0. References

N/A

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?

When looking at both sets of data we see that it is not normally distributed ruling out the use of the t-test. For this reason we will use a Mann-Whitney U test. We operate on a two-tail test because it is not clear as to whether ridership would increase or decrease. The null hypothesis is that the distributions of the ridership when raining is equal to the ridership when it's not raining. The p-critical value is 0.01.

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.

The Mann-Whitney U test is applicable because the distribution of ridership data for both samples are not normally distributed. This test provides us a good alternative to the t-test in the event that the distribution of the samples are not normal.

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 results from the Mann-Whitney U test are below:

Mean of Ridership without Rain	1090.28
Mean of Ridership with Rain	1105.45
U statistic	1924409167
P-Value	0.025

The results of the test show that the outcome is statistically significant.

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

These results show that the ridership of system is higher when it is raining and that these results are statistically significant. This is seen by looking at the means of each of the samples: Mean with Rain > Mean without Rain. Also, looking at the p-value of 0.025 we see it is statistically significant because it is greater that the p-critical value of 0.01.

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?

The OLS model was used to perform linear regression on the samples.

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

The input variables used were rain, precipi, hour, meantempi, and meanpressurei. A dummy variable was also used when looking at the UNIT column.

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.

The choice of my input variables were based in part on trial and error. Features that seemed like they could most likely affect the outcome of ridership were placed in the model, then gradually, features were removed to view their effect on the R^2 coefficient. This narrowed down the choices to specific features that made the largest impact.

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

The parameters of the non-dummy features are as follows:

rain	8.096
precipi	22.491
Hour	65.357
meantempi	-11.114
meanpressurei	-338.464

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

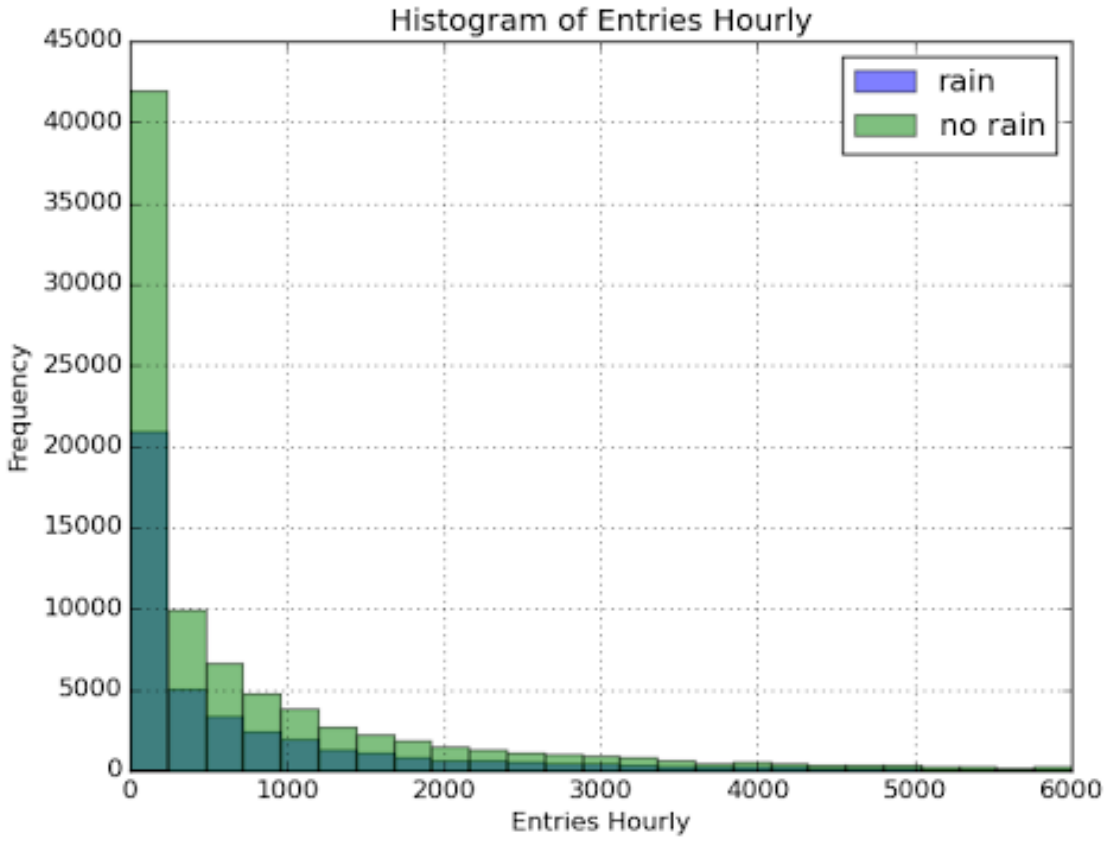
The models R^2 value is 0.4796.

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 shows how well our model predicts the outcome. In this case, the value of R^2 shows that 53% of the variation is residing in the residual and is not accounted for in the model. For this data set having an R^2 value > 0.4 for a linear model is acceptable.

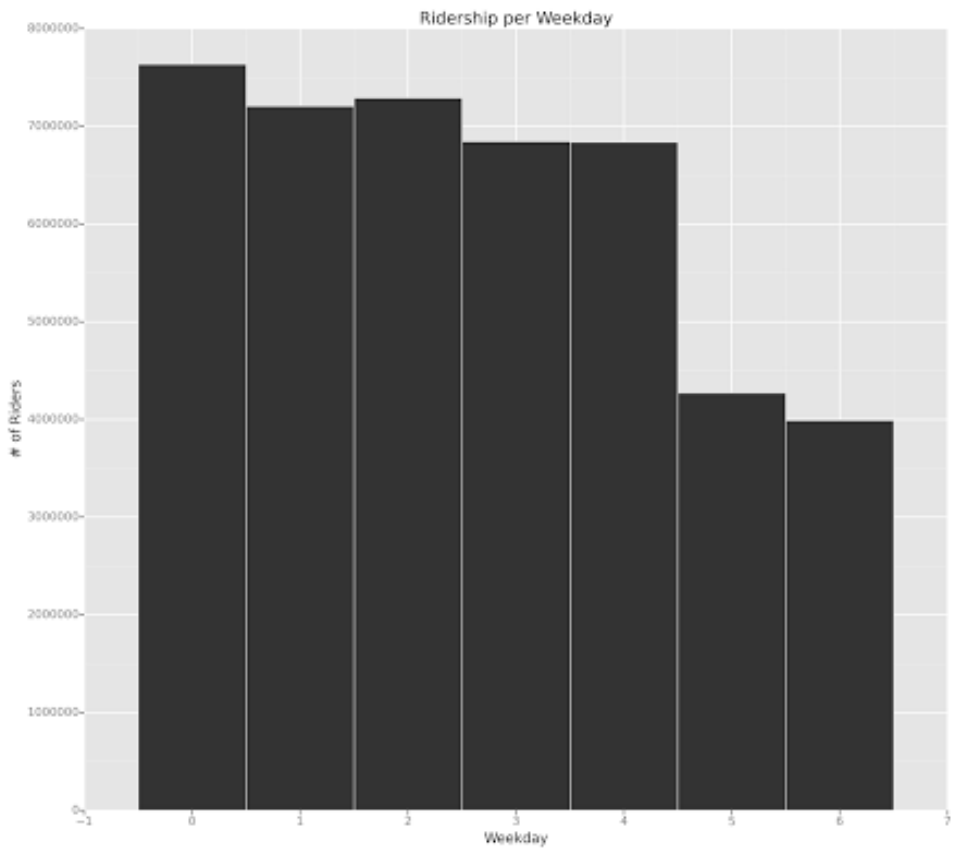
3. Visualization

3.1. Histogram of entries per hour for rainy and non-rainy days.



Looking at this distribution we see that both on rainy and non-rainy days that the distributions are similarly distributed and negatively skewed. The distribution is not normally distributed in either case.

3.2. Ridership by day of week.



Looking at the distribution of ridership we see a large dip in ridership for Saturday and Sunday. This may be attributed to people spending time at home and not making the trip to work.

4. Conclusion

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?

It is clear that more people ride the NYC subway while it is raining. The first set of data to bring us to this conclusion is the Mann-Whitney U test. Using this test we were able to determine that the mean of the ridership when it was raining was greater than the mean of the ridership when it was not raining. Furthermore, we determined that this difference was statistically significant. From here we were able to predict ridership using a linear regression and the ordinary least squares model. The predictions were able to acceptably model ridership using some key factors such as rain. We were able to determine that rain was a key factor in creating the ridership predictions through our models.

4.2. What analyses lead you to this conclusion?

The initial analysis used to determine statistical significance was a two-tailed Mann-Whitney U test. This test was used because the distribution of the datasets were not normal. This allowed us to conduct a test to check whether the means of the two distributions were statistically significant. We found that the mean ridership while it rained, 1105, was greater than the mean ridership while it didn't rain, 1090. We received a p-value of 0.025, proving that the difference in means is statistically significant. From this point a linear regression was done using the ordinary least squares method. We were able to find a set of features, which included rain, that gave us a coefficient of determination greater than 0.4. This showed that our model acceptably predicted the outcome.

5. Reflection

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

Some key shortfalls come around the use of our analysis. The Mann-Whitney U test's key disadvantage is that it is difficult to make a qualitative statement about the actual difference between populations. For this reason, confidence intervals cannot be determined. Another shortfall occurs in the use of ordinary least squares to perform the linear regression. OLS always finds the optimal solution when performing linear regression but does not account for confidence intervals between our thetas. This could cause an issue when performing our analysis.