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 on-line documentation on Python and R packages, I used during writing this paper. I also used Q&A portals like StackOverflow, or specialized blog posts to solve specific code issues or get coding tips, especially in Python. An extensive list of references is given at the end of this paper.

1 Statistical Test

In this section I tried to determine if weather conditions influence subway ridership. Dataset contained indicators of rain and fog, I decided to check if rain has a statistically significant influence over number of passengers using subway. Testing if foggy weather influences ridership would be interesting as well, however the number of datapoints with fog is only 419, which around 1% of entire dataset.

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?

Distribution of number of entries with and without rain is not normal, so it is incorrect to use t-tests. I used **Mann Whitney U-test** to asses if rain has statistically significant influence on subway usage, since it does not make any assumptions about underlying distribution. More detailed explanation of chosen test is given in section 1.2 below.

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%**.

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.

To answer the question above, I chose column **ENTRIESn_houry** as an indicator of number of passengers using subway. In order to decide which statistical test to use here, let's try to determine the distribution of **ENTRIESn_houry** when it rains and when it does not. Two histograms of these observations are plotted on Figure 1.

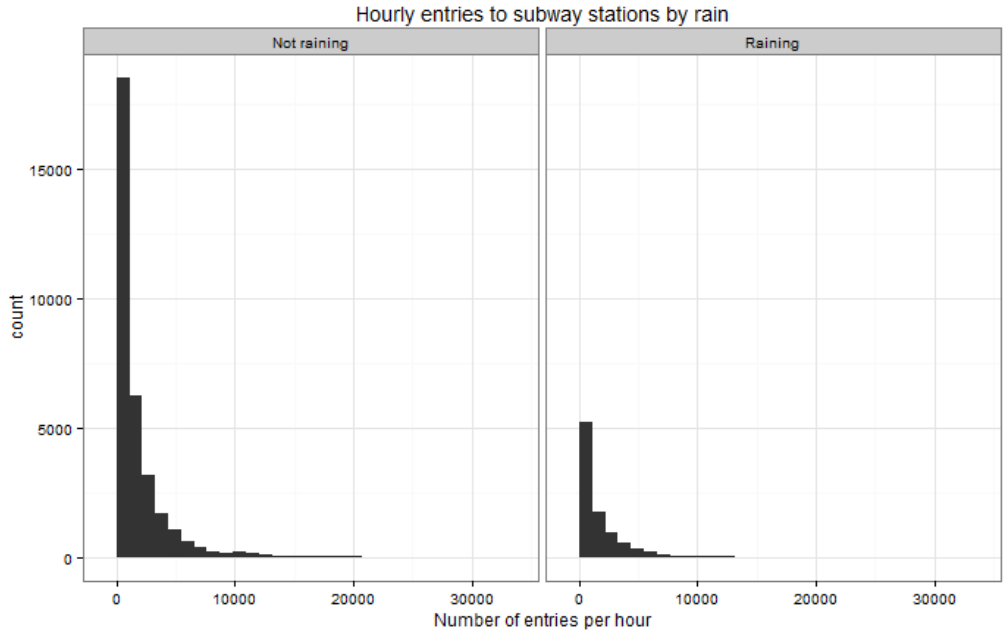

Figure 1 – Distribution of ENTRIESn_hourly variables during rainy and clear days

We can see clearly that distribution of two observations is not normal. It does not have a bell-shaped curve. Histogram suggests that this variable has exponential distribution. Thus we cannot use t-test for this distribution.

An alternative to t-test is a **Mann Whitney U-test**. It is non-parametric, which means that it does not make any assumptions about underlying distribution. I decided to use this test to determine if usage of subway is different during rainy and non-rainy days.

1.3 What results did you get from this statistical test?

Performed statistical test gave following results:

	Value
P-value	~0
Mean hourly entries, rain	2028
Mean hourly entries, no rain	1845
Mean hourly entries, overall	1887

1.4 What is the significance and interpretation of these results?

Performed statistical test is very significant since it produce extremely low value, much less than p-critical value. This means that the probability of null hypothesis is true is almost non-existent. Performed test strongly suggests that usage of New York Subway, as measured by number of passengers entering subway stations, is different when it rains, compared to when it doesn't rain. We can also see that ridership during rain increases.

2 Linear Regression

In this section I tried to find a way to predict hourly entries and estimate effectiveness of found prediction model. In order to select best approach and number of features, I selected 7 variables for testing regression models:

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

This choice was made purely on intuition. Detailed reflection on how I selected these seven variables is given in section 2.3 below.

After that 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 coefficient. I plotted mean absolute errors and R-squared coefficients of my models on Figure 2 and Figure 3 below.

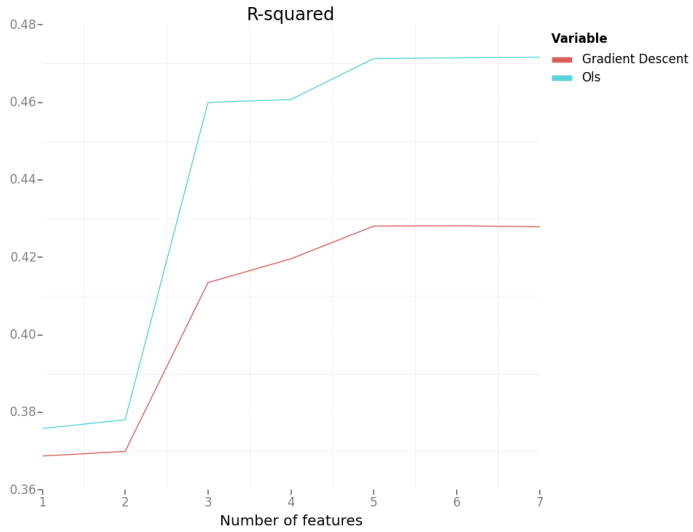

Figure 2 - R-squared coefficient of various regression models

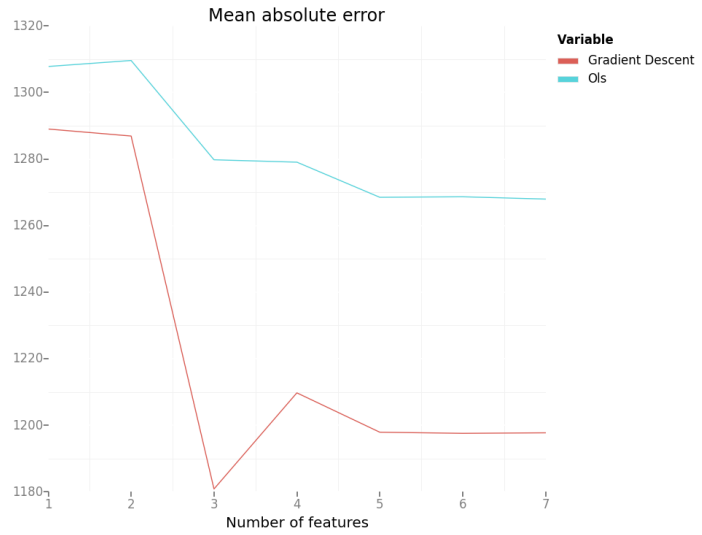

Figure 3 - Mean absolute error of various regression models

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model

Based on my findings above, I decided to choose ordinary least squares approach to predict ENTRIESn_hourly in my regression model. This approached yielded highest R-squared coefficient, compared to an alternative approach, which was gradient descent. It is also worth noting that minimal absolute error was in a model with gradient decent and three features, contrary to intuition that increased number of features improves prediction quality.

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

Largest R-squared coefficient was produced by Ordinary least squared model with five parameters: 'rain', 'precipi', 'hour', 'meantempi', 'day_week'. As suggested in exercises on linear regression in Intro do Data Science course, I added dummy feature 'UNIT' in both OLS and gradient descent models. I used feature normalization for gradient descent, but not on OLS.

2.3 Why did you select these features in your model?

As I mentioned above, I used intuition to select initial features to test how selected features influence efficiency of regression model. These features are distinct from each other. For this reason I did not use mean daily parameters, like meanpressurei etc. I also did not use timestamp variables and categorical variables with large number of factors (ex. 'conds' that contains different weather conditions). I also excluded station coordinates and coordinates of weather station, since their influence in ridership seems weak.

Of selected seven variables I chose variables that:

a. Produce largest R-squared in regression model
b. Of two models with similar R-squared I chose the one with less number of variables

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

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

2.5 What is your model's R-squared (coefficients of determination) value?

R-squared parameter for ordinary least squared model with variables 'rain', 'precipi', 'hour', 'meantempi' and 'day_week', dummy-variable 'UNIT' and an intercept is **0,471**.

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?

In order to assess whether the linear model is appropriate to predict ridership for this dataset, let's start with plotting a histogram of residual errors of our selected model.

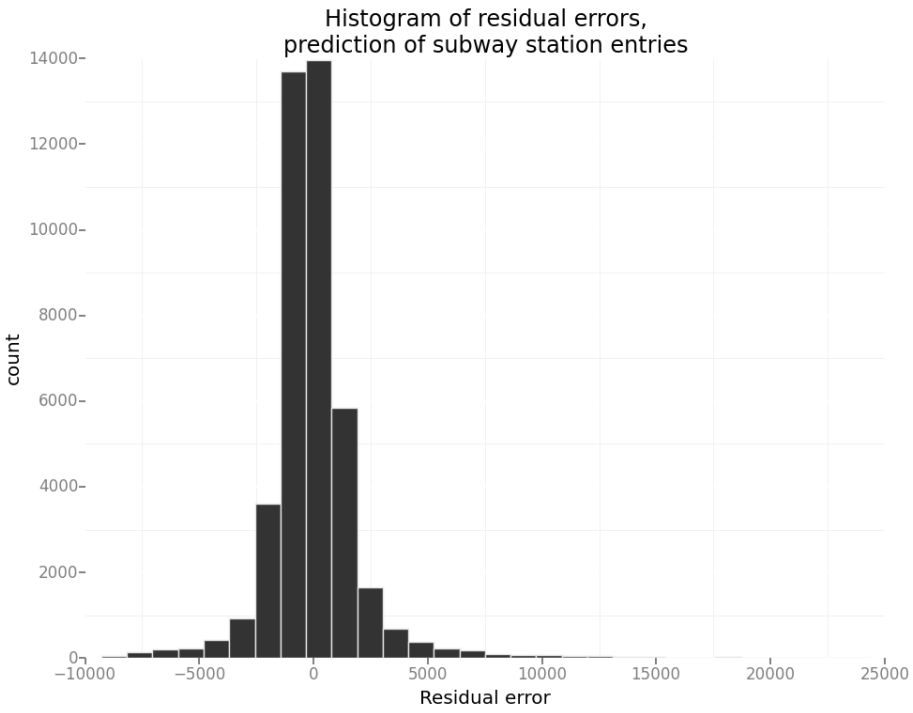

Figure 4 – Histogram of residual errors of chosen regression model

We can see that residual errors look like it follows normal distribution. We might be tempted to determine that the distribution is normal. However this distribution has very long tails since it has a few very large residuals. Q-Q plot of residual errors looks like this:

Figure 5 – Q-Q plot of residual errors in linear prediction model of subway entries

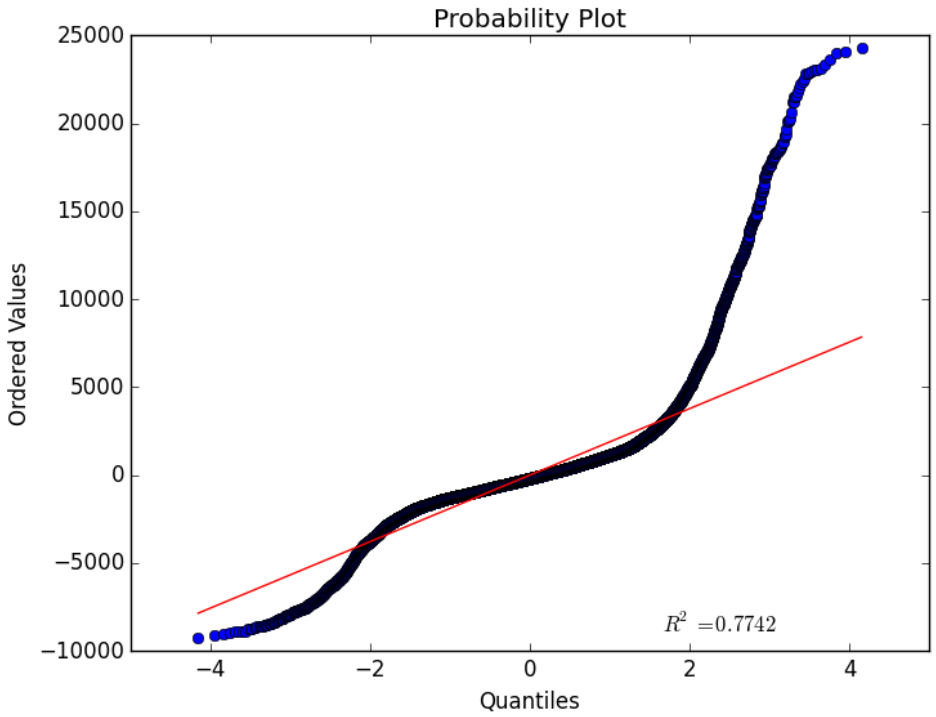

Figure 6 – Q-Q plot of residual errors in linear prediction model of subway entries

It shows that although residual error of chosen model look like they have a normal distribution, in fact this distribution is not normal. Thus we can see that linear model with selected variables is not a good choice for this dataset.

Now let's look at R-squared and mean absolute error to estimate how effective is selected regression model.

The R-squared value in my regression model is small, **0,471** to be precise. It means that less than half of response variable variation is explained by current regression model. Thus, regression model designed in this paper is not appropriate for this dataset. This is supported by mean absolute error of this model, **1268**, which is 67% of mean hourly entries.

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 7.

From this histogram we can clearly see that number of data points when it's not raining is significantly more than on 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. A more rigorous attempt to determine how rain influences subway usage is described in section 1 above.

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. I defined passenger turnover as a sum of columns ENTRIESn_hourly and EXITSn_hourly. This visualization is displayed in Figure 8.

This map 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.

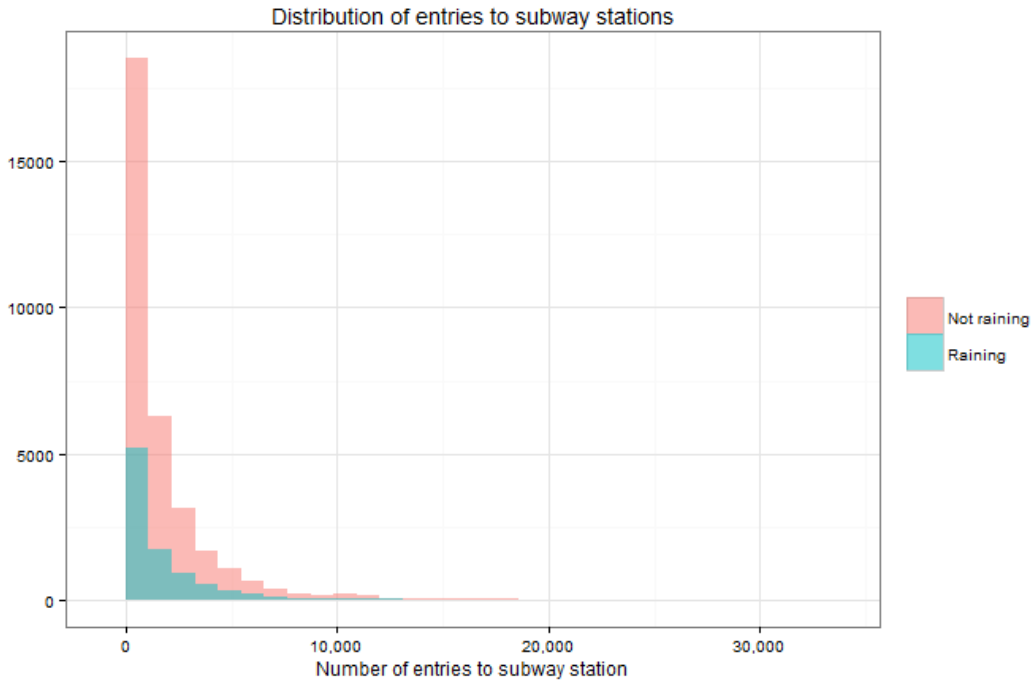

Figure 7 – Distribution of entries to subway stations in New York City

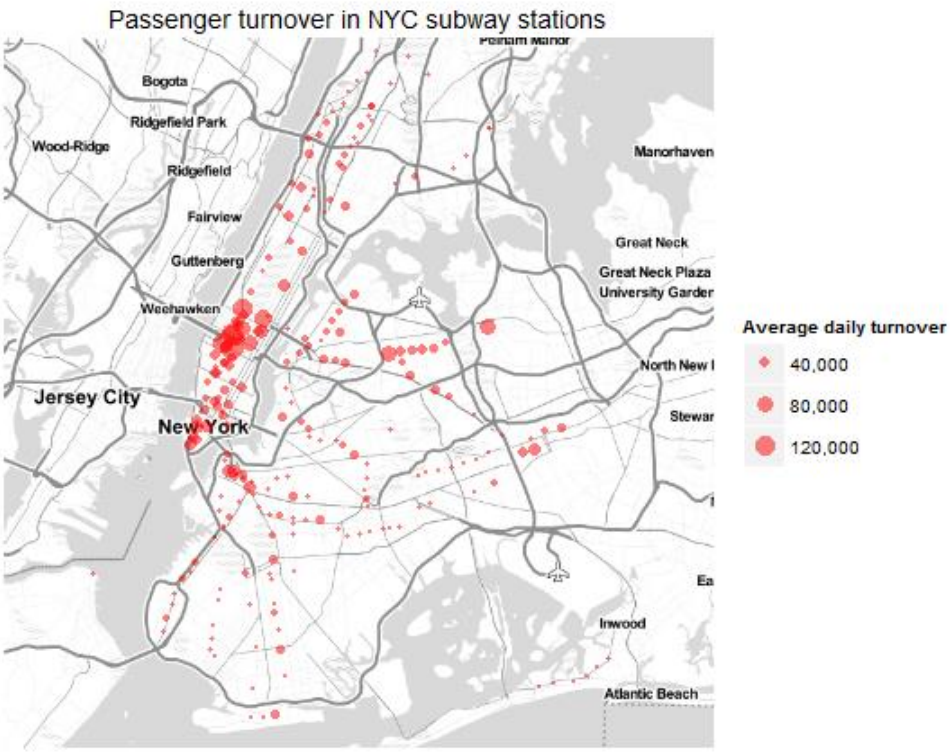

Figure 8 - Daily passenger turnover in New York City subway stations

4 Conclusion

In this paper I have found several insights to ridership patterns in New York City subway. In particular, rain influences significantly on number of passengers using subway: subway usage increases on rainy 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 largest R-squared coefficient. However quality of that prediction model was not sufficient: mean absolute error was more than 60% of median, and R-squared coefficient was around 47%.

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

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?

Statistical analysis performed in sections 1.3 and 1.4 above shows that we can almost certainly state that more people ride the NYC subway when it's raining.

4.2 What analyses lead you to this conclusion?

Mann-Witney U test performed in section 1.3 showed with that subway usage in NYC increases during rainy days with great confidence. We can also see that from linear regression model developed in section 2 above. Feature parameter for 'rain' variable is 106.5. Since there are only two possible values for this variable, 1 and 0, we can see that rain increases subway usage, given all other values are the same.

5 Reflection

When I developed linear regression model, I tried to see which of two approaches, OLS or linear regression, produces better results. I also wanted to determine if adding more features to the model increases model efficiency.

Not surprisingly, regression models perform differently, based on selected approach. 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.

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 temporal 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 better performance of my regression model. Using non-linear function in gradient descent could produce more effective model too. Also I chose features based purely on intuition. Using more formal methods for feature selection 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.

6 List of sources

1. Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux. **The NumPy Array: A Structure for Efficient Numerical Computation**, Computing in Science & Engineering, **13**, 22-30 (2011), DOI:10.1109/MCSE.2011.37
2. Wes McKinney. **Data Structures for Statistical Computing in Python**, Proceedings of the 9th Python in Science Conference, 51-56 (2010)
3. H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.
4. D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf
5. Chang, Winston. R Graphics Cookbook. Beijing: O'Reilly, 2013. Print.
6. http://stats.stackexchange.com/questions/116315/problem-with-mann-whitney-u-test-in-scipy
7. https://github.com/dkahle/ggmap/blob/master/man/get_stamenmap.Rd
8. http://stackoverflow.com/questions/5352099/how-to-disable-scientific-notation-in-r
9. http://barsukov.net/programming/2014/07/26/endomondo-code/