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

During this assignment I have found useful the following references:

- Lesson and Problem Set 2:
  - stackoverflow answer about MAX selection in SQL (https://stackoverflow.com/questions/612231/how-can-i-select-rows-with-maxcolumn-value-distinct-by-another-column-in-sql)
  - This specific post at the forum (https://discussions.udacity.com/t/clarification-for-problem-set-2-5-fixing-turnstile-data/4563/4)
  - Differences between strftime and strptime (https://www.quora.com/In-Python-what-is-the-difference-between-strftime-and-strptime)
- Lesson and Problem Set 3:
  - Blog post about using OLS for linear regression in python (http://connor-johnson.com/2014/02/18/linear-regression-with-python/)
- Lesson and Problem Set 4:
  - ggplot blog documentation (http://blog.yhathq.com/posts/ggplot-for-python.html)
  - Group dataframe entries depending on column value: (https://stackoverflow.com/questions/17926273/how-to-count-distinct-values-in-a-column-of-a-pandas-group-by-object)

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 analyze the NYC subway data. In the problem set, I used the `scipy.stats.mannwhitneyu()` function to apply it. This particular implementation returns the p-value of a one-tailed T-test.

The Null Hypothesis was: There is no difference in the number of entries at the Subway between rainy and non-rainy days. That is, that the rainy days population mean (μ_R) is equal to the non-rainy population mean (μ_N): $\mu_R = \mu_N$ or $\mu_R - \mu_N = 0$

The p-critical value was 0.05, or a confidence interval of 95%. The test p-value was 0.249, so the result would have been statistically significant in a two-tailed test using that confidence interval.

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 a non-parametric non-paired test that doesn't assume that the data follows a normal distribution, and compares two samples to see if their populations are statistically significant different.

As we had two exclusive variables and independent observations that do not follow a normal distribution, the test is fitting.

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.**

Mean entries per rainy days: 1105.4463767458733

Mean entries per non-rainy days: 1090.278780151855

Mann-Whitney Statistic: 1924409167.0

p-value: 0.024999912793489721

critical p-value (95% confidence interval): 0.05

1.4 **What is the significance and interpretation of these results?**

These results confirm that there is a significant difference between the two populations (use of the subway between rainy and non-rainy days) tested, and we can reject the null hypothesis. As the mean on rainy days is greater than on non-rainy days, we can confirm that people in NYC use the subway more on rainy days than on 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:

2.1.a OLS using Statsmodels or Scikit Learn

2.1.b Gradient descent using Scikit Learn

2.1.c Or something different?

I used the OLS implementation in Statsmodels (Assignment 3.5), as well as the Gradient Descent using Scikit Learn (Optional Assignment 3.8)

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

Rain, precipi, Hour, Maxpressurei, Meantempi and UNIT as dummy.

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 used the following features in the model:

Rain and precipi: It was the principal hypothesis of this problem set and it was intuitive.

Hour: I introduced this variable because I thought that rain during peak hour will probably affect the subway ridership going and coming back from work.

Maxpressurei: it improved my R^2 by one point when I introduced it.

Meantempi and fog: I introduced this variables in the model because it felt intuitive that days with fog and with lower average temperatures people would use more the Subway than go walking as are harder days ‘psychologically’.

I used the dummy variable ‘UNIT’ as was exemplified because when I experimented removing or changed it the R^2 value decreased.

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

rain	-9.993638	maxpressurei	-392.765135
Hour	58.023348	meantempi	-13.348751
precipi	-41.239875	intercept	13088.98957
fog	201.681613		

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

The model R^2 is 0.480015221228.

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?**

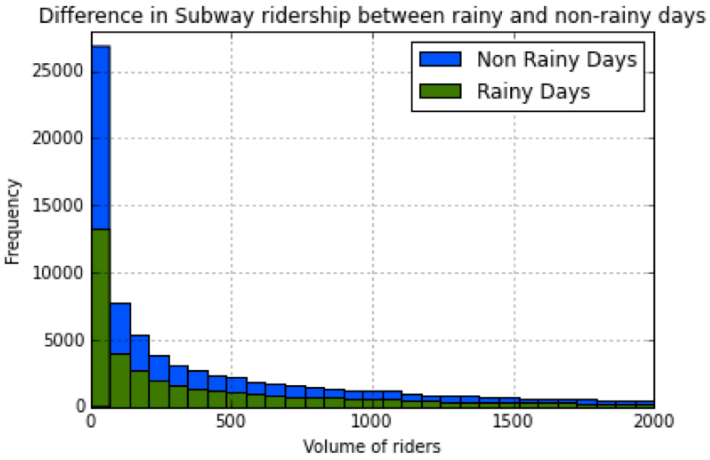
The R^2 value is a measure of the fit of our model to the data. An R^2 of 0 would mean that our model do not help explain the entries we record, that the variables we used do not affect the outcome. A perfect fit of 1.0 would mean that our model perfectly explains the observations we make, that the variables we tested completely explain the behavior of our data.

Honestly, as per my background in biological sciences with infinite variables and uncontrollable processes, I can affirm that a R^2 value of almost 0.50 is amazing. However, first we would have to compare this value to a Null Hypothesis (the R^2 to a median value of all the observations) to know if it explains the behavior better than the Null Hypothesis. Furthermore, with a little bit of experimentation and data manipulation I am sure we could find a better model to explain this dataset.

Section 3. Visualization

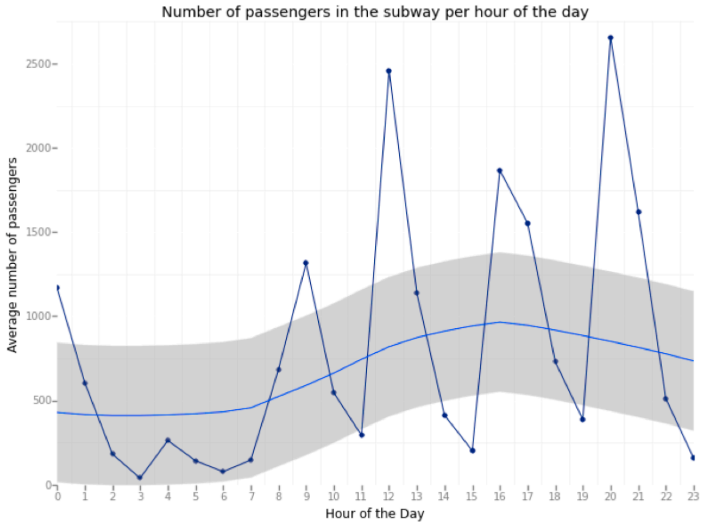
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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



In this figure we can see that, in general, there are more entries in our data of non rainy days that rainy days, and that the observations follow an skewed distribution.

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.



In this figure we can see how the use of the Subway is dependent on the hour of the way, clearly showing the peak hours when NYC people go to work, to eat, and home.

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?

From the analysis we have made, more people ride the NYC subway when it's raining. However, the difference, albeit statistically significant, is rather small: our data reports an average of approximately 15 more entries in rainy days in comparison to non-rainy days. Taking into account that on our dataset an average of more than 1000 people used the Subway every day, 15 more people is only a 1.5% increase.

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.

The Mann-Whitney U test reported that there was a significant difference in the means of the two samples, and in consequence, that we could reject the Null Hypothesis that their populations are equal. Therefore, we can confirm that more people ride the NYC Subway on rainy days.

Furthermore, we created a linear model that included weather variables (mainly if it rain or not and its quantification) to successfully predict the NYC ridership. For this, we can affirm that the rain affects the use of the subway.

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.

The dataset used in this analysis is undoubtedly small to generalize its conclusions to the daily NYC subway use. For example, we have much fewer entries with rainy days than non rainy days, as well as only one month (May of 2011), a timespan that could (but probably is not) be representative of the yearly weather conditions and NYC Subway ridership. Interestingly, the variable 'thunder' of our regression, which is normally directly related to the magnitude of the precipitation, is always 0, so we cannot evaluate if NYC people are more prone to ride the subway if they hear thunders.

In this project we used a non-parametric test, and although the result was significant, non-parametric tests are usually less powerful than parametric test, and with large enough datasets (for example, a yearly or two year span data) we could have used a parametric test that assumed normality to increase the statistical power of our analysis. Moreover, we used the OLS regression model directly to our data, without knowing if the data fits this type of linear regression. We did not do any prior transformation such as logarithmic to better prepare the data for fitting.