P1: Analyzing the NYC Subway Dataset

Do more people ride the NYC subway when it is raining or when it is not raining?

Section 1. Statistical Test

1.1 Which statistical test did you use to analyse 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 the Mann-Whitney U to analyze the NYC subway data. The Mann-Whitney U test is a two-tailed test. The null hypothesis is that there is no statistical difference between riders of the NYC subway system when it is raining or when it is not raining. The p-critical value is 0.05 for the Mann-Whitney U test.

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 to the dataset because the dataset does not follow a normal distribution. In such a case, the Mann-Whitney U test makes no assumption about the distribution of the dataset.

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 p-value is 0.04999 and the mean for the number of riders who ride on rainy days is 1105 vs. 1090 for the number of riders who ride on non-rainy days.

1.4 What is the significance and interpretation of these results?

Since the p-value is less than the p-critical value it means that the null hypothesis can be rejected. In other words, the two populations are not the same.

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:

I used Statsmodel's Ordinary Least Squares (OLS) method.

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, weekday (custom), mintempi, meanwindspdi, minpressurei, mindewpti, yearday (custom), fog, and precipi as features in the model. I used dummy variables based on UNIT as part of the features.

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.

Initially, I selected the UNIT dummy variables because they had the biggest impact on the R^2 value. Then, I started experimenting with other features. Finally, I decided to test each feature independently for it's impact on R^2, sort that list, and decide from that list which features I wanted to use for the final model. I decided not to select features that appeared to be derived / highly correlated with another feature (e.g. rain and precipi; meantempi and mintempi/maxtempi).

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

const	1.095357e+03
Hour	4.643451e+02
weekday	-1.827942e+02
mintempi	-1.052444e+02
meanwindspdi	-5.518353e+01
minpressurei	-4.369554e+01
mindewpti	-4.468369e+01
yearday	7.650390e+01
fog	-1.176000e+01
precipi	1.051127e+01

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

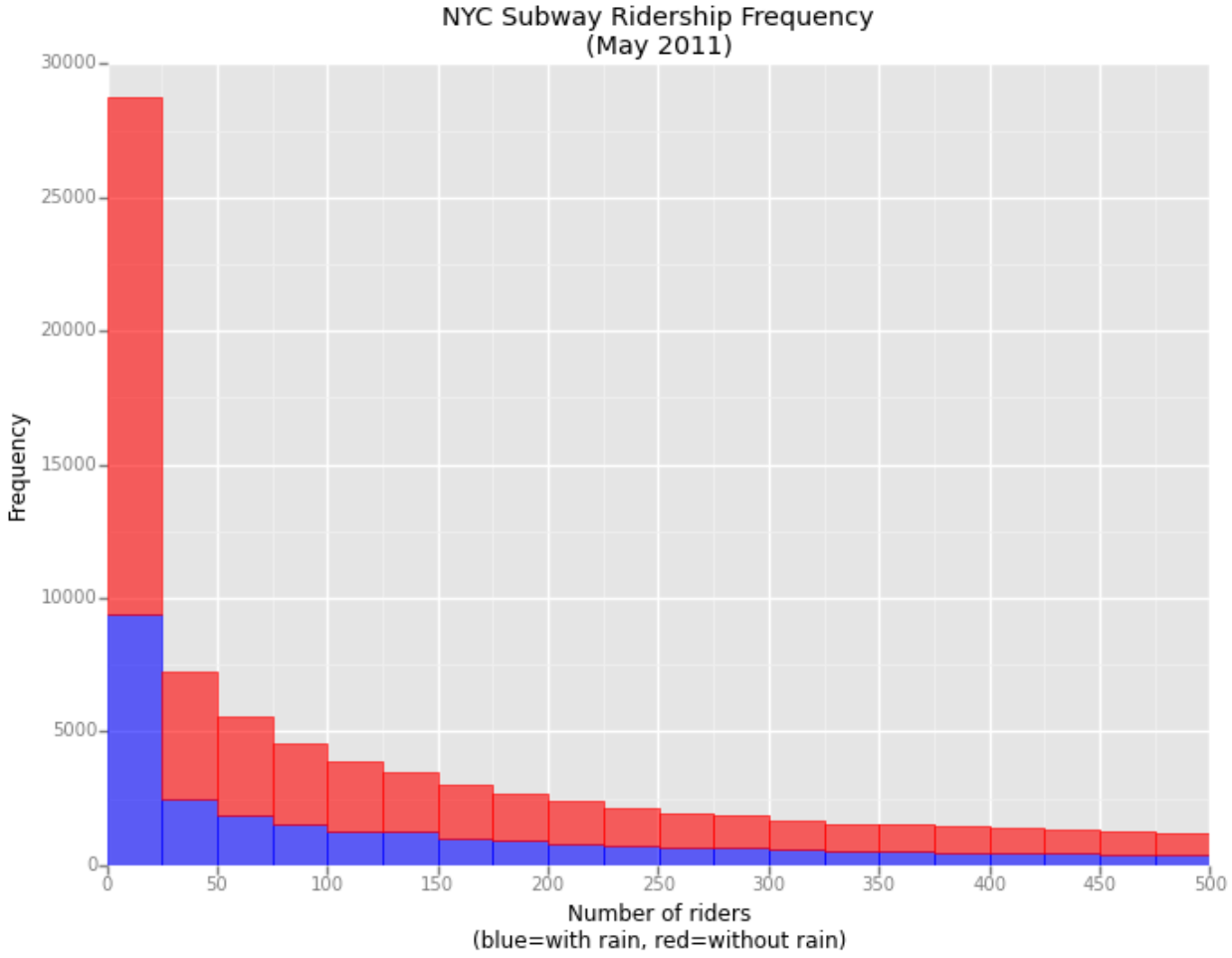
R^2 (with 'precipi' feature) is 0.463868040791
R^2 (without 'precipi' feature) is 0.463852059416

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?

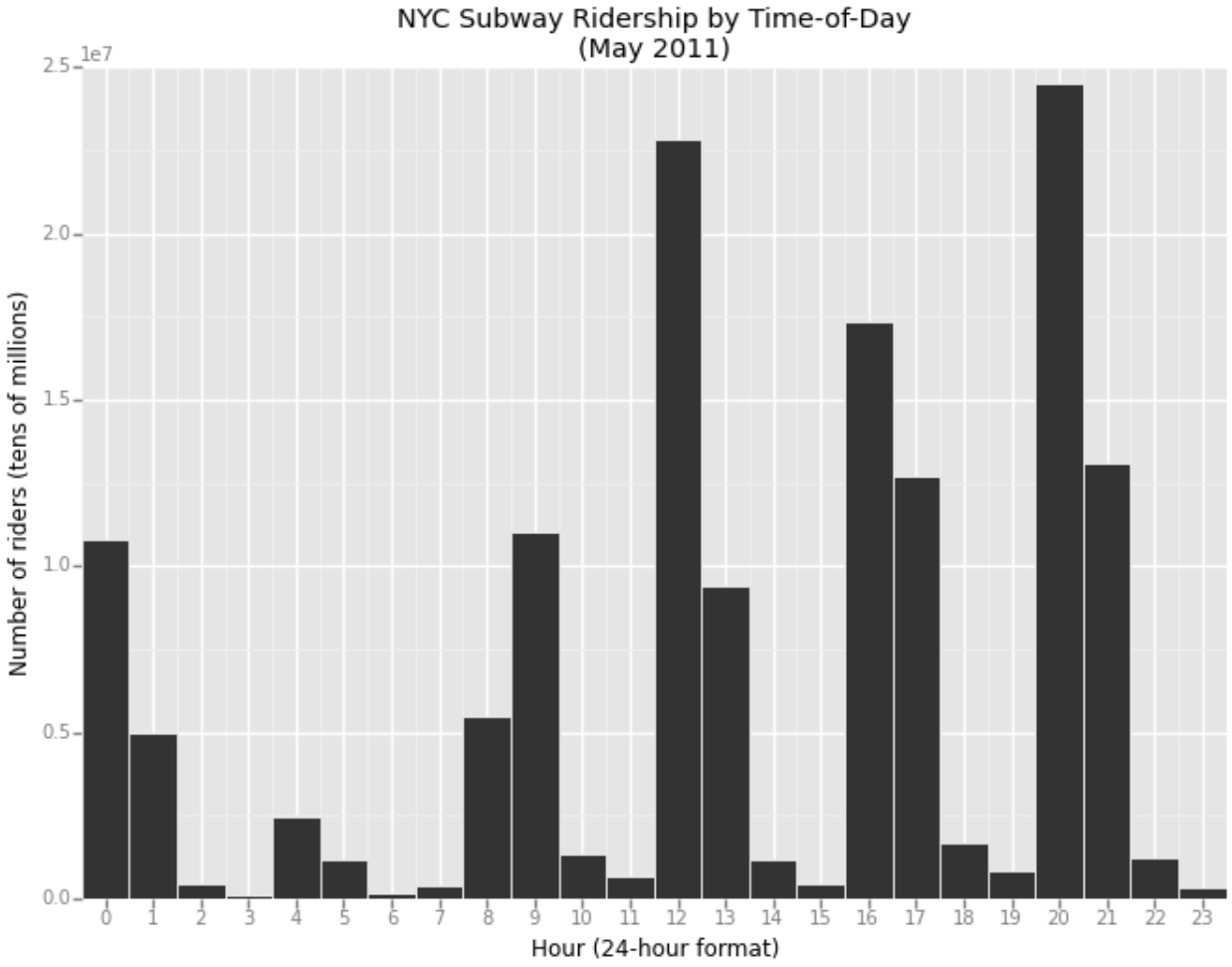
This R^2 value means that 46% of the variance of the ridership on the NYC subway can be explained by the features that were selected. I think that this linear regression model is appropriate for the dataset because it shows that ridership is dependent on location, time, day of the week/year, and some weather factors. Intuitively, this makes sense to me.

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.



3.2 One visualization can be more freeform:



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?

Yes, from the analysis and interpretation of the data there are slightly more people who ride the NYC subway when it is raining. However, it is worth noting that this finding is correlation and not causation.

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.

The Mann-Whitney U test showed that the two populations are not identically distributed. Looking at the means and medians of people who ride the NYC subway when it is raining versus when it is not raining we see that rainy day ridership statistics are slightly larger (mean: 1105 vs 1090; median: 282 vs 278). To validate this conclusion, the linear regression using OLS shows a positive weight influence (1.05) when ‘precipi’ is added to the features selected. R^2 goes from 0.463852 to 0.463868 with the addition of the ‘precipi’ feature. Another way of stating this finding with R^2 is that .0016% of the variance in ridership can be attributed to rain.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. **Dataset,**
2. **Linear regression model,**
3. **Statistical test.**

The potential shortcomings of this analysis can be reflected upon from three perspectives: dataset, linear regression model, and statistical test. The dataset only contains data for the month of May 2011. Ideally we'd have data for the full year to test ridership on rainy days in other months. Another shortcoming of the dataset is that the times in which the ridership were recorded are not consistent. As such, it's possible that rain could be missed in cases where many hours have passed (e.g. 8 hours between reads on R422 on 5/3/11) or vice versa. Some shortcomings of a linear regression model, such as OLS, are primarily that they require a linear fit and that outliers can negatively impact the model. It's possible that with some of the features selected that a non-linear fit would be more appropriate. In regards to statistical tests, the Mann-Whitney U test was helpful in showing that the two populations were not identical, however, did not provide much more information than that. Other statistical tests can provide more information if the dataset meets the requirements (e.g. normal distribution for Welch's T test).

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

The turnstile unit that a person enters from seems to be such a significant factor in determining ridership numbers that I think it would be interesting to correlate that to locations in the city and see how it relates to traffic patterns. Another interesting thing to note from our dataset is that entries and exits are not equal. So, that makes me wonder whether these missing exits are "turnstile jumpers", data loss, or a leak in the system somewhere.