Analyzing the NYC Subway Dataset Documentation

Release 1.0

Ignacio Toledo

January 03, 2015

CHAPTER

ONE

OVERVIEW

This project consists of two parts. In Part 1 of the project, we have completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.

This document addresses part 2 of the project, where we answer a set questions to explain our reasoning and conclusions behind our work in the problem sets.

The main purpose of the project is to analyze the ridership behavior for the New York City subway. The dataset used contains a sample taken from the month of May 2011, using the publicly available turnstile data from MTA. The turnstiles in different stations of the system report the absolute number of entries and exits at certain hours for a given time interval. The improved dataset that we use reports the number of entries for time intervals of 4 hours, so it present us with 6 daily reports by turnstile.

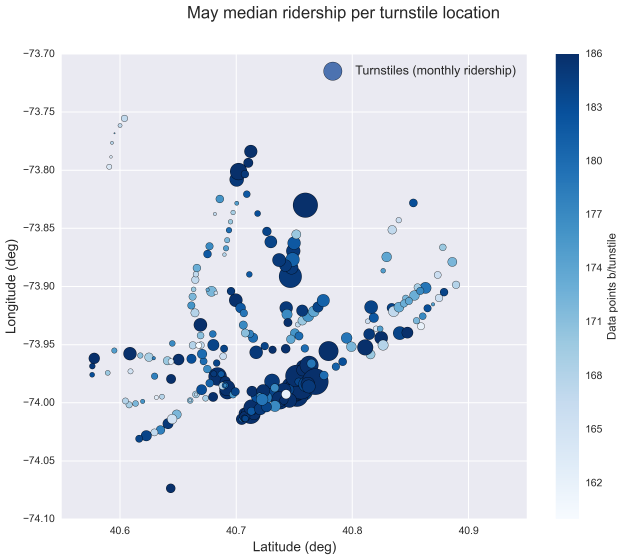

Figure 1.1: Turnstiles' locations from the improved dataset.

Besides the information provided by the NYC subway, the dataset also includes weather information taken from several weather stations within the NYC area: each turnstile, depending on the location in NYC, is merged with the weather information of the closest weather station, thus providing temperatures, wind speed, pressure, conditions, precipitations, etc.

The project focuses one main question: Does the weather conditions, specifically precipitations, affect the NYC subway ridership? To answer this question we will use exploratory tools, statistical tests and visualizations. Also,

we will try to fit a model to the data by choosing certain predicting features; will the use of the precipitation variable improve the fit?

1.1 Supporting Material

Within the project github repository you will also find an ipython notebook, where most of the work done was recorded for reference.

1.2 Some remarks about the datasets used

For this project we use the data set provided at Data Analyst Nanodegree's portal for Project 1. The description of the variables can be found on...

However, after the exploratory and data analysis, we created another dataset by further munging the improved dataset. The basic idea was to smooth out features that might be caused by individual turnstiles or measurements. To do this, I grouped the data by time stamp and aggregated the entries by hour by adding all the entries. Also, the precipitation information for each time stamp was included by means of two columns:

- `rain_hour`: indicator (0 or 1) for precipitations for the particular date date and time. It is 1 if for any of the stations the conditions were Rain, Light Rain, Hard Rain or Light Drizzle at that moment.
- `rain_day`: indicator (0 or 1) for precipitations for the particular day of the report. If at any station of our turnstiles the conditions reported precipitations during the day the value is set to 1.

CHAPTER

TWO

STATISTICAL TEST

In lecture 3 and its problem set, the following question was given *Do rainy days affect the ridership of the NYC subway?* To answer this problem we began by

creating two samples from our data:

- Sample A (*No rain*) is a subgroup containing the entries where no rain was reported, using the information of the *rain* variable ($rain = 0$)
- Sample B (*Rain*) is a subgroup with the entries where some precipitation was reported by means of the *rain* variable ($rain = 1$)

By studying the distributions, using histograms, we were able to characterize both data samples. We found out that both samples have a similar shape, clearly not normal, and positively skewed (*figure 2.1*).

Figure 2.1: Ridership distribution comparison between rainy and dry days.
Please note the logarithmic scale on axis Y. It was used to allows us to study the visualization with more detail.

Because of the non-normal distribution we decided to use the median as measure of average for the samples:

- Sample A, days without precipitation, show a **median ridership of 901 passengers per hour.**
- Sample B, rainy days, report a **median of 945 passengers per hour.**

To assess the significance of this result, that rain seems to increase ridership in the NYC system by a small amount, we will use a non-parametric test.

2.1 Statistical Test Used

The Mann Whitney U test is chosen to assess the statistical significance of this result. The null hypothesis in our case is that both populations are equal, or that there is no significant deviation on both populations medians (two-tailed hypothesis).

2.2 Justify the Statistical Test

The Mann Whitney U test, or Wilcoxon rank-sum test, is chosen because of characteristics of our samples: we can't use a parametric test because the distributions do not seem to follow any particular and well known probability distribution which we could use to make inferences that could directly report the significance of any difference between both populations.

The U test is particularly powerful to assess the significance of the difference between the median of two samples that have similar distributions. The assumptions that our data samples must comply with are basically:

- All observations of both groups are independent
- The responses are ordinal (so we can use the ranking algorithm of the U test).

2.3 Results

We used the scipy implementation of the Mann Whitney U test (scipy.stats.mannwhitneyu). The results from the test are:

- $U = 150678745.0$
- $p = 1.91 \cdot 10^{-6}$

But the user should be aware that scipy reports p-value for a one-tailed hypothesis, so we multiply by 2 to get the significance for our hypothesis:

- $p = 3.82 \cdot 10^{-6}$

2.4 Interpretation and discussion

The interpretation, given the result from the U test, is that the the ridership is not the same for rainy days than non-rainy days, with a significance higher than 95% ($p < 0.05$). Furthermore, from the descriptive statistics of our samples we can conclude that the ridership tends to be higher in rainy days.

However we have limited ourselves here to follow the procedure suggested by the lectures, assuming that observations of both groups are independent and there no other factors that might wrongly induce this result. Even when the data sample we use for the project has been through a more complete wrangling, there are still some issues that might affect the results:

- There is missing data for several turnstiles. From the original sample of 240 turnstiles, only 52 have complete data for May;

- As discussed on the forums, some precipitation data is missing from some weather stations. But more important, we are using the variable *rain* to create our samples: this variable indicates if there was some precipitation at anytime of the day in the particular turnstile. Is the appropriate variable to use to build the subgroups?

Let's look with more detail at these problems.

2.4.1 Missing data

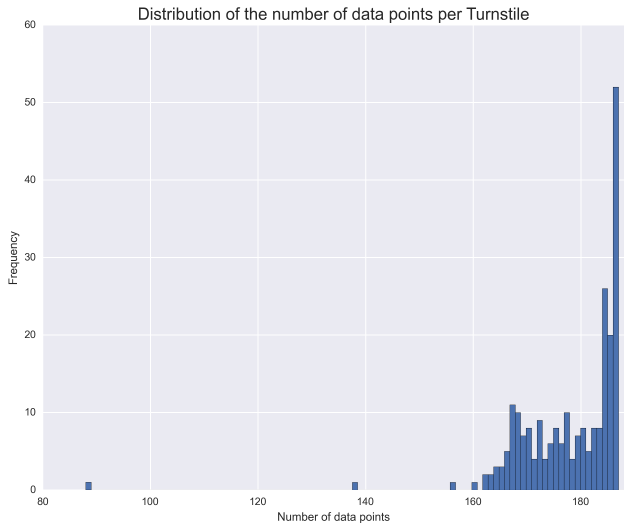

Figure 2.2: Number of data points (measurements) by turnstile on our improved dataset.

Figure 2.2 shows some turnstile have missing data for the month of May; with 31 days and 6 daily reports it is expected that a complete monitored turnstile should have 186 measurements. This is the case for 52 turnstiles, but a 185 turnstiles have a number of measurements between 160 and 185. 3 turnstiles had less than 160 entries, and after inspections they have been removed because of the huge amount of missing data or time stamps reporting 0 entries. Of the 185 turnstiles with incomplete data, there was one case where at all time stamps the number of entries was 0, which was also removed as it does not add any information to our analysis (even when in other cases it could give further information).

CHAPTER

THREE

LINEAR REGRESSION

3.1 Linear regression algorithm(s)

3.1.1 Gradient descent

3.1.2 OLS (with statsmodels)

3.2 Features used

3.3 Feature selection: why?

3.4 Results: R Squared

3.5 Interpretation and limits

CHAPTER

FOUR

VISUALIZATION

4.1 Ridership distribution with weather

4.2 Supporting visualizations

CHAPTER

FIVE

CONCLUSION

CHAPTER

SIX

REFLECTION

6.1 Shortcomings and limitations

6.2 Insights