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?

The Mann-Whitney U-test was used to analyze the NYC subway data and compare ridership on rainy and non-rainy days. A two-tail test was ideal for this analysis. The null hypothesis was that the two populations of rainy and non-rainy days were the same, that is, there was no correlation with ridership. The p-critical value used was 0.05.

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 non-parametric Mann-Whitney test was applicable given the non-normal distribution of rainy and non-rainy populations as illustrated in Section 3.1.

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.

Statistical metric	Values
Mean for rainy population	1,105.45
Mean for non-rainy population	1,090.28
Percent difference in mean (rainy vs. non-rainy)	1.4%
U-statistic	1,924,409,167.0
p-value (one-sided)	0.024999912
p-value (two-sided)	0.049999825

1.4 What is the significance and interpretation of these results?

The null hypothesis is rejected. While the percent difference between the means are very small (1.4%), the U-statistic is very, very large at 1,924,409,167.0, which is close to its theoretical maximum, and the two-sided p-value of 0.049 is less than the p-critical of 0.05. There is a statistical difference between the ridership on rainy and non-rainy days.

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?

A machine learning algorithm using OLS with statsmodel was used to train the linear regression coefficients.

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

The features used were rain, precipitation, mean wind speed, hour of the day, mean temperature, and weekday. The dummy variable used were the turnstile *unit* identifications. Linear regression was not applied to the unit parameters.

2.3 Why did you select these features in your model?

The default feature set of rain, precipitation, hour and mean temperature were preserved as they maintained a higher R-squared. Two further features, mean wind speed and weekday, were included as they were likely drivers for ridership; there was a boost in R-squared by including mean wind speed and weekday features.

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

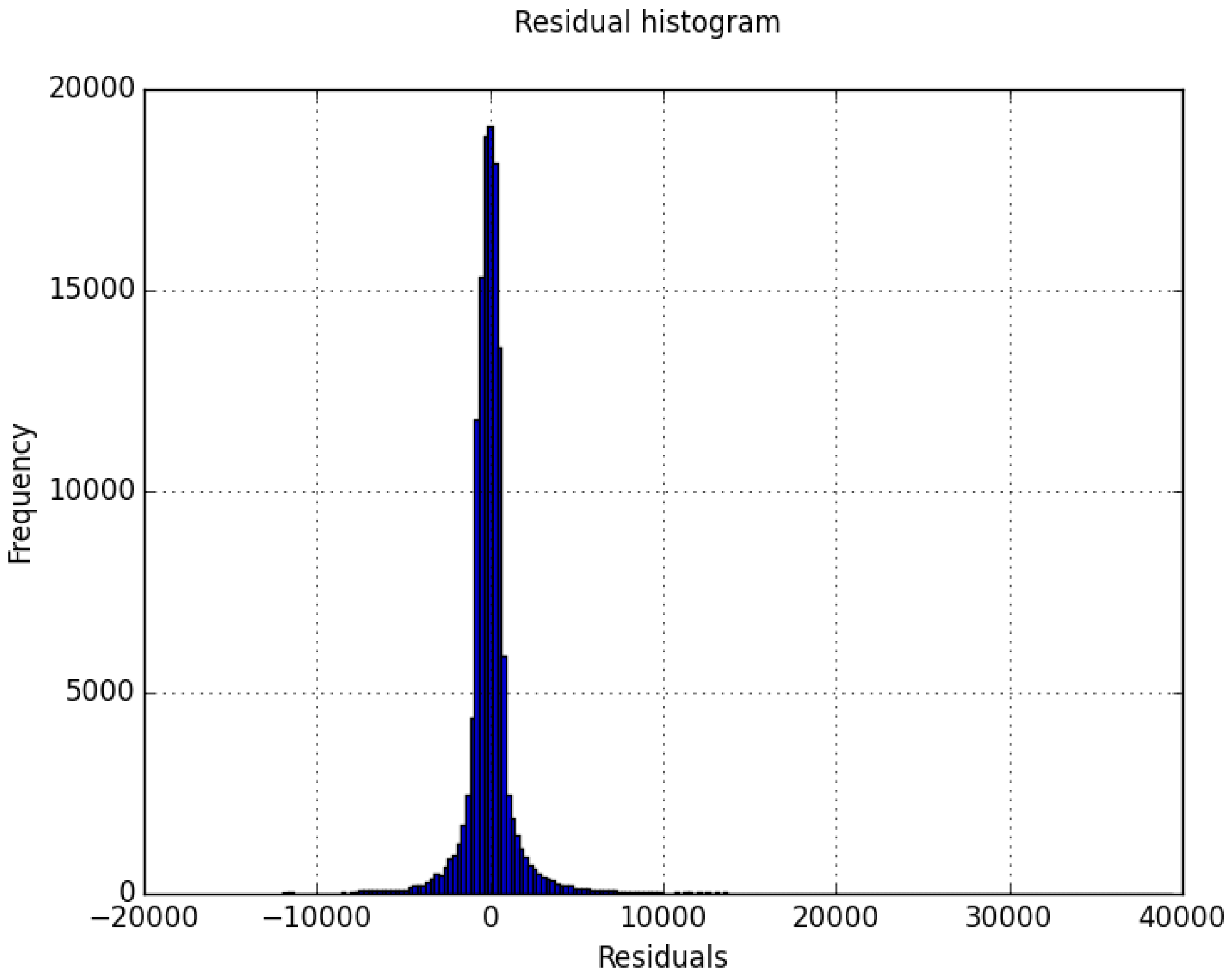
[3124.375 3393.75 3663.0625 ..., 1047.0625 1047.0625 1047.0625]

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

0.463491462688

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?

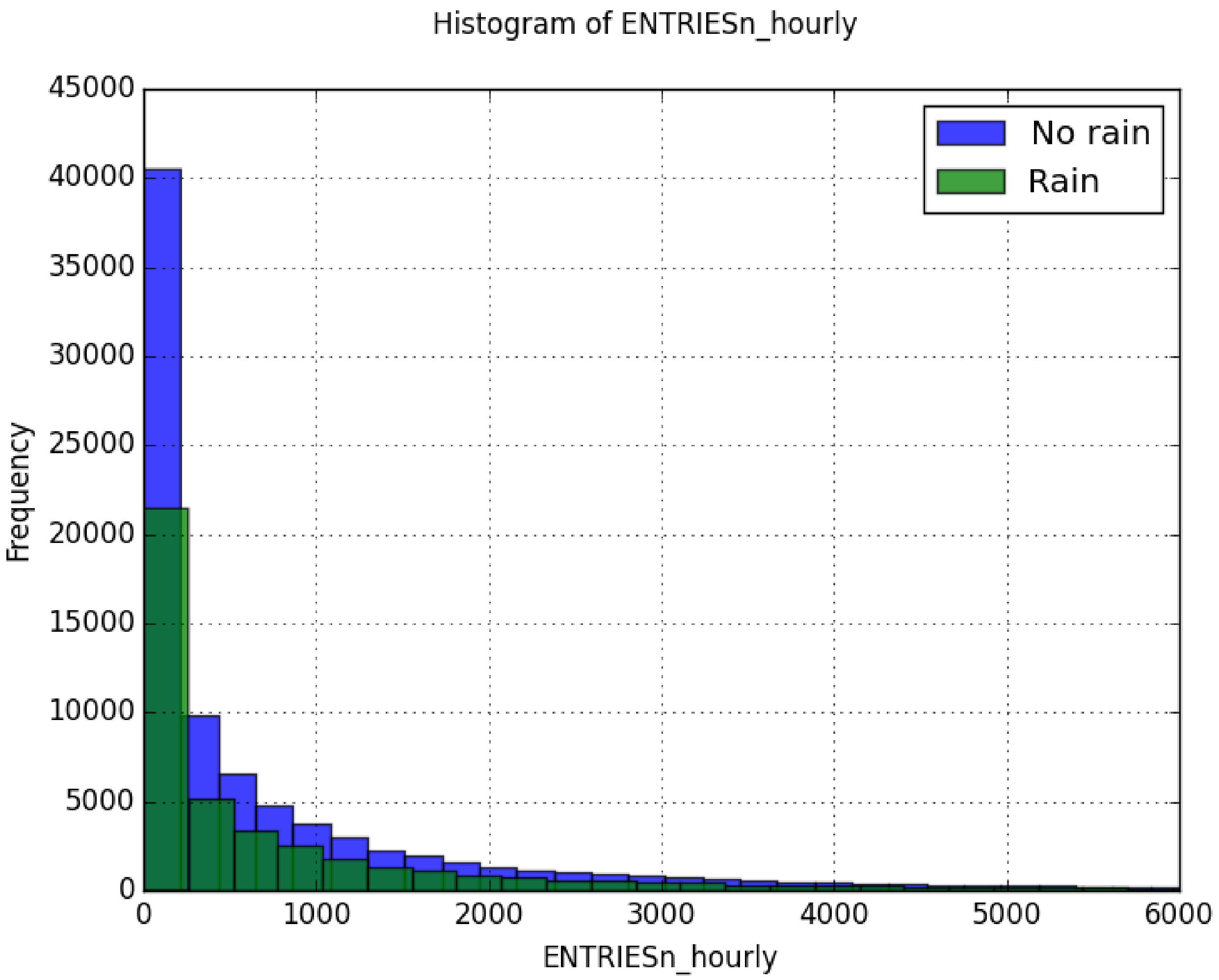
The R2 value of 0.463 (46.3%) is a low measure for the goodness of fit of variability in our regression model. This model may be appropriate for *ball-park* estimations. The adequacy of the linear regression can also be examined through the distribution of residuals, the difference between the predicted and the actual values.



The histogram of the residuals has long tails, which suggests that there are some very large residuals - a reason to question the linear regression model. Perhaps, advance studies on the drivers of readership coupled with polynomial regressions may be prove to be a better fit.

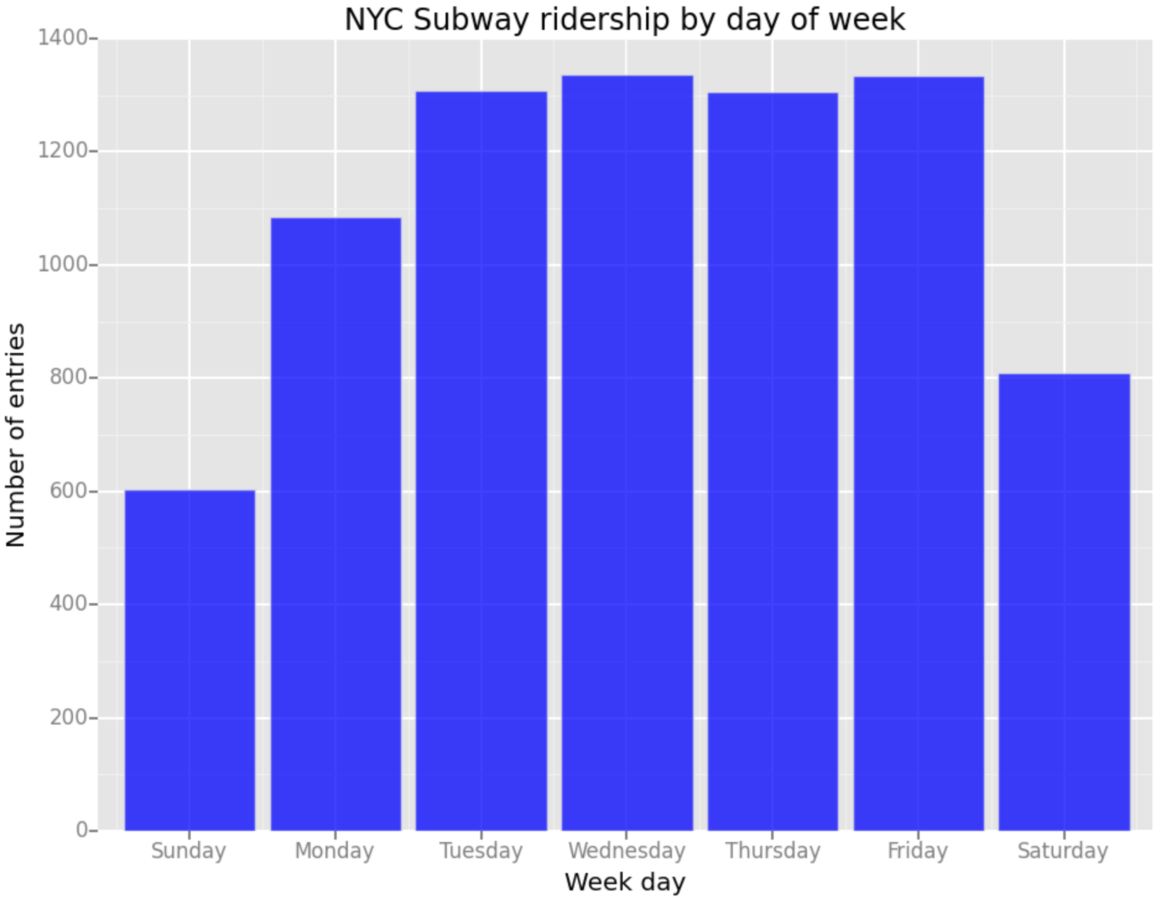
Section 3. Visualization

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



The above illustration depicts the non-normal distribution of ridership between rainy and non-rainy days. It should be noted that the two populations are not of equal sizes. Section 1.3 dives deeper into the statistical analysis of these two populations.

3.2 One visualization can be more freeform.



The above diagram shows the mean trend in ridership over the course of a week. It can be observed that ridership is high during the weekdays (Monday to Friday) and the lowest on the weekends (Saturday and Sunday). This is inline with the thought that people would use the subway mostly during the normal working days. Hence, weekdays were used as a feature in Section 2.3.

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?

Based on the Mann-Whitney U-test producing a very high U-statistic and a p-value of 0.049, it can be concluded with high certainty that more people ride the NYC subway on rainy days. It is very likely and understandable that rain is a motivator for people to seek refuge by using the subway instead of other modes of travel, such as walking and biking.

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 was used to draw the above conclusion despite the percent difference between the means of rainy and non-rainy populations being only 1.4% The positive coefficient for the rain feature parameter further indicates that rain contributed to the increase in ridership. Yet, the low R2 of 46.3% indicates that there is a fairly weak correlation to 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.

When inspecting the dataset, it can be noted that there was more entries than exits, as well as more non-rainy days than rainy days. This could have likely happened due to errors/miscounts on some turnstiles, and the vast variability in ridership by *units*. It is likely that taking a larger dataset over the course of multiple months and normalizing the data by subway locations could have resolved these issues.

Furthermore, it should be noted that many variables included in the dataset that might be very closely related such as minimum, mean and maximum temperature. It may be difficult to disentangle the effects of such similar features and running the risk of problems with collinearity. This can cause some linear regression algorithms to give incorrect results.

The quantitative analysis used in this project had some shortcomings. Notably, the low R2 value. As discussed in Section 2.6, the inclusion of more features through a more expansive study or polynomial regressions could have increased the accuracy of the model. The machine learning algorithm used was unsupervised. It may have been beneficial to split the dataset into a trained and untrained sample sets.

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

The machine learning algorithm used here was very interesting and had a very steep learning curve. It'd be interesting to see if I can revisit this analysis once I'm done with the machine learning course. I would have liked to use gradient descent and cost analysis a little more with this dataset.

References

- Machine Learning with Python - Linear Regression
- U-statistic
- SciPy documentation
- Stats model documentation
- ggplot documentation
- Multicollinearity