References

https://en.wikipedia.org/wiki/Linear_regression#Introduction_to_linear_regression
https://en.wikipedia.org/wiki/Mann–Whitney_U_test
https://en.wikipedia.org/wiki/Gradient_descent
https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-to-interpret-the-constant-y-intercept
https://en.wikipedia.org/wiki/Multicollinearity#Definition
http://www.statsoft.com/Textbook/Multiple-Regression#cresidual

Section 1

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?

Since we have two samples which are not normally distributed and come from the same population, we used the Mann-Whitney U-Test with a two tailed p-critical value of 0.05.

Ho = The values in the two samples (raining and not raining) are not significantly different (ie. Greater or lesser)
Ha = The values of one sample are significantly different (ie. Greater or lesser) than the values in the other sample

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.

Since we have two samples which are not normally distributed and come from the same population, we can use the Mann-Whitney U-Test.

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.

Using data returned, we can get a P value of approximately 0.0499. The mean number of entries for sample with rain is approximately 1105.44. The mean number of entries for sample without rain is approximately 1090.28.

1.4 What is the significance and interpretation of these results?

We get a P value of approximately 0.0499. Since this is less than our p-critical value of 0.05, we reject the null hypothesis.

Section 2

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

I used the gradient descent algorithm from lesson #3.

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

I used hour and precipitation as features. Unit (station) was used as a dummy variable.

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.

All the features I tested were based on intuition. When it is very foggy outside, people may be more inclined to use the subway for safety reasons. In the case of rain, people are probably more likely to use their cars in order to avoid walking in the rain. Other weather factors such as mean wind speed and mean temperature will also likely have an effect on ridership. Finally, the specific station (some are in busier locations than other) will also likely have a huge effect on the variation of ridership.

Ultimately, after some exploration, I removed fog, mean wind speed, rain and mean temperature because their effect on ridership was not significant. In general, the simpler the model the better. I decided to only use hour and precipitation as features.

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

2.75247864e+01 (Precipitation)
4.67686750e+02 (Hour)

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

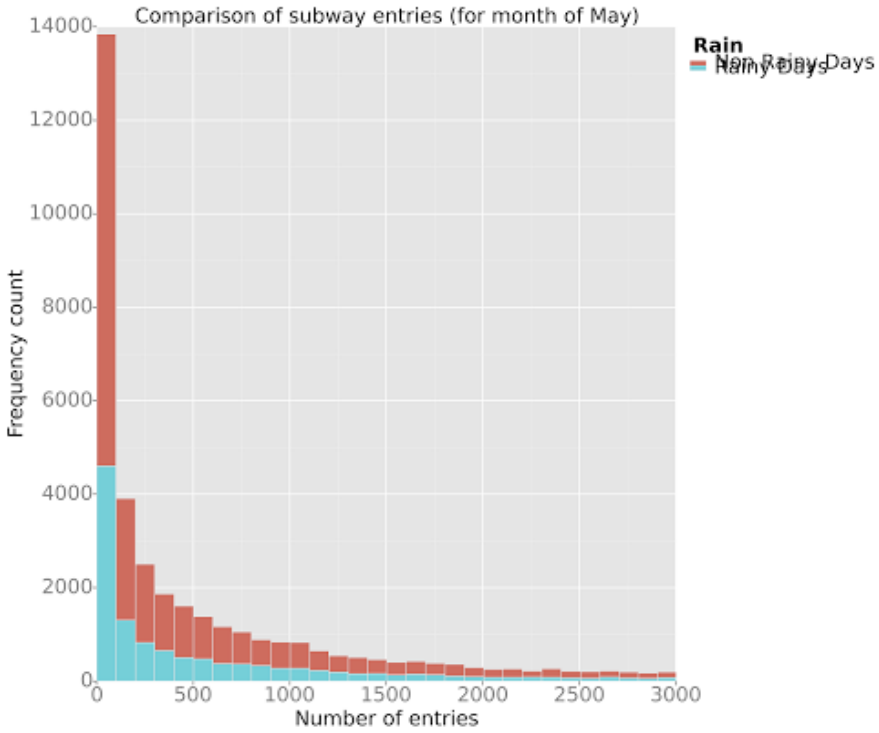
Your r^2 value is: 0.463307661665. Good job! Can you make it even better?

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?

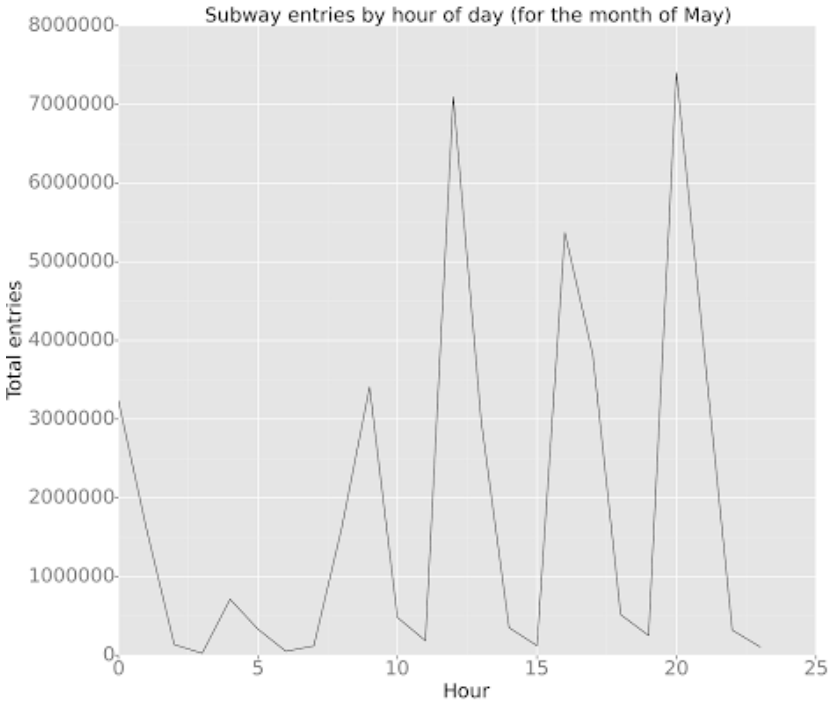
R^2 is the proportion of the variance in the entries per hour that is predictable from the independent variables. The higher R^2, the higher percentage of points our line of best fit passes through. As such, we can say that our features have a positive correlation with ridership and can be a reasonable predictor of ridership.

In addition, our R^2, tells us that our model can explain approximately 46% of the original variability. This means we are left with 54% residual variability. Given this, our model is still a reasonable predictor of ridership.

Section 3



For majority of hours in a day, the number of entries for subway stations is between 0 and 500.



From the line graph, we can see ridership reaches a global maximum around 8pm at night. In addition, ridership rises and falls in a relatively cyclical pattern. There are sharp peaks approximately every 5 hours before dropping drastically. After 10am, non-peak times are all relatively the same level. The same can be said for peak times.

Section 4

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?

From our model, we see that rain does not have a significant effect on ridership. In fact, the most important variables to consider are the time of day and specific station.

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.

From the Matt-Whitney U test, we know we can confidently reject the fact that the difference in the mean entries on rainy and non-rainy days is due to random sampling.

From the coefficients of our model, we can see that precipitation level has an effect of approximately 2.75247864e+01. The effect of hour on our model is even more pronounced with a coefficient of 4.67686750e+02.

Section 5

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

In addition, since I used gradient descent, our cost function might have multiple local minima. Our algorithm might be trapped in a local minimum that isn't actually the global minimum. To adjust for this, we should run our algorithm on random data each time.

Using regression, there is always a chance that you are over fitting the line since generally the more features you add the higher the Coefficient of Determination. That is why in addition to paying attention to the R^2 value, I focused on the theta/percentage change of each variable.

Our model also had a risk for multicolinearity error. Even with removing most of my features, our coefficients still seem impossibly large ([2.75247864e+01, precipitation] and [4.67686750e+02, hour]). Our dummy variables are generally the place which will cause the most error. Using UNIT as a dummy variable adds hundreds of dummy variables to the model. In this case, adding UNIT was a probably a good idea only because specific stations probably have a huge effect on ridership variation.

The Mann-Whitney U-Test is susceptible to error when samples are drawn from two populations with similar means but with different variances (as in our case). It's better to use the t-test if possible in these cases.