Intro to Data Science final project - NYC Subway Analysis

Diego Lins de Freitas

Introduction

The New York City Subway System is managed by Metropolitan Transportation Authority and delivered over 1.71 billion rides in 2013. It is an important public transit systems of New york. The question that this project is trying to answer is "how does rain affect ridership in the New York City Subway system?". With predictions for the ridership in the subway it is possible for MTA provide better services allocating resources to attend the demand. This project aims to use statistical and visualization tools, machine learning techniques to to build a model to make predictions.

Data Wrangling

In order to make predictions upon data it is necessary to collect it, clean it and prepare it to build a good model. The data of the ridership from period of may 2011 is used in combination with the weather data, both available at:

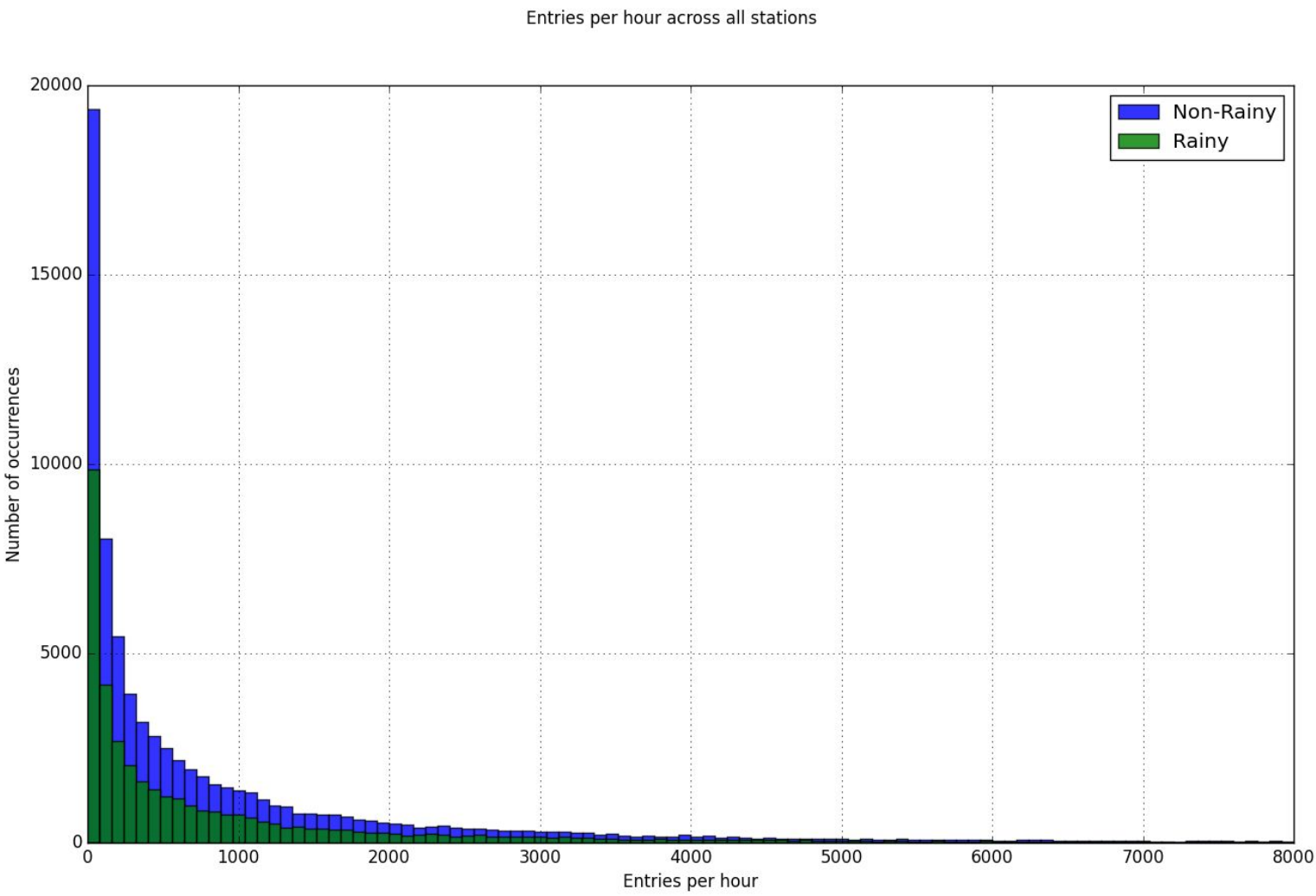
http://web.mta.info/developers/data/nyct/turnstile/turnstile_110507.txt

The data combined can be found at

https://www.dropbox.com/s/meyki2wl9xfa7yk/turnstile_data_master_with_weather.csv

Data Analysis

It is necessary to check if the research question is a valid question. In order to do that, an analysis in the data is done to verify if there is a any correlation between rainy days and the volume of the ridership and apply a statistical test. Two samples of the data were taken from the available data: the ridership of a rainy day and the ridership of a non-rainy day. The entries where the value for column ENTRIESn_hourly is 0 or greater than 8000 was removed to improve visualisation.



The histogram of the samples show that the data is not normally distributed, so the Welch's t-test cannot be applied, then the Mann-Whitney U-Test was used to test the data

Mean of Rainy Day	Mean of Non-rainy day	U	p	2 * p
1105.4463767458733	1090.278780151855	1924409167.0	0.0193096344	0.03861927

It was defined a P critical value of 0.05. Since i am testing two distributions and Mann-Whitney U-Test give one side p-value it is necessary to double the p-value. With a two-sided p-value of 0.038 is lower than the P critical value, it means that there is a difference between rainy and non-rainy days.

Machine Learning

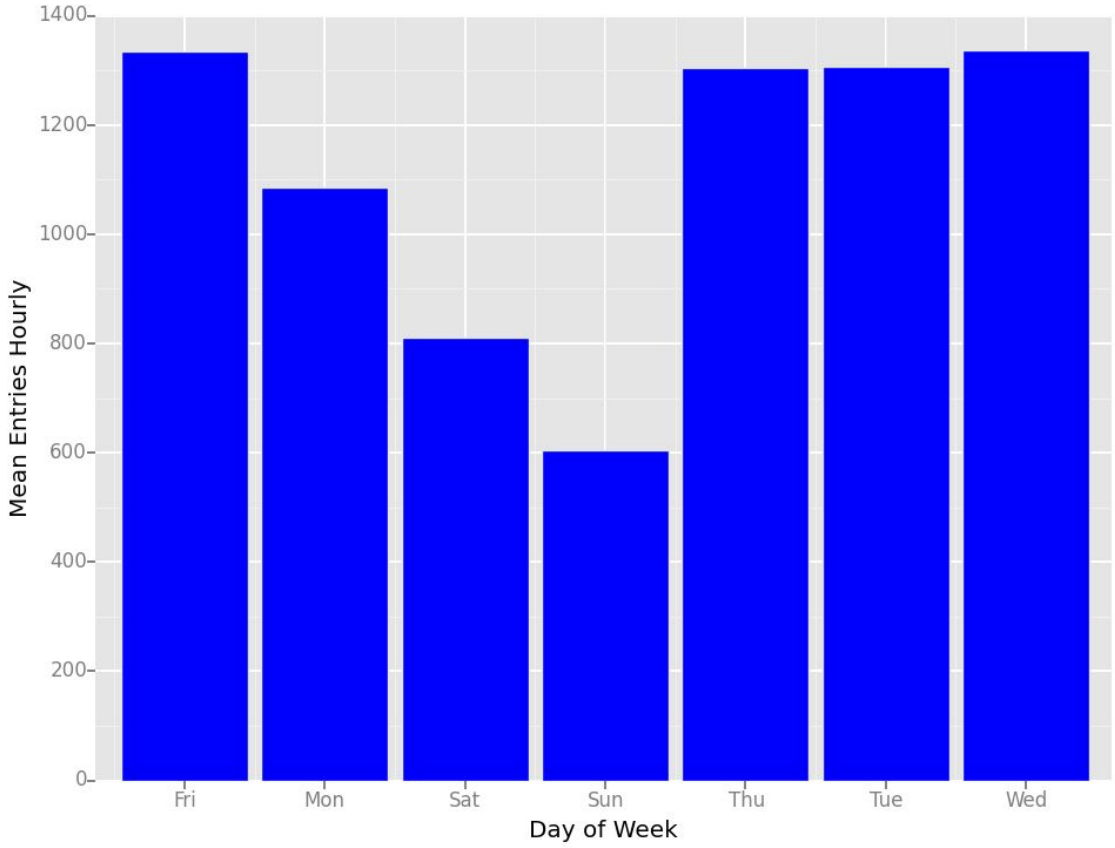
Linear regression using Gradient descent algorithm is used to create the model. First i selected only a one feature that indicates if it is a rainy or non-rainy day. The data is not normalized.

Features	**rain**
Theta	200.36432936
R^2	0.401786414273

This value of R^2 is not good and i tried other features in order to improve the R^2.

Features	**rain**	**Hour**	**meantempi**
Theta	256.00414382	608.11187034	107.68016786
R^2	0.461129645864		

The table above shows the best results a can get with the available data as is. Tried to guess other features that a can create based in the existing data and realized that the day of the week should be a good one. the following graph show that some day of the week has a little difference in the ridership.



Then i put the week day as a feature in the model. The week day was extracted from the column DATEn using the standard ISO 8601 where the days are represented from 1 through 7, beginning with Monday and ending with Sunday. The predictions gets a little better with the new feature as shown below.

Features	**rain**	**Hour**	**meantempi**	**dayofweek**
Theta	248.60649279	607.45532691	113.25132085	-289.64641964
R^2	0.474350279029			

Following is the result for without the rain as a features:

Features	**Hour**	**meantempi**	**dayofweek**

Theta	606.5228954	5.20641571	-293.84983304
R^2	0.467902779175		

MapReduce

Map Reduce is not used in this project because data set is too small and the algorithm to implement gradient descent could become very complex. It would be better to have more than one month of data to train the model. A good example of application of map reduce with a large dataset for this problem would be to calculate the average value of the for other types of weather that is not explicit in the dataset using the combination of the columns **fog** an **rain**. Below is the output of a simple mapreduce program to calculate that average:

weather	**Average**
fog-norain	1315.57980681
fog-rain	1115.13151799
nofog-norain	1078.54679697
nofog-rain	1098.95330076

Conclusions

There is a shortcoming with gradient descent. If we used a larger dataset that span not only through a month, but years, with a more complex model using other variable, it will take much more time to run the program because gradient descent is very slow.
Other algorithm could be used to compare the performance between gradient descent, like ordinary least squares, since i think a value of 0.47 for is not good enough to rely on. The result of the algorithm without the rain as a feature show a difference in the R^2 of 0.0064475 that confirms the results of the Mann-Whitney U-Test, so we can say that the rain affects the ridership of the New York City Subway system.

Source Code

https://github.com/diegofreitas/ds-nycssystem