Short Questions to Analyzing the NYC Subway Dataset

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?

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

3. What results did you get from this statistical test?

4. What is the significance of these results?

Section 2. Linear Regression

1. What approach did you use to compute the coefficients theta and produce prediction in your regression model:
    a. Gradient descent (as implemented in exercise 3.5)
    b. OLS using Statsmodels
    c. Or something different?

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

3. Why are these features appropriate?

4. What is your model's R^2 (coefficients of determination) value?

5. What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model is appropriate for this dataset, given this R^2 value?

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., scatterplots, 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 be two histograms of ENTRIESn_hourly for rainy days and non-rainy days

2. One visualization can be more freeform, some suggestions are:

a. Ridership by time-of-day or day-of-week
b. How ridership varies by subway station
c. Which stations have more exits or entries at different times of day

Section 4. Conclusion

Please address the following questions in details, and 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?

2. What analyses lead you to this conclusion?

Section 5. Reflection

Please address the following questions in details, and your answers should be 1-2 paragraphs long.

1. Please discuss potential shortcomings of the data set and the methods of your analysis.

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