Analyzing the NYC Subway Dataset

Questions

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?

The statistical test used was the **Mann-Whitney U test.**

We performed the statistical set assuming that the condition rainy or the condition dry day will produce a variability of the results (amount of ridership), and not only an increase or only a decrease in the amount of ridership (therefore, **two tailed**).

Null Hypothesis $\mathbf{H_0}$: the ridership figures will be equal for the two different conditions (technically speaking, the populations are the same).

Alternative Hypothesis $\mathbf{H_a}$: the ridership figures will be not equal for the two different conditions (technically speaking, the populations are not the same).

Being *X* the given random draws from the "rainy day" ridership population and *Y* the given random draws from the "dry day" ridership population, we could express the Null ($\mathbf{H_o}$) and Alternative ($\mathbf{H_{1)}}$ Hypothesis as follows:

$$H_0: \ P\,(\,X > Y\,) \ = \ 0.5$$

$$H_1: \ P\,(\,X > Y\,) \ \neq \ 0.5$$

two-tailed-test

We performed the test with a p-critical value of 0.05, or 5%.

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.

We plotted the number of entries for a rainy and the number of entries for a dry day (section 3 of this document); it is observable that neither follows a **Normal distribution**; of course, these results are only indicative (visual). In order to prove it mathematically, we would have to perform a Shapiro-Wilk test. We do not know what distribution they follow, but we assume it is not Normal.

Having taken for good the visual interpretation, and without further investigating the variances, sample size…etc. we have to rule out the Welch`s test and perform the **Mann-Whitney U test**.

Possible Welch's test

- Samples independent (assumed)
- Sample sizes not equal (true)
- Variances (assumed not equal)
- Do not follow a Normal distribution (assumed from plotting) but we do not really know which distribution they follow

Unknown dist.?
Better Mann-Whitney U 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.

The results from the Mann-Whitney U test are:

- **Mean "rainy day"** = 1105.4463767458733 ~ **1105**
- **Mean "dry day"** = 1090.278780151855 ~ **1090**

- **U value**: 1924409167.0
- **pvalue**: 0.0049999825586978 ~ **0.005** (two-tailed)

1.4 What is the significance and interpretation of these results?

As **(p value 0.005 < p critical value 0.05) --> Reject Null Ho** (or accept Ha) @ 95 % Confidence.

Because the p-value obtained (0.025) is less than the critical p-value (0.05) we reject the Null hypothesis with a 95 % confidence; statistically speaking, we could say that **there are differences in the amount of ridership observed in the NYC Subway between a rainy day and a dry day** (technically speaking, the populations are different).

Apart from that, we cannot make conclusions with only the mean figures apart from saying that on average there are a few more entries on a rainy day compared to a dry day; only 15 entries more on average (out of a little bit more than a thousand), not an important difference.

On the other hand, the U value calculated is very high and that points in the same direction of the conclusion drawn from the p value (reject the null hypothesis).

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 computed the parameters or weights (coefficients Theta) and produced prediction for ENTRIESN_HOURLY in our regression model following the **OLS from StatsModels** approach.

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

The following input variables were used in the model: '**fog**' (foggy day or not), '**rain**' (rainy or dry day), '**precipi**' (rain precipitation), '**Hour**' (hour of the day) and '**meantempi**' (average temperature). All of them are numerical variables.

The variable '**UNIT**' (turnstile device) was used as a dummy variable in the model. Its nature is categorical, therefore not imputable directly to the model but still important. In this way, it is converted to 0, 1 values (numbers) so the variable can be taken into account.

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.

After trial and error with combinations of variables, we ended up with the highest R^2 with that set of variables.

All are "weather" variables, except 'Hour', and for example a very foggy-persistent day with a low temperature could have the same effect on the subway entries as a rainy day with low precipitation. Introducing 'fog' as a variable caused an increase in R^2. We tried numerous other combinations but we could not manage to increase R^2.

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

```
                Coef
----------------------
 const        1386.4841
 fog            99.0142
 rain          -40.4623
 precipi        10.8960
 Hour           67.3920
 meantempi      -8.3850
```

2.5 What is your model's R2 (coefficients of determination) value?

R^2 value is **0.479852626372 ~ 0.48**

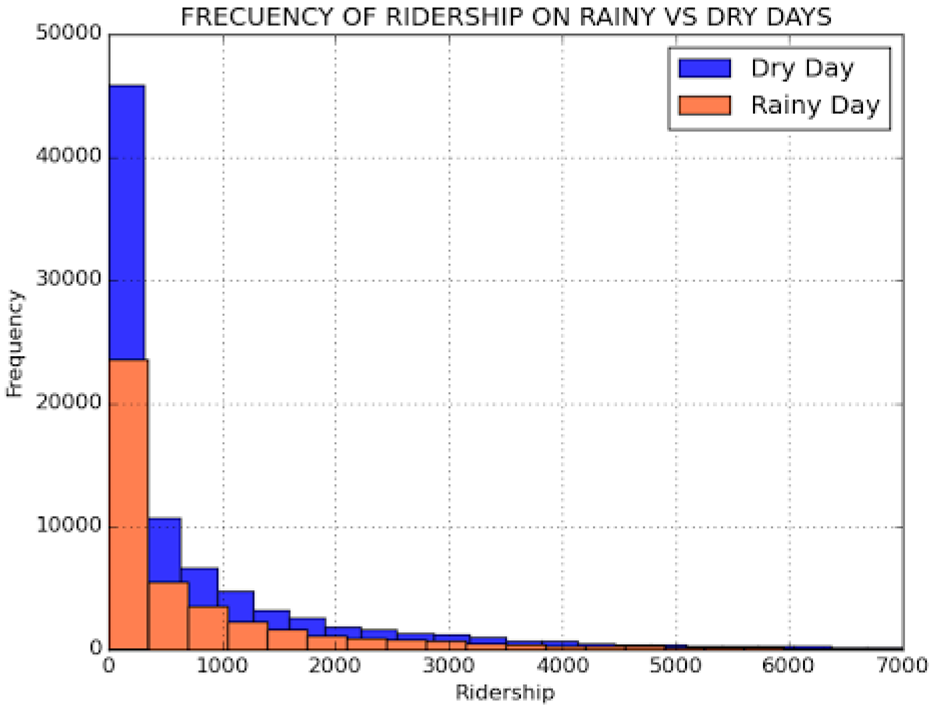
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

The 0.48 R^2 explains a 48% of the variation within the model; with this result we cannot say that the model is very accurate, or in other words that the goodness of fit is not so high for the model.

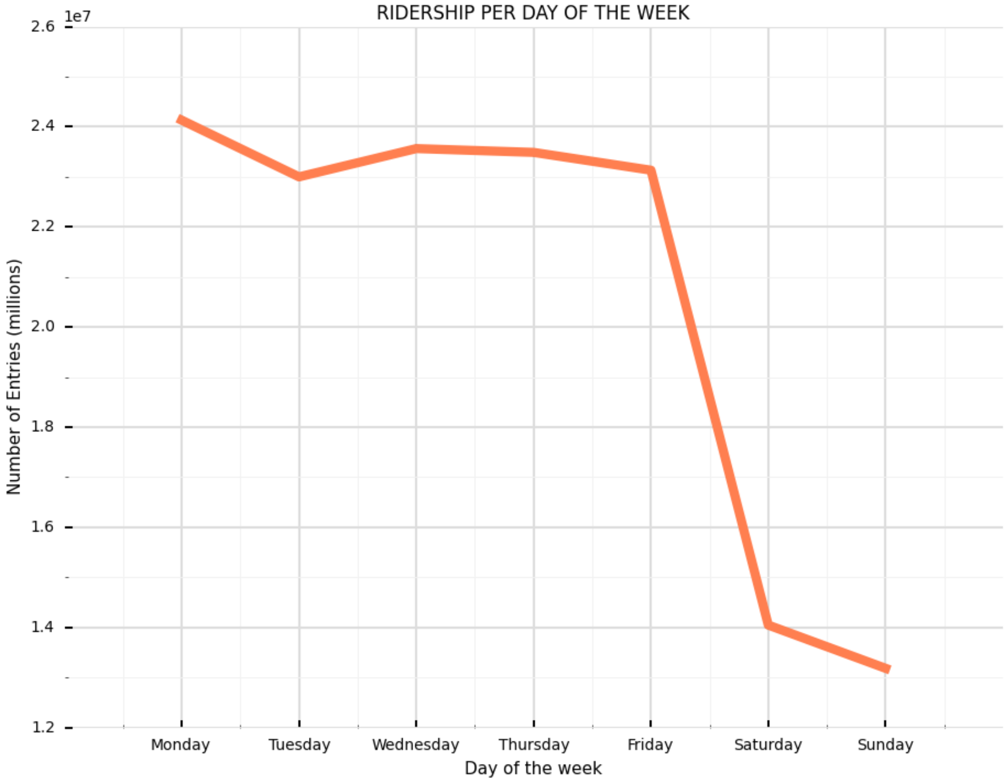
We think this model alone could be appropriate for this dataset, just being on the safe side. However, we have to take into account what will be the use and the nature of the study: for instance, a 0.48 R^2 for a critical medical study where people lives were at stake this R^2 should be unacceptable. However, maybe for other kind of studies it can be accepted (perhaps in our case).

Section 3. Visualization

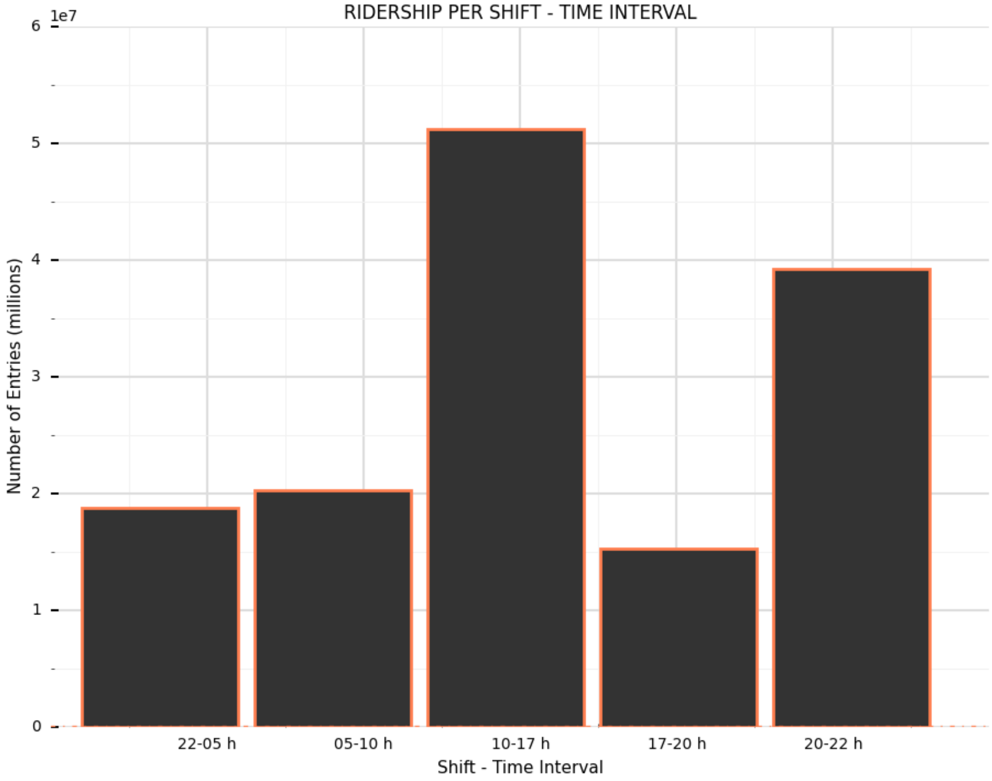
We present here several visualization regarding the NYC subway turnstile dataset.



The graphic from above shows the comparison of the ridership frequency ('ENTRIESn_hourly' variable) for a rainy day and for dry (no rainy) day. It is clear that both distribution do not follow the Normal distribution pattern. Please, bear in mind that only rained in about 1/3 of the days accounted for in the dataset so it is normal that the frequency of the rainy days is superior to the frequency of the dry days.



The graphic above shows the number of entries (millions) in the subway net by day of the week. It is clear that most entries are made during labor days and that there is an acute descend during the weekends. Could it be that in most cases the usage is work/business related? Entries on Saturday are more than on Sunday: perhaps on Saturdays the people go out more (open stores, theaters…etc) and decide to stay home on Sundays? We would need to have additional data in order to draw conclusions, data like demographics, fields related to postal addresses, lines (aggregation of turnstiles or stations)…etc.



Usually the rush hours (05-10h morning and 17-20h evenings) concentrate most rides in a subway net, however the passengers use this mean of transportation more during the central hours of the day (10-17h), and during the late evening (20-22h), do they go out for dinner or do they work long hours?.

Section 4. Conclusion

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?

We could say, although not categorically, that more people ride the NYC subway when it is raining.

We could also say, based on the Mann Whitney test, that there is a difference in riding between a rainy day and a dry day.

The model, while not being total irrelevant, does not explain much either, with a R^2 about 48%.

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.

To form the opinion expressed in the last question, we studied the mean of the two sets of data, we performed a Mann Whitney test, and we run a linear regression.

Based only on the mean figures, we could say that on average more people ride the subway on a rainy day than on a dry day, about 15 more people, but that is only about 1.36 % more people. The difference is so small that we should be cautious. We should study the procedure about how the data was collected, exogenous events that could have affect the results (for example, it could happen that all Madison Square Garden basketball games were in rainy days and that moved up the mean for rainy days….).

Regarding the Mann Whitney test, although the results were clear (the populations are different) we did not formally test the distributions for Normalcy (Shappiro).

About the linear regression, the R^2 is 48%. Although this is not bad it could have been better. We tried to increase the R^2 values testing several different mix of variables and also introducing a dummy variable.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis:

In our opinion the dataset is **correct** for the purposes of the study; maybe an increase of the sample size would have been better to get more accurate results. Anyway we found some **inconsistencies**: for example, there are about 27 million more entries than exits. That should raise concern about how the data was collected, although that difference did not a priori affected our study. Also, the TIMEn variable shows some inconsistencies being most entries recorded at a xx:00:00 format (by the whole hour) while others with a “by the second” 23:59:12 format. Is some cases one hour worth of entries not really one hour worth of entries?

It is crucial to study the **methodology**. What is consider a rainy day? How was it measured? What if in north NY it was raining and not in South NY (within the subway area)? Is that consider a rainy day or not? The same reasoning could apply to the fog variable, for example.

The data was take in May, which is consider a normal month of the year with kids are still attending school, people not yet on holidays and with “not so bad-little good” weather. We consider that month a good choice for a sample but it would have been better to study a whole year worth of data.

The **Mann Whitney U** test was applied here however a basic condition was not fully tested: the normalcy of the data distribution. Based solely on the plot the Mann Whitney U test is sufficient because it is more appropriate when dealing with skewed data (like our case).

The **OLS method** used to perform the linear regression have some shortcomings: the outliers effect (any data value that has a dependent value that differs a lot from the rest of the data will have a disproportionately large effect, due to the squaring effect of least squares), the non-linearity effect (in practice, most systems are not linear or in other words in real world relationships tend to be more complicated than simple lines), the use of too many variables (not in our case), the dependence among variables (if the variables are correlated to each other, for example fog and rain, rain and temperature, hour and temperature…), data noise (inconsistent values, for example the difference found in entries – exits), the wrong choice of features (for example, the variable mix we used to feed the model may not be the optimal one).

Section 0. References

http://blog.yhathq.com/posts/ggplot-for-python.html

https://pypi.python.org/pypi/ggplot/

http://pandas.pydata.org/pandas-docs/stable/visualization.html

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

https://en.wikipedia.org/wiki/Ordinary_least_squares

https://en.wikipedia.org/wiki/Gradient_descent

http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

http://stackoverflow.com/questions/20625582/how-to-deal-with-this-pandas-warning

http://pandas.pydata.org/pandas-docs/stable/api.html

http://stackoverflow.com/questions/21733893/pandas-dataframe-add-a-field-based-on-multiple-if-statements

http://bconnelly.net/2013/10/summarizing-data-in-python-with-pandas/

http://social-metrics.org/python-pandas-cookbook/

http://www.stackoverflow.com (by the hundreds)