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, 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 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.

Cross Validated -- T-test for non normal when N>50?
Wikipedia – Mann Whitney U Test
Engineering statistics handbook – Are the model residuals well-behaved?
eHow – the disadvantages of linear regression
StackOverflow – how to adjust x axis in matplotlib

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?

ANSWER:
I used Mann Whitney – U Test. This test returns a one-tail P value, however, I used a two-tail test, thus I doubled the result of the p value.
The null hypothesis [H_0: P(mean entries with rain > mean entries without rain) = 0.5]
My chosen p-critical value is 0.05

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.

ANSWER:
The Mann-Whitney U test is applicable to the set as it does not make any assumptions related to the distribution, and in fact, It has greater efficiency than the t-test on non-normal distributions (which is the case in our dataset). Also, other

requirements/assumptions of the test seem to be met:
1. The sample drawn from the population is random
2. Independence within the samples and mutual independence
3. Ordinal measurement scale

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.

ANSWER:
Mean with rain: 1105, Mean without rain: 1090, p-value: 0.025
Note: in this case I will be using a two-sided test, thus I need to double the value of the p-value obtained from the Mann-Whitney – U test (~0.05) to compare it against the p-critical value of 0.05

1.4 What is the significance and interpretation of these results?

ANSWER:
Given that the obtained p-value for a two-sided test is (just) below the critical value of 0.05, we reject the null hypothesis in favour of the alternative hypothesis, which suggests that the population with rain is different than the population without rain. In other words, more people take the subway, on average, during rainy days than they do when there is no rain. This result is statistically significant

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

ANSWER:

I opted for the Gradient descent method

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

ANSWER:
I chose 'rain', 'precipi', 'Hour', 'meantempi' and 'UNIT' as the dummy variable

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 R^2 value.”

ANSWER:
I chose those features as I thought that these were aspects that would probably have an impact on the subway ridership. In a sense, I used my personal experience as a starting point. I know that I decide whether to take the subway in function of the temperature and if it’s raining. Also, I considered that the ridership would be affected by the hour of the day as people normally end their workday around the same time and a number of them take this sort of public transport. Lastly, not all stations are equally busy, some of them will be more used than others.

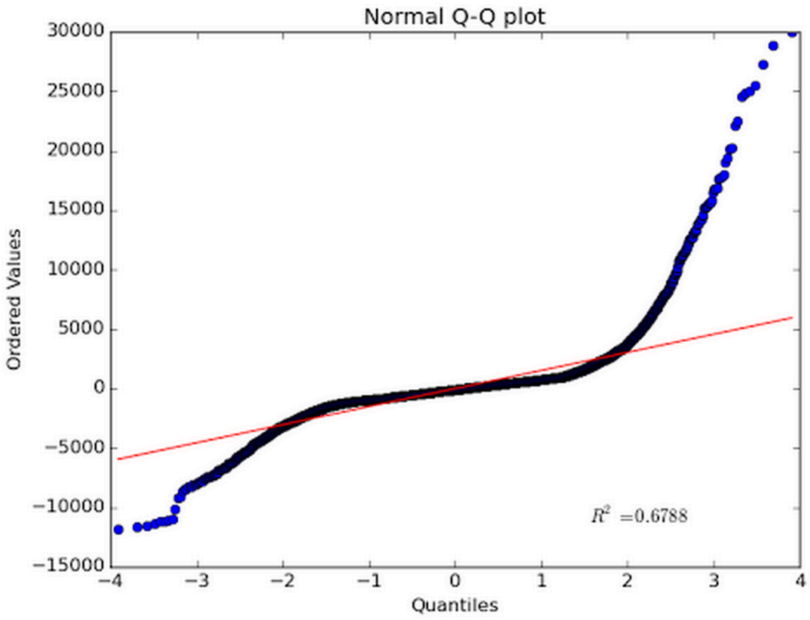
2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
ANSWER:
‘rain’ = 2.92
‘precipi’ = 14.65
‘Hour’ = 467.71
‘meantempi’ = -62.22

2.5 What is your model’s R^2 (coefficients of determination) value?
ANSWER:
The R2 value I obtained was: 0.464

2.6 What does this R^2 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 R^2 value?
ANSWER:
The R^2 is telling us that the model explains around 46% of the variance in the data. The number might seem somewhat low (as the maximum is 100%). However, in this case we are trying to model a human behavior, thus the value we might expect is probably not in the high range (>60%). On the other hand, a low R^2 is most problematic when you want to produce predictions that are reasonably precise.

Another point of concern here is related to the residuals. When we do a residual analysis using a histogram, we find that there is a long tail particularly on the positive direction, these large residuals suggest that a linear model might not be the best fit. This is further reinforced by the graphical result from the QQ plot (depicted below) which shows that residuals are not normally distributed.

All in all, the result of R^2 and the residuals suggest that perhaps the relationship of the data might not be linear, thus our current linear model might not be the most accurate in predicting ridership but can be a good first approximation. Nonetheless, we should explore other model alternatives that are not linear if we want more accurate explanatory power.

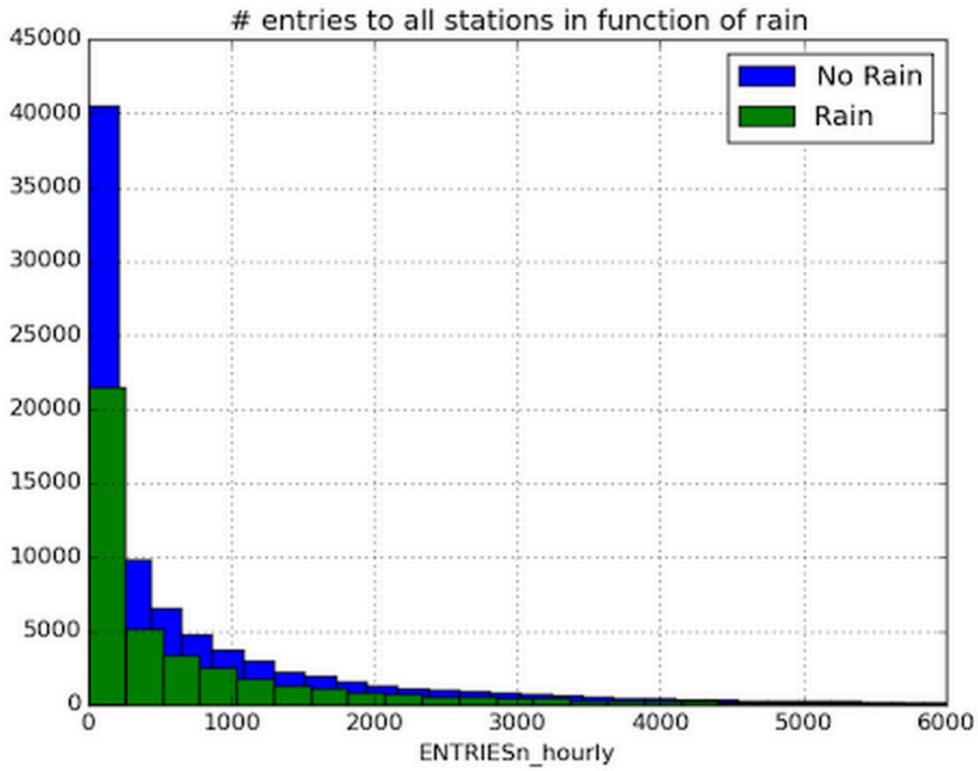


Section 3. Visualization

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.

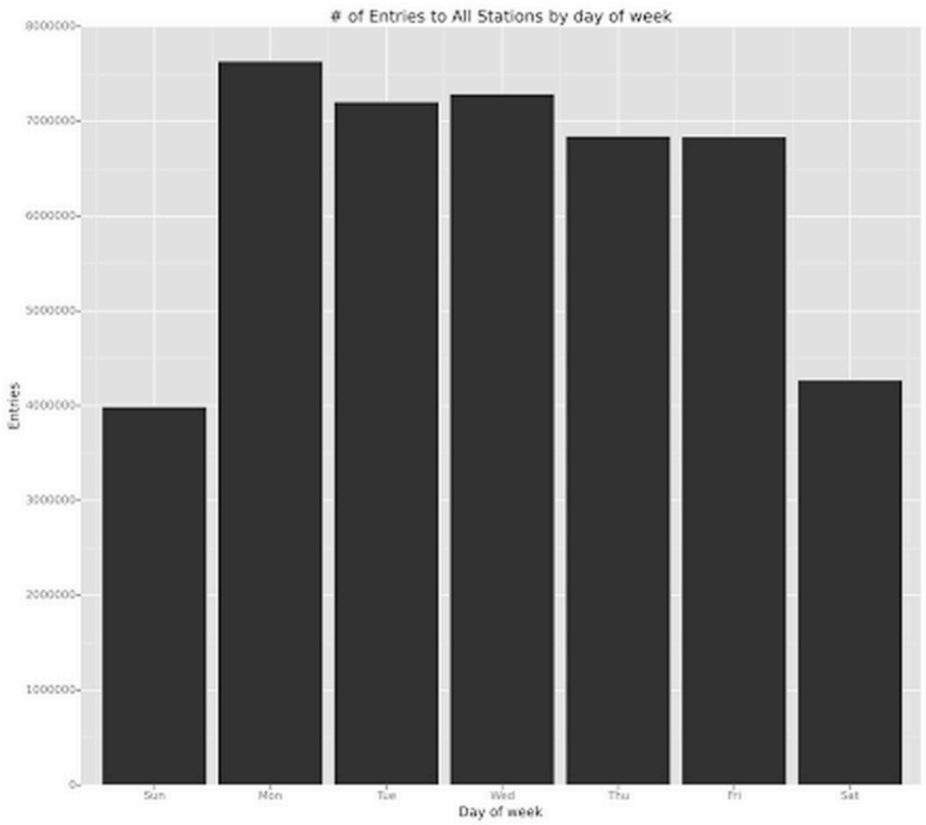


- COMMENTS:

  The histogram shows that data is not normally distributed, it has a noticeable positive skew. In other words, most of the entries per hour (i.e. ENTRIESn_hourly variable) are below 10,000 but there are a few observations where the number of entries per hour are almost 60,000. This distribution is seen across both type of datapoints, those when it rained and those when it didn't.

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



COMMENTS:

The graph above depicts the total number of entries per day of the week. We can see that weekends (Sat-Sun) have the lowest volume of ridership. In contrast, Monday is the day of the week when the number of people using the subway is at its peak.

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?
ANSWER:
Based on the analyses conducted to the basic NYC subway data set, we can confirm that on rainy days more people choose to take the subway compared to non rainy days. A detailed discussion of the analyses used to derive this conclusion are found in the next answer.

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.
ANSWER:
The analyses that lead to this conclusion are the following:

Firstly, the statistical test applied to the mean of hourly entries showed that the difference in means, between rainy and non rainy days, is statistically significant (1105 vs 1090 with a p-value below 0.05). Secondly, we fit a linear regression model whose output suggested an increment in ridership (2.92) when it rainy and moreover, an increment based on the amount of rain (as measured by the variable 'precipi' which had a value of 14.65). The goodness of fit (R^2) from this model is on the lower half but it can still be deemed relevant. Also, the graphical analysis of residuals suggests that our model is appropriate in the sense that there is no hidden structure in the data.

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**,

   ANSWER:

   The first potential shortcoming of the dataset we have been provided is that the period of analysis is only one month long (May 2011). The issue here is that we may not be taking into account a possible seasonality effect. For example, it could be the case that, for some reason, people take the subway on rainy days comparatively more than they would do any other month of the year. Or perhaps in winter they take the subway most of the time irrespective of the rain because it's too cold to walk. Similarly, having only a single year might pose a similar problem in terms of seasonality. Perhaps there's a change in attitude towards taking the subway on rainy days, or new technologies (better raincoats?) have been adopted which affect the propensity to take the subway on a rainy day.

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

   ANSWER:

   Linear regression models have a number of shortcomings, some of which can be ameliorated. For example, the regression models are sensitive to outliers. In the case of our dataset this could mean that perhaps there are a few observations (hourly entries) which for some reason were too high on a particular rainy day (maybe a concert). This could affect the values that the model outputs. Another potential shortcoming of the model could be that some features have a non-linear relationship. Consequently, if we try to fit a linear model to a variable that has an inherently non-linear relationship, the values obtained will probably not be accurate.

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

No additional comments at this moment…