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, and 4 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.

http://pandas.pydata.org/pandas-docs/stable/10min.html

http://pandas.pydata.org/pandas-docs/stable/dsintro.html

http://docs.ggplot2.org/0.9.3.1/geom_bar.html

https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

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?

To analyze the data I used a Mann-Whitney U-Test. Since the data was not normally distributed a Welch’s T-Test would not work.
I used a two-tail test P value as I was testing to determine if NYC Subway ridership was influenced by rain.

The two populations we are comparing is the Number of Entries to the NYC when it is not raining compared to when it is
Null Hypothesis: The populations are equal when compared.
Alternate Hypothesis: The two populations are unequal. $\alpha = 0.05$
The p-critical value I used was 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.

Since the data is not normally distributed I could not use Welch's T-Test to analyze the data. So I used the Mann-Whitney U-test as it was a better fit based on the type of data we were comparing. For the Mann-Whitney U-test the data needs to come from independent groups and the independent variable (rain in this case) should be from two categorical groups. For the test the dependent variable can be continuous or ordinal and our data fit these needs. Additionally, as stated before, the data needs to be not normally distributed which is the case with this dataset.

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.

The statistical data of the Mann-Whitney U-Test for the subway data comparing number of entries for rain versus not rain is as follows:

With rain mean: 1105.446
Without rain mean: 1090.279
U Value: 1923309167
P value: 0.0249999

1.4 What is the significance and interpretation of these results?

Since I used the Mann-Whitney U-Test in python which provides a P value for a one sided test and the evaluation is for a two sided test the P value given needs to be doubled. This means the two-sided P value is 0.0499998 which means the criteria to be significant. Since the p value found is less than 0.05 and we will reject the null hypothesis. Based on this we can see that rain causes a statistically significant change in ridership though it is just within the significance range.

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. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

I chose to go with OLS using Statsmodles.

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 variables of rain, precipi, Hour, meantempi, maxpressurei, and meanpressurei
I used UNIT 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.”

I chose rain and precipi suspecting that not only if it were raining, but also how much it was raining would play a part in the tendency for people to choose to ride the subway. I chose Hour and the time of day would likely play a part in people’s decision making process. I selected meantempi and the mean temperature would likely be a factor for people deciding if walking in the rain would be acceptable to them. And I selected maxpressurei and meanpressurei because I suspected that they played a role in the changes of weather even though people are often unaware of the values and found that adding these to my model improved the results of my R2 value.

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

rain	-3.140027
precipi	32.495022
hour	65.356154
meantempi	-10.611657
maxpressurei	335.557511
meanpressurei	-653.501337

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

The R2 value for my model is 0.47959

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 R2 value of 0.47959 means that the linear regression model explains 47.959% of the original variability. This leaves 52.041% of the residual variability unexplained by the model. Since we are looking for fit for a model of human choice being close to 0.50 is actually fairly good. However, we can see that the model is far from a perfect predictor, likely due to the many other factors that would contribute to a person's decision to ride the subway.

Using this linear model to predict ridership is not the best fit considering the fairly low R2 value achieved. With so much residual variability unaccounted for it is hard to see that this model is a good fit to the data. The histograms show that residuals have long tails which lends itself to questioning the fit of this linear regression.

Section 3. Visualization

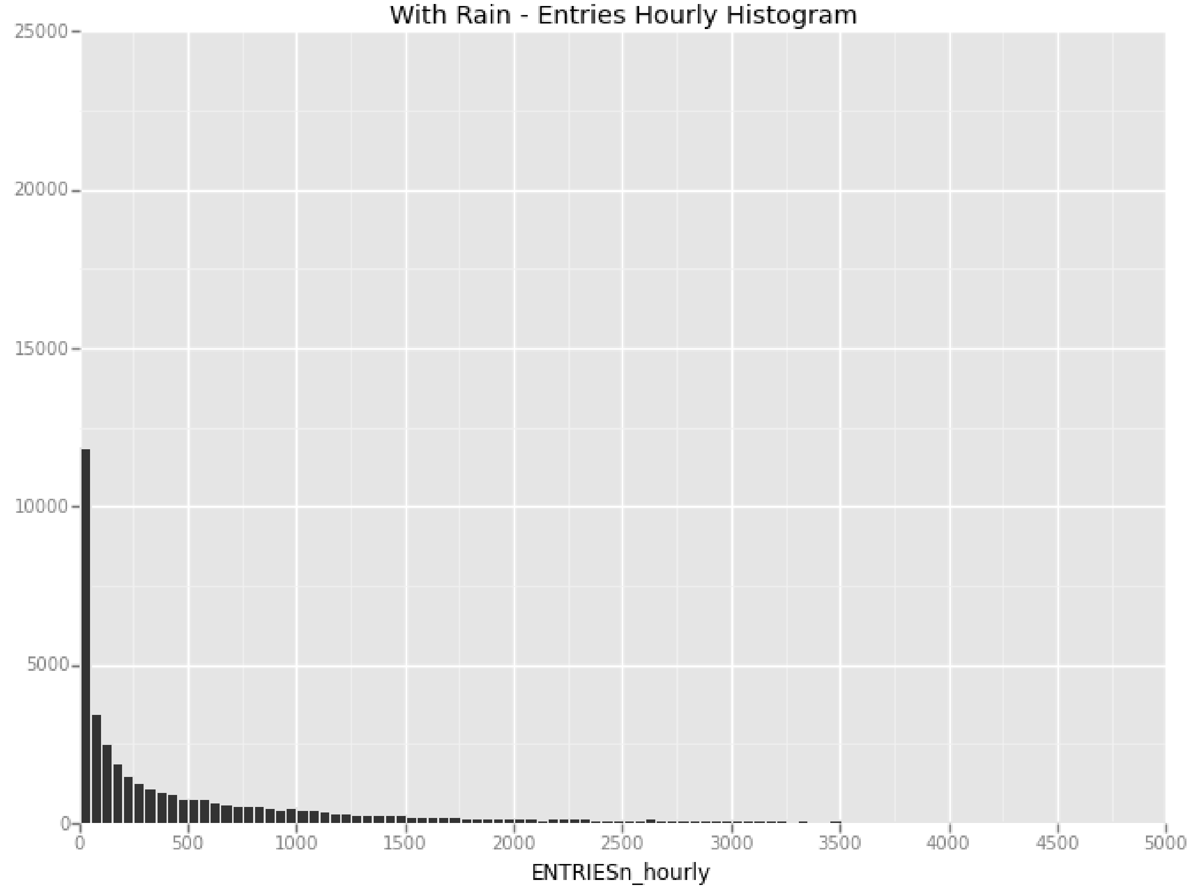
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.

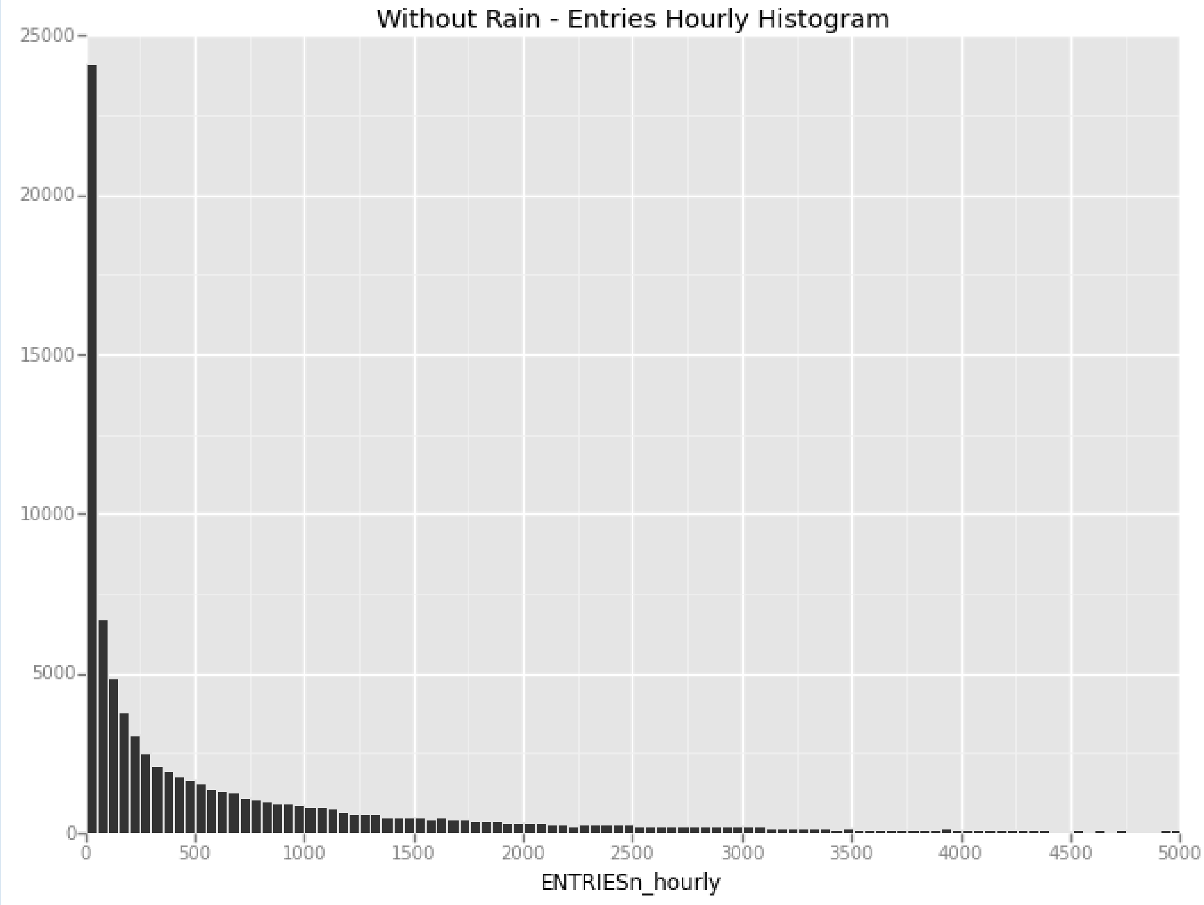
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- 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.



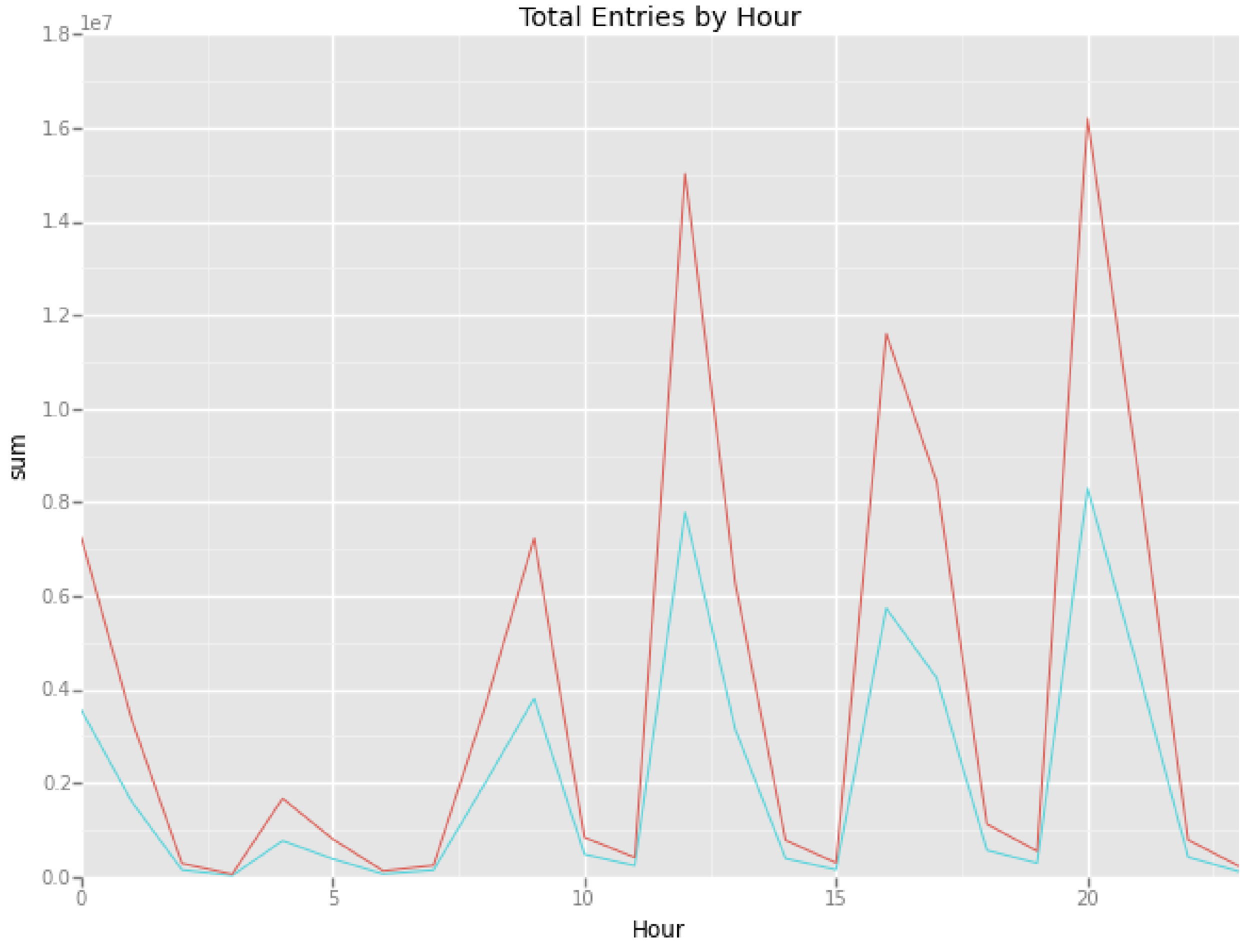
This histogram shows the number of times that a particular number of Entries can be found within the data. In this case I have separated the graphs between those when it was raining and when it is not. This particular graph represents the Entries during times when it was raining. As one would expect the count for the Entries goes down as the number of entries goes up, busy ridership days are not as common.



This histogram is like the one above but represents the non-raining days. The axes are the same as the previous graph to allow for direct comparison. There are more entries in this graph as the data has more non-rain days. As one would expect the count for the Entries goes down as the number of entries goes up, busy ridership days are not as common.

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



This graph shows total ridership by hour. The red lines represent the times of no-rain and the blue are those totals for rainy times. Since there were more non-rain times than rainy times there are more riders for the non-rain times. Note that the sums on the left are shown based on 1e7 so are larger numbers than it appears. It is notable that the peak times match up regardless of the weather conditions.

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?

> As a result of the analysis I conclude that more people ride the NYC subway when it is raining. While the overall ridership of the data is higher when comparing non-rainy total entries to rainy total entries this is a result of their being more clear days. Considering

that the number of rainy data points only accounts for 10% of the overall data, it is significant the amount of riders shown for wet days. This conclusion is backed up by the statistical analysis

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.

To perform the statistical analysis of the data I used both a Whitney-Mann U Test and a linear regression. The Whitney-Mann U Test resulted in a p value of 0.0249999 which is smaller than the critical p value of 0.025 since this is a one tailed test. With a value lower that our critical value we can reject the null hypothesis and conclude that rain plays a statistically significant role in the different mean values we see in the data. The mean for rainy data points is 1105.446 and for non-rainy data points are 1090.279.

The linear regression used considers the condition of rain, the hour of the day, the amount of precipitation, mean temperature, as well as max and mean pressure. Using these features I gained an R2 value of 0.479591051573. Since we are trying to gain a predictive model for human behavior we cannot expect to get as high an R2 value as we might for a more predictable system. Our R2 value is high enough to consider it predictive of the system.

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.

The dataset shows only certain times for each Unit and are not consistent from one to another which lends itself to be challenging to compare data well. Also there is no way to know how long someone rode on the train, which could be a useful side angle for the analysis; do people ride longer when it is raining. Perhaps something that considered if it were dark out or how hard it was raining. Additionally the dataset only covers one month, it would seem that a year’s worth of data would result in a different result. Also with many different variables in the dataset that are so closely related there could easily be confounding data making accurate predictions challenging.

As for the analysis the linear regression could likely be better if other factors were used. The linear regression for this model is not a strong fit with an R2 value of only 0.47959. With so much residual variability it leaves the system difficult to predict.

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