**Assessment Task 2: Data exploration and preparation**

**Student Name: Khushbu Oswal**

**Student Number: 14098887**

**1a. Initial Data Exploration**

1. **Identification of Attribute Types**

**1.1) Attribute Name:** age

**Attribute Type:** Ratio

**Justification -** Each value represents a measurement on a continuous scale and numerical representation of a quantity. A zero indicates an absence or a lack of quantity that can be compared. **The value of age is a ratio variable because it has a value of zero.** **As seen in the histogram (Figure2) and boxplot (Figure4), the most common age range recorded in the dataset is 26 to 34.**

**1.2) Attribute Name:**job

**Attribute Type:**Nominal

**Justification - Data** is non-quantitative, meaning it does not have a number, and it is qualitative data. The job attribute type is nominal. It delivers only descriptive information to differentiate given values from the other values. There is inadequate information to order values from highest to lowest and vice versa. It can, therefore, neither be ranked nor compared. The data can only be categorized. For example, we cannot compare someone who is retired from work or a career with a technician.

**1.3) Attribute Name:**marital

**Attribute Type:**Nominal

**Justification - The** Marital attribute type is nominal. Data is non-quantitative, meaning it does not have a number, and it is qualitative data. The data is categorized into these three categories married status, divorce status and single status. It delivers only descriptive information to differentiate given values from the other values. There is inadequate information to order values from highest to lowest and vice versa. For example, divorced or single people cannot be compared; if they are single, it does not mean they will qualify for higher preferences than someone who is divorced. It can, therefore, neither be ranked nor compared.

**1.4) Attribute Name:**education

**Attribute Type:**Ordinal

**Justification -**The education level (university. degree > professional. course > high. school > basic.9y > basic.6y > basic.4y.) is an ordinal categorical data. Classification is based on quantitative data samples. It involves placing descriptive information into order values from highest to lowest and vice versa. Therefore, it has adequate information to rank and compare. For example, a person with a university degree will get higher preferences in the market than a high school certificate.

**1.5) Attribute Name:**default

**Attribute Type:**Nominal

**Justification -**Data is non-quantitative, meaning it does not have a number, and it is qualitative data. The default attribute type is nominal. It delivers only descriptive information to differentiate given values from the other values. The data can only be categorized. For example, It is categorized as - No/unknown (not sure whether it is a yes or no). There is inadequate information to order values from highest to lowest and vice versa. It can, therefore, neither be ranked nor compared.

**1.6) Attribute Name:**housing

**Attribute Type:**Nominal

**Justification -**The housing attribute type is nominal. Data is non-quantitative, meaning it does not have a number, and it is qualitative data. The data is categorized into these three categories yes/no. It delivers only descriptive information to differentiate given values from the other values. The data shows the number of individuals who have a housing loan and who do not have a loan. It provides inadequate information to order values from highest to lowest and vice versa. It can, therefore, neither be ranked nor compared.

**1.7) Attribute Name:**loan

**Attribute Type:**Nominal

**Justification -**Data is non-quantitative, meaning it does not have a number, and it is qualitative data. The loan attribute type is nominal. It delivers only descriptive information to differentiate given values from the other values. It is categorized into no/yes. The data shows the number of individuals who have a personal loan and who do not have a loan. It provides inadequate information to order values from highest to lowest and vice versa. It can, therefore, neither be ranked nor compared.

**1.8) Attribute Name:**contact

**Attribute Type:**Nominal

**Justification -**The contact attribute type is nominal. Data is non-quantitative, meaning it does not have a number, and it is qualitative data. The data is categorized into these two categories telephone/cellular. It delivers only descriptive information to differentiate given values from the other values. A preference of contact is only labelled. It provides inadequate information to order values from highest to lowest and vice versa. It can, therefore, neither be ranked nor compared.

**1.9) Attribute Name:**month

**Attribute Type:**Nominal

**Justification -**Data is non-quantitative, meaning it does not have a number, and it is qualitative data. The month attribute type is nominal. It delivers only descriptive information to differentiate given values from the other values. It is categorized into Jul/Aug/Dec/Sep/May/Jun/Oct/Nov/ Mar/Apr. It provides inadequate information to order values from highest to lowest and vice versa. It can, therefore, neither be ranked nor compared.

**1.10) Attribute Name:**day\_of\_week

**Attribute Type:**Nominal

**Justification -**The day\_of\_week attribute type is nominal. Data is non-quantitative, meaning it does not have a number, and it is qualitative data. The data is categorized into Fri/Mon/Thu/Tue/Wed. It delivers only descriptive information to differentiate given values from the other values. It provides inadequate information to order values from highest to lowest and vice versa. It can, therefore, neither be ranked nor compared.

**1.11) Attribute Name:**duration

**Attribute Type:**Interval

**Justification -**Data isquantitative (numerical) data. Data follows some ordered scale when divided into categories and separated using an equal distance measure. Moreover, interval data lacks a 'true zero,' meaning a measure of zero on an interval scale does not represent absence. As a result, zero is simply another variable on an interval scale. For example, the fact that someone has not been contacted today does not mean he has not been contacted in the last few days. However, we can count the last contact duration (maybe in seconds) of the last few days. Therefore, it will include the last contact duration.

**1.12) Attribute Name:**campaign

**Attribute Type:**Interval

**Justification -**The campaign is an interval attribute. Data isquantitative (numerical) data. Campaign values followed some ordered scale when divided into categories and separated using an equal distance measure. Moreover, interval data lacks a 'true zero,' meaning a measure of zero on an interval scale does not represent absence. As a result, zero is simply another variable on an interval scale. For example: Contacted with the participant during the campaign.

**1.13) Attribute Name:**pdays

**Attribute Type:**Ratio

**Justification -**The value of pdays is a ratio variable. Each value represents a measurement on a continuous scale and a numerical representation of a quantity. A zero indicates an absence or a lack of quantity that can be compared. For example, an individual was not previously contacted.

**1.14) Attribute Name:**previous

**Attribute Type:**Ratio

**Justification -**The value of previous is a ratio variable. Each value represents a measurement on a continuous scale and a numerical representation of a quantity. A zero indicates an absence or a lack of quantity that can be compared. For example:  It does not include contacts with the participant before the campaign.

**1.15) Attribute Name:**poutcome

**Attribute Type:**Ordinal

**Justification -**The poutcome (with given values as success > failure > non - existent) is an ordinal categorical data. Classification is based on quantitative data samples. It involves placing descriptive information into order values from highest to lowest and vice versa. Therefore, it has adequate information to rank and compare.

**1.16) Attribute Name:**emp.var. rate

**Attribute Type:**Ratio

**Justification -**The value of emp.var.rate is a ratio variable. Each value represents a measurement on a continuous scale and a numerical representation of a quantity. A zero indicates an absence or a lack of quantity that can be compared.

**1.17) Attribute Name:**cons.price.idx

**Attribute Type:**Interval

**Justification -**The cons.price.idx is an interval attribute. Data isquantitative (numerical) data. cons.price.idx values followed some ordered scale when divided into categories and separated using an equal distance measure. Moreover, interval data lacks a 'true zero,' meaning a measure of zero on an interval scale does not represent absence. As a result, zero is simply another variable on an interval scale.

**1.18) Attribute Name:**cons.conf.index

**Attribute Type:**Ratio

**Justification - The** value of cons.conf.index is a ratio variable. Each value represents a measurement on a continuous scale and a numerical representation of a quantity. A zero indicates an absence or a lack of quantity that can be compared. For example: economic growth does not increase or decrease it remains the same.

**1.19) Attribute Name: euribor3m**

**Attribute Type:**Ratio

**Justification -**The value of euribor**3m** is a ratio variable. Each value represents a measurement on a continuous scale and a numerical representation of a quantity. A zero indicates an absence or a lack of quantity that can be compared.

**1.20) Attribute Name:**nr. employed

**Attribute Type:**Interval

**Justification -**The nr.employed is an interval attribute. Data isquantitative (numerical) data. nr.employed values followed some ordered scale when divided into categories and separated using an equal distance measure. Moreover, interval data lacks a 'true zero,' meaning a measure of zero on an interval scale does not represent absence. As a result, zero is simply another variable on an interval scale.

**1.21) Attribute Name:**subscribed

**Attribute Type:**Nominal

**Justification - Data** is non-quantitative, meaning it does not have a number, and it is qualitative data. The subscribed attribute type is nominal. It delivers only descriptive information to differentiate given values from the other values.

Subscribed - Subscribed a term deposit/ not subscribed a term deposit

There is inadequate information to order values from highest to lowest and vice versa. It can, therefore, neither be ranked nor compared. The data can only be categorized.

1. **The Summarising Properties for the Attributes**

**Frequency and distribution Data Visualisations**

2.1 Age Chart, pie chart, radar chart

Description automatically generated

Figure 1: Pie Chart for Age

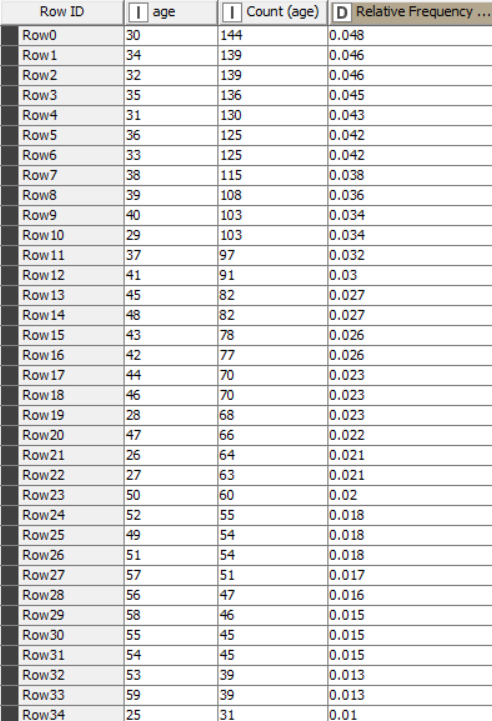
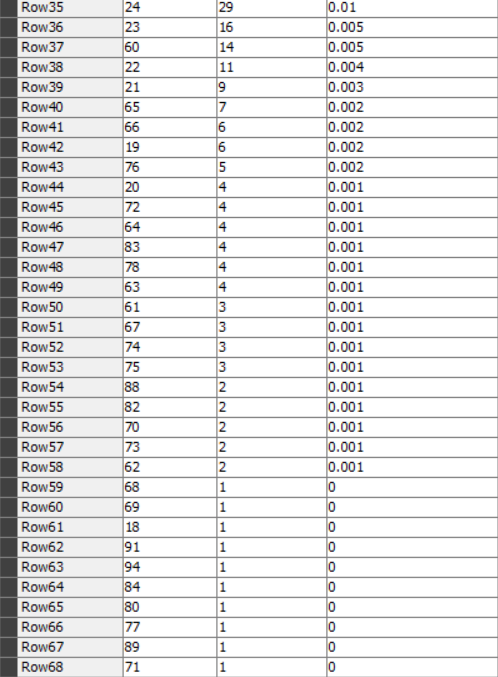
 

Figure 2: Value Count and Relative Frequency for Age

Chart, box and whisker chart

Description automatically generated

Figure 3: Box Plot for Age

Using Figures 1 and 2 above, it is clear that the highest number of occurrences of adults in their thirties is a decade or period of stability and establishing status which is 144, making up a row count of 4.8% of the age data provided. On the other hand, from the Value count table, it is evident from the data provided that the ages 18, 68, 69, 71, 77, 80, 84, 89, 91, and 94 hold the lowest number of occurrences, which is 1, making up a row count of 0.033% (1/3000 \* 100). The value count table indicates that people over 60 and under 22 are nearly the lowest occurrence number. The Histogram shows the distribution of age as well. From boxplot we depict the median of the age is 38. According to the box plot the data points that differs significantly from other data points are identified by using this formula: Outliers < Q1- 1.5\*IQR(32 – 1.5\*15 = 9.5) and Outliers > 47 – 1.5\*IQR(47 – 1.5\*15 = 24.).

**Table 1: Age**

|  |  |
| --- | --- |
| Statistics | Value |
| Mean | 39.9467 |
| Median | 38 |
| Mode | 30 |
| Range | 76 |
| Geometric Mean | 38.663 |
| Geometric Standard Deviation | 1.289 |
| Mean Absolute Deviation | 8.409 |
| Minimum Value | 18 |
| Quantile | 38 |
| Sum | 119840 |
| Median Absolute Deviation | 7 |
| Maximum Value | 94 |
| Standard Deviation | 10.4983 |
| Variance | 110.215 |
| Skewness | 0.8987 |
| Kurtosis | 1.2498 |
| Q1 | 32 |
| Q3 | 47 |
| Interquartile range | 15 |
| Percentile | 37.871 |

2.2 Job

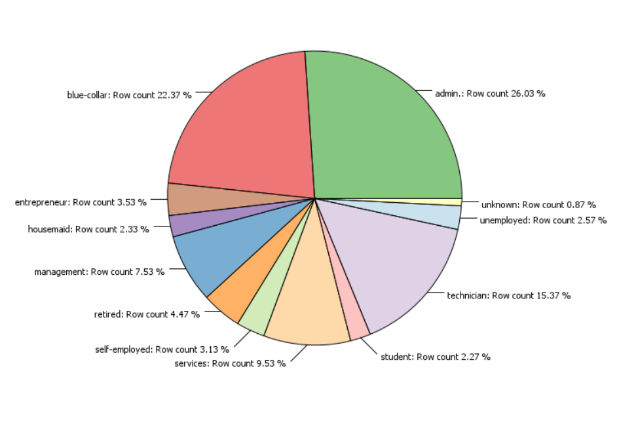


Figure 4: Pie Chart for Job

Chart, bar chart

Description automatically generated

Figure 5: Bar Chart for Job



Figure 6: Value Count and Relative Frequency for Job

The bar chart and pie chart show the information related to the occupations of people of different ages with different education levels. It is noteworthy that so many people have the admin profession, according to the dataset, which has a row count of 26.03%. Have 781 highest occurrences and a relative frequency of 0.2. It shows that 0.87% is the missing data in the dataset, with a count of 26 and a relative frequency of 0.009. The least common occupation is student and housemaid, with 68 and 70 occurrences, a row count of 2.27% and 2.33%, and a relative frequency of 0.023.

**Table 2: Job**

|  |  |
| --- | --- |
| Statistics | Value |
| Mode | admin |

2.3 Marital

Chart, pie chart

Description automatically generated

Figure 7: Pie Chart for Marital

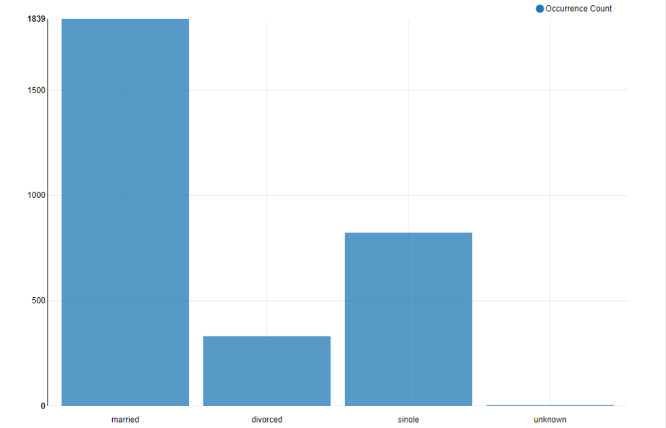


Figure 8: Bar Chart for Marital

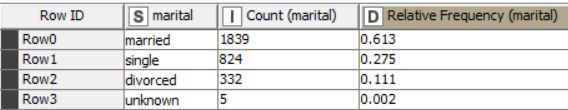


Figure 9: Value Count and Relative Frequency for Marital

The dataset contains data on four key indicators: the distribution of the population by marital status, the proportion of married, divorced, single, and unknown (not sure about the categorization of the data). Figures 7,8, and 9 replicate that married status has the highest occurrences in the dataset of 1839, making up 61.3% of row count with a relative frequency of 0.613. The anonymous data does not describe if it is divorced, married, or single. The lowest occurrence in the dataset is 332 of divorced status with a 0.111 relative frequency and 11.07%.

**Table 3: Marital**

|  |  |
| --- | --- |
| Statistics | Value |
| Mode | Married |

2.4 Education

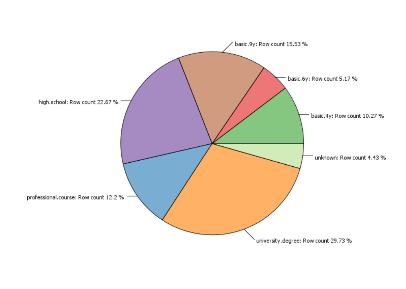


Figure 10: Pie Chart for Education

Chart, histogram

Description automatically generated

Figure 11: Bar Chart for Education

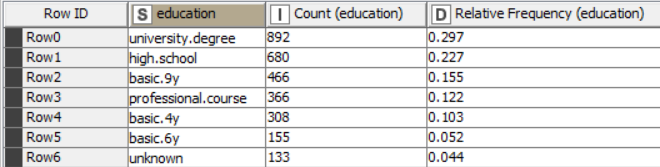


Figure 12: Value Count and Relative Frequency for Education

Figure 10,11 and 12 shows the given education level, which indicates that most of the population has a university degree, with the highest occurrence of 892, making up a row count of 29.73. In contrast, there are fewer people with basic.6y education has the lowest count of 155, making up a row count of 5.17%. There is no indication of what falls into which category in the unknown data.

Looking at the summary table we can identify that the most common education level is

university.degree. From the sample of data, we can understand that the education column is the

education that the person has completed or drop out from. Because in one of the rows of data, a person

is 40 years old but has an education of basic.6y, that means he dropped out from school at year 6. It

doesn’t make sense if the person is 40 years old and is still in Year 6.

Table 4: Education

|  |  |
| --- | --- |
| Statistics | Value |
| Mode | University.degree |

2.5 Default

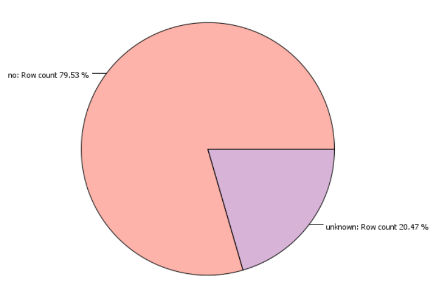


Figure 13: Pie Chart for default

Chart, bar chart, histogram

Description automatically generated

Figure 14: Bar Chart for Default

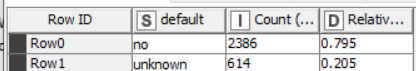


Figure 15: Value Count and Relative Frequency for Default

Figures 13, 14, and 15 above outline the individuals who have not paid payments on a loan or credit card according to the account terms. Two thousand three hundred eighty-six individuals have not made the payments and have a relative frequency of 0.795. 20.47% of row count unknown data is not defined.

2.6 Housing

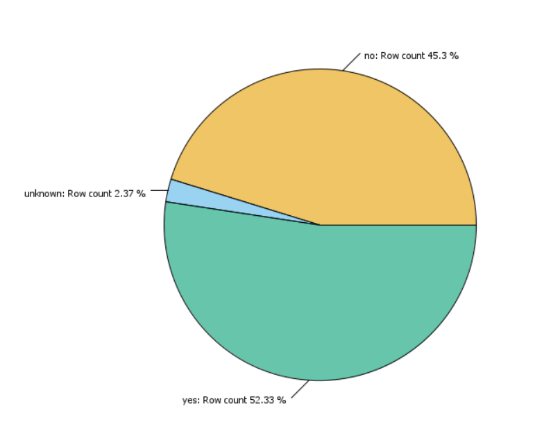


Figure 16: Pie Chart for Housing

Chart, bar chart

Description automatically generated

Figure 17: Bar Chart for Housing

Table

Description automatically generated

Figure 18: Value Count and Relative Frequency for Housing

As shown in figures 16, 17, and 18, individuals with housing loans and those without housing loans are outlined. According to this dataset, 52.33% (count = 1570) individuals have a housing loan, and 45.3% (count = 1359) individuals do not have a housing loan. Therefore, most individuals have a housing loan.

2.7 Loan

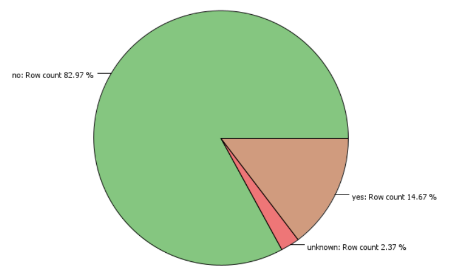


Figure 19: Pie Chart for Loan

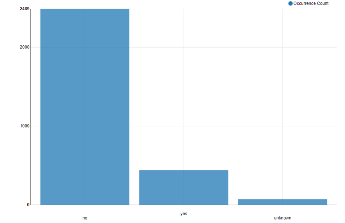


Figure 20: Bar Chart for Loan

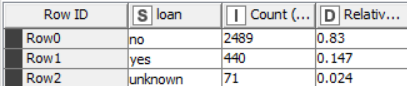


Figure 21: Value Count and Relative Frequency for Loan

As shown in figures 20, 21, and 19, individuals with personal loans and those without personal loans are outlined. According to this dataset, 14.67% (count =440) individuals have a housing loan, and 82.97% (count =2489) individuals do not have a housing loan. Therefore, most individuals do not have a personal loan

2.8 Contact



Figure 22: Pie Chart for Contact



Figure 23: Bar Chart for Contact

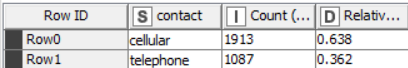


Figure 24: Value Count and Relative Frequency for Contact

As shown in figures 23, 24, and 22, it appears that the majority of participants prefer to be contacted through their cellular phones. A relative frequency of 0.638 and a count of 1913 make up a row count of 63.77% who prefers to use a cellular phone.

2.9 Month

Chart, pie chart

Description automatically generated

Figure 25: Pie Chart for Month

Chart, bar chart

Description automatically generated

Figure 26: Bar Chart for Month



Figure 27: Value Count and Relative Frequency for Month

Using Figures 25, 26, and 27 above, it is clear that the highest number of occurrences of the last contact was in May, with a 1015 count and a row count of 33.83%. On the other hand, from the Value count table and the pie chart, it is evident from the data provided that December holds the lowest number of occurrences, which is 17, making up a row count of 0.57%. The Pie chart indicates that September and October have an almost similar count of last contacted month of year which is approximately 43.

Table 4: Month

|  |  |
| --- | --- |
| Statistics | Value |
| Mode | May |

2.10 **Day\_of\_week**

Chart, pie chart

Description automatically generated

Figure 28: Pie Chart for Day\_of\_week

Chart, bar chart, histogram

Description automatically generated

Figure 29: Bar Chart for Day\_of\_week

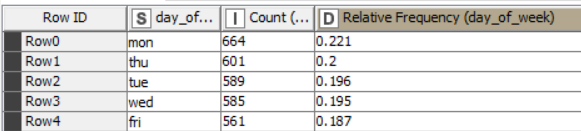


Figure 30: Value Count and Relative Frequency for Day\_of\_week

The dataset contains data on five key indicators: the distribution of the last contact of the week by Friday, Monday, Thursday, Tuesday, and Wednesday. Figures 29,30 and 28 replicate that last contact day of week that has highest occurrences on Monday with a count of 664 and making up a row count of 22.13%. The lowest occurrence in the dataset is 561 of Friday of week with a row count and relative frequency of 18.7% and 0.187.

2.11 Duration

Diagram

Description automatically generated

Figure 31: Box Plot for Duration

Figures 31 and 32 represent the duration of the last contact of individual. From boxplot we depict the median of the duration is 180. According to the box plot the data points that differs significantly from other data points are identified by using this formula: Outliers < Q1- 1.5\*IQR (106 – 1.5\*222.5 = -227.75) and Outliers > 47 – 1.5\*IQR (328.5 – 1.5\*222.5 = -5.625).

**Table 5: Duration**

|  |  |
| --- | --- |
| Statistics | Value |
| Mean | 260.527 |
| Median | 180 |
| Mode | 83 |
| Range | 3,781 |
| Geometric Mean | 176.436 |
| Geometric Standard Deviation | 2.516 |
| Mean Absolute Deviation | 173.084 |
| Minimum Value | 4 |
| Quantile | 180 |
| Sum | 781582 |
| Median Absolute Deviation | 94 |
| Maximum Value | 3785 |
| Standard Deviation | 272.059 |
| Variance | 74,016.136 |
| Skewness | 4.174 |
| Kurtosis | 33.158 |
| Q1 | 106 |
| Q3 | 328.5 |
| Interquartile range | 222.5 |
| Percentile | 180.547 |

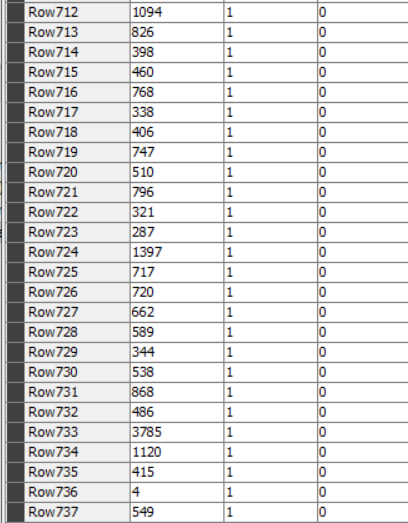
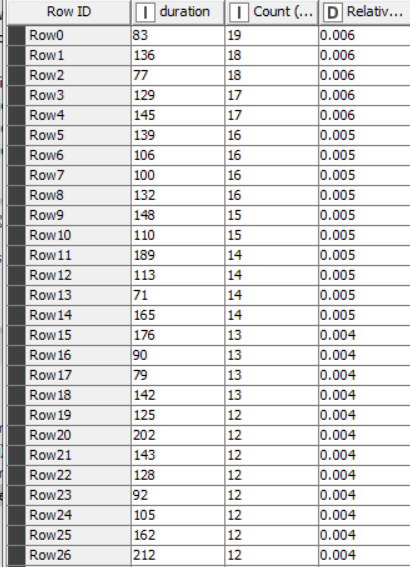


Figure 32: Value count for Duration

2.12 Campaign

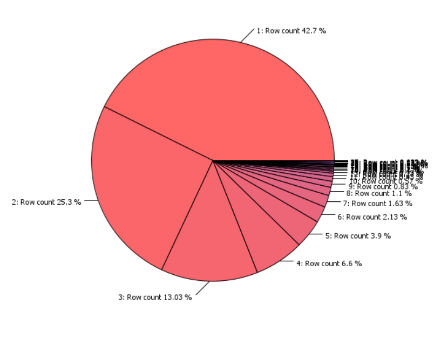


Figure 33: Pie Chart for Campaign

A picture containing chart

Description automatically generated

Figure 34: Histogram for Campaign

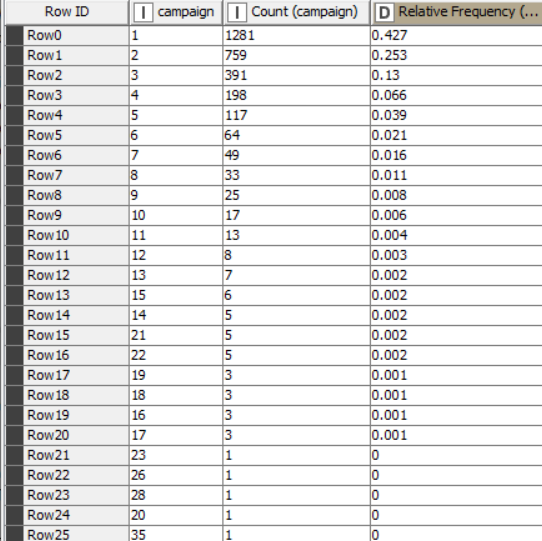


Figure 35: Value Count for Campaign

A picture containing chart

Description automatically generated

Figure 36: Box Plot for Campaign

Using Figures 33 and 36above, it is clear that the highest number of occurrences is in the campaign 1 with a count of 1281 and a relative frequency of 0.427. The Histogram shows the distribution of campaign with a highest number is in the interval 0-3. From boxplot we depict the median of the campaign is 2. Outliers < Q1- 1.5\*IQR(1 – 1.5\*2 = -2) and Outliers > 47 – 1.5\*IQR(3 – 1.5\*2 = 0).

**Table 6: Campaign**

|  |  |
| --- | --- |
| Statistics | Value |
| Mean | 2.594 |
| Median | 2 |
| Mode | 1 |
| Range | 34 |
| Geometric Mean | 1.929 |
| Geometric Standard Deviation | 2.011 |
| Mean Absolute Deviation | 1.662 |
| Minimum Value | 1 |
| Quantile | 2 |
| Sum | 7782 |
| Median Absolute Deviation | 1 |
| Maximum Value | 35 |
| Standard Deviation | 2.736 |
| Variance | 7.486 |
| Skewness | 4.049 |
| Kurtosis | 24.289 |
| Q1 | 1 |
| Q3 | 3 |
| Interquartile range | 2 |
| Percentile | 1.984 |

**2.13 Pdays**

A picture containing chart

Description automatically generated

Figure 37: Pie Chart for Pdays

Table

Description automatically generated

Figure 38 : Value count for Pdays

Figures 37 and 38 represent the number of days that pass by after the client was last contacted from a previous campaign. The majority of clients (2891) are contacted after 999 days from a previous campaign making up a row count of 96.37% and a relative frequency of 0.964. One client was contacted after 11 days had passed from a previous campaign.

**Table 7: Pdays**

|  |  |
| --- | --- |
| Statistics | Value |
| Mean | 962.9197 |
| Median | 999 |
| Mode | 999 |
| Range | 998 |
| Geometric Mean | 823.831 |
| Geometric Standard Deviation | 2.718 |
| Mean Absolute Deviation | 69.539 |
| Minimum Value | 1 |
| Quantile | 999 |
| Sum | 288759 |
| Median Absolute Deviation | 0 |
| Maximum Value | 999 |
| Standard Deviation | 185.848 |
| Variance | 34,539.33 |
| Skewness | -4.958 |
| Kurtosis | 22.602 |
| Percentile | 998.998 |

**2.14 Previous**

Chart, pie chart

Description automatically generated

Figure 39: Pie Chart for Previous

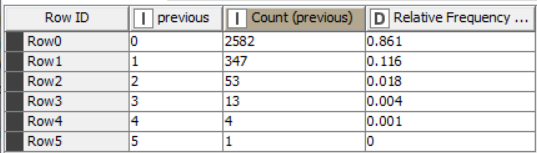


Figure 40: Value Count for Previous

Figure 39 and 40 depicts the number of contacts performed before the campaign. The highest number of contacts performed before the campaign is five, and contact with only one client. The lowest number of contacts is 0, meaning 2582 clients were not contacted before the campaign. This describes the campaign's failure because many individuals were not contacted before the campaign.

**Table 8: Previous**

|  |  |
| --- | --- |
| Statistics | Value |
| Mean | 0.171 |
| Median | 0 |
| Mode | 0 |
| Range | 5 |
| Geometric Mean | 0 |
| Mean Absolute Deviation | 0.294 |
| Minimum Value | 0 |
| Quantile | 0 |
| Sum | 513 |
| Median Absolute Deviation | 0 |
| Maximum Value | 5 |
| Standard Deviation | 0.475 |
| Variance | 0.226 |
| Skewness | 3.546 |
| Kurtosis | 16.621 |
| Percentile | 0.002 |

**2.15 Poutcome**

Chart, pie chart

Description automatically generated

Figure 41: Pie Chart for Poutcome

Chart, bar chart

Description automatically generated

Figure 42: Bar Chart for Poutcome

Table

Description automatically generated

Figure 43: Value Count for Poutcome

Figures 41 and 42, and 43 show that the previous marketing campaign failed. This is also justified above with the help of the previous attribute. Most (86.07%) of the poutcome attribute result is non-existent. Success received in the previous marketing campaign was 3.17% only.

**2.16 Emp.var.rate**

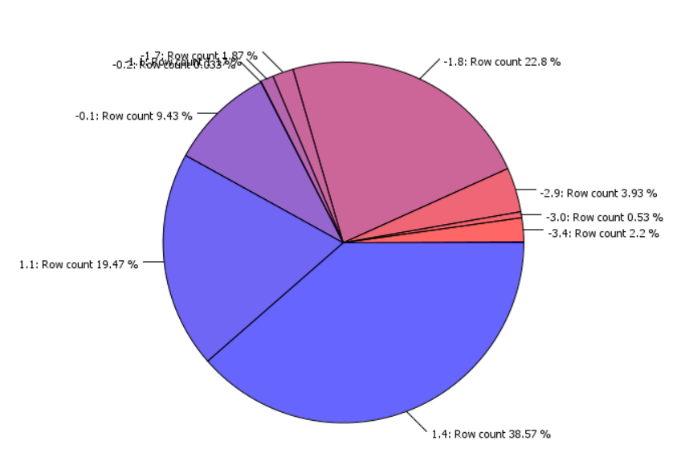


Figure 44: Pie Chart for Emp.var.rate

Chart, box and whisker chart

Description automatically generated

Figure 45: Box Plot for Emp.var.rate

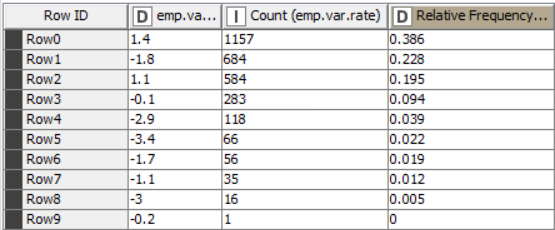


Figure 46: Value Count for Emp.var.rate

The distribution of employment variation rate is shown by value count, Box plot, and pie chart. Using figures 44 and 45, we depict that when the employment variation rate is high, the count is 1157 with a relative frequency of 0.386. The median of the employment variation rate is 1.1 by analysing the boxplot.

**Table 9: Emp.var.rate**

|  |  |
| --- | --- |
| Statistics | Value |
| Mean | 0.085 |
| Median | 1.1 |
| Mode | 1.4 |
| Range | 4.8 |
| Mean Absolute Deviation | 1.41 |
| Minimum Value | -3.4 |
| Quantile | 1.1 |
| Sum | 254.2 |
| Median Absolute Deviation | 0.3 |
| Maximum Value | 1.4 |
| Standard Deviation | 1.558 |
| Variance | 2.426 |
| Skewness | -0.717 |
| Kurtosis | -1.079 |
| Q1 | -1.8 |
| Q3 | 1.4 |
| Interquartile range | 3.2 |
| Percentile | 0.684 |
|  |  |

**2.17 Cons.price.idx**

Chart, pie chart

Description automatically generated

Figure 47: Pie chart for cons.price.idx

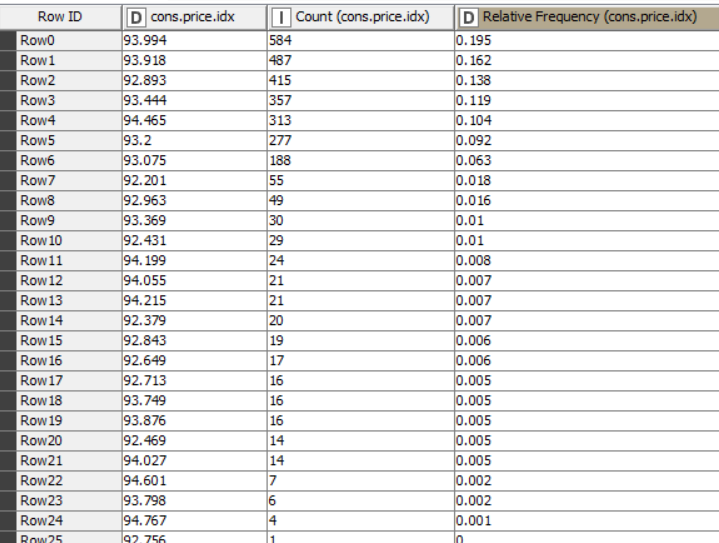


Figure 48: Value Count for cons.price.idx

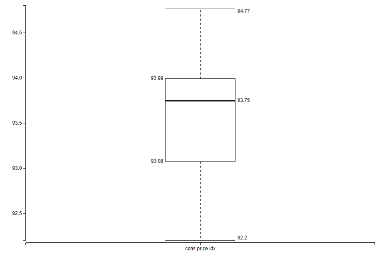


Figure 49: Box Plot for cons.price.idx

The distribution of consumer price index is shown by value count, Box plot, and pie chart. Using figures 47,48 and 49, the highest consumer price index is 94.767 with a count of 4 and relative frequency of 0.001. And the median depicted from the boxplot is 93.749. The highest number of occurrences is 584 with a consumer price index of 93.994.

**Table 10: cons.price.idx**

|  |  |
| --- | --- |
| Statistics | Value |
| Mean | 93.575 |
| Median | 93.749 |
| Mode | 93.994 |
| Range | 2.566 |
| Mean Absolute Deviation | 0.504 |
| Minimum Value | 92.201 |
| Quantile | 93.749 |
| Sum | 280,724.768 |
| Median Absolute Deviation | 0.306 |
| Maximum Value | 94.767 |
| Standard Deviation | 0.571 |
| Variance | 0.326 |
| Skewness | -0.25 |
| Kurtosis | -0.821 |
| Q1 | 93.08 |
| Q3 | 93.99 |
| Interquartile range | 0.91 |
| Percentile | 93.634 |

**2.18 Cons.conf. idx**

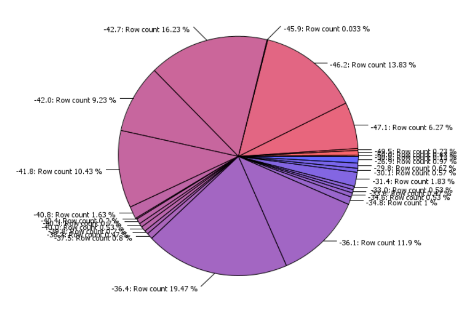


Figure 50: Pie chart for cons.conf.idx

Chart, box and whisker chart

Description automatically generated

Figure 51: Box Plot for cons.conf.idx

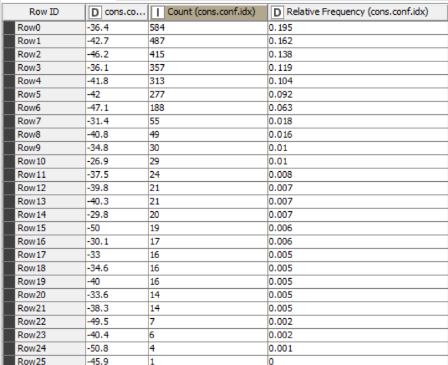


Figure 52: Value Count for cons.conf.idx

The pie chart, box plot, and value count show an individual's economic growth decrement in this dataset. Most (584) individual has a consumer confidence index of -36.4. The lowest consumer confidence index (consumption of money and spending money) is -50.8. In a comparison of the other data point, -26.9 is the highest consumer confidence index, with a count of 29. Outliers < Q1- 1.5\*IQR(-42.7 – 1.5\*6.3 = -52.15) and Outliers > 47 – 1.5\*IQR(-36.4 – 1.5\*6.3 = -45.85).

**Table 11: cons.conf.idx**

|  |  |
| --- | --- |
| Statistics | Value |
| Mean | -40.481 |
| Median | -41.8 |
| Mode | -36.4 |
| Range | 23.9 |
| Mean Absolute Deviation | 3.893 |
| Minimum Value | -50.8 |
| Quantile | -41.8 |
| Sum | -121,441.7 |
| Median Absolute Deviation | 3.893 |
| Maximum Value | -26.9 |
| Standard Deviation | 4.55 |
| Variance | 20.705 |
| Skewness | 0.292 |
| Kurtosis | -0.421 |
| Q1 | -42.7 |
| Q3 | -36.4 |
| Interquartile range | 6.3 |
| Percentile | -39.962 |

**2.19 euribor3m**

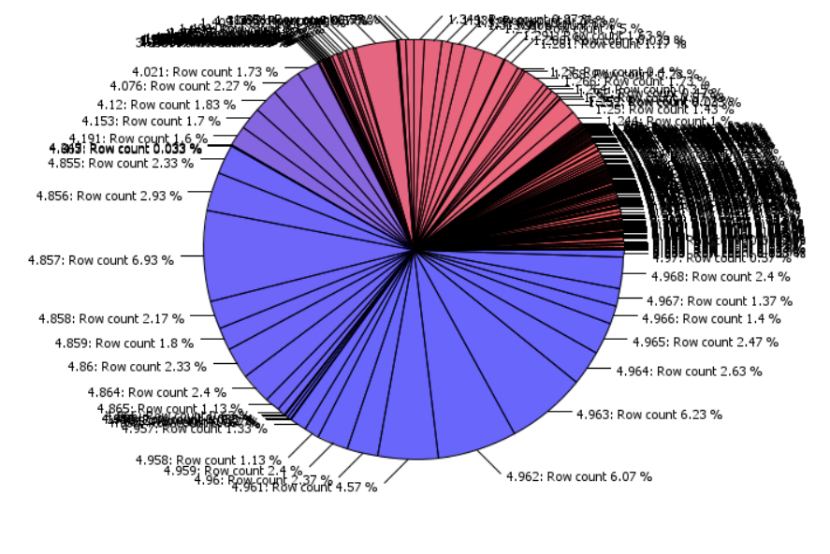


Figure 53: Pie chart for euribor3m

Chart, bar chart

Description automatically generated

Figure 54: Histogram for euribor3m

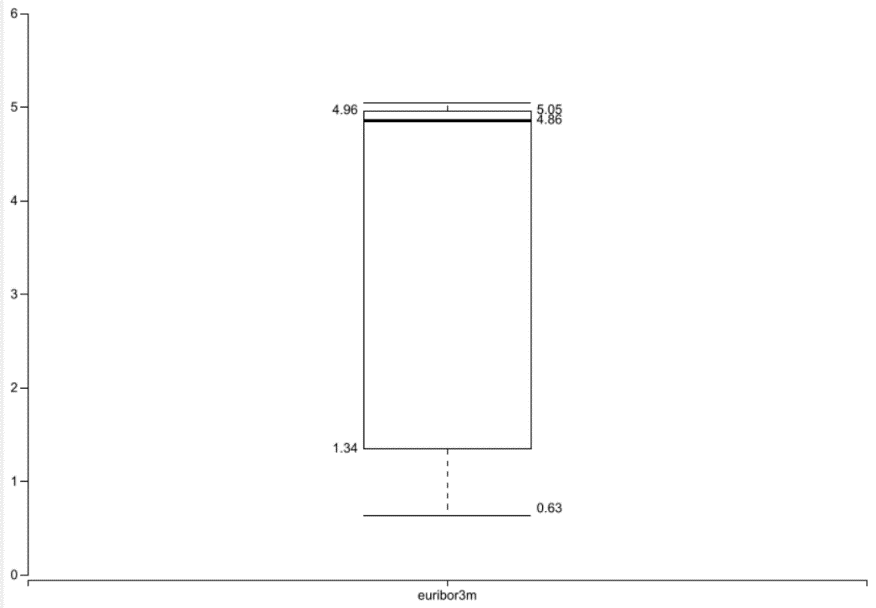


Figure 55: Box Plot for euribor3m

The highest number of counts of the rate for deposits in euros for three months is 208, with a relative frequency of 0.069.The histogram shows the highest distribution between 4.621 and 5.064. The Box Plot shows the median -41.8 of Euribor and has a range of 23.9.

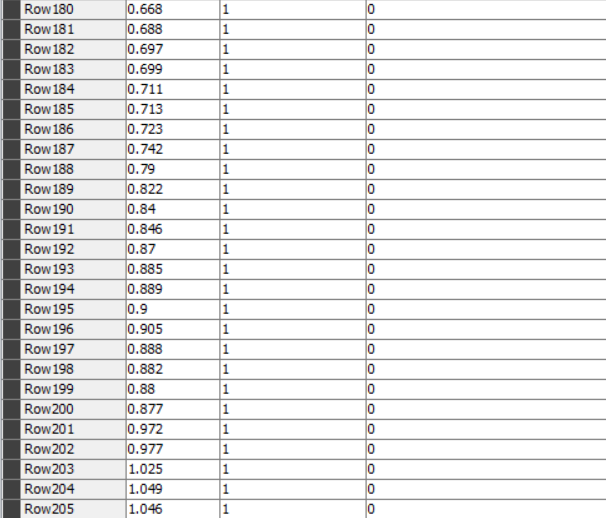
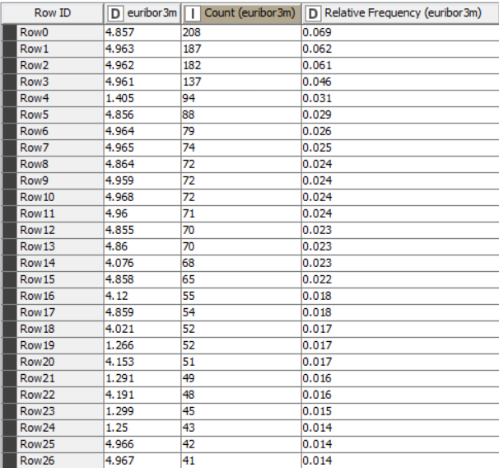


Figure 56: Value Count for euribor3m

**Table 12: euribor3m**

|  |  |
| --- | --- |
| Statistics | Value |
| Mean | -40.481 |
| Median | -41.8 |
| Mode | -36.4 |
| Range | 23.9 |
| Mean Absolute Deviation | 3.893 |
| Minimum Value | -50.8 |
| Quantile | -41.8 |
| Sum | -121,441.7 |
| Median Absolute Deviation | 3.893 |
| Maximum Value | -26.9 |
| Standard Deviation | 4.55 |
| Variance | 20.705 |
| Skewness | 0.292 |
| Kurtosis | -0.421 |
| Q1 | -42.7 |
| Q3 | -36.4 |
| Interquartile range | 6.3 |
| Percentile | 4.4 |

**2.20 nr. employed**

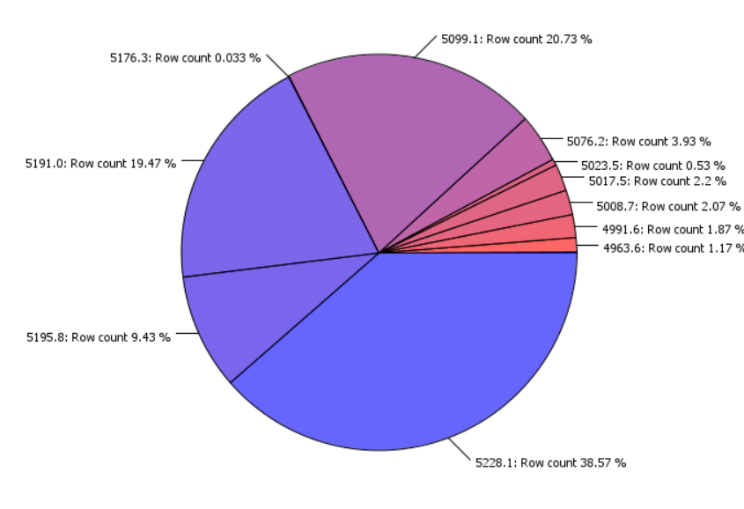


Figure 57: Pie Chart for nr.employed

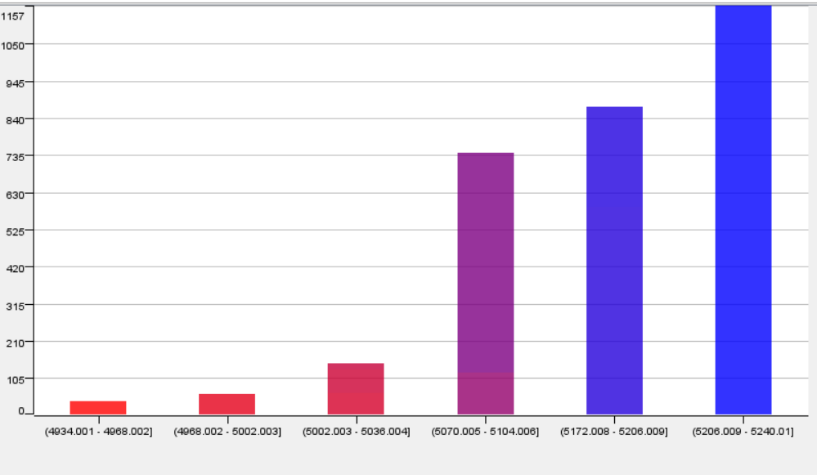


Figure 58: Histogram for nr.employed

Chart, box and whisker chart

Description automatically generated

Figure 59: Box Plot for nr. Employed



Figure 60: Value Count for nr.employed

Figures 58, 59, 60, and 57 show the dataset's number of employments. The highest number of counts is 1157, and the number of employees is 5,228.1. The histogram shows the distribution of the dataset. The highest bar is between the interval 5206 – 5240. The boxplot's median is 5191, and the range is 264.5.

**Table 13: nr. employed**

|  |  |
| --- | --- |
| Statistics | Value |
| Mean | 5167.333733333308 |
| Median | 5191.0 |
| Mode | 5228.1 |
| Range | 264.5 |
| Mean Absolute Deviation | 61.461726666666685 |
| Minimum Value | 4963.6 |
| Quantile | 5191.0 |
| Sum | 15,502,001.2 |
| Median Absolute Deviation | 37.1 |
| Maximum Value | 5228.1 |
| Standard Deviation | 71.33236744919029 |
| Variance | 5088.306645906301 |
| Skewness | -1.039 |
| Kurtosis | -0.02 |
| Q1 | 5099.1 |
| Q3 | 5228.1 |
| Interquartile range | 129 |
| Percentile | 5187.597654915942 |

**2.21 Subscribed**

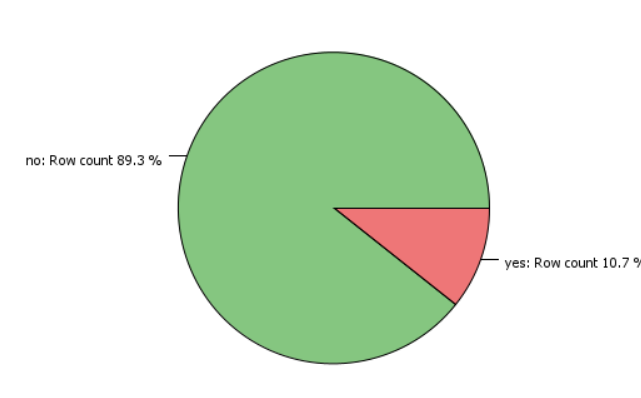


Figure 61: Pie Chart for Subscribed

Chart, bar chart

Description automatically generated

Figure 62: Bar Graph for Subscribed

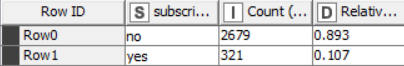


Figure 63: Value Count for Subscribed

Figures 61, 62, and 63 above outline the individuals who have subscribed to the term deposit and who have not subscribed to the term deposit. The number of individuals who have subscribed to the term deposit is 321(10.7%), and those who have not is 2679(89.3%). Most of the individuals have not subscribed a term deposit.

1. **Exploration of Dataset**

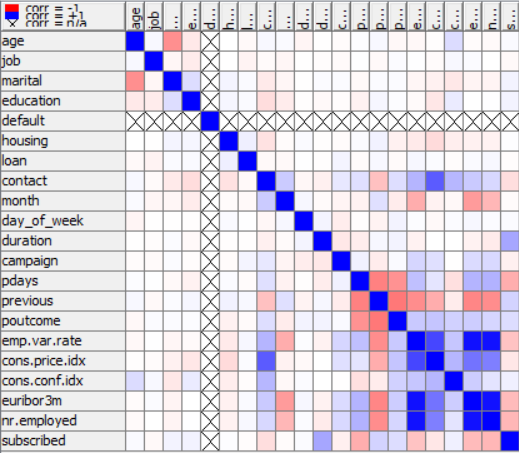


Figure 64- Rank correlation matrix for the given dataset

The Spearman correlation matrix measures the monotonic association between two dataset attributes. It measures whether one variable increases or decreases with another. Blue squares represent the positive correlation, while red squares represent the negative correlation. The Cross sign represents the missing value, but in the matrix, as shown, there is no cross sign.

Positive correlation between the number of employees and employment variation rate, which states the strong relationship between these two attributes. The consumer confidence index and employment variation rate are directly proportional. Plays and previous as a given attribute of the dataset describe the negative correlation between them. The number of days that passed by after the client was last contacted is inversely proportional to the number of contacts performed before the campaign. Missing values are found in the dataset. Weak relationship between the number of contacts performed during the campaign and the outcome of the previous marketing campaign. Overall negative correlation (inversely proportional relationship) between the attribute is more than the attribute which are directly proportional to each other.

Graphical user interface

Description automatically generated

Figure 65 - Scatter Matrix for age marital and housing

The scatter matrix shows the relationship between the age, marital and housing. The married individuals below the age of 60 do have housing loan and personal loan. The people above the age of 60 does not have housing and personal loan. Thus the relation between the marital and age is inversely proportional to each other according to the rank correlation matrix.

Chart

Description automatically generated

Figure 66 - Scatter Matrix for previous and poutcome

The scatter matrix shows the relationship between the number of contacts performed before the campaign and the outcome of the previous marketing campaign. From the plot, it is expected that if the number of contacts with the participants is low, then the campaign outcome will be low. Therefore, it shows the negative correlation between the previous and poutcome of the dataset according to the rank correlation matrix.

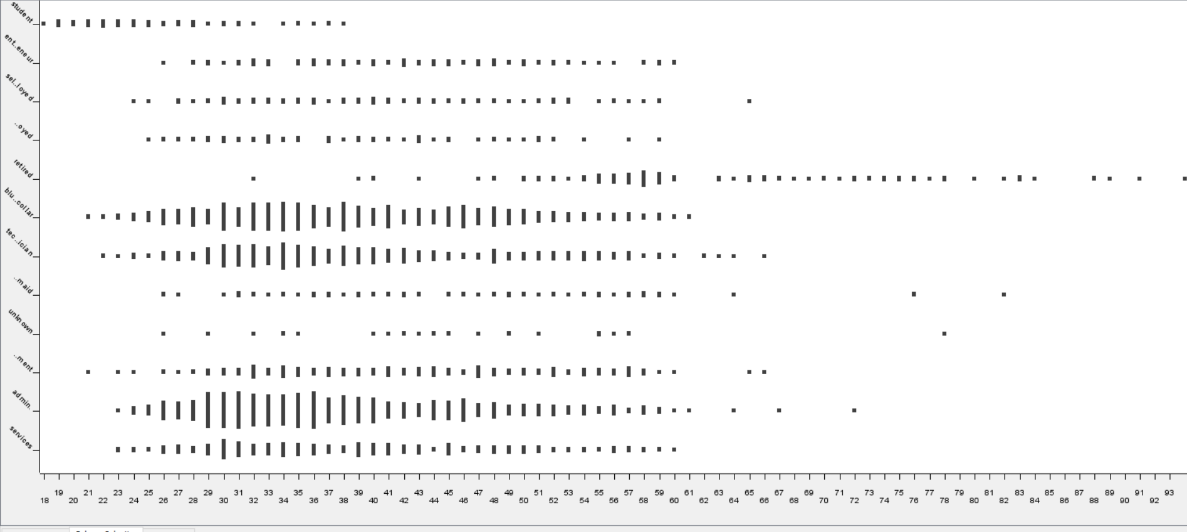


Figure 67 - Scatter plot for age and job

Figure 67 above shows the relationship between the age and job of an individual. The distribution of the student in this dataset is below the age of 39. Based on the scatter plot, it is evident that most individuals retired at the age of 45 or older. From the mid-60s onward, people stop working based on this dataset.

Table

Description automatically generated

Figure 68 - Scatter plot for age and education

Figure 68 replicates the relationship between the age and education of an individual. From the plot, we can see that the age range for the education level is basic.4y is highly scattered on the plot. Some of the data points do not make sense as it shows that the people in their mid–age or old – age still have an education level of basic.4y.

Graphical user interface

Description automatically generated

Figure 69 - Scatter Matrix for age duration and job

This plot shows that people below 60 who do different jobs have a long duration of contact because they do work. On the other hand, people above the age of 60 do not have that much-extended contact duration because they are retired from their fields, so they are not contacted related to job matters. The scatter matrix shows the relationship between age, duration, and job. The data points are more scattered on the plot showing the relationship between the last contact duration of the individual and the job of the individual below the age of 60.

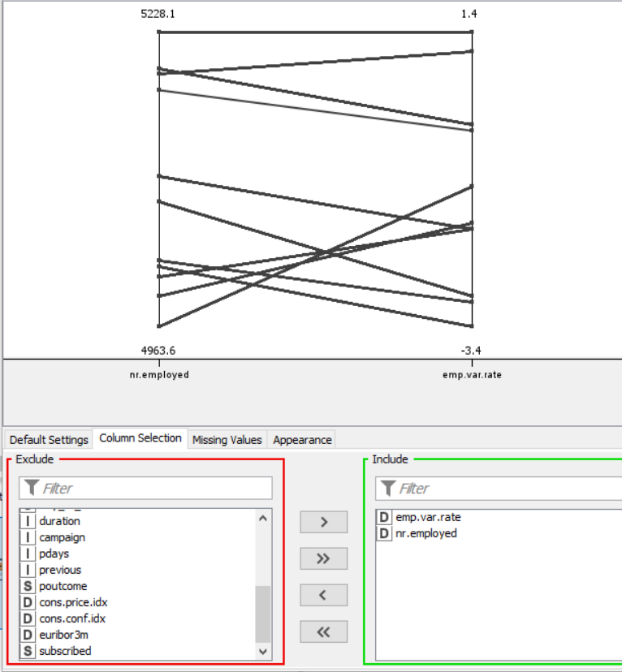


Figure 70 – Parallel coordinates for emp.var.rate and nr.employed

Figure 70 shows the direct relationship between the employment variation rate and the number of employees in the given dataset. It states that these two attributes are directly proportional to each other. Employment variation rate increases as the number of employees increases. Thus, this shows the positive correlation between these attributes according to the rank correlation matrix.

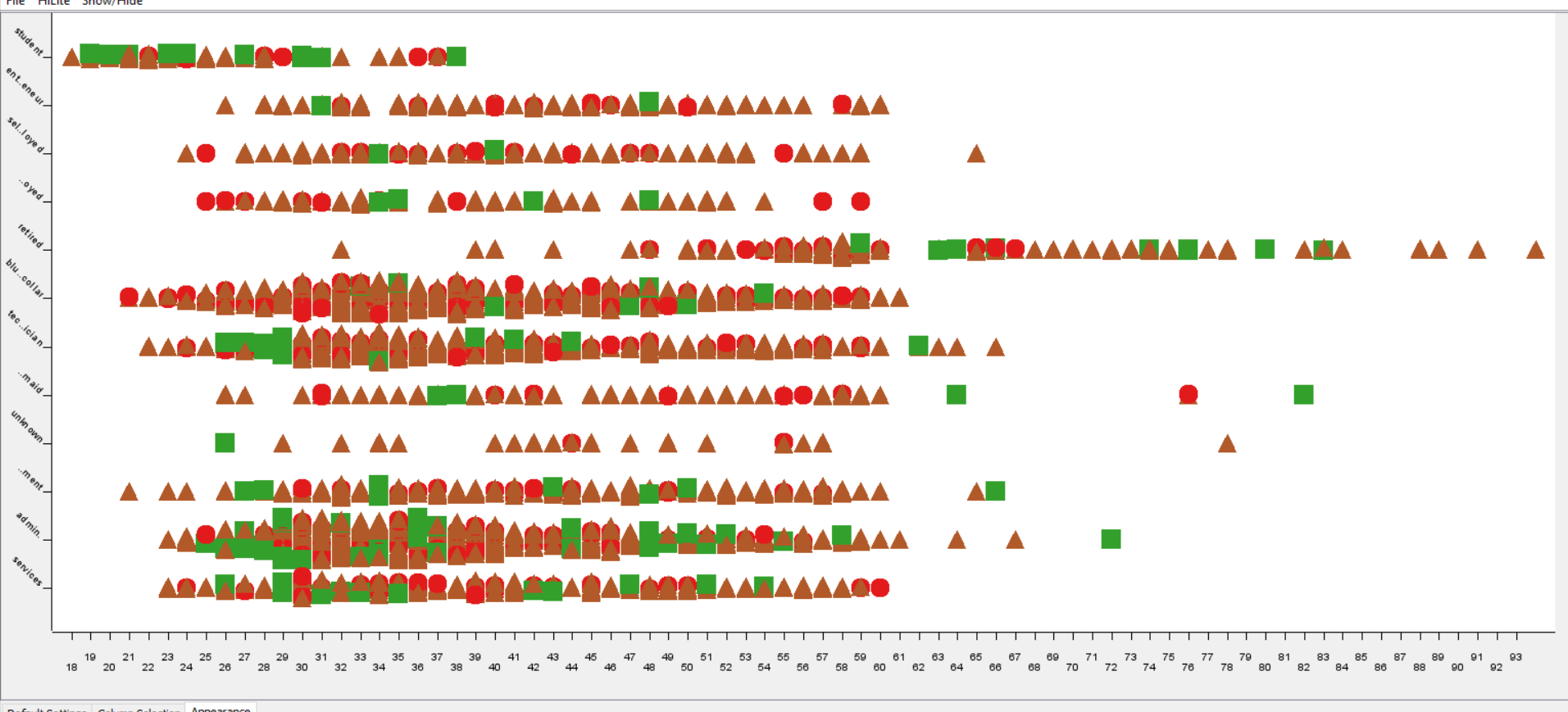


Figure 71: clustering for campaign, pdays and previous

Figure 71 shows the relationship between the number of contacts performed during the campaign(campaign), number of days that passed by after the client was last contacted from a previous campaign(pdays) and number of contacts performed before the campaign(previous). So here diagram shows 3 clusters' values. Cluster\_0 value is represented by rectangle with a data sample of 109, Cluster\_1 value is represented in the form of a circle with a data sample of 367 and cluster\_2 is represented by the triangle with a 2524 data samples. Figure 72 shows the values of the centroid of each sample. Figure 71, it is easily depicted the data samples of cluster\_2 are more than cluster\_0 and cluster\_1.

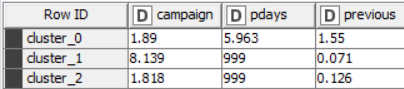


Figure 72: Centroids

**1b. Data Pre-processing**

1. **Binning**

Data contains outliers and errors. There is a way to deal with noisy data. A process called Binning is being used for this dataset. Binning is a way to group a number of more or less continuous values into a smaller number of "bins". There are two different ways to do binning:

1. Equi- interval (equal - width) binning

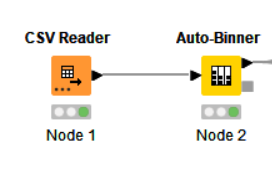
This is the process where it splits the whole range of numbers into equal-sized intervals.

Here we use the CSV reader node to read the dataset. To perform binning, we use the auto-binner node. Binning is done on an equi-width basis. First, we connect the executed CSV reader node to the auto-binner node. Then, to perform configuration, we choose the campaign attribute for binning. Bin naming is based on the "Numbered" option, and then there is an option for the number of bins. The number of rows in the dataset is 3000. A particular number of bins can provide different distributions in a histogram, so 29 bins have been chosen and performed the binning as equi-width(width). We used this method to identify a number of bins => bins= 2\* ∛n and then we selected the most appropriate bin around this value. As there will always be outliers, the focus was to have a minimal bin range and cover as many possible values as possible. Then execute it, and the bin dataset has been binned.

Graphical user interface, application

Description automatically generated

**The summary of the workflow:**



**Binned data is shown below:**

Table

Description automatically generated

**Histogram for binned data:**

A picture containing chart

Description automatically generated

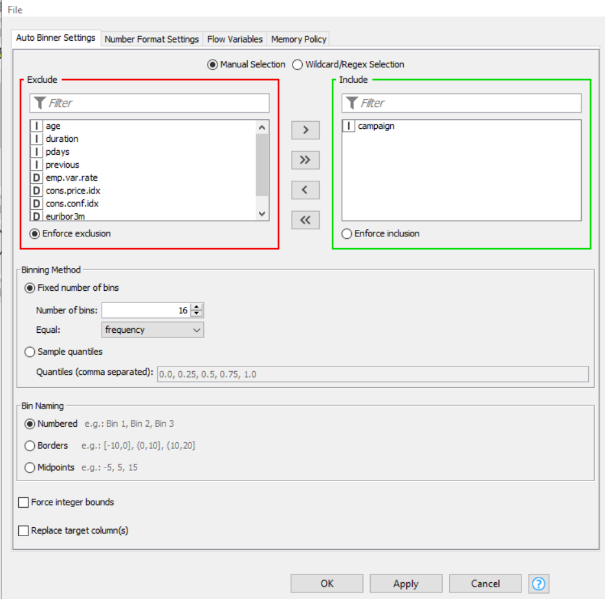
**The summary of the workflow:** Diagram

Description automatically generated with medium confidence

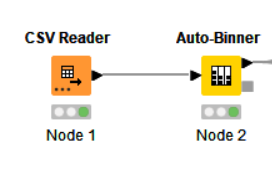
b) Equal- frequency (equi - depth) binning:

This is the process where it keeps a fixed number of data samples in the bins.

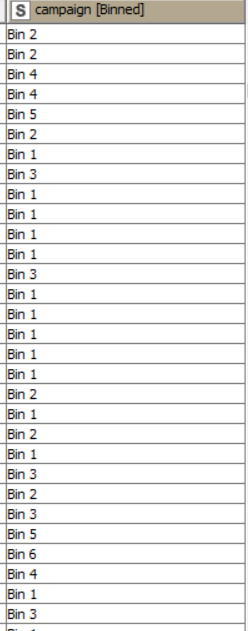
Here we use the CSV reader node to read the dataset. To perform binning, we use the auto-binner node. Binning is done on an equi-depth basis. First, we connect the executed CSV reader node to the auto-binner node. Then, to perform configuration, we choose the campaign attribute for binning. Bin naming is based on the "Numbered" option, and then there is an option for the number of bins. A particular number of bins can provide different distributions in a histogram, so 16 bins have been chosen and performed the binning as equi-depth(frequency). This number of bins has been selected in comparison with the other bins. For example, randomly taking 21 and 25 bins shows the same distribution of samples in the bins as 16 bins. One hundred eighty-seven samples in each bin. But there could be 1 or 2 bins that could have a higher number of samples than the other, or maybe it does have a lower number of samples than the other. Then execute it, and the bin dataset has been binned.



**The summary of the workflow:**



**Binned data is shown below:**



**Histogram for binned data:**

Chart, bar chart, waterfall chart

Description automatically generated

**The summary of the workflow:**

Diagram

Description automatically generated with medium confidence

1. **Normalization**

The normalization process involves scaling values with a specific range to a different preferred order.

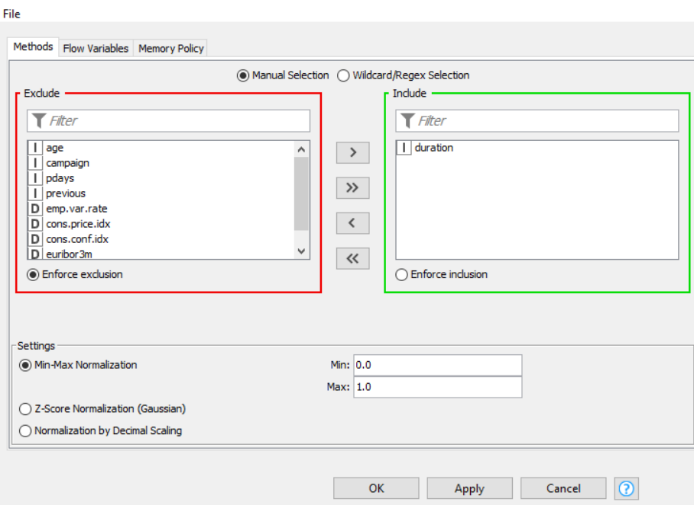
a) Min-Max Normalization

Here we use the CSV Reader node to read the dataset. In order to perform normalization, a normalizer node has to be used. The specific attribute where we need to perform normalizer is the duration attribute. Connect the CSV Reader to the normalizer node. We select the min-max normalization under the settings option in the configuration process. By default, the configuration process's Max and min value is specified.

A mathematical formula brings the maximum and minimum values between 0.0 and 1.0. This is the standard procedure for the max and min values. After execution, the duration value lies between 0.0 and 1.0. This is used to reduce the computation time and help the model identify the patterns.

Diagram

Description automatically generated



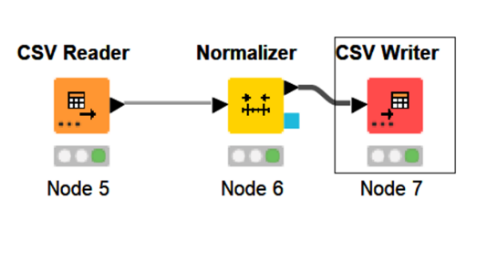
Normalized data by using min-max normalization is shown below:

A picture containing table

Description automatically generated

1. Z-Score Normalization

Here we use the CSV Reader node to read the dataset. In order to perform normalization, a normalizer node has to be used. The specific attribute where we need to perform normalizer is the duration attribute. Connect the CSV Reader to the normalizer node. We select the z-score normalization under the settings option in the configuration process. After the execution the data is normalized.



Graphical user interface, application

Description automatically generated

Normalized data by using z-score normalization is shown below:

Table

Description automatically generated

1. **Discretisation**

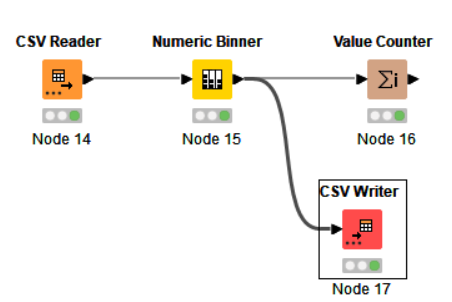
Reducing the volume of values for a given continuous attribute by grouping it into fixed intervals and producing the same or similar analytical results. This process is part of data reduction, replacing numerical attributes with nominal ones.

To perform Discretization, connect executed CSV reader to the Numeric Binner. The specific attribute where we need to perform Discretization is the age attribute. To define the range, the maximum and minimum value of the attribute is 94 and 18. In the configuration, we choose the age by double clicking on it and then clicking add button to discretise that attribute. It will create the bin, and then we will define the range according to the three bins.

Bin 1 - Adult => [-infinity,44.0]

Bin 2 - Mid-age => (44,65]

Bin3 - Old-age => (65, -infinity]



Graphical user interface, text, application

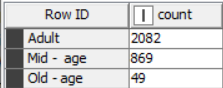
Description automatically generated

Discretised data is shown below:

Table

Description automatically generated with medium confidence

Value Count is shown below



1. **Binarization**

It maps a continuous or categorical attribute into one or more binary variables.

To perform binarization, connect executed CSV reader to the One-to-Many node. The specific attribute where we need to perform binarization is the marital attribute. In the configuration we choose marital attribute to perform the binarization. When we execute it, we can see the processed data, which includes four columns married, divorced, single, and unknown.

Diagram

Description automatically generated

Graphical user interface

Description automatically generated

**Binarized data is shown below:**

A screenshot of a computer

Description automatically generated with low confidence

**1c. Summary**

Below is a summary of the report's significant findings. Further research on the datasets could be valuable to explore the data further based on the conclusions of these studies.

Individuals who are in their **30s** should be called by cellular. Data indicates that **10.77%** of marketing campaigns fail, which is higher than success. Therefore, if the outcome of the marketing campaign should be a success, then the campaign members need to contact the individuals before and during the campaign to succeed.

There are **89.3%** of the individuals who have not subscribed to the term deposit. If an individual does not want a high risk in the investment process, then that participant should subscribe to the term deposit as term deposit are a low-risk way to invest money and to earn a fixed rate of interest.

The married customer maybe needs housing loan from the organisation, so the organisation wants to make a promotion of their product loan, the primary target van be a customer with marital status married.

Most **(584)** individual has a consumer confidence index of **-36.4** shows the decrement in the economic growth. So, by promoting economic growth will help the country to grow. Economic growth can be developed by giving volunteering jobs to the unemployed people. By removing unclear laws.

As a result of this, **3000** rows of value with varying ranges of attributes and well distributed frequencies and outliers for almost every numerical attribute, **clustering** and grouping data, and forming associations for analysis based on different attributes were partially beneficial. Resolving the issue and identifying the missing data would help reveal contacts made before the campaign that correspond to those made during the campaign, thus indicating the campaign's failure.