Analyzing the NYC Subway Dataset

Questions & Answers

Overview

Answers are inline. I used the IMPROVED DATASET.

Section 0. References

- https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
- http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
- http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html

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?

I used a non-parametric test, the Mann-Whitney U-test. Since I assumed that more people ride the subway on rainy days, I used a one-tail P value. The null hypothesis is, that the ridership on dry days is not higher than on rainy days. My p-critical value is <.025.

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 does not make assumptions on the distribution of the data, i.e. the data has not to be normally distributed. It just gives the probability that the samples are from the same population. So a t-test, for instance, would not be applicable. Further, requirements for sample size (>20) are met.

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.

```
rain: 1105.4463767458733
no rain: 1090.278780151855
p: 0.024999912793489721
```

1.4 What is the significance and interpretation of these results?

On a significance level of .025, the ridership on rainy days is higher .

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:

At first I used gradient descent with code from the old lesson #3, then OLS from Statsmodels.

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, rain, station and day_week, the latter two being dummy variables, because they have ordinal values.

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.

At first I asked myself under which circumstances I would use the subway and included additional variables. Then I added other features at random. Later I tried to leave the stations out, because that section of the code did not invite to experiment, and saw the first bigger change in R^2. So I started over and included only the stations plus one feature and compared the resulting R^2s. I settled on stations, hours, day_week and rain, which is a hybrid approach between intuition and increasing R^2.

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

The parameters are:

const	3135.6513
hour	855.0956
rain	13.5237

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

R-squared: 0.434

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?

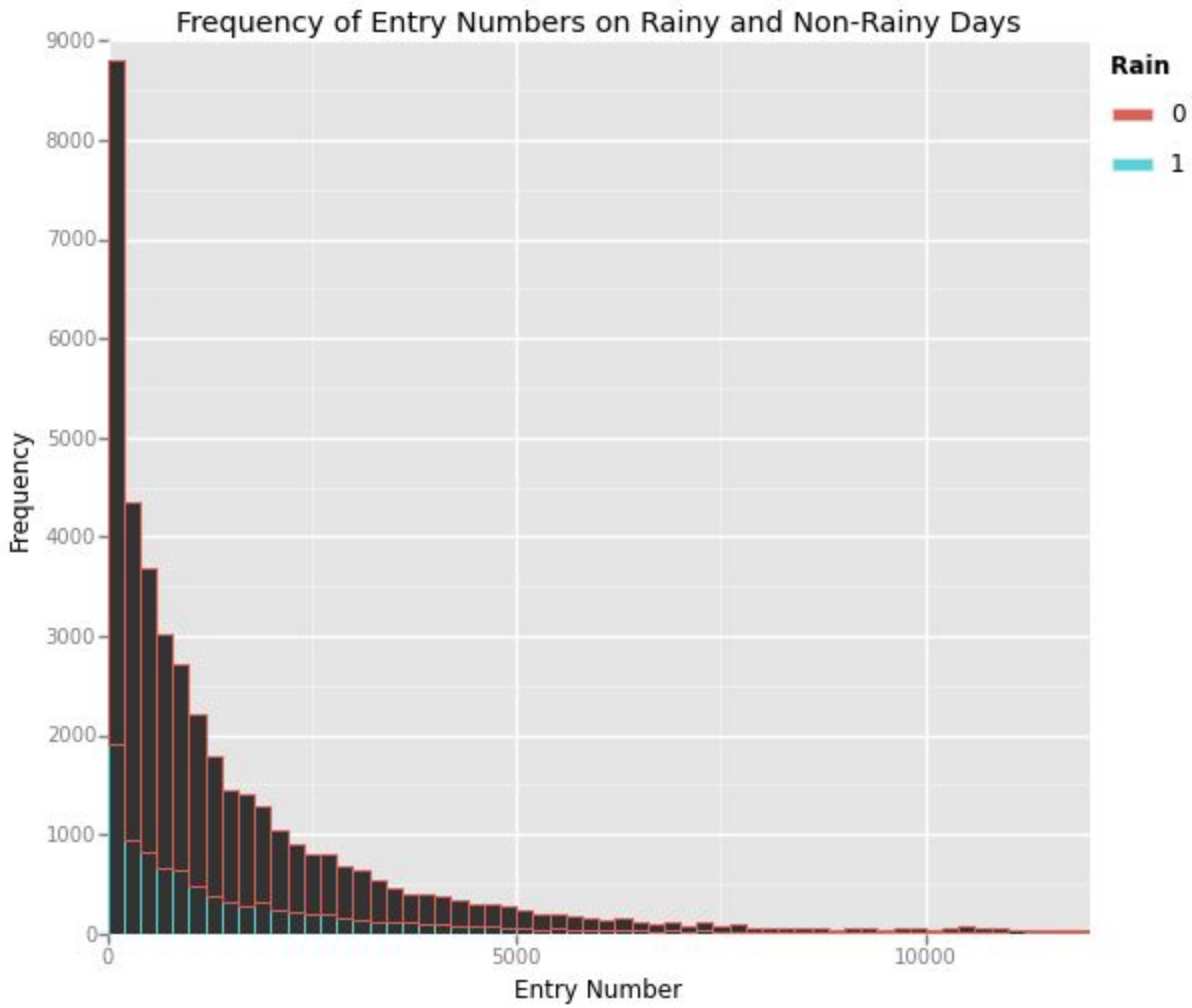
It means that I can predict ridership moderately well. Given that the dataset does not include all information which might have an impact on ridership. for instance public events, school holidays or construction sites, I think that the model is appropriate.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

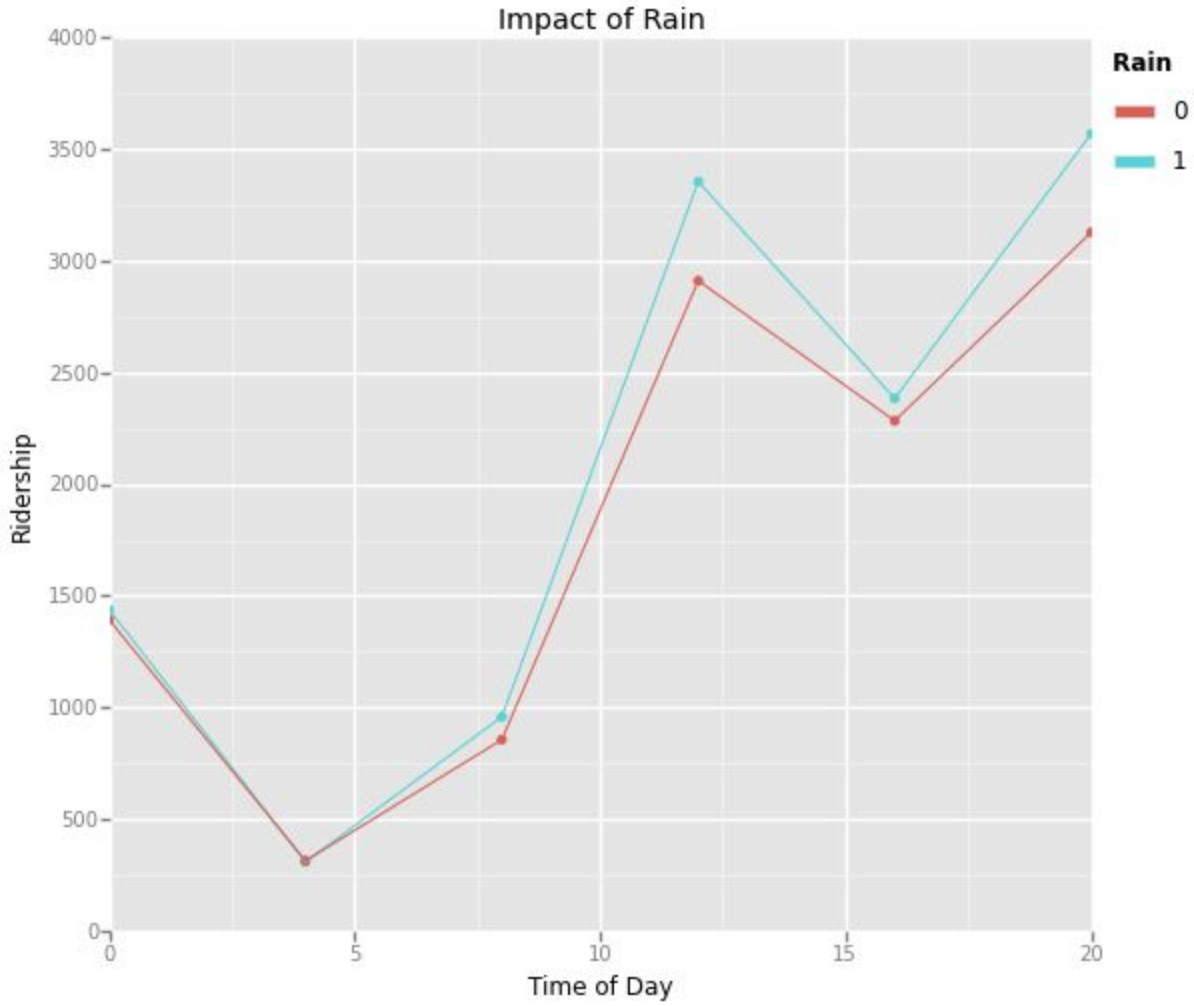
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for

non-rainy days.



There are much more rows with few entries independent of rain, and fewer rows for rainy days in general. Seems to be a power-law distribution.

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.



People tend to use the subway more often on rainy days, especially during morning and evening commute hours. The time

interval is 4 hours, so that I can only estimate at which time the rain has the highest impact.

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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?

More people ride the NYC subway when it is raining. Rain is not the major influence on ridership, though. Other factors, such as the station or the time and day seem to have a bigger impact on how many people use the subway. There may be even features with a higher effect on passenger volume, which are not included in the dataset.

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.

As shown in section 1, slightly more people use the NYC subway on rainy days. The difference is statistically significant on an alpha level of .025. This is supported by the visualization from section 3 (Impact of Rain), where the influence of the rain seems to be stronger at certain times of the day, probably the commute hours.
While rain has an effect on the subway usage, other factors turned out to be more important in the linear regression, as discussed in section 2. From the information available in the dataset, the station, day of the week and time of the day had a bigger influence on the ridership.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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

From my experience, school holidays, seasons or sports events have a major effect on subway usage. These were not included in the dataset. Further, the cleaning of the data may have introduced errors.
The histogram in section 3 shows, that most of the entry numbers are rather low, but few are very high, hinting at a power-law distribution. I tried to run the linear regression on subsets of the data, chosen by entry numbers. R^2 fell, when I chose only subsets with low, medium or high entry numbers for linear regression. Linear regression might not be the best choice to model this situation.