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.

[1] System-specific parameters and functions for python: https://docs.python.org/2/library/sys.html

[2] Windows command line: http://www.microsoft.com/resources/documentation/windows/xp/all/proddocs/en-us/ntcmds.mspx?mfr=true

[3] Windows redirection command: http://www.microsoft.com/resources/documentation/windows/xp/all/proddocs/en-us/redirection.mspx?mfr=true

[4] Python Logging: https://docs.python.org/2/howto/logging.html

[5] t-distribution wikipedia: http://en.wikipedia.org/wiki/Student%27s_t-distribution

[6] One and Two Sample t-tests: http://www.stat.columbia.edu/~martin/W2024/R2.pdf

[7] Two sample t-tests examples:
https://www.ma.utexas.edu/users/mks/statmistakes/2samplevs1sampletest.html

[8] T-test Wikipedia: http://en.wikipedia.org/wiki/Student%27s_t-test

[9] Welch's t-test: http://en.wikipedia.org/wiki/Welch%27s_t_test

[10] Sawilowsky, Shlomo S. (2005). "Misconceptions Leading to Choosing the t Test Over The Wilcoxon Mann–Whitney Test for Shift in Location Parameter". Journal of Modern Applied Statistical Methods 4 (2): 598–600.

[11] Mann-Whitney U Test:
http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html

[12] Matplotlib: http://matplotlib.org/users/legend_guide.html

[13] Histogram bin setting:
http://stackoverflow.com/questions/6986986/bin-size-in-matplotlib-histogram

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?

I used Mann-Whitney U test to analyze the NYC subway data.

I used two-tail p value, since I wanted to know if being rainy or non-rainy would have effects on the number of entries in NYC subway.

The null hypothesis is:

H_0: The number of entries of the day with rain and the day without rain are drawn from the same population.

Using a 95% confidence, the p-critical value is 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.

According to the histogram of the data with rain and without rain, we could see that the data are not normal.

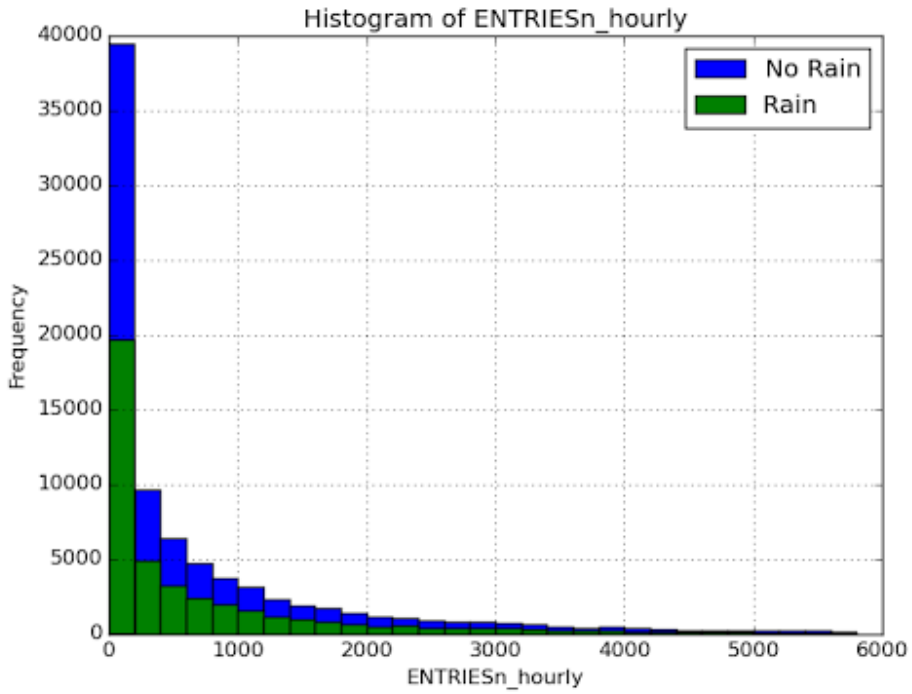

Figure 1 Histogram of data w/ or w/o rain

Mann-Whitney U test is a Non-parametric test, which doesn't assume normality in the data. Therefore it is applicable to this dataset. And according to some researches [10], we should this kind of non-parametric test under this condition to gain more statistical power.

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 mean for the number of entries with rain is 1105, while the mean for the number of entries without rain is 1090.

The one-tail p-value is 0.024999912793489721. So the two-tail p-value is 0.0499. Using a 95% confidence level, we conservatively reject the null hypothesis, and conclude that the number of entries in rainy days and non-rainy days are drawn from different populations.

1.4 What is the significance and interpretation of these results?

H_0: The number of entries between the day with rain and the day without rain are drawn from the same population.

H_a: The number of entries between the day with rain and the day without rain are drawn from different populations.

Since the data don't meet the normal assumption, Mann-Whitney U test could be applied to perform the hypothesis testing. From the Mann-Whitney U test, we have the test statistic as 1924409167.0 and one-tail p-value as 0.024999912793489721. So the two-tail p-value is 0.0499. With a 95% confidence level, we reject the null hypothesis, and conclude that the number of entries in rainy days and non-rainy days are drawn from different populations.

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:

a. Gradient descent (as implemented in exercise 3.5)

b. OLS using Statsmodels

c. Or something different?

a. Gradient descent

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

Features I used include: 'rain', 'Hour', 'day_week', 'weekday' and 'UNIT'.

Yes, I used 'UNIT' as dummy variables.

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 R^2 value."

Variable 'UNIT' is added to the model, since adding it could dramatically improve the R^2 value, which indicates the number of entries heavily depends on the location of the subway.

Variable 'Hour' is added to the model, since based on common sense, different time slots would have different level of ridership. Also, through experimentation, adding 'Hour' could greatly improve the R^2 value.

Variable 'rain' is added to the model, since based on intuition, people will use subway more often on unfavorable weather conditions. By the way, variable 'thunder' is not added, since its value is always 0. So it cannot provider useful information. Variable 'fog' is excluded, since it doesn't contribute much to the model. To avoid overfitting, it is excluded.

Variable 'day_week' and 'weekday' are added to the model, since based on intuition, people may choose different transportation methods on different days.

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

Hour: 855.609

rain: 16.816

day_week: 81.371

weekday: 490.035

2.5 What is your model's R^2 (coefficients of determination) value?

R^2 = 0.48219.

2.6 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?

This R^2 value means that the goodness of fit is fair.

R^2=0.48219 means that our model have explained about 48% of the original variability (in the values of number of entries), and are left with about 52% residual variability.

Given our application is to predicting the ridership of subway, for this R^2=0.48219, I think this model is appropriate, at least it has explained about half of the variability in the Y values. Using this model, we could make useful predictions.

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.

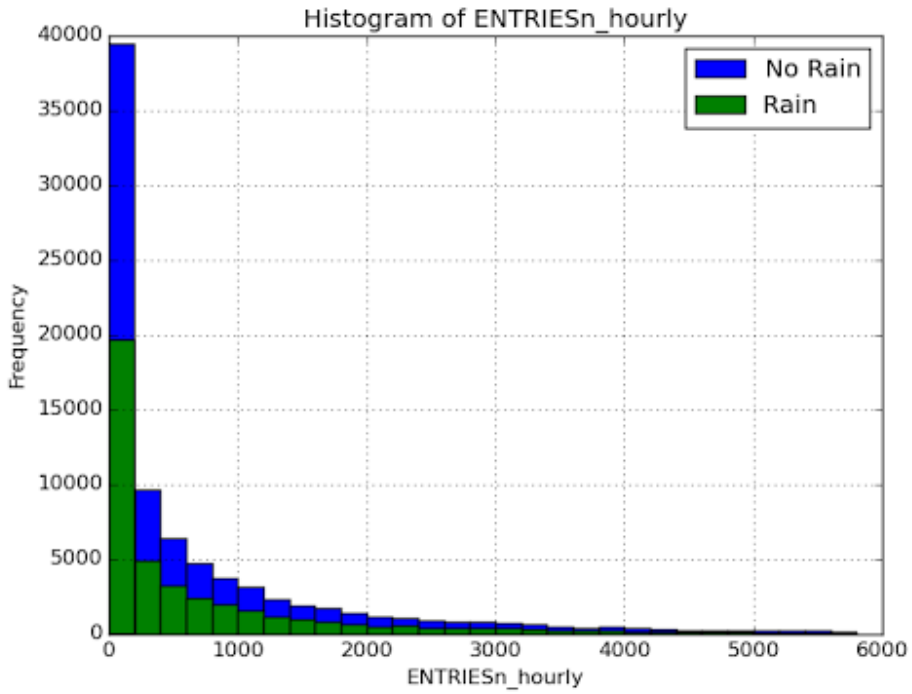

Figure 2. Histogram of Number of Entries w/ or w/o Rain

From the above figure, we could see that the both data for rainy and non-rainy days are right-skewed. We could use log transformation to first transform the data to be normally distributed. If we want to do hypothesis test on this dataset, we'd better use non-parametric tests.

We may also notice that for each number of entries, non-rainy days always have higher frequencies. However, we could conclude that non-rainy days has larger number of entries, since we may have more days without rain than with rain.

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

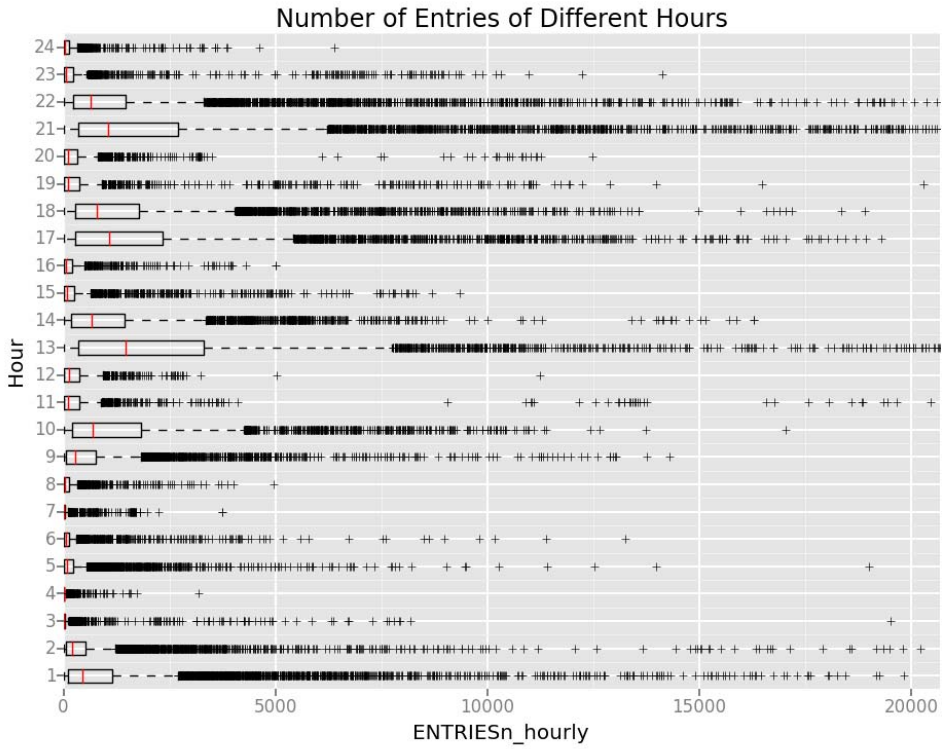

Figure 3

Using the “turnstile_data_master_with_weather.csv” dataset, the above boxplot is obtained.

From the above figure, we could tell that the average number of entries are much higher at time 10:00, 13:00, 14:00, 17:00, 18:00, and 21:00.

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?

According to the Mann-Whitney U test, I find that more people ride the NYC subway when it is raining.

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 one-tail p-value of the Mann-Whitney U test is 0.024999912793489721, and the two-tail p-value is 0.0499, which is less than 0.05. And the mean when it is raining is 1105, while the mean for the number of entries without rain is 1090. Although the difference is not dramatic, I think we could safely conclude that more people ride the NYC subway when raining.

One the other hand, from the result of linear regression, the coefficient for rain is 16.816, which means if it is raining, the number of people riding NYC subway will increase by 17 or 18 people. So linear regression reinforces the conclusion that when raining more people will ride the NYC subway.

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 dates in the datasets (either the original one or the enhanced one) are from 05/01/2011 to 05/30/2011. It is very difficult to draw accurate conclusions given such a short period. Moreover, we have several conditions of weather: rain, fog, thunder, etc. But we do not have any data with thunder (variable thunder always being 0). So one possible improvement is to collect more data in different months, e.g. January, March, August, October, December, etc.

Another possible improvement is to add several more feathers in the dataset. For example "in_holiday" and "out_holiday", where "in_holiday" stands for the holidays on which more people tend to ride the subway, such as New Year Eve, and "out_holiday" stands for the holidays on which people tend to leave the city, such as Thanksgiving.

Overall speaking, the p-value of the Mann-Whitney U test is not very small and the R^2 value of the linear regression is not very high, which indicate that our model or conclusion may not be very robust. We could derive the same conclusion by looking at the residual plot shown below. We can see from Figure 4 that there are many large errors outside the interval [-200,200]. Therefore, the prediction power of the linear model is limited.

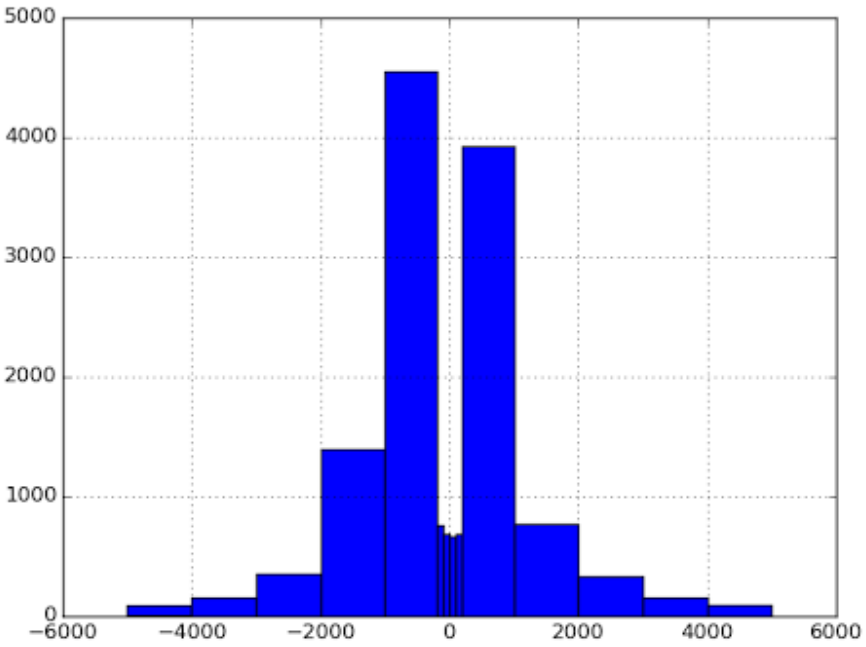

Figure 4

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