Methods of data analysis to study the effectiveness of scientific journal promotion

Olha V. Korotun[1], Tetiana A. Vakaliuk[1,2,3,4], Tetiana M. Nikitchuk[1] and Mariia O. Korotun[1]

[1]*Zhytomyr Polytechnic State University, 103 Chudnivska Str., Zhytomyr, 10005, Ukraine*
[2]*Kryvyi Rih State Pedagogical University, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*
[3]*Institute for Digitalisation of Education of the NAES of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine*
[4]*Academy of Cognitive and Natural Sciences, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*

Abstract

The article is devoted to studying the effectiveness of promoting the scientific journal "Journal of Edge Computing" among the scientific community. The article uses statistical methods and machine learning to analyse data collected during the distribution of invitations to review the journal. The purpose of the study was to conduct a comprehensive analysis of the data to determine the effectiveness of sending letters to foreign and domestic researchers with an invitation to view the journal's page, to interest them in the list of research areas of the journal and to invite them to publish in it in the future. The article describes in detail the stages of the analysis, from data collection and cleaning to model building and interpretation of the results. The results allowed us to identify the countries whose researchers are most interested in the journal to focus on them in the future. The study results may be useful for other scientific journals seeking to expand the geography of countries and attract new authors to publish in their scientific journals.

Keywords

edge computing, Journal of Edge Computing, data analysis, R programming language

1. Introduction

In today's world of technology, a new paradigm of edge computing is becoming increasingly widespread, aimed at moving computing resources closer to the data source. This is due to the speed of data processing, reduced network load, continuity of data processing and efficient use of computing resources. This trend helps to open up new opportunities in various areas of human life. The purpose of this study is to conduct a comprehensive analysis of the collected data from the scientific journal "Journal of Edge Computing" [1] to identify critical trends for reaching and attracting foreign and domestic researchers for further visits and publication in the journal and to outline the prospects for research in this area.

Since this journal was established relatively recently, the authors of the article had the idea to analyse the letters of invitation sent to scholars to visit and view the website of this journal (https://acnsci.org/jec) and the feedback received from them. Such an analysis would allow for the collection and interpretation of the information received about the interest of scholars. It is essential to understand the breadth of the audience, namely who viewed the journal, how many of them were from which countries, and the reasons for this, as well as the difficulties that could arise in this regard. This will help to make the journal's content more relevant and exciting for many scholars and encourage them to publish in the journal.

CS&SE@SW 2024: 7th Workshop for Young Scientists in Computer Science & Software Engineering, December 27, 2024, Kryvyi Rih, Ukraine

korotun-o@ztu.edu.ua (O. V. Korotun); tetianavakaliuk@gmail.com (T. A. Vakaliuk); tnikitchuk@ukr.net (T. M. Nikitchuk)
https://acnsci.org/vakaliuk/ (T. A. Vakaliuk)
0000-0003-2240-7891 (O. V. Korotun); 0000-0001-6825-4697 (T. A. Vakaliuk); 0000-0002-9068-931X (T. M. Nikitchuk); 0009-0000-3780-0421 (M. O. Korotun)

2. Literature review

If we consider studies using data analysis, there is a relatively large number of them. In [2], based on the qualitative data analysis of the following stages: data collection, data cleaning, data presentation and results determination, as well as quantitative data analysis using correlation and regression, a study on the use of social science textbooks with augmented reality (AR) based on Android in secondary school is presented. The authors note that such technology will significantly improve the educational process and make it more interesting, practical and interactive.

Katahira [3] highlight the widespread problem of data heterogeneity in various scientific fields based on their clustering. They describe it as the process of dividing data into several groups (clusters) so that objects in one cluster are more similar to each other than objects from different clusters. As a result, we can reveal hidden structures in the data and better understand the reasons for differences in observations.

The R/LinekdCharts tool for simplifying the analysis of large amounts of data is described in [4]. Its advantages include interactive visualisation (creation of interconnected charts), efficiency (creating visualisations with minimal code), and flexibility (overview charts, detailed graphs). Data analysis using the R programming language and many practical examples are presented in [5]. RStudio is used as an integrated development environment (IDE). Davis [6] also describes data analysis methods using the R programming language. The authors cover exploratory data analysis, spatial data analysis, statistics and modelling, and effective presentation of results.

3. Methodology

The preparatory stage of analysing a particular data set involves finding the main metrics of descriptive statistics, namely calculating the median and arithmetic mean, finding quartiles and interquartile ranges, and calculating covariance and correlation coefficients.

At the main stage of the study, we will use data mining methods and machine learning algorithms to build a regression model and cluster the data. Regression is supervised learning, while clustering represents unsupervised learning of the built models.

A regression model allows you to predict the relationship between the dependent variable and the independent variable. This can be done by building a basic regression model, namely a linear model, according to the following formula (1).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{1}$$

where y is the dependent variable, $(x_1, x_2, ..., x_k)$ are independent variables, and u is the random deviation.

This model will make it possible to predict the number of visits (dependent variable – visitors) based on the emails sent out (independent variable – mailings).

To assess the quality of the model, we use the coefficient of determination (R^2). This statistical measure will allow us to understand how well the model fits the data set under study. It is calculated using the following formula (2):

$$R^2 = \frac{\sum\limits_{i=1}^{n} (Y_{\text{pr}} - Y_{\text{av}})^2}{\sum\limits_{i=1}^{n} (y_{\text{fact}} - Y_{\text{av}})^2} \tag{2}$$

where Y_{pr} is the predicted value of the dependent variable; Y_{av} is the average value of the dependent variable; Y_{fact} is the actual value of the dependent variable. Measured from 0 to 1, the closer to 1, the better the model is built. In addition to the coefficient of determination, there are other metrics for evaluating models, such as AIC, BIC, RMSE.

Researchers use various clustering methods to divide data into several groups. Let us consider the main ones and describe their features: k-means method, in which each object in the dataset belongs

to only one cluster, the number of clusters must be determined in advance, for example, this can be done using the elbow method; hierarchical clustering allows you to build a hierarchical structure of clusters, the number of clusters can be determined by the built dendrogram; DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method builds clusters based on data density, separating areas of clusters with low data density, after building a visualisation of the clustering performed by this method, we can determine the number of clusters formed in the existing data set.

For the present study, we chose the k-means method, which allows us to divide the available data into several clusters based on their similarity. This method is based on minimising the sum of the squared distances between the data and the found cluster centre (3).

$$\sum_{i=1}^{n} d\left(x_i, m_j\left(x_i\right)\right)^2 \tag{3}$$

where d is the metric, x_i is the i-th data object, $m_j(x_i)$ is the centre of the cluster to which x_i is assigned at the j-th iteration. Let us present the algorithm of the iterative k-means clustering method (algorithm 1).

Algorithm 1 K-Algorithm for clustering data using the k-means method.

Require: Data points $X = \{x_1, \ldots, x_n\}$, number of clusters k
Ensure: Cluster assignments and centroids
Initialize k centroids $\{\mu_1, \ldots, \mu_k\}$ randomly from the data points
while centroids have changed **do**
  // Assign points to the nearest centroid
  for each point $x_i \in X$ **do**
    $cluster[i] \leftarrow \arg\min_j ||x_i - \mu_j||^2$ {Assign x_i to the closest centroid}
  end for
  // Update centroids
  for $j = 1$ to k **do**
    $C_j \leftarrow \{x_i \in X : cluster[i] = j\}$ {Points assigned to cluster j}
    $\mu_j \leftarrow \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$ {Recalculate centroid of cluster j}
  end for
end while
return Cluster assignments and centroids

The described mathematical apparatus is available in the R language, as there are specially built-in functions that will facilitate the implementation of this study and allow us to draw certain conclusions.

4. Results

To analyse the generated data set, we will apply the following generally accepted scheme consisting of the following steps (figure 1).

According to the above scheme, let's describe the work that needs to be done at each stage of data analysis.

1. Data collection includes identifying the purpose, sources, and the actual data collection process. To obtain the data, we selected scientific articles from the open collection of published scientific research ScienceDirect in similar areas of the JEC journal and collected email addresses of scientists from around the world, namely, their names, article titles, and countries, which allowed us to form a data table (figure 2). This process was carried out daily during June 2024 and added an average of about 60-70 records.
2. Data cleaning involves detecting and eliminating errors in the data and processing missing values. If necessary, data standardisation can be performed. For the obtained data set, the analysis was carried out in the RStudio environment using the R programming language.

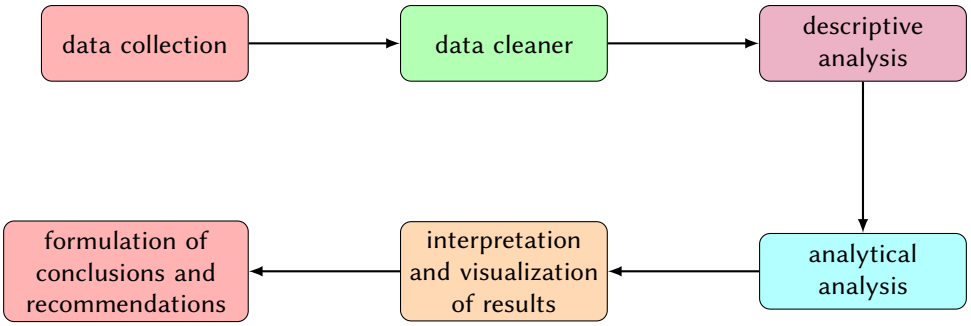

Figure 1: Scheme of analysis of the generated data set.

email	fullname	title	country
13845902468@163.com	Limei Yan	Task optimization and scheduling of distributed cyber–physical system based on improved ant colony algorithm	China
abiola@vt.edu	Abiola A. Akanmu	TOWARDS NEXT GENERATION CYBER-PHYSICAL S	Virginia
ahaleem@jmi.ac.in	Abid Haleem	An integrated outlook of Cyber–Physical Systems for Industry 4.0: Topical practices, architecture, and applications	India
albous@mail.ntua.gr	Alexandros Bousdekis	A human cyber physical system framework for operator 4.0 – artificial intelligence symbiosis	Greece
amitkrtyagi025@gmail.com	Amit Kumar Tyagi	Cyber Physical Systems: Analyses, challenges and possible solutions	India
amitkrtyagi025@gmail.com	Amit Kumar Tyagi	AARIN: Affordable, accurate, reliable and innovative mechanism to protect a medical cyber-physical system using blockchain technology	India
anumba@ufl.edu	Chimay J. Anumba	TOWARDS NEXT GENERATION CYBER-PHYSICAL S	Florida
ashutosh.trivedi@colorado.edu	Ashutosh Trivedi	Secure-by-construction synthesis of cyber-physical systems	USA
aswathy.su@gmail.com	S.U. Aswathy	AARIN: Affordable, accurate, reliable and innovative mechanism to protect a medical cyber-physical system using blockchain technology	India

Figure 2: Table of generated data.

3. The descriptive analysis includes an overview of the available data in the dataset, determination of its structure, data visualisation, and calculation of statistical indicators (measures of central tendency, standard deviation, quartiles, interquartile range, etc.).
4. Analytical analysis is the stage where the machine learning methods used, namely regression, classification, clustering, etc., are usually presented.
5. Interpretation and visualisation of the results include determining the main conclusions after analysing the data set and building the necessary results graphs for clarity.
6. At the stage of formulating conclusions and recommendations, the results of the data analysis are described, and recommendations for further data analysis are written.

The study analysed the effectiveness of sending letters to scientists and the feedback they received in the form of reviewing the journal using the following steps presented in table 1.

Let us describe in detail the stages of data analysis of the obtained dataset for the scientific journal "Journal of Edge Computing".

In the first stage, it was necessary to form the required dataset; for this purpose, a table with many records was built from the data collected daily in the above form. At the end of the month, we manually cleaned the data set and deleted those records with undeliverable email addresses from the table. The final dataset for further analysis is shown in the form of a data table consisting of four columns, namely: country number, country name (country), number of letters sent to scientists from this country (mailings) and number of journal views received from this country (visitors), the first records of the resulting table are shown in figure 3.

Table 1
Steps to determine the effectiveness of emails to scientists.

№	Stage name	Description
1	Comparison of countries by the number of letters sent	We compare how many letters were sent to different countries, which is aimed at identifying those countries that should be focused on in the future
2	Comparison of countries by the number of journal views	We compare how many scientists from different countries have viewed the journal, which is aimed at identifying those countries that are most interested in the journal's research topics
3	Calculation of the conversion rate	We will calculate the conversion rate for each country in order to assess the effectiveness of the letters sent to scientists from different countries
4	Use of statistical significance tests	We use statistical significance tests to find out whether there is a statistically significant difference between the number of views or conversion rates in different countries
5	Using machine learning models (building a linear regression model)	Machine learning models can be used to predict how many scientists from a particular country will view a journal if an email is sent to them
6	Country segmentation (data clustering)	Countries can be segmented based on their distribution into groups with similar characteristics, for example, the number of letters sent or the number of views received

Let us load and view the obtained data set in the RStudio environment using the R programming language. For this, we install and connect the necessary packages for working with data.

Let us use the dim() and print() functions of the R language to get the following information about the data set: the table consists of 4 columns and 53 rows; let us look at the first records of the table to check if the required data set is loaded correctly into the dset variable (figure 4). From the figures, we can see that the loaded set in dset matches the built data table.

№	country	mailings	visitors
1	Algeria	2	4
2	Australia	14	7
3	Austria	4	9
4	Azerbaijan	0	1
5	Bahrain	0	1
6	Bangladesh	1	8
7	Brazil	5	3
8	Canada	20	11
9	China	186	4
10	Denmark	3	1

Figure 3: Final dataset.

Now let us move on to consider the data set; you need to determine its internal structure, this we use the str() diagnostic function pay special attention to the data types; in our case, the data types and the values entered into the table cells are the same, no further manipulations are required.

In the next stage, we will conduct a descriptive analysis of the available data set and display its descriptive statistics. The presentation of such statistics will contribute to a better understanding of the available data. For numerical columns of data, the minimum and maximum values will be displayed, which allows you to understand the range of data values, median, mean, and quartiles of data. For this, we use the summary() function, which provides basic statistical information for each data set column

```
> dim(dset)
[1] 53  4
> print(dset)
    №        country mailings visitors
1   1        Algeria        2        4
2   2      Australia       14        7
3   3        Austria        4        9
4   4     Azerbaijan        0        1
5   5        Bahrain        0        1
6   6     Bangladesh        1        8
7   7         Brazil        5        3
8   8         Canada       20       11
9   9          China      186        4
10 10        Denmark        3        1
```

Figure 4: Checking for the correct loading of the required data set.

```
> summary(dset)
       №          country             mailings         visitors
 Min.   : 1   Length:53          Min.   :  0.00   Min.   :  1.00
 1st Qu.:14   Class :character   1st Qu.:  0.00   1st Qu.:  1.00
 Median :27   Mode  :character   Median :  3.00   Median :  4.00
 Mean   :27                      Mean   : 10.13   Mean   : 12.57
 3rd Qu.:40                      3rd Qu.:  6.00   3rd Qu.: 10.00
 Max.   :53                      Max.   :186.00   Max.   :195.00
```

Figure 5: Descriptive statistics of the data.

(figure 5).

If you look at the result, you can see that in the mailings and visitors columns, the values of the first and 3rd quartiles are small and close, which means that most of the values in these columns are in this range, and this is confirmed by the interquartile range for these columns (figure 6).

```
> IQR(dset$mailings)
[1] 6
> IQR(dset$visitors)
[1] 9
```

Figure 6: Finding interquartile ranges for columns.

Let us look at the maximum values of these columns. We see that they are significantly large, indicating that there are so-called outliers in the data (abnormally large values in the available data set). In order to visually verify the preliminary results of the study, we will build dot plots of the data distribution by country based on the number of letters sent and received journal visits figure 7. Looking at the obtained diagrams in figure 8-9, it was found that the most significant emissions in terms of the number of sent letters are observed in China, India and the United States, which may be due to a sufficiently large number of users since these countries are the largest in terms of population in this dataset, and the peculiarities of their active behaviour on the network and access to it. In terms of the results of log views, outliers were found in India, Nigeria, the United States, and Ukraine. India, Nigeria, and the United States are countries with large populations, which automatically increases the potential scientific audience of the journal, and there is also a different rate of growth in the level of education and interest in scientific research. As for Ukraine, the scientific validity of the research, the accessible form of presentation, and the credibility of the authors of publications can contribute to the journal's popularity.

```
> dset%>%ggplot(aes(country,mailings))+
+   geom_point()+
+   theme(axis.text.x = element_text(angle = 90))
> dset%>%ggplot(aes(country,visitors))+
+   geom_point()+
+   theme(axis.text.x = element_text(angle = 90))
```

Figure 7: Creating separate dot plots for mailing and visits.

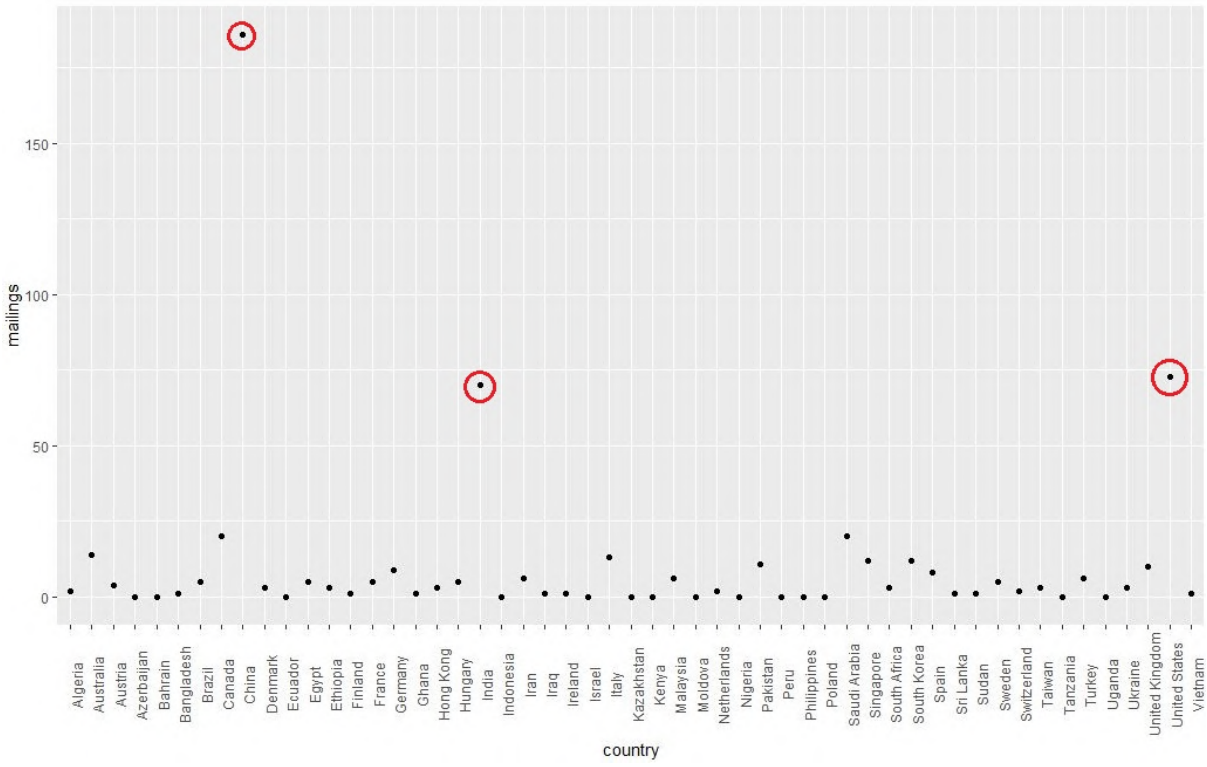

Figure 8: Separate dot graphs for mailing and visiting.

In order to understand the distribution of data in mailings and visitors, it is advisable to build boxplots (a box with a moustache) that reflect the distribution of data in these columns and their variability and asymmetry (figure 10).

Let us describe the result of visualising the boxplots, which is a refinement of the result obtained earlier using the summary() function: pay attention to the location of the box, as it is not in the centre of the graph, but at the bottom, which indicates the asymmetry of the distribution of emails sent and views received; the middle line in the boxes reflects the median of the data, we can see that for mailings it is slightly higher than the average value, and for visitors, on the contrary, slightly lower; the edges of the boxes representing the first and third quartiles confirm the small inter-quartile range calculated earlier, i.e. how scattered the data is in the set itself, so the values for mailings and visitors do not differ significantly; next, we describe the upper whiskers of the box, which extend to the maximum values, in mailings it is about 15, in visitors – about 12; the points on the graph that are located well above the upper whiskers are called outliers, in mailings outliers are the values at 23, 73, 75 and 186, in visitors – at 41, 53, 130 and 195. So, we can summarise that both of these columns have similar values, their distributions are similar and skewed, outliers are present, and we need to calculate the standard deviation for these columns to determine the variability of the data.

Let us calculate the standard deviation for mailings and visitors. This will give you an idea of how far a typical value in a column is from the mean (figure 11) if you don't take into account outliers.

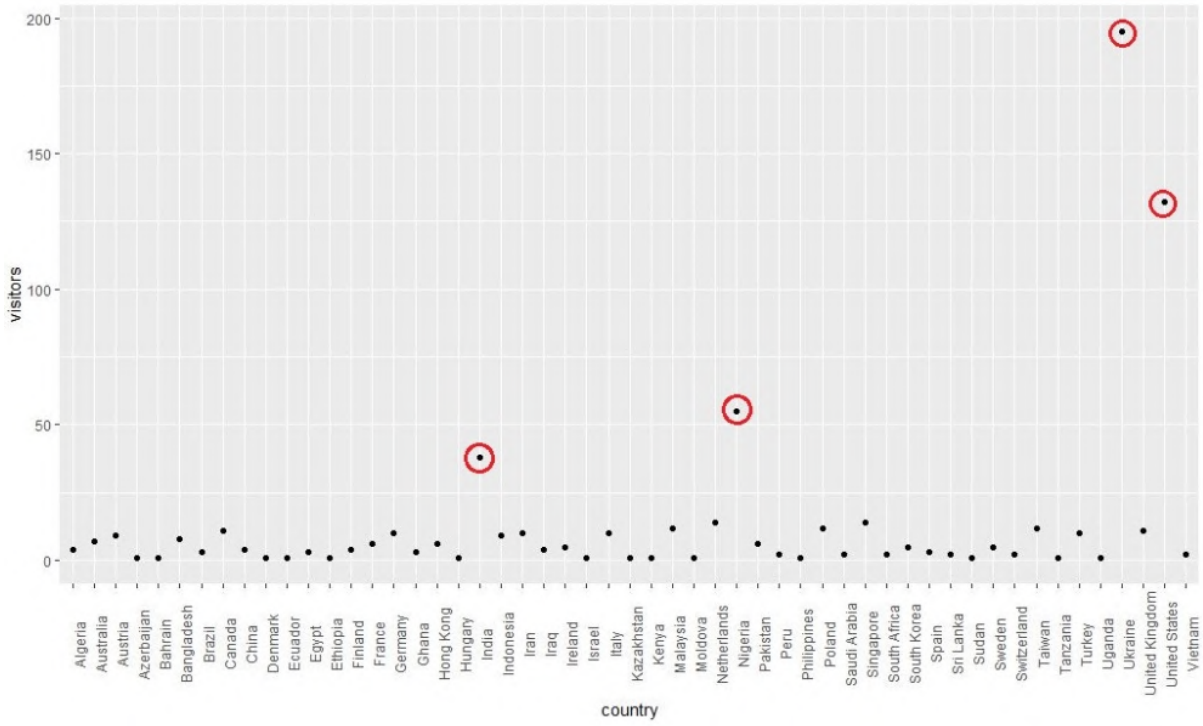

Figure 9: Separate dot graphs for mailing and visiting.

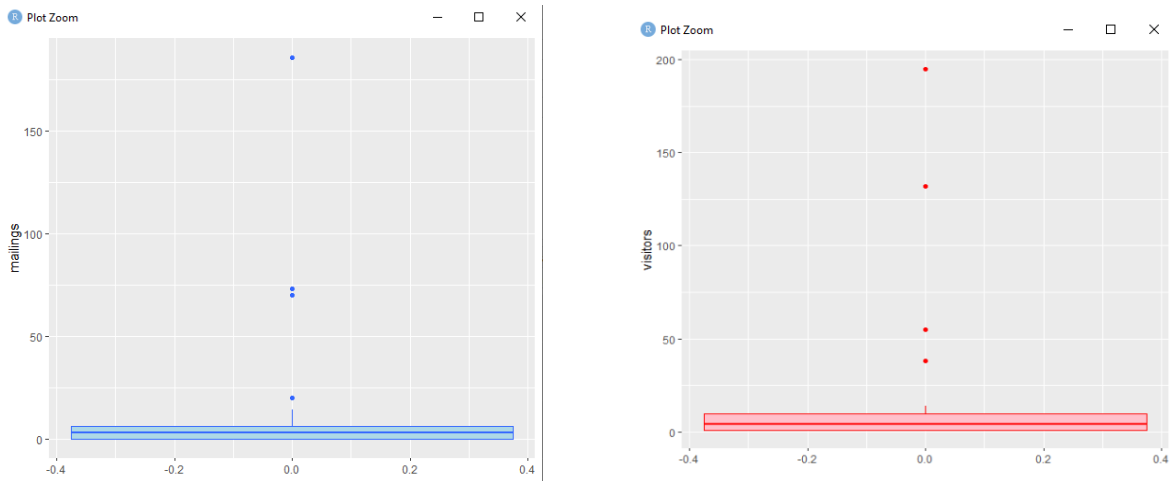

Figure 10: Boxplots of the mailings and visitors columns of the studied dataset.

```
> sd(dset$mailings)
[1] 5.172512
> sd(dset$visitors)
[1] 4.19198
```

Figure 11: Calculating the standard deviation for mailings and visitors.

The standard deviations obtained are small, meaning we have a low level of data variability. Let us see if the value of sending emails and the number of visits to the magazine change in tandem, so let us calculate the covariance coefficient (figure 12).

The value of this coefficient is relatively high, indicating a strong relationship between the values of mailings and visitors, i.e. they change in the same direction, and an increase in visitors accompanies an

```
> print(cov(dset$mailings, dset$visitors))
[1] 160.3084
```

Figure 12: Calculation of the covariance coefficient.

increase in mailings. Let us check how strongly the number of mailings and the number of visits to the journal are related by calculating the Spearman correlation coefficient (figure 13).

```
> cor(dset$mailings, dset$visitors, method = "pearson")
[1] 0.1766738
>
> cor(dset$mailings, dset$visitors, method = "spearman")
[1] 0.5242531
```

Figure 13: Calculation of Spearman's correlation coefficient.

The obtained correlation coefficient is closer to 1, which indicates a strong linear relationship, so we can build a model of the relationship between the dependent variable visitors and the independent variable mailings using linear regression, which is one of the most common machine learning methods (figure 14).

```
> relation <- lm(dset$visitors ~ dset$mailings)
> print(relation)

Call:
lm(formula = dset$visitors ~ dset$mailings)

Coefficients:
  (Intercept)  dset$mailings
      10.5321         0.2007
```

Figure 14: Using linear regression, build a model of the relationship between the dependent variable visitors and the independent variable mailings.

Let us write the formula of the resulting model in the form of the following equation:

$$visitors = 0.2 * mailings + 10.53 \tag{4}$$

The built model can predict future visits to the journal's page depending on the number of letters sent to scientists in a particular country.

Let us calculate the conversion rate for each country as the number of letters sent by the number of visits to the journal's page – this will allow us to assess the effectiveness of the mailing to each country; for this rate, we will add another column "Conversion" to the data set (figure 15).

Let us take a look at the list of countries with a conversion rate of more than 100% (figure 16).

Thus, these are scientists from those countries who spent much more time sending letters than received a positive result in the form of journal views, which should be considered in the future.

As a follow-up to the above analysis, it is advisable to use a statistical significance test, such as a t-test, to compare the mean values of the two columns (mailings, visitors) to determine whether there is a statistically significant difference between the number of visits and the number of mailings or conversion rates in different countries. In R, there is a particular function t.test() that displays the t-statistic (t), degree of freedom (df), p-value, and confidence interval. If we apply the t-test to mailings and visitors, we get the following result in figure 17.

The calculations allow us to draw the following conclusion: the value of the t-statistic (t=-0.41427) is small, which indicates the acceptance of the null hypothesis, namely, the closeness of the average values

```
> dset$convertion <- round((dset$mailings / dset$visitors) * 100, 2)
> print(dset)
# A tibble: 53 × 5
     `№` country    mailings visitors convertion
   <dbl> <chr>         <dbl>    <dbl>      <dbl>
 1     1 Algeria           2        4       50
 2     2 Australia        14        7      200
 3     3 Austria           4        9       44.4
 4     4 Azerbaijan        0        1        0
 5     5 Bahrain           0        1        0
 6     6 Bangladesh        1        8       12.5
 7     7 Brazil            5        3      167.
 8     8 Canada           20       11      182.
 9     9 China           186        4     4650
10    10 Denmark           3        1      300
# i 43 more rows
```

Figure 15: Conversion rate for each country.

```
> dset%>% filter(country == "Australia" |
+                  country == "Brazil" |
+                country == "Canada" |
+                country == "China" |
+                country == "Denmark" |
+                country == "Egypt" |
+                country == "Ethiopia" |
+                country == "Hungary" |
+                country == "India" |
+                country == "India" |
+                country == "Pakistan" |
+                country == "Saudi Arabia" |
+                country == "South Africa" |
+                country == "South Korea" |
+                country == "Spain")
# A tibble: 14 × 5
     `№` country      mailings visitors convertion
   <dbl> <chr>           <dbl>    <dbl>      <dbl>
 1     2 Australia          14        7      200
 2     7 Brazil              5        3      167.
 3     8 Canada             20       11      182.
 4     9 China             186        4     4650
 5    10 Denmark             3        1      300
 6    12 Egypt               5        3      167.
 7    13 Ethiopia            3        1      300
 8    19 Hungary             5        1      500
 9    20 India              70       38      184.
10    33 Pakistan           11        6      183.
11    37 Saudi Arabia       20        2     1000
12    39 South Africa        3        2      150
13    40 South Korea        12        5      240
14    41 Spain               8        3      267.
```

Figure 16: List of countries with conversion rates above 100%.

of the two columns and the absence of a significant difference between them, which is also confirmed by the value of the p-value coefficient.

To segment the countries by the number of emails sent and the number of visits to the magazine,

```
> t.test(dset$mailings, dset$visitors)

        Welch Two Sample t-test

data:  dset$mailings and dset$visitors
t = -0.41427, df = 102.35, p-value = 0.6795
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.087258   9.219333
sample estimates:
mean of x mean of y
 10.13208  12.56604

> t.test(dset$mailings, dset$convertion)

        Welch Two Sample t-test

data:  dset$mailings and dset$convertion
t = -1.99, df = 52.199, p-value = 0.05184
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -354.6204    1.4612
sample estimates:
mean of x mean of y
 10.13208 186.71170

> t.test(dset$visitors, dset$convertion)

        Welch Two Sample t-test

data:  dset$visitors and dset$convertion
t = -1.962, df = 52.257, p-value = 0.0551
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -352.231364    3.940043
sample estimates:
mean of x mean of y
 12.56604 186.71170
```

Figure 17: The result of the t-test for mailings and visitors.

we will use k-means clustering of the data set. This is one of the most common, simple, effective, and flexible methods of cluster data analysis and allows grouping data based on their similarity. The clustering will be carried out by three clusters (figure 18).

Let us display the number of countries in each cluster (figure 19).

Look at the countries' lists in each cluster (figure 20).

After clustering by the k-means method, the following results were obtained, which allowed us to divide the countries into three clusters based on the number of letters sent and visits to the journal. Each cluster has its characteristics, which allow us to conclude the level of interest of scientists from different countries in this journal.

Cluster 1 (green triangles) includes countries where the number of letters sent and visits to the journal is small, which indicates a low level of interest of scientists in the journal's topics; figure 21 shows that there were few mailings and, accordingly, few visits. In the countries of this cluster, the journal's promotion activities are likely to be ineffective, so it may be worth revising the journal's promotion strategy in these countries or focusing on the audience from other clusters.

```
> k <- 3
> kmeans_results <- kmeans(dset[,3:4], centers = k)
> dset$cluster <- kmeans_results$cluster
> print(dset)
# A tibble: 53 × 6
     `№` country    mailings visitors convertion cluster
   <dbl> <chr>         <dbl>    <dbl>      <dbl>   <int>
 1     1 Algeria           2        4         50       1
 2     2 Australia        14        7        200       3
 3     3 Austria           4        9         44.4     3
 4     4 Azerbaijan        0        1          0       1
 5     5 Bahrain           0        1          0       1
 6     6 Bangladesh        1        8         12.5     1
 7     7 Brazil            5        3        167.      1
 8     8 Canada           20       11        182.      3
 9     9 China           186        4       4650       2
10    10 Denmark           3        1        300       1
# i 43 more rows
```

Figure 18: Clustering by three clusters.

```
> library(broom)
> kmeans_results %>%
+   augment(dset) %>%
+   count(cluster)
# A tibble: 3 × 2
  cluster     n
    <int> <int>
1       1    32
2       2     4
3       3    17
```

Figure 19: Number of countries in each cluster.

The second cluster (blue square), on the contrary, includes countries with abnormally high values either in the number of mailings or in the number of visits, so you can see that in some countries (USA, Ukraine), the interest in the journal is high with a small number of mailings, and in some countries (India, China), on the contrary, it is low, so we do not see the expediency of sending mailings to these countries in the future. For the countries in this cluster, a more detailed analysis is likely to be required to understand the reasons for such deviations from the general trend, and it may be necessary to adjust the strategy for each country separately.

The third cluster (red circles) includes countries with average values that reflect both a high level of interest in the journal, for example, countries such as Austria, the Netherlands, Turkey, Malaysia, etc., and a low level of interest, for example, Australia, Canada, Pakistan, etc. Also, it includes countries such as Poland and Nigeria, to which no mailings were made. The countries in this cluster are the most heterogeneous in terms of interest, so to build an effective journal promotion strategy, it is necessary to further segment this cluster by other criteria, such as language, region, scientific interests, etc.

To sum up, based on a detailed analysis of the number of letters sent and visits to the "Journal of Edge Computing" in different countries, we could divide them into groups according to the level of interest in the topics of the scientific publication. This country segmentation will allow the editorial team to define the journal's target audience more accurately and focus marketing efforts on countries with high interest. However, although the clustering results give us a general idea of the distribution of countries by level of interest, they do not explain why certain countries have a high, medium or low level of interest in the journal. To better determine the reasons for the difference in interest among researchers,

```
> claster1
# A tibble: 32 × 6
     `№` country    mailings visitors convertion cluster
   <dbl> <chr>         <dbl>    <dbl>      <dbl>   <int>
 1     1 Algeria           2        4       50         1
 2     4 Azerbaijan        0        1        0         1
 3     5 Bahrain           0        1        0         1
 4     6 Bangladesh        1        8       12.5       1
 5     7 Brazil            5        3      167.        1
 6    10 Denmark           3        1      300         1
 7    11 Ecuador           0        1        0         1
 8    12 Egypt             5        3      167.        1
 9    13 Ethiopia          3        1      300         1
10    14 Finland           1        4       25         1
# i 22 more rows
# i Use `print(n = ...)` to see more rows
> claster2
# A tibble: 4 × 6
     `№` country       mailings visitors convertion cluster
   <dbl> <chr>            <dbl>    <dbl>      <dbl>   <int>
1     9 China              186        4     4650          2
2    20 India               70       38      184.         2
3    50 Ukraine              3      195        1.54       2
4    52 United States       73      132       55.3        2
> claster3
# A tibble: 17 × 6
     `№` country        mailings visitors convertion cluster
   <dbl> <chr>             <dbl>    <dbl>      <dbl>   <int>
 1     2 Australia            14        7      200          3
 2     3 Austria               4        9       44.4        3
 3     8 Canada               20       11      182.         3
 4    16 Germany               9       10       90          3
 5    22 Iran                  6       10       60          3
 6    26 Italy                13       10      130          3
 7    29 Malaysia              6       12       50          3
 8    31 Netherlands           2       14       14.3        3
 9    32 Nigeria               0       55        0          3
10    33 Pakistan             11        6      183.         3
11    36 Poland                0       12        0          3
12    37 Saudi Arabia         20        2     1000          3
13    38 Singapore            12       14       85.7        3
14    40 South Korea          12        5      240          3
15    46 Taiwan                3       12       25          3
16    48 Turkey                6       10       60          3
17    51 United Kingdom       10       11       90.9        3
```

Figure 20: Lists of countries in each cluster.

additional research is needed to identify the factors influencing users' decisions to visit the journal and study the scientific publications presented. As for the factors, we can assume the following: cultural peculiarities that may affect the perception of information and the choice of information sources; the level of economic development of the country, respectively, access to information technology; access to the Internet, especially its speed, the cost of access to the network, which may limit users to online resources; language barriers, since these are English-language publications; thematic relevance as the correspondence of the journal's topics to the interests and needs of scientists from different countries; availability of other similar scientific journals.

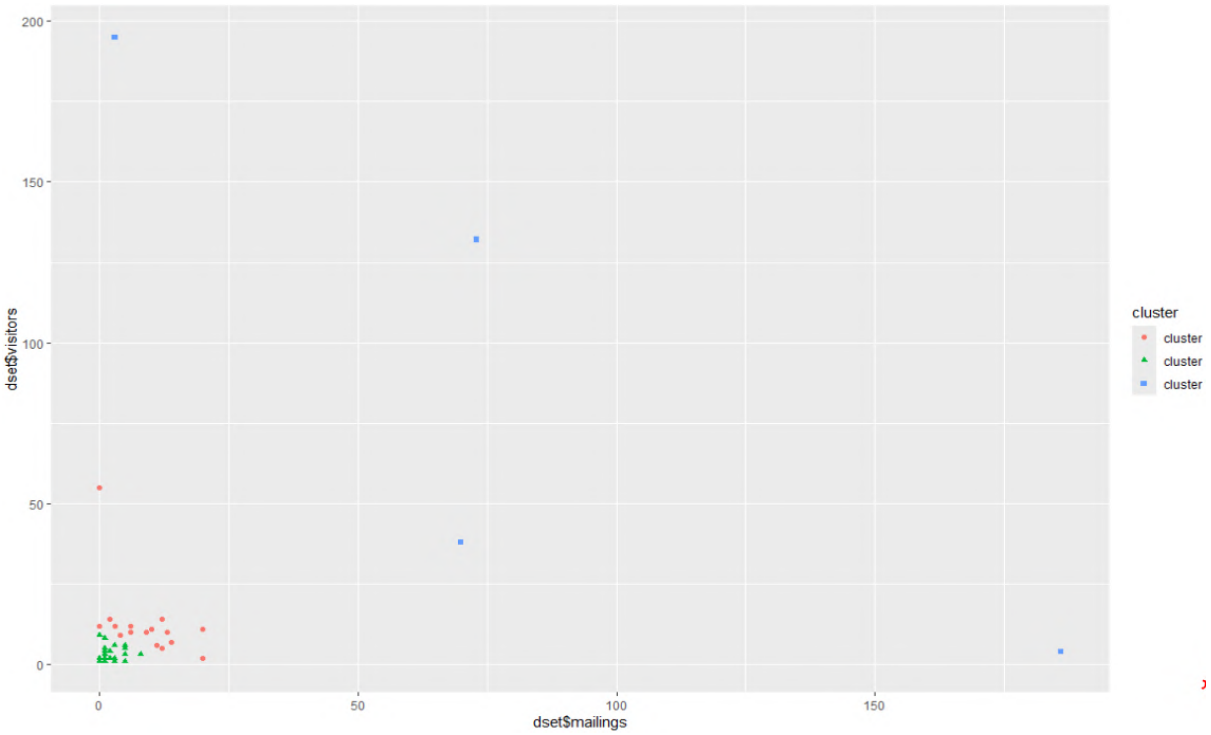

Figure 21: Visualisation of clustered data.

5. Conclusions

The study conducted a comprehensive analysis of the data set collected from the mailing lists by the scientific journal "Journal of Edge Computing" editors, which was aimed at establishing the effectiveness of electronic invitations sent to foreign and domestic researchers for review and further publication. It revealed the list of countries of residence of researchers who have expressed interest in the journal. We can state that there is a positive correlation, i.e. the more invitations were sent, the higher the number of visits to the journal. In some countries, there is a very high or low interest in the journal, so the analysis highlighted the differences in the level of involvement of scientists from different countries in reviewing the journal. As for the results obtained from the data clustering, countries were grouped into appropriate clusters depending on the level of interest of scientists. For the future, a linear model was built to predict the preliminary result of the interest of scientists from different countries, which will allow the journal's editorial board to attract more interested countries and effectively conduct a marketing campaign to attract scientists. To carry out this study, the R programming language and various statistical methods were used to clean the data, conduct descriptive and analytical data analysis, and build visualisations, which allowed for efficient and detailed analysis with the required results. Thus, the study has shown the effectiveness of the measures taken to send out electronic invitations.

Declaration on Generative AI: During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using this service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] T. A. Vakaliuk, S. O. Semerikov, Introduction to doors Workshops on Edge Computing (2021-2023), Journal of Edge Computing 2 (2023) 1–22. doi:10.55056/jec.618.

[2] N. Ratmaningsih, A. Abdulkarim, D. S. Logayah, D. N. Anggraini, P. Sopianingsih, F. Y. Adhitama, M. A. Widiawaty, Android-Based Augmented Reality Technology in the Application of Social

Studies Text-books in Schools, Journal of Advanced Research in Applied Sciences and Engineering Technology 48 (2024) 29–50. doi:10.37934/araset.48.1.2950.
[3] K. Katahira, Evaluating the predictive performance of subtyping: A criterion for cluster mean-based prediction, Statistics in Medicine 42 (2023) 1045–1065. doi:10.1002/sim.9656.
[4] S. Ovchinnikova, S. Anders, Simple but powerful interactive data analysis in R with R/LinkedCharts, Genome Biology 25 (2024). doi:10.1186/s13059-024-03164-3.
[5] K. Imran, W. N. Arifin, T. M. H. T. Mokhtar, Data Analysis in Medicine and Health Using R, 2024. URL: https://bookdown.org/drki_musa/dataanalysis/.
[6] J. Davis, Introduction to Environmental Data Science, 2023. doi:10.1201/9781003317821.