Section 0. Citations

I mostly just used the Udacity exercises that I had done and the videos to help me with this project. However, the R documentation and the ggplot2 library for R were very useful to create my visualizations.

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? What is the null hypothesis? What is your p-critical value?

I used a two-tailed Mann-Whitney U test. I am comparing the hourly subway entries on days when it rains to days when it doesn't rain. My null hypothesis is that whether it rains or not does not affect the hourly entries into the subway. My p-critical value is 0.05.

1. 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 applicable to this dataset because it is a non-parametric test. Looking at the distributions of the dataset shows that the distribution is not normal, but instead heavily skewed towards lower values. Welch's T-test is a parametric test, assuming that the distributions of the samples are normal. Therefore, the Mann-Whitney U test is more applicable in this situation than Welch's T-test.

1. 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.

I got a p-value of about 0.02499. The mean for days when it was raining was 1105.45, while the mean for days when it wasn't raining was 1090.29

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

The p-value of 0.249 is given as a one-tailed value. Doubling this, I get the two-tailed value of 0.4998, just under my p-value of 0.05. This allows me to reject the null hypothesis and decide that whether it is raining significantly affects the number of hourly entries into the subway.

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?

I used OLS using Statsmodels.

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

I used the hour and whether the day of the week is a weekday. I used UNIT as a dummy variable.

1. 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.

I used hour because I thought that the time of day would probably influence how many people got onto the subway. For example, there would probably be more people on the subway in the morning than in the middle of the night. When I included hour in my model, my R2 value increased by 0.084 from if I just used weekday.

I used weekday because I knew from the statistical test that a lot more people use the subway on weekdays then weekends. When I included weekday in my model, my R2 value increased by 0.0134 from if I just used hour. This isn't as much as I expected, but it was enough to keep it in my model.

I tested a lot of other features, but none of them resulted in my R2 value increasing by more than 0.01. I decided not to include these features to keep my model simple.

I used UNIT as a dummy variable because I thought that it wouldn't work to directly compare different units with each other. This is because the units are at different stations, and also units that are at more crowded entries into the subway would have more entries than other units even within the same station. Including this in my model increased my R2 value by 0.3776 over using only hour and weekday.

1. What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

My parameters were 123.39 for hour, 981.37 for weekday, and a parameter for each UNIT dummy variable which I will not list here. The intercept was -99.12.

1. What is your model’s R2 (coefficients of determination) value?

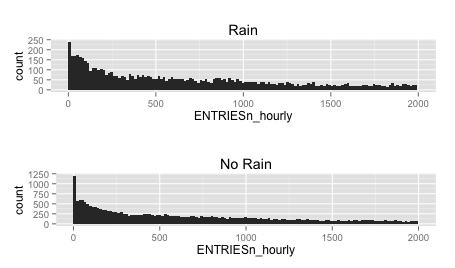
My model's R2 value is 0.4814.

1. What does this R2 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 R2 value?

I think that this model is not that appropriate to predict subway ridership. The R2 value is under 0.5, meaning over half of the variation was not due to any of the features that I tested. I tested almost all of the features in the spreadsheet. Therefore most of the cause for the variation is probably from unknown features not included in the dataset. My model provides only a basic fit for ridership and probably shouldn't be used to make solid predictions.

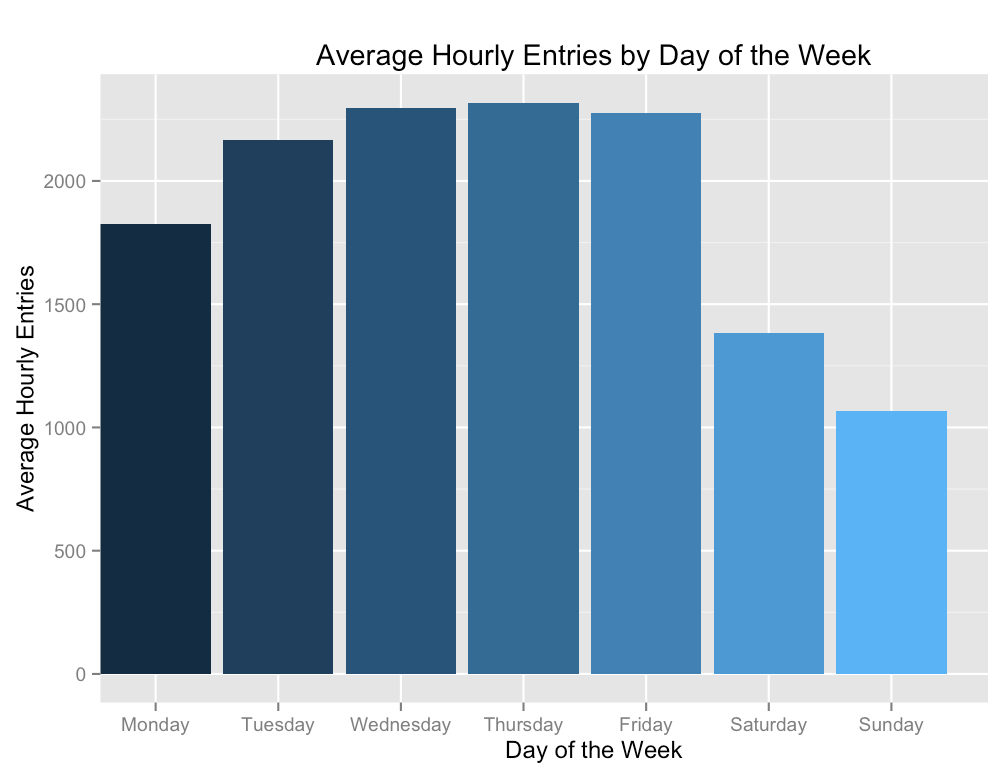
Section 3. Visualization

1. One visualization should contain two histograms: one of ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days.



These histograms show how skewed the data is and how it does not follow a normal distribution. It also shows that relative to the size of the data, rainy days seem to have more higher values, which hints at a higher overall subway usage.

1. 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.



The above chart is insightful because it breaks down the average hourly entries by day of the week. The chart clearly shows that weekdays have more hourly entries than weekend days by a long shot. An interesting observation is the low (compared with other weekdays) hourly entries on Monday. This could be something to be investigated further.

Section 4. Conclusion

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?

From my analysis of the data, more people ride the NYC subway when it is raining.

1. What analyses lead you to this conclusion?

The Mann-Whitney U test that I did in problem set three lead me to this conclusion. I got a p-value of slighly under 0.025 in my tests, and so the two-tailed p-value is slightly under my p-critical value of 0.05. Therefore, I am able to reject the null hypothesis and conclude that whether the day is a weekday or a weekend day does affect mean hourly entries into the subway.

Section 5. Reflection

1. Please discuss potential shortcomings of the methods of your analysis.

The dataset had some shortcomings-

* 1. There were only five 'hours' in the day: 4, 8, 12, 16, and 20. In reality there should be 24 hours.
  2. The dataset was collected in the span of one month. Subway ridership could be considerably different at different times of the year.

My analysis also had one major shortcoming-

My r2 value was under 0.5, meaning that I could not find the source of most of the variation of subway entries. I am not sure if this is because I didn't test enough or because there weren't enough features in the dataset to explain all of the variation.

1. Do you have any other insight about the dataset that you would like to share with us?

Despite all of the weather features in the dataset, weather did not seem to make much of a difference in subway ridership according to my linear regression. The biggest source of variation in my dataset were the individual turnstile units, accounting for 37.8% of variation. The next biggest were hour and weekday, with 8.4% and 1.3%, respectively. No other variables had a greater than 1% impact on variability. So although the weather (especially rain) does clearly affect subway ridership, it is not a good way to predict it.