Analyzing the NYC Subway Dataset

Project for Data Analysis Nanodegree

Author: Nikita Barsukov

Overview

This is my first project for a Data Scientist nanodegree, which is a series of courses on various topics of Data Science created by Udacity. In this project I analyzed a dataset about ridership in New York City subway. Dataset was provided by Udacity.

In particular, I tried to find out which weather conditions influence ridership, predict amount of people entered a station, and visualize this dataset. Python code used to calculate all the statistical figures, and to generate plots, is available on my personal GitHub account: https://github.com/nikita-barsukov/intro_to_ds/

References

Since the supporting course used Python as a programming language, and Python is a relatively new programming language for me, I used various on-line resources a lot. Time and again I consulted these on-line manuals:

- Beginners tutorial on Python, https://wiki.python.org/moin/BeginnersGuide/Programmers
- On-line reference for NumPy package, http://docs.scipy.org/doc/numpy/reference/index.html
- On-line documentation for Pandas package, http://pandas.pydata.org/pandas-docs/version/0.15.1/
- Documentation for GGplot for Python, https://github.com/yhat/ggplot

More extensive list of references is given at the end of this paper.

Section 1. Statistical Test

First I tried to determine if weather conditions influence subway ridership. Dataset contained indicators of rain and fog, I decided to check if rain and/or fog has a statistically significant influence over subway ridership.

I decided to use overall passenger turnover, which is sum of entries and exists over a certain point in time, as a measurement of ridership in subway. Since its distribution is non-parametric, as suggested in “Intro to Data Science”, a Udacity course that accompanies this project, I used **Mann Whitney U-test** to asses if rain and fog have statistically significant influence on ridership.

Both for rain and fog, I used **two-sided P-value**, with **null hypothesis** being **two samples** (passenger turnover with and without respective weather condition) **come from the same distribution**. I chose **P-critical** value to be **1%**.

	Rain	*Fog*
P-value	~0	0.00015
Mean passenger turnover	Rain: 3488 / No rain: 3179	Fog: 2796 / No fog: 3253

Mean passenger turnover for entire dataset is 3248

Performed statistical tests strongly suggests that for both weather condition, ridership is different with and without it. We can also see that ridership during rain increases. Ridership during fog on the other hand decreases. This may be related to the fact that there are more service interruptions during fog.

Section 2. Linear Regression

In this section I tried to find a way to predict hourly entries and asses prediction found prediction model. I also aimed to test effectiveness of several approaches to building regression model, namely:

- Adding more prediction variables;
- Using ordinal least squares and gradient descent.

I carefully selected which columns to use in my regression models. The four features from a linear regression assignment from accompanying course were the obvious choice. Then I added fog, day of the week and wind speed. I decided not to add variables like maximum temperature, since essentially it might have the same effect as chosen variables, mean temperature in this case.

At the end I selected 7 variables for testing regression models:

- 'rain',
- 'precipi',
- 'hour',
- 'meantempi',
- 'day_week',
- 'fog',
- 'wspdi'

Then I created an iterator over these variables. In this iterator I selected all the variables up until current iterator position, and used them as feature variables for ordinary least squares model and for gradient descent model. Then I calculated predictions of hourly entries, and calculated mean absolute error and R-squared parameters. At the end I got two arrays of mean absolute error, one for OLS and one for gradient descent, and two arrays for R-squared. All arrays were of length 7, which corresponds to maximum number of used variables.

As suggested in exercises on linear regression in Intro do Data Science course, I added dummy feature in both OLS and gradient descent models. I used feature normalization for gradient descent, but not on OLS.

I plotted mean absolute errors and R-squared coefficients of my models on Figure 1 and Figure 2 below.

We can see that regression models perform differently, based on what criterion is selected. Models with gradient descent have significantly lower mean absolute errors. That means that predictions of hourly entries differ less on average from actual hourly entries. However R-squared coefficient, which shows how close data is fitted to regression line, is visibly better for OLS models.

Generally, adding more features increases model performance. It is important to note that the improving rate is not even. Both OLS and gradient descent models had a leap in efficiency after adding third variable. These four variable s were default variables used for training regression models in exercises for accompanying course ‘Intro to Data Science’. Also it worth noting that adding two last parameters, ‘fog’ and ‘wspdi’ had almost no effect on our fitness parameters.

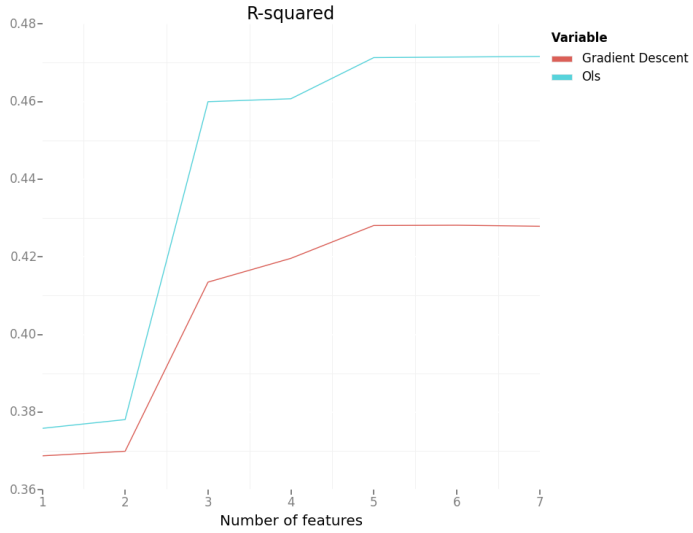

Figure 1 - R-squared coefficient of various regression models

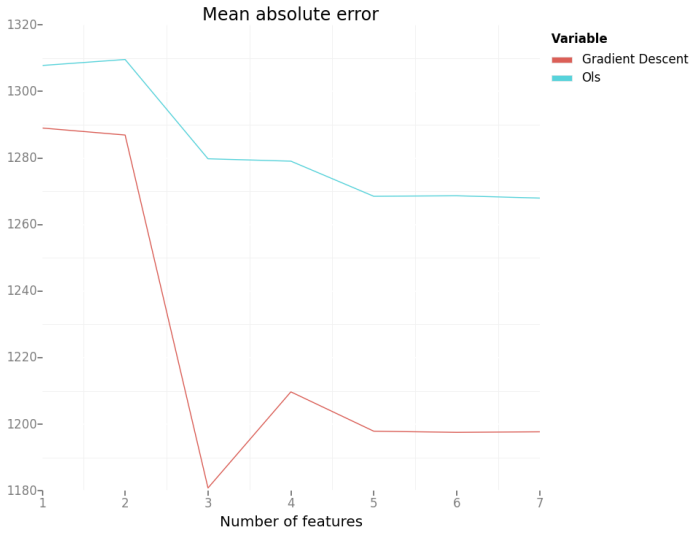

Figure 2 - Mean absolute error of various regression models

It is also worth noting that minimal absolute error was in a model with gradient decent and four features, contrary to intuition that increased number of features improves prediction quality.

Based on the above I would choose ordinary least squares with five parameters: 'rain', 'precipi', 'hour', 'meantempi', 'day_week'. Parameters of these non-dummy features are as follows:

Table 1 - Feature parameters of linear regression with least R-squared coefficient

Parameter	Value
'rain'	106.49817262
'precipi'	-3662.4343221
'hour'	122.22245505
'meantempi'	-15.61404268

'day_week'	-147.27794206
Intercept	2011.08917196

Section 3. Visualization

In this part I was required to make two visualizations. One should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days. I decided to make one plot with two histograms, colored by rain indicator. This visualization is displayed in Figure 3.

Second visualization was up to a student. I decided to plot average daily passenger turnover of New York subway on a map of New York City. Passenger turnover is defines as sum of columns ENTRIES_n and EXISTS_n. This visualization is displayed in Figure 4

Because of my limited knowledge of plotting capabilities of Python graphical libraries, I created both charts using R, specifically ggplot2 and ggmap library.

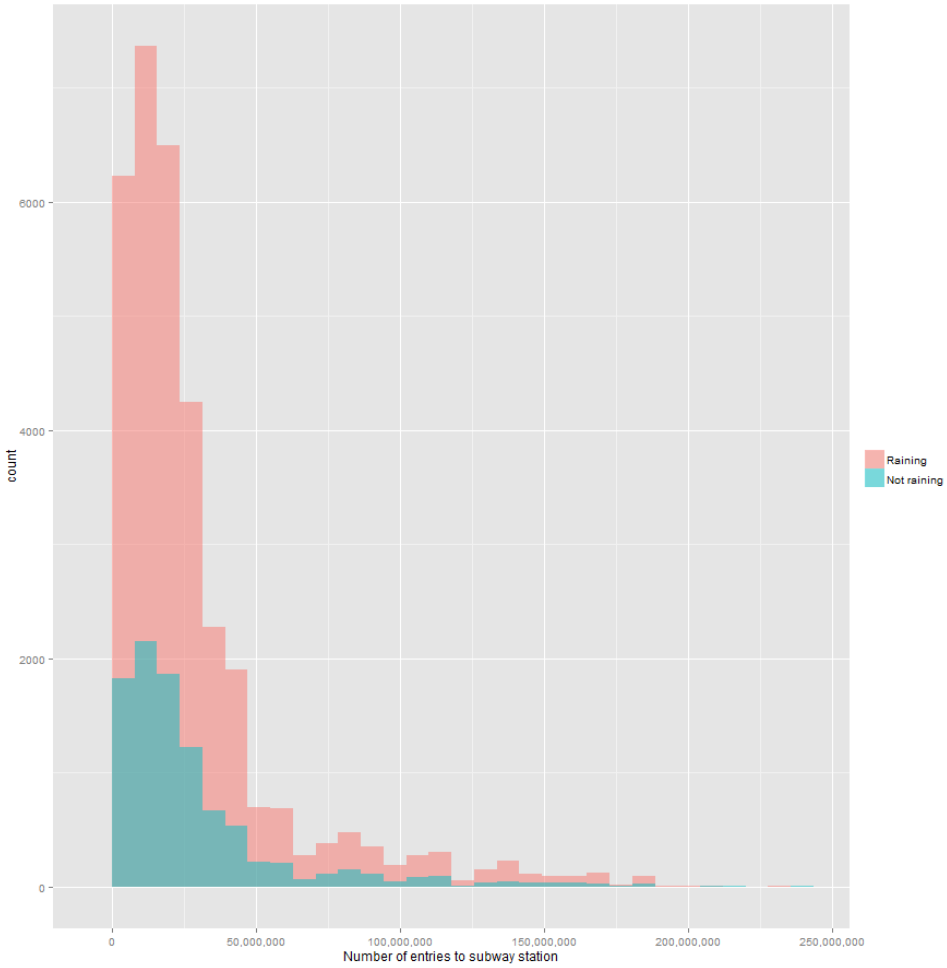

Figure 3 - Distribution of entries to subway stations in New York City

From the histogram above we can clearly see that number of data points on rainy days is significantly more than on not rainy days. However from this plot only it is hard to estimate which days have more ridership. More appropriate plot would be a line plot of average ridership by hour, grouped by rainy and non-rainy days.

Figure 4 - Daily passenger turnover in New York City subway stations

Map above shows broad ridership patterns in NYC subway. Largest passenger turnover have stations in mid-Manhattan, close to Central park. There are also several stations with large passenger turnover on end stations, especially north of Bronx. These stations can be transit stations to popular sub-urban train lines.

Section 4. Conclusion

In this paper I have found several insights to ridership patterns in New York City subway. Weather conditions, particularly fog and rain, influence greatly on number of passengers using subway. Ridership increases on rainy days, and decreases on foggy days.

I have also tried to develop a prediction model of number of entries bases on various features in the dataset. I tried several models with various number of features, using ordinary least squares algorithm and linear gradient descent algorithm. Regression model using 5 features with ordinary least squares had best R-squared coefficient.

Two additional visualizations revealed geographical patterns in ridership in New York City subway. Most passengers use stations in middle Manhattan, and at the end of subway lines.

Section 5. Reflection

There are several shortcomings in this research, relating both to the research itself and to dataset used.

Dataset contains ridership information from November 2011, which is almost four years old as of date of writing. Updating this dataset would add new stations built during these years. Ridership patterns might change as well. Having larger timespan, perhaps several months or even a year, would greatly improve quality of dataset. We will see not only patterns within a day or week, but within whole year as well.

My research has several potential shortcomings too. For example, I did not test linear gradient descent with larger number of iterations, this can yield even better performance of my regression model. Also I chose features based purely on intuition. Using more formal methods for feature selection, like principal component analysis would add rigor to my regression model.

Topics for further analysis might include elements of spatial analysis. For example we could try to find turnover not just by station, but by route, and answer questions like “Where do passengers go in the morning on a weekday”. We can also juxtapose the dataset with other publicly available data. Using ridership data with, for example, crime rate in the area, residential density or average income might give us interesting insights into how new yorkers use subway.