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?

Welch's *t*-test. One Tail P value. The null hypothesis to check whether two samples means are equal.

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.

This statistical test let us assume that both sample does not have the same sample size or variance.

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.

t : 5.0428827476194309
p-values: 4.6414024316324798e-07
mean of sample 1: 2028.1960354720918
mean of sample 2: 1845.5394386644084

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

We reject the null hypothesis and both sample means are not equal, our p value is less than our p critical of 95%.

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:

Gradient descent

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

'rain', 'fog', 'wspdi', 'hour', 'meanpressurei', 'meantempi'. Yes, UNIT

3. Why did you select these features in your model? We are looking for specific reasons that lead you to believe that

rain - since we are investigating rain day versus non rain day, it best to include it as our feature

fog - foggy weather may affect people's traveling behaviour

wspi - I observed that when I added in wind speed, it increases my overall R^2 value.

hour - hour should be included in the feature as entries_hourly is related to it.

meantempi - Temperature may be related to rain or non rain and including this feature daily average tempi, also increases my R^2 value.

meanpressurei - I used this feature daily average pressurei , because it might be related to the temperature.

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

0.46156312679019895

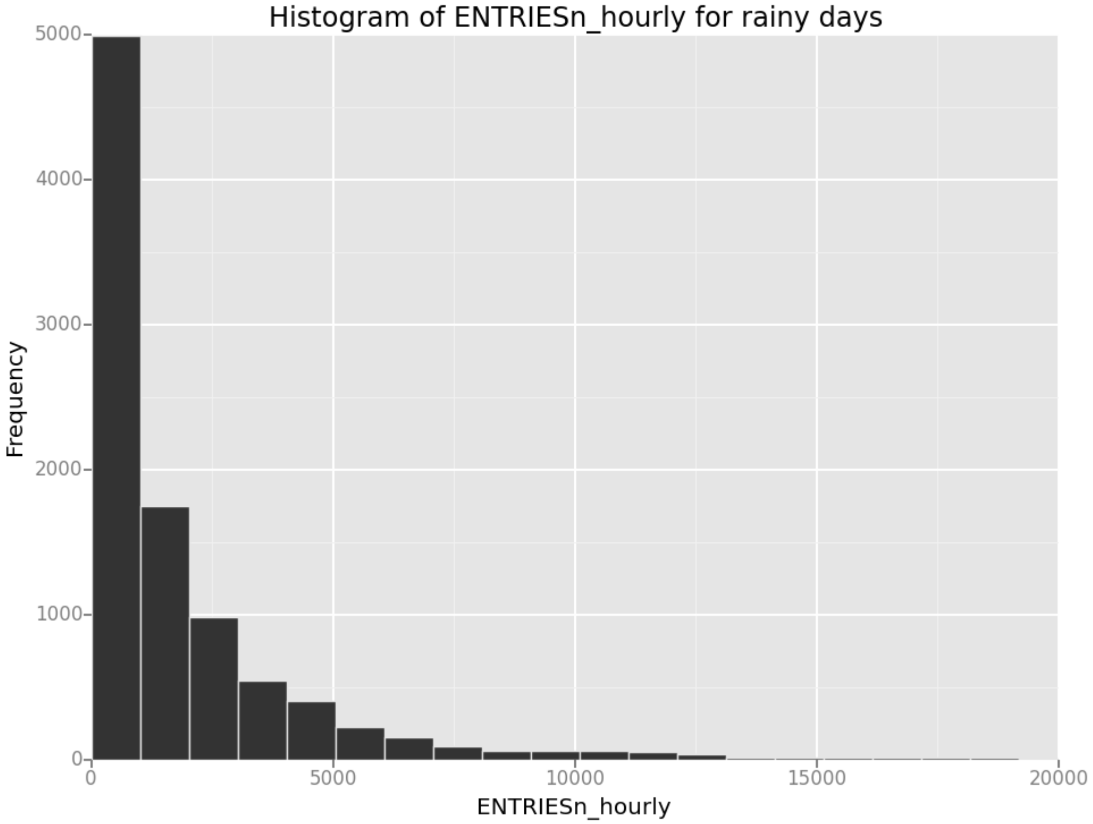
5. What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

Its hard to tell. Even though we get a high R2 value which is greater than zero. That does not mean that the model fits the data. The coefficient estimates and predictions may be biased.

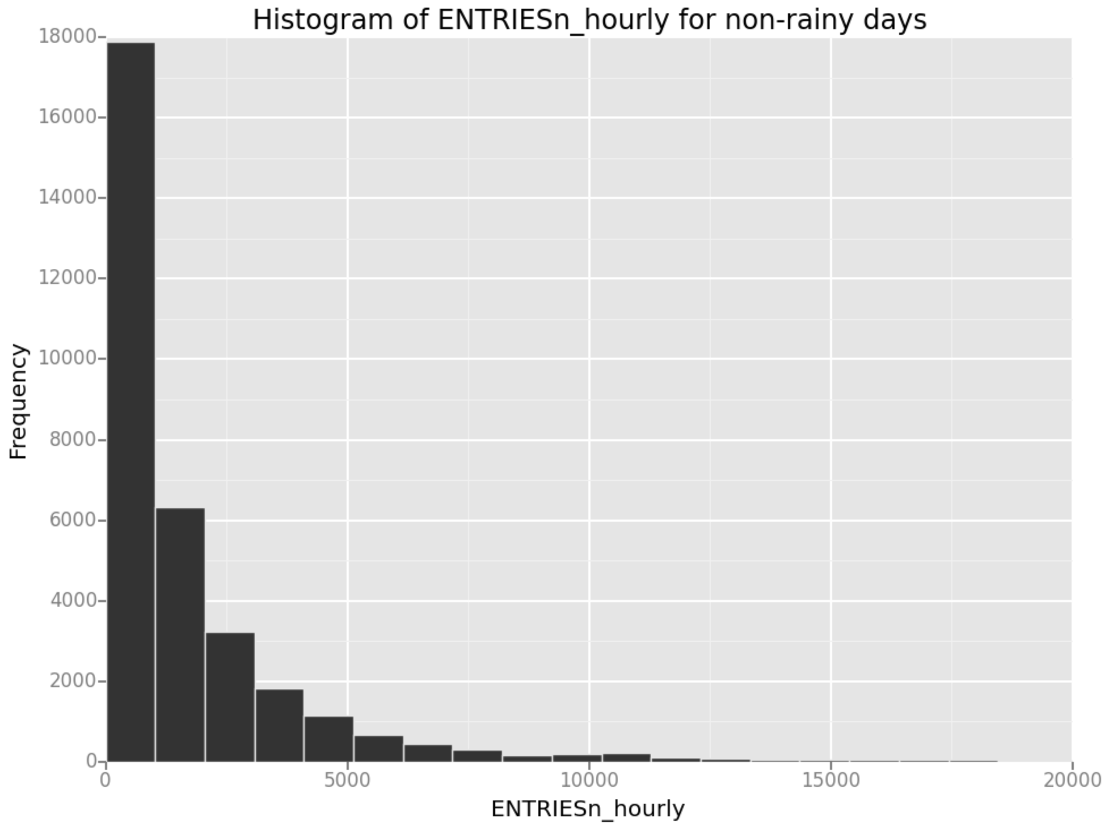
Section 3. Visualization

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. One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

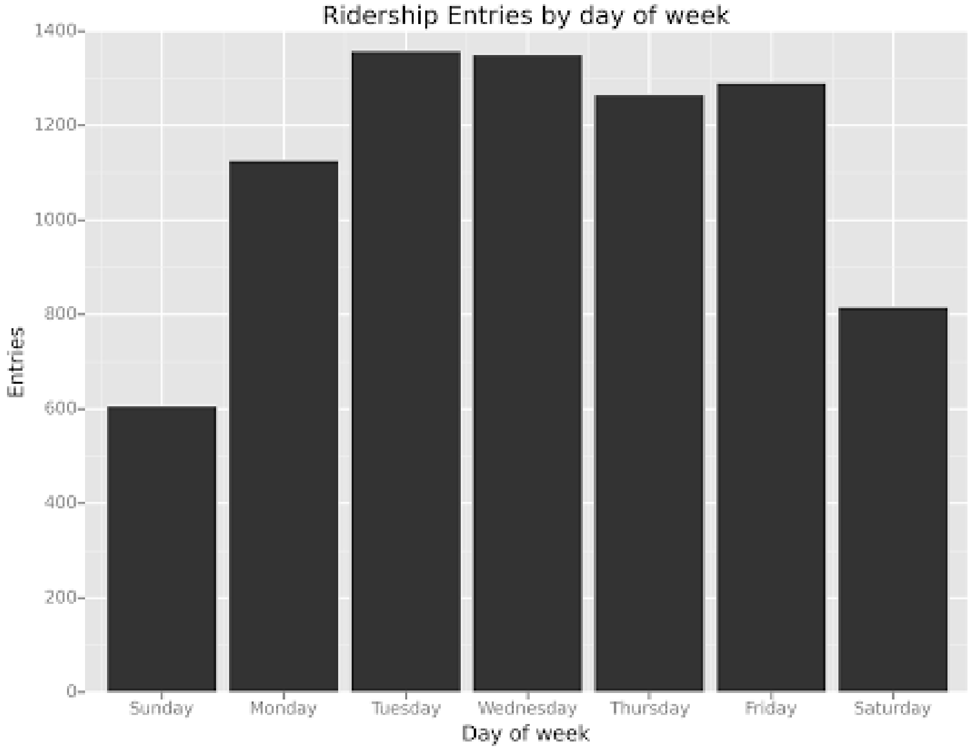


The frequency are much lower when compare to non rainy days.



There are more entries per hour during non rainy days.

2. One visualization can be more freeform:



More people ride the subway on weekday compare to weekends.

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?

Nope.

2. What analyses lead you to this conclusion?

From the histogram, we can see that on rainy day's frequency on the first bin is way lower than non rainy day's frequency. Also, by changing the bin sizes. non-rainy day frequency is still higher than rainy day.

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 data set might not be consistent like some turnstiles have reported Entriesn_hourly for a few hour intervals, or miss report of data for some weather station. Gradient descent is easy and faster to implement and it is scalable. If we were to change our model we can still use it back.

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