Analyzing the NYC Subway Dataset

Sheng Weng

Section 1. Statistical Test

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?

   I used the Mann Whitney U-test to analyze the NYC subway data. The critical P value for this test would be 5%. I used a two-tail P value because we cannot predict before collecting data whether rain will be related with higher or lower ridership. The null hypothesis is that there is no difference between the number of entries on rainy days and non-rainy days.

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.

   Because the distribution of the dataset is non-normal, and the Mann Whitney U-test is a non-parametric test that could be applied to non-normal data.

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.

   The two-tail p-value is 0.024999912793489721*2 ≈ 0.0499, which is smaller than the critical value of 0.05. The means for rainy days and non-rainy days are: 1105.4463767458733, 1090.278780151855

4. What is the significance and interpretation of these results?

   We can see that the means are different between the rides on rainy days and non-rainy days. Because the resulting p-value 0.0499 is smaller than the critical value of 0.05, we can also infer that the null hypothesis is not true. So there must be statistical difference between the number of entries on rainy days and non-rainy days.

Section 2. Linear Regression

1. What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

I used the gradient descent.

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: 'rain', 'precipi', 'Hour', 'mintempi', ‘fog’. I used ‘UNIT’ as a dummy variable.

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 R^2 value.”

   I chose rain and precipi for the same reason, because I thought people might intend to use the subway more often when it’s rainy outside. I chose hour because I thought the number of riders largely depends on specific times of a day. I used feature fog because I thought bad weather would make people ride subway more often. I used mintempi because if it’s very cold outside, riding a subway would be a good choice for transportation.

4. What is your model’s R^2 (coefficients of determination) value?

   It’s 0.464905540741.

5. 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?

   After evaluating the residual plots, we are assured that the regression model is valid. The R^2 value is larger than 0.2, which means that the regression model is reasonable to predict ridership. But there must be a better model to get more symmetrical spread residuals, at the same time achieve R^2 value as close as one so that the predictions would be more accurate.

Section 3. Visualization

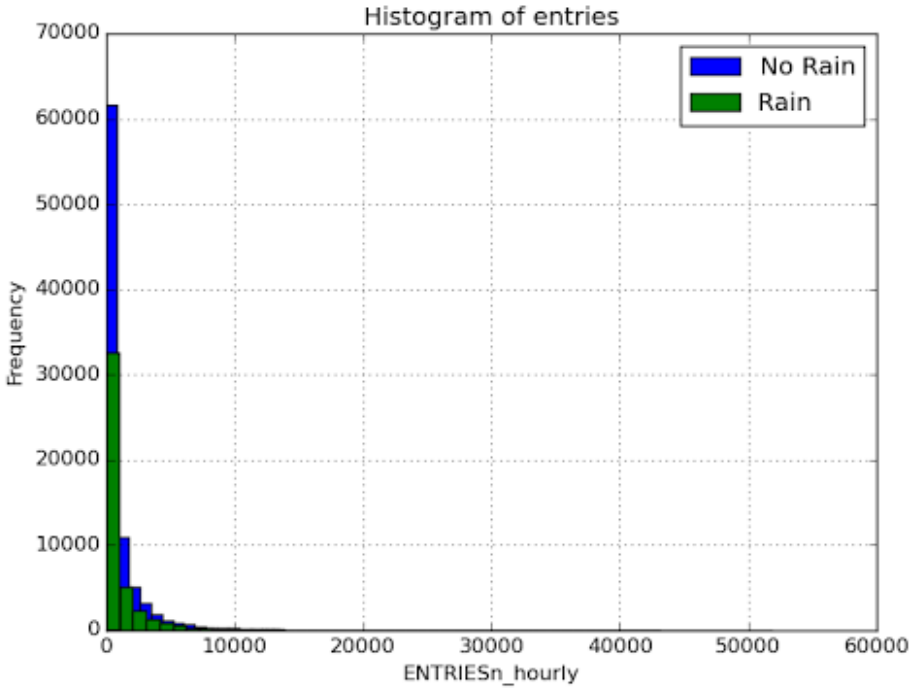
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

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.

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 different plots.

   For the histogram, 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, you might have one interval (along the x-axis) with values from 0 to 1000. The height of the bar for this interval will then represent the number of records (rows in our data) that have ENTRIESn_hourly that fall into 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.

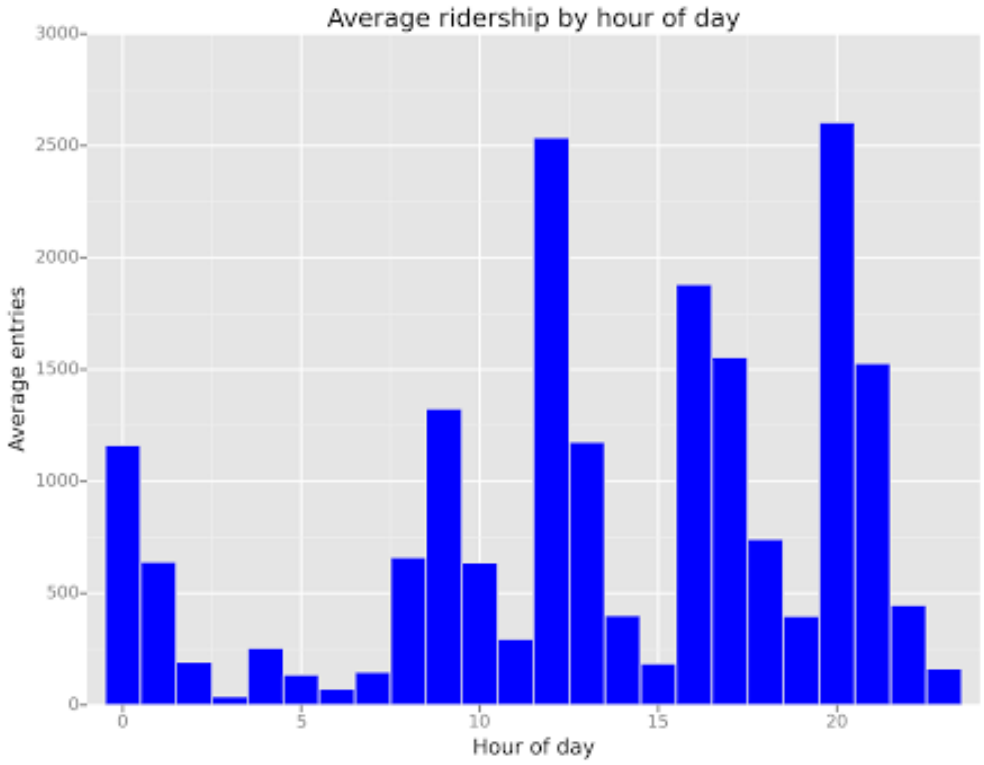


As we can see, although the trend and distribution of the two histograms look similar, the one for ‘no rain’ has much larger frequency when the ENTRIESn_hourly is relatively small. The overall frequency for ‘no rain’ at different ENTRIESn_hourly are also larger than that of for ‘rain’. This is due to fewer samples for ‘rain’.

2. One visualization can be more freeform, some suggestions are:

   a. Ridership by time-of-day or day-of-week

   b. Which stations have more exits or entries at different times of day



The second plot shows the distribution of ridership for each hour, which is represented by the average ridership bar at different time of day. There are two obvious peaks of ridership that happen at noon and 8pm. I think this may due to people are off for lunch and dinner.

Section 4. Conclusion

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

1. From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining versus when it is not raining?

   Yes, according to the result there exists statistical difference between the number of riders on rainy days and non-rainy days. The mean for rainy days is larger than non-rainy days, so more people would ride the NYC subway when it's raining.

2. What analyses lead you to this conclusion?

   First, the mean of the rides on rainy days is larger than that of on non-rainy days. Second, we use Mann Whitney U-test to test the null hypothesis of equal number of ridership for rainy days and non-rainy days. The resulting p-value 0.0499 is smaller than the critical value of 0.05, which indicates that there must be statistical difference between the number of ridership on rainy days and non-rainy days.

Section 5. Reflection

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

1. Please discuss potential shortcomings of the data set and the methods of your analysis.

   First, we need much more data to make the conclusion more convincing because right now we only have the ridership data collected in May. Second, the time of recording data is not consistent. We'd better record the data at the same time everyday, for example, every half an hour.

   In terms of the analyzing methods, I think there are more factors other than weather that can affect people's decision on whether to use the subway or not. Sometimes people have to ride a subway because the road is closed. Sometimes there might be more visitors riding the subway when it's holiday season. From the dataset we also don't know if it's raining all-day or just for a short while. If it's not raining in rush hours it may have little impact on the number of entries.

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

I think one thing that may improve the quality of the dataset is that it would be better to know the percentage of the number of ridership in the number of people who actually go outside on a specific day. Because on a rainy day people tend to stay at home, so the total number of people who are outside might be smaller than that on a sunny day. However, among those who have to go outside a large proportion of them may choose to ride the subway.