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.

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 used the Mann-Whitney-U-Test to analyse the subway data and used the two-tail p value. The null-hypothesis is, that both samples - rainy and non-rainy - are drawn from the same population with the same distribution. With a significance level of alpha = 0.05 and a two sided test my p-critcal value is 0.025 (alpha/2 for each side)

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.

The Mann-Whitney-U-Test is applicable to the dataset because it does not assume any particular distribution. The Welch-T-Test for example assumes normal distribution, but the graphical analyses does not conclude normal distribution, therefore I did not assume any distribution and decided to to use the Mann-Whitney-U-Test for the stated reason.

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.

The Mann-Whitney-U-Test - also called Wilcoxon Test - was performed with R, since scipy gave an error.

W = 153635121, p-value = 5.482e-06, mean_norain = 1845.539439, mean_rain = 2028.196035

1.4 What is the significance and interpretation of these results?

The p-value is way smaller than the two-sided alpha level of 0.025 (0.05 / 2 for each side). This means that we can reject the null-hypothesis and conclude that both samples are statistically different. On average there are 183 more people riding the subway on rainy days in those samples.

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:

We want to predict the number of hourly entries and use the a linear regression to achieve this. Linear regression means that there we assume a linear relationship between the independent variable of hourly entries and the dependent variables. For example, if rain has a positive linear relationship with the ridership as hinted in the statistical test above, then more rainy days will result in more people taking the subway. Based on the models learnt so far in the Intro to Data Science course I have choosen to implement Ordinary-Least-Squars (OLS) models using statsmodels. This model is easy to interpret though other models may explain more of the variance

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

Yes. 'UNIT', 'weekday' and 'conds' as modeled as dummy variables, because they are categorical variables and represent qualitative data, not quantitative data. Once modeled as dummy variable we can use them in a quantitative model.

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.”

I have chosen 'hour' because mass transportation in a big city plays likely a role for commuting to work. People use the subway to go to work in the morning and come back home after work in the evening. Commuting to work is the same reason why I included 'weekday'. I used the specific weather conditions per unit 'conds' because it is more specific than 'rain' and I believe weather conditions will likely play a role because it can make the subway a more convenient choice compared to other forms of transportation. 'UNIT' greatly improved my R^2, although some of them are insignificant.

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

I have modeled 'UNIT', 'weekday' and 'conds' as a dummy variable. The coefficient for 'hour', the only non-dummy variable in my model, is 123.6972.

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

R^2 is 0.484 and adjusted R^2 is 0.481

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 adjusted R^2 of 0.481 means, that the independent variables in OLS model can predict 48,1% of the variation of the dependent variable of hourly entries. That's nearly half of the variance. Other models than linear may yield a better R^2, but the advantage of a linear model is the interpretability. A

linear relationship keeps the model simple and understandable whereas more complex models need more interpretation.

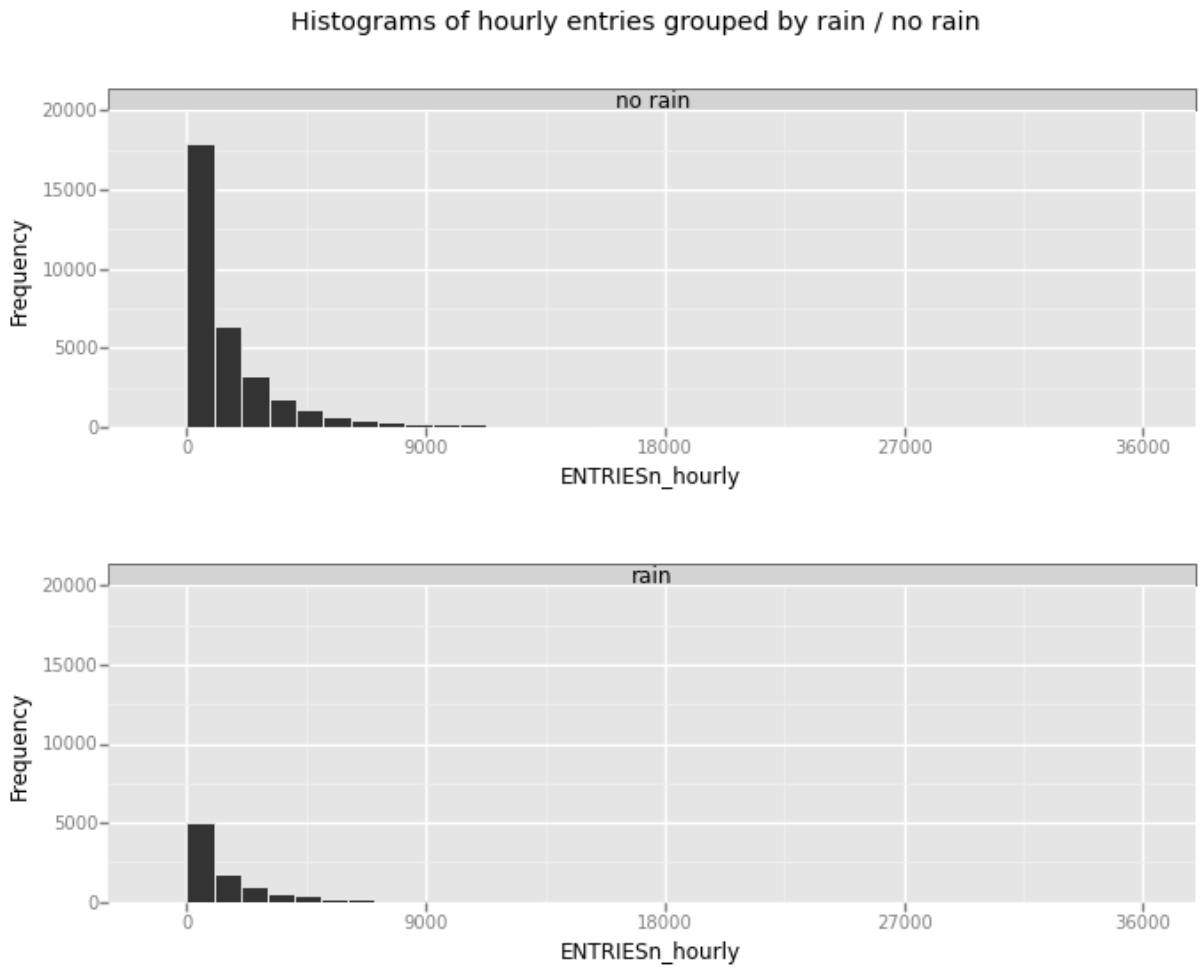
Section 3. Visualization

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.

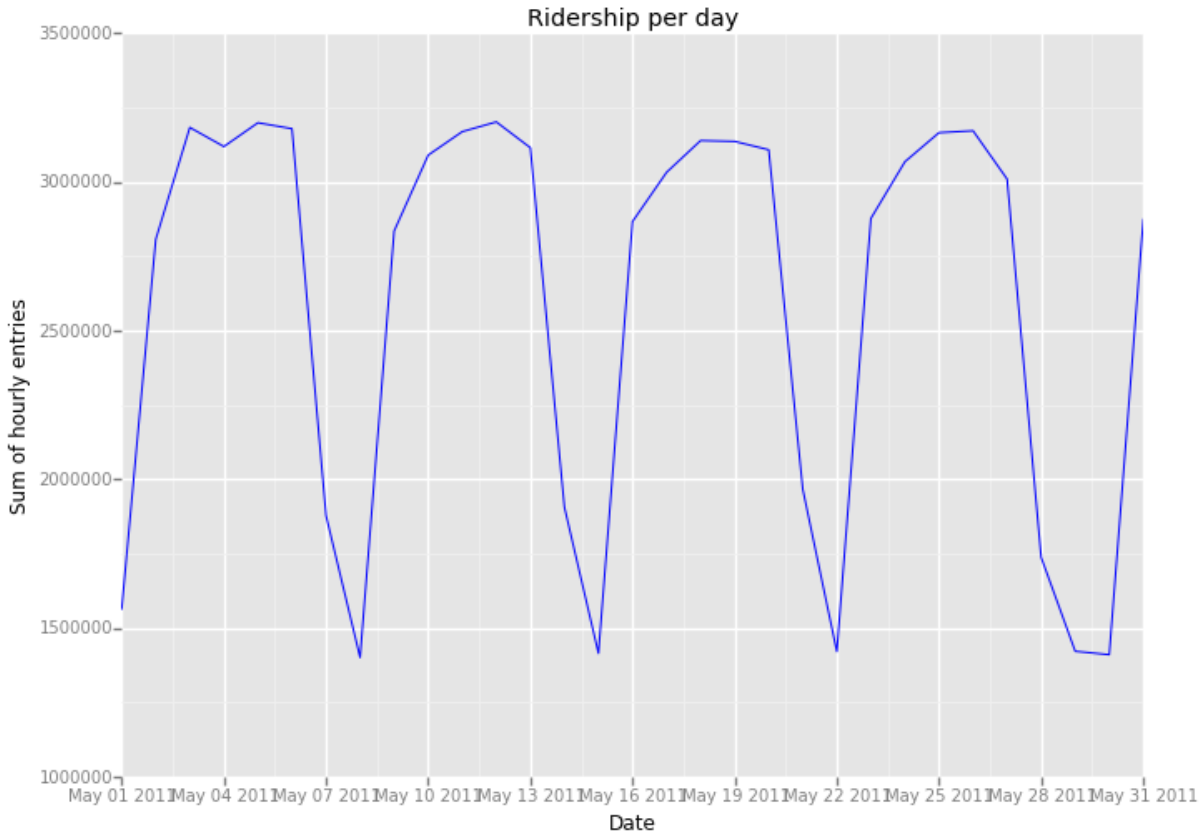
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- 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.



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 day



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?

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.

There is a statistically difference between the amount of riderships on rainy days and those on non-rainy days. The Mann-Whitney-U-Test showed that both samples are not from the same population and therfore have different distributions. The compared means show that on average there are 183 more people riding the subway on rainy days in those samples. Based on those results, people do ride the NYC subway more when it is raining.

The linear regression though showed a more complex relationship between specific weather conditions and ridership. Suprisingly, the coefficient for the different rain types differ in their direction. While light rain has a positive coefficient, light drizzle, rain and heavy rain have a negative coeficient, meaning in this model less people are riding the subway during those weather conditions. With a adjusted R^2 of 0.481 the Oridnary-Least-Squares model used here can predict 48,1% of the variation of the dependent variable of hourly entries, leaving 51,9% of the variance unexplained due to other

factors like unkown variables. Further examinations with more complex models than the OLS should be conducted to improve the R^2.

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

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

The data consists only of one month in may 2011. Predictions may be not accurate for colder months. For example, those rainy conditions that had a negative coefficient could turn positive in winter months. The weekday variable had a significant impact. I strongly suggest to include additional information on public holidays in New York City, as they may have a similar effect since people don't need to go to work.

Given the adjusted R^2 of 48,1%, there is room for improvement because more than half of the variance is not explained by the variables. A possible solution could be the use of non-linear models. They might not be as eays to interpret, but may explain more of the variance left.