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, and 4 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**

Unfortunately I did not write down all the sources while working on the project.

# **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?

The statistical test used to analyze the NYC Subway data: **Mann-whitney U test**

One-tail or a two-tail P value : two-tailed, since it is not known which data set would be higher or lower.

Null hypothesis: That the two populations are the same(Rain has no correlation with ridership)

p-critical value: 5%

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.

**Mann-Whitney U test** was used because the data is not normally distributed. When we have such kind of data, a non-parametric test(**Mann-Whitney U test**) is best suited to give us the best answer. Another benefit of **Mann-Whitney U** test is that it does not assume any underlying probability. For comparison, **Welch’s** two-sample t-test assumes the the data is normally distributed. In order to make sure that neither data set is normally distributed a **Shapiro-Wilk test** could have been conducted as well.

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.

“Good job! Your calculations are correct.  
Here's your output:  
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)“

Mean entries with rain: 1105.446  
Mean entries without rain: 1090.279  
U-statistic: 1924409167.0  
p-value: 0.025

1.4 What is the significance and interpretation of these results?

By comparing the **means** there is ~ 1,4% (exact - 1.39%) entries in the subway when it rains then when it is not.However, that would not be enough to draw conclusions, since the number is too small. The **U-statistic** is very high which is almost the maximum value -1937202044.0( ½ of the product of the number of values in each set).The Null Hypothesis would be True if the Ustatic was ½ of the maximum value. The p-value is 0.025, which satisfies the p-critical value. By taking in consideration all of the information, we can conclude with 95% confidence that the Null Hypothesis is false(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?

**Gradient Descen**t **algorithm** was used to train the **regression coefficients**. I kept the mean normalization feature scaling , and used the default values of learning rate ( **Alpha** = 0.5, **N** = 75 iterations). The given values reached the local minimum, which was confirmed by the plot of **cost history** VS **N of iterations**.

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

Rain,precipitation,mean wind speed,hour,mean temperature. Dummy variable used was UNIT( turnstile location/identification number- categorical). UNIT variable is boolean(0,1) features with prefix ‘unit’. Each data point have a 1 in the feature that it belonged to.

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.

There were not better R^2 values than the one used in the experimentation(rain, precipitation, hour, and mean temperature). Noticed increase of r^2 value after including wind speed.

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

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

Your r^2 value is 0.479247705439

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?

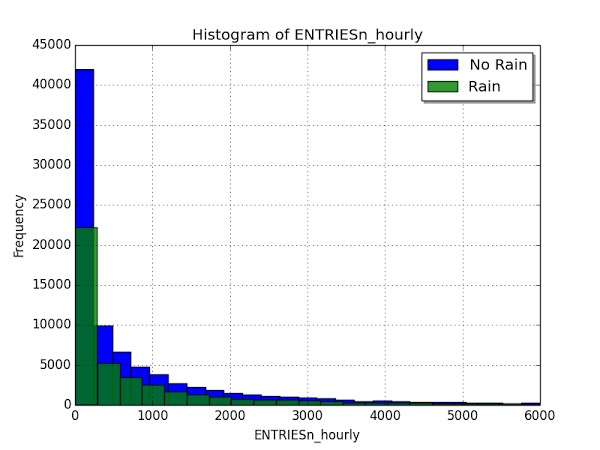
**R^2** is the the percentage of **variance** that is explained , and is measure of goodness of fit. My model explains only **~48%** of the variation. Based on the result the model would not be sufficient to make major conclusions. By including more features or utilizing polynomial regressions better results can be drawn. However, that might lead to over-fitting and the the model might not work on new dataset. **Regularization** could be good way to solve the over-fitting.

# **Section 3. Visualization**

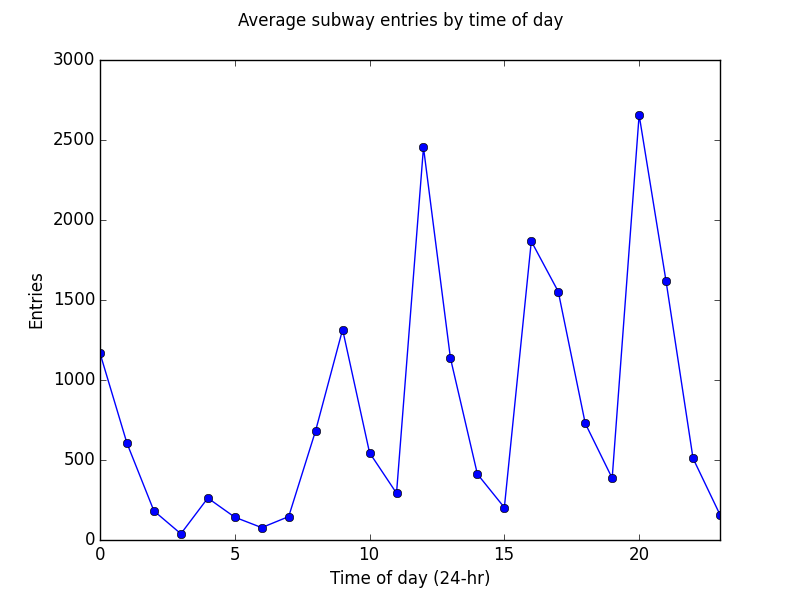
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.



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.

* 

# **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?

By simply looking at the means of both data sets, it is is insufficient to make conclusions due to the variance.

However, Mann-Whitney U test does confirm that the two sets are statistically different. With the results given from the **Mann-Whitney U test**, we can conclude that more people ride the NYC subway when it is 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.

Since positive coefficient of the rain parameter is observed, it can be concluded that the presence of rain to some extend drives the increase of ridership. The ~48% **R^2** leads toward the conclusion that the above statement might not have been the case of all data points.However, the small residual is has relatively high accuracy, having in mind the objectives. Overall, even though the means of the two datasets are not significantly different, the **Mann-Whitney U test** does show statistically significant change in number of people riding the NYC subway when it is raining vs when it is not. And I can conclude that the the rain drives the ridership of the subway.

# 

# **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.

While working with the data I noticed that there are more entries than exist. This can have effect on both rain and no rain data sets.

An increased data size and being normalized by location ID could increase the confidence of Mann -Whitney U test and the regression model. Based on the results with ‘UNIT’ column, it could be seen that some stations were more active than others.The Mann-Whitney U test skipped this, and only looked in the entries for rain and no rai. Maybe if examining the same station over a period of time might give more accurate results.

For the purpose of this study the linear regression model was adequate. Over all the subway ridership has limitations having in mind that there is certain amount of people that can ride it at the same time, The inclusion of more features as mentioned in one of the sections could have increased the accuracy of the model.

Considering the extreme, subway ridership certainly has an asymptotic limit; only so many riders can get on the subways! As mentioned in Section 2.6, the inclusion of more

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

Not at this moment