Suresh Aswathnarayana
Suresh.aswathnarayana@live.com

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? What is the null hypothesis?

   Mann-Whitney U-Test. One-tail test. Null hypothesis that people use the subway more when it is raining.

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

   The NYC data set is not uniformly distributed and the Mann- Whitney U-Test is applicable to such non uniform distributions.

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.

   With_rain_mean= 1105.4464

   Without_rain_mean= 1090.2788

   U = 1924409167.0

   P= 0.024

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

   p= 0.024 is less than alpa = 0.05 and more than alpha = 0.01 hence the p = 0.024 is approximately 98%. The mean difference between the with_rain and without_rain is very small. Hence we reject the null hypothesis and say that people do use the subway less when it is raining.

Suresh Aswathnarayana
Suresh.aswathnarayana@live.com

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)
2. OLS using Statsmodels
3. Or something different?

   Gradient descent Algorithm

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

   Features used ['rain', 'precipi', 'Hour', 'meanwindspdi', 'meandewpti', 'meantempi'] and Dummy Variable [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 R2 value."

   I used these features as the inclusion of my hour, meanwindspeed, meandew, feature in my model improved my Pearson's coefficient from 0.40 to 0.47. These features provided good correlations to the target function.

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

   0.4741

5. What does this R2 value mean for the goodness of fit for your regression model?

   R2 of 0.4741 is a good fit using a linear regression model.

Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

No a polynomial model with mean feature scaling would give us a better model.

Section 3. Visualization

Suresh Aswathnarayana
Suresh.aswathnarayana@live.com

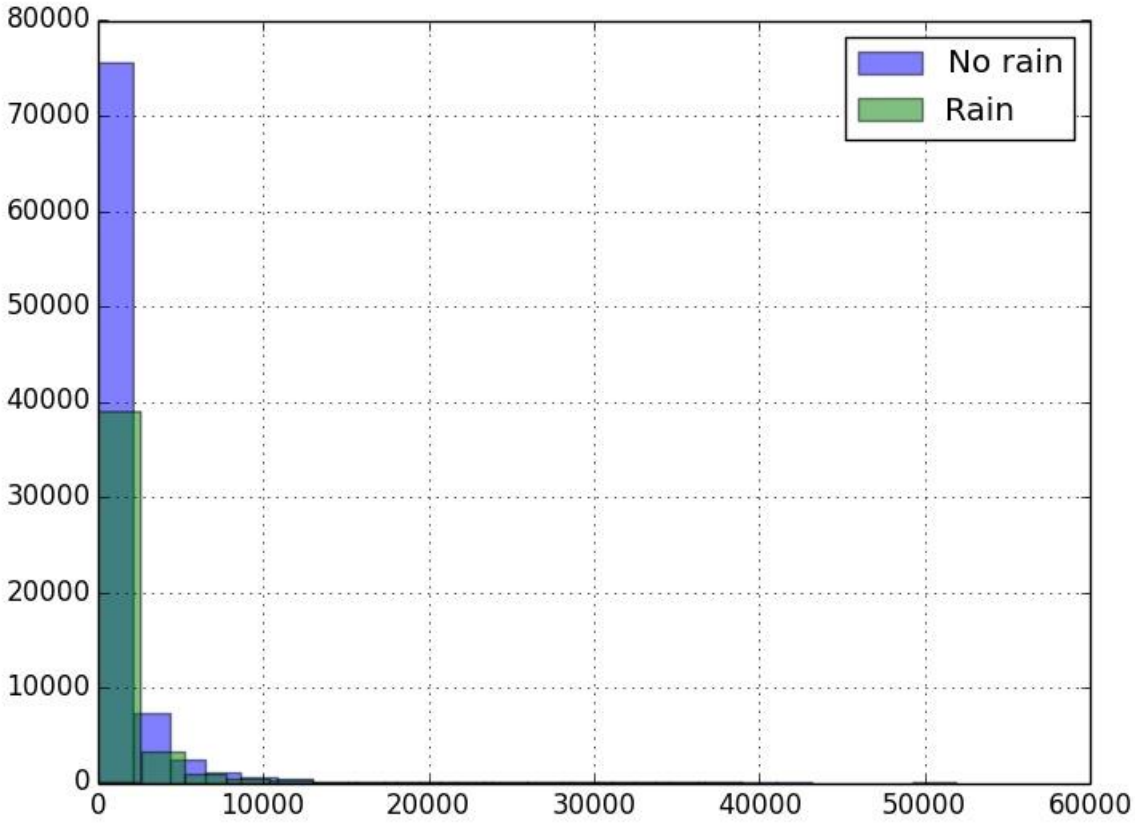
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

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.

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 different plots.

For the histogram, 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, you might have one interval (along the x-axis) with values from 0 to 1000. The height of the bar for this interval will then represent the number of records (rows in our data) that have ENTRIESn_hourly that fall into this interval.

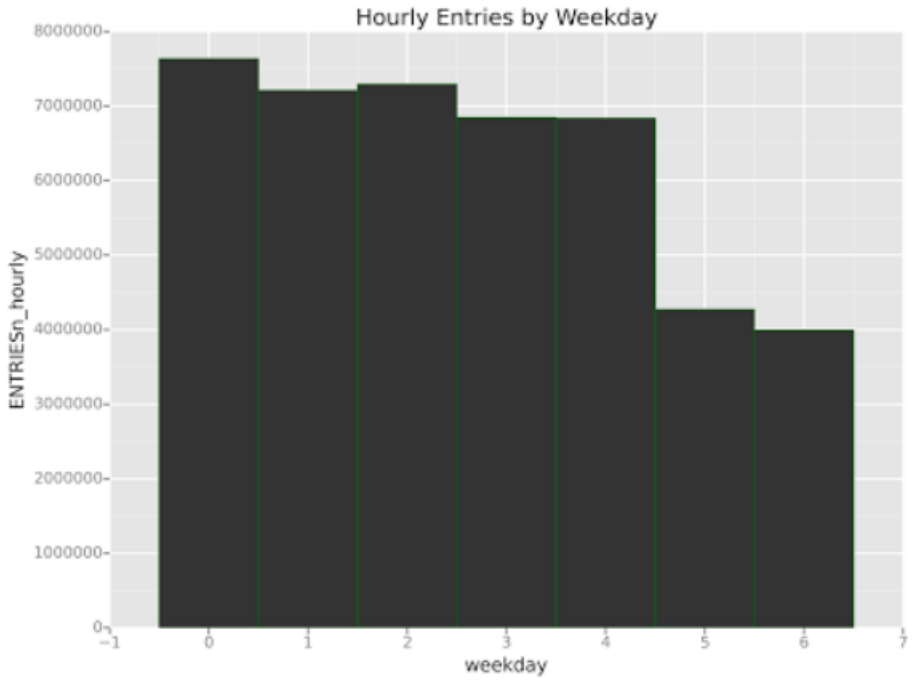


Suresh Aswathnarayana
Suresh.aswathnarayana@live.com

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.

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



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 mean for the ridership during a rainy day was more than the mean ridership during no rain days. But looking at the histogram plot one can also conclude that the ridership during a no rain day is more than the ridership during a rainy day. The small probability value of 0.024 which is less than the alpha of 0.05 for a one tail test shows that the population medians are different although looking at the histogram figure the normal distribution overlap one another. We can conclude that the ridership is more during no rain day as compared to a rain day by looking at the histogram plot.

2. What analyses lead you to this conclusion?

Suresh Aswathnarayana
Suresh.aswathnarayana@live.com

The low probability value of Mann-whitney test and the histogram plot helps us make the conclusion that the ridership on a non-rainy days is more than the ridership during rainy days.

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.

   Most of our analyses were based on linear regression models. A polynomial or a multivariate fit or polynomial fit would have been better in achieving better predictive capabilities.

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

   No.