EDA – Data preprocessing

1. Missing data

The Figure 1 produces a graphic that shows which variables are most likely to be missing together. By default, the missing values are represented by white cells in the graphic. Hence, it indicates that there is no missing data in the whole dataset with 45,211 observations.

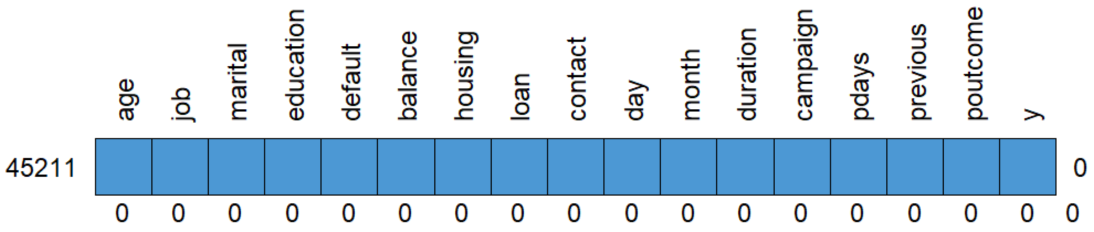

Figure 1: Md.patern function for detecting missing values

2. Reducing noise

In the data manipulation step, a Z-score threshold of 3 is defined, which will be used to identify and remove data from variable (Balance and Duration) points that fall outside of this range. This chunk code demonstrates how to perform noise reduction on a variable in a data frame using Z-scores, which can help improve the accuracy of subsequent analyses.

In order to improve the accuracy of subsequent analyses that rely on the Balance and Duration variables and make it easier to identify patterns and relationships with other variables, the chunk code of reducing noise. Regarding the specific code provided, the reductions in the number of observations in the Balance and Duration variables after reducing noise are relatively small compared to the total number of observations in the original dataset. Particularly, after reducing noise in the Duration variable, the observations are 44,248 (see Figure 2); after reducing noise in the Balance variable, the observations are 43,524 (see Figure 3). This suggests that there was not a significant amount of noise in the data that needed to be removed. However, even a small reduction in noise can improve the accuracy of subsequent analyses, and this could have important implications for understanding the relationship between the Balance and Duration variables and other variables in the dataset.

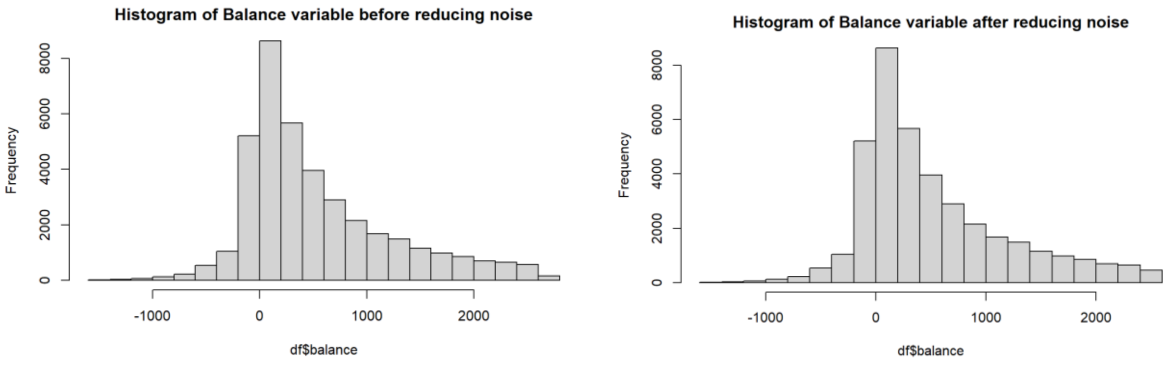

Figure 2: Reducing noise for Balance variable.

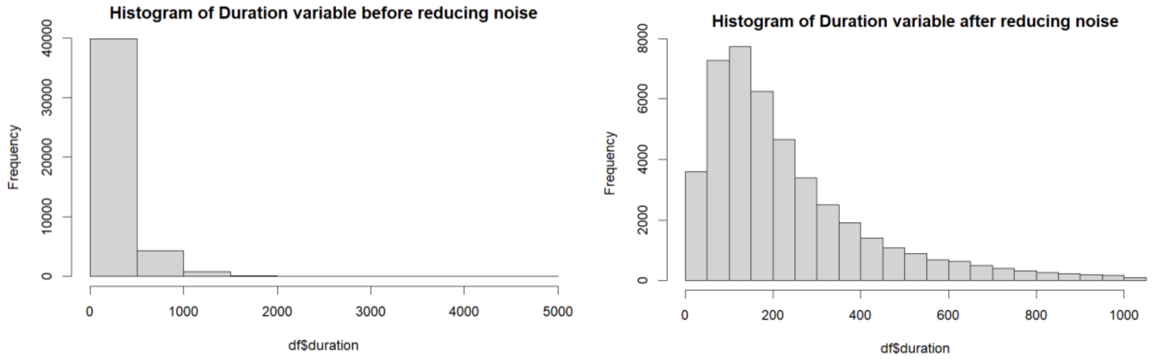

Figure 3: Reducing noise for Duration variable.

3. Convert the "duration" column to minutes.

Originally, the Duration variable is shown in numerical seconds. It is better to convert it into minutes for further analysis.

4. EDA - Descriptive & Frequencies

4.1.For numerical variables

The graphs in Figure 4 illustrates the descriptive statistics of the numerical variables.

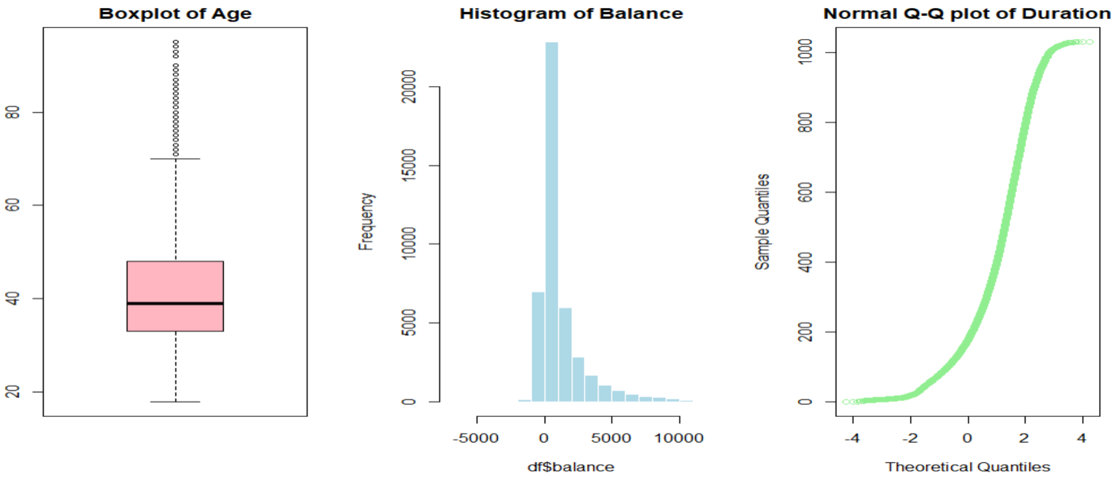

Figure 4: Visualization of three numerical variables: age, balance, and duration

The data and visualization show the mean, standard deviation, coefficient of variation and distribution of the three numerical variables (age, balance, and duration) as follows:

The figures above represent the descriptive statistics of three variables - age, balance, and duration. These statistics help us understand essential characteristics of each variable, such as the central tendency, variability, and relative variability.

Table 1: Statistical analysis for variables: age, balance, and duration.

Variable	Mean score	Standard Deviation	Coefficient of variation
age	40.93621	10.61876	25.93978
balance	1362.272 Median is 448	3044.766	223.5064
duration	258.1631	257.5278	99.75393

Starting with age, the mean score of 40.93621 indicates that the average age of the sample or population is around 41 years old. The standard deviation of 10.61876 suggests that the ages are moderately spread out from the mean, with some individuals being much older or younger than the average. The coefficient of variation of 25.93978% indicates that the age data has a moderate degree of relative variability, with the standard deviation being about one-quarter of the mean. This suggests that the age range is not too wide, but there is still some variability in the data. As shown in the box plot above, there are extreme outliers that fall outside outer fences which vary from 70 years old.

Moving on to the variable balance, the mean score of 1362.272 indicates that the average balance for the sample or population is around $1,362. The median balance of 448 suggests that the distribution of balance is skewed, with some individuals having much higher balances than others. The standard deviation of 3044.766 is quite large, indicating that the data points are spread out from the mean by a considerable amount. The coefficient of variation of 223.5064% is also quite high, suggesting that the balance data has a high degree of relative variability. This means that the balance range is quite wide, with some individuals having much higher balances than others. Furthermore, the histogram above indicates the unevenly distribution of the variable without any outlier.

Finally, for the variable duration, the mean score of 258.1631 indicates that the average duration of the sample or population is around 258 seconds or about 4.3 minutes. The standard deviation of 257.5278 suggests that the duration data is quite spread out from the mean, with some calls being much shorter or longer than the average. The coefficient of variation of 99.75393% indicates that the duration data has a moderate degree of relative variability, with the standard deviation being around one-third of the mean. This means that the duration range is not too wide, but there is still some variability in the data. Moreover, the Normal Q-Q plot above identifies suspected outliers where the expected values deviate from the reference line.

4.2.For categorical demographic variables

The marital variable

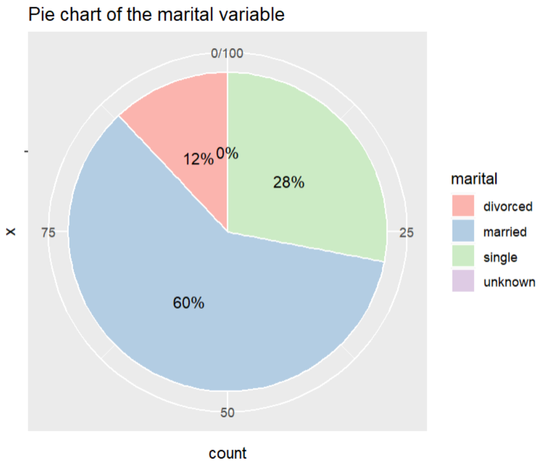

Figure 5: Pie chart of the marital variable

The percentages in the Figure 5 represent the proportion of individuals in the data set who fall into each category of the marital variable.

- 12% of the individuals in the data set are divorced.
- 60% of the individuals in the data set are married.
- 28% of the individuals in the data set are single.
- 0% of the individuals in the data set have an unknown marital status.

These percentages provide insight into the distribution of marital status among the individuals in the data set. For example, the majority of individuals are married, while a smaller proportion is single or divorced. Additionally, the fact that there are no individuals with an unknown marital status suggests that this variable is well-defined and complete in the data set.

For the job variable

The bar chart in Figure 6 of job types shows the distribution of job types among the dataset. The highest job types are blue-collar, management, and technician, while the lowest ones are unknown, student, housemaid, and entrepreneur. The highest job type, blue-collar, has more than twice the number of counts compared to the job types such as admin and services. This visualization provides a clear understanding of the most common job types in the dataset and highlights the importance of the top three job types in this dataset.

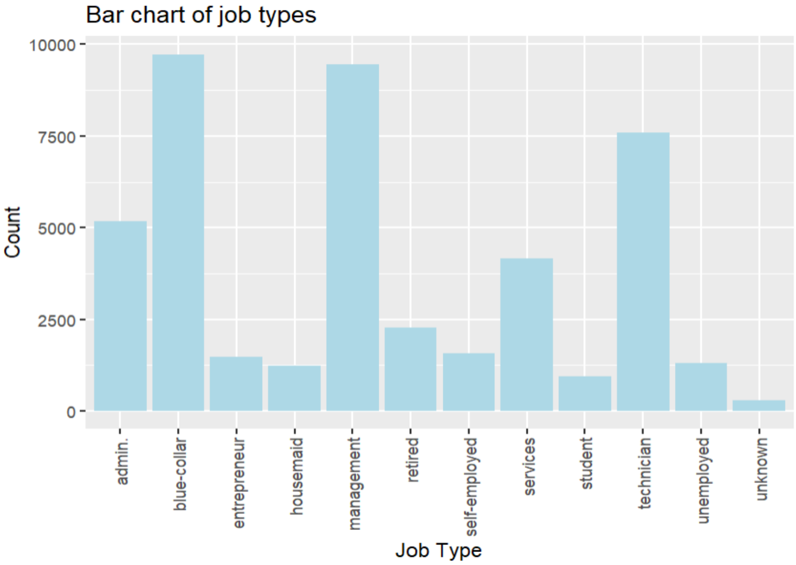

Figure 6: Bar chart of the job variable

For the education variable

This bar chart in Figure 7 shows the distribution of education levels among the individuals in the dataset. The highest number of individuals have completed secondary education, followed by those with tertiary education. The third highest category is primary education, while the smallest category is unknown education level.

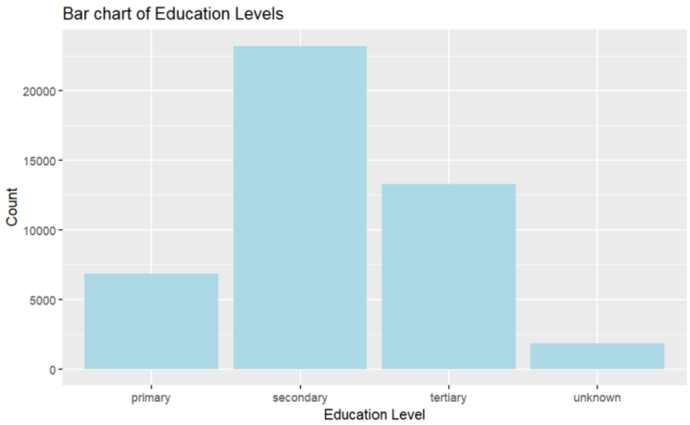

Figure 7: Bar chart of the education level variable

4.3.For the dependent Y variable

The given information in the Figure 8 presents the response rate of a survey question that asks individuals if they have subscribed a term deposit. Out of all the respondents who participated in the survey, 88.30152% answered "no" and 11.69848% answered "yes" to this question. This suggests that a large majority of respondents are not willing to use banking services in the near future, while only a small proportion expressed interest in using these services.

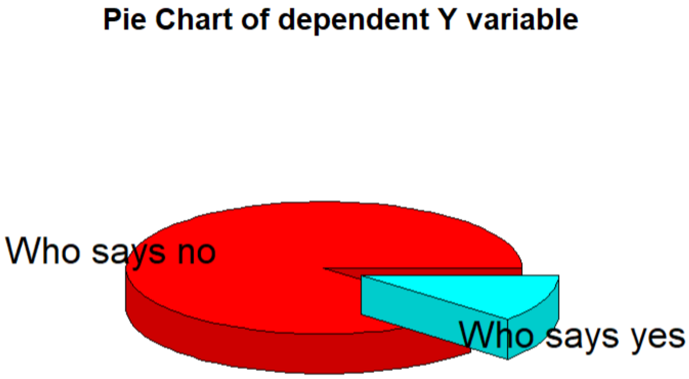

Figure 8: Pie chart of the dependent Y variable