David Broadwater

Udacity Data Analyst Nanodegree Cohort 2

Project 1

Short Questions to Analyzing the NYC Subway Dataset

Analyzing the NYC Subway Dataset: Short Questions

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course. This document addresses part 2 of the project.

All of the results presented here were calculated using the Udacity IDE and subway dataset (versus the “improved dataset” available for local use).

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?

We used the Mann-Whitney U test, as implemented in SciPy. In SciPy, this meant the default P-value calculated was a one-tail value, so I doubled it to produce the standard two-tail value. The null hypothesis for this case (using the Mann-Whitney U test) is that there is no difference in the distribution of subway ridership values on days it rained compared to days that it did not rain. Since the alternative to the null hypothesis could either be higher or lower ridership values on rainy days versus non-rainy days, a two-tailed test is appropriate. Our p-critical value for this case is the usual value of 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.

This test is applicable to the data set because the Mann-Whitney U Test does not assume our data is from a particular probability distribution; this is good because it is easy to tell from a histogram that our data does not have a normal distribution. A Shapiro-Wilks test for normality produces a p-value of 0.0 for both populations (indicating non-normal distributions), but since this test may not be accurate for populations larger than 5000 (according to the warning displayed when running the test) we cannot rely on those results as our population sizes are 87847 entries without rain and 44104 entries with rain. The Mann-Whitney U test is also applicable here because it tests exactly what we are trying to determine: if either population is more likely to have higher values than the other.

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.

Performing the Mann-Whitney U test on this dataset produced a two-sided p-value of 0.049999825586979442 (just slightly below our p-critical value). The mean ENTRIESn_hourly value for ridership without rain was 1090.278780151855 and the mean value for ridership with rain was 1105.4463767458733.

1.4 What is the significance and interpretation of these results?

Since our p-value was below our p-critical threshold, the Mann-Whitney U test suggests one distribution is more likely to have higher values than the other. The difference in mean values for each sample confirms this, and seems to suggest that ridership increases on days with rain, since the mean values are slightly larger.

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:

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

For the results presented here, I implemented the OLS approach using 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 ‘rain’, ‘Hour’, ‘mintempi’, ‘maxtempi’, ‘meanwindspdi’,‘fog’, and ‘precipi’ as features for my model. I also used the “default” dummy variable ‘UNIT’ included in the “starter” code.

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.”

I used rain because that is part of the main relationship that we’re trying to test, and I suspected that rain would affect the amount of people riding the subway.

I used Hour because I thought that the time of day would have a large impact on the ridership totals (e.g., lower ridership during the late evening when people are mostly sleeping).

I used mintempi because I suspected ridership would increase (or at least be impacted) if it was especially cold outside, since people would probably be less likely to walk outside instead. This of course depends on the assumptions made on the order of preferred modes of transit; depending on availability and cost concerns, I could see taxi ridership increasing on cold days.

I used maxtempi for most of the same reasons I used mintempi, except assuming the opposite: that ridership would increase if it was warm outside.

I used meanwindspdi because I thought that if it was especially windy outside, people would be less likely to walk outside.

I used fog because I assumed if it was foggy outside people would be more likely to take the subway than walk.

I used precipi because I thought that the amount of rain might have an impact on how many people decided to take the subway (e.g., people might be less inclined to "tough it out" with an umbrella or raincoat if it was raining especially hard, as indicated by a large precipi value).

I used UNIT because it was included in the default code and because it made sense that the unit collecting the turnstile information (and the station, by proxy) could have an impact on the ridership. This was included as a dummy variable because it is a categorical feature, and we wanted to denote in our model if an entry belonged to a certain category or not.

In the end, all of the features I selected were ultimately included because they also resulted in a larger R2 value than the R2 value calculated without them, and they could be included without exceeding the 30 second server calculation limit.

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

```
Using StatsModels OLS, the coefficients are:
rain              10.046302
Hour              62.239983
mintempi         -15.976535
maxtempi           3.794069
meanwindspdi      33.752174
fog              274.059647
precipi          -81.110211
```

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

My calculated R2 value using StatsModels OLS was 0.485414650479.

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 R2 value indicates the goodness of fit is not particularly good. While it does seem to be able to predict ridership to some degree (since the R2 value is non-zero), it is still not close to our ideal R2 value of 1. Based on this result, I do not think this model is appropriate for predicting subway ridership; a more complex

model or different features are likely needed.

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.

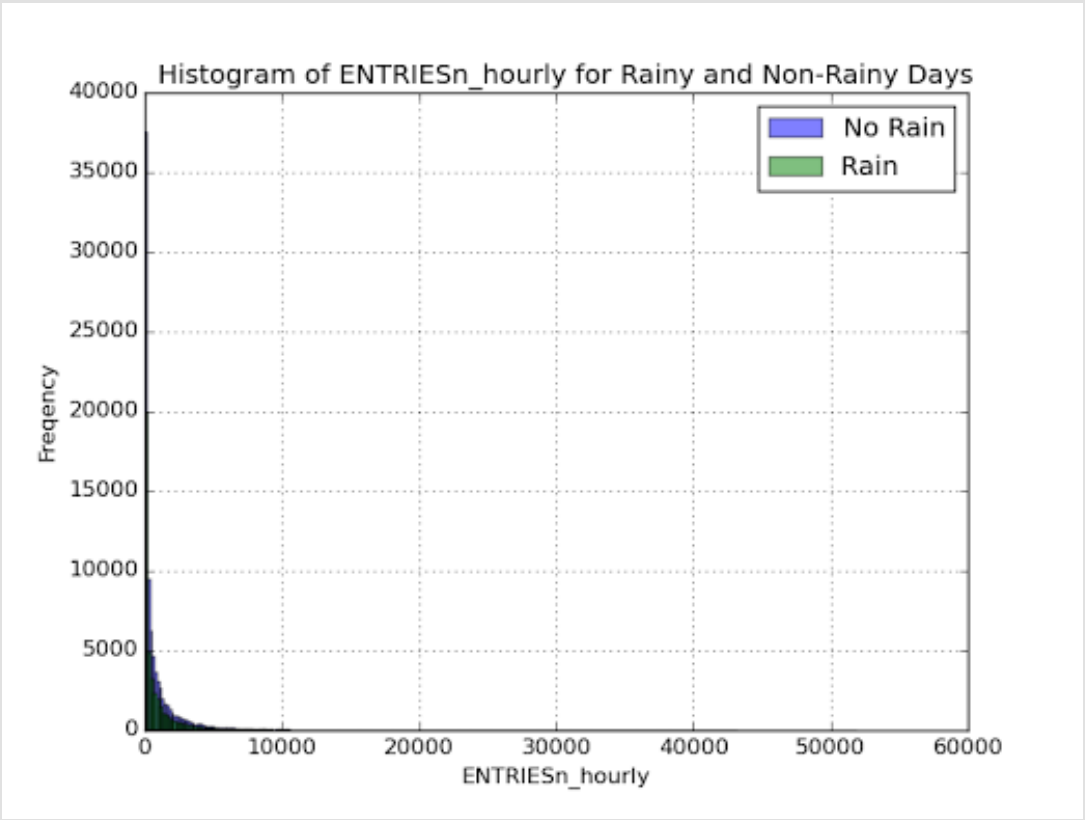
You can combine the two histograms in a single plot or you can use two separate plots. If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.

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.

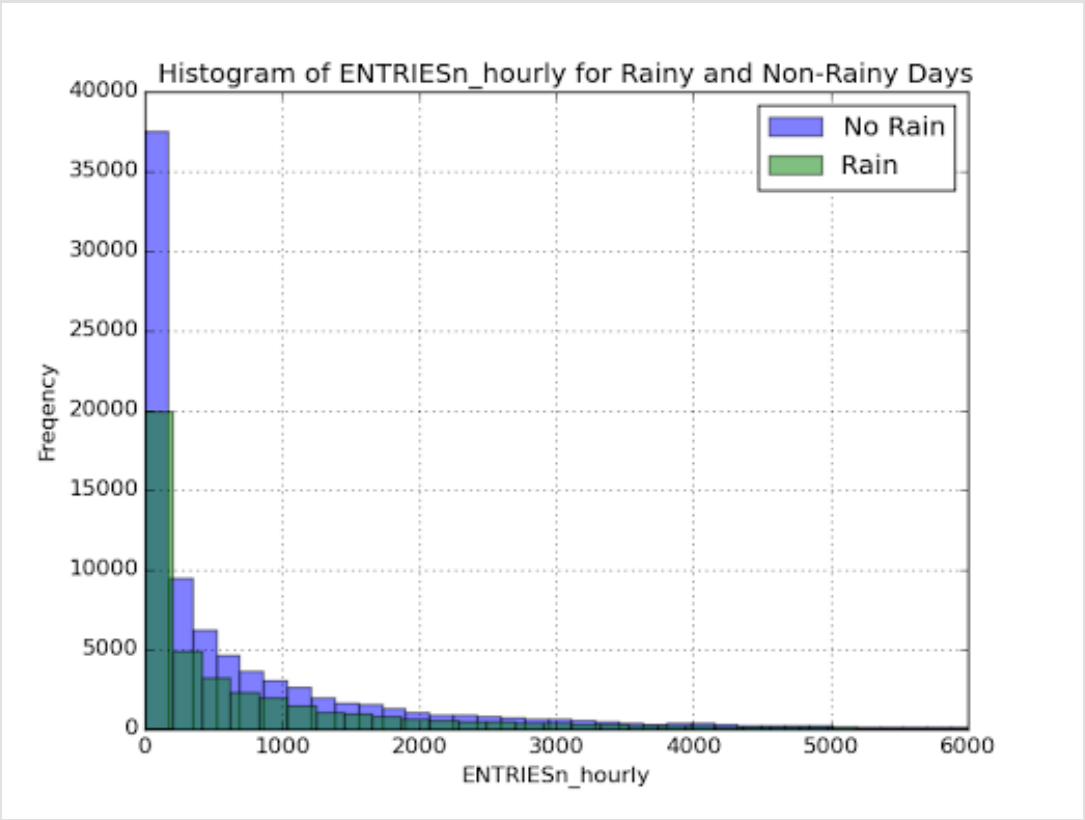
Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.

This figure shows the number of times each ENTRIESn_hourly value occurs within each population (days with rain and days without rain). A quick glance at the figure indicates that there are outliers in the populations, since the values that occur most often are all well below 10000, but our plot goes out to 60000. A calculation of the maximum values for each population reveals a maximum value of 51839 entries on

days with rain and a maximum value of 43199 on days without rain.



As suggested in Problem Set 3.1, I reduced the maximum x-axis value for the plot by a factor of 10 to 6000 in order to more clearly see the distribution of values. It is easy to see that the distributions are not normal, since the histograms do not have the “bell-curve” shape of a normal distribution. For both populations, the frequency of each ENTRIESn_hourly value decreases as the values themselves increase, with the smallest values occurring most often. In both cases, this is likely due to the frequency of the readings throughout the day (i.e., if the readings were taken less frequently, we would probably have higher frequencies of larger values).

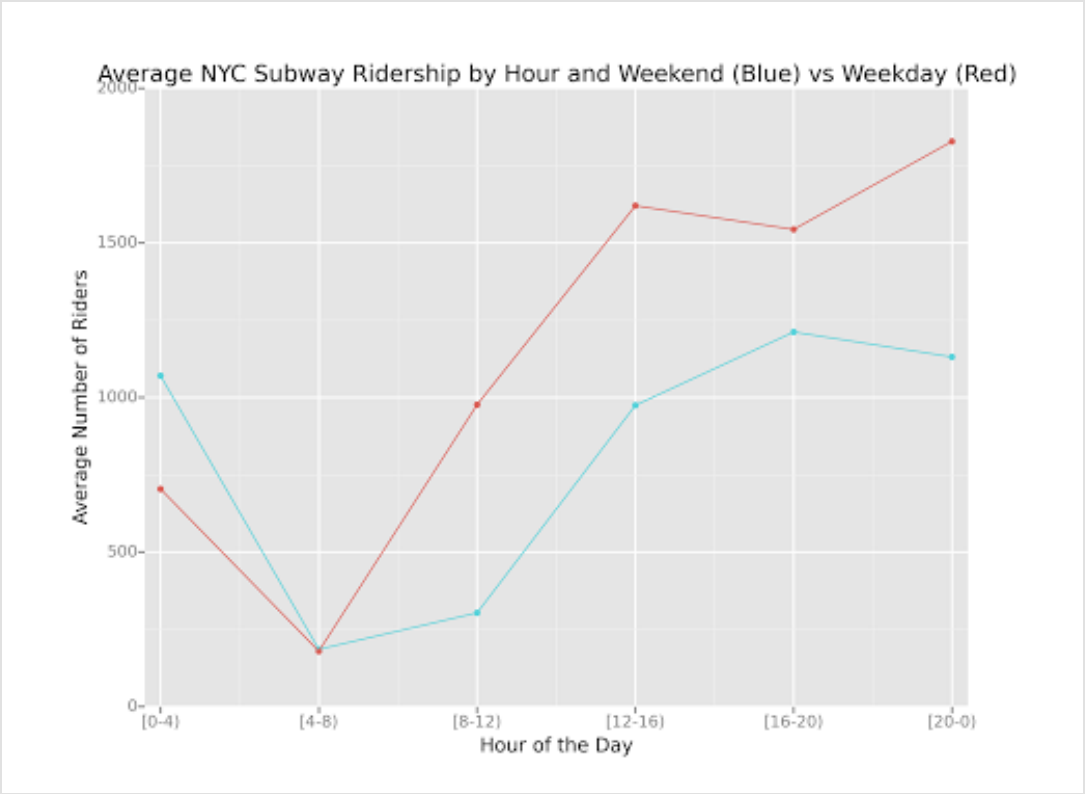


We can also note from the figure that entries on days with rain appear to occur less frequently than days without rain. This is confirmed in our data, with 87847 entries on days without rain and 44104 entries on days with rain. Given the data we have, this appears to indicate that in general there are more days without rain than with rain. Because the number of entries is smaller on days with rain, it also means that any outliers would have a larger effect on some of our statistics such as the mean, since there would be fewer “normal” entries to balance them out.

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

This figure shows subway ridership by time of day for weekdays and the weekend. First, I combined each of the entries into four-hour bins and then averaged them to help account for the variability in the hourly readings (mainly due to when the data happened to be recorded). I then split the binned ENTRIESn_hourly data into two groups: those recorded on a weekday (Monday-Friday) and those recorded on a weekend (Saturday-Sunday). It is important to note that these values correspond to the hour when the readings were taken, not necessarily when they actually occurred. The entries were mostly read every four hours, so some of the entries in a

reported value could have been from 3:59 hours earlier. Also note that the labels indicate half-open intervals for each bin (e.g., “[4–8)” indicates 4:00–7:59).



Looking at the weekday data, reported ridership entries peak during the hours of 8:00–11:59 PM. One possible explanation for this is people either going out to dinner or coming home from dinner and work at the end of the day, again since the readings taken during those time could include entries from the previous 3:59 hours. Weekday reported ridership is at its lowest during the hours of 4:00–7:59 AM, as expected (since that is probably when most people are sleeping). According to the NYC MTA website, the subway does actually run 24 hours a day. Reported ridership then increases with each time bin, when it dips slightly during the hours of 4:00–7:59 PM, likely corresponding to a slight dip during mid-afternoon (factoring in the temporal intervals in the readings), before the evening rush. Overall, reported ridership is higher during the evening hours than during the daytime. Some possible explanations for this are people going out to dinner or running errands after work further away from their normal commute area, or the fact that people may be less likely to walk instead of taking the subway when it is dark outside due to safety concerns.

While weekend reported ridership is lower than weekday reported ridership, much of the same trends hold. Reported weekend ridership is lowest during 4:00–7:59 AM, as on weekdays, and increases throughout the day. However, the peak average reported ridership occurs during the hours of 4:00–7:59 PM, compared to 8:00–11:59 PM on weekdays. This is likely due to the fact that most people don't work on weekends, so they can go out to eat or run errands earlier in the day. One other interesting insight is that average reported ridership is higher during the hours of 12:00–3:59 AM on weekends than on weekdays, likely due to an increase in nightlife and late-night activities on the weekend (particularly Saturday).

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?

From the data we have and methods described in this course, more people ride the subway when it is raining than when it is not raining. However, it does not appear to be a significant change, since the difference in average daily ridership on rainy days compared to non-rainy days is very small (roughly 15 entries, an increase of 1.39% in average daily ridership compared to non-rainy days). This could be due to the fact that people are pretty set in their daily transportation routine, regardless of the weather outside. It also could be the result of differences in the way rain affects the NYC population. For instance, someone who is not cost-averse to taking a taxi might choose to take a taxi of walking to a subway station in the rain. Conversely, someone who normally walks (but would consider a taxi cost-prohibitive), might choose to walk part of the way in the rain to a subway station to get to their destination.

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.

There are a few different results that support the assertion that more people ride the train when it is raining than when it is not raining. First, the Mann-Whitney U Test

results indicate that we're more likely to see higher values in one of the populations than the other. Taking a look at the mean, median, and maximum ENTRIESn_hourly values for each shows larger values for the ridership population when it is raining than when it is not:

```
Median
Rain:           282.0
Without Rain:   278.0

Mean
Rain:           1105.446376745873
Without Rain:   1090.278780151855

Max
Rain:           51839.0
Without Rain:   43199.0
```

Each of these results support the conclusion that there is an increase in ridership on rainy days. The smaller relative population size of the ridership on rainy days means it is more susceptible to the effects of outliers on our descriptive statistics than the ridership on non-rainy days, but that doesn't necessarily mean that those outliers are not valid values. However, identifying and examining those outliers using the interquartile range, a box plot, and additional background information about those particular readings would be prudent in a real-world setting (but is outside the scope of this course).

While our OLS linear model didn't have a particularly good R2 value, the coefficient it produced for rain as an input variable also suggested an increase in ridership on days with rain. This is because the coefficient was positive, indicating a positive linear relationship between ridership and whether or not it was raining on that given day.

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:

- Dataset
- Linear regression model
- Statistical test

First, the data could have been recorded better. While obviously all of the data wrangling was intentional and for educational purposes, it's always easier if the data already comes in the format we need to do our analysis. If we even would have had hourly readings, we could have come up with better insights overall about ridership trends, especially if we used the weather data to determine rain status on a smaller time scale than days. By grouping the populations based on days that it rained (or did not rain), we potentially included some ridership values that occurred when it was not raining outside at that particular time. For instance, if it only rained for an hour on a given day, all of the entries recorded on that day would be included in the "days with rain" population, even though most of the day it was not actually raining. In this case, the time the rain occurred during the day would also likely affect the potential impact on ridership (e.g., did the rain occur in the middle of the night when ridership is generally lower? Or did it occur during peak traffic in the evening?).

The linear regression model we used did not appear to be a great method for predicting subway ridership, given the relatively low R2 value we calculated for it. This suggests that perhaps subway ridership is too complex to accurately model linearly with the features we had and used. This is likely due to the limitations in our dataset noted earlier, as well as the interdependencies in some of the features. Using a more advanced machine learning algorithm on this dataset might produce better results.

Also, the mean and median values for both populations are very close to one another, suggesting the effect of rain might not be that significant. Given more time and information, a more thorough investigation of the potential outliers in the two populations would also give more insight.

Finally, more information about the other types of transportation in NYC (and how many people use them) would also provide more insight into our analysis. For instance, how popular is the bus system? Roughly how many people take taxis?

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

No, I've addressed everything I wanted to above.