Katherine Tansey

Project P1: Analyzing the NYC Subway Dataset

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?

Answer: I analyzed the data using a Mann-Whitney U test for non-parametric data. The output of Mann-Whitney U test was a one-tailed P-value. I converted this to a two-tailed P-value by multiplying the returned p-value by 2. The use of a two-sided p-value was done as a one-sided p-value assumes in advance that rain will not be associated with lower ridership, whereas a two-sided p-value allows for both directions (that rain results in lower or higher ridership). Our null hypothesis is that the distributions are the same, while the alternative is that the distributions between the two groups are different. The p-critical value was set to p = 0.05, and if the resulting p-value was less than or equal to the p-critical value than the null hypothesis would be rejected for the alternative hypothesis.

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.

Answer: The Mann-Whitney U test is a non-parametric test. This means that the test does not assume the data is drawn from any particular underlying probability distribution. Since our data was not normally distributed, the Mann-Whitney U test is an optimal choice.

1.3 What results did you get from this statistical test?

Answer: The results from the Mann-Whitney U test were as follows: U = 1924409167.0, one-tail p-value = 0.025. The two-tailed p-value was 0.05. The mean for hourly entries when it is raining was 1105.446, and the mean for hourly entries when it is not raining was 1090.279.

1.4 What is the significance and interpretation of these results?

Answer: The resulting two tailed p-value from the Mann-Whitney U test was 0.05, which is less than or equal to are p-critical value. We therefore reject the null hypothesis that the distributions are the same and accept the alternative that that they are different. This means that hourly entries when it is raining are different from the hourly entries when it is not raining.

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: Gradient descent (as implemented in exercise 3.5), OLS using Statsmodels, Or something different?

Answer: I used gradient descent to compute the coefficients theta.

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

Answer: I used the following features as input variables: rain, precipi, Hour, meantempi, meanwindspdi, UNIT. UNIT was the only dummy variable included in the model.

2.3 Why did you select these features in your model?

Answer: I included the rain variable (whether it was raining or not) as work in the previous section showed this to lead to increased NYC subway ridership. Furthermore, based on this information I decided to include the precipi variable, as potentially ridership might be influence by the amount of precipitation that was occurring. I included mean temperature (meantempi) as I thought it would alter ridership on the NYC subway, with people potentially using the subway more when it was cold out. I included mean wind speed (meanwindspdi) as I thought that people may use the subway more when it was very windy. I included time of day (Hour) into the model as I thought that peak hours (commuting times) would lead to increased ridership. Lastly, I included UNIT (subway station) into the model as I thought that ridership might be different depending on stations, with some being more highly used than others. Furthermore, inclusion of UNIT dramatically improved my R^2 value.

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

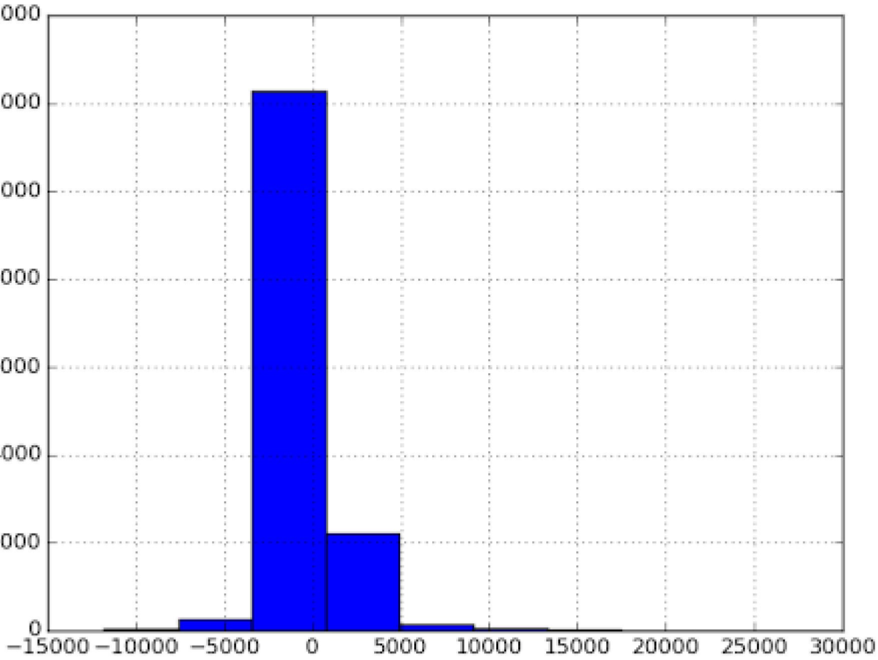
Answer: The coefficients for the non-dummy features in my linear regression model are: rain = 7.636, precipi = 4.059, Hour = 463.325, meantempi = -43.149, meanwindspdi = 55.165.

2.5 What is your model's R2 (coefficients of determination) value?

Answer: The R^2 of my model was 0.464.

2.6 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?

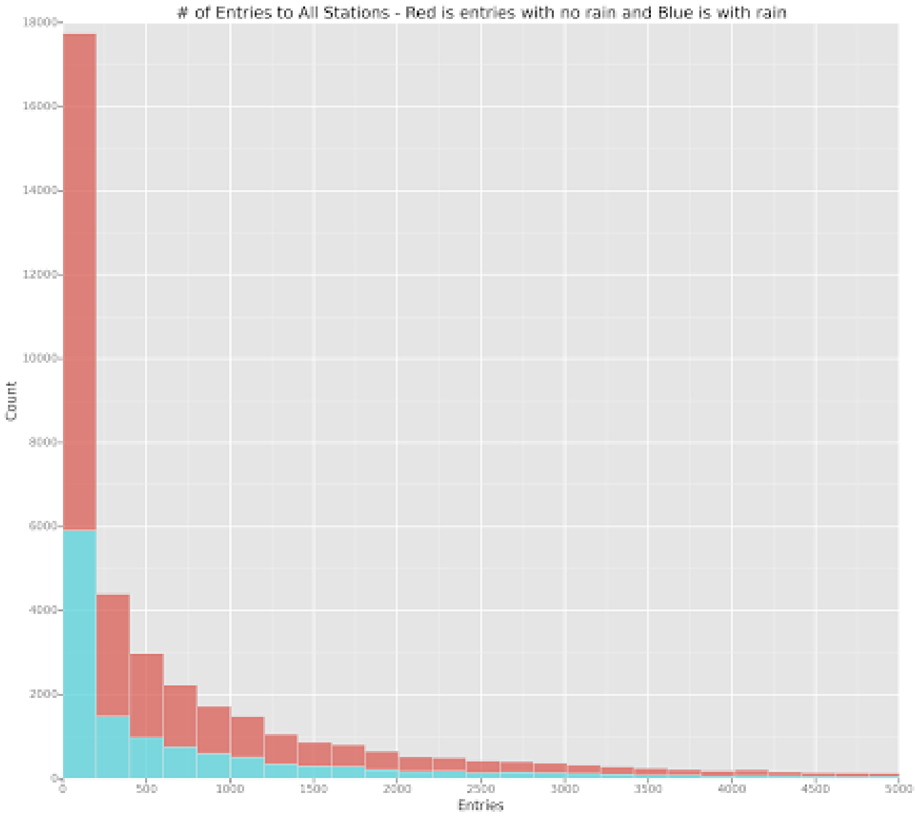
Answer: R^2 from a linear regression model is the proportion of variance in the dependent variable that is explained by the input variables (features). It can be thought of as how well the data are fitted to the regression line. An R^2 of 1 would indicate that the regression model perfectly fits the data. The R^2 of my model is only 0.464. This means that we are only explaining 46.4% of the variance in hourly entries into the NYC subway using all of the input variables included in the model. Furthermore, the goodness of fit of our model can be assessed by plotting the residuals (see histogram below). Here we can see the histogram has long tail indicating very large residuals, which tells us that some data points have large errors in the prediction from the linear model. I would conclude from this information that we have a relatively poor goodness of fit and thus the linear model to predict ridership may not be appropriate for this dataset.



Histogram of the residuals from the linear regression model.

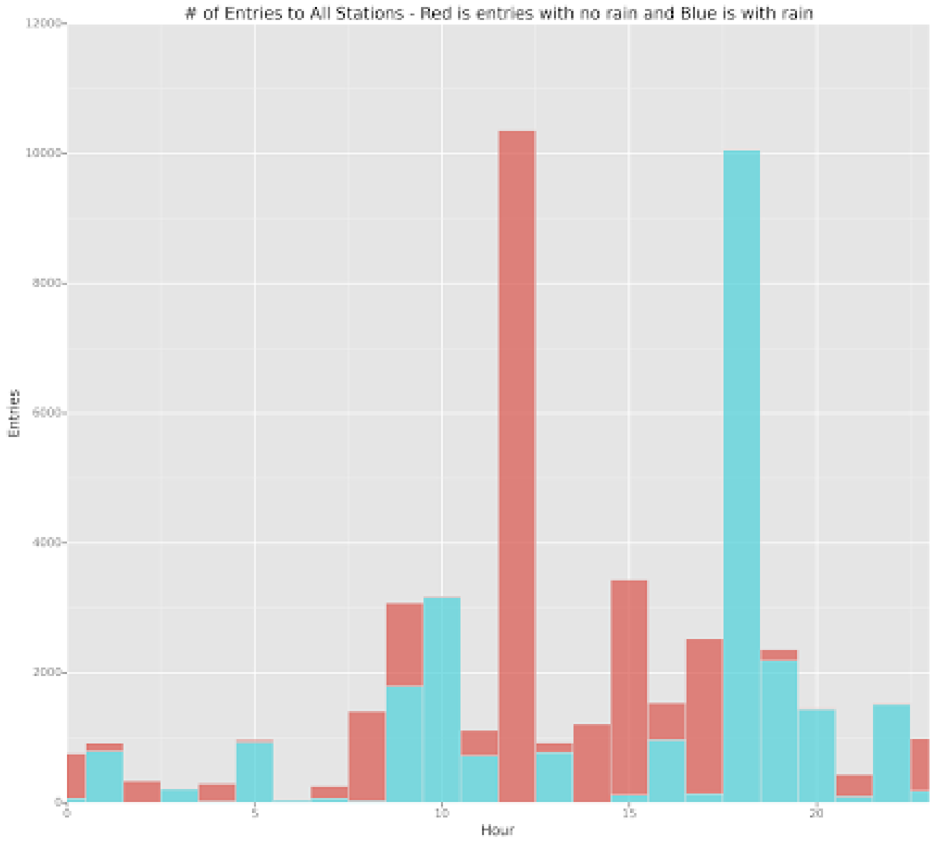
Section 3. Visualization

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



Answer: The plot above shows the total number of entries hourly into the NYC subway by whether or not it is raining. Red colored bars indicate no rain and blue colored bars indicate rain. There appear to be fewer entries when raining, but this maybe due to sample sizes differences between the two groups, meaning there are likely to be more non-rainy days than there are rainy days and thus more hourly entries into the subway on non-rainy days due to their larger number, and this is not accurately reflected in the current plot.

3.2 One visualization can be more freeform.



Answer: Plot depicts number of entries (y-axis) by time of the day (hour, x-axis). Color of the bars indicates whether or not it was raining with red indicating no rain and blue indicating rain. We can see that entries into the subway varies by hour, with peaks on non-rainy days at mid-day (roughly noon), and a peak around nightly commuting times (about 6pm) on rainy days.

Section 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?

Answer: From the analysis undertaken it does appear that people are more likely to take the NYC subway when it is raining. This conclusion is mainly drawn from the results of the Mann Whitney U test, which allowed us to reject the null hypothesis (there is no difference in ridership) for the alternative hypothesis (ridership differs depending on whether or not it is 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.

Answer: The average ridership is increased on rainy days (mean = 1105.446) compared to non-rainy days (mean = 1090.279). Using a Mann Whitney U test, I found that this difference in the mean level of ridership was borderline statistically significant (two-tailed p-value=0.05; p-critical value = 0.05).

Since our two-sided p-value was equal to the p-critical value, the null hypothesis is rejected and the alternative accepted, meaning more people take the subway when it is raining than when it is not..

This result is not apparent in the histogram plotting the total number of entries hourly into the NYC subway by whether or not it is raining. This is most likely due to sample sizes differences between the two groups, meaning there are likely to be more non-rainy days than there are rainy days and thus more hourly entries into the subway on non-rainy days due to their larger number, and this is not accurately reflected in the histogram. A better representation of the data would be a plot containing the mean and standard deviation for the number of hourly entries by whether or not it was raining.

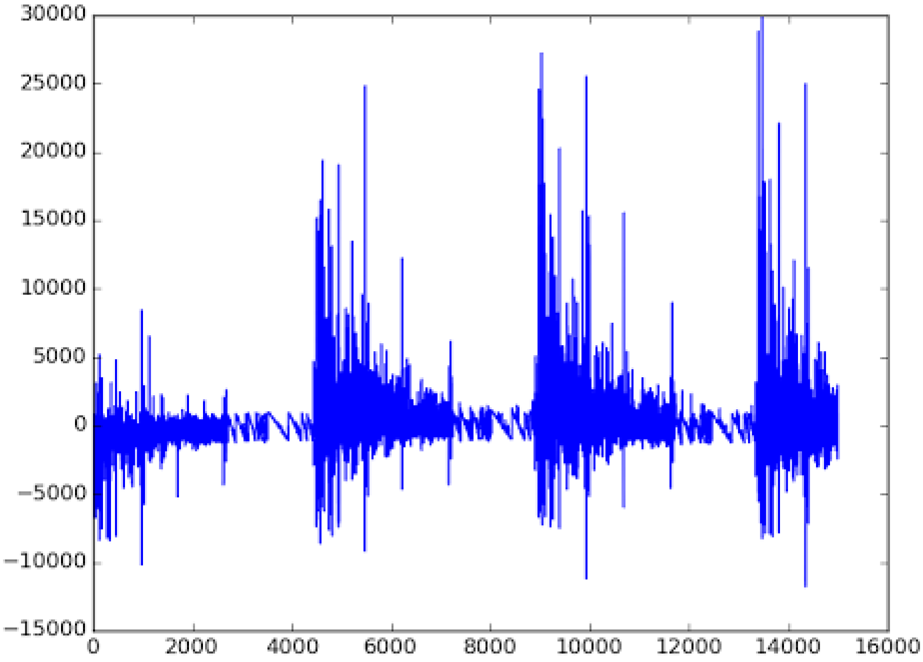
Furthermore, while people do appear to take the subway more often when it is raining, this does not explain a large proportion of the variance in NYC subway ridership, even when other variables were included into a linear regression model. The maximum R^2 value that was obtained was only 0.464. Therefore, while people are more likely to take the subway when it is raining, this information does not explain the entire variance in NYC subway ridership.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test.

Answer: The data used was only a subset extracted from a particularly time period. Using more data, from a wider time period (entire year or years) may offer different insight into the relationship between weather and NYC subway ridership. It may be possible that the month we are examining does not offer enough information for the tests or models being performed, or that this month is somehow an anomaly or outlier in the overall relationship that could only be clear using a wide time period. This is particularly true for the analysis between NYC subway ridership and rain. The resulting two-sided p-value was at the p-critical value. While I did reject the null hypothesis, this result may be strengthened or weakened by inclusion of further data which would increase the number of observations for both rainy and non-rainy days, giving us greater insight into the true nature of the relationship between these variables.

Moreover, the linear model used may not have been appropriate for the data. The histogram of the residuals included above (question 2.6) showed some very large residuals values, suggesting that the goodness of fit from the linear model was not optimal. Plotting the residuals per data point, a cyclical pattern can be seen in the data indicating some non-linearity in the data. This means that for some data points, the model does not account for their variation as well as other data points. This indicates that within the data there is another factor having an effect on these data points that has not been accounted for in the model. This effect may potentially be non-linear, and therefore the use of a linear model may be incorrect for the dataset, and the assumptions of the model are being violated.



Plot of the residuals from the linear model per data point.