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, 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.

[1] https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

[2] https://en.wikipedia.org/wiki/Welch's_t_test

[3] http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

[4] http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html

[5] http://stackoverflow.com/questions/4136244/matplotlib-pyplot-how-to-enforce-axis-range

[6] http://stackoverflow.com/questions/15928539/matplotlib-how-to-make-the-marker-face-color-transparent-without-making-the-li

[7] http://stats.stackexchange.com/questions/124995/how-do-i-interpret-the-p-value-returned-in-scipys-mann-whitney-u-test

[8] http://mustafaotrar.net/istatistik/non-parametrik-mann-whitney-u-testi/

[9] https://www.cscu.cornell.edu/news/statnews/stnews68.pdf

[10] http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

[11] http://www.medicine.mcgill.ca/epidemiology/joseph/courses/EPIB-621/fit.pdf

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 Mann Whitney U-test because our data is non-normally distributed. By using this method we want to test whether these two distributions(rainy days vs. non rainy days) are the same or not. So our P-value should be two sided and our null hypothesis should be: "H_0: The distribution of the entries are the same between rainy and non rainy days." and the alternative hypothesis should be: "H_A: The distribution of the entries are not same between rainy and non rainy days."

P-value is 0,0498 (By Mann Whitney U test we get a one-sided p-value, which is 0,0249. For making it a two-tail P-value we have to double this number.) and 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.

In problem set 3:1 Explotary Data Analysis question, we see that histograms are not normally distributed, which means we don't have enough information about the distributions. Thus we can't use Welch's t-test.

Therefore, we have to test whether these two distributions are the same or not. We have to use Mann Whitney U-test because this method can be applied on unknown distributions and according to Wikipedia it is nearly as efficient as t-test on normal distributions.

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 mean of entries with rain = 1105,4464

The mean of entries without rain = 1090,2788

The Mann Whitnet U-stat = 1924409167

P-value = 0, 0498 (two sided)

P-critical = 0,05 (α=0,05 so we have 1-α=0,95 significance level)

As a result P-critical value is bigger than P-value so we have to reject H_0 that the distributions are the same.

1.4 What is the significance and interpretation of these results?

In 95% significance level, we reject H_0 hypothesis because our P-value is less than the P-critical value so we can say that the distribution of the entries are different between rainy and non rainy days.

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:

1. OLS using Statsmodels or Scikit Learn

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

Features:

- rain
- precipi
- Hour
- meantempi

Dummy variables:

- UNIT

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."

For this problem, i think there are some features that are logical to add to the model.

- Hour of day (‘Hour’): At some hours people will use subway more.
- Temprature (‘meantempi’): I think people use subway more when it is colder.
- Rain (‘rain’): When it is raining, people will prefer using subway more instead of getting wet
- Precipitation (‘precipi’): Like rain, any kind of water vapour including drizzle, snow, graupel and more, will affect ridership positively

When I add these features to the model, my R^2 value becomes 0,4792

Also some other variables that i think they are reasonable:

- Fog (‘fog’): Fog reduces visibility so people might decide to use subway more instead of driving cars
- Wind speed (‘meanwindspdi’): When it is windy, especially tourists might decide to use the subway more

However, when I add these features to the model, none of them increased my R^2 value drastically. For every feature i add, increase in R^2 is very small which means these extra features (‘fog’ and ‘meanwindspdi’) are not significant for the model. I think the variation in the model is best described with the first 4 feature (‘Hour’,’rain’,’precipi’ and ’meantempi’).

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

Coefficient of rain: 29,4645

Coefficient of precipi: 28,7264

Coefficient of Hour: 65,3346

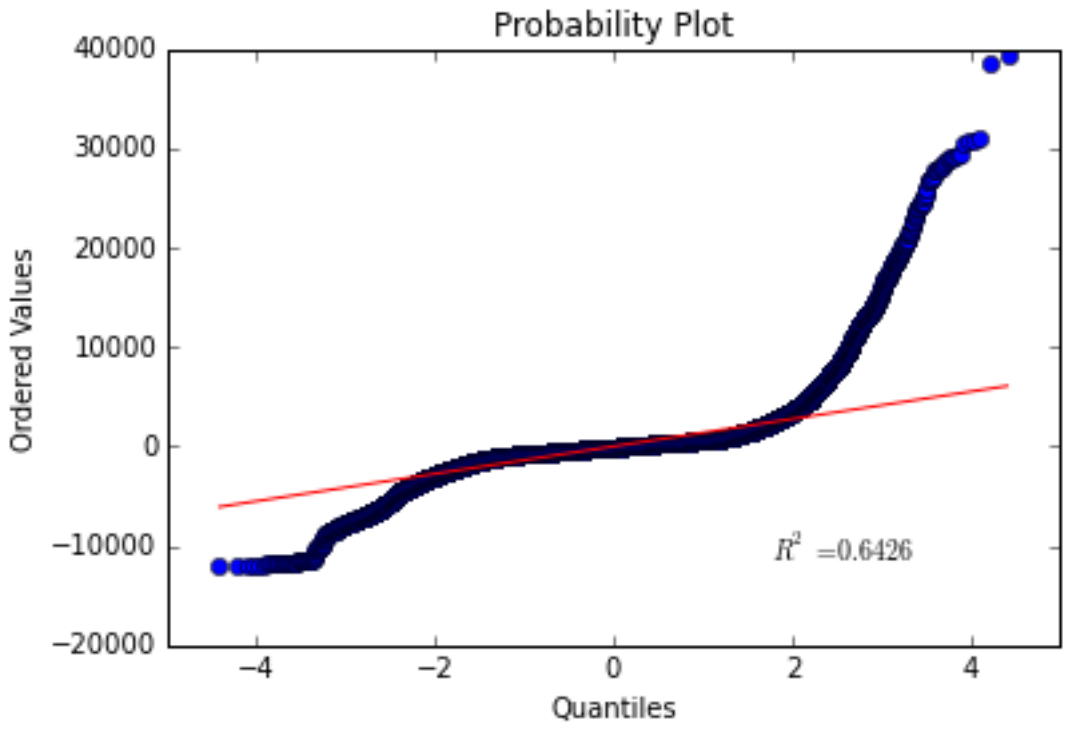
Coefficient of meantempi : -10,5318

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

R^2 = 0,4792

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 thisR^2 value?

The R^2 value explains only 48% of the total variability of our model. This percent indicates the data is not close enough to the fitted regression line. We can also say that the proposed model does not improve prediction over the mean model [9]. However, according to Minitab Blog: “In some fields, it is entirely expected that your R-squared values will be low. For example, any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%. Humans are simply harder to predict than, say, physical processes.”[10]. Therefore, instead of looking only at R^2 value for the conformity of our model, we have to check the normality of the residuals. By using scipy.stats.probplot, we get the plot shown below.



This plot shows us the residuals are not normally distributed because most of the points in the plot are far away from the red line. Also the R^2 value, which is 0,0642 (calculated within the plot), proves that residuals are not normally distributed. Now we can conclude that this linear regression model is not appropriate for the given dataset.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

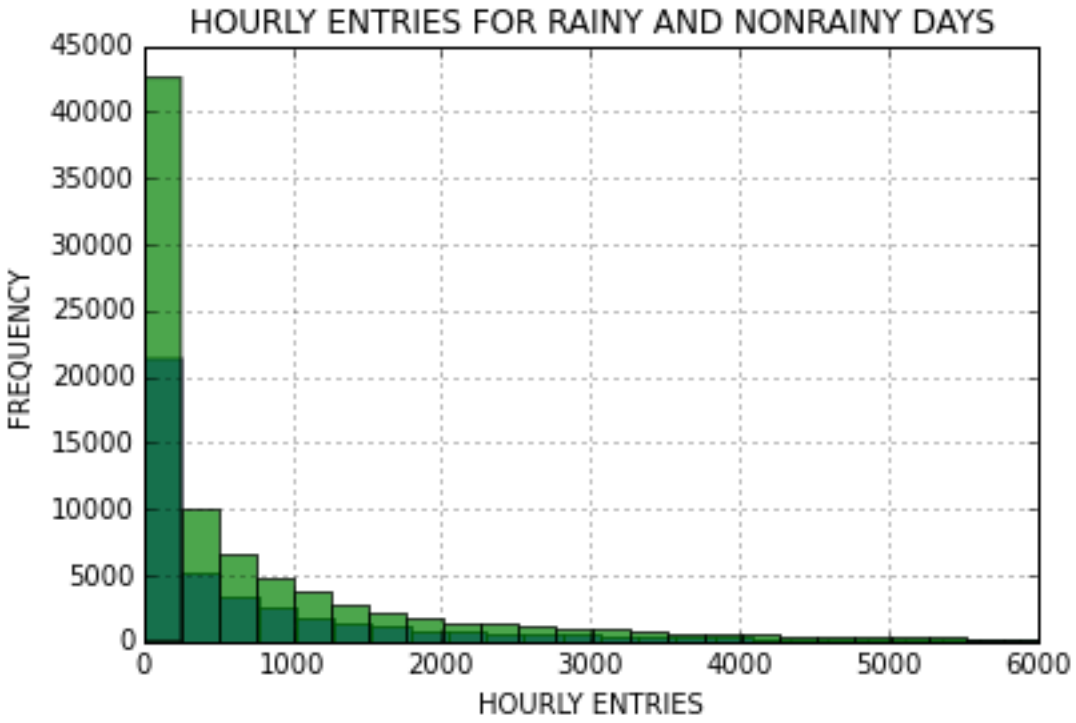
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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

- 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,

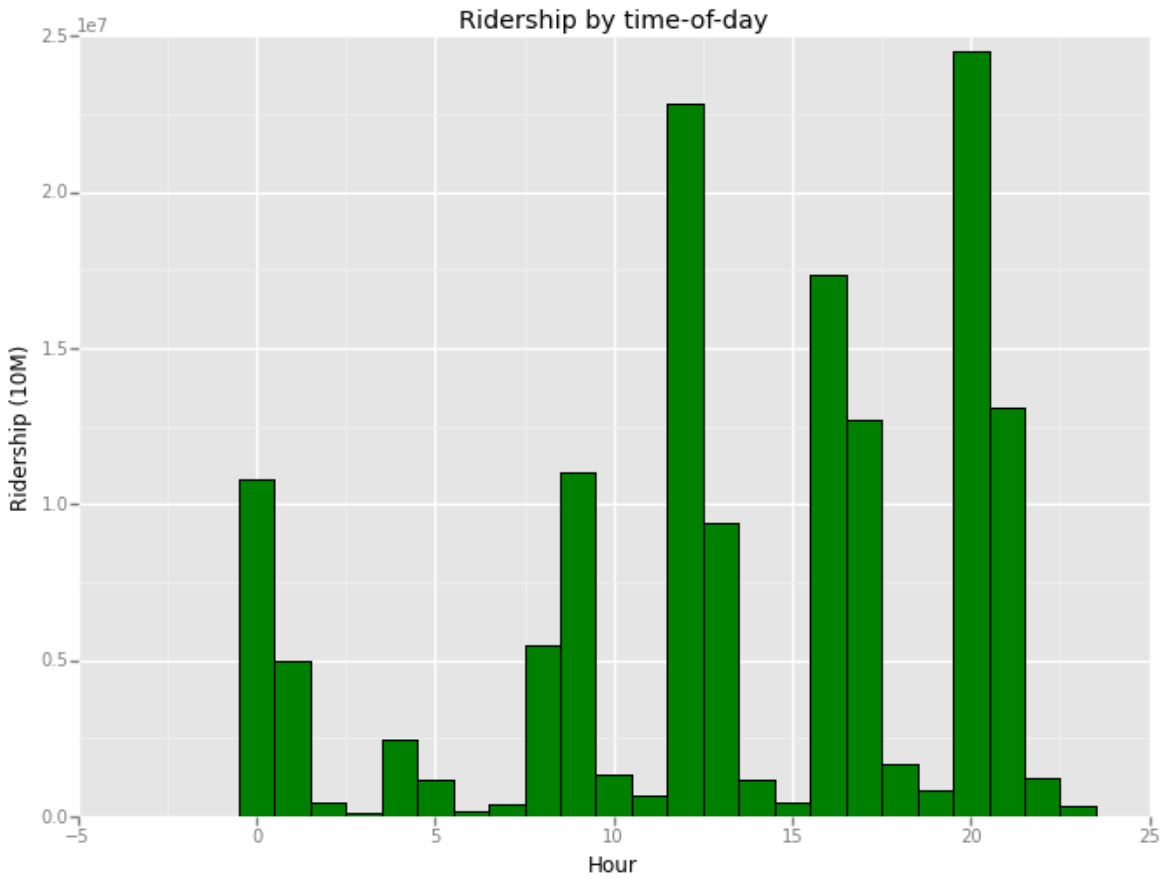
- each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



This histogram represents hourly entries for rainy and nonrainy days. Intervals represents the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurance on the y-axis. Dark green bars indicates frequency of rainy days and light green bars shows frequency of non rainy days. Also x-axis has been truncated at 6000.

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



This bar graph shows ridership by time of day. On x-axis, hours are represented in 24 hour format. On y-axis, total number of ridership is represented (x10 million). For example, at 20:00 there are 24505996 entries.

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?

According to Mann Whitney U-test, in 95% significance level, these two datasets have nonnormal probability distributions and are not the same which means rejecting null hypothesis and accepting the alternative hypothesis. We also know the means so we can say that more people ride the NYC subway when it is raining.

In contrast, we have developed a regression model which doesn't answers the question: "Do more people ride the NYC subway when it is raining or when it is not raining?". This model tries to answer which attributes mostly effect the ridership.

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.

Our statistical test is about testing weather these two datasets have the same distribution or not, but Mann Whitney U-test does not provide any other information about these datasats' distribution or about their parameters. By using only U-stat we can't find the dataset that has bigger mean but the P-value explains us much information. After calculating the two sided P-value, we reject the null hypothesis because our P-value is less than the P-critical value. We know these two datasets are different and we also know the means so we conclude saying that people ride the NYC subway more when it is raining.

Our linear regression model is about explaining what features effect people to ride the NYC subway. In our model rain feature has a coefficient of +29,4645, which means everytime it is recorded raining, ridership increases by 29,4645. We can interpret this as rain positively affects ridership. However it is not an proper model to explain most of the variability of our model. By using this model we can explain only half of the variability which means only half of our forecasts are close enough to our actual data. Also the residual plot is not normally distributed. Making predictions using this linear regression model about whether people ride subway more in rainy days or not will yield an error.

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.

In my opinion, both dataset and methods used for analysis have some shortcomings. First of all, our linear regression model is not necessary to answer the question: "Do more people ride the NYC subway when it is raining or when it is not raining?" Moreover the residuals are not normally distributed (as explained before), which means actual values are highly differ from forecasted values. So we can't say that his model is appropriate for the given dataset. For this project linear regression part is just for practising, and not for answering our main question. Secondly, the given dataset is not randomly choosen and the date period is not long enough. It covers only 1 month (May) and 30 consecutive days in the year 2011. So our analysis rely on this old, limited and short time span.Lastly, SciPy Mann Whitney U-test does not provide

enough information about the test results. With only using SciPy Mann Whitney U-test we can't conclude which dataset has bigger mean. Some additional calculations are needed to find the dataset that has bigger mean. Moreover its documentation is not well prepared as well. To use this methed, additional research is needed.

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