Michael Hood
January cohort
drmichaelallenhood@gmail.com

Intro to Data Science - Project 1

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 a Mann-Whitney U-test and interpreted the P-value as a two-tailed test to see if there was a difference in ridership when raining vs. the ridership when not raining.

The null hypothesis is that the ridership is the same regardless of whether or not it is raining:

$$Ho: \text{There is no difference between subway ridership when it is raining or not raining.}$$
$$HA: \text{There is a difference between subway ridership when it is raining or not raining.}$$

I used p-critical value of $\alpha = 0.05$

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.

I looked up more information about the Mann-Whitney U test and learned that the following assumptions should be satisfied if I want to properly apply the test to the data:

1. I assume that ENTRIESn_hourly for each sample do not follow a normal distribution.
2. I assume that the ridership values are ordinal, i.e. 1200 riders can be determined to be greater than 1000 riders.
3. I assumed that both samples are independent and unpaired.
4. The data values are continuous. (close enough, you cannot have 1.5 riders)
5. Both sample distributions have roughly the same shape (see Figure 1):

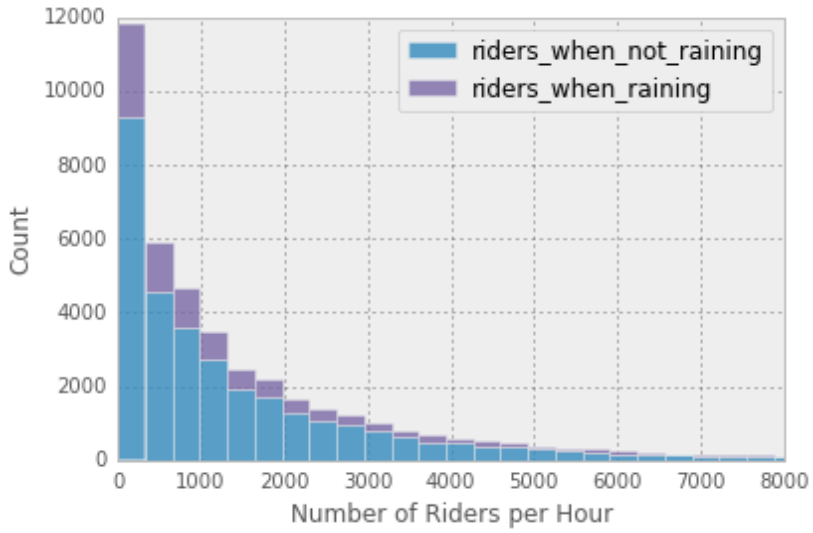


Figure 1: Histogram of riderships when raining and not raining.

I believe that all of the above assumptions are satisfied, so I am confident the Mann-Whitney U test is appropriate.

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.

```
U, p = scipy.stats.mannwhitneyu(rain_df.rain, rain_df.no_rain, use_continuity=True)

mean_riders_rain, mean_riders_no_rain = rain_df.rain.mean(), rain_df.no_rain.mean()

mean_riders_rain, mean_riders_no_rain
Out[20]: (2028.1960354720918, 1845.5394386644084)

U, p
Out[21]: (284493459.5, 0.0)
```

1.4 What is the significance and interpretation of these results?

The extremely low p-value of 0.0 means that there is effectively a zero percent chance that the difference in ridership values between the two groups can be attributed to random chance. Therefore, we reject the null hypothesis, and can conclude that the distribution of riders when raining is distinct from the distribution of riders when not raining.

This can be interpreted that there is a difference in ridership on the NYC subway depending on whether or not it is raining. Looking at the mean ridership of 2028 riders per hour when raining, versus 1846 riders per hour when not raining, indicates that the difference is in the direction of more riders when 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?

I used the implementation of gradient descent that I submitted for problem 3.5. I also used `np.linalg.lstsq` to create a multiple linear regression but I will not focus on that because I was only able to achieve an r^2 of 0.03. I spent far more of my time building a model using gradient descent, so that is what I will base my discussion on.

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 the following features:

1. `'hour'` - This is the hour of the day as an integer, e.g. 0-24.
2. `'precipi'` - This is the amount of rainfall in inches
3. `'UNIT'` - I used units as dummy variables
4. `'weekday'` - I used the fact of whether or not it was a weekday 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.

Here are my reasons for including the following features:

1. `'hour'` - It seems reasonable that more people will ride the subway during the daytime than at night. Additionally, there are certain times of day that are busier than others, such as morning and evening rush hour. The time of day was the second strongest predictor of overall ridership.

2. `'precipi' -` The level of precipitation did not do much to increase the predictive power of my model. Incorporating this feature provided a slight boost of about 0.01 to r^2 .
3. `'UNIT' -` The visualization in Section 3 (figure 3) shows that location is an extremely important factor in determining the ridership. Incorporating the unit - as a proxy for physical location - by using dummy variables made a significant boost to r^2 of nearly 0.3.
4. `'weekday' -` I thought to include this feature while I was conducting mann-whitney u tests on the ridership. I noticed a difference in mean ridership for both rainy and non-rainy days of about several hundred people. This makes sense, as I think it is reasonable to assume that less people will be rushing around to and from work on the weekends. Accounting for whether or not the day was a weekday improved my r^2 by about 0.08.

I experimented with many other variables such as `'fog'`, `'tempi'`, and `'wspdi'` but there was no meaningful improvement in r^2 when they were incorporated. I even developed a method of incorporating the distance of a turnstile unit from it's associated weather station. None of these resulted in a meaningful improvement in r^2 .

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

$$\theta_{rain} = -70.45500581 \text{ and } \theta_{hour} = 850.37168499$$

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

This model has an r-squared of 0.481912.

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 coefficient of determination, r^2 , is a measure of how much of the total variation between the observed and predicted ridership values can be accounted for by my model. In this case, about 48% of the variation can be explained by my model. If the model were appropriate, I would have expected something closer to 0.7 or 0.8. While it is possible that I could have extracted some additional features from the dataset that had more predictive power, I do not believe that the underlying relationship between that weather is a very useful method of predicting ridership levels. Therefore, I do not think that my linear model is appropriate. It may be that there is some other linear model - based on non-weather features - that is more appropriate.

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.

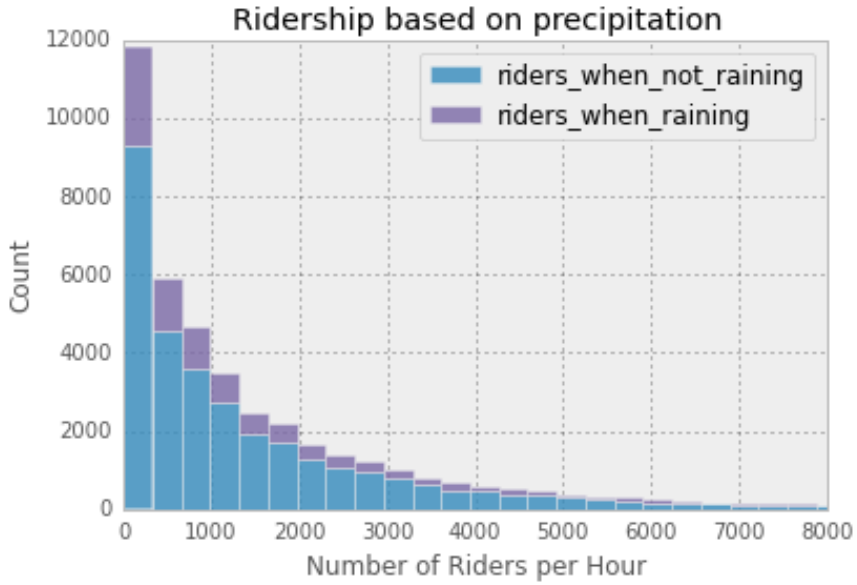


Figure 2: This visualization shows a clear bias toward greater ridership when there are rainy conditions. It is interesting to note that as we go right, toward the higher ridership levels of 5000 riders per hour and greater, the difference in count between raining and not raining becomes less and less until it is almost imperceptible.

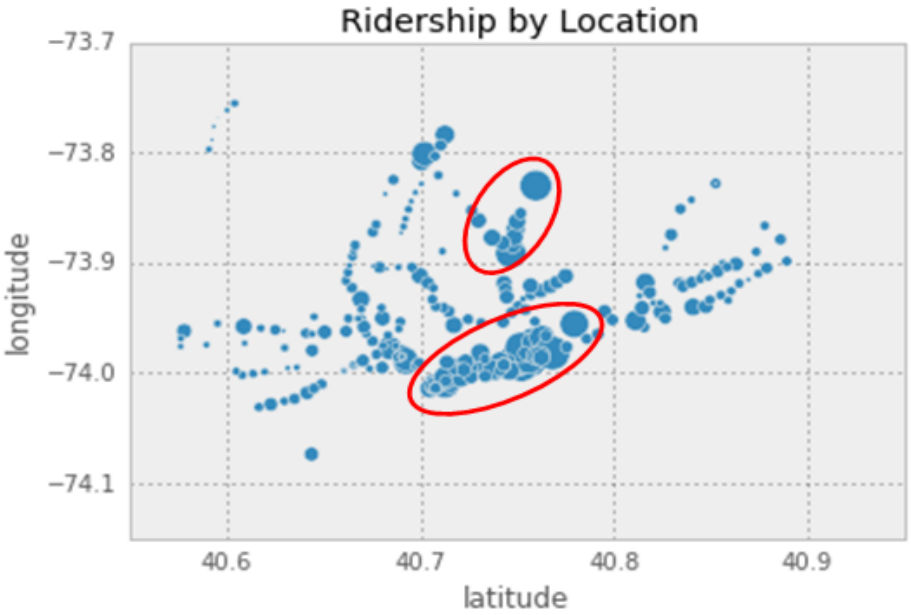


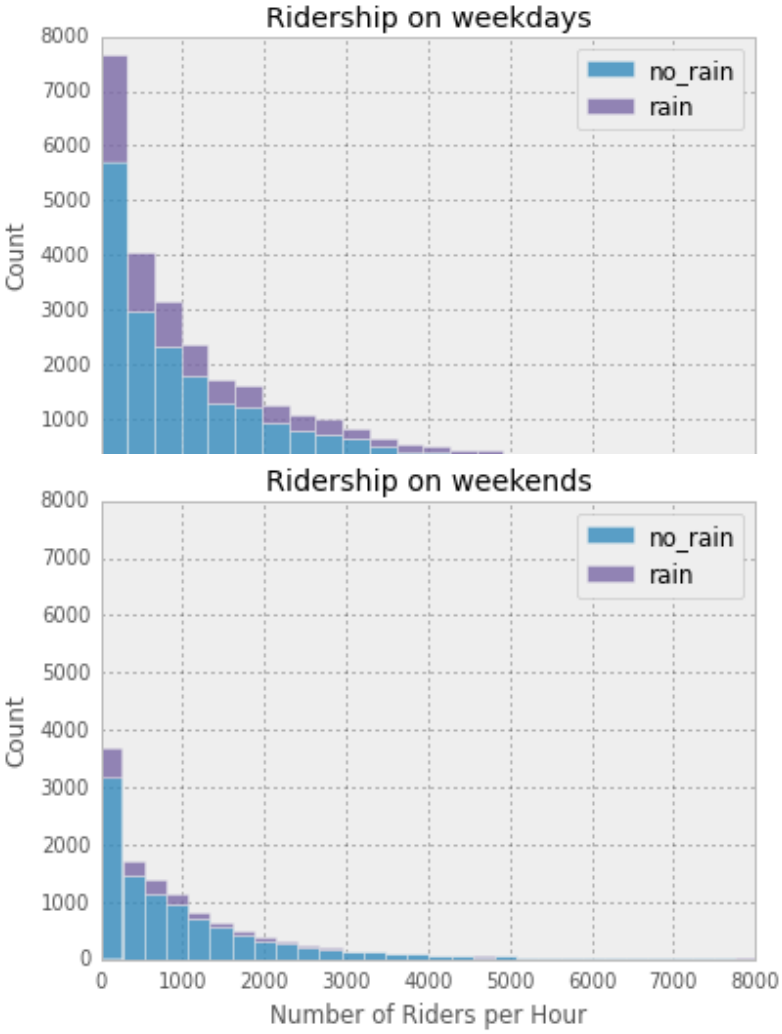
Figure 3: This visualization shows where each turnstile is by latitude and longitude. The size of each point is determined by the cumulative sum of riders for that unit over the month. This visualization demonstrates a significant “clumping,” wherein the bulk of the total riders fall into a few small regions near the center of the map and there are numerous outlying stations with low numbers of riders. A nice feature - that I could not figure out how to include - would be to overlay this plot on top of a map of NYC.

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?

The short answer is yes, more people ride the subway when it is raining than when it is not raining. However, the increase in ridership is not evenly distributed. By referring back to Figure 1, we can see that the increase in ridership when raining primarily manifests itself when the number of riders per hour is 5000 or less. For high traffic numbers - greater than 5000 riders per hour - the increase in riders is almost imperceptible. The increase in traffic is most obvious for the ridership levels of [0, 333], [334, 666], and [667, 1000] per hour. These lower traffic numbers see increases of 15-20% when it rains.

Out of curiosity, I segmented the ridership by weekend and weekday and plotted the histograms of ridership levels in order to see if the ridership difference between raining and non-rainy days was as pronounced on the weekend as it was on the weekdays.



On weekdays, the increase in ridership is about 20-25% for the lower numbers of riders, with the higher numbers of riders seeing relatively little increase in frequency.

On weekends, the increase in ridership is far less, about 10-15% for the lower numbers of riders.

From a business perspective this is useful information because it means that no additional capacity needs to be added to the NYC subway system. If the subway can already handle a high volume of riders, then there is no reason to expect a major increase in the frequency of high volume periods when it is raining. The only increase is in low volume periods, which the subway can already handle.

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 two previous histograms do a good job of showing that an interesting difference is present between the weekend and weekday ridership levels when raining. To test these difference for statistical significance, I again used the Mann-Whitney U test.

The tests show that there is a clear difference, both tests are statistically significant with p-values = 0.0, but it is interesting to note that the ridership difference when raining is far more significant during the weekday than it is during the weekend.

Here are the respective test statistics and some useful descriptive statistics as well:

```
For Weekdays the Mann-Whitney U test results are:
U stat = 175495347.500000 and p-value = 0.000000
The mean no_rain = 2227.961266 and the mean rain = 2133.569694
The median no_rain = 1042.500000 and the median rain = 1028.500000
```

```
For Weekends the Mann-Whitney U test results are:
U stat = 13258790.500000 and p-value = 0.000000
The mean no_rain = 1091.611276 and the mean rain = 1226.057557
The median no_rain = 612.000000 and the median rain = 674.000000
```

The U-stat of 175,465,347 for the weekdays is approximately 13 times bigger than the U-stat of 13,258,790 for weekends. These tests merely confirm our observations from the histograms - that there is an increase in ridership when it is raining, manifesting in a greater frequency of low numbers of riders, and that the difference is more pronounced on the weekdays than the weekends.

Section 5. Reflection

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

Limitations of the Dataset: I would have liked to see data over a longer period of time. If there are trends, then they would definitely be more pronounced over a period of several months than they are with just a single month. This would have given us more samples of days when it was raining. Of the 42,649 datapoints, 9,585 were for when conditions were rainy. Additionally, the dataset did not include information on whether or not a certain turnstile unit was operational or not. There might be periods when some units were closed and others were open, or maybe they are down for maintenance.

Limitations of the Analysis: One limitation of my analysis is that I did not check for bad data. For example, I could have checked for NaN values. Looking back, I could have devised some sort of checks that looked for NaN values, and filled them in or discarded the data. Also, there may have been erroneous values, either values that were absurdly high or unusually low. Perhaps, a particular turnstile failed to reset its entry counter, and it was incorrectly recording data. These potential outliers could have skewed my analysis.

There are two other features that I think warranted more investigation and possible inclusion in my model. The first feature is the day of the week. I used whether a day was a weekday or not as a feature that was successful at improving r-squared, however, I neglected to investigate if being a given day - such as Monday or Friday - had some additional predictive power over being merely a weekday. The other feature that I wanted to experiment more with was using the turnstile distance from a weather station. I think that I can create some sort of model that would scale or combine weather readings to get a better guess of the actual weather at a given turnstile. This may have enabled the use of other weather variables like fog and windspeed to obtain a better r-squared.

Another benefit to having data over a longer time period would be that it would help dampen the effects of surges from holidays and summer tourism. I also neglected to check for special events like major sporting events, holidays (Memorial Day), etc. These could have certainly made a difference in my analysis.

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

I do not have any other insights to contribute.

Thanks for making a great course, I'm learning a bunch and having fun! Please let me know if you need me to clarify anything in my report or supply any substantiating code or explanations.

Email: drmichaelallenhood@gmail.com