**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. 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

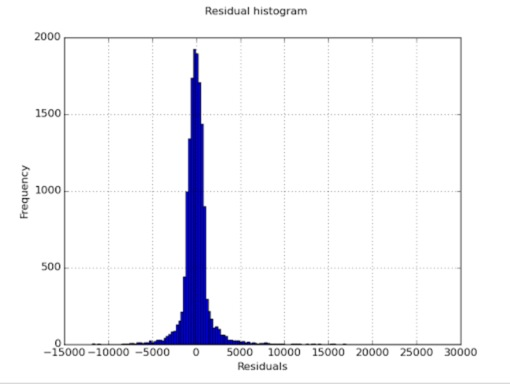
Finally weather can play an important aspect of ridership so included temperature and rain

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

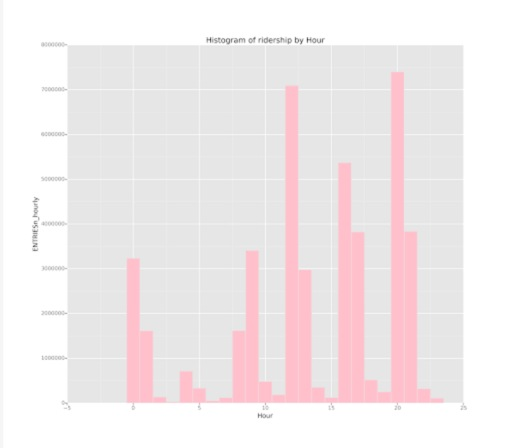
R2 = 0.52

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?

* R-square is 1 minus the ratio of residual variability. When the variability of the residual values around the regression line relative to the overall variability is small, the predictions from the regression equation are good. R-square value is an indicator of how well the model fits the data
* For current problem, the R2 value using all features is 0.52. This means that we have explained 52% of the original variability, and are left with 48% residual variability
* This does not represent strong linear relationship between dependent and independent variables; the linear model is not appropriate for this dataset.
* Also looking at the residual plot (below), it can be inferred that although the histogram is of normal distribution but long tail is the indicator of greater residuals

****

* On adding ‘rain’ as variable gives same R2 value indicating weak linear relationship
* Finally, the graph between Entries and Hour further elucidate the weak linear relationship between dependent and independent variables.



Therefore I believe that instead of simple straight line, a polynomial model might have predicted 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.

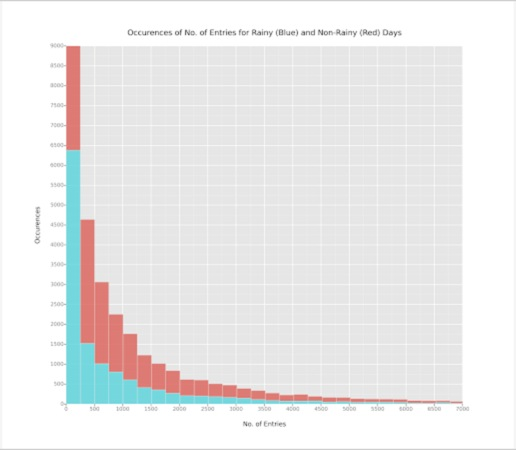
You can combine the two histograms in a single plot or you can use two separate plots.

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

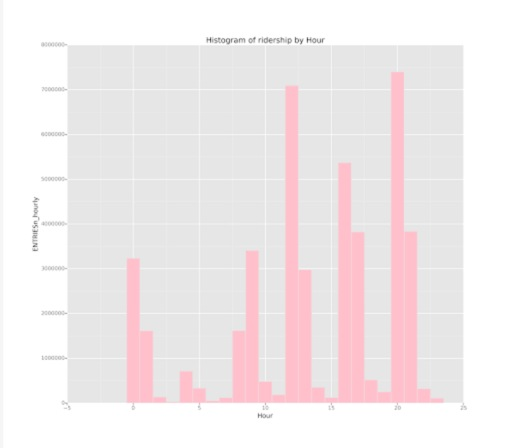


3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

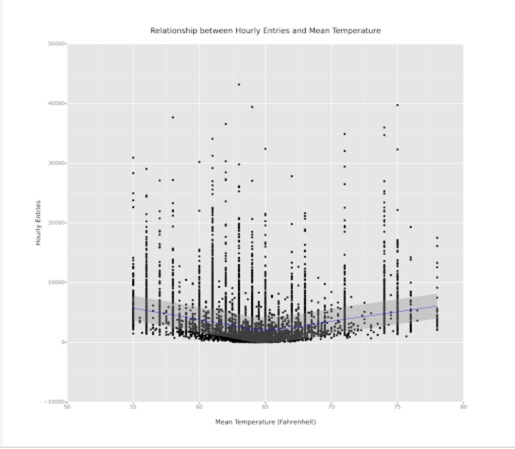
Ridership by time-of-day

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

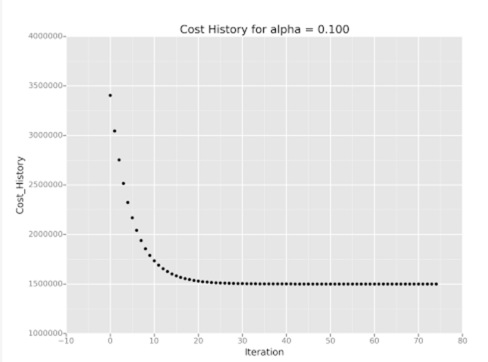
4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

**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 R2 value of 0.52. Also the plot of Cost history for alpha = 0.1 is as below



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**

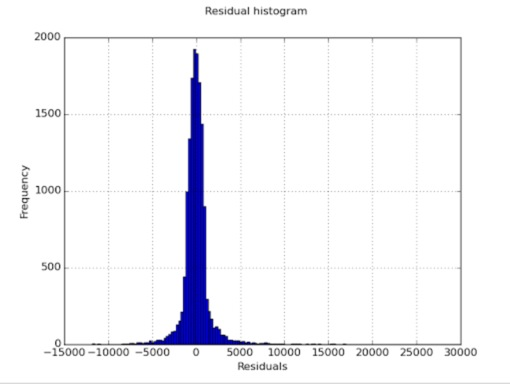
Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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 –

1. 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
2. It seems ‘fog’ condition does not have any effect on the ridership as the feature “fog” shows a similar R2 value as feature “rain”
3. Also rain = 1 can mean a 10 minute rain or 2 hour rain thereby can give different result
4. The R2 value achieved is 0.52 the long tails for residual plot suggest lesser explanation of X-Y relation. Thus using polynomial regression and weighted analysis of feature could have increased this value.
5. Also the long tails of residual histogram shows inappropriateness of Linear Regression as perfect model to capture relationship of variables



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