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. Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value?

The statistical test I chose to use is two tails Mann-Whitney U-Test.

2. Why is this statistical test applicable to the dataset?

Since we cannot know in advance if rain is associated with raiders' numbers we will use the two tail test, and set the P-value to 0.05. In general, a two-tailed test is the more conservative approach.

Mann-Whitney U-Test is a non parametric test that performs well for the fallowing situations:

1. For distributions that are not normal. The distributions of entries are not normal.
2. For groups of items that have different averages. The average numbers of raiders during rainy days are larger than the averages numbers of raiders during non rainy days.
3. For groups with non equal number of measurements. The data set include less rain days than non rain days.

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.

P-value = 0.025,

Avg raiders during rainy days =1105

Avg raiders during non rainy days =1090

4. What is the significance and interpretation of these results?

The results of the statistical test indicate that the difference in the averages of the 2 groups is significant and not happen by chance. the confidence level of the test is more than 95 %. Based on this result we can say that more people use the subway during rainy hours.

Section 2. Linear Regression

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) yes
2. OLS using Statsmodels. yes
3. Or something different? Yes, Robust linear model (RLM). RLM is a form of regression model design to be not overly affected by violations of assumptions by the data generation process. It is known that regression models are highly sensitive to outliers and may give misleading results, especially when the outlier results from non normal measurements error. For example: if the variance is not constant but depends on the variable, or in the presence of an outliers that are caused by different data generation process. Simple least square model will be dragged by the outliers. The variance of such data is artificially inflated, therefore will mask the outliers in the data. The RLM model I used here applies the maximum likelihood estimation suggested by Huber T to down weighting outliers in the input. The RLM model proved to be superior compare to non robust methods when the outliers are in the input variables.

2. What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

I use the fallowing parameters 'precipi', 'Hour', 'meantempi', 'meandewpti', 'rain' and the 2 dummy variables ‘Date’ and ‘UNIT’

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.”

To select the features that contribute best to the model I fallow the next steps: First, calculating the model coefficients for each feature separately this gives the prediction power of each feature. I ordered the features according to the coefficient size. Start with the feature having the smallest coefficient, I check whether adding feature with similar coefficients increase or decrease the power of the model. If adding the next feature increases the power of the model I add the feature to the model if decrease or did not change I remove the feature that contribute less when combines with other features. In this way I include more and more features. I stop adding features when adding new feature (any one of those left out of the model) cause a reduction in the power of the model.

4. What is your model’s R^2 (coefficients of determination) value? R^2=0.4947

5. 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?

The goodness of fit parameter, tell us how much of the variation in X can be explain by the variation in Y, or vice verse. In this case the model can explain almost 50% of the daily variation in the number of riders.

The linear model did not perform very well, as I mention above, only 50% of the variations in rider’s numbers can be explained by the model. Whether this model is appropriate or not depends on the motivation for that model. For example if it is part of a research that want to explains how different parameters affect the variations in the entries numbers of NYC subway riders, this model give fair result. If the model needs to be a part in decision making model, Im sure we can build a better model easily, only by adding other parameters and or building more sophisticated model.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class

(e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

1. 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.

2. 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 different plots. Please note that a histogram of ENTRIESn_hourly is not the same as a bar chart with hours along the x-axis (which shows how ridership varies across hours of the day).

Remember to increase the number of bins in the histogram (by having larger number of bars, each with smaller width). The default bin width is not sufficient to capture the variability in the two samples.

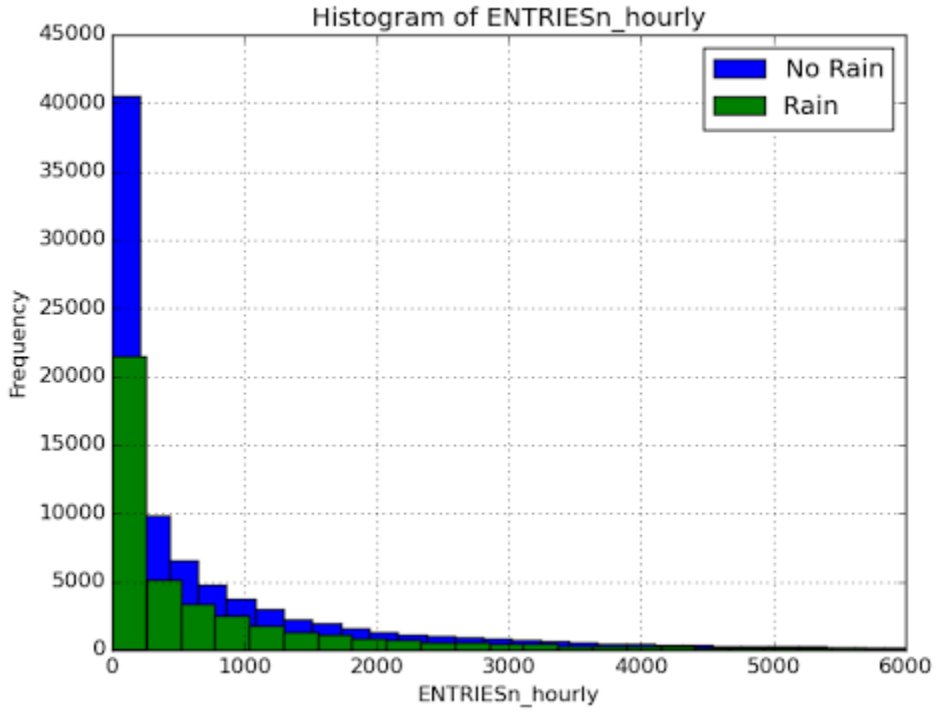

Figure 1: Frequency of hourly entries over all stations for may 2011; rainy hours are in green, non rainy hours are in blue (see legend). Looking closely we can see the difference between the two groups. Take this figure at face value we can see that the total number of entries during non rainy days are (~95000000) higher then total number of entries during rainy days (~48000000).

This difference in total number of raiders, might be due to the simple fact that in the data set, we had more non-rainy (~87000) hours than rainy hours (~44000).

2. One visualization can be more freeform, some suggestions are:
1. Ridership by time-of-day or day-of-week
2. How ridership varies by subway station
3. Which stations have more exits or entries at different times of day

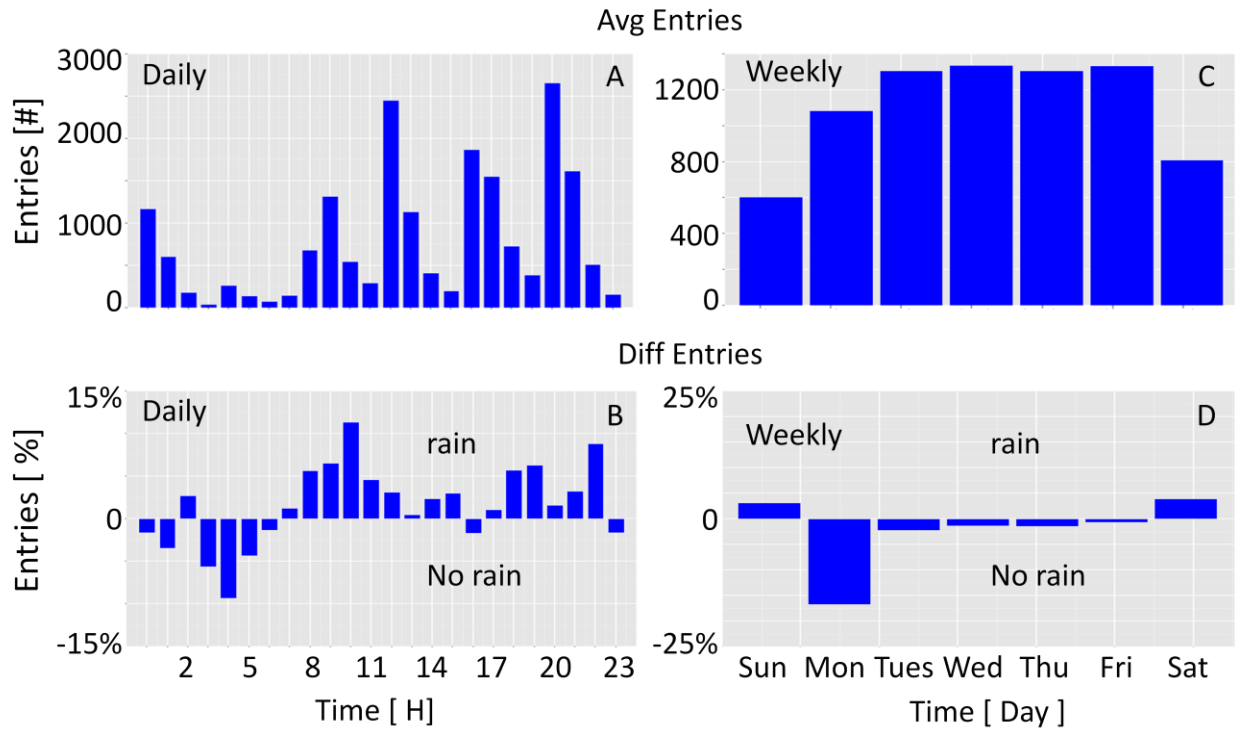

Figure 2: A. Average number of entries for each hour during the day. The daily changes in the number of entries are very noticeable. B. The difference between the average number of entries during rainy and non rainy hours, as a percent of the hourly average Entries. C, D. Same as in A and B, but the average was done on different days in the week. As for the hourly distribution, we find systematic changes between the days of the week. However, the percent daily difference indicates that rainy conditions reduce the average numbers of daily raiders, this is in contrast to the result of Mann-Whitney U-Test.

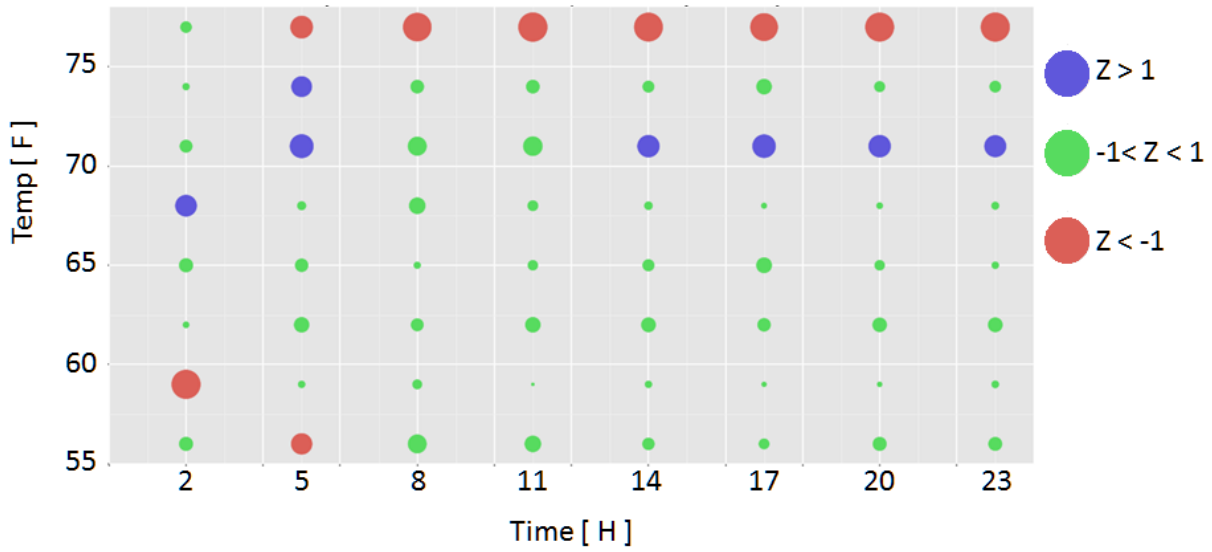

Figure 2 : Normalized number of entries (radius of the circle) as function of temperature (Y-axis) and the time in day (X-axis). To mask out the hourly change in the entries number (see figure 2A) I look at the Z-score. The average entries at a given hour is reduce from the average entries of each hour and each temperature, than divided by the standard deviation of the given hour {no_entries(h,T) – avg(no_entries(h))}/ std(no entries(h)). Colors indicate values that are larger (blue) or smaller (red) than -1 standard deviation (green indicate deviation smaller than 1 standard deviation).

Section 4. Conclusion

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

1. From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining versus when it is not raining?

The average numbers of raiders that use the subway during rainy hours in May 2011, are slightly higher (~1.3%) than during non rainy hours. The statistical test I applied on the data indicates that this small difference is significant (see section 1, question 3). In figure 2B, I show that changes imply by rainy conditions are not regular, but depend on the hour of the day. In case of a rain during the active hours (07:00- 22:00), the number of raiders increase by almost 15%. In case of a rain during the non-active hours (23:00-06:00), the number of raiders deceases by similar percent (Figure 2B). Numbers of raiders change also due to other features, for example, hour of the day (figure 2 A) or day of the week (Figure 2C), this changes can reach to more than 50 %. The linear regressions model show that weather condition explains only small portion of

the changes in the number of raiders ($R^{\wedge 2} < 0.1$), while other feature such as station number are more dominant.

2. What analyses lead you to this conclusion?

I explain that in the previous answer

Section 5. Reflection

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

1. Please discuss potential shortcomings of the data set and the methods of your analysis.

The objective of this project is to find how weather conditions especially rain is associates with raiders' number in the NY subway. Toward that, it is important to obtain broad prospective about different factors that are associates with changes in raiders' number. For example, number of raiders' changes considerably on hourly or daily bases (figure 2A, C), I can assume that similar changes can be also found on a monthly base. In this context the main shortcoming of the dataset is the small number of measurements, especially the fact that the data was obtained during only one month. Larger data set might reveal whether temperature or other features are associates with raiders number, or explain the strange result we obtain in Figure 2C.

The linear regression model could explain almost 50% of the changes in raiders' number when combining the optimal features included in this data set. However, most of the model power was based on the stations name, daily and hourly changes, while the weather condition explain less than 10% of the changes. Further investigation should include larger data set, this might produce significant information using similar analysis as I show in figure 2 and figure 3.

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