ANALYZING THE NYC SUBWAY DATASET

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, since it does not assume equal population size, variance and data distribution. For this analysis, a two-tail P value was used, and the null hypothesis is that both populations are the same.

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 test is applicable because each group measurement is independent from the other.

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 the number of people taking the subway on the population on rainy days is different than the population on non-rainy days. This is noticeable due to the p value of 0.0249, and since the test is one-tailed by default, 0.0249 times 2 is smaller than 0.05, therefore we must reject the null-hypothesis that both populations are the same.

1.4 What is the significance and interpretation of these results?

It is possible to say with a 95% certainty that the mean of people taking the subway on rainy days is different than the mean of people taking the subway on sunny days, for the entire population.

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)

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 used were: precipi (indicates the amount of precipitation in this day and location), meanwindspdi (mean wind speed for that location), meantempi (mean temperature for this location) and TIMEn (hour of the day for location).

Dummy variables had to be used to include categorical variable UNIT.

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."

The precipi variable is a more detailed version of rain and snow, because not only it contains information whether there was or was not (0 or 1), but also a higher number means heavier rain.

Windspeed was selected because it seemed logical to me that if people would take the subway more when it rains, they would do it when it is windy as well.

Mean temperature was used for it seemed logical to me that people would take the subway more frequently on cold days.

Hour of day seems a nice feature because it will help count rush hours, when people go and come back from work.

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

8.15 ; 54.6 ; -43.4 and 463.8.

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

R^2 value is 0.4643.

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?

Since the closer to 1 the better, this linear model does not seem a 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.

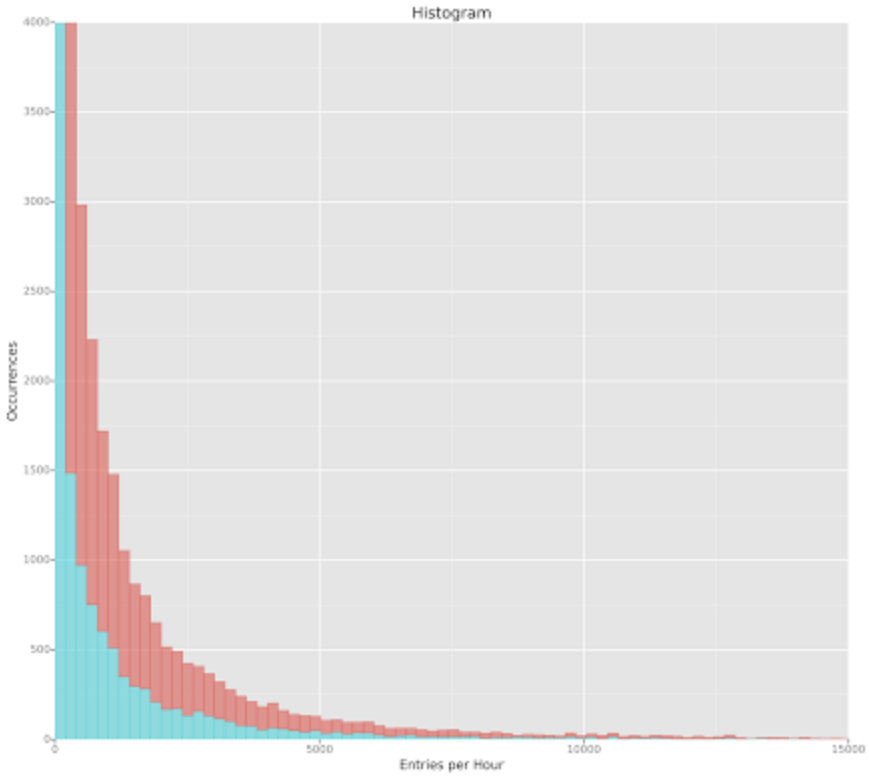
You can combine the two histograms in a single plot or you can use two separate plots.

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.

First of all, I had many problems inserting the color labels for those graphics. Somehow, it does not seem to be working properly on the web environment of Udacity, since the same code lines generated them when running on my computer.



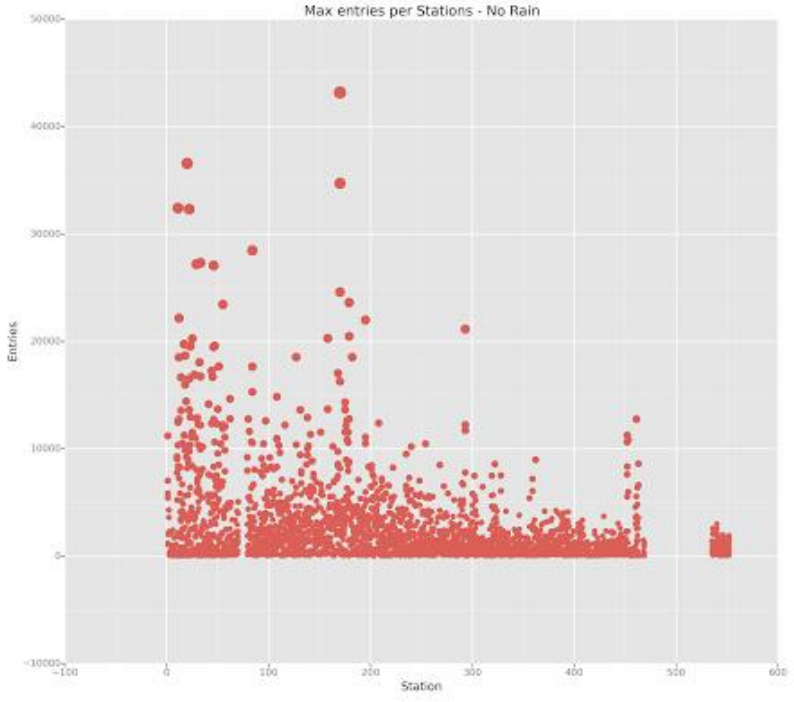
This is the Histogram of Entries per hour with a few changes: bin size has been resized to 200, the max number of entries per hour has been limited to 15,000 and the count of occurrences has been capped to 4,000. With this graphic we can see that the distributions of both rainy and non rainy days are practically the same. However, the difference in the area of both, which would correspond to the total number of entries, cannot be used as comparison since the number of non rainy days might be much greater than the number of rainy days measured.

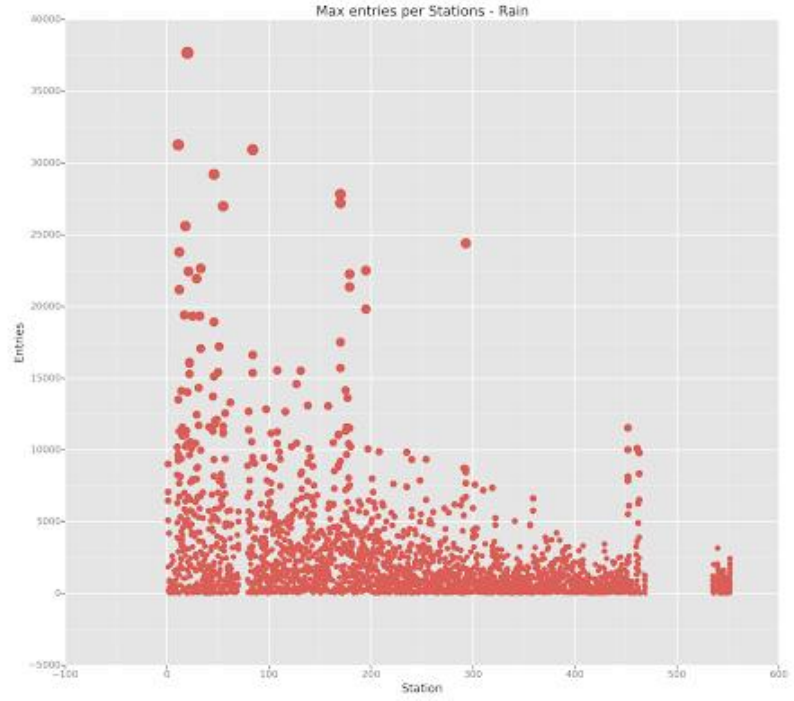
3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

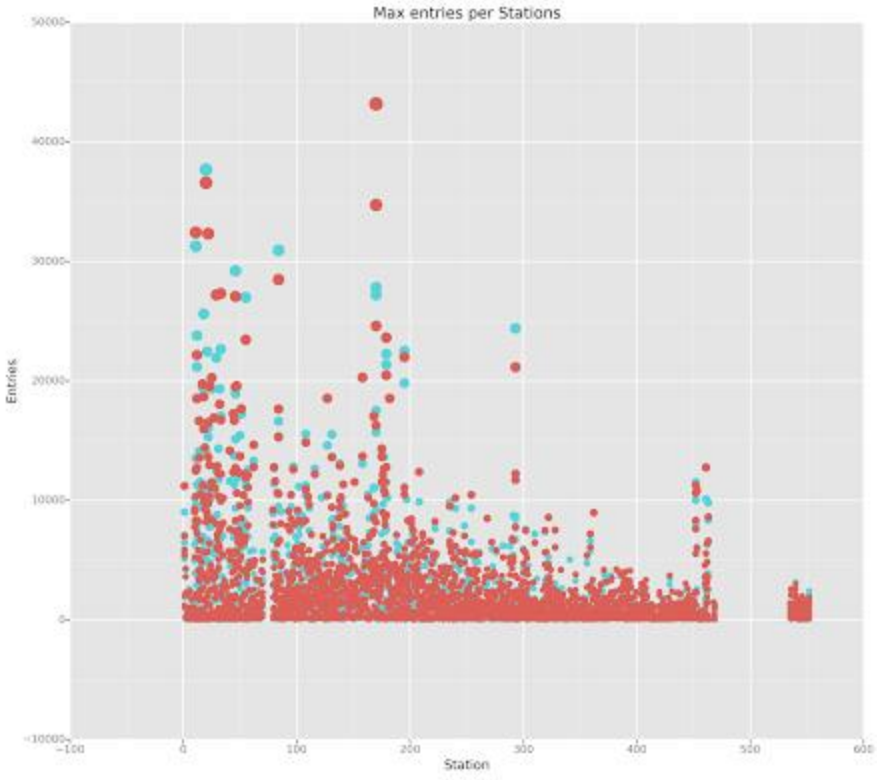
Ridership by time-of-day

Ridership by day-of-week

First part of the analisys consist of observing the Max values for each station, in order to chck if it may have any impact. As it is observable, they look similar, but non-rainy days has a greater value (above 40,000). This graphics should be used together with the next ones, of average.

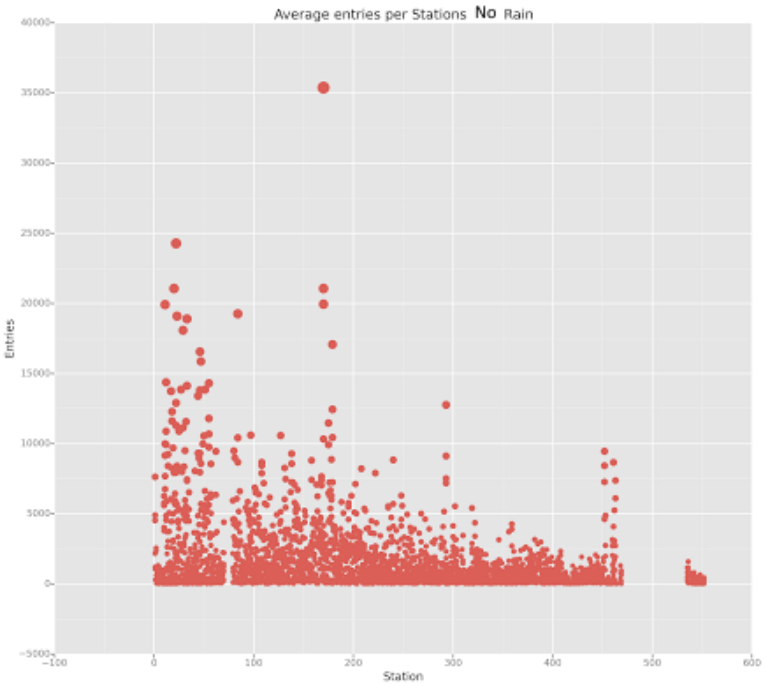

Max entries per Stations - No Rain
Entries
Station
50000
40000
30000
20000
10000
0
−10000
−100
0
100
200
300
400
500
600

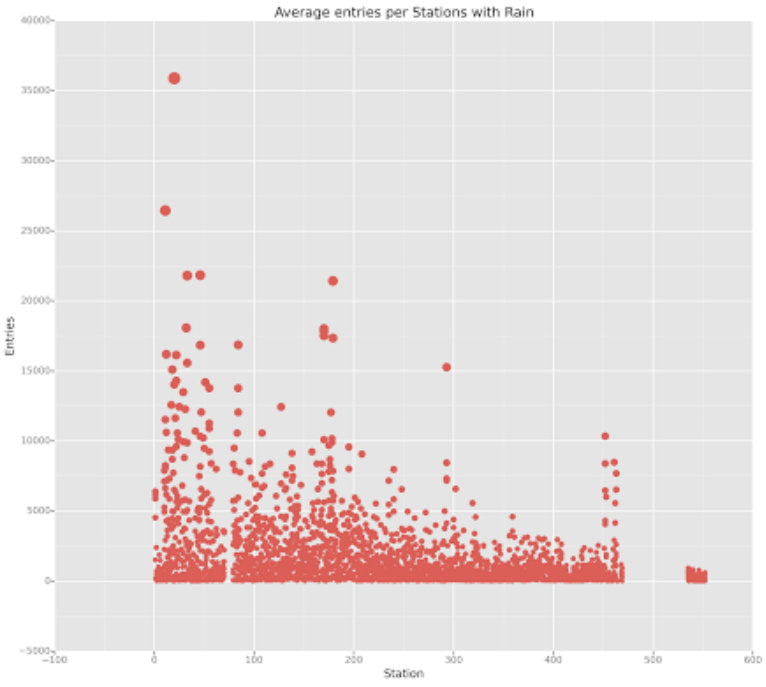

Max entries per Stations - Rain
Entries
Station
40000
35000
30000
25000
20000
15000
10000
5000
0
−5000
−100
0
100
200
300
400
500
600

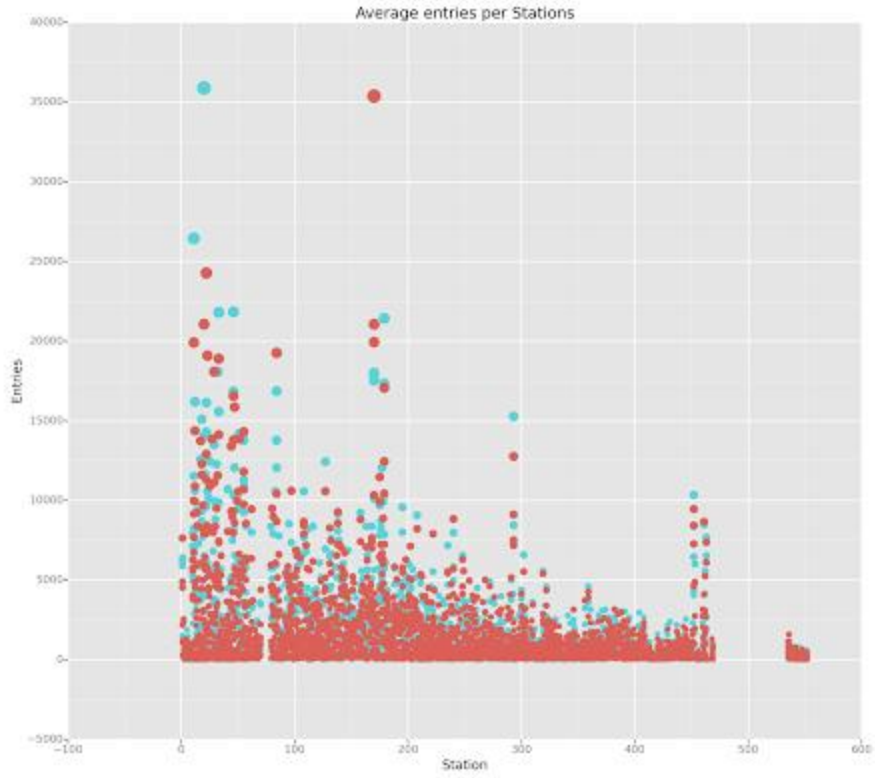


The last graphic shows both rainy and non rainy days on the same graph.

The next graphics show average values for each station, in rainy and non rainy days. To use the station name as X axis, the letter part of their names was removed, leaving only their number. For an example, 'R012' becomes '12'.

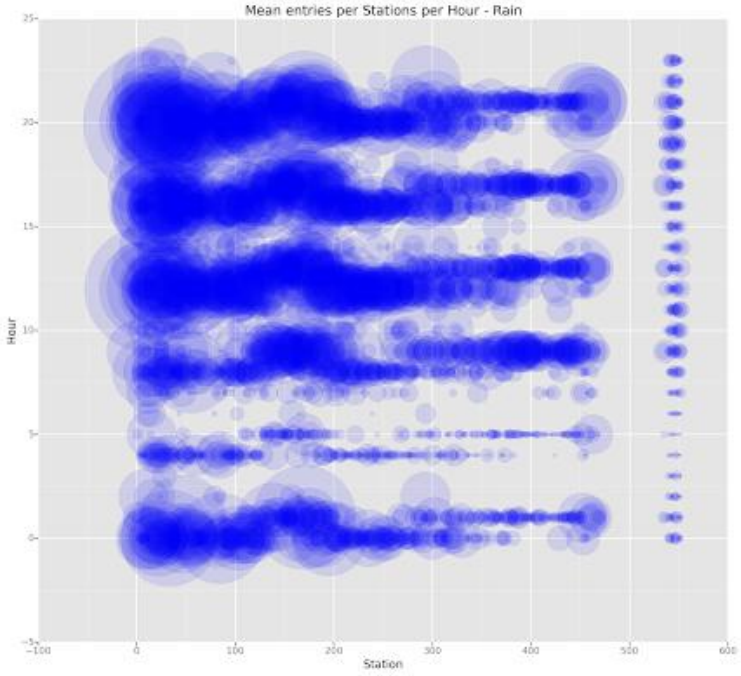

Average entries per Stations No Rain
Entries
Station
40000
35000
30000
25000
20000
15000
10000
5000
0
−5000
−100
0
100
200
300
400
500
600

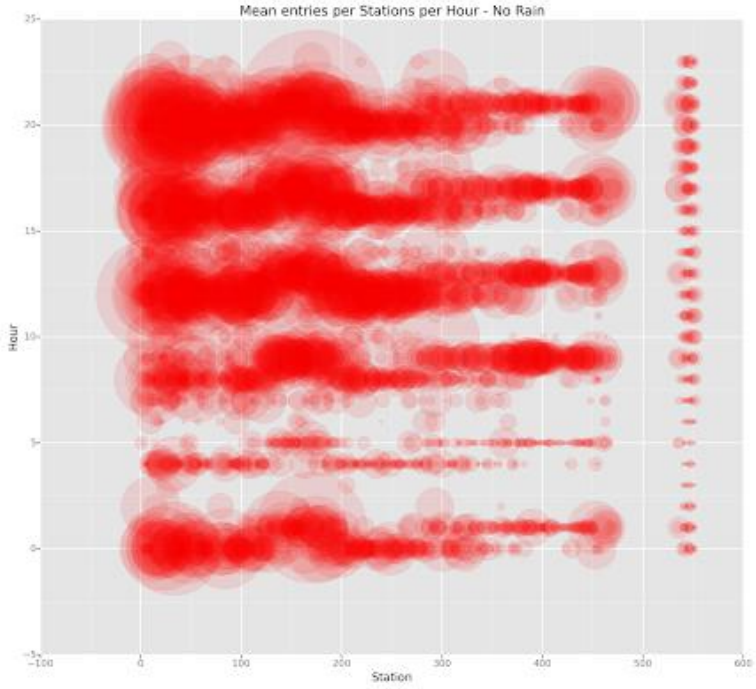

Average entries per Stations with Rain
Entries
Station
40000
35000
30000
25000
20000
15000
10000
5000
0
−5000
−100
0
100
200
300
400
500
600



Those last graphs does not offer a good view of which one might be bigger. On average, it seems some stations have more users on rainy days, as other have more users on non rainy days. This could imply, for an example, that stations where user count increases on rainy days are those closer to the center of the city, where people that would otherwise go walking or by bicycle now take the subway. On the other hand, those where usage increase in non rainy days are those from people that have cars and live further, and choose their cars over subway on rainy days.

The next graphs, however, have shown to me a slight tendency of a majority of non rainy days users. The graphs consists of the mean number of entries, per station, per hour. The size of each point is proportional to the number of entries, therefore, the bigger the circle, the higher the number of entries. Since it was difficult to see one ball overlapping the other, I used a jitter like graph to ease the visualization. Therefore, the more intense the color is, and the bigger the area painted, the higher the overall mean. Also, the presence of lighter areas means alone high means, as the darker areas mean a higher value shared by stations.


Mean entries per Stations per Hour - Rain
Hour
Station
25
20
15
10
5
0
−5
−100
0
100
200
300
400
500
600


Mean entries per Stations per Hour - No Rain
Hour
Station
25
20
15
10
5
0
−5
−100
0
100
200
300
400
500
600

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 my observations, more people use the subway 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.

According to the statistical analysis, there is significant evidence to support the populations are different. Also, from the linear regression, precipitation and windspeed seem to have a direct impact on entries per hour (however, it is not as significant as the hour of day). The final analysis was performed using graphs as support material.

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.

Despite the fact that graphs support the results, they do not definitely proof less people take the subway on rainy days. According to linear regression, the most significant aspect to determine entries per hour is the hour of day, which is observable in the last 2 graphs, as 3 massive lines seems to form across all stations around midday, 4pm and 8pm. The precipi category, which seem more specific to the rain, as it specifies how much it is raining instead of a binary value, has a smaller value in the linear regression than the other cathegories.

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

My only observation is that we could use some parts of the exercises in class to prepare this report automatically. Maybe it is part of the test being able to produce such reports as well, in this case please disregard, but otherwise it would be more practical if this report were produce more automatically.

Sources used as support material:

[1] - http://ocw.mit.edu/resources/res-6-009-how-to-process-analyze-and-visualize-data-january-iap-2012/lectures-and-labs/MITRES_6_009IAP12_lab3a.pdf