ANKARA SCIENCE UNIVERSITY, RESEARCHER
Vol. 2, No. 1, July 2022
e-ISSN: 2717-9494
Research Article/Araştırma Makalesi
Doi: https://doi.org/10.55185/researcher.1135824

Big Data Reduction and Visualization Using the K-Means Algorithm

Hakan AKYOL[1], Hale Sema KIZILDUMAN[2], Tansel DÖKEROĞLU[3*]
[1]Çankaya University, Graduate School of Natural and Applied Sciences, Ankara, Türkiye; ORCID: 0000-0002-5695-8790
[2]Çankaya University, Graduate School of Natural and Applied Sciences, Ankara, Türkiye; ORCID: 0000-0002-6449-771X
[3]Çankaya University, Software Engineering Department, Ankara, Türkiye; ORCID: 0000-0003-1665-5928
*Corresponding Author: tdokeroglu@cankaya.edu.tr

Received: 25 June 2022; Accepted: 30 June 2022

Reference/Atıf: H. Akyol, H. S. Kızılduman and T. Dökeroğlu, "Big data reduction and visualization using the K-means algorithm", Researcher, vol. 02, no. 01, pp. 40-45, Jul. 2022, doi:10.55185/researcher.1135824

Abstract

A huge amount of data is being produced every day in our era. In addition to high-performance processing approaches, efficiently visualizing this quantity of data (up to Terabytes) remains a major difficulty. In this study, we use the well-known clustering method *K*-means as a data reduction strategy that keeps the visual quality of the provided huge data as high as possible. The centroids of the dataset are used to display the distribution properties of data in a straightforward manner. Our data comes from a recent Kaggle big data set (Click Through Rate), and it is displayed using Box plots on reduced datasets, compared to the original plots. It is discovered that *K*-means is an effective strategy for reducing the amount of huge data in order to view the original data without sacrificing its distribution information quality.

Keywords: big data, data reduction, visualization, *k*-means

1. Introduction

Data visualization is the way of representing your data using graphical/visual elements to perceive and analyze your data in shorter times and more meaningfully [1]. By utilizing visual components such as charts and graphs, data visualization tools ease to identify and analyze trends, outliers, and patterns in data. However, big data analytics come with new problems and research opportunities for the visualization of the data [2]. Dealing with large volumes of data is far more difficult than dealing with small amounts of data [3]. Enrico and Antonio present a detailed survey about the recent developments and research areas of big data analytics and visualization in their study. Studies in this area have still been continuing [4][5].

In this study, we maintain the visual quality of the box plots (which give information about the distribution of the data) while reducing the size of big data. During this study, we clustered the data with the *K*-means algorithm and used the obtained centroids in our visual elements [6]. Thus, we obtain similar plots with fewer data while keeping the data distribution information of the big data [7].

2. Data Reduction Techniques

This section briefly explains the data reduction techniques we have used in our study.

Randomized Data Reduction: We employ the randomized data reduction approach to compare the performance of the results obtained using the *K*-means algorithm. *n* many data instances are chosen at random from the large data collection, and graphs are drawn using this data. During this procedure, no sampling approach is employed. This serves as a benchmark for evaluating the quality of our *K*-means algorithm outcomes. In order to be fair with our comparisons, we take the same size random values and *K* values.

Data Reduction using the *K*-means clustering algorithm: The algorithm aims to divide data instances into *K* clusters, with each trial belonging to the cluster with the cluster centroid. *K*-means clustering minimizes within-cluster variances. This technique is computationally hard. However, heuristic

algorithms can quickly report near-optimal solutions easily. Therefore, it can be used to select the most representative data instances to give information about the distribution of the big datasets.

3. Experimental setup and evaluation of the results

The datasets we have used in our experiments are Click Through Rate (CTR) from Kaggle [8]. The prediction of advertisement CTR is an important challenge in the field of computational advertising. Increasing the accuracy of advertising CTR prediction is crucial for improving precision marketing efficacy. The dataset discloses large anonymised advertising datasets. There are one million instances in this big dataset.

The visual elements in our study are produced with a PC having an i7 processor, 16 GB RAM, 64-bit operating system, and 8 GB Intel(R) HD Graphics 630 + 4GB NVIDIA GeForce GTX 1050 graphics card. Pycharm IDE is used. The python version is Python 3.9.12. The packages used are pandas, numpy, statsmodels.api, matplotlib.pyplot, seaborn and sklearn.cluster *K*-Means.

In Figure 1, the visualization of the *city* column of the dataset (with 10, 100, 1000, 10000, and original data sizes) are presented. The data is obtained from the first rows of the original dataset. As the selected datasets get bigger, they represent the distribution of the original dataset in a better way. Dataset with 10 instances has the biggest deviation in terms of median values from the original dataset's median. As it can be seen the lowest and highest values of the data are different from the original dataset. In Figure 2, we give the visualization of K-mean results (centroids) with 10, 100, 200, and 500 (this was the biggest *K* value we can get during our experiments) size datasets. The distribution of the datasets is matched. Because the number of instances is very few, all the data cannot be seen in the plot. However, with *K*=500, a plot very similar to the original data is obtained.

Figure 3 gives the distribution of the *city* and *device_size* columns of the dataset (with randomly selected 100, 1000, 10000, and original data sizes). Figure 4 gives the data distribution of the *city* and *device_size* column of the dataset by producing the data with the *K*-means algorithm using *K*=10, *K*=100, *K*=200, and *K*=500. Although the frequency of the data cannot be seen in Figure 4, a better visualization than selecting random slices of data is provided. As the value of K increases, better plots are available. Figures 5 and 6 present the *gender* and *device_size* data visualizations from our dataset in the same way and the reader can easily see the higher quality of the plots with the results of *K*-means. The plot with *K*=500 is almost the same as the original dataset's plot whereas *K*=10 cannot match the upper values of the *Gender* axis. Approximately using 0.0005% of the original dataset, we have drawn clear plots of the big data with the *K*-means algorithm. The execution time of the *K*-Means algorithm is reasonable up to 100 centroids.

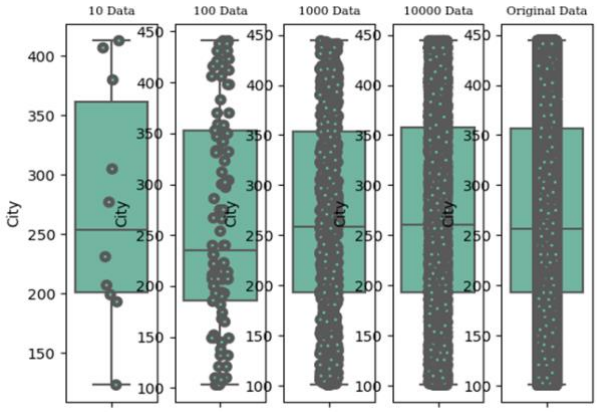

Figure 1: The Distribution Visualization of The *City* Data of The Dataset (with randomly selected 10, 100, 1000, 10000, and original data sizes).

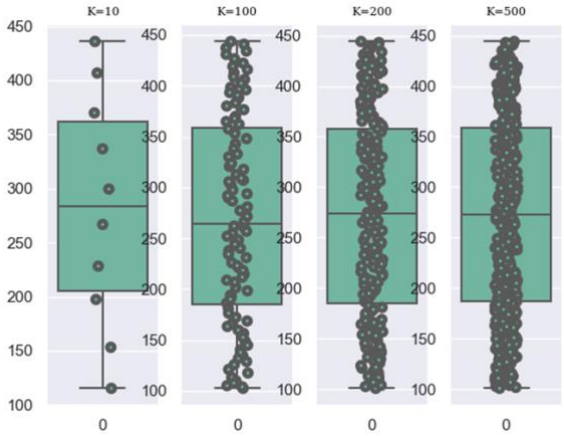

Figure 2: The Distribution of The *City* Data of The Dataset by Producing the Data with the *K*-Means Algorithm. *K*=10, K=100, *K*=200, and *K*=500 are presented in the respective columns.

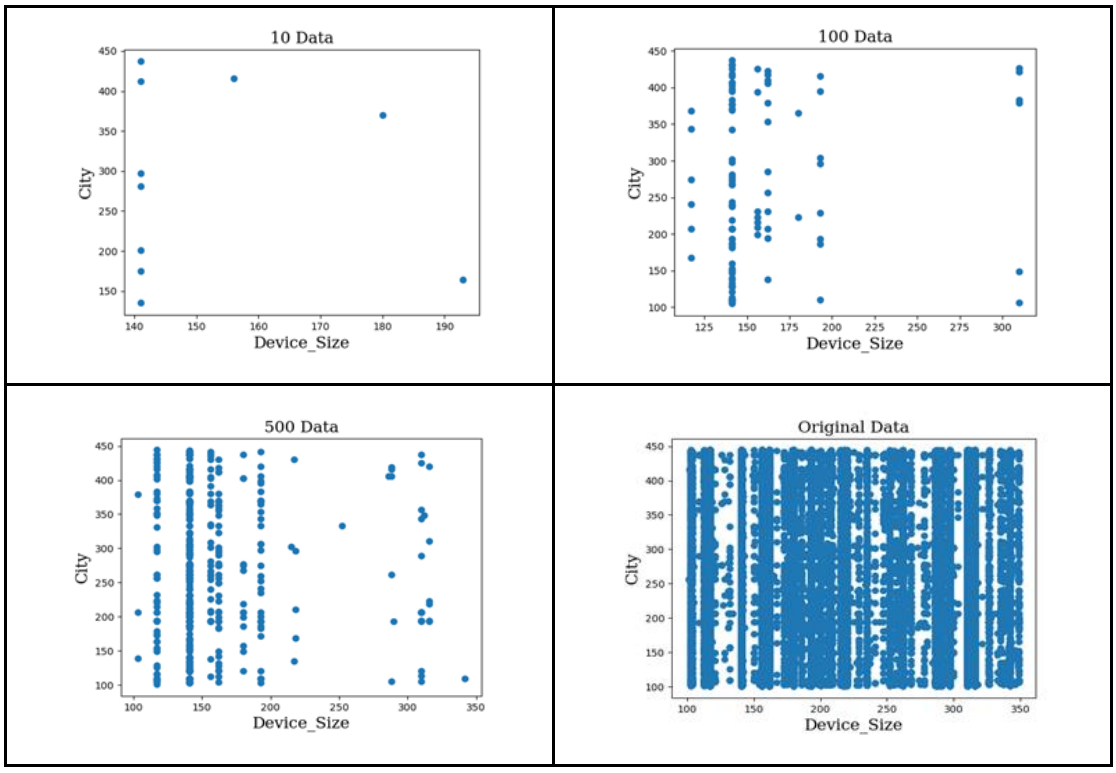

Figure 3: The Data Distribution of The *City* and *Device_Size* Data of The Dataset (with randomly selected 100, 1000, 10000, and original data sizes).

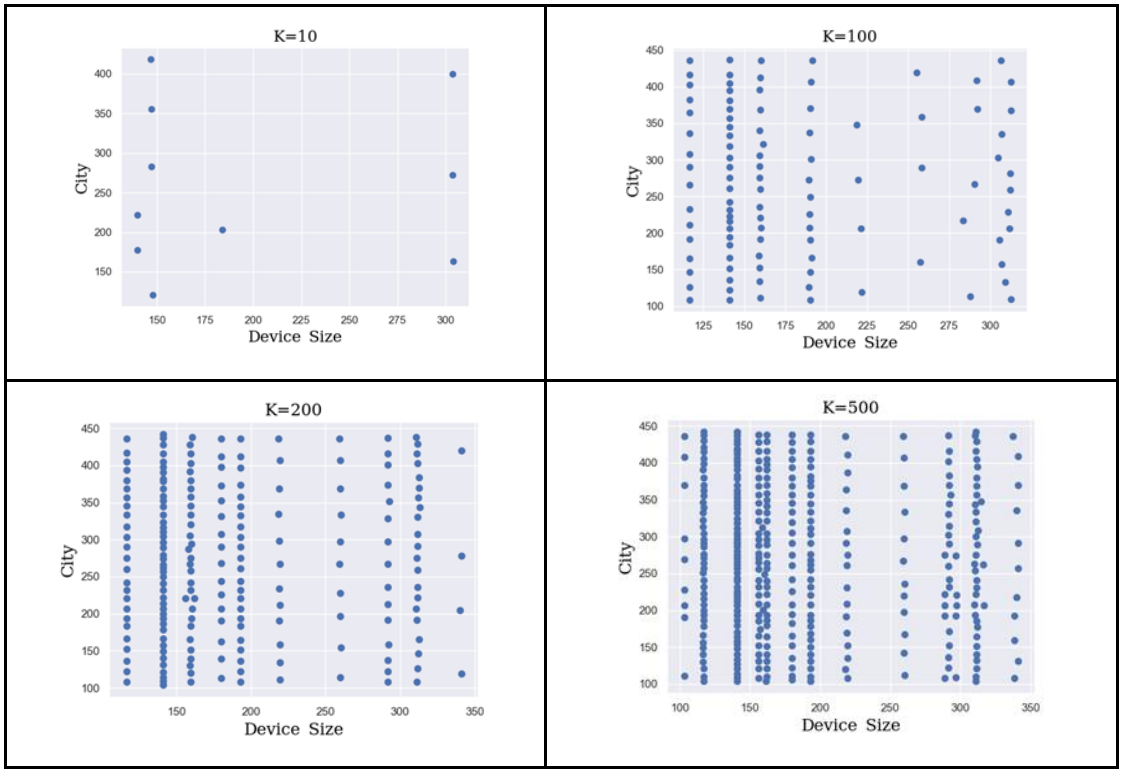

Figure 4: The Data Distribution of The City and *Device_Size* Data of The Dataset by Producing the Data with the *K*-Means Algorithm. *K*=10, *K*=100, *K*=200, and *K*=500 are presented in the respective columns.

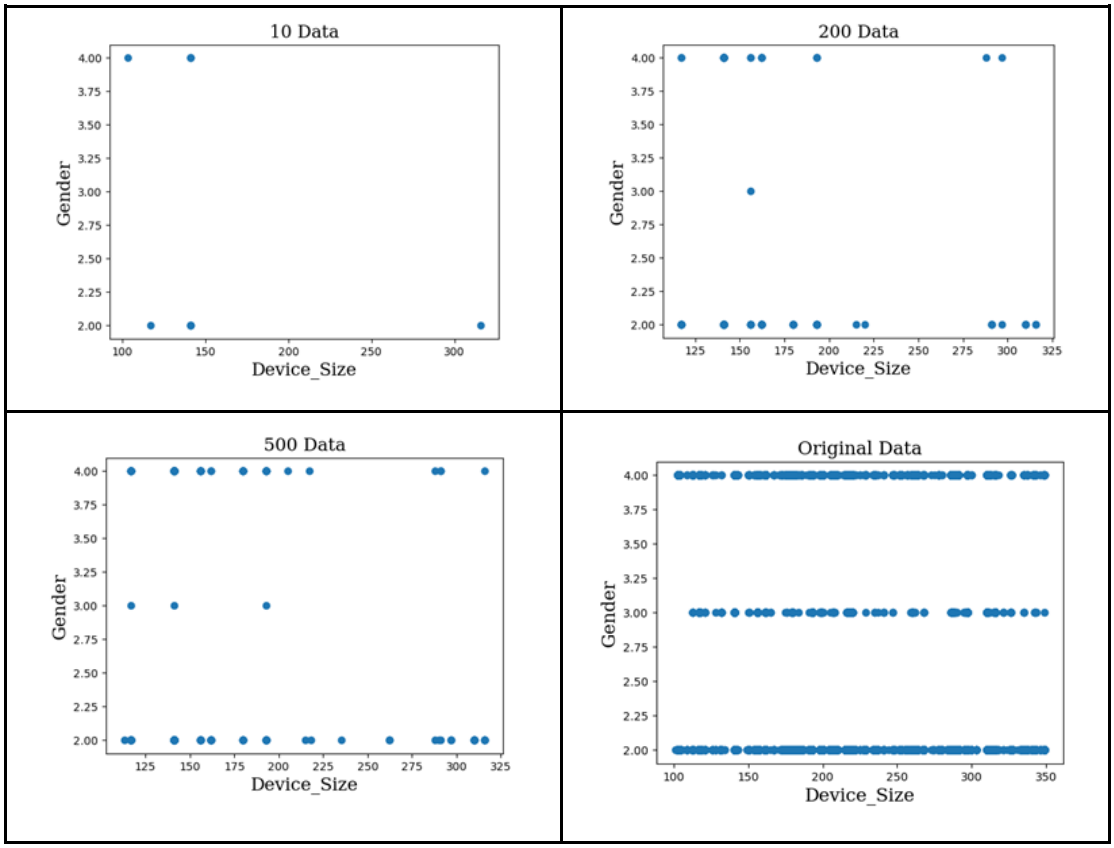

Figure 5: The Data Distribution of The *Gender* and *Device_Size* Data of the Dataset (with randomly selected 10, 200, 500, and original data sizes).

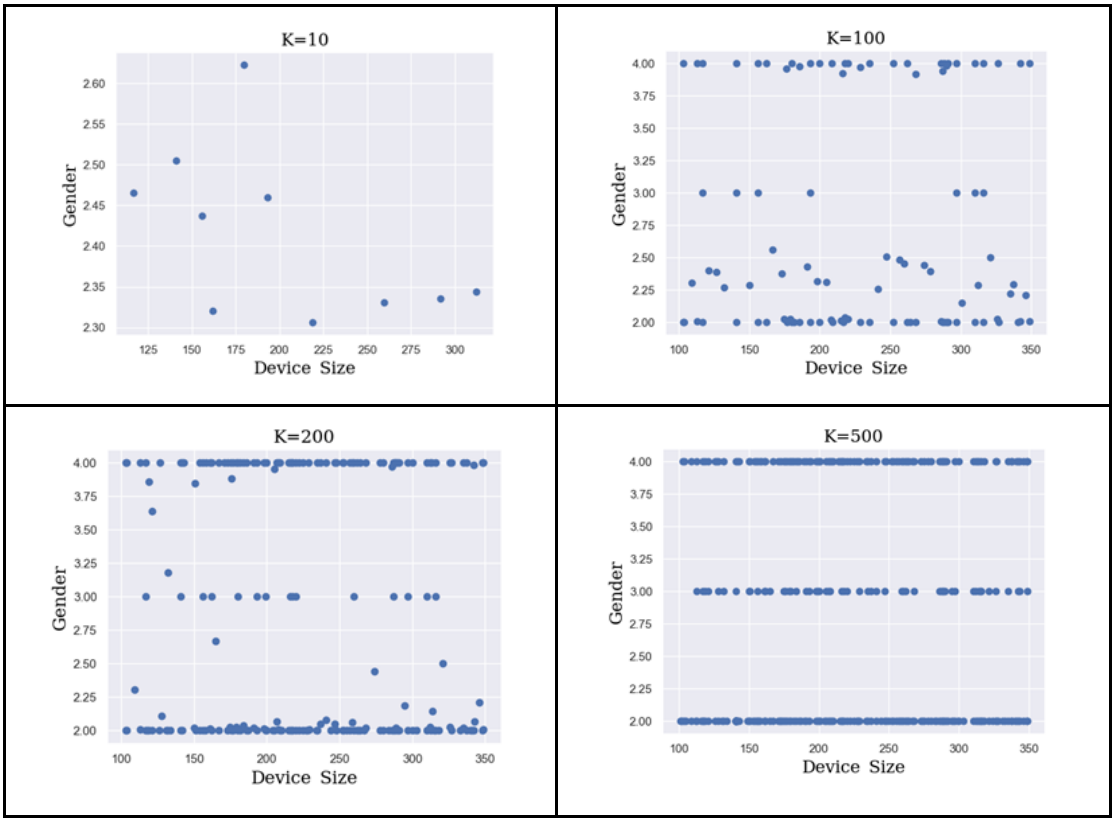

Figure 6: The Data Distribution of The *Gender* and *Device_Size* Data of The Dataset by Producing the Data with the *K*-Means Algorithm. *K*=10, *K*=100, *K*=200, and *K*=500 are presented in the respective columns.

4. Conclusion and future work

Although *K*-means is a clustering algorithm to set the best set of centroids, in this study, it is used as a technique to select/reduce the most indicative data instances so as to visualize big data sets. From the results of our dataset, we have observed the distribution information of the dataset can be kept with a smaller set of data instances obtained as centroids using the *K*-means algorithm. This visualization problem is still a hot topic for researchers. According to the behavior of the datasets, visualization will always be a critical issue for decision-makers. To the best of our knowledge, the method we propose here is the first application of the *K*-means algorithm to the visualization of big data to represent the original data with a reduced set.

In our future work, we intend to study with much bigger datasets and use a big data visualization tool such as Tableau, QlikView, or Microsoft Power BI. We will compare the visual elements of the reduced sets of original big data sets and try to keep the quality and informative features of the data as high as possible.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Friendly, M. (2008). A brief history of data visualization. In Handbook of data visualization (pp. 15-56). Springer, Berlin, Heidelberg.

[2] Keim, D., Qu, H., & Ma, K. L. (2013). Big-data visualization. IEEE Computer Graphics and Applications, 33(4), 20-21.

[3] Andrienko, G., Andrienko, N., Drucker, S., Fekete, J. D., Fisher, D., Idreos, S., ... & Sharaf, M. (2020, March). Big data visualization and analytics: Future research challenges and emerging applications. In BigVis 2020-3rd International Workshop on Big Data Visual Exploration and Analytics.

[4] Agrawal, R., Kadadi, A., Dai, X., & Andres, F. (2015). Challenges and opportunities with big data visualization. In Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems (pp. 169-173).

[5] Ali, S. M., Gupta, N., Nayak, G. K., & Lenka, R. K. (2016). Big data visualization: Tools and challenges. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I) (pp. 656-660). IEEE.

[6] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. Pattern recognition, 36(2), 451-461.

[7] Dokeroglu, T., Deniz, A., & Kiziloz, H. E. (2022). A Comprehensive Survey on Recent Metaheuristics for Feature Selection. Neurocomputing.

[8] Click-Through Rate (CTR), https://www.kaggle.com/datasets/louischen7/2020-digix-advertisement-ctr-prediction, 2022.