Analyze the NYC subway data

Section-1. Statistical Test

1.1 Q. Which statistical test did you use to analyze the NYC subway data?
Answer: I use Mann-Whitney U-Test to analyze the MTA subway data.

Q. Did you use a one-tail or a two-tail P value?
Answer: A two-tail p-value was selected since an appropriate initial question, given the results of the Weather-Related Data section of the *Data Exploration* supplement, is simply whether or not there is a statistically significant difference between the populations

Q. What is the null hypothesis?
Answer: The Mann-Whitney U test is a nonparametric test of the null hypothesis that the distributions of two populations are the same.

Q. What is your p-critical value?
Answer: p-critical value is 0.05

1.2 Q. 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:
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. Based on the characteristics of the data, Welch's t-Test or Mann-Whitney U-Test may be used to check the null hypothesis that the mean of two populations is the same against an alternative hypothesis. The Welch's t-Test should meet the following assumptions:

a. Both samples are drawn from normal population
b. The two samples are independent.

Therefore, I examined if the data I used for analysis were normally distributed. First, the histograms for the number of entries per hour for days on rainy days and non-rainy days showed they were not normal distribution. Second, 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 ($p < 0.05$, Table 1). 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 Q. 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:
Mean with no rain: 1105.5
Mean with rain: 1090.3
U-statistic: 1924409167.0
P value: 0.0249

1.4 Q. What is the significance and interpretation of these results?
Answer: Comparing the means yields 1.4% more subway entries when it rains. This statistic alone is insufficient in drawing conclusions or correlation. The U-statistic has a high value, very close to the maximum value of 1937202044.0, or half the product of the number of values in each data set. A U-statistic of half the maximum would indicate that the null hypothesis is true. Of note, the p-value 0.025 satisfies the p-critical value, and the conclusion can be drawn with 95% confidence that the null hypothesis is false and that ridership is different with vs. without rain. Rain is likely a factor to cause differences in subway ridership

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?

Answer: Both gradient descent (GD) and OLS models where used to run linear regression on the NYC subway dataset. Both models look for linear relationships between the features and the predicted values or NYC subway rides.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
Answer: In the GD model the features used were: rain, precipitation (precipi), hour of the day (Hour), mean temperature (meantempi) and dummy variables for individual station (UNIT). In the OLS model, the features are used where: rain, mean temperature (meantempi) and dummy variables for stations (UNIT) and dummy variables for hours of day (Hour).

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 R^2 value."

Answer: I maintained rain, precipitation, hour, and mean temperature because out of experimentation, I was unable to find R^2 values that were better. Broadening the hypothesis that "people use the subway more often when it's raining" to "people use the subway more often

when there's bad weather outside," I also included wind speed. I saw a slight increase in my R^2 values.

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

Answer: [-4.24736210e+00, 8.60518981e+00, 4.64832083e+01, 4.64163466e+02, -3.19114921e+01, 1.08898857e+02]
Coefficients are refer to following features rain, precipi, Hour, meantempi, meanpressurei and meanwindspdi

2.5 What is your model's R^2 (coefficients of determination) value?
Answer: The R squared for the GD model is 0.461. The R squared for the OLS is 0.525

2.6 What does this R^2 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 R^2 value?
Answer: The R squared for the OLS is 0.525 which means we can explain about 52.5% of the data variability with the model. In other words, our model lets us predict NYC subway entries with 52% accuracy.

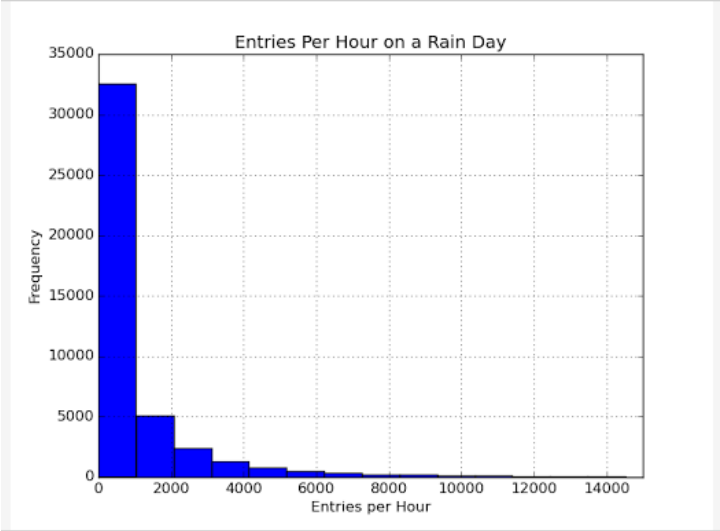
Section-3. Visualization

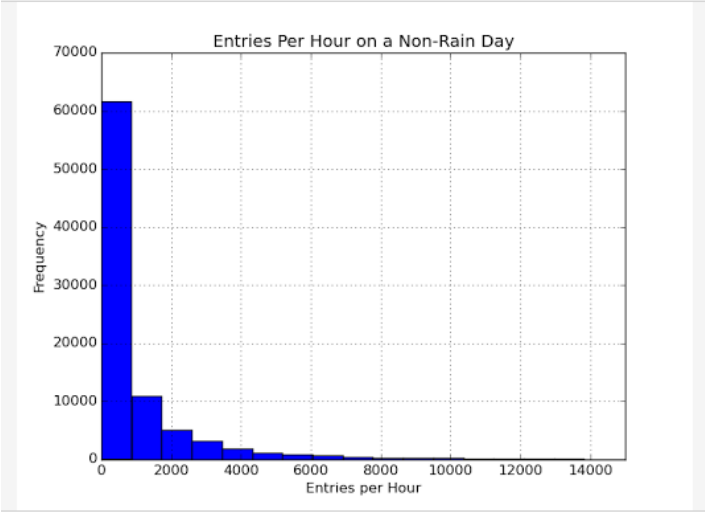
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

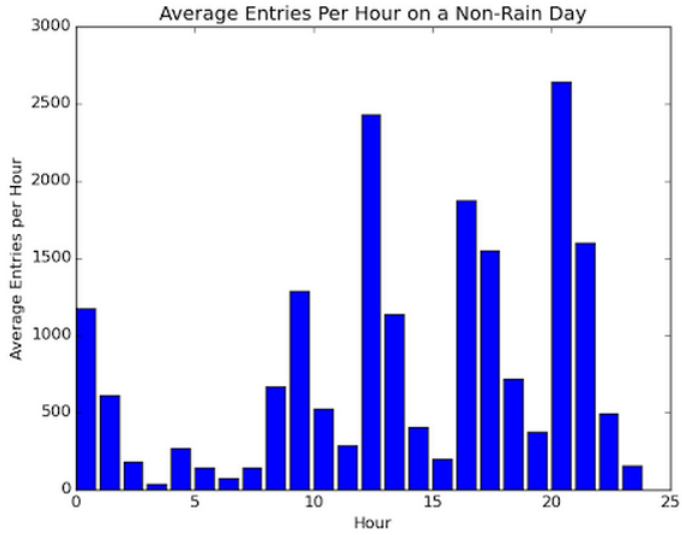
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

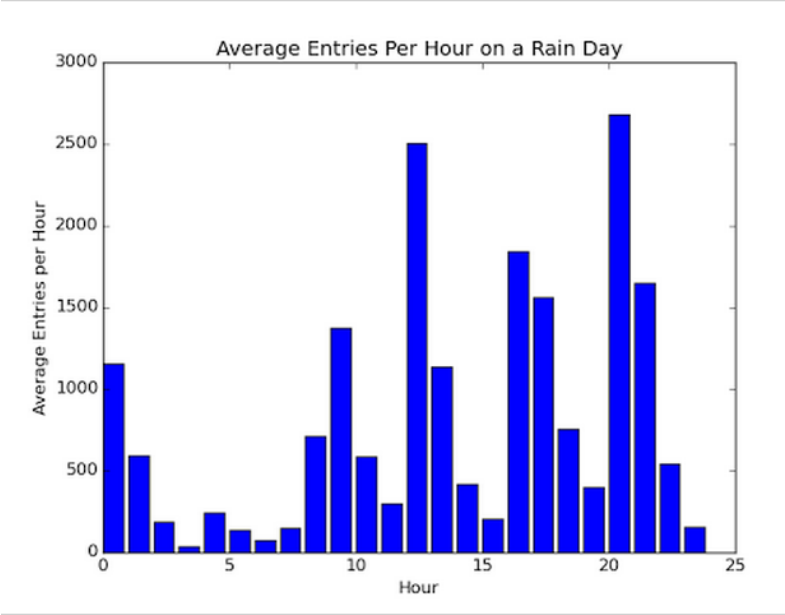
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, 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.


Entries Per Hour on a Rain Day
Frequency
Entries per Hour
0
5000
10000
15000
20000
25000
30000
35000
0
2000
4000
6000
8000
10000
12000
14000

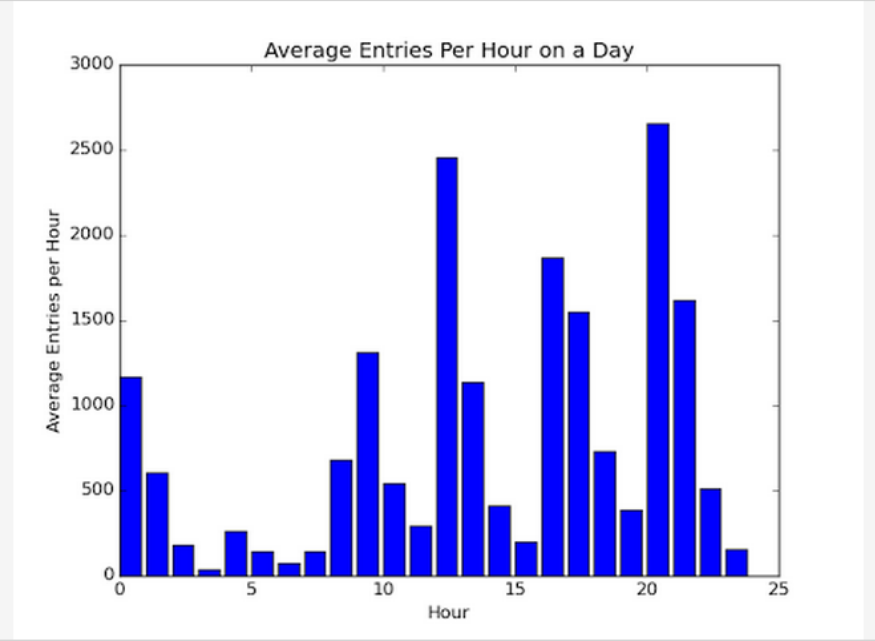

Entries Per Hour on a Non-Rain Day
Frequency
Entries per Hour
0
10000
20000
30000
40000
50000
60000
70000
0
2000
4000
6000
8000
10000
12000
14000


Average Entries Per Hour on a Non-Rain Day
Average Entries per Hour
Hour
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
25



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



By plotting the average number of subway entries at each hour, it's clear that there are several peaks throughout the day, with the most prominent ones being at noon and 8pm. Interestingly, these peaks are larger than those during rush hours (8-9am and 5-6pm). It raises some intriguing questions about the demographics and characteristics of NYC subway riders: assuming a 9-5pm

workday, why are more subway entries occurring at 5pm vs. 9am? Are more people going out to lunch (12pm) and dinner (8pm), or is that there work schedule? Without any demographic data, it would be impossible to determine these questions from the current data set.

Section 4. Conclusion

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: On average, between 15 and 100 more people ride the NYC subway on a rainy day compared to a non-rainy day. These numbers come from using simple mean comparison, and linear regressions with Gradient Descent and OLS. In the mean comparison, we see a difference of 15 entries per hour, while in the gradient descent model the theta for the rain variable was 104.5. Given that the rain variable is a boolean the interpretation of the theta is that when it rains, the model predicts on average 104.5 more people will ride the subway.

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: The positive coefficient for the rain (0 or 1) parameter indicates that the presence of rain contributes to increased ridership. This may have not been the case for all data points, with the R^2 being approximately 46%; however, the small residuals show relatively high accuracy, given our objectives. Although the means of both data sets are not that different from each other, the Mann-Whitney U test did indicate that there was a statistically significant change in ridership for rain vs. no-rain. It is conscientious to claim that rain increases subway ridership

Section 5. Reflection

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 data set under consideration was limited to a single month in the late spring / early summer of a particular year. As a result, among countless other possible factors for which the available data did not account, precipitation may in fact have an increased impact on the number of entries at other times of the year (e.g., during the winter months). Thus, the data set was limited by its temporal locale.

The linear regression model was created, while having very high r and R2 values did not, based on residual analysis, adequately model the data. There is in fact *not* a linear relationship between the explanatory and response variables under consideration; thus, a non-linear model would likely be more appropriate for the current data set.

The statistical tests that were employed seemed effective (as long as sample sizes were kept small enough). However, it's unclear how traditional statistical tests relate to massive data sets.

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

Assuming all statistical tests and learning models were implemented and interpreted correctly, it became clear that computational power was very important in data science, not due to the ability merely to apply methods to data, but in the ability to repeat numerous tests on random samples of data, which, at least in the case of this analysis, encouraged more confidence in test/model results.

Reference

http://ggplot.yhathq.com/docs/index.html (GGPlot)

http://graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm (GraphPad)

http://statsmodels.sourceforge.net/devel/generated/statsmodels.regression.linear_model.OLS.html (statsmodels.regression.linear_model.OLS)

https://www.udacity.com (Data Analysis with R, Intro to Data Science, Udacity classes for project one)

http://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/ (Introduction to Probability and Statistics)

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

http://en.wikipedia.org/wiki/Statistical_significance

http://en.wikipedia.org/wiki/Probability_density_function