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**

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.

**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?
* The Mann-Whitney U test was used to analyze the NYC subway data. I used a two-tailed P value. The null hypothesis is that the mean of the entries with rain is equal to the mean of the entries without rain. The p-critical value is 0.0498.
  1. 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.

1. The distributions of the data sets for the two conditions are not normal, hence the Mann-Whitney test is applicable.
   1. 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.

* Results from the statistical test are as follows:
* P-value: 0.0498
* Mean with rain: 1105.4464
* Mean without rain: 1090.2788
  1. What is the significance and interpretation of these results?
* The two-tailed p-value is less than 0.05. This means that the null hypothesis that the mean number of hourly entries with rain is equal to that without rain is rejected.

**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. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

Ans. 1. OLS using Statsmodels. OLS is more accurate than Gradient descent.

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

Ans. Apart from dummy variables, 'rain' and 'Hour' were the features used in the model. The following are the dummy variables used in the model:

unit\_R170

unit\_R022

unit\_R179

unit\_R029

unit\_R012

unit\_R046

unit\_R023

unit\_R047

unit\_R293

unit\_R055

unit\_R084

unit\_R011

unit\_R178

unit\_R452

unit\_R001

unit\_R108

unit\_R010

unit\_R195

unit\_R021

unit\_R102

unit\_R018

unit\_R033

unit\_R175

unit\_R177

unit\_R051

unit\_R138

unit\_R017

unit\_R057

unit\_R168

unit\_R080

unit\_R050

unit\_R127

unit\_R025

unit\_R141

unit\_R163

unit\_R044

unit\_R019

unit\_R031

unit\_R248

unit\_R032

unit\_R083

unit\_R461

unit\_R240

unit\_R235

unit\_R208

unit\_R014

unit\_R053

unit\_R020

unit\_R142

unit\_R081

unit\_R028

unit\_R222

unit\_R131

unit\_R105

unit\_R113

unit\_R144

unit\_R167

unit\_R079

unit\_R463

unit\_R035

unit\_R132

unit\_R176

unit\_R300

unit\_R158

unit\_R164

unit\_R097

unit\_R101

unit\_R116

unit\_R027

unit\_R110

unit\_R201

unit\_R160

unit\_R041

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.”
* ‘rain’ is the first variable that comes to mind to decide whether people might use the subway or not. If it rains, then people might decide to use the subway. Also, ‘Hour’ of the day is another variable that helps in predicting usage of subway as people would use the subway more often in the mornings and the evenings as compared to late night or during the afternoon. All the dummy variables were included as they categorize the data and hence would help in identification.
* To filter out the dummy variables, I ran a full model with all the dummy variables. The resulting model summary gave the t-values and p-values of the dummy variables. Inclusion of dummy variables with higher t-values improved the r-squared significantly. I sorted the variables in descending order of t-values and included dummy variables in the model till the r-squared value crossed the threshold mentioned of 0.4.

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

* Coefficients of the non-dummy variables:

‘const’: -18.9345

‘rain’: 62.6158

‘hour’: 59.1683

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

0.401

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 R2value?

* R2 value of 0.401 means that the model explained 40.1% of the data fit the regression model line. This linear model to predict ridership is not appropriate for this dataset, given the R2 value. It could be that some variables given in the dataset which could increase the R2 value have not been identified. There is also a chance that some important variable has not been captured during data collection too.

**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.

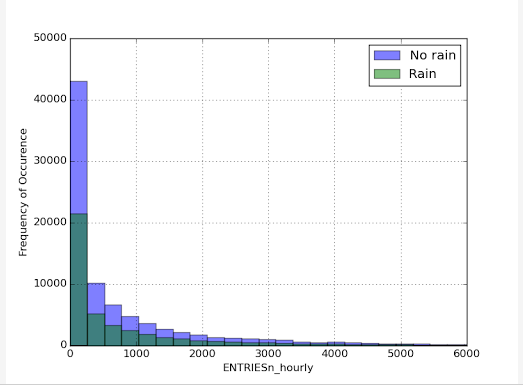
3.1 One visualization should contain two histograms: one of  ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days.

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.



* A histogram of the hourly number of entries is plotted when there is rain and when there is not. The histograms are not normally distributed and have their peaks near 0.

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



* The graph shows the plot of the mean pressure variation versus dates. It shows how the mean pressure variation varies on different days.

**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?

Ans. More people ride the NYC subway when it rains as compared to when it does not. The low two-sided p-value of 0.0498 indicates that the null hypothesis that the mean subway number of hourly entries with rain is equal to that without rain is rejected. As per the linear model, the positive coefficient value of 62.61 for variable ‘rain’ shows that approximately 63 more people would ride the NYC subway when it rains. The effect is not as significant though as compared to those of the dummy variables. For the dummy variables included in the model, the range of the coefficients vary from -1862.95 to 8953.54 which means that they are more important in deciding the number of people riding the NYC 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.

Ans. As mentioned above, this inference was drawn from both the p-value of the statistical test which helped in rejecting the null hypothesis that the mean subway number of hourly entries is the same with and without rain, and the regression analysis which highlighted the fact that subway ridership increases by 62.61 on average when it rains.

**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:

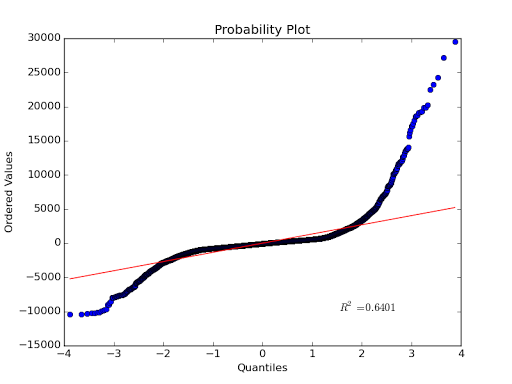
1. Dataset,
2. Analysis, such as the linear regression model or statistical test.

Ans.

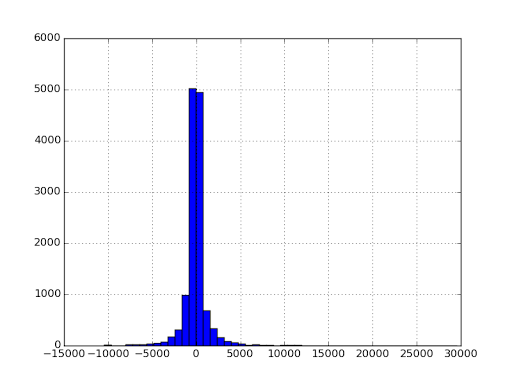
The shortcomings of the methods are as follows:

1. The data is not captured for all the hourly intervals for each of the unit types indicating that the data is not complete. This causes a hindrance in building reliable models for our analysis.

2. The Q-Q plot of the residuals clearly shows large tails and hence is an indication of non-normality.



The residual plot is also not normal. This shows that the model is not very appropriate. Non-linear terms may need to be incorporated into the model to address it.



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

Ans.