Analyzing the NYC Subway Dataset

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. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

**Section 0. References**

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

**Section 1. Statistical Test**

* 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 u-test for the subway data. At first, I used the default one-tailed test and compared it to a pre-established p-critical value of 0.05. The null hypothesis was that there was no difference in ridership between the two populations. My p-critical value was 0.05. Since the one-tailed P-value came out to 0.025, I first thought that we could reasonably conclude that the null hypothesis is not true.

However, upon reflection, I now see that using a one-tailed test wasn’t the correct approach because, in this context, the one-tailed test relies on the assumption that ridership would never be greater when there is no rain, at worst it would be equal. We can’t reasonably make this assumption. Doubling the p-value to represent a two-tailed test gives p=0.05, which is equal to p-critical, and therefore we can’t reject the null hypothesis based on this test.

* 1. 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.

We had to use the mann-whitney test rather than a t-test because we could not assume a normal (or any other) particular type of distribution for the ridership by hour for either the ran- or no-rain case.

* 1. 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.

I got:

With-rain mean: 1105.4463767458733

Without-rain mean: 1090.278780151855

U: 1924409167.0

P: 0.024999912793489721)

* 1. What is the significance and interpretation of these results?

Mean ridership with rain is 1105, which is ( (1105-1090) / 1090) \* 100 = 1.37% higher than the ridership without rain. The U and P-values tell us that this difference is not statistically meaningful. Since the two-sided p-value of 0.05 is not less than the pre-established p-critical value of 0.05, we cannot reasonably be sure that the null hypothesis (populations are the same) is false.

**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:

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

I used Gradient descent, as implemented in exerceise 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? I did not use any dummy variables. I used: rain, precipitation, hour, and mean temperature.

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 selected these features based mostly on intuition, but I added them one by one until the model began to work. I figured that mean temperature would have a strong influence because everyone likes to get out of the cold. I selected rain and precipitation for similar reasons and hour thinking that there would be noticeable peak times. I selected hour because I suspected that there would be peak load times.

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

2.92 14.65 467.70 -62.21

Rain – Precip. - Hour - Mean Temp

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

I got: r^2 value is 0.463968815042

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?

R^2 of 46% means that 46% of the variance is explained by this model. I don’t think this was the best model since we’re not explaining a lot of the variance.

**Section 3. Visualization**

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

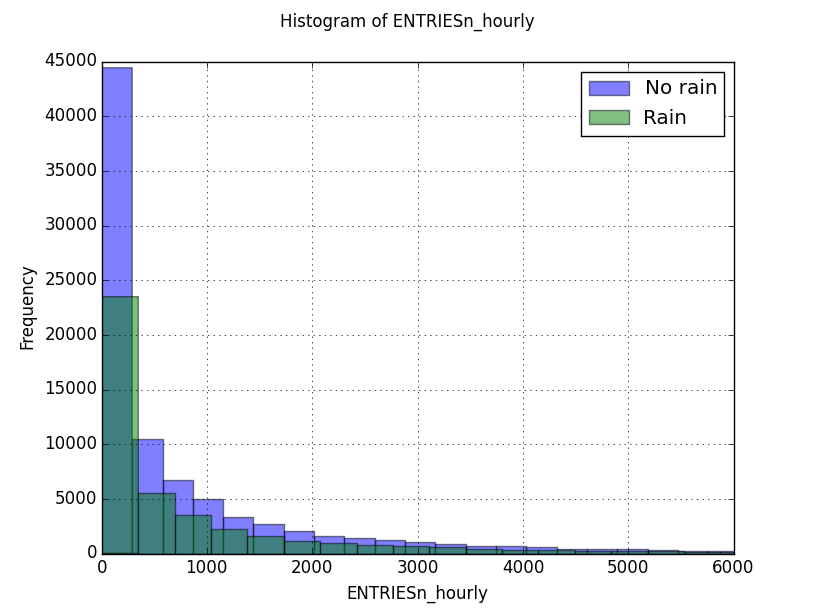
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

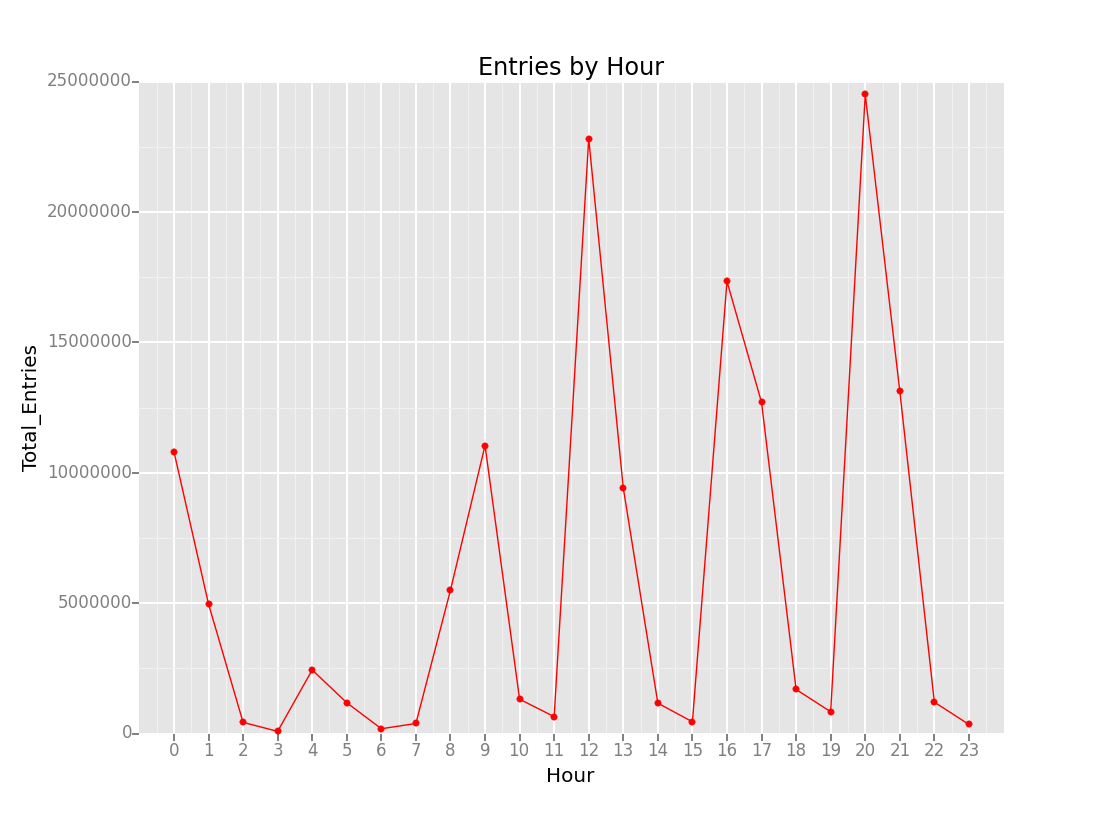
3.1 One visualization should contain two histograms: one of  ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days.

* 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.

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. Some suggestions are:

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

As far as entries per hour of rain VS non-rain, these are not normal distributions. They look more like one-sided gausians. Since only the month of May was considered, and May is a rainy month in New York, there appear to be more rainy day samples than non-rainy day sample.



I took the recommendation from my feedback and plotted entries per hour. There are huge outsized peaks at 12-noon, 4pm, and 8pm. Thinking about it now, these peaks are probably still there whether it rains or not. These huge numbers of riders, possibly commuting and therefore not likely to choose to or to not ride based on rain, might well render the rest of the rain / no rain analysis meaningless.

Looking at the above plot now, in conjunction with the Gradient descent coefficient for hour, which was 467 compared to the rain-no rain coefficient of 2.92, even if you normalize the scale of hour (say, divide it by 23, so it varies from 0 to 1), it would be a coefficient of 20, meaning that the hour of the day predicts ridership much much more powerfully (maybe 10x more powerfully) than rain.

**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?

Our data analysis does not positively indicate whether more people ride the subway when it’s raining or not. Technically, the analysis indicates that an equal number of people ride the subway when it is raining or not raining. I would caution that this analysis might only apply to the month of May, which are the days our samples are drawn from. Ridership in the winter months, for example, when it is often brutally cold, might show more variance based on precipitation or rain.

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 Mann-Whitney U test was used to compare the populations of Entry Counts when raining to Entry Counts when not raining. The test produced a two-tailed p value of 0.05. This is not a strong indication that the two populations are different.

In the same exercise, we computed the means of the two populations. The mean for ridership during rain was higher, but not by a huge margin. When we subjected the same data to Linear Regression and tried to establish predictive coefficients for some of the data on Entry Count, we found that both rain and precitpitation had positive theta coefficient values, when considered along with some other factors. This is a possible indication that rain can influence Subway ridership entries per hour but it is by no means conclusive. Many other factors, such as hour of entry, probably dwarf rain as a predictor.

**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.

As I mentioned above, I feel that the biggest weakness in our analysis is that all the samples are drawn from May. May has a lot of temperature and precipitation variation, which is good, but ridership stats in May based on weather may not be applicable to a month with less temperature and precipitation variations like January, for example, or August.

I think subway ridership is probably more complex than a linear test could show. Using a curve-fitting method that included polynomials of 2nd and maybe third order on some of the variables might ultimately give a better match. I’d also suggest that the particular month, or at least the season inquestion, should be one of those variables to consider.

Another thing to consider, particular to New York, is that many observant Jewish people might not ride the Subway during the Sabbath, from Friday at Sundown to Saturday at Sundown. Stations in heavily Jewish neighborhoods may see less ridership during these hours regardless of meterological conditions. If I were analyzing this again, I might consider only Monday through Thursday data.

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

So my viewpoint on whether rain did or did not affect subway ridership changed completely between this revision and my original submission. My insight from this is that I had a tendency to draw a hasty conclusion. I may also have been hasty in revising this conclusion. I hope one day to have a depth and breadth of experience with data science that can lead me towards better analysis quickly!