John Gritch
Intro to data science, Project:
Analyzing the NYC Subway Dataset, revision 1

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 the Mann-Whitney Test with a two-tailed P value. The null hypothesis was: Taking our data as random samples of the larger distributions of turnstile entries on rainy and non-rainy days the probability of the mean ranks of the rainy sample being greater than the mean ranks of the non-rainy sample is 0.5.

More informally the null hypothesis is roughly that there is little to no shift between the distributions of turnstile entries between rainy and non-rainy days.

P-value: **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.

While not a precondition for a Mann-Whitney test, especially at our sample size, it should be noted that both samples were not normally distributed. The distribution free and non-parametric Mann-Whitney test was an applicable test as the data fit the preconditions for valid use, namely:

- The subway turnstile data, if not functionally continuous, was definitely ordinal as the dependent variable was a count of the number of flesh and blood people who passed through the turnstiles.

- Our dependent variable of turnstile entries is separated into two distinct groups by our independent variable of rain or no-rain.

- The observations within the individual groups (rainy and non-rainy) are assumed to be independent, because for the most part person Y's decision to take the subway does not affect person Z's decision on any given day.

- It is also assumed that the groups (rainy and non-rainy) are independent of each other. To me it is less clear that this is a valid assumption as it seems reasonable that we are measuring the same flesh and blood people over the course of the month and whether a person rides the subway on rainy days could affect whether they ride on non-rainy days, or vise versa.

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.

Mean of ENTRIESn_hourly on rainy days: 1105.4463767458733
Mean of ENTRIESn_hourly on non-rainy days: 1090.278780151855

Mann-Whitney U statistic: 1924409167.0
P-value: 0.024999912793489721 * 2 = **0.0499998255869794**

1.4 What is the significance and interpretation of these results?

An informal interpretation of the results is that there is very little chance (less than 5%) that we would have seen the test results that we did if the ridership on this part of the subway was the same on rainy versus non-rainy days. More formally we can say that the distributions of the two groups are shifted by a measurable amount and that amount is described by a Mann-Whitney U statistic of 1924409167.0 and a resultant P-value less than 0.05.

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

Gradient descent (as implemented in exercise 3.5).

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

Features:
1. Hour
2. mintempi
3.EXITSn_hourly

Yes, I used the UNIT column as a dummy variable.

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.

- Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”
- Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R2 value.”

1. I used Hour because it made intuitive sense to me that entries through the turnstiles would ebb and flow during the hours people are most likely to be traveling, e.g. the morning and evening rush hours.

2. I used mintempi because I thought people might be more likely to use the subway if “it was cold outside”. Mintempi did not have a huge effect on my predictive abilities, but I kept it in because it helped the model to a small degree.

3. I included EXITSn_hourly because while certainly in an hour's time frame some spots in the city will

have more entries than exits (e.g. business districts at 8 A.M.) or vise versa, it could be seen from the data that entries matched exits to a large degree. I also found EXITSn_hourly had by far the largest effect on raising my R^2.

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

The normalized weights of hour, mintempi, and EXITSn_hourly are 2.51203828e+02, -4.36310443e+01, and 1.04107862e+03, respectively.

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

The R^2 was 0.605650362405.

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?

This R^2 value means that about 60% of the variation in ENTRIESn_hourly can be explained by the chosen features and dummy variable. I do think a linear model is appropriate for this data set as a plot of the residuals is very close to normally distributed. However, the residuals are skewed a little bit to the left so there is some structural tendency for the model to underestimate the predicted values.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

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.

- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.

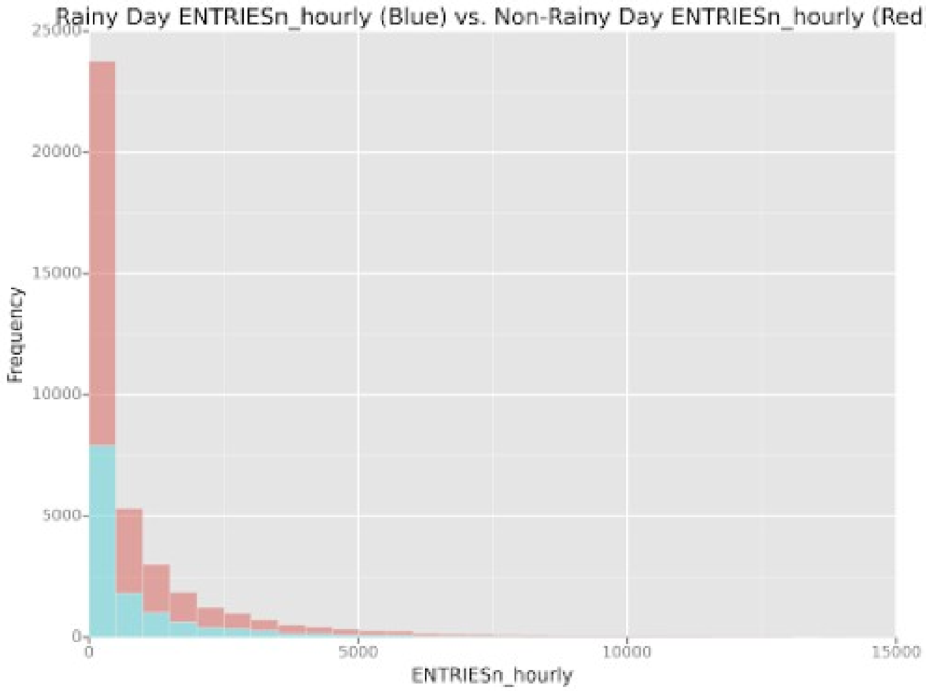


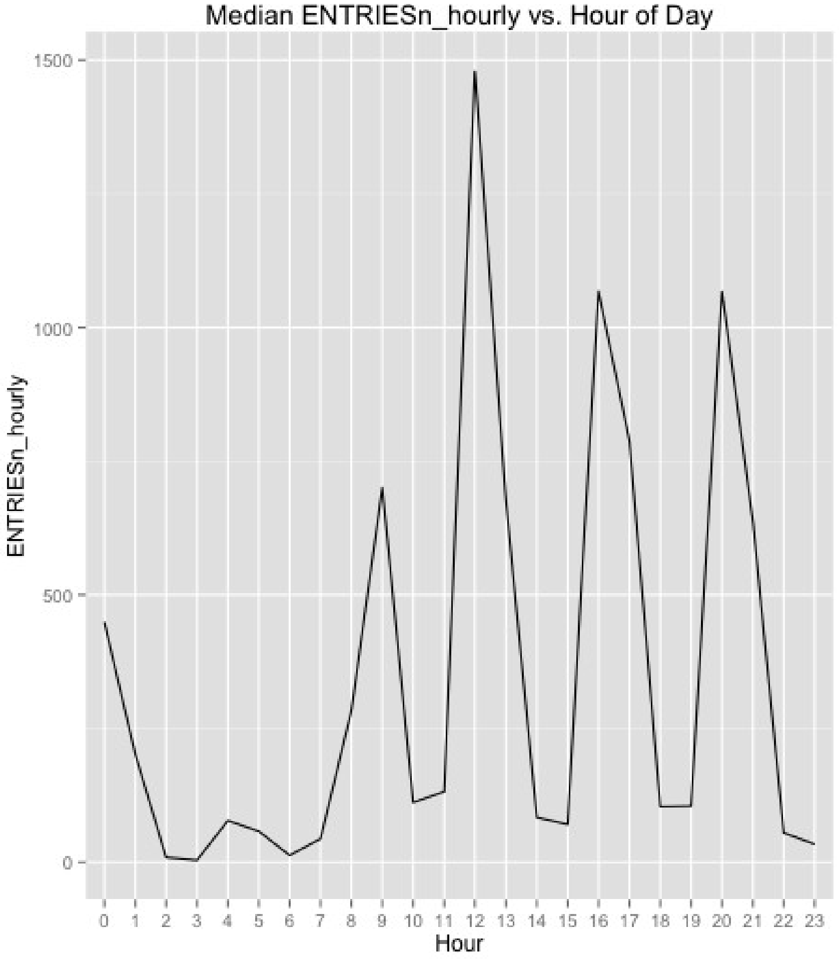
Figure displayed with a bin size of 500 and with an X-axis truncated at 15,000 entries per hour. I did not include a legend (or guide) because I encountered a known problem with ggplot and the online interpreter. This figure shows that during days with at least some rain the frequency of the absolute counts of ENTRIESn_hourly between 500 and 5000 was roughly half or less than half of the same measure on days with no rain.

From this graph we can also see that the distributions of ENTRIESn_hourly on rainy and non-rainy days is roughly similar. This similarity opens up more options for choosing an appropriate statistic, but more generally it would be hard to get an accurate sense of what is going on without having at least a rough idea of these distributions.

It is not shown in this figure, but the long right tail appears to extend asymptotically to about 45,000 and there is no second peak or massing of data past the truncation point of 15,000.

3.2 One visualization can be more freeform. Some suggestions are:

- Ridership by time-of-day
- Ridership by day-of-week



This is a graph of the median ENTRIESn_hourly vs. the hours of the day for the month of May. The graph shows that there are a few short periods of steady low ridership, but that most of the day is spent ascending to or descending from a local maximum. These local maximums are 9 A.M., 12 P.M., 4 P.M., 8 P.M., and 12 A.M. Most of these local maximums seem fairly intuitive, a morning rush hour commute, lunch hour traveling, an early evening rush hour commute and a late evening commute. The midnight maximum may be people returning from swing shift work or arriving for night jobs.

However, it should also be noted that this data includes weekends so non-work based traveling is influencing the data as well. It's possible the midnight (or any other time) data may have two distinct patterns based on workdays versus non-work days and we are seeing the mathmatical median that might not reflect levels of ridership seen on any given day in the real world.

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?

There are a few valid ways (it seems to me) to answer this question.

What we know for sure is that over the course of May more people entered the turnstiles of our sample on days when there was no rain as opposed to days when there was rain. However, this is mostly a function of it having rained only 10 days out of the 30.

However, on days where there was rain, the turnstiles in our sample had more entries on average than when there was no rain (details below in 4.2).

It should be noted that our data could only distinguish between days that had some rain and those that did not. There was no way to examine ridership during the moments (or hours, or even four hour blocks) when it was actively raining and those when it was not.

Furthermore, what I do not know is if one) we can extend the findings from our sampled turnstiles to the NYC subway system in general and two) if we can extend our findings past the month of May. With these reservations (and with those discussed in Section 5) for the purpose of this exercise my answer is yes, more people ride the subway when it is raining then when it is not.

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.

If we repurpose the Mann-Whitney test (from problem set 3) into a directional test taking the general form that the distribution of x (rain) is “greater” than than the distribution of y (no rain) one set of hypotheses could be:

Null: A sample from the population distributions of ENTRIESn_hourly on rainy days has an equal chance of being greater than a sample made from the population of ENTRIESn_hourly on non_rainy days as vise versa.

Alternative: The population distribution of ENTRIESn_hourly on rainy days is greater than on non-rainy days.

For this test we get a P value of 0.249 (rounded to the thousandths place). Setting a significance level of alpha = 5% we can reject the null hypothesis in favor of the alternative. The median and IQR of ENTRIESn_hourly on rainy days is 939.0 and 2129.0, respectively. The median and IQR of ENTRIESn_hourly on non-rainy days is 893.0 and 1928.0, respectively.

We can extract some supporting evidence that rain has a positive effect on subway ridership (or at least

ENTRIESn_hourly) from various OLS linear regression results. The model with the strongest predictive ability, that I found, used EXITSn_hourly and rain as features and had an R^2 of 0.534914453416. However, these two features are very likely correlated with each other to a high degree so the coefficient of rain in this case (356.5455) is suspect. However, I used several combinations of features in the OLS linear regression models and the presence of rain always had a positive (more ENTRIESn_hourly) effect on the models.

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, including:

1. Dataset,
2. Analysis, such as the linear regression model or statistical test.

My first reservation is with the origin of the data. Not having an API key with the MTA it's hard to get details on exactly how they acquire and package their data. While their data is probably accurate it may not mean what I think it means in any number of ways. However, putting those reservations aside a very serious shortcoming of the original dataset, at least as far as how I used it, lies with the ENTRIESn_hourly values. These values (if the downloadable dataset is meant to mimic the ENTRIESn_hourly we created in the problem set) were created by finding the difference between two adjacent rows. My first reservation is that ENTRIESn_hourly is really the number of entries in a four hour block of time.

Secondly, the blocks do not span the same periods of time in the day so they can not be compared straight across. Not only are the blocks not synchronized across the hours (i.e. each day having 6 four hours blocks all starting at the same time), some of the blocks start at seemingly random points between hours such as 00:42:48. One effect of these dissimilar blocks is that a period of high traffic may be split across two blocks.

In my mind this throws into serious question the validity of the conclusions we drew from the Mann-Whitney test, both forms of linear regression and the visualizations. I think the Mann-Whitney test would have been better served if we had summed up all the entries for the day and then ran our tests off those cumulative numbers (and as suggested by Charlotte Turner making sure to aggregate by UNIT as well as rain and no-rain). As long as the full 24 hours of the day were represented this would have avoided the situation of possibly splitting up a busy four hour block into two not so busy four hour blocks, or vise versa. As the data stands there is no telling how this mix and match affected our results.

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

If I was to further study this data set I would explore why mean atmospheric pressure seems to have more predictive power on ENTRIESn_hourly than the temperature measurements (mean temp, min temp, max temp).