# Analyzing the NYC Subway Dataset

## by Jianniao Cai

### Answers

## 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.

- **Difference between one-tailed and two-tailed tests**
- **Counting Rows**
- **pandas.DataFrame.append**
- **pandas.get_dummies**
- **SQL Group by**
- **Mann-Whitney U Test wiki**
- **SGDRegressor**
- **R2 value**
- **Goodness of fit**

## 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?

- I have used Mann-Whiney-U-test to analyze the NYC subway data
- A two-tail P value was used, because we don't know which sample mean is higher than the other sample mean.
- The null hypothesis is that there is no difference between the mean of entries with rain and the mean of entries without rain.
- My 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.

- Because the data is not normally distributed and two samples come from the same population.

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.

- P-value: 0.0249999
- The mean of entries with rain: 1105.4463767
- The mean of entries without rain: 1090.2787802

1.4 What is the significance and interpretation of these results?

- A significance level is set as usually to 0.05.
- Because the p-value is equal to the significance level, i.e., two samples are significantly different at 5% level, the null hypothesis 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:
OLS using Statsmodels or Scikit Learn
Gradient descent using Scikit Learn
Or something different?

- I used OLS with help of Statsmodels to get the coefficients theta in my regression model.
- I also used gradient descent with help of Sickout Learn in my another regression model.

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

- In my first model I have used 'rain','precipi','Hour' as the features.
- In my second model I have used 'rain','meanwindspdi','Hour','meantempi','meanpressurei' as the features.
- In both models ‚UNIT' is used as 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.

- I used rain because I thought that when it is very rainy, people might want to use subway to avoid rain.
- I used hour because I thought that the public traffic always faces heavy rush hours as people come and leave their jobs.
- I used meanwindspdi, meantempi, meanpressurei, precipi because they can drastically improve the R2 value.

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

- In my first model:
  - the parameter of rain:  -17.60236319
  - the parameter of precipi: 65.74602064
  - the parameter of Hour: 59.49372653

- In my second model:
  - the parameter of rain: 75.99248368
  - the parameter of meanwindspdi: 61.63935065
  - the parameter of Hour: 24.84255451
  - the parameter of meantempi:-12.28326225
  - the parameter of meanpressurei: 68.0064754

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

- In my first model R2 value is 0.4784122

- In my second model R2 value is 0.4137407

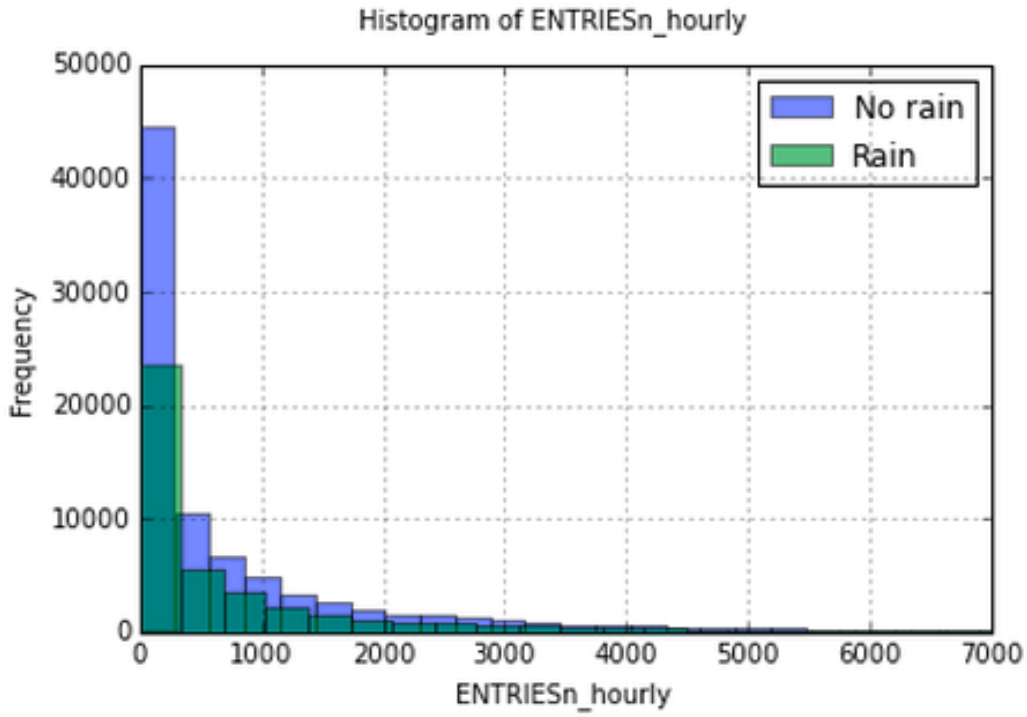
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

- In my first model R2 value means that 47.8% of the total variance in „ENTRIESn_hourly“ is explained by the regression model.
- In my second model R2 value means that 41.4% of the total variance in „ENTRIESn_hourly“ is explained by the regression model.
- The both R2 values are higher than 0.4, which is appropriate enough for this dataset.

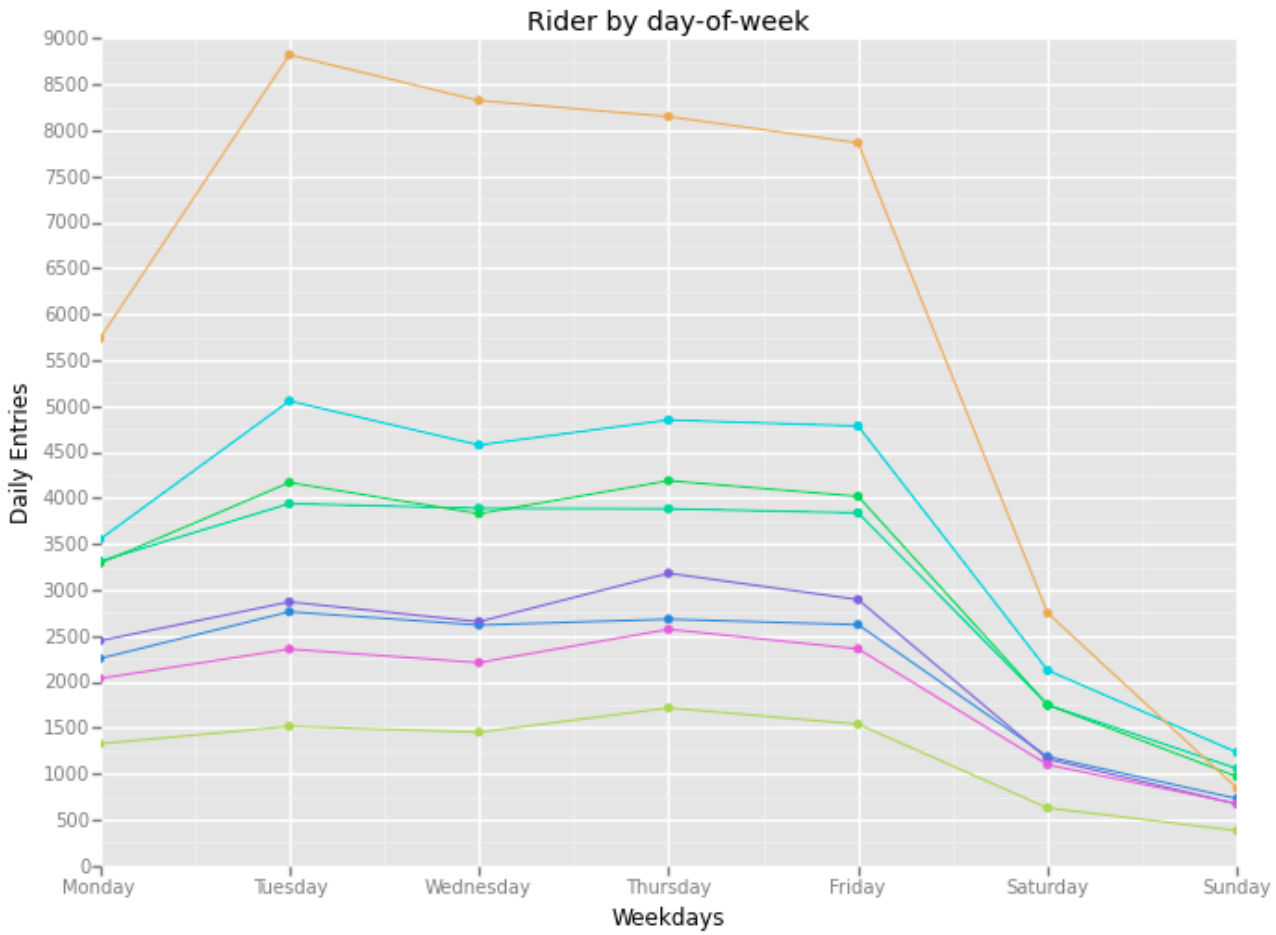
## Section 3. Visualization

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.



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.




## Section 4. Conclusion

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?

- From my analysis, more people ride the subway when it is not raining.

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.

- From the first linear regression model, we found the parameter of rain is negative, which means the influence of rain on the entries is negative. That's, if it is raining, the entries will decrease.

- From the first graph we can see that, the frequency of entries when it is not raining is always higher than when it is raining. Especially, at the not-rush hour the frequency of entries when it is not raining is almost twice more than when it is raining.
- Therefore, we can conclude that people prefer not to take the subway when it is raining, especially at the not-rush hour.

## 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:
Dataset, Analysis, such as the linear regression model or statistical test.

- Dataset: We only get the entries data of 6 hour per day. We cannot correctly calculate the daily entries of every weekday.
- Analysis:
  - the R2 values of both models should be improved.
  - it will be better, if we can choose the ten representative subway stations in the visualisation instead of randomly choosing ten stations.