ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/medsci.2021-84335

Water Consumption Per Capita and COVID-19 Indices: An Analysis with Machine-Based Learning Methods

Kişi Başına Düşen Su Tüketimi ve COVID-19 İndeksleri: Makine Öğrenimi ile Yapılan Analiz

Ahmet AKGÜL[a,c], Kerem ŞENEL[b], Sinem ERYİĞİT[c], Saygın TÜRKYILMAZ[d], İbrahim SAYIN[d]

[a]Department of Gerontology, İstanbul University-Cerrahpaşa Faculty of Health Sciences, İstanbul, TURKEY
[b]Department of Healthcare Management, İstanbul University-Cerrahpaşa Faculty of Health Sciences, İstanbul, TURKEY
[c]İst-GETAM, Gerontechnology Center, İstanbul, TURKEY
[d]Clinic of Otolaryngology Head and Neck Surgery, Bakırköy Training and Research Hospital, İstanbul, TURKEY

ABSTRACT Objective: To evaluate the relationship between water use of countries and coronavirus disease-2019 (COVID-19) data in order to establish the effect of water dynamics on COVID-19 pandemics. **Material and Methods:** The country-based water consumption per capita (WC) and COVID-19 indices [total cases per 1 million population (CI-C) and deaths per 1 million population (CI-D)] collected from the Worldometer website and the Global Competitiveness Scores from The Global Competitiveness Report 2019 of World Economic Forum which was accessed at the day of May 30th, 2020. The relationship between water consumption and COVID-19 incidences was evaluated with "machine learning" methods. The statistical analyses were performed with the use of R software, version 4.0.0 (R Project for Statistical Computing). **Results:** For a total of 138 countries, we found a positive correlation between WC and CI-C (R 0.13). For the first 20 developed countries, we found a negative correlation between WC and CI-C and CI-D (R -0.18). The USA data is completely different from other 19 analyzed countries. The correlation coefficient becomes -0.44 when the USA is excluded, compared to -0.18 for all 20 countries including the USA. **Conclusion:** A negative relationship between water consumption and COVID-19 incidences was found at least for a relatively homogeneous group which is comprised of mainly developed countries. Among the developed countries, USA exhibits different characteristics. In contrast, when all 138 countries were analyzed, a positive relation was found. The results can be used from various disciplines since the water and its relations with micro-systems serve one of the most important issues for our present time and future.

Keywords: Water; coronavirus; pandemics; socioeconomic growth; machine-based learning

ÖZET Amaç: Bu çalışmanın amacı, ülkelerin su tüketimlerinin koronavirüs hastalığı-2019 [coronavirus disease-2019 (COVID-19)] salgınları üzerindeki etkisini belirlemek için su kullanımı ile COVID-19 verileri arasındaki ilişkiyi değerlendirmektir. **Gereç ve Yöntemler:** Worldometer web sitesinden, ülkelerdeki kişi başına düşen su tüketimi [water consumption (WC)] ve COVID-19 indeksleri [1 milyon nüfus başına toplam vaka (COVID-19 indices-Case "Cl-C") ve 1 milyon nüfus başına ölüm (COVID-19 indices-Death "CI-D")] elde edildi. Dünya Ekonomik Forumu'nun 2019 Küresel Rekabet Edebilirlik Raporu'ndan Rekabet Edebilirlik Puanları ile ülkelerin gelişmişlik durumları 30 Mayıs 2020 tarihli erişim sonucuna göre sıralandı. Su tüketimi ile COVID-19 vakaları arasındaki ilişki makine öğrenimi yöntemleriyle değerlendirildi. İstatistik analiz için R Software, version 4.0.0 (R Project for Statistical Computing) kullanıldı. **Bulgular:** Toplam 138 ülke için WC ve CI-C (R 0.13) arasında pozitif bir korelasyon bulundu. İlk 20 gelişmiş ülke için, WC ile CI-C ve CI-D (R -0.18) arasında negatif bir korelasyon bulundu. ABD verileri, incelenen diğer 19 ülkeden tamamen farklı idi. Korelasyon katsayısı, ABD hariç tutulduğunda -0.44 olurken, ABD dâhil edildiğinde ise 20 ülke için -0.18 olarak bulundu. **Sonuç:** Gelişmiş ülkeler içinde -ABD farklı özellikler sergilemektedir- homojen bir şekilde su tüketimi ile COVID-19 vakaları arasında negatif bir ilişki vardır. Buna rağmen 138 ülkenin tamamı incelendiğinde su tüketimi ile COVID-19 vakaları arasındaki ilişki pozitiftir. Su ve mikro çevre ilişkileri günümüz ve geleceğimiz için en önemli konulardan biri olarak birçok farklı disiplin içinde kullanılmaya ve araştırılmaya değerdir.

Anahtar Kelimeler: Su; koronavirüs; pandemiler; sosyoekonomik büyüme; makine öğrenimi metodu

Water in terms of quality and quantity previously demonstrated to be the cause of endemic diseases.[1] For the current coronavirus disease-2019 (COVID-19) pandemic, access to water is an essential issue in order to provide hygiene and sanitation.[2] The COVID-19 pandemic may be considered as one of

Correspondence: Ahmet AKGÜL
Department of Gerontology, İstanbul University-Cerrahpaşa Faculty of Health Sciences, İstanbul, TURKEY/TÜRKİYE
E-mail: ahmet.akgul@istanbulc.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Medical Sciences.

Received: 07 May 2021 ***Accepted:*** 03 Jun 2021 ***Available online:*** 11 Jun 2021

the biggest exemplifying experiences due to testing the ability of relatively unknown microorganisms and their multidimensional interaction with our planets macromolecular system. The interaction of a novel virus with human cells as well as the whole body itself, organization of health services, consumption of natural resources like water during this pandemic give valuable data regardless of time for humanities future. In order to shed light on the subject matter, we planned to investigate the relationship between the water use of countries and COVID-19 data to establish the effect of water dynamics on COVID-19 indices.

MATERIAL AND METHODS

DATA SOURCES

We obtained the country-based water consumption and COVID-19 data from the Worldometer website, and the global competitiveness scores (GCS) from "The Global Competitiveness Report 2019" of World Economic Forum.[3,4] Countries which have a data in both data sources were included for the analysis. Countries with missing data in website or countries which were not ranked in The Global Competitiveness Report 2019 were excluded.

VARIABLES ASSESSED

From the Worldometer website, we obtained the following data elements for each country on the day of May 30th, 2020: daily water consumption per capita (liters) (WC); total cases per 1 million population (COVID-19 indices-Case, CI-C); and deaths per 1 million population (COVID-19 indices-Death, CI-D). The ranking of countries was done with the GCS. The overall performance of a country is reported as progress scores on a 0 to 100 scale. A score of 100 represents the "frontier", i.e. an ideal state.

STATISTICAL ANALYSIS

The statistical analyses were performed with the use of R software, version 4.0.0 (R Project for Statistical Computing). We calculated Pearson correlation coefficients as a preliminary analysis to explore if there is any relationship between WC and COVID-19 incidences. We used the GCS to obtain a homogeneous group of countries that are presumed to suffer less from data collection problems.

The WC, CI-C, and CI-D were scaled and standardized to have zero mean and unit standard deviation. First, a distance matrix was generated to provide a visual depiction of the similarities and dissimilarities between countries. Then, the results of the K-means cluster analysis were presented as a two-dimensional plot of the first two principal components and pairwise charting of the original variables; i.e., WC vs. CI-C and CI-D, respectively.

K-means cluster analysis was employed to group similar countries together in terms of WC, CI-C, and CI-D. Hartigan-Wong algorithm, which defines the total within-cluster variation as the sum of squared Euclidean distances between items and the corresponding centroid, has been used.[5] The optimal number of clusters was obtained by using the average silhouette method.

In order to perform a formal cluster analysis to explore if there is any significant pattern among this group of countries, we have used k-means cluster analysis. K-means is a popular unsupervised machine learning algorithm in order to group similar data points into clusters and discover underlying patterns.[6] The algorithm pinpoints centroids, i.e., the imaginary or real locations representing the centers of the clusters. Then, every data point is allocated to the nearest cluster. There are various distance metrics that can be used with the algorithm, such as the Euclidean, Manhattan, Chebyshev, and Minkowski distances. The Euclidean distance is usually the default and the most common distance metric. To determine the optimal number of clusters, various methods, such as the elbow method and the average silhouette method, can be used.[7] We prefer to use the average silhouette method since it provides an objective estimate for determining the optimal number of clusters. The average silhouette approach measures the quality of a clustering by determining how well each lies within its cluster object. Higher average silhouette widths correspond to better clustering. The method computes the average silhouette of observations for different values of k. The optimal number of clusters is the value of k that maximizes the average silhouette.

RESULTS

In total 138 countries have comparable data from both data sources and analyzed. In order to understand if there is a relationship between WC and COVID-19 incidences, first, Pearson correlation coefficients have been calculated for all the countries in our sample, i.e. a total of 138 countries. The correlation coefficient between WC and CI-C is calculated as 0.13. When we substitute CI-C with CI-D, the correlation coefficient goes down to 0.02. These figures point to a low positive correlation between water consumption and COVID-19 incidences for evaluated 138 countries.

Considering that the reported figures are particularly problematic for developing and underdeveloped countries, we have decided to limit our study to a relatively more homogeneous group, i.e. the top 20 countries in terms of the GCS. For these countries, the correlation coefficients between WC and CI-C/CI-D were calculated as -0.18. The USA dynamics were completely different from other 19 developed countries. When the USA was excluded from the first 20 countries analysis; correlation coefficient becomes more evident and goes to -0.44. This suggests a negative relationship between water consumption and COVID-19 incidences, at least for a relatively homogeneous group which is comprised of mainly developed countries.

Figure 1 shows a graphical representation of the distance matrix of countries calculated from WC, CI-C, and CI-D. If the color of a box is green (smaller distance), it means that the corresponding two countries are similar. A red box, on the other hand, is an indication of greater distance and dissimilarity. The USA stands as an outlier since the distance between the USA and any another country on the list is always greater than the overall average (red boxes). The Netherlands seems to be another outlier, having only slight resemblances to France and Germany.

The average silhouette method shows that using 8 clusters is the optimal choice (Figure 2). Since we have three variables, representation with a single two-dimensional plot is only possible if we use principal components. In total 82.3% of the total variation in

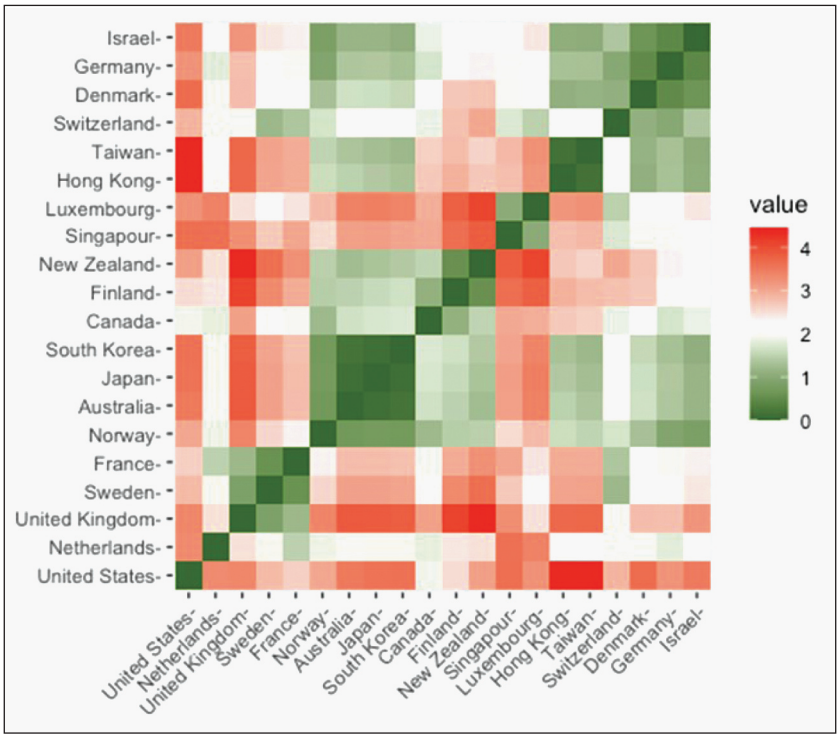

FIGURE 1: Distance matrix calculated from country-based water consumption per capita, COVID-19 total cases per 1 million population and COVID-19 total deaths per 1 million population.

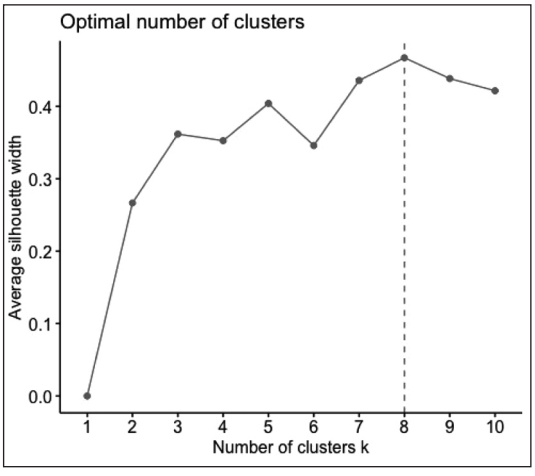

FIGURE 2: The results of the average silhouette method for k-means clustering of the countries in terms of daily water consumption per capita, total cases per 1 million, and deaths per 1 million (average silhouette width for determining optimal number of clusters).

variables can be explained by the first two principal components. Another observation is that the USA and the Netherlands comprise clusters on their own (Figure 3). This is in line with our visual observation from the distance matrix in Figure 1. For further analysis, we need plots between original variables to interpret the results. Figures 4 and 5 depict WC versus CI-C and CI-D, respectively.

The USA comprises a cluster on its own. Figure 4 and 5 shows that the highest WC is in the USA. Despite its high WC, the USA also ranks high in terms of CI-C. This is contrary to our previous finding, i.e. the slightly negative correlation observed between WC and CI-C. This result suggests that the aforementioned correlation is probably stronger if the USA is excluded from the sample. Indeed, the correlation coefficient becomes -0.44 when the USA is excluded, compared to -0.18 for all 20 countries including the USA.

Figure 5 shows that CI-D is also high in the USA, but not as high as CI-C. In fact, CI-C and CI-D in the USA are roughly 1.5 and 1 standard deviations above the mean, respectively. Canada, Finland, and New Zealand comprise another cluster. The above average WC in these countries is accompanied by below average CI-C and CI-D. Norway, Australia, South Korea and Japan are clustered together. These countries have slightly above average WC whereas their CI-C and CI-D are below average. Another country that comprises a cluster on its own is the Netherlands. WC for this country is around the mean. On the other hand, there is an interesting result when the other variables pertaining to COVID-19 are ex-

FIGURE 3: K-means cluster plot with two principal components.

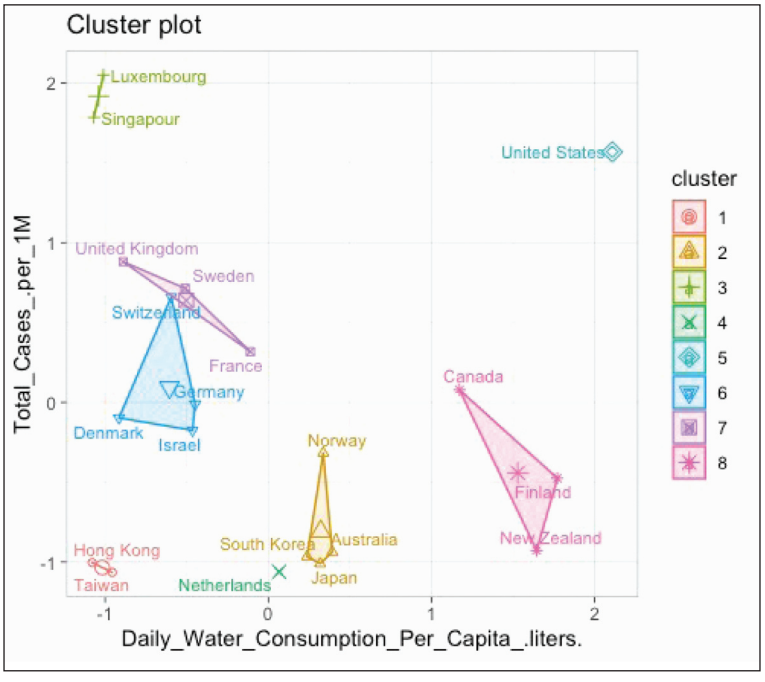

FIGURE 4: The country-based water consumption per capita (liters) and COVID-19 total cases per 1 million population comparison.

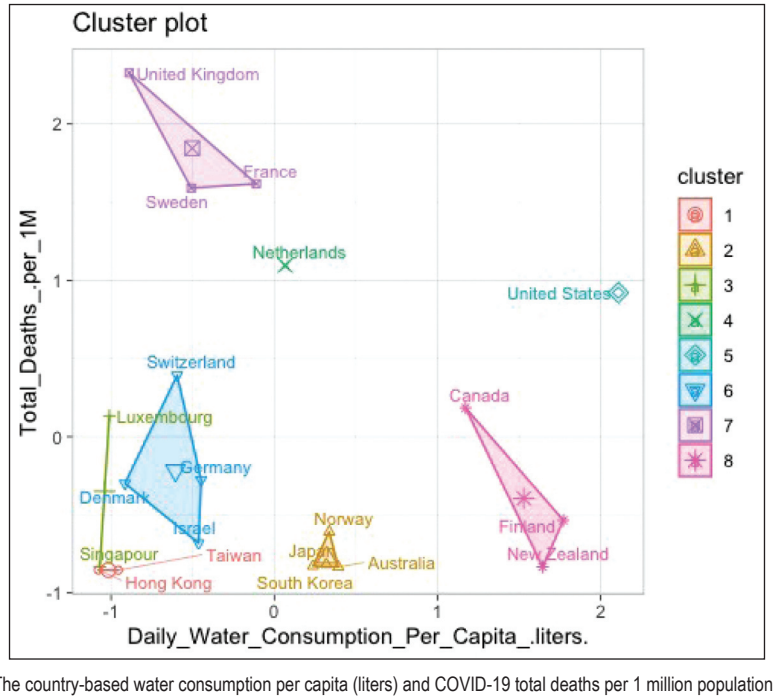

FIGURE 5: The country-based water consumption per capita (liters) and COVID-19 total deaths per 1 million population comparison.

amined. While CI-C is roughly 1 standard deviation below the mean, CI-D, on the other hand, is roughly 1 standard deviation above the mean. The relatively high death toll in the Netherlands suggested by our results has already made the headlines.[8]

Taiwan and Hong Kong comprise another cluster. Despite their below average WC, their CI-C and CI-D are also below average. Although the measures implemented against COVID-19 in these countries can be deemed to be rather intrusive and somewhat

excessive when compared to Western countries, these measures seem to have paid off.[9,10]

Singapore and Luxembourg are clustered together. Although these countries rank very high in terms of global competitiveness score, their populations, particularly the population of Luxembourg, are much lower compared to other countries. Therefore, their results may not be meaningful for comparison. Germany is clustered with Denmark, Switzerland, and Israel. The WC in these countries is slightly below the mean, whereas their CI-C and CI-D are around the mean levels. Finally, the United Kingdom, France, and Sweden are clustered together with their below average WC and above average total cases and CI-D.

DISCUSSION

For a total of 138 countries, we found a positive correlation between WC and CI-C (R 0.13). For the first 20 developed countries, we found a negative correlation between WC and CI-C and CI-D (R -0.18). The USA data is completely different from other 19 analyzed counties. The correlation coefficient becomes -0.44 when the USA is excluded, compared to -0.18 for all 20 countries including the USA.

Water is the single most important requirement for developing advanced civilizations. It's impact on life is beyond being a source, instead it lies behind the governance of systems and serves as a key between the community and the ecosystem. Among all multidimensional impact on life two points are the most important. The intake of water for bodily composition and health and the need of water for sanitation. The water use per capita data may be a sign for availability of both which will be basis for how nations provide and use water. The effect on community's water consumption on their susceptibility to infectious diseases is a hardly complicated point for evaluation. However, COVID-19 pandemic may be a model to work on. To understand the interaction between water consumption and infectious diseases, we designed this study which is resulted in interesting findings.

The data for the opposite correlation between water consumption for the first 20 developed countries needs an emphasis. The water that used in developed countries were mostly for industrial and settlement purposes. Besides most of these countries have water regulation strategies and attempts were given to limit the use of water and protect the resources. However according to our correlation limiting water use increases the pandemic effected case number which will mean that an undefined minimum level of water consumption per capita needed for preventing societies immunity. This means that consuming less water always not necessarily to mean a healthy community. This observation is also supported from the USA water consumption data. The USA is the different one and USA itself is clustered in a different cluster from machine based learning system. The USA water consumption data is high; like its data on COVID-19 related disease and deaths. In opposite saying using high amounts of water does not mean that it is good for community's health and the high consumption does not necessarily mean better immunity and better sanitation. As outlined previously a minimum set of water consumption data may require.

When the GCS decrease the correlation closed to 0 point and in the result of 138 counties the correlation becomes different as COVID-19 seems to be increased with consumption. Our explanation for this finding is different since the global competitiveness score decrease the water sanitation and supply issues within the countries started to decrease and the reporting of infected cases from health systems and governments more likely to be under the desired levels. This means that we need to improve the acquisition of big data from all over to world and has a long way for a uniformed world that we can offer all world citizens similar services and similar conditions. In other way of view; the first 20 developed countries and other 127 countries have completely different characteristics which means that they cannot be put in the same box and compared.

On the other hand, the data for COVID-19 is imperfect and incomplete, particularly for developing and underdeveloped countries. Most developing

countries suffer from an acute lack of COVID-19 testing capacity, and they either collect low-quality data or do not record deaths at all (Figure 1).[11] Furthermore, death tolls are occasionally revised in many countries, which dispute the reliability of reported figures.[12,13]

It is obvious that the susceptibility to pandemic cannot be explained from single data. However this view can be a basis for evaluating complex mechanisms of our evolution. There are lot of factors related to susceptibility to infection starting from genetic predisposition to local regulation rules.

Although these results point to a negative correlation between the variables, there may also be other factors in play. Initially, Britain signaled that it would pursue a mitigation approach to allow the virus to circulate more widely and begin to create "herd immunity." Later, a mathematical simulation performed by the epidemiologist Neil Ferguson and his COVID-19 Response Team at Imperial College London, the government ultimately went with a full lockdown, an all-in attempt to suppress the spread of the virus as much as possible.[14] Sweden, on the other hand, maintained polices geared toward attaining herd immunity. Sweden's policy and the ambivalent stance of the British government, particularly at the onset of the pandemic, might have exacerbated the frequency of COVID-19 incidences.

The limitation is although there may be a relationship between water consumption and COVID-19 incidences and deaths, one should not overlook the other factors that may be in play. The level of restrictions and preventive measures, population density, and demographic factors may all have an impact on the spread and mortality of the disease. It should also be noted that the WC may not be an accurate indicator in terms of bodily water intake and/or personal hygiene. Industrial use of water, cultural differences pertaining to showering habits, and the availability of sufficient water and its impact on water usage and saving practices may all contribute to the discrepancies among usage statistics for different countries.

CONCLUSION

Our results indicate a direct relationship between water consumption and COVID-19 indices for the analyzed 138 countries. When the first 20 developed countries are taken into consideration, the relation becomes inverse. The USA acts differently from the first 19 developed countries for water consumption and COVID-19 indices data.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

All authors contributed equally while this study preparing.

REFERENCES

1. Armitage R, Nellums LB. Water, climate change, and COVID-19: prioritising those in water-stressed settings. Lancet Planet Health. 2020;4(5):e175. [Crossref] [PubMed] [PMC]
2. Akanda AS, Johnson K. Growing water insecurity and dengue burden in the Americas. Lancet Planet Health. 2018;2(5):e190-e191. [Crossref] [PubMed]
3. WorldoMeter [İnternet]. © Copyright Worldometers [Erişim tarihi: 30 Mayıs 2020]. Coronavirus Reported Cases and Deaths by Country, Territory, or Conveyance. Erişim linki: [Link]
4. Schwab K. The global competitiveness report. 2019;21. [Cited: 2 June 2020]. Available from: [Link]
5. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. J R Stat Soc Ser C. Applied Stat. 1979;28(1):100-8. [Crossref]
6. Towards Data Science [İnternet]. [Erişim tarihi: 31 Mayıs 2020]. Understanding K-means Clustering in Machine Learning. Erişim linki: [Link]
7. UC Business Analytics R Programming Guide [İnternet]. [Erişim tarihi: 7 Haziran 2020]. K-means Cluster Analysis. Erişim linki: [Link]
8. The Guardian [İnternet]. © 2020 Guardian News & Media Limited or its affiliated companies [Erişim tarihi: 8 Haziran 2020]. Could the "liberal" Dutch have learned from Taiwan's approach to coronavirus? Erişim linki: [Link]
9. The Guardian [İnternet]. © 2020 Guardian News & Media Limited or its affiliated companies [Erişim tarihi: 8 Haziran 2020]. Test and trace: lessons from Hong Kong on avoiding a coronavirus lockdown. Erişim linki: [Link]
10. TReNDS [İnternet]. [Erişim tarihi: 20 Mayıs 2020]. In Low-Income Countries Fundamental Data Issues Remain for COVID-19 Response. Erişim linki: [Link]
11. Live Science [İnternet]. [Erişim tarihi: 2 Haziran 2020]. China increases Wuhan's COVID-19 death toll by 50%. Erişim linki: [Link]
12. Npr [İnternet]. © 2020 npr [Erişim tarihi: 2 Haziran 2020]. Moscow Doubles Last Month's Coronavirus Death Toll Amid Suspicions of Undercounting. Erişim linki: [Link]
13. Reuters [İnternet]. © 2020 Reuters [Erişim tarihi: 2 Haziran 2020]. Spain revises coronavirus death toll down by nearly 2,000. Erişim linki: [Link]
14. The Washington Post [İnternet]. [Erişim tarihi: 8 Haziran 2020]. A tale of two epidemics: Scientists in Sweden and Britain fight over who took the right public health path. Erişim linki: [Link]