Caio Pumar Freitas

June 2017

Analysis of the NYC subway dataset

This report goes through the multistage process of dealing with the NYC subway dataset on the year 2011 using the Python Programming Language(-V 2.7.10) and appropriate libraries to: (1) Wrangle data, (2)Find out its shape and general trends through the use of statistical tests, (3) Do regression ("shallow" learning) to extrapolate future numbers in subway ridership.

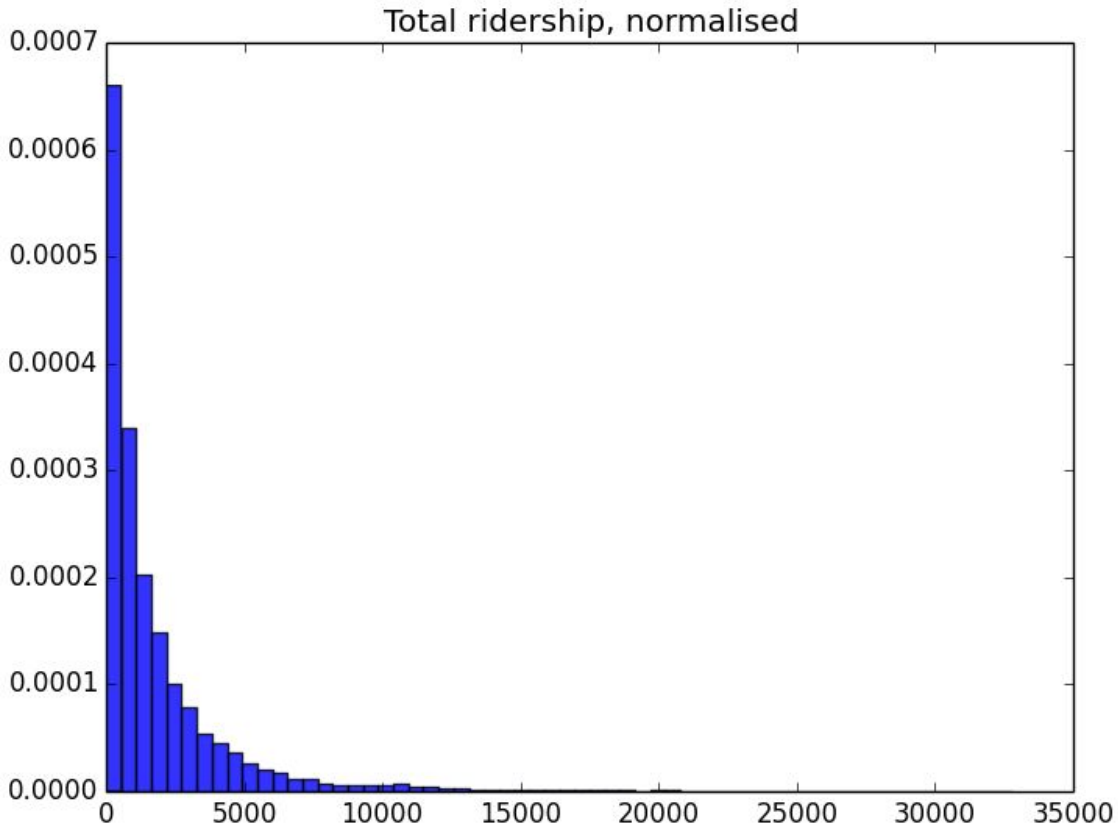
Statistical Tests

The first step was loading the data in a Dataframe object and rearranging rows and columns to make it easier to extract the exact information that would be analysed. In general, that entails dividing the 'ENTRIESn_hourly' column into parts that represent different parameters that could interfere with ridership, such as rain, fog, which day of the week it is, etc.

Although it would be arguably simpler to just plot the histogram and check that 'ENTRIESn_hourly' is absolutely not normal, a normality test(Shapiro-Wilk test) was also done which rejected the Null Hypothesis (The distribution **is** normal) with a $p < e-07$.

In order to compare ridership under different circumstances the respective slices of 'ENTRIESn_hourly' were compared with a Mann-Whitney U Test, which does not assume normality. One of the comparisons done was between subway entries when raining or not. The results ($p < 0.000554348$) determine that the ridership distributions get changed on rainy days, which in turn suggests that a regression might be useful to predict future behaviour, e.g. *if it*

rains the NYC subway might expect more or fewer passengers, or might expect a peak on a different time of the day.



Linear Regression

An ordinary least squares linear regression was done using Python's library statsmodel. Out of the total number of variables present on the dataset, a fraction of them were selected to optimize the fit. Evidently, data that was not numerical had to be cut off from the regression and also information like the distance (in latitude and longitude) from the weather station that collected the data to the city of New York. With the parameters calculated a $R^2 = 0.44$ was

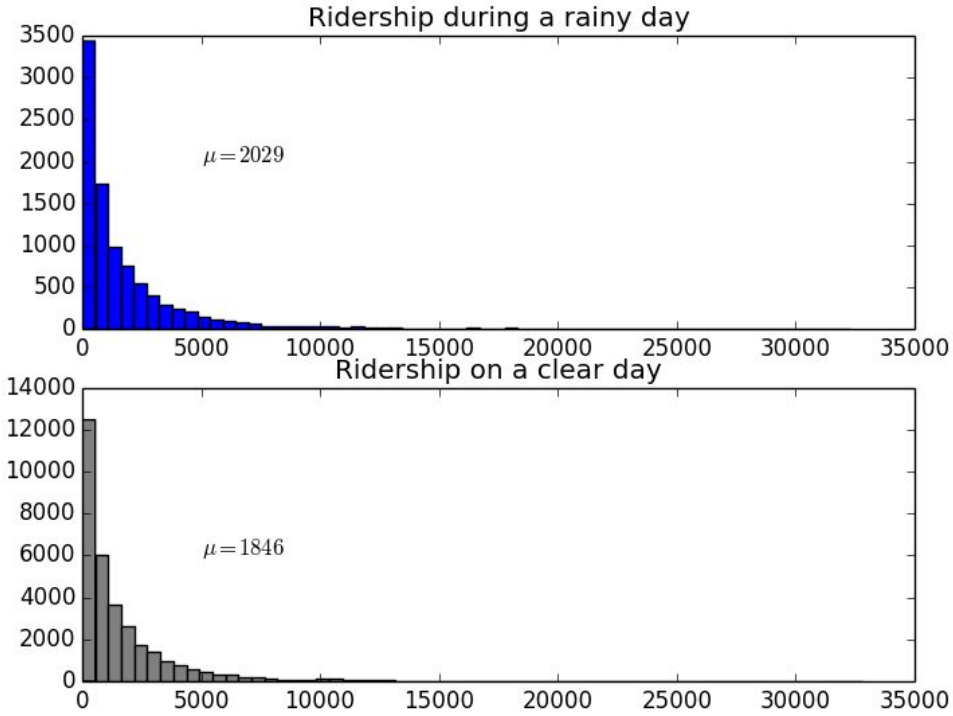
achieved when column choice was experimented upon. Namely, the final parameters used were(dummy variables have a *):

"EXITSn_hourly; hour; day_week; weekday*; latitude; longitude; fog*; precipi; pressurei; rain*; tempi; wspdi; meanprecipi; meanpressurei; meantempi; meanwspdi"

A R^2 value is a measure of how well the line is fit to the data, it ranges from 0 to 1 and not always the largest R^2 is desirable. You might be overfitting and that actually decreases the predictive power of the model. Especially in cases where the data comes from human sources, you can't get as good a fit as a dataset that describes a physical process.

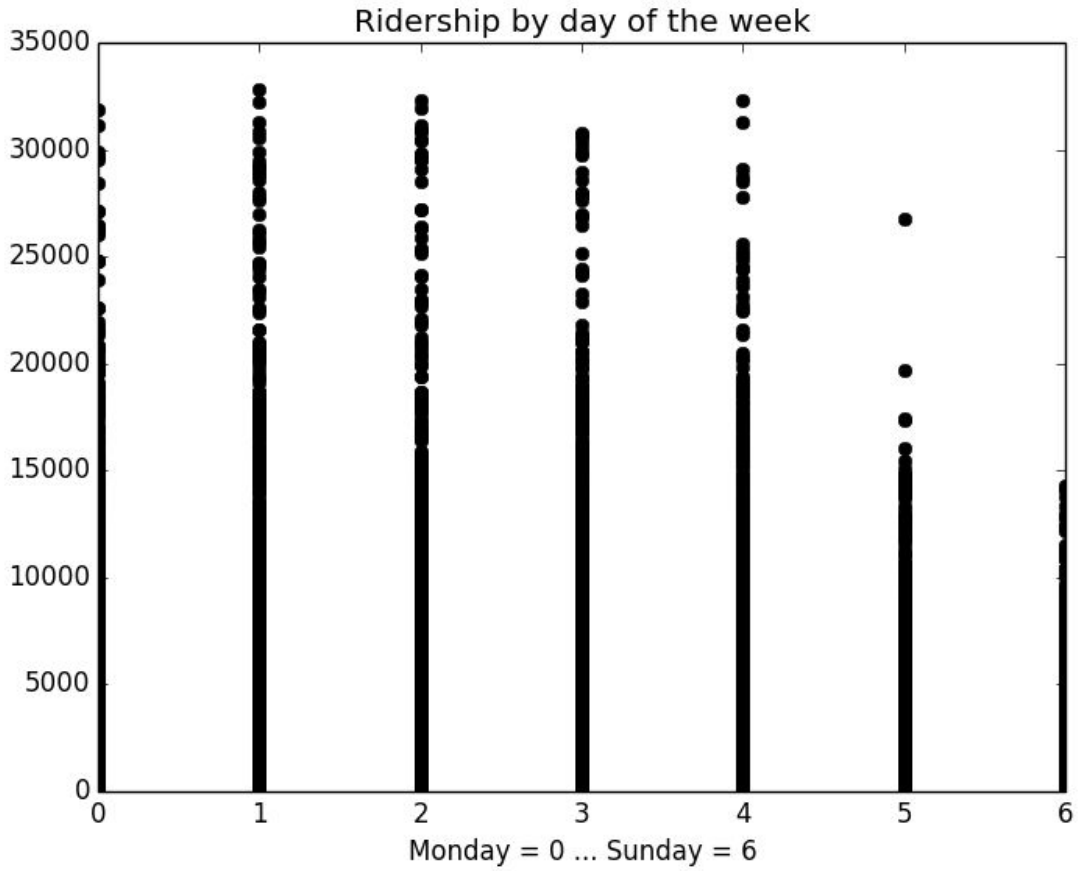
Visualization

Being visual creatures that we humans are, no amount of data can bring out an epiphany the way a nice plot can, especially from a layman's perspective:



The plot above suggests that more people ride the subway when it rains, an expected result.

Also that there are more clear than rainy days in NYC, thankfully!



Another intuitive result, new yorkers ride the subway more often in the work days.