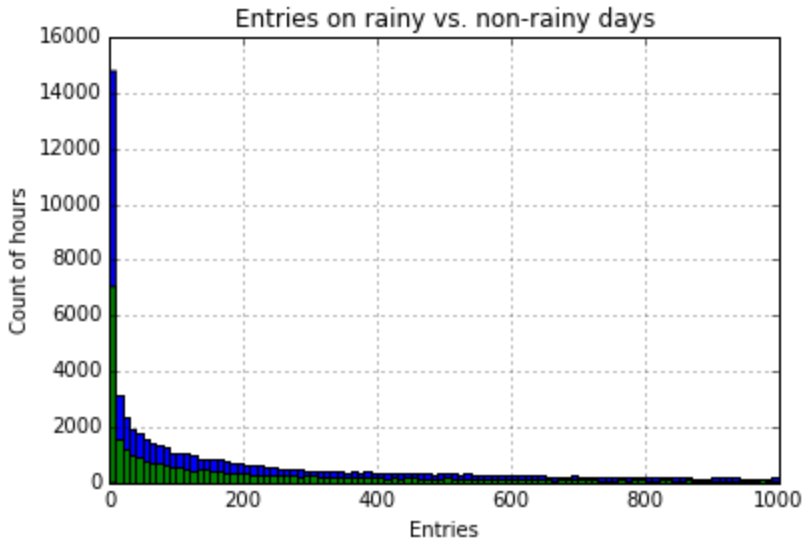
46.3% (48.7% when submitted on Udacity and tested against the Udacity test dataset).

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 does not fit the data well, because it is only able to explain 48.7% of its variation. Therefore, the model is not appropriate for this dataset.

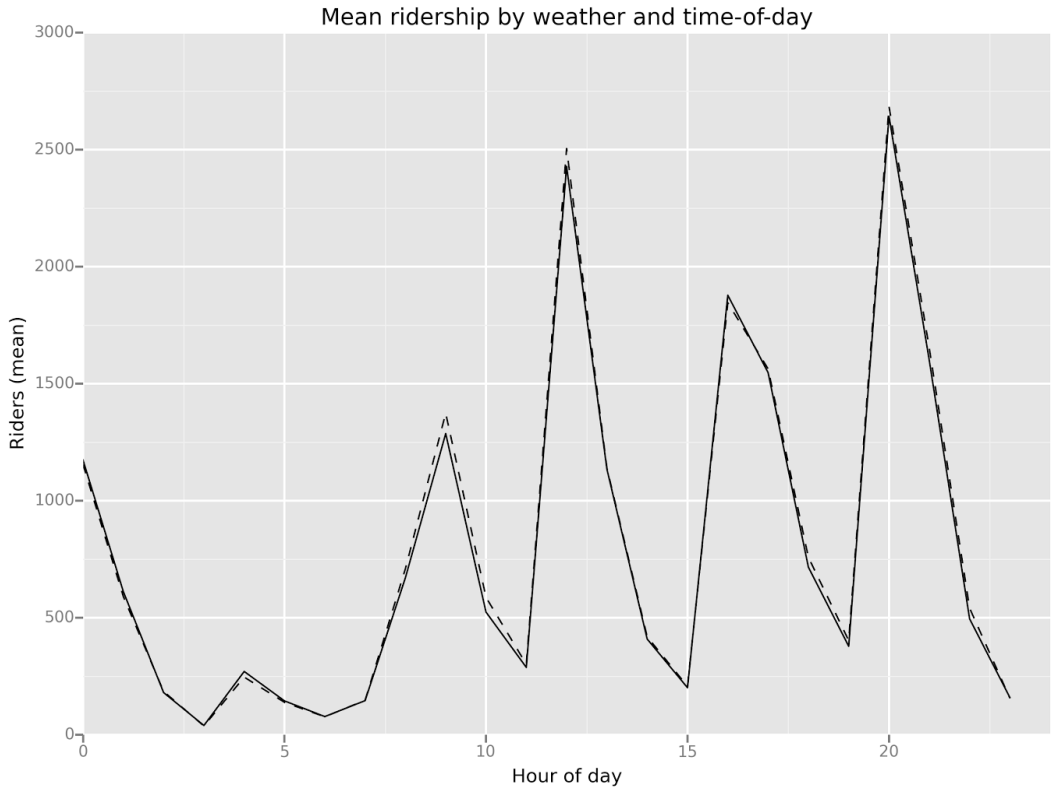
Section 3. Visualization

Figure 1. Distribution of ridership on rainy days (green) and non-rainy days (blue):



Ridership on rainy and non-rainy days have a similar positively skewed distribution .

Figure 2. Ridership by rain (dotted for rain, solid for non-rainy) and time-of-day



There are peak hours throughout the day (12, 16, 20) during the vast majority of ridership occurs. There are also hours (9, 1, 17, 21) during which a moderate amount of ridership occurs. This graph provides important context for the difference in ridership on rainy and non-rainy days; though there are more riders on rainy days vs. non-rainy days, the hourly variation in ridership does not change.

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?

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.

More people are likely to ride the subway when it's raining. A Mann-Whitney-U test showed a statistically significant difference between rainy day hourly ridership and non-rainy day hourly ridership of 15 people per hour per station, with a U statistic of 1924409167 and a p value of .05. Most likely, people ride the subway more often when it's raining to avoid walking or biking outside in the rain. This difference may have real world applications. Further analyses would examine differences in rainy day ridership by station. If certain stations have increases in ridership of 100 people or more, then perhaps those stations would benefit from extra trains for rainy days in order to avoid delays.

However, it's unlikely that ridership and weather have a linear relationship, or that weather causes increased ridership. The linear model including weather variables has poor predictive power (R^2 of < 50%). There could be another variable that is causing the increase in ridership that correlates with rain.

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.**

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

The dataset is a sample of total subway data, so it's possible that the dataset is not representative of the underlying population.

Another type of linear regression such as Ordinary Least Squares, as well as non-parametric analyses, could better capture the relationship between weather and ridership. I tried most of the features in the dataset, but none with an effect > 0.5% were weather-related. Judging by the results of the t-test and also by intuition, weather should have some influence on subway ridership.

There is no variation in mean hourly ridership based on rain, which provides more evidence that rain is not causing increased ridership, but that rain is correlated with another variable that is causing increased ridership.

It was unclear to the author of this project how to correlate the variables in this dataset, given that many of the variables have non-normal distributions, and calculating the Pearson correlation coefficient requires that both variables are normally distributed (Scipy). So that information is missing from the analysis.