Analyzing the NYC Subway Dataset

Answers

Answers to the short questionnaire of Udacity's nanodegree in Big Data Foundation..

E-mail: sonamlpu657@gmail.com

References

The most relevant contents and source of information, I have used is Udacity course materials. Most of the code needed to complete this project was provided by Udacity to help the student complete the different problems and exercises of the Big Data course. On top of that code base. The main source code file is http://localhost:8888/notebooks/Anaconda3/Untitled.ipynb and can be accessed on GitHub (https://github.com/Sonam06/NYC-SUBWAY-DATA-ANALYSIS).

Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data?

I ran three Mann-Whitney U tests to analyze the effect of dichotomic independent variables `'time'`, `'rain'` and `'weekday'` on dependent variable `'ENTRIESn_hourly'`. The Mann-Whitney test is the non-parametric equivalent test to the two independent samples t-test, and it is used to compare differences between two groups when the assumptions of the t-test are not met. This is the case, as visual inspection of the histograms for these three variables showed clear signs of non-normality.

Did you use a one-tail or a two-tail P value? What is the null hypothesis?

I used a two-tail p-value as direction of the difference in ridership caused by the presence or absence or rain cannot be predicted beforehand. Two-tail tests are also more conservative than one-tail tests, which reduces the probability of Type I error (rejecting a null hypothesis that is actually true.

What is your p-critical value?

Since I ran the tests against the whole dataset provided (`turnstile_weather_v2.csv`), which has 42,649 entries, I selected a p-critical value of 0.01 instead of the typical 0.05. The reason is that with such a large sample sizes, even small differences are likely to be identified as statistically significant. In these cases it is recommended to use a smaller critical p-value to limit the risk of having false positives.

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 appropriate for all these three cases (fog, rain and weekday) because the next criteria are met:

- Observations are independent
- Sample sizes are greater than 20
- There are two groups to compare

- The dependent variable is measured at continuous or ordinal level

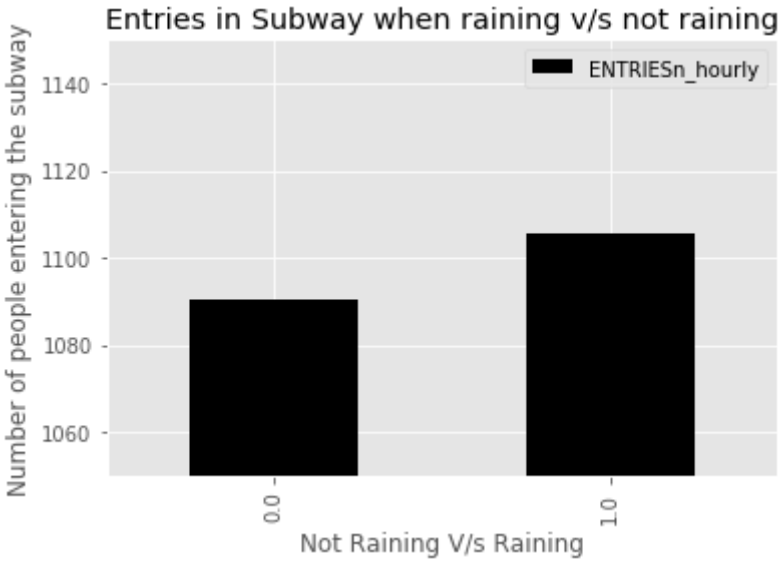
Also, this test is appropriate because the distributions of `ENTRIESn_hourly` split by `rain`, `fog` or `weekday` are not normal.

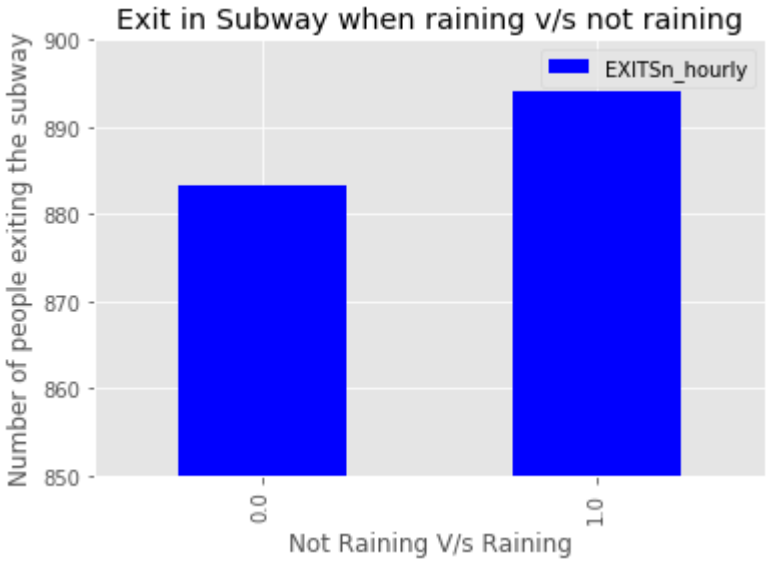
1.3 What is the significance and interpretation of these results?

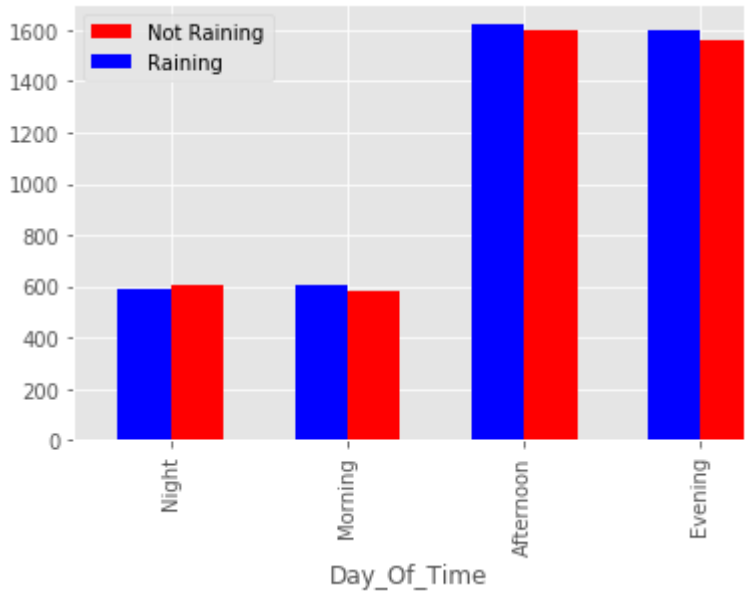
All these three tests found significant differences. That is, the likelihood of observing these data being the null hypotheses true is inferior to .001. However, the actual impact of fog and rain in ridership is very small, compared to the impact of 'weekday', which has the strongest influence.

Visualizations

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







Conclusion

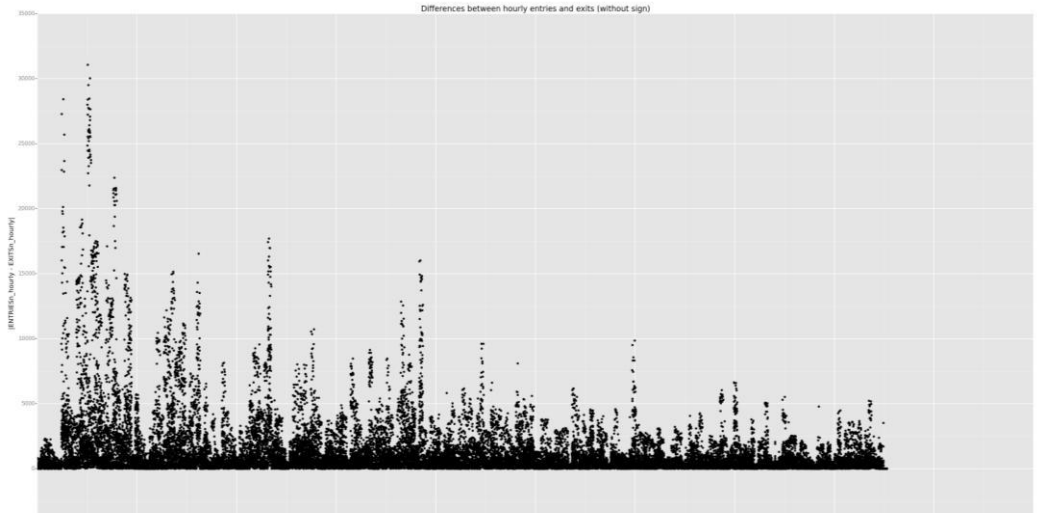
In absolute figures, there are less people riding the subway when it is raining because there are less rainy days in the data set provided than rainy days. As a result, the not-rainy series of the histogram shown in Figure 4 is taller than the rainy days series.
Also, the people entering the subways is more in working days as compared to holidays.

However, relative ridership frequency increases when it is raining, as proved by the results of the Mann-Whitney U test. This test shows a difference of 182.65 hourly entries (in average) in favor of rainy days, being this difference statistically significant ($p<.001$).

Reflection

One of the main shortcomings I have found in the data set is the limited number of dates considered. The dataset contains entries corresponding to the period from 28th April 2011 to 2nd June 2011, both days included (36 days in total). All dates are concentrated in spring, when precipitations and temperature are different than in other seasons (winter, autumn and summer). Also, not including data from other years may be a source of bias (2011 could have been a particularly sunny or rainy year). To reach solid conclusions, data entries should be distributed uniformly across different years and the 12 months of the year.

A detailed analysis of the first 5,000 elements showed recurrent large differences between hourly entries and exits. The difference between hourly entries and exits was calculated and plotted, showing a similar pattern to the residuals' (see Figure 7). This unexplained effect could be introducing noise into the model.



Difference between hourly entries and exits, unsigned, plotted for each data entry. Unusually large differences are also concentrated on the left side.

3.2