Analyzing the NYC Subway Dataset

Look at NYC Subway data and figure out if more people ride the subway when it is raining versus when it is not raining.

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.

[1]: Average rainy/snowy days in new york: http://www.weather-and-climate.com/average-monthly-Rainfall-Temperature-Sunshine,New-York,United-States-of-America.

[2]: Blogpost about the analysis of the ridership of the new york subway. http://rainydaysny.blogspot.dk/

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?

In order to answer the very broad question about whether more people ride the subway on rainy days compared to non-rainy days I used a Mann-Whitney rank test. The test was performed as one-tailed p and a null hypothesis saying, that there are no difference in subway ridership counts on a rainy compared to non rainy days or the ridership count is higher on non-rainy days. 0.5 was selected as the p-critical value.

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 rank test does not make any assumption about the samples or population probability distribution. The histogram in section 3 indicates that you can not make a suitable assumption about the dataset distribution. In addition the Mann-Whitney rank test does not require the samples sizes to be equal. The sample size for non rainy days is approximately 3 times the size of the rainy days sample size. No assumption about the distribution probability and that the unequal sample sizes for rainy and non-rainy days ridership counts makes Mann-Whitney rank test suitable as a statistical 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.

U	p-value	Rainy day Mean	Not rainy day mean
153635120.5	2.741e-06	2028.20	1845.54

1.4 What is the significance and interpretation of these results?
The result of the Mann-Whitney rank test shows that there is a significant amount of more people riding the subway on a rainy day compared to non-rainy day. The p-value is far below the p-critical value of 0.05 which proves a strong significance in this test.

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: Gradient descent (as implemented in exercise 3.5) OLS using Statsmodels, Or something different?
Gradient descent.

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 five normal features: rain, tempi, day_week, hour and fog. In addition two dummy variables was selected as features namely UNIT and conds.

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."
My first approach was to reasoning about which feature that seems relevant for this kind of model. In addition I experimented with several of the variables by trying different combinations and then looking at the R^2.
My reasoning for the feature list is as follows:
rain: I think if it rains people tends to use the subways as an alternative to walking and using a bike, though people in the periphery of the city that has a car available might decide to use that instead. However if it rains some people might also wait to a non rainy day to do their errands.
tempi and fog: This features are selected based on both reasoning and data exploration. If it is foggy or the temperature is low people might prefer the subway instead of driving or walking, though the inclusion of these features only increase the R^2 with approximately 0.00002. So basically the two features are mainly included in the model because it improves the R^2 value.
day_week: Some days are more busy transportation wise. I think people do have different travelling habits in the weekday compared to the weekend. e.g. if the main population ridership is working people, the weekends might have fewer riders in general. E.g. if the rain affects the ridership count i would imagine this effect should be higher in the week days.
hour: There are peak hours in which the largest population use transportation e.g. whether it is to or from work. This means that at daylight there are more riders compared to dark hours.
UNIT: Some UNIT locations are more busy than others. I could have included the "station","latitude" or "longitude" however this feature seems to have a better resolution due to more units per station and thereby makes these feature redundant.
conds: I think this feature tells us what the subway riders experience and decide on right before they decide to use the subway. e.g. by looking out the window they decide to choose the subway because the sky gives the impression that it is going to rain.

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

1.01811928e+02 -4.43392842e+00 -2.87967834e+02   8.53609486e+02 -2.40844034e+01

2.5 What is your model's R^2 (coefficients of determination) value?
0.474150197797

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 tells us something about how well the regression model is able to predict the ridership count of the subway. As the value approaches 1 the model tends to be stronger for the dataset and thereby producing better predictions. As the value approaches zero the model predictions becomes weaker. Since the value is closer to zero than one, I would suggest that this linear regression model build with gradient descent and the selected features is not suitable to predict the amount of subway ridership count.

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.

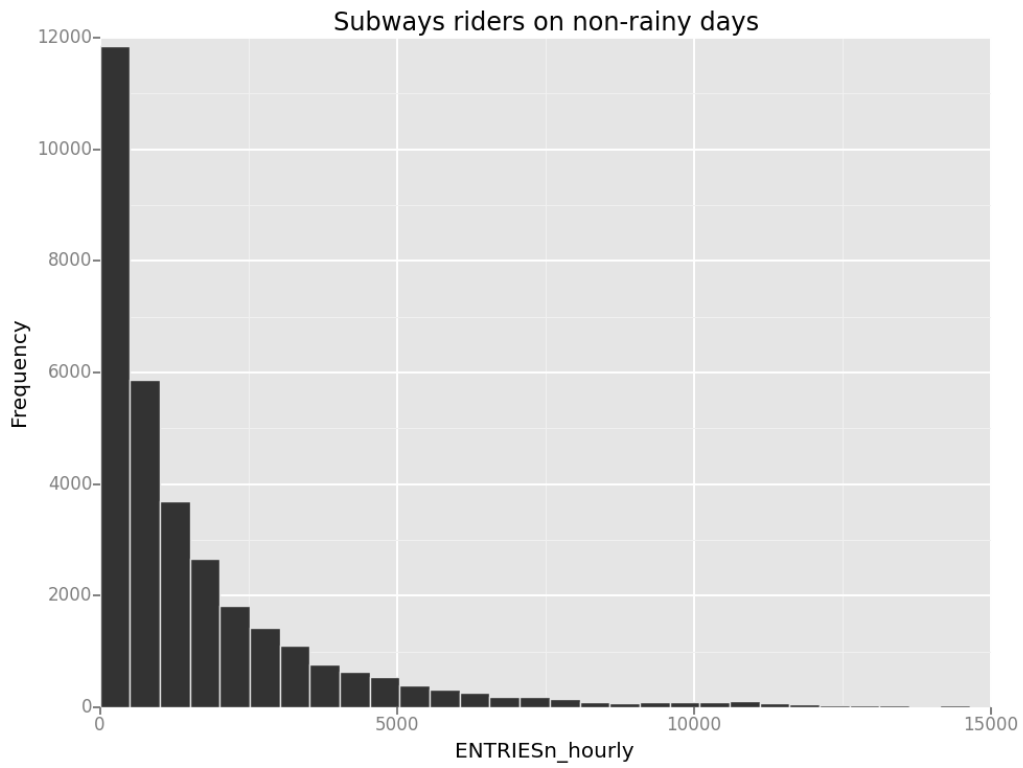

figure 1: Shows the histogram of the ridership count on non-rainy days.

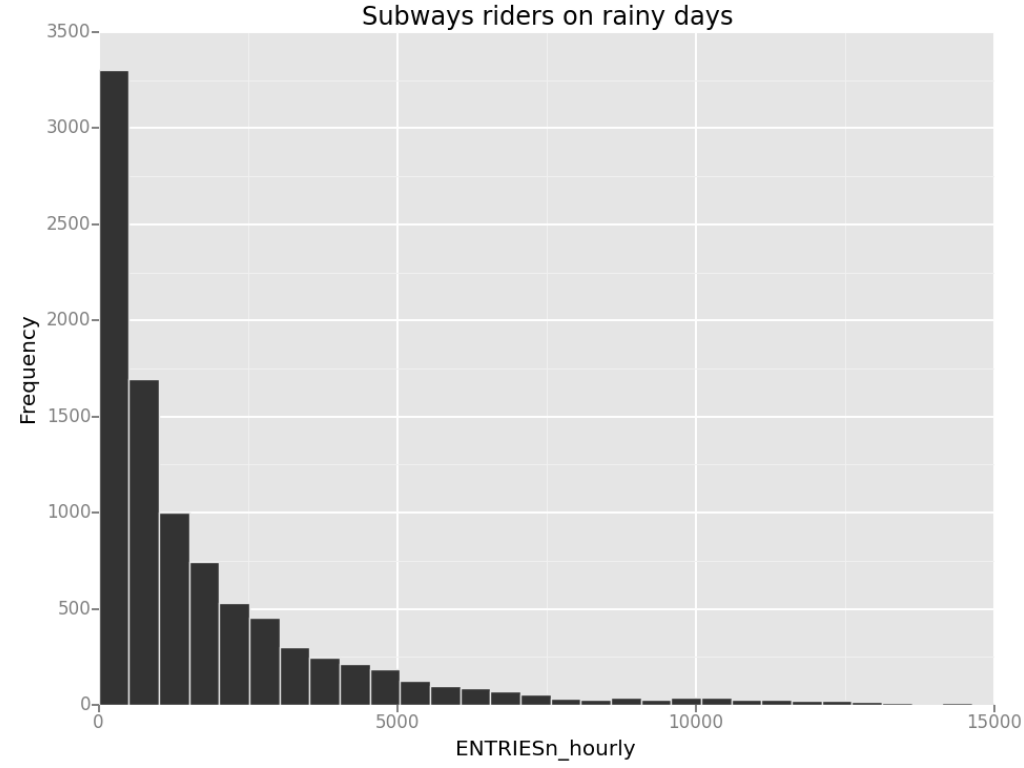

figure 2: Shows the histogram of the ridership count on rainy days.

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. Some suggestions are: Ridership by time-of-day, Ridership by day-of-week

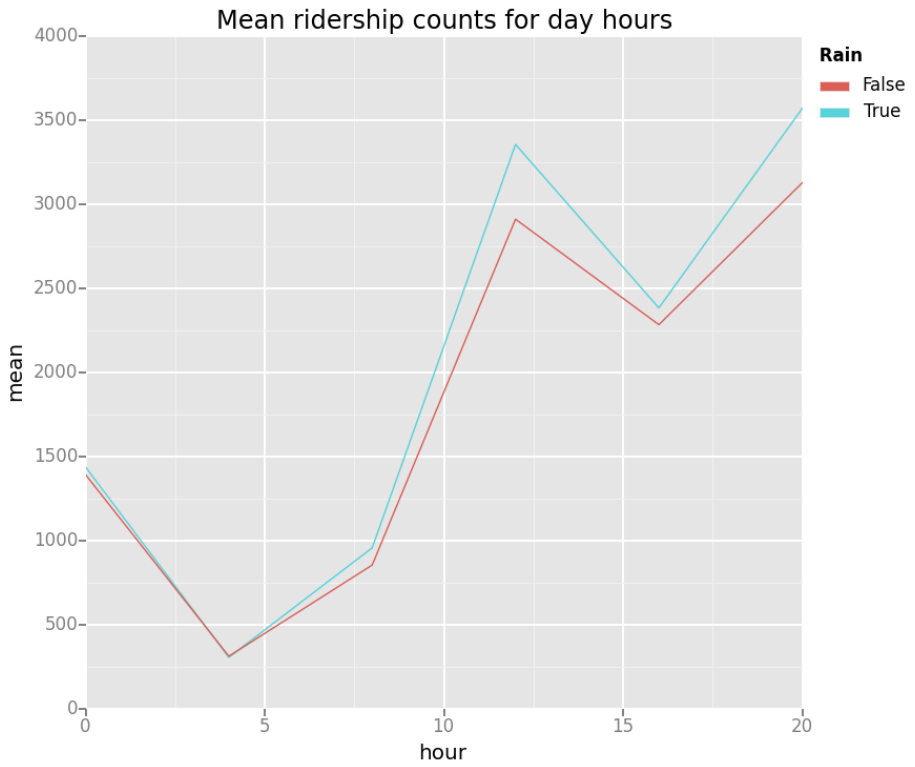

Figure 3: Shows the relationship between the day hour and ridership count mean for rainy and non-rainy days.

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.
4.1 and 4.2 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?
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 are some results that indicate that more people ride the subway when it is raining. However, there are also results that indicate no difference in riderships on rainy days compared to non-rainy days.
By looking at the results from the statistical test in section 1, it has a clear significant evidence that more people ride the subway on rainy days. However if you apply the same test and group the data by station the results only shows significant more riders on rainy days for 3 stations out of 256, namely 181 ST(p=0.0270206610386), 191 ST (p=0.0249503364134) and WILSON AVE (p=0.0467716461052). In addition, if you apply the same test but without weekend days no station shows significant more ridership counts on rainy days.

On the other hand if you group the data by hour as in figure 3, there are significant more riderships at 3 clock hours: 8 (p=2.22720386696e-06), 20 (p=0.0125946249159) and 12 (p=2.63981530698e-11). However 16 o clock has a p=0.0621640881076 which is relatively close to the critical value and outside peak hour(compared to 12 and 20 in figure 3) thus suggest that at daylight rain increase the ridership count in general.
If we look at the results from the linear regression from section 2 the model produced a R^2 = 0.474150197797 and without the rain feature the result is R^2 = 0.473395676914 which is a remarkable small difference. This result might lead to suggestions, that the rain feature does not have an impact on the ridership count. Though i think this small impact i due to unsuitable model design.

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.
My intuition tells me that more people ride the subway on rainy days, however this dataset has several shortcomings in terms of providing enough evidence to support my intuition. One shortcoming is the fact that the data is only collected in May, though it seems to be the most rainy month[1], however it would be appropriate to collect data for an entire year. Imagine the data was only collected in December, then my intuition would be, that there would be no difference between the ridership for snowy and rainy days.
Another shortcoming is sample size difference. If we look at figure 3 again and especially 16 o clock, i would expect to get a significant result on the statistical test if the sample sizes are larger. According to the results from the analysis described in section 1 which shows significant more ridership on rainydays, I would expect that this would also be the case station wise. Perhaps a larger sample size would provide that result.
In general the dataset still has a lot of potential to provide interesting statistical results. There are still a lot of variables that i did not consider when performing the analysis, for instance the ‘weekDay’ and the ‘cons’ variable.
The result from my linear regression model provided some poor results in the prediction of ridership count. This suggest trying a different method than gradient descent and other kinds of machine learning approaches, which already has been suggested elsewhere [2].
Finally i would like to give my thought on the asked question about whether rainy days increase the riderships of the subway. As mentioned before, the results only apply for the ridership in May, which might be an unusual month in terms of ridership patterns. Also the dataset might provide more insights if the resolution in the hour timestamp was increased. This could elaborate on the precipitation and ridership pattern by investigating how the rain distribution is on the particular day and thereby provide a clear idea about what a rainy day is and how many subway riders the rain actually affects. If the rain is only distributed in the early morning hours this might not have an effect on the peak hours at daylight.

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