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

1.1 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? What is your p-critical value?
I used a two-sided P-test to evalutate the hypothesis that the train ridership of NYC subway users is affected by the weather. The null hypothesis was that ridership (dependent variable) had no affected by the weather (independent variable). I was able to reject the null hypothesis with a confidence of P <0.1 using a two-tailed p-value test.

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.
We had to run this parametric test as the underlying data samples were not normally distributed and therefore ruled out the use of a T test.

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.

- We can reject the null hypothesis because ridership does change depending on the weather.
- P < 0.1
- Mean With Rain = 1105.45
- Mean Without Rain = 1090.28

1.4 What is the significance and interpretation of these results?
The significance was P < 0.1, which indicates that there is a correlation between the change in weather and the ridership on the subway.

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:

I used the Gradient Descent model to model my predictions

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

I used the following input variables Rain, precipi, hour, meantempi. I also used a dummy variable for the Unit column to run the test.

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 choose to use the above features as they were the ones I felt gave the best proxies for rain. I choose to use variables for that predicted rain as my initial P-test results indicated that there maybe a correlation between the rain and train ridership. The analysis resulted in an r square value supporting the notion that rain may affect ridership

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

They indication to what extent the variable helps explain the variance in the data and therefore holds more explanatory power for the model.

After normalizing them the coefficients in this example were:

Variable	Weight
Rain	0.336605
Hour	10.918090
Meantempi	64.258833
Precipi	0.170004
Maxpressurei	30.033902
Maxdewpti	57.276658
Meanwindspdi	5.545358

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

R^2 =0.461129068126

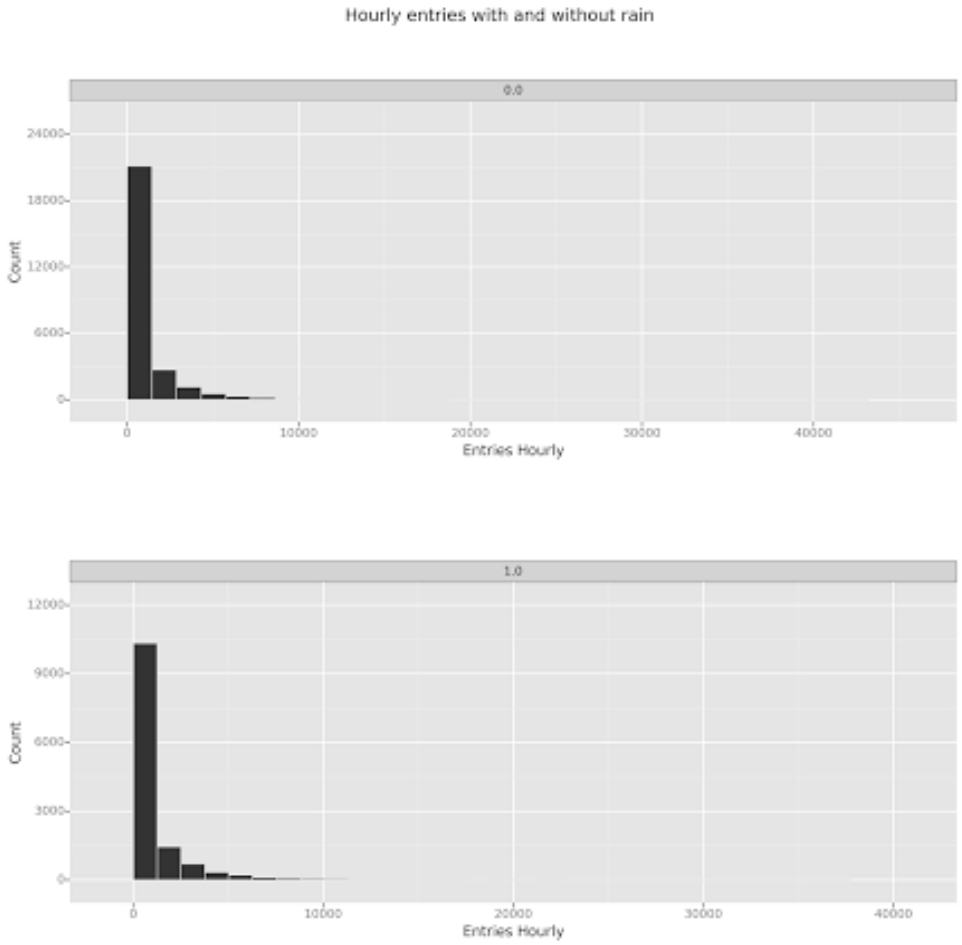
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

The R^2 value indicates the level of variance that can be explained by the model as specified. In this case it would explain about 46% - which is ok considering it is a very basic linear regression model. Obviously employing more advanced analytical and statistical methods would result in much higher explanatory values.

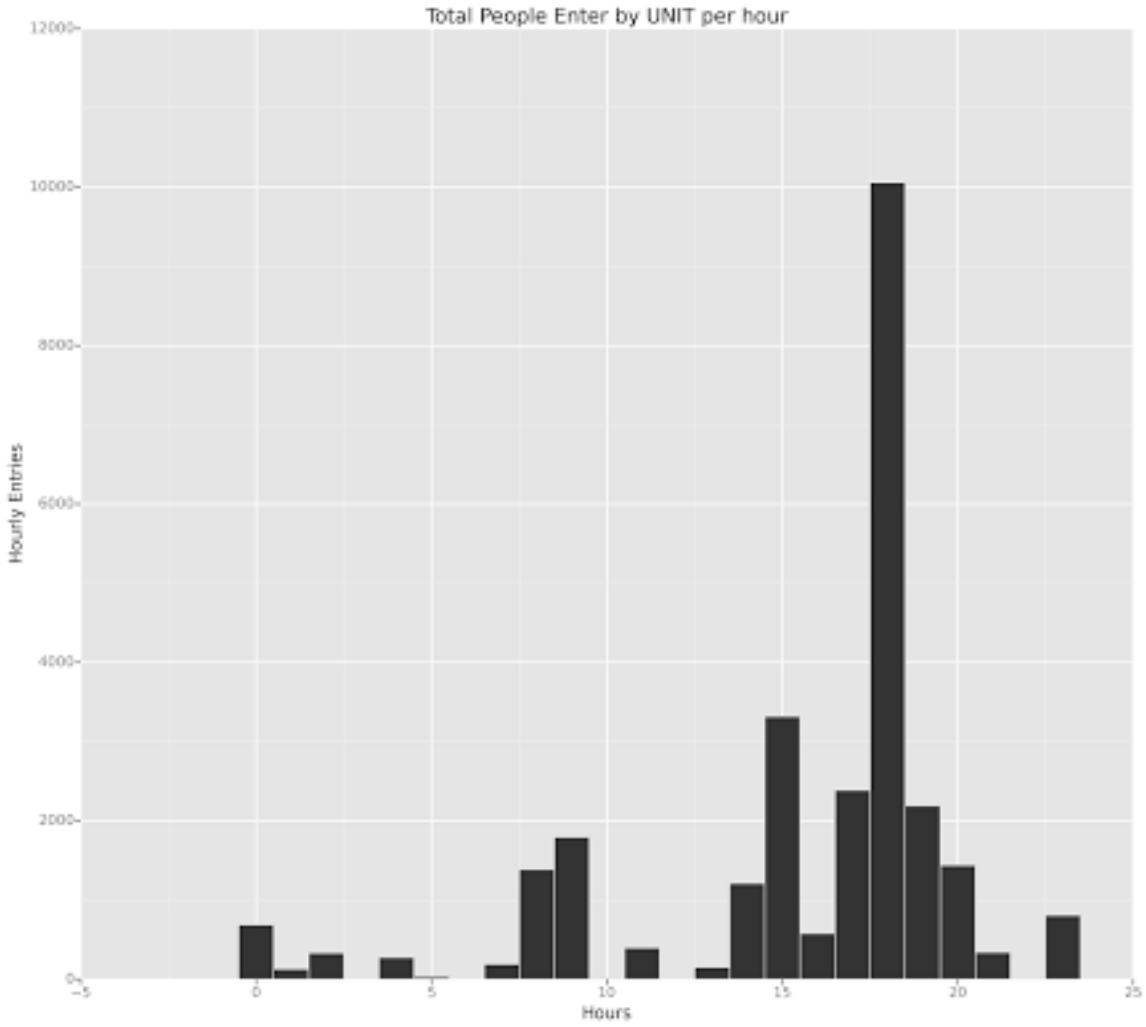
Section 3. Visualization

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.

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. Some suggestions are:



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?
After having run different models against the data and getting similar results from these tests I feel that we can say that there is a some type of correlation between the weather and ridership. However I would also say that we cannot yet say that this is a definite correlation until further data analysis is performed.

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.

After checking the distribution of the data it was determined that there is a variation of ridership depending on the weather. Upon further analysis using the Mann-Whitney U-Test it was confirmed, at a confidence interval of $p < 0.1$, that when it rains there are more subway riders. Finally a gradient descent model was run on the data using rain, perception, hour of the day and the mean temperature as the variables it was noted that the coefficient for rain was substantially more significant than the other variables, which suggests that when it rains there is an increase in subway ridership.

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, Linear regression model, Statistical test.

The main shortcoming of the analysis, I feel, is that I was not able to assess with a significant enough confidence, R^2 , the exact weather based variable(s) from our dataset that likely influenced riderships, as per the initial rejected hypotheses results from the two tailed t-test. I feel that this may have been a shortcoming due mostly to the dataset & the linear regression model.

The dataset was limited to only a single point in time, May 2011, which could bias the results quite significantly as we did not look at even 1 full year's worth of data. I also feel that the severity of rain would have had a higher impact on the ridership but unfortunately the dataset was limited to a static rain or no rain result.

The statistical analysis run on the data was sound and we were able to reject the hypothesis with a significant confidence, with the only real improvement here being data points to assess. The results of the Gradient Decent model however were not as encouraging as the initial statistical test results we ran and I therefore think this could have been improved. This could have been done by calculating better Thetas for our tested coefficients, if we were to still use Gradient Descent. I would also suggest running a different regression model, e.g. ordinary least squares regression, to see how that model works on our dataset.

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

It appears that day of the week also has an effect on the ridership of the subway and that week days tend to have a higher number of hourly entries. I feel that given our analysis above and this new possible insight could allow us to build a more rigorous model by including the day of the week as a factor in our statistical analysis.