Assignment on
DATA MINING

Prepared By:
Md Monsur Ali
Matriculation Number: 24547
Course: Data Mining
Faculty: Communication and Environment
Semester: WS2017 / 18
Hochschule Rhein-Waal(HSRW)

Prepared to:
Tim Stockheim
Faculty of Communication and Environment
Hochschule Rhein-Waal(HSRW)

Date of submission: 28.01.2018

Abstract:

Classification and patterns extraction from customer data is very important for business support and decision making. Timely identification of newly emerging trends is very important in business process. Large companies are having huge volume of data but starving for knowledge. To overcome the organization current issue, the new breed of technique is required that has intelligence and capability to solve the knowledge scarcity and the technique is called Data mining. The objectives of this paper are to procure energy for the next year at the wholesale market and charge customers once a year for their energy consumption. Basically, this is an Energy Trading Company and it have seven years customers energy consumption data. Using data mining tools, clustering, analyzing to get some information regarding customer consumption, how much energy company will buy for their customer. In the first phase, cleansing the data and developed the patterns via Rstudio and K means clustering algorithm. Using clustering, it answers question no a. In the second phase, filtering data and plotting chart to answer question b, c and d

Table of Contents

List of Tables & Figures

1.Introduction:

For a successful business, identification of high-profit, low-risk customers, retaining those customers and bring the next level customers to above cluster is a key task for business owners and marketers. Traditionally, marketers must first identify customer cluster using a mathematical mode and then implement an efficient campaign plan to target profitable customers. This process confronts considerable problems. The Energy Trading Company supply electricity to the customer where most of the customers are households and they have a long-term relationship with the company. The business model is to procure energy for the next year at the wholesale market and charge customers once a year for their energy consumption. It need the value "meter reading" only once a year for billing. But there are also other reasons for meter reading so often it have a few more readings from the metering device of the customer. The difference to the last value is the amount of consumption in the time between the two dates. If the company buys too much or not enough energy in advance the difference is procured with an additional fee against the spot market. Therefore, it is very important to accurately estimate how much energy customers will require in the future. For existing customers, the common understanding is that the expected energy amount for the next year is the energy amount required the customer required the year before. Buying energy is the responsibility of the Procurement Department so they do most of the calculation but sometimes they need data analysis.

2.Methods and Materials:

2.1Cluster:

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

Below some clustering algorithms name are shown:

i. Hierarchical clustering
ii. k-means clustering
iii. Distribution-based clustering
iv. Density-based clustering

2.2K-means clustering:

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycentre's of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words, centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function.

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \qquad \text{…….. (1)}$$

where μi is the mean of points in Si. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{\mathbf{x},\mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|^2 \qquad \text{…….. (2)}$$

2.3Boxplot Chart:

In descriptive statistics, a box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. The spacings between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data, and show outliers. In addition to the points themselves, they allow one to visually estimate various L-estimators, notably the interquartile range, midline, range, mid-range, and trimean. Box plots can be drawn either horizontally or vertically. Box plots received their name from the box in the middle.

2.4RStudio:

R is a powerful language and environment for statistical computing and graphics. It is a public domain (a so called "GNU") project which is similar to the commercial S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S, and is much used in as an educational language and research tool. The main advantages of R are the fact that R is freeware and that there is a lot of help available online. It is quite similar to other programming packages such as MatLab (not freeware), but more user-friendly than programming languages such as C++ or Fortran. You can use R as it is, but for educational purposes we prefer to use R in combination with the RStudio interface (also freeware), which has an organized layout and several extra options.

RStudio layout:

The RStudio interface consists of several windows

i. **Bottom left:** console window (also called command window). Here you can type simple commands after the ">" prompt and R will then execute your command. This is the most important window, because this is where R actually does stuff.
ii. **Top left:** editor window (also called script window). Collections of commands (scripts) can be edited and saved. When you don't get this window, you can open it with File → New → R script Just typing a command in the editor window is not enough, it has to get into the command window before R executes the command. If you want to run a line from the script window (or the whole script), you can click Run or press CTRL+ENTER to send it to the command window.

iii. **Top right:** workspace / history window. In the workspace window you can see which data and values R has in its memory. You can view and edit the values by clicking on them. The history window shows what has been typed before.

iv. **Bottom right:** files / plots / packages / help window. Here you can open files, view plots (also previous plots), install and load packages or use the help function.

2.4Libraries:

R can do many statistical and data analyses. They are organized in so-called packages or libraries. With the standard installation, most common packages are installed. To get a list of all installed packages, go to the packages window or type library () in the console window. If the box in front of the package name is ticked, the package is loaded (activated) and can be used. There are many more packages available on the R website. If you want to install and use a package (for example, the package called "geometry") should:

- **Install the package:** click install packages in the packages window and type geometry or type install.packages("geometry") in the command window.
- **Load the package:** check box in front of geometry or type library("geometry") in the command window.

3.Data and Results:

For this study, customers electricity consumption of data of the company has been taken. Using these data, customers have been clustered using RStudio tool. The first steps in the clustering process involve selecting the data set and filtering data for applying the algorithm. K-means clustering algorithm were used to these data and step by step answered all a, b, c, and d questions. Below a short description are presented,

3.1Data Filtering:

First, both datasets are filtered by pod_id and merged. After that, all columns name was changed to recognize them very well. To shift bill reading and actual date dataset, it used lag. Shifting them, it used to make differences reading and date. It needed yearly electricity consumption of customers that's why calculated yearlyconsum column. Finally remove all null row and save all dataset as finaldataset. Below

	podid	date	billedreading	mrtype	mrcategory	country	postalcode	city	LONG	LAT	LAGbilledreading	LAGdate	energyconsumption	dateofconsumption	yearlyconsum	regioncluster
1	P000001	2011-09-15	40326.000	11	2	DE	97355	Wiesenbronn	10.30000	49.75000	33859.000	2010-03-01	6467.000	563	4193	3
2	P000001	2012-09-22	44608.000	11	2	DE	97355	Wiesenbronn	10.30000	49.75000	40326.000	2011-09-15	4282.000	373	4190	3
3	P000001	2013-09-10	49034.000	11	2	DE	97355	Wiesenbronn	10.30000	49.75000	44608.000	2012-09-22	4426.000	353	4576	3
4	P000001	2014-09-21	53371.000	11	2	DE	97355	Wiesenbronn	10.30000	49.75000	49034.000	2013-09-10	4337.000	376	4210	3
5	P000001	2015-10-02	57736.000	11	2	DE	97355	Wiesenbronn	10.30000	49.75000	53371.000	2014-09-21	4365.000	376	4237	3
6	P000001	2016-09-15	62236.000	11	2	DE	97355	Wiesenbronn	10.30000	49.75000	57736.000	2015-10-02	4500.000	349	4706	3
7	P000001	2017-03-31	64916.000	11	2	DE	97355	Wiesenbronn	10.30000	49.75000	62236.000	2016-09-15	2680.000	197	4965	3
8	P000002	2010-05-31	56695.000	7	1	DE	81375	MÃ¼nchen	11.57550	48.13740	64916.000	2017-03-31	-8221.000	-2496	1202	3
9	P000002	2011-04-16	61590.000	10	1	DE	81375	MÃ¼nchen	11.57550	48.13740	56695.000	2010-05-31	4895.000	320	5583	3
10	P000002	2017-04-25	85260.800	11	2	DE	81375	MÃ¼nchen	11.57550	48.13740	61590.000	2011-04-16	23670.800	2201	3925	3
11	P000003	2010-04-20	67389.000	1	1	DE	92353	Postbauer-Heng	11.35720	49.30530	85260.800	2017-04-25	-17871.800	-2562	2546	3
12	P000003	2011-04-20	70319.000	10	1	DE	92353	Postbauer-Heng	11.35720	49.30530	67389.000	2010-04-20	2930.000	365	2930	3
13	P000003	2012-04-18	73086.000	10	1	DE	92353	Postbauer-Heng	11.35720	49.30530	70319.000	2011-04-20	2767.000	364	2775	3
14	P000003	2013-04-19	75941.000	10	1	DE	92353	Postbauer-Heng	11.35720	49.30530	73086.000	2012-04-18	2855.000	366	2847	3
15	P000003	2015-04-30	81411.008	3	3	DE	92353	Postbauer-Heng	11.35720	49.30530	75941.000	2013-04-19	5470.008	741	2694	3
16	P000003	2016-04-27	84220.000	10	1	DE	92353	Postbauer-Heng	11.35720	49.30530	81411.008	2015-04-30	2808.992	363	2824	3
17	P000003	2017-04-13	86817.000	10	1	DE	92353	Postbauer-Heng	11.35720	49.30530	84220.000	2016-04-27	2597.000	351	2701	3
18	P000004	2010-04-01	153490.000	1	1	DE	91619	Obernzenn	10.46670	49.45000	86817.000	2017-04-13	66673.000	-2569	-9473	3
19	P000004	2010-07-15	154757.000	7	1	DE	91619	Obernzenn	10.46670	49.45000	153490.000	2010-04-01	1267.000	105	4404	3
20	P000004	2011-08-10	160290.000	11	2	DE	91619	Obernzenn	10.46670	49.45000	154757.000	2010-07-15	5533.000	391	5165	3
21	P000004	2012-01-31	162640.000	11	2	DE	91619	Obernzenn	10.46670	49.45000	160290.000	2011-08-10	2350.000	174	4930	3

Showing 1 to 21 of 21,872 entries

Figure 1: Sample Dataset

3.2Finding:

Question a

For question a, k-means clustering algorithm was used and clustering was four. It took LONG and LAT parameters for clustering because LONG and LAT are indicated the location of customers. Four clustering made four regions group for all customers. It works like, inputting dataset and give number of four clusters. Then it calculates centroid and also check distance from centroid. It makes group based on minimum distance. This work repeat for four times and four number is final clustering. With that four clustering, is four regional differences. Showing the regional differences, it used ggplot to visualize. Below the ggplot graph is shown:

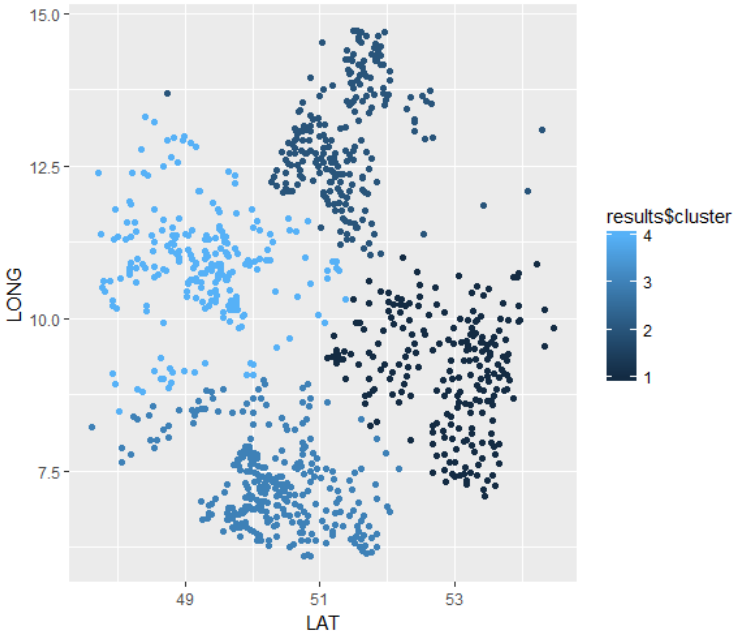


Figure 2: Different regions

In this figure, X axis show LAT, y axis shows LONG and dots are shown as customers. All dots are compared with four deafferents colour and it indicate region differences.

Question b

For comparing the energy consumption in the different regions, in 2015, it used to mutate to add a new column. In this column, four clusters (1, 2, 3, 4) was added according to clustering. Then, reformat the dataset according to date format and separate all 2015-year data. After that, it used boxplot chart to visualize the energy consumption in the different region. Below the boxplot chart is shown,

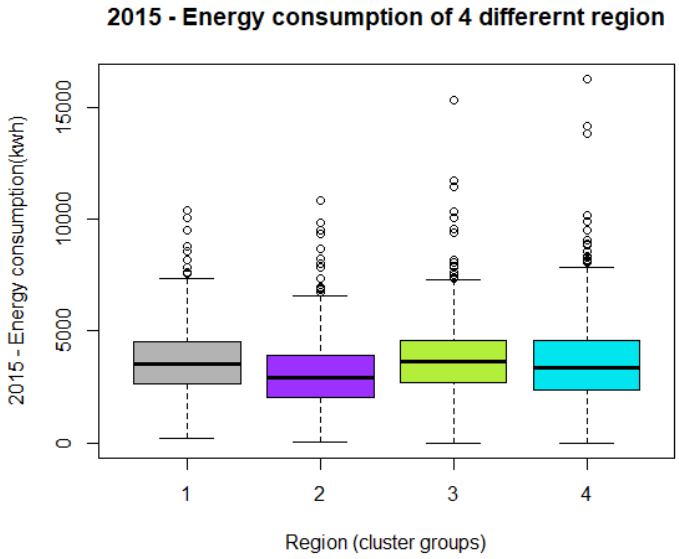
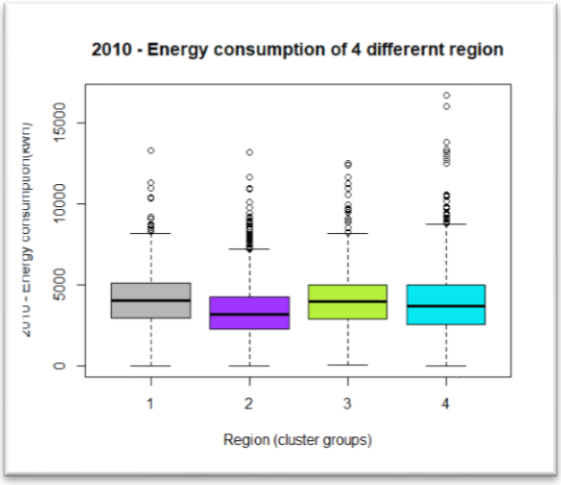


Figure 3: Energy consumption of different regions in 2015

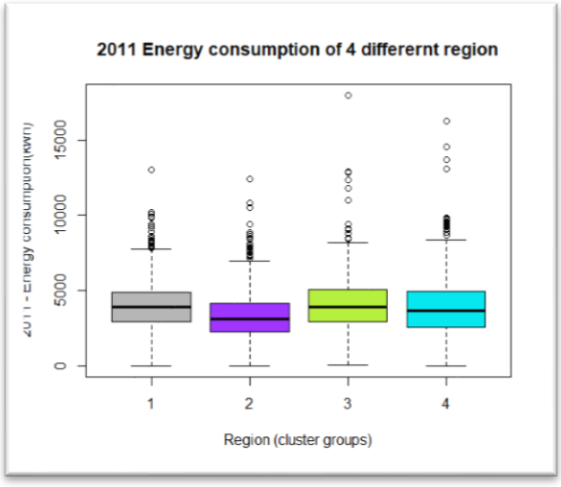
In this boxplot chart, it is clearly shown that the number of one and three regions have almost same consumption than other two regions. Number four region is achieved the second position of consumption where number two region has lowest consumption.

Question c

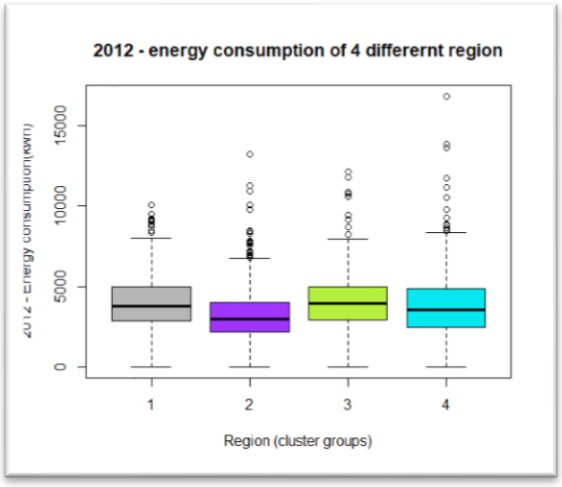
For number question c, the energy consumption of these long-term customers is rising or declining, it checked 2010 to 2017 consumption separately with four different regions. It separated every year from dataset and draw boxplot chart. Below it shown all description



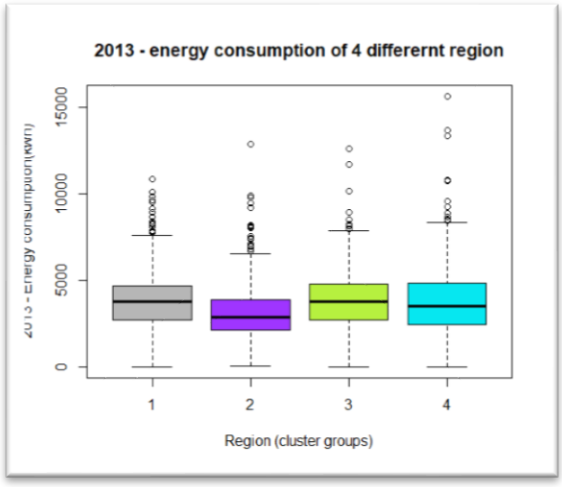
(1)



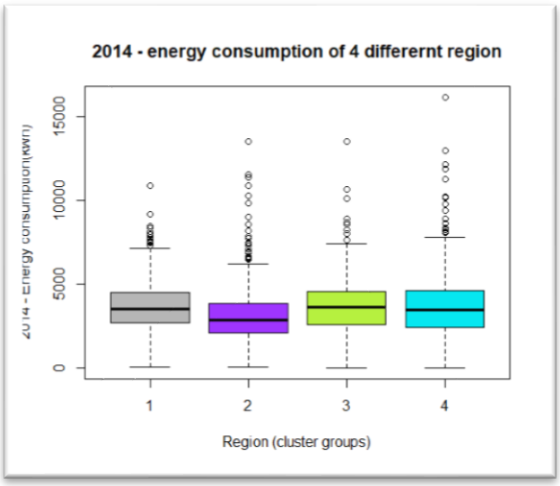
(2)



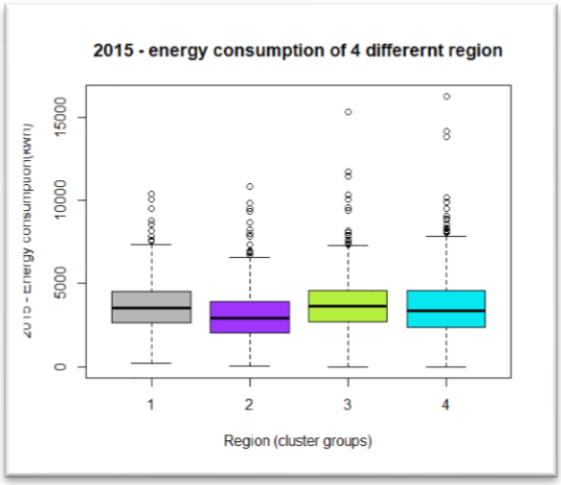
(3)



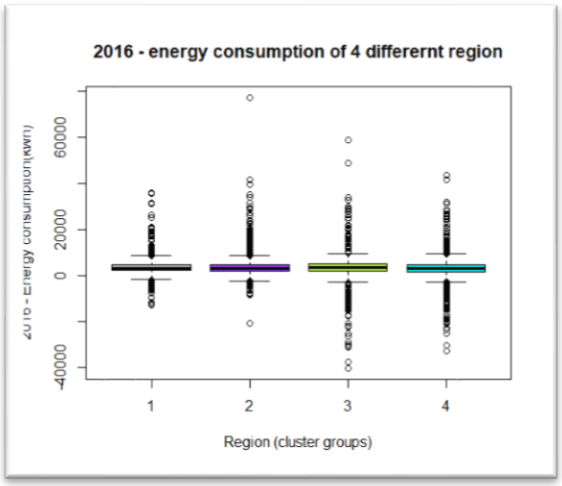
(4)



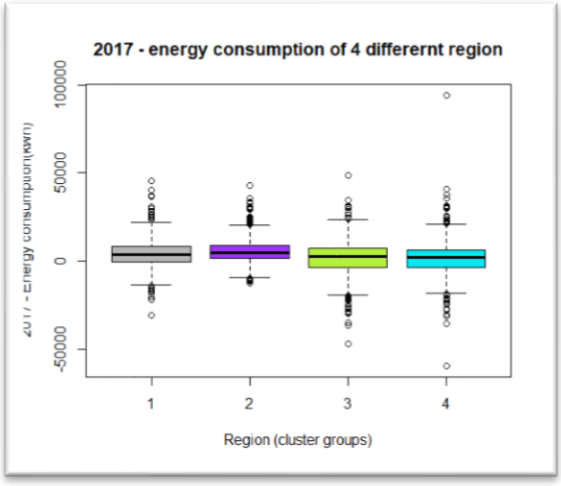
(5)



(6)



(7)



(8)

Figure 4: 2010-2017 years energy consumption of four different region

In 2010 and 2011, energy consumption was slightly decreasing in 2011 in every region but 2012 energy consumption was little increase. Between 2013 to 2016, energy consumption was continuously decreasing. But the interesting fact was that 2016-year energy consumption was decreasing surprisingly huge. In 2017, energy consumption was increasing. Procure department should buy less energy for the next year, because customers consumption is decreasing per year.

Question d:

Codes are separately sent as .text and. r format.

3.Conclusion:

Data mining is an important process to discover knowledge about customer behavior towards business offerings. It explores the unknown credible patterns those are significant for business success. The consumption of customers was decreasing almost every year and regional differences were also found. Procure department needs to buy less advance electricity for their customer.

Reference

1. “A (very) short introduction to R” Paul Torfs & Claudia Brauer Hydrology and Quantitative Water Management Group Wageningen University, The Netherlands. Published: 3 March 2014
2. https://en.wikipedia.org/wiki/K-means_clustering
3. https://en.wikipedia.org/wiki/Cluster_analysis
4. https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html
5. https://en.wikipedia.org/wiki/Box_plot