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.

Intro to Data Science (Udacity)
Intro to Statistics (Udacity)
Programming foundations with Python (Udacity)
Shapiro–Wilk test - Wikipedia, the free encyclopedia
SQL tutorials – W3schools.com

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

Null hypothesis is that the two population means - ridership on rainy days and ridership on non-rainy days is equal.

p-critical value 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.

The data I analyzed include two populations: ridership on rainy days and ridership on non-rainy days. These two populations have possibly unequal variances and sample size. The Welch's t-Test should meet the following assumptions:
a. Both samples are drawn from normal population
b. The two samples are independent.
On examining distribution I found, the histograms for the number of entries per hour for days on rainy days and non-rainy days showed were not normal distribution.
Also, the Shapiro-Wilks test, which is a test to check if a sample came from a normally distributed population, was against the null hypothesis that the populations were normally distributed ($p < 0.05$). Taken together, the data did not meet the assumptions for Welch's t-Test. Thus, I chose Mann-Whitney U-Test, which can be used for data with both normal and non-normal 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 mean of ENTRIESn_hourly on rainy days was 1105.45 and the mean of ENTRIESn_hourly on non-rainy days was 1090.28.
The MannWhitneyU-Test results showed that the p_value for the test was 0.024 (0.048 for two-tailed value).

1.4 What is the significance and interpretation of these results?

Because the p_value (0.024) (in case of 2-tailed, p value = 0.048) is less than 0.05, I concluded that the null hypothesis was rejected and the ridership on rainy days and non-rainy days were significantly different

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 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?

I tested the correlation between ridership and the following features – rain, Hour, Weekday, Holiday, Peak hours, max temp, min temp, UNITS

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 was intrigued to check impact of different hours of day on ridership so included Hours and Peak hours
Similarly, holidays can affect the ridership i.e. weekends and weekdays may have different impact on usage of subway
Finally weather can play an important aspect of ridership so included temperature and rain

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

R2 = 0.53

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?

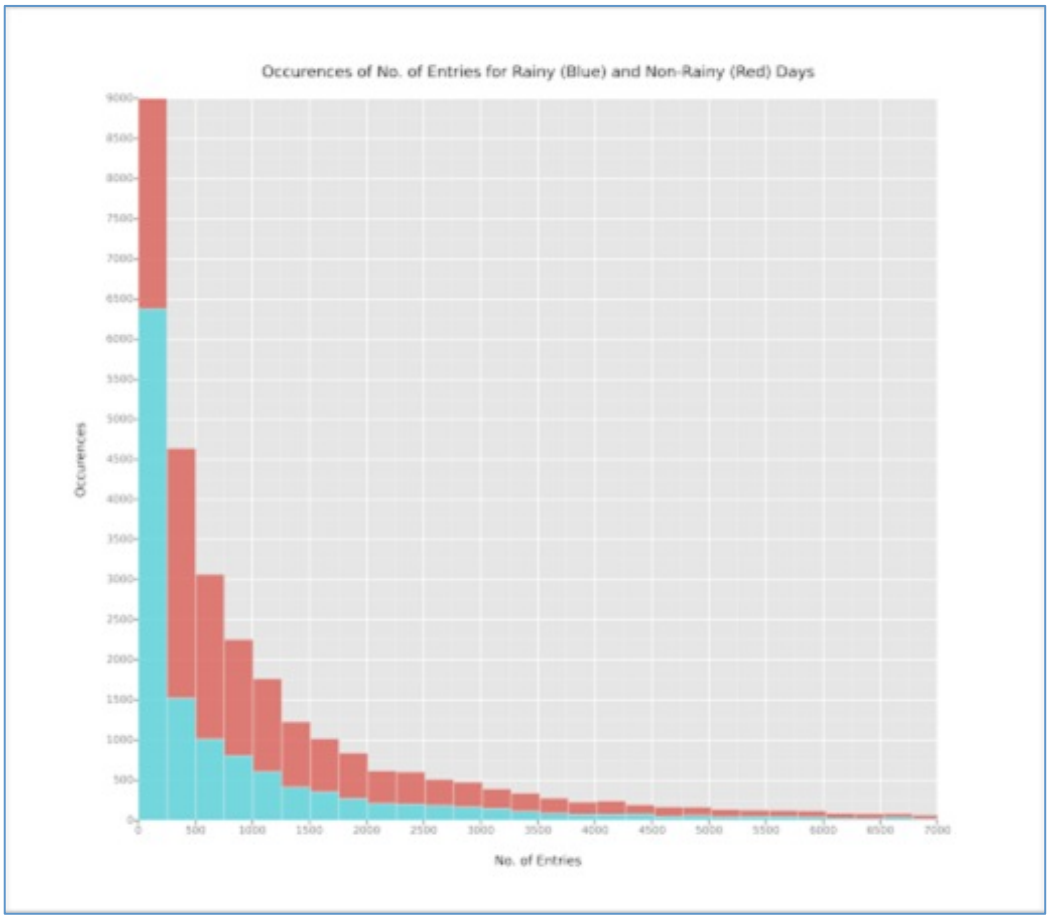
The R2 value using all features is 0.53, which show a positive correlation between these features and ridership in NYC subway.
Particularly, there is less correlation between "Hour" and ridership (R2 = 0.017), high positive correlation between "EXITSn_hourly" and ridership (R2 = 0.53) and negative correlation between each weather feature and ridership.
In this test, the overall R2 value is 0.53, which suggests a relative high correlation between features and ridership.
Based on this R2 value, this linear model is a good model for this dataset and may be served as a predictive model for future dataset.
Having said so, I also believe that instead of simple straight line, a polynomial model might have predicted the better

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.

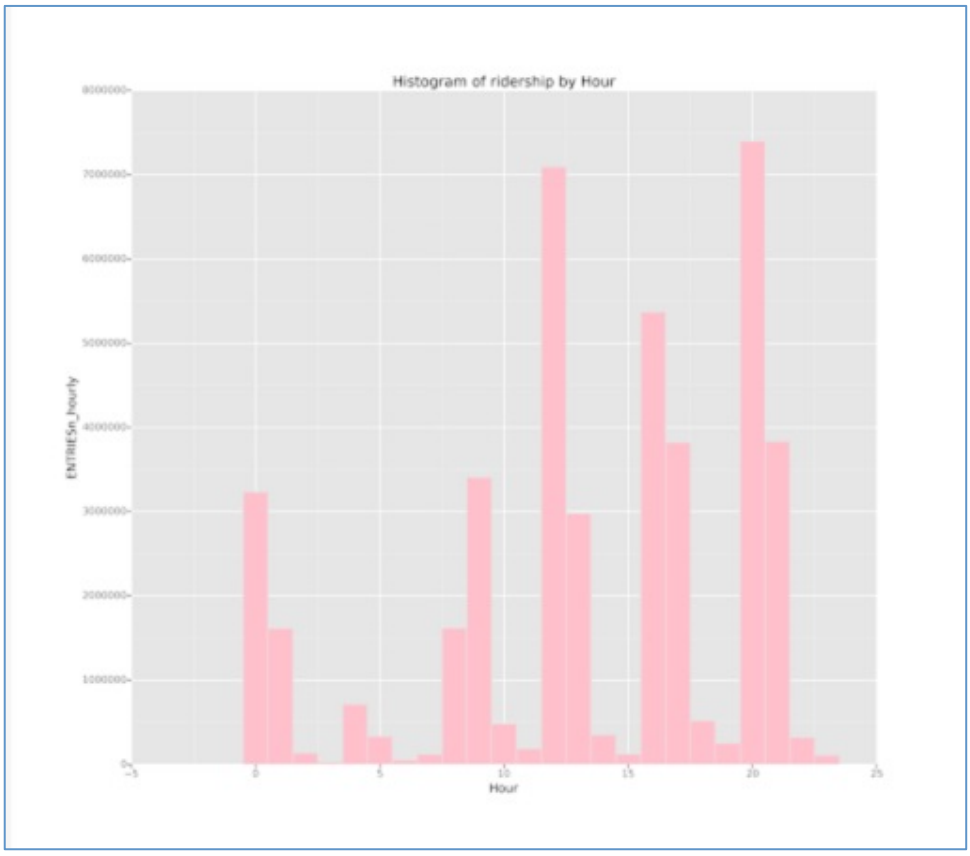
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.
You can combine the two histograms in a single plot or you can use two separate plots.
If you decide to use to two separate plots for the two histograms, please ensure 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.

I illustrated occurrences of entries for rainy (blue) and non-rainy (red) days cumulatively

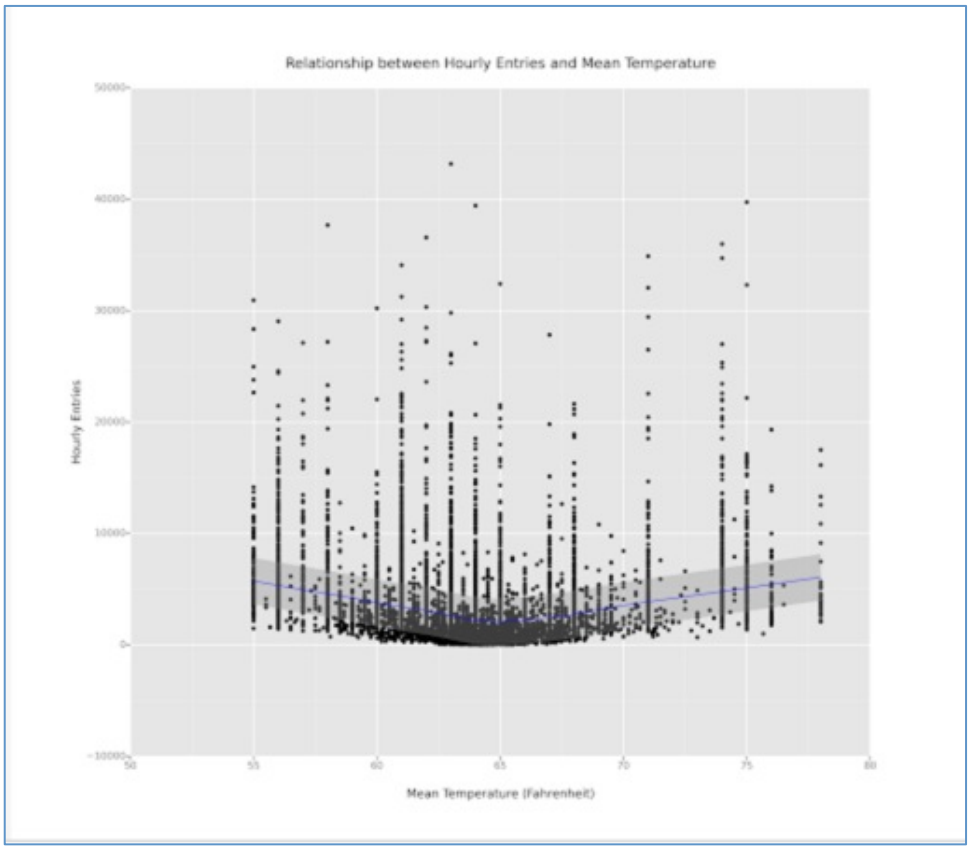


3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:
Ridership by time-of-day
Ridership by day-of-week

I investigated and illustrated the ENTRIESn_hourly per hour. As can be seen, for particular hours (12pm, 20pm) we have very high ridership(>7000000)



In addition, I thought to explore relationship between mean temperature and hourly entries into subway. Though the correlation is not strong but a trend is being shown. To explore further I tried to add smoother through GLM



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?

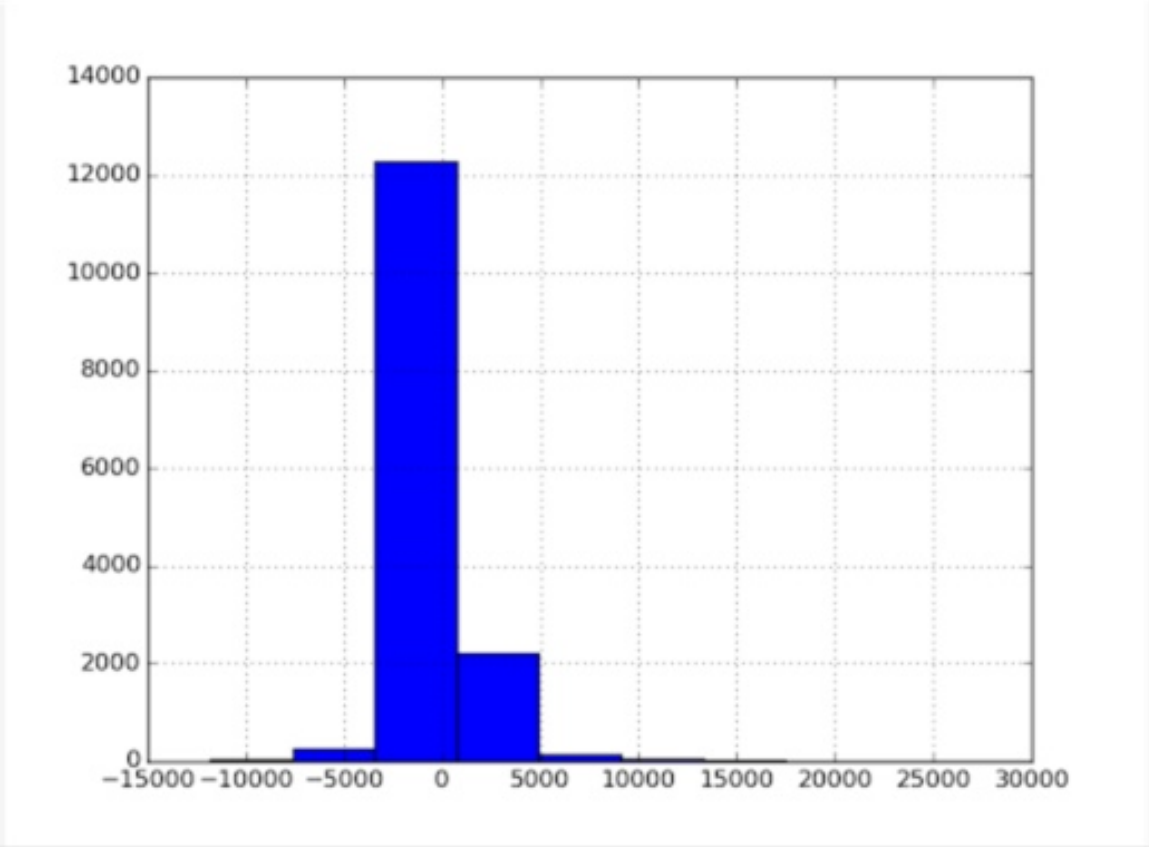
Yes, more people ride the NYC subway when it is raining than when it is not raining

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.

Data Analysis - In this analysis I tested the null hypothesis- the ridership on rainy days and non rainy days are the same, and an alternative hypothesis- the ridership on rainy days and non-rainy days are not the same at an alpha level of 0.05.
The Mann-Whitney U-Test results showed a p_value of 0.024 and for 2-tail P-test it will be 0.48, which was less than 0.05. Because the p_value (0.048) was less than 0.05, I concluded that the null hypothesis was rejected and the ridership on rainy days and non-rainy days were significantly different.

The mean of ENTRIESn_hourly on rainy days (1105.45) was slightly higher than the mean of ENTRIESn_hourly on non-rainy days (1090.28).

Data Modeling – The model created using linear regression with gradient descent had an R^2 value of 0.53. But more importantly, plotting the residuals showed that there was little difference between the predicted values and original values



Thus, the results supported that there were more people ride the NYC subway when it was raining versus when it was not raining.

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.

Few shortfalls, which I feel with the data, are –

a. The dataset consists data for one year and that too of particular month of May. By providing historical data or different months with different rainfall can help modeling better
b. It seems 'fog' condition does not have any effect on the ridership as the feature "fog" shows a similar R^2 value as feature "rain"
c. Also rain = 1 can mean a 10 minute rain or 2 hour rain thereby can give different result
d. The R^2 value achieved is 0.53 but using polynomial regression and weighted analysis of feature could have increased this value

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

I tried to explore relationship of mean temperature (meantempi) with ridership and found that for a temp of 75 F or below, there was significantly higher ridership than for temperature above 75 F