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-Whitnet U statistics.*

*Null Hypothesis: two sample (ridership on rainy vs non-rainy days) comes from the same population.*

*I used two-tail P value since the deviation of the estimated value was possible in both direction.*

*Null hypothesis is the general statement that two variables do not have any relationship. In this case: there is no relationship between ‘rain’ and ‘ENTRIESn\_hourly’*

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.

*Data don’t seem to be clear fit for normal distribution.*

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.

*Null hypothesis was rejected and we assume that there is a relationship between ‘rain’ and ‘ENTRIESn\_hourly’.*

*with\_rain\_mean: 1105.44*

*without\_rain\_mean: 1090.27*

*U: 1924409167.0*

*p: 0.025*

*2\* p = 0.05*

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

*2\* p <= critical value*

*2 \* 0.25 <= 0.05*

*We can confidently (95% confidence level) assume that rain has impact on ridership as there are not enough evidence that two samples are coming from the same population. We can assume there is a relationship between ‘rain’ and ‘ENTRIESn\_hourly’.*

# 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:
2. Gradient descent (as implemented in exercise 3.5)
3. OLS using Statsmodels
4. Or something different?

*I used OLS from Gradient descent*

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

*X = df[['rain', 'precipi', 'fog', 'meantempi']]*

*Features: Rain, precipi, meantempi, fog*

1. 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 suspected each would be related to our dependent variable. More specifically, rain, precipi, fog and meantempi are all weather conditions and weather can influence people’s decision whether to choose publication transportation or other means of transportation such as car, bike, walking…*

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

R2  = 0.07

*Coefficients of independent variable:*

*Constant: 5.34640305*

*'rain': 21.65567364*

*precipi': 420.88114853*

*‘fog': -52.42722999*

*'meantempi': 1100.60865797*

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

*It shows that to what extend our prediction is close to the data or in other words how well are model is fitted to data and can explain the variance in data.*

1. Do you think this linear model to predict ridership is appropriate for this dataset, given this R2value?

*The model 2.4 is not very effective in explaining the variance in data as R-Squared is very close to zero which suggests that the model cannot fit the data at all.*

*I did more investigation to use other variables to be able to fit the data better. I use ‘Hour’, ‘weekday’, and ‘Unit’ (to get day of week, I used the DATEn variable to find out which day of the week it was and the created a dummy variable to say whether it is weekday or not ).*

*I rebuilt the model using these dummy variables to check if I can fit the better and it worked! I could get a much better R-squared = 0.517 which is pretty good! But still not enough!*

*R-squared tells us to what extend ‘Hour’, ‘rain’, ‘UNIT’ and ‘Weekdays’ can explain the variance in the data or in other words to what degree these variables are related to ridership.*

*The reason I chose these variables is because, I think:*

*‘Hour’: Time of the day is important factor is ridership. There are certain hours that people are more likely to use subway… like rush hours*

*‘DATEn’: Depending on which day of the week it is, or if it is weekend or not, can help to make a model to fit the data better. Because I suspect that during weekdays people are more like to use subway to get to work than weekends.*

*‘UNIT’: Not all subway stations are the same crowded. Some specific ones might be more crowded than others. Maybe the ones in downtown area, or the ones that are key locations… this can be helpful factor as well to consider.*

*Since there are so many variables (dummy ones) I won’t talk about each coefficient separately. There are 492 variables! However very briefly:*

*Weekday coef.: 2.68961565e+02. This shows that whether it is weekday or not has positive correlation with ridership*

*Hours coef.: -2.91547569e+01 -1.24267409e+02*

*-5.41004944e+01 -4.05784502e+01 -2.56446745e+02 -2.35631547e+02*

*-7.37690633e+01 -8.94638106e+01 -1.70537186e+02 3.73209065e+01*

*-2.31678283e+01 -4.19070751e+01 2.85019322e+02 2.07459257e+00*

*-9.52615516e+01 -4.01223213e+01 1.45835026e+02 9.44080846e+01*

*1.52452865e+01 9.10131189e+00 3.53897750e+02 1.44442572e+02*

*-1.50851379e+01 -3.67036412e+01*

*This is the coefficient of dummy hour variable. It starts from 00:00 (midnight) to 11:00pm. Overall it says that ridership is higher in rush hours (9 in the morning or late afternoon and evening have positive correlation with ridership)*

# Section 3. Visualization

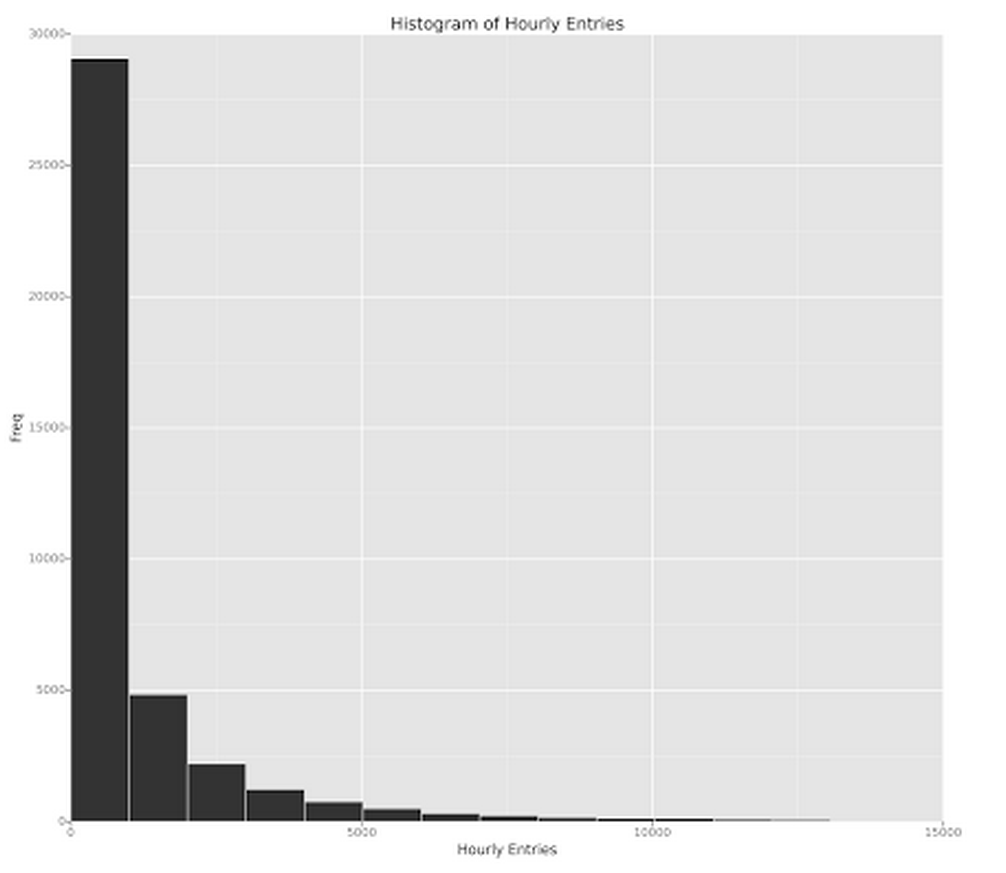
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

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.

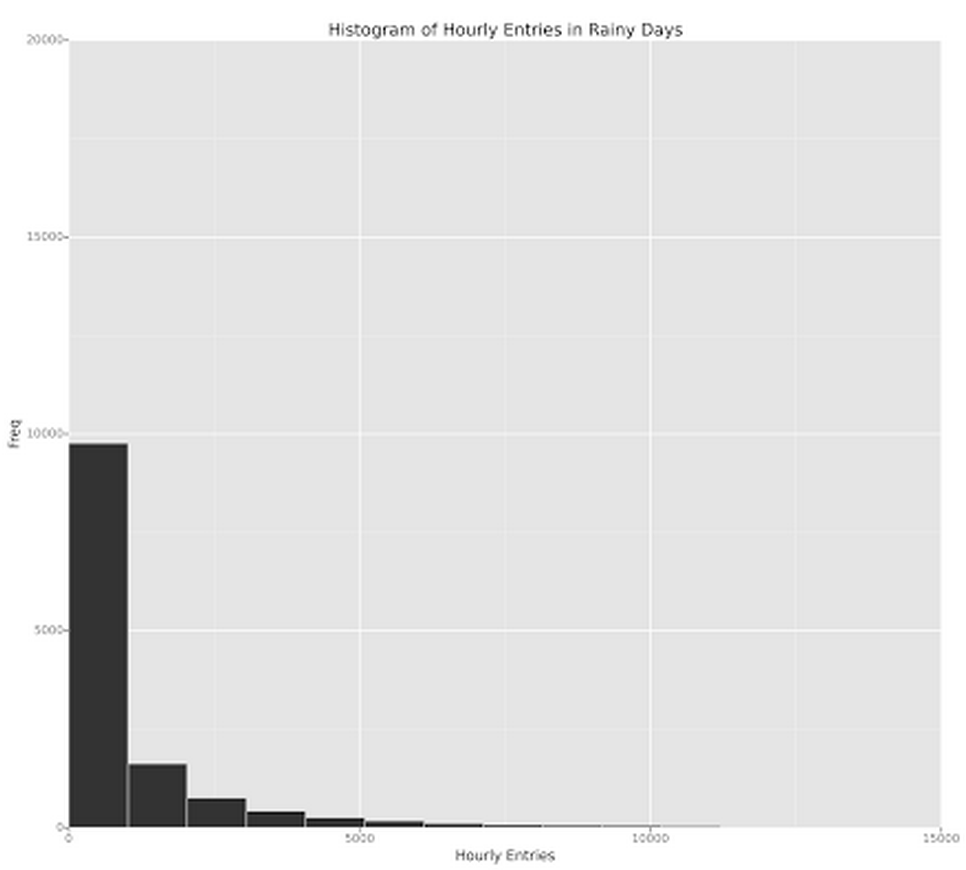
1. One visualization should contain two histograms: one of  ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days. You can combine the two histograms in a single plot or you can use two different plots.

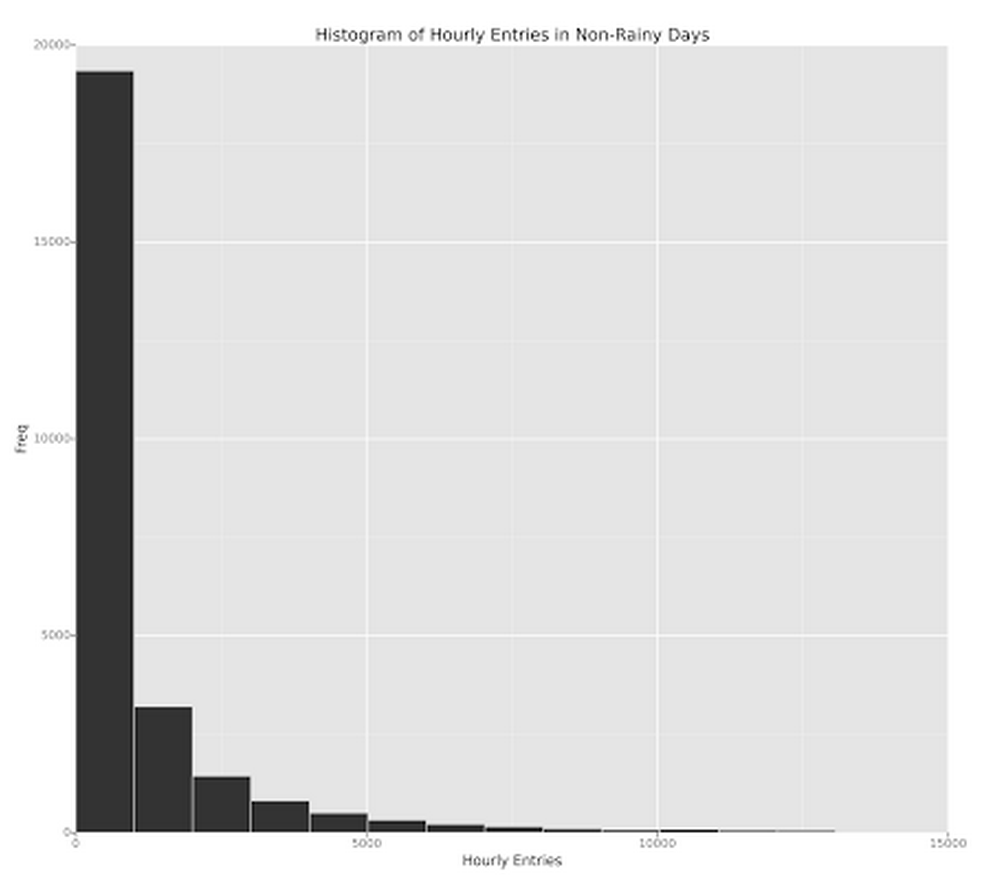
For the histogram, you should have intervals representing the volume of ridership (value of ENTRIESn\_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, you might have one interval (along the x-axis) with values from 0 to 1000. The height of the bar for this interval will then represent the number of records (rows in our data) that have ENTRIESn\_hourly that fall into this interval.

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.



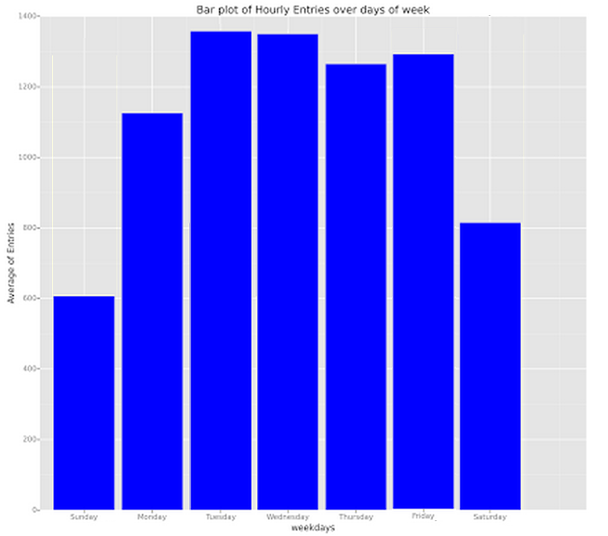
*The usual case for subway is hourly entries between 0 to 1000 people!*





*Based on the above two histograms, we can conclude that in both rainy and non-rainy days typical average ridership is between 0 to 1000. However in non-rainy days it is even more frequent.*

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



*This bar plot shows the average of entries hourly per each day of week. In weekend it is in its lowest*

# 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?

*Yes. There are different methods to test this. One way is to model the data and see the impact of ‘rain’ on the ridership. Is it positive or negative? Is it significant or no?*

*Another method is using t-test or Mann-Whitney U statistics and compare the mean of each group (with rain or without) and then check the hypothesis.*

*Here I would show the evidence using latter method:*

*Mean for ridership when it is raining: 1105.44*

*Mean for ridership when it is not raining: 1090.27*

*Here is the result of Mann-Whitney U statistics*

*U: 1924409167.0*

*p: 0.025*

*0.5 <= 0.05 (2 \* p <= critical value)*

*With such a small P we can conclude that the two samples (ridership on rainy vs non-rainy days) are not coming from same population (with 95% confidence level) meaning ‘rain’ has impact on ridership and since the mean for rainy days is higher, the impact should be positive or in other words samples of ridership on rainy days tends to be larger than the other one!*

1. What analyses lead you to this conclusion?

*Using Mann Whitney U-stats the null hypothesis is rejected and we can assume that the distribution of the two sets are different from each other. P value for Mann-Whitney stat is 0.024 (2\* p is 0.05) which is less than/equal to 0.05 (critical value) therefore the mean is significantly different.*

*Since rainy group has higher mean and also ‘rain’ has a positive correlation with ridership, we can say that people tend to ride the NYC subway more when it is raining.*

# 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.

*Data set was not complete:*

* + *We didn’t process all the possible data as we had limitation in computation, therefore only subset of the data was chosen.*
  + *Hours of collecting data were not really consistent. It would be better if we had data for every hour.*
  + *It was only limited to specific month so we cannot conclude about general behavior for the whole year.*
  + *Linear regression doesn’t seem to be the best method to fit the data. As mentioned in section one with using many variable (temp, pressure, fog, wind, Hour, day, Unit…) I got .51 for R-Squared which suggests that the model is not a great predictor. Maybe polynomial would be a better option, or regression tree!*

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

Published by [Google Drive](https://docs.google.com/)–[Report Abuse](https://docs.google.com/abuse?id=16T3kirC0IxvtfxlZb7n5kOz5xFF_JTwrG31J2OZj8KM)–Updated automatically every 5 minutes