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.

- Mann-Whitney documentation: http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

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 the Mann-Whitney test to analyze the NYC subway data because the data was not normally distributed. The Mann-Whitney test is a one-sided test and resulted in a one-sided p-value of 0.25. Because our p-critical value was 0.05, indicative of a 95% confidence level, we can reject the null hypothesis, which stated that the distributions of both populations (with rain and without rain) are identical.

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.

The Mann-Whitney test was applicable to this dataset because the data was not normally distributed. Because of this, the t-test was not applicable, but the Mann-Whitney could be used as it does not assume that the data is drawn from any particular underlying probability distribution.

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 Mann-Whitney test returned the following values:
- Mean of entries on rainy days: 1,105
- Mean of entries of non-rainy days: 1,090
- p-value: 0.025

1.4 What is the significance and interpretation of these results?

The above results tell us that there is a statistically significant difference between entries on rainy vs. entries on non-rainy days. The p-value tells us that the difference between the means of the rainy and non-rainy days is because the populations do in fact differ and can be explained by at least one statistically significant difference.

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:

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

I employed gradient descent to compute the coefficients theta and produce the 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?

In my model, I used the following variables as my features:
- Rain
- Hour
- Mean Temp
- Fog
- Mean Wind Speed

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.

I selected these variables based on intuition and data exploration. Primarily, I tried to consider what would cause someone to use the subway instead of other modes of transportation, and I concluded that it must be due to weather: rain, fog, wind speed, and temperature. I also selected Hour because it had a signifiant effect on my r^2 value, specifically a difference of 0.04.

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

The coefficients or theta values of the non-dummy features are found by inspecting the theta calculations from the theta gradient descent procedure and aligning the outputs with the columns in the features array. This gives us the following weights:
- rain: -10.2
- Hour: 439.4
- meantempi: -47.5
- fog: 57.0

- meanwindspi: 57.8

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

My model outputs a R2 value of 0.46

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?

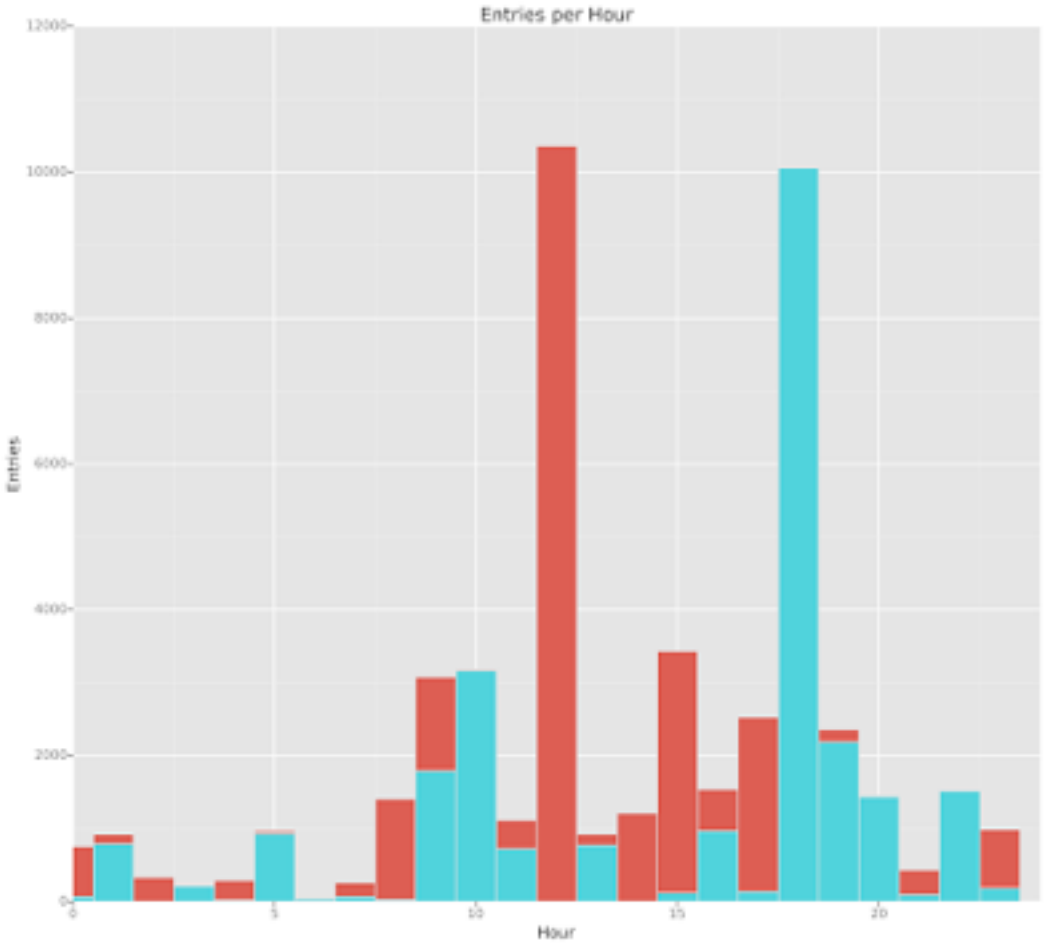
This R2 value of 0.46 means that the model has a pretty good fit, i.e. the linear regression fits the data about 46%. I think this linear regression model is appropriate for this dataset because the R2 value explains a fair amount of the data or about half of the standard deviation. R2 is not the only value to look at though. Looking at other data, e.g. the residuals, helps us better appreciate the goodness of fit: the residuals gives us a normal distribution of values with a large amount at 0, indicating that the model was a fairly good fit.

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.

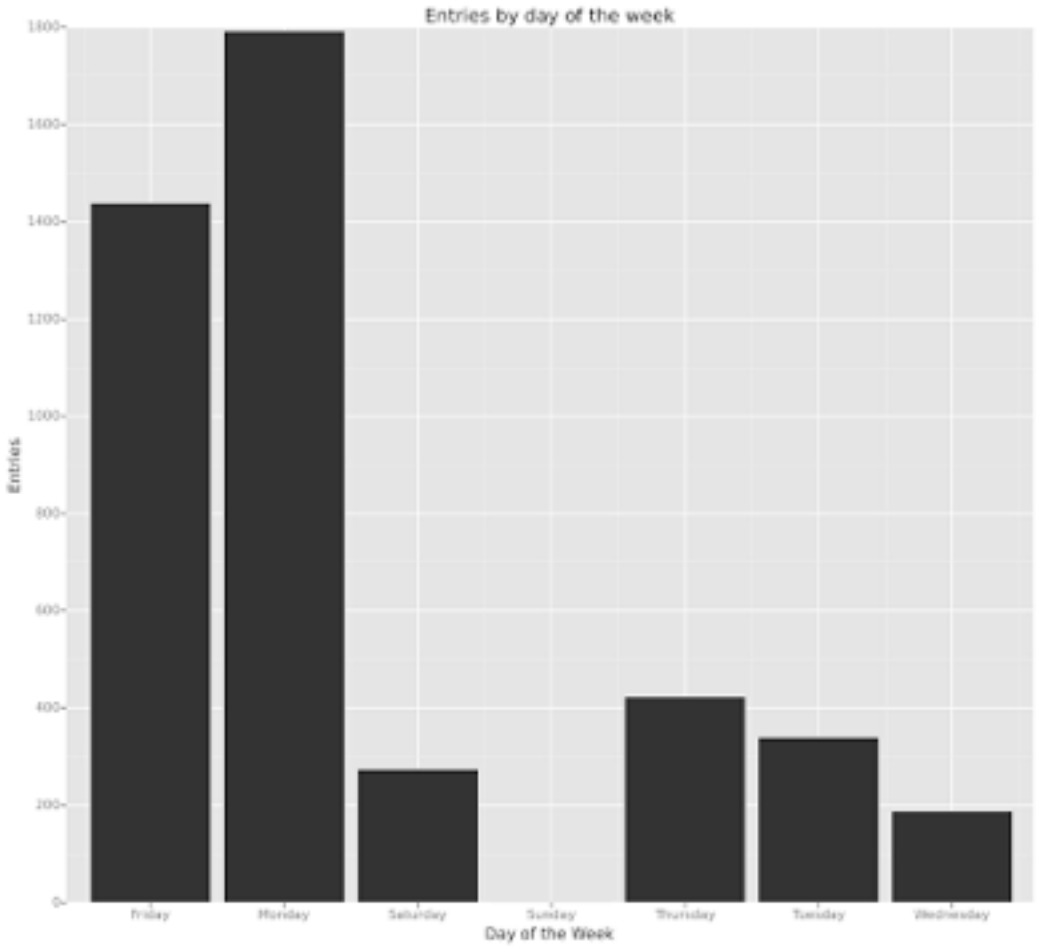
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

The visualization below shows the number of entries per hour for rainy vs. non-rainy days with red for non-rainy and blue for rainy days. This visualization shows that ridership primarily occurs between 9a and 8p and the peak occurring at 12p for non-rainy days and 6p for rainy days.



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

The below chart depicts the ridership by day of the week. This shows that ridership primarily occurs on Monday and Friday with very little occurring on the weekends, none on Sundays in fact.



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?

From analyzing the data, more people do in fact ride the subway when it is raining. We know this because the mean value of ridership on rainy days is greater than the mean value of the ridership on non-rainy days and the Mann-Whitney test indicates that this difference is statistically significant.

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 statistical test indicates that the rainy day ridership and the non-rainy day ridership are different populations. This conclusion coupled with the higher mean value for rainy day ridership indicates that more people ride the subway on rainy days. Additionally, the linear regression model found a negative correlation, i.e. coefficient, between non-rain and ridership.

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: Dataset, Analysis, such as the linear regression model or statistical test.

The analysis has potential shortcomings because our current analysis uses only a partial dataset, i.e. one month's worth of data. Additionally, I don't think we can reply solely on a single statistical test or just the outputs of the linear regression model. Ideally, we'd combine the analysis with a couple other statistical tests and models.

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