Sampurna_Simkhada_Report.docx

by sampurna simkhada

Submission date: 24-Mar-2025 08:48PM (UTC+0545)
Submission ID: 2623586845
File name: Sampurna_Simkhada_Report.docx (612.38K)
Word count: 1964
Character count: 10987

School of Computing, Creative Technology and Engineering

Student ID	7766895
Student Name	Sampurna Simkhada
Module Name & CRN	Applied Machine Learning
Level	5
Assessment Name & Part No.	Part 1
Project Title	Case Study 1:Bank Marketing Campagain
Data of Submission	3/24/2025
Course	Applied Machine Learning
Academic Year	2025

Table of Contents

Introduction:

Banking will continue to be an extraordinarily high-profile field which is the arena for competition for new customers, along with existing ones, in view of their (banks') displaying ranges of possibilities and proposals through marketing that are set at the most targeted readers. There is no point in trying to gain brand recognition and business growth without marketing that is therefore for sure the main factor in this matter, but without the new ideas it will not be the case nevertheless (Fan, 2023). The present approach will bring about very much more awareness of the brand and will be a more effective force promoting business growth than it used to be; however, the conventional means that still stay in use now are already superfluous (K and Raja, 2018) The new era of customers specifically demands real-time personalization, seamless experiences, online communication with up-to-date content, and additional sideline services to fulfil their needs on time.

Literature Review:

Banks' data-driven marketing strategies have been an appealing theme for the specialists nowadays due to the fact that the exact segmentation is of the first order of importance and the poll of the individual, which would usually include the consumer's preferred products, the location of the consumer in the customer segment cloud and the consumer's past banking experience, would be the first data point, using which they could predict customer bank balance (Fan, 2023).

R is a statistical computing and visualization machine of knowledge, has become a powerful and efficient tool for analysis of difficult data such as those from bank marketing campaigns. The purpose of this literature review is to outline the main approaches and methods which are employed in the analysis of such datasets, with a special emphasis on four particular issues, namely, dealing with missing values, identifying the outliers, applying principal component analysis (PCA), implementing data visualization, and producing the charts. (Jolliffe and Cadima, 2016)

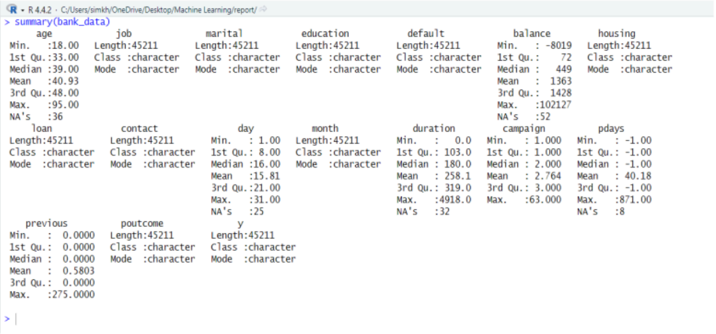
Data visualization plays a vital role in understanding complex datasets and communicating insights effectively. The use of histograms, bar plots, and boxplots is common in the analysis of marketing campaign data (Nordmann et al., 2022). Visualizations can reveal hidden patterns and trends that are not immediately obvious from raw data. In this analysis, histograms and bar plots are used to visualize the distribution of variables such as customer balance and job roles, while boxplots are used to explore the relationship between age groups and call duration (Wang, Makedon and Chakrabarti, 2004). Scaling and Normalizing data is a crucial part to be in order before feeding it into any machine learning algorithm. The code exhibits the way of testing the quality of standardized numeric variables by using z-scores that provide a guarantee for each factor to have the same influencing role in the following analysis or prediction. (Chen, 2021) Standardization is needed the most when you have to pile up individual measurements like distance, money, or temperature, etc, together as the standard method brings them on the same scale. (Brownstein, Adolfsson and Ackerman, 2019).

The application of R in analysing bank marketing campaign data provides a systematic approach to data exploration, encompassing preprocessing, outlier detection, visualization, and dimensionality reduction Singh, 2019).. Previous studies have highlighted the importance of these methods in delivering relevant and insightful analyses of complex datasets.By utilizing tools such as MICE for missing data imputation, IQR for outlier detection, PCA for dimensionality reduction, and advanced visualization techniques, analysts can derive valuable insights from bank marketing data (Srinivasan, 2023).

Exploratory Data Analysis :

Summary:

We can obtain a summary of the data frame by using the summary() function.



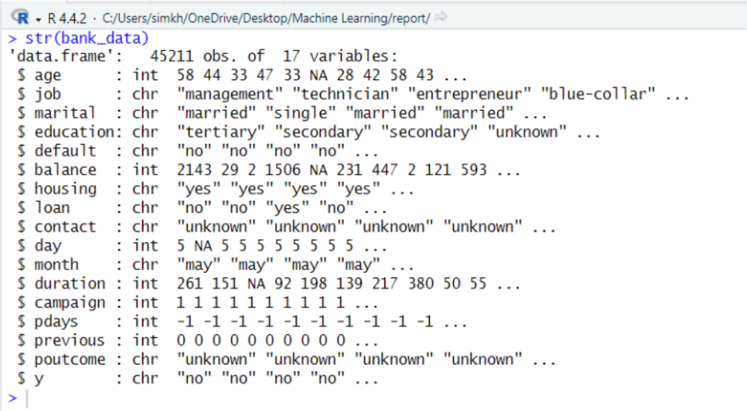
```
R 4.4.2 · C:/Users/simkh/OneDrive/Desktop/Machine Learning/report/
> summary(bank_data)
      age             job              marital           education           default             balance          housing         
 Min.   :18.00   Length:45211       Length:45211       Length:45211       Length:45211       Min.   : -8019   Length:45211      
 1st Qu.:33.00   Class :character   Class :character   Class :character   Class :character   1st Qu.:    72   Class :character  
 Median :39.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :   449   Mode  :character  
 Mean   :40.93                                                                               Mean   :  1363                     
 3rd Qu.:48.00                                                                               3rd Qu.:  1428                     
 Max.   :95.00                                                                               Max.   :102127                     
 NA's   :36                                                                                  NA's   :52                         
     loan             contact               day            month             duration         campaign          pdays        
 Length:45211       Length:45211       Min.   : 1.00   Length:45211       Min.   :   0.0   Min.   : 1.000   Min.   : -1.00  
 Class :character   Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.: -1.00  
 Mode  :character   Mode  :character   Median :16.00   Mode  :character   Median : 180.0   Median : 2.000   Median : -1.00  
                                       Mean   :15.81                      Mean   : 258.1   Mean   : 2.764   Mean   : 40.18  
                                       3rd Qu.:21.00                      3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.: -1.00  
                                       Max.   :31.00                      Max.   :4918.0   Max.   :63.000   Max.   :871.00  
                                       NA's   :25                         NA's   :32                        NA's   :8       
    previous           poutcome              y            
 Min.   :  0.0000   Length:45211       Length:45211      
 1st Qu.:  0.0000   Class :character   Class :character  
 Median :  0.0000   Mode  :character   Mode  :character  
 Mean   :  0.5803                                        
 3rd Qu.:  0.0000                                        
 Max.   :275.0000                                        

> 
```

Figure 1: Summary of the dataset

Figure 1 shows the summary which reflects the Maximum value, Minimum value, Mean, Median, first quartile, third quartile and for numerical variables. Categorical variables the summary reflects length, Class and Mode.

Data Type:

We can find the data type by the help of str() function.


R 4.4.2 · C:/Users/simkh/OneDrive/Desktop/Machine Learning/report/

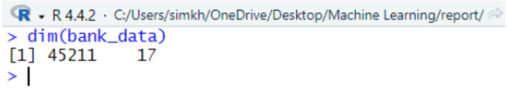
```
> str(bank_data)
'data.frame':   45211 obs. of  17 variables:
 $ age      : int  58 44 33 47 33 NA 28 42 58 43 ...
 $ job      : chr  "management" "technician" "entrepreneur" "blue-collar" ...
 $ marital  : chr  "married" "single" "married" "married" ...
 $ education: chr  "tertiary" "secondary" "secondary" "unknown" ...
 $ default  : chr  "no" "no" "no" "no" ...
 $ balance  : int  2143 29 2 1506 NA 231 447 2 121 593 ...
 $ housing  : chr  "yes" "yes" "yes" "yes" ...
 $ loan     : chr  "no" "no" "yes" "no" ...
 $ contact  : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ day      : int  5 NA 5 5 5 5 5 5 5 5 ...
 $ month    : chr  "may" "may" "may" "may" ...
 $ duration : int  261 151 NA 92 198 139 217 380 50 55 ...
 $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ y        : chr  "no" "no" "no" "no" ...
> 
```

Figure 2: Structure of data frame

Figure 2 shows the structure of the dataframe which has 7 integer variables and 10 character variables .

Dimension:

We can predict the data dimension with the help of the dim() function.


R 4.4.2 · C:/Users/simkh/OneDrive/Desktop/Machine Learning/report/

```
> dim(bank_data)
[1] 45211    17
> 
```

Figure 3: Dimension of the data frame(bank_data)

Figure 3 shows that the data frame has 45211 number of rows(observations) and 17 columns(variables) .

Histogram:

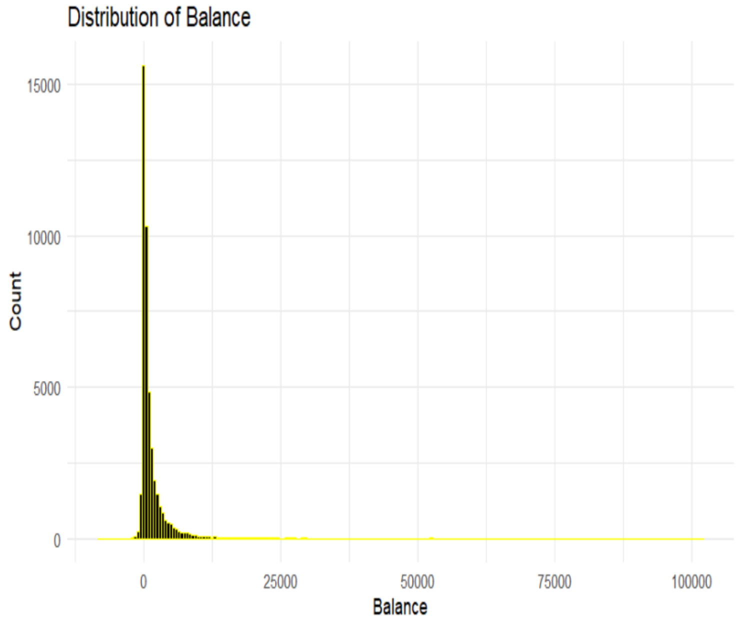


Figure : histogram showing distribution of balance

The chart shown above represents a distribution of bank account balances in a bank's campaign data set. The x-axis shows the balance amounts (from 0 to 100,000), while the y-axis provides the number of accounts up to 15,000. The data is heavily skewed to the right, indicating that the highest number of accounts cluster near the zero-balance mark, i.e. most accounts have very little money (below 5,000). The number starts to decline significantly as the balance tops 25,000; consequently, the amount of accounts with higher balances goes down too.

A oversized tail to the right also appears with an extreme outlier near 100,000, meaning some account holders have extremely high balances and only constitute a tiny proportion of the total accounts. The majority of clients expect their data to be in a reasonable range but some may have a bigger volume of money. A few people with the largest portion of the money are at the top of the income distribution while the rest of the population has much smaller amounts, which is a certain kind of fallacy usually seen in financial statics.

10

Principal Component Analysis(PCA):

PCA is a dimension reduction technique in machine learning.

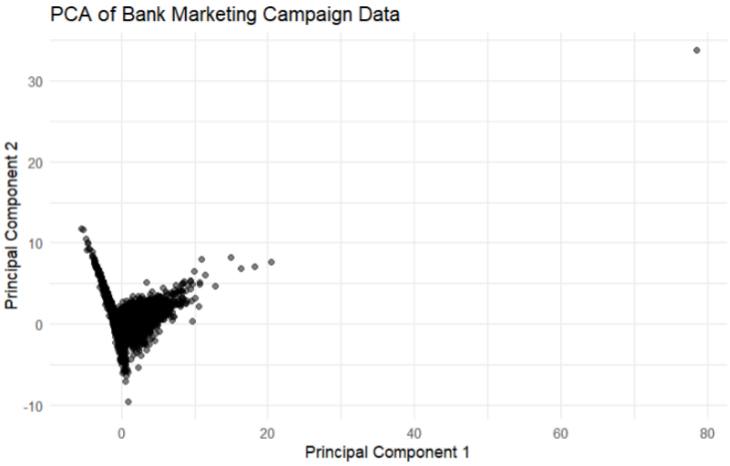


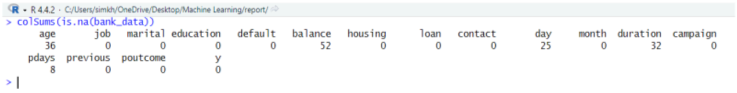
Figure: PCA visualization as Scatterplot

The PCA results scatter plot on bank marketing data is shown so that the x-axis represents PC1 and the y-axis stands for PC2. The greater part of the data lies in the lower-left corner, which indicates its homogeneity, whereas a wider distribution along PC1 would reflect more variability. For instance (PC1: 80, PC2: 30) is the location of one of the few of their kind. The diagram reveals that there are no clear groups, but the widespread shot along PC1 could be a pointer to the direct patterns of customer behavior

Data Pre-processing

Missing Values

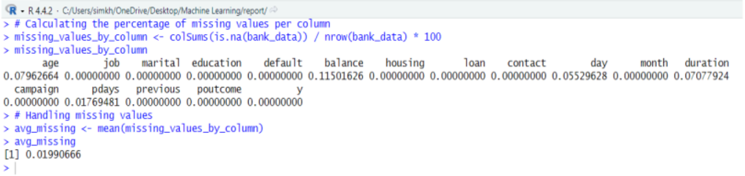
Missing values can be calculated by using is.na() function.



R 4.4.2 · C:/Users/simkh/OneDrive/Desktop/Machine Learning/report/

```
> colSums(is.na(bank_data))
      age       job   marital education   default   balance   housing      loan   contact       day     month  duration  campaign 
       36         0         0         0         0        52         0         0         0        25         0        32         0 
    pdays  previous  poutcome         y 
        8         0         0         0 
> |
```

Figure 6 : missing values per column

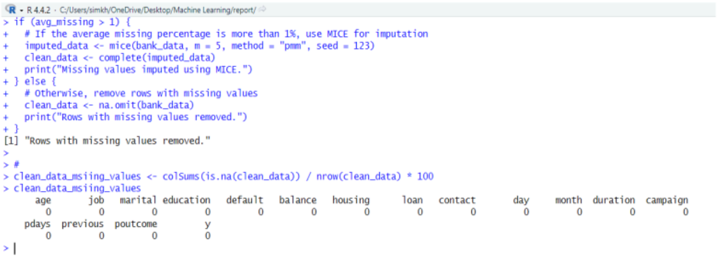


R 4.4.2 · C:/Users/simkh/OneDrive/Desktop/Machine Learning/report/

```
> # Calculating the percentage of missing values per column
> missing_values_by_column <- colSums(is.na(bank_data)) / nrow(bank_data) * 100
> missing_values_by_column
       age        job    marital  education    default    balance    housing       loan    contact        day      month   duration 
0.07962664 0.00000000 0.00000000 0.00000000 0.00000000 0.11501626 0.00000000 0.00000000 0.00000000 0.05529628 0.00000000 0.07077924 
  campaign      pdays   previous   poutcome          y 
0.00000000 0.01769481 0.00000000 0.00000000 0.00000000 
> # Handling missing values
> avg_missing <- mean(missing_values_by_column)
> avg_missing
[1] 0.01990666
> 
```

Figure 7 :Missing values by percentage

Figure 6 shows gives the number of missing values where age has 30 missing values, balance has 52 missing values, day has 25 missing values, duration has 32 missing values where other variable have 0 missing values .Here are almost 153 missing values in this dataset . the missing data set in percentage is also not above or equal to 1 % for each row. The average percentage of missing values in each column is also only 0.19% which is minimal. So, I decided to remove those missing values form the data set.



R 4.4.2 · C:/Users/simkh/OneDrive/Desktop/Machine Learning/report/

```
> if (avg_missing > 1) {
+   # If the average missing percentage is more than 1%, use MICE for imputation
+   imputed_data <- mice(bank_data, m = 5, method = "pmm", seed = 123)
+   clean_data <- complete(imputed_data)
+   print("Missing values imputed using MICE.")
+ } else {
+   # Otherwise, remove rows with missing values
+   clean_data <- na.omit(bank_data)
+   print("Rows with missing values removed.")
+ }
[1] "Rows with missing values removed."
> 
> #
> clean_data_msiing_values <- colSums(is.na(clean_data)) / nrow(clean_data) * 100
> clean_data_msiing_values
      age       job   marital education   default   balance   housing      loan   contact       day     month  duration  campaign 
        0         0         0         0         0         0         0         0         0         0         0         0         0 
    pdays  previous  poutcome         y 
        0         0         0         0 
> |
```

Figure 8 : Removing missing values form the dataset

In the above figure the na.omit() function is used to remove the rows with the missing values in R. After the removing the rows with the missing values, We have checked the missing values for each column is 0. So, it has no missing values.

Outliers:

To detect outliers, I used univariate analysis with the Interquartile Range (IQR) method as well as boxplots to visualize the outliers in each numerical column.

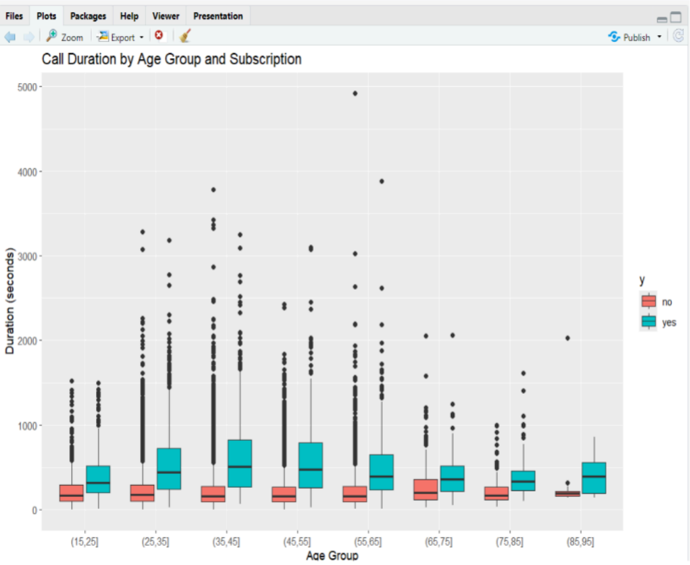
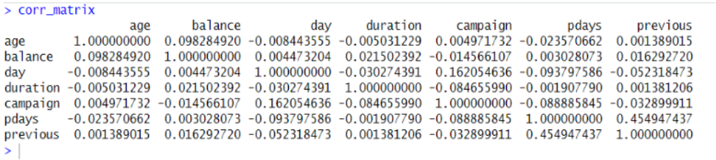


Figure 9 : Boxplot to detect outliers

The figure illustrates call duration (measured in seconds) by age category for subscribers ("yes," teal) and non-subscribers ("no," red). Across every age range, subscribers have longer call durations when compared to non-subscribers. The median call duration for subscribers is highest among middle-aged age groups (35-45, 45-55). The age group 85-95 has the greatest difference in median call duration for subscribers vs. non-subscribers, while the age group 75-85 has the lowest call duration for both subscribers and non-subscribers. There are outliers, including a subscriber from the 55-65 age group with a call duration of nearly 5000 seconds (over 80 minutes), which the graph indicates are extreme examples. Call durations are positively skewed, with most call durations being relatively short and a few call durations being exceptionally long, skewing the distribution to the right.

Multicollinearity:



```
> corr_matrix
                  age      balance          day     duration     campaign        pdays     previous
age       1.000000000  0.098284920 -0.008443555 -0.005031229  0.004971732 -0.023570662  0.001389015
balance   0.098284920  1.000000000  0.004473204  0.021502392 -0.014566107  0.003028073  0.016292720
day      -0.008443555  0.004473204  1.000000000 -0.030274391  0.162054636 -0.093797586 -0.052318473
duration -0.005031229  0.021502392 -0.030274391  1.000000000 -0.084655990 -0.001907790  0.001381206
campaign  0.004971732 -0.014566107  0.162054636 -0.084655990  1.000000000 -0.088885845 -0.032899911
pdays    -0.023570662  0.003028073 -0.093797586 -0.001907790 -0.088885845  1.000000000  0.454947437
previous  0.001389015  0.016292720 -0.052318473  0.001381206 -0.032899911  0.454947437  1.000000000
> |
```

Figure 10: corelation between the variables.

Figure 10 shows the correlation between most of the variables is moderate and does not show a strong collinearity with each other. So, no need it.

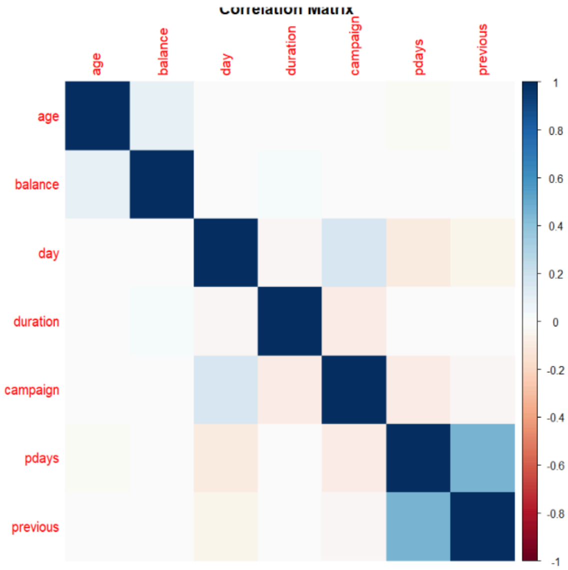


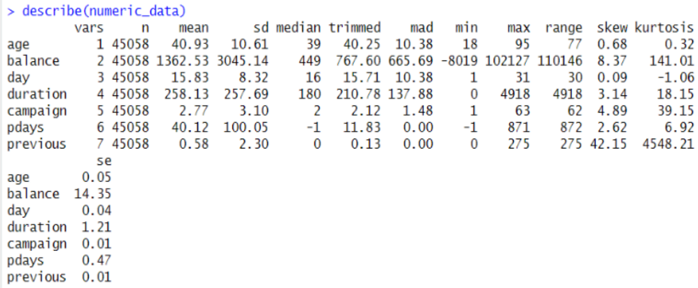
Figure 11 : Heatmap

A heatmap correlation matrix is a useful tool that represents how different banking variables are dependent on each other.

One may notice that every variable is perfectly related to itself, and it is depicted with the dark blue colour shaping a diagonal of the correlation matrix. However, most variables have weak or even no correlation as it is displayed via the light-coloured squares, which indicates the non-influence of each other. The most potent relation in "pdays" and "previous" is the one with a moderate positive correlation (around 0.6), therefore, it indicates that they tend to move in the same direction. Also, there is a weak positive correlation (around 0.2-0.3) between "day" and ”campaign” . Some pairs such as "duration" and "campaign" show the reverse relationship, meaning they have very weak negative correlations (light pink). All in all, high multicollinearity presents no challenge in the dataset, with the majority of the variables yielding fresh data.

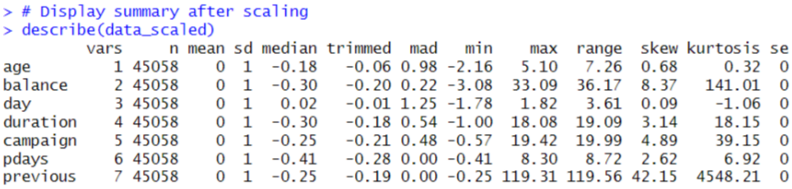
Scaling:

The given data needs to be scaled it can be done through the scale() function.



```
> describe(numeric_data)
         vars     n    mean      sd median trimmed    mad   min    max  range  skew kurtosis
age         1 45058   40.93   10.61     39   40.25  10.38    18     95     77  0.68     0.32
balance     2 45058 1362.53 3045.14    449  767.60 665.69 -8019 102127 110146  8.37   141.01
day         3 45058   15.83    8.32     16   15.71  10.38     1     31     30  0.09    -1.06
duration    4 45058  258.13  257.69    180  210.78 137.88     0   4918   4918  3.14    18.15
campaign    5 45058    2.77    3.10      2    2.12   1.48     1     63     62  4.89    39.15
pdays       6 45058   40.12  100.05     -1   11.83   0.00    -1    871    872  2.62     6.92
previous    7 45058    0.58    2.30      0    0.13   0.00     0    275    275 42.15  4548.21
            se
age       0.05
balance  14.35
day       0.04
duration  1.21
campaign  0.01
pdays     0.47
previous  0.01
```

Figure 12: Data summary before scaling.



```
> # Display summary after scaling
> describe(data_scaled)
         vars     n mean sd median trimmed  mad   min    max  range  skew kurtosis se
age         1 45058    0  1  -0.18   -0.06 0.98 -2.16   5.10   7.26  0.68     0.32  0
balance     2 45058    0  1  -0.30   -0.20 0.22 -3.08  33.09  36.17  8.37   141.01  0
day         3 45058    0  1   0.02   -0.01 1.25 -1.78   1.82   3.61  0.09    -1.06  0
duration    4 45058    0  1  -0.30   -0.18 0.54 -1.00  18.08  19.09  3.14    18.15  0
campaign    5 45058    0  1  -0.25   -0.21 0.48 -0.57  19.42  19.99  4.89    39.15  0
pdays       6 45058    0  1  -0.41   -0.28 0.00 -0.41   8.30   8.72  2.62     6.92  0
previous    7 45058    0  1  -0.25   -0.19 0.00 -0.25 119.31 119.56 42.15  4548.21  0
```

Figure 13 : Data after scaling

Figure 12 shows the data summary of the data before scaling. After using the scale() function in the dataset the mean becomes 0 and the standard deviation becomes 1 as seen in figure 13 above.

References:

1. Singh, D. (2019) *R: A programming language for data analytics*. SSRN Electronic Journal. Available at: https://www.academia.edu/93310135/R_Language_for_Data_Analytics (Accessed: 24 March 2025).

2. Fan, B. (2023) 'Banking and finance marketing strategies in the age of big data', *SHS Web of Conferences*, 154, 02018. Available at: https://doi.org/10.1051/shsconf/202315402018 (Accessed: 24 March 2025).

3. Nordmann, E. et al. (2022) 'Data visualization in R: A beginner's guide for non-R users', *Advances in Methods and Practices in Psychological Science*, 5(2), pp. 1–36. Available at: https://doi.org/10.1177/25152459221087691 (Accessed: 24 March 2025).

4. Brownstein, N.C., Adolfsson, A. and Ackerman, M. (2019) 'Clusterability analysis using R's datasets package', *Data in Brief*, 25, 104004. Available at: https://doi.org/10.1016/j.dib.2019.104004 (Accessed: 24 March 2025).

5. Wang, Y., Makedon, F. and Chakrabarti, A. (2004) 'R-Histograms: Spatial representation for complex objects', *Proceedings of the 12th ACM Multimedia Conference*, pp. 356–359. Available at: https://doi.org/10.1145/1027527.1027610 (Accessed: 24 March 2025).

6. Jolliffe, I.T. and Cadima, J. (2016) 'Principal component analysis: Advances and applications', *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202. Available at: https://doi.org/10.1098/rsta.2015.0202 (Accessed: 24 March 2025).

7. K, D. and Raja, S.R. (2018) 'Digital marketing's role in modern banking', *International Journal of Science Development and Research*, 3(7), pp. 130–135. Available at: https://www.ijsdr.org/papers/IJSDR1807030.pdf (Accessed: 24 March 2025).

8. Chen, L.P. (2021) *Introduction to data science: R for analysis and prediction*. Boca Raton, FL: Chapman and Hall/CRC.

9. Srinivasan, P. (2023) 'Outlier detection in R using IQR', *Medium*. Available at: https://medium.com/@psrinivasan028/steps-to-detect-outliers-using-interquartile-range-iqr-with-r-1f9ed7895d96 (Accessed: 24 March 2025).

10. Fan, B. (2023) 'Big data's impact on banking marketing strategies', *SHS Web of Conferences*, 154, PESD 2022. Available at: https://doi.org/10.1051/shsconf/202315402018 (Accessed: 24 March 2025).

11. OpenAI (2025) *ChatGPT* [AI model]. Available at: https://chat.openai.com (Accessed: 24 March 2025).

Sampurna_Simkhada_Report.docx

ORIGINALITY REPORT

SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
12%	6%	4%	10%

PRIMARY SOURCES

1	Submitted to The British College Student Paper	3%
2	Submitted to Leeds Beckett University Student Paper	1%
3	nccur.lib.nccu.edu.tw Internet Source	1%
4	www.coursehero.com Internet Source	1%
5	Submitted to Mancosa Student Paper	1%
6	statsvetforening.se Internet Source	1%
7	Submitted to Massey University Student Paper	1%
8	Submitted to University of Sydney Student Paper	1%
9	Bohan Fan. "Research on the Marketing Strategy of Banking and Finance Business Given Big Data Technology", SHS Web of Conferences, 2023 Publication	1%
10	lib.dr.iastate.edu Internet Source	1%
11	Submitted to Liverpool John Moores University Student Paper	<1%

12	digitalcommons.lsu.edu Internet Source	<1 %
13	personen.utwente.nl Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		