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.

# 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?

**Answer**: Mann-Whitney U test. Two-tail. The null hypothesis is that the two ridership samples are from the same population. My p-critical value is 0.1.

* 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.

**Answer**: From the below data visualization, I concluded that the data of the any one of two samples (ENTRIESn corresponding to rainy days and non-rainy days) is not normally distributed. However, the commonly used T-test needs the assumption of normal distribution and thus it is not applicable. Then I realized I could use non-parametric test and finally I chose Mann-Whitney U test which does not assume the data is normally distributed. Besides, the numbers of observations in the two samples are both much larger than the recommended minimal number, i.e. 20. So, based on these reasons, I concluded that Mann-Whitney U test is applicable to the dataset.

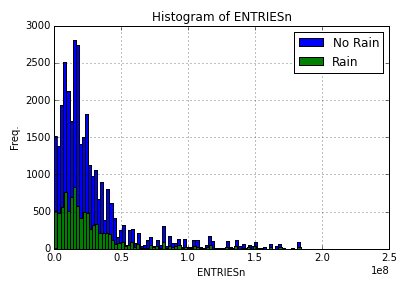


Figure 1

* 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.

**Answer**: The samples corresponding to raining and not-raining have means equal to 28158345.4741 and 28009333.3351, respectively. The obtained two-tail p-value is 0.541464814081.

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

**Answer**: The chosen significance level is 0.1. Since the obtained two-tailed p-value is 0.541464814081 which is smaller than the chosen significance level 0.1, the null hypothesis is rejected. In other words, the two ridership samples (corresponding to raining and not-raining) are statistically from different populations, at the significance level of 0.1.

# 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?

**Answer**: I used OLS 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?

**Answer**:

1. Features (input variables): 1) 'rain', 2) 'fog', 3) 'meanprecipi', 4) 'pressurei', 5) 'meantempi', 6) 'wspdi', 7) 'day\_week', 8) 'weekday'
2. Dummy variables: 1) 'UNIT', 2) 'conds'

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.”

**Answer**: The features I used can be grouped into 3 kinds as follows.

1. **Features about weather**: on the whole, I thought the willingness of people to take subway can be greatly influenced by the weather.
   1. Rain: From the above hypothesis test, I concluded rain takes affects on ridership. Besides, in my intuition, people might give up walking and turn to subway in raining days.
   2. Fog: Driving car on the road or walking outside in very foggy days could be inconvenient or even dangerous.
   3. meanprecipi, pressurei, meantempi, wspdi: when these weather conditions are very high or very low, the weather could be bad. They also influence people’s prediction of weather and thus people’s choices on transportation tools and activity plans. The reason of using meanprecipi instead of precipi is that meanprecipi can get me a higher R^2. Using mean or not for the other 3 features is because of similar reasons.
   4. 'conds': As what I said above, weather may affect people’s willingness to take subway, so I counted it in. While this feature is categorical thus hard to be used directly, I used it by using dummy variables.
2. **Features about location**: On the whole, I thought different location of the station and the turnstiles could lead to the difference passenger volume.
   1. ‘UNIT’: This feature is categorical thus hard to be used directly, I used it by using dummy variables.
3. **Features about time**: On the whole, passenger volume should be changing over the date.
   1. 'day\_week', 'weekday': On weekday, people might stay at home/office in most of time on that day, while they might like go outside on weekend.

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

**Answer**:

1) 'rain': -8.46628476e+05

2) 'fog: 2.04854828e+06

3) 'meanprecipi': 2.50831612e+07

4) 'pressurei': -3.18138545e+06

5) 'meantempi': 3.91747951e+04

6) 'wspdi': -1.61059177e+04

7) 'day\_week': -1.14047506e+05

8) 'weekday': -1.89845202e+05

Intercept is 109833804.443

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

**Answer**: 0.929129588989

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?

**Answer**: The closer R2 is to 1, the better the model, while the closer R2 is to 0, the poorer the model. The R2 for my model is 0.929129588989 which is very close to 1, so if judging purely from R2, the model is good enough. Let us look at the plot of residuals as shown below (the left figure). From the figure, we could roughly conclude that the errors are roughly normally distributed and that most of the error values are in range [-1e-7, 1e+7]. From the statistics of **absolute** value of residuals (the right figure), the mean is around 4.15e+6 and 75% is around 4.58e+6. Both of them are relative small compared to the mean of the real ridership which is around 2.8e+7. Based on the above analysis, I think the model is appropriate for the dataset.

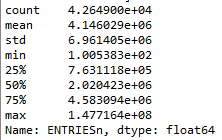
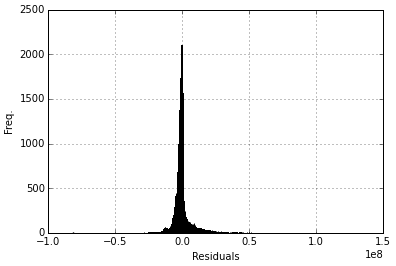


Figure 2 Figure 3

# 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.

**Answer**: The histogram I am going to show is actually the same with the one shown in Section 1.

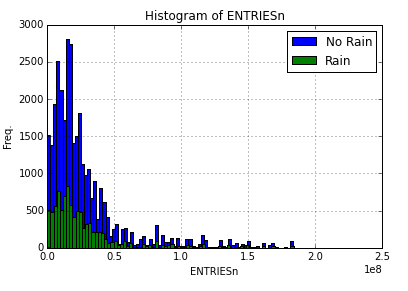


Figure 4

**Short description**: From the histogram, we could conclude that the ridership for rainy day or for non-rainy day is not normally distributed. While their distributions look similar. However, from the result of hypothesis test in Section 1, they are statistically from different populations, at the significance level of 0.1.

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

**Answer**: Below is the histogram plot of Ridership by day-of-week.

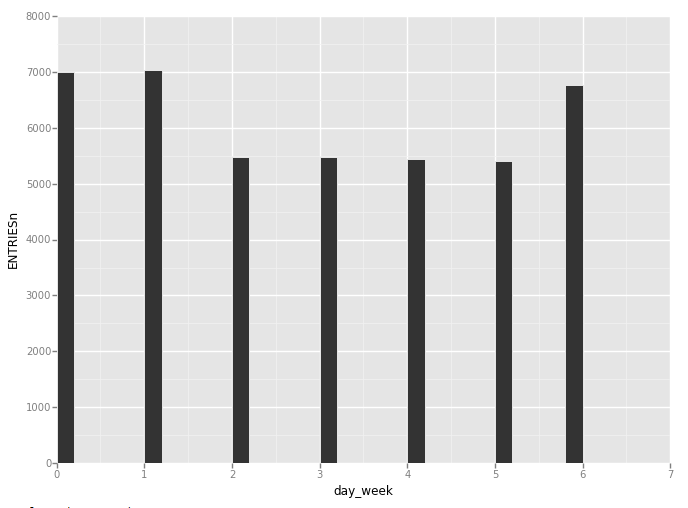


Figure 5

**Short description**: Judging from the ridership over day of week, NYC subway ridership is relative high on Monday, Tuesday, and Sunday, when compared to other day of week. Another observation is that the ridership is quite stable over Monday, Tuesday, and Sunday, and also quote stable over Wednesday to 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?

**Answer**: **More** people ride the NYC when it is **not** 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.

**Answer**: From hypothesis test result shown in Section 1, the two samples for rainy days and for non-rainy days are statistically from different populations, at the significance level of 0.1. Thus I concluded that rainy or not does have affects on the ridership volume. Then from the linear regression results, I obtained the coefficient of feature 'rain' is a negative value -8.46628476e+05, thus I concluded that for a certain day, 'rain'=0 leads to roughly around 8.46628476e+05 more ridership than when 'rain'=1, according to the negative coefficient.

# 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.

**Answer**:

1. Dataset: features can have relationships, and different features may be of the same aspect, so we need to reduce the features duplicated or strongly-related. PCA might be a method to do that to fine tune the data.
2. Statistical test: If the chosen significance level is 0.05, then the hypothesis test fails to reject the null hypothesis due to the obtained two-tailed p-value 0.54 is a little bigger than 0.05. So setting a significance level 0.05 leads to different final hypothesis test results compared to my chosen significance level 0.1.

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