Analyzing the NYC Subway Dataset

Notes: The answers are highlighted by grey shade.

Short 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 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?

Mann Whitney U-test is used here to analyze the NYC subway data.
The null hypothesis is 'the mean of entries to the subway station in rainy day is equal to that without rain.'
Two-tail P-value is used with p-critical value of 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.

Because Mann Whitney U-test is efficient to test if two non-normal populations are of same mean. From the exercise 3.1 figure, the histogram, it is easily to tell the populations are not normally distributed. So I didn't use Welch's t-test here (But it is valid to use here because they are large sample sizes.). Their shapes are similar in the histogram. So I use the Mann Whitney U-test here as my statistical test.

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 result:
P-value is 0.05.
Mean of ridership in rainy day is 1105.4.
Mean of ridership in non-rainy day is 1090.3.

1.4 What is the significance and interpretation of these results?

P-value = 0.05, which is equal to our p-critical value of 0.05. We reject the null hypothesis at the 95% confidence level that the mean of entries to the subway station in rainy day is equal to that without rain. This indicates the means of these two samples

are significantly different. We could interpret this result as ridership of subway in NYC is significantly influenced by weather of rain.

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?

OLS is used here to compute the coefficients theta and produce prediction for ENTRIESn_hourly in my regression model.

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

The features that I used in my model are 'rain', 'fog', 'Hour', 'meantempi', 'UNIT'. 'UNIT' is transferred to dummy variables as part of the features.

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."

Features in my model are chosen for reasons listed below.
'rain': I used 'rain' as a feature because driving and walking are more difficult in rainy day and part of these people would choose taking the subway.
'fog': I used 'fog' as a feature because more people would choose taking the subway in foggy day than normal day.
'Hour': I used 'Hour' as a feature because there is relatively stable peak hour for the public transportation.
'meantempi': I used as a feature because people who usually walk outside may choose taking the subway if the temperature is too low or too high.
'UNIT': 'UNIT' is also an important feature because the ridership is also relevant to the location of the station.

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

‘rain’: -39.5.
‘fog’: 218.1.
‘Hour’: 62.3.
‘meantempi’: -13.0971.

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

The value of R2 is 0.484.

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?

This R2 value means that the variability of subway ridership in NYC are 48.4% explained by variability of features of ‘rain’, ‘fog’, ‘Hour’, ‘meantempi’ and ‘UNIT’ in my regression model and are left with 51.6% residual variability.

As this R2 value is not high, I don’t think this linear model to predict ridership is appropriate for this dataset. We need to find more features that are related to the ridership of subway.

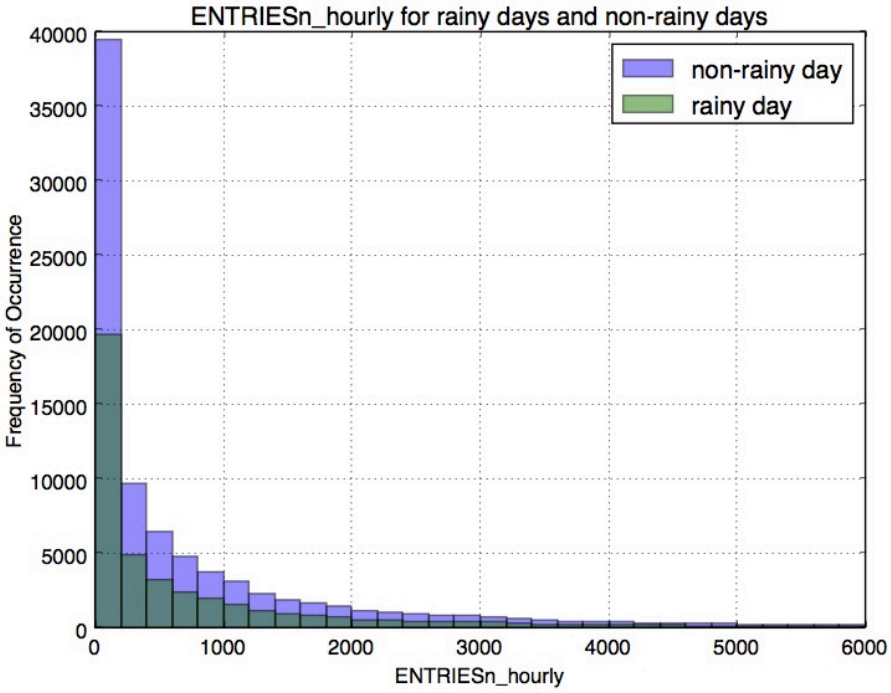
Section 3. Visualization
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.

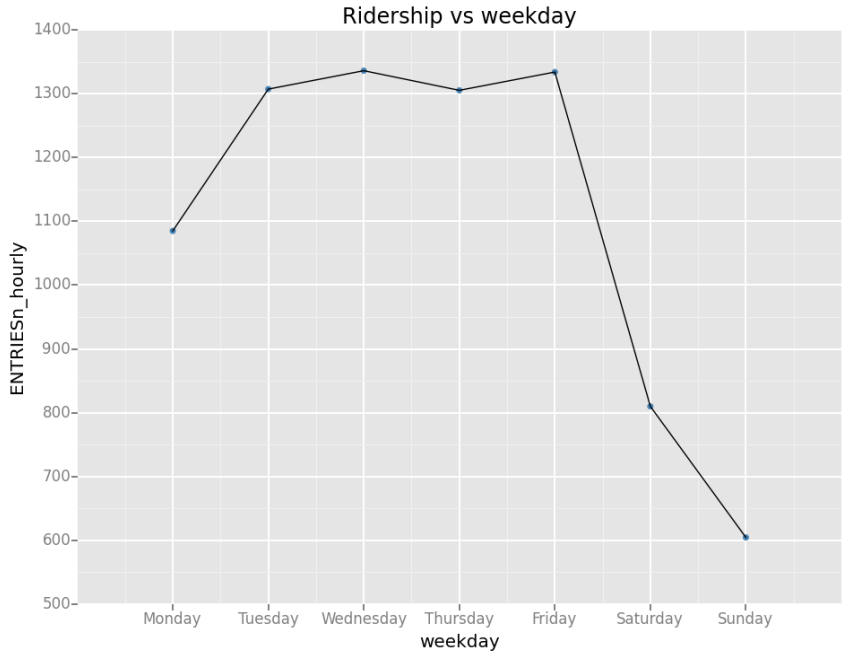
From this figure we could find that the distributions for subway ridership in non-rainy days and rainy days are similar.

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:

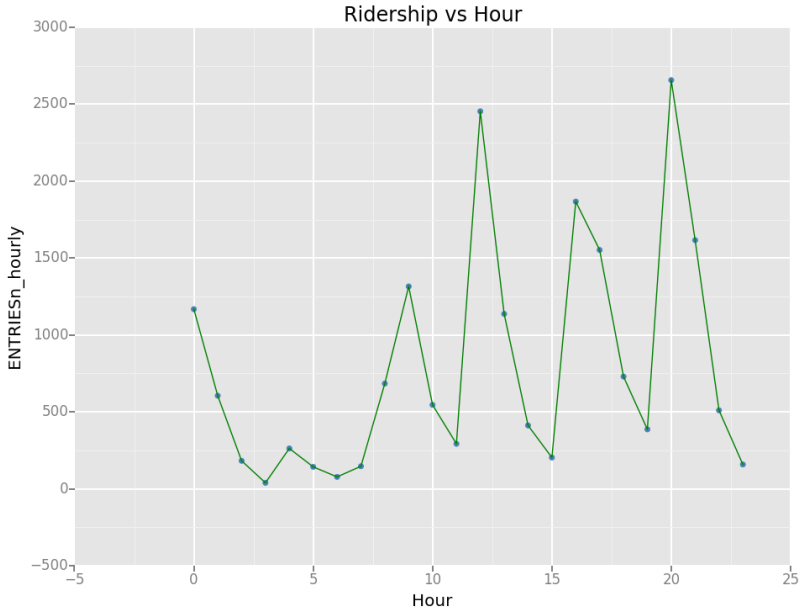


- Ridership by time-of-day

- Ridership by day-of-week



From the figure above, subway ridership on weekend are relatively less than that on weekdays.



The figure above shows that 8:00, 12:00, 16:00 and 20:00 are the peak time for subway transportation in a day.

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?

I did a Mann Whitney U-test to analyze this dataset with the null hypothesis that the mean of entries to the subway station in rainy day is equal to that without rain and get p-value = 0.05, which equals to our p-critical value of 0.05. We reject the null hypothesis at the 95% confidence level that the mean of entries to the subway station in rainy day equal to that without rain. This indicates the means of these two samples are significantly different. We could interpret this result as ridership of subway in NYC when it is raining is significantly different from that 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.

```
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
rain          -39.1162     41.257     -0.948      0.343      -119.989    41.757
Hour           62.2581      2.486     25.047      0.000        57.386    67.131
meantempi     -13.0971      2.669     -4.907      0.000       -18.329    -7.865
fog           218.0823     51.681      4.220      0.000       116.777   319.387
```

(statistical results of non-dummy features in my regression model)

As p-value of ‘Hour’, ‘meantempi’ and ‘fog’ is less than 0.05, these features influence the subway ridership significantly. Combined with their coefficients, we could conclude that:

1. Hourly entries to subway is 62 more when the hour increases by 1.
2. Hourly entries to subway is 13 more as the temperature decreases by 1.
3. Hourly entries to subway is 218 more when foggy.

Here, ‘fog’ has the highest coefficient, showing that ‘fog’ has the largest effect on the ridership. However, in my model, I didn’t consider multicollinearity. In fact, ‘fog’ is likely influenced by the feature ‘Hour’ or ‘meantempi’. In my experiment that I dropped or added some features in the linear regression model, the coefficient of ‘fog’ changes greatly, but the coefficient of ‘Hour’ and ‘meantempi’ remain steady. So the interpretation for the ‘Hour’ and ‘meantempi’ are relatively reliable. But the interpretation for ‘fog’ should be further evaluated.

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 was limited to May 2011. The weather (rain, fog, temperature, etc.) doesn't change wildly in May so that the dataset can't expose the relationship between wider range of weather and subway ridership in NYC. Bigger dataset across the whole year is needed in this topic.

My regression model doesn't consider multicollinearity and doesn't analyze the residual. In fact, variables in this dataset may exist positive or negative influences on each other. R-squared value, 0.484, is not as high as expected. This shows that more features should be taken into consideration and the feature of 'rain' should be dropped. To get the better model with appropriate features, stepwise method could be used.

For the statistical test I used, there is an assumption that the two population are from the same distribution. From the exercise 3.1 figure, we could tell they are nearly identically distributed. But more tests should be used to demonstrate. Otherwise, the conclusion of the current test is possibly unreliable.

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