Assessment Task 1 – SAS Predictive Business Analytics

DATA EXPLORATION

Group 1
Manjyot Joher | 12897981
Vi Nguyen | 13592629
Aiden Ye Yint Hlyan | 14017432
Su Myat Than Cin | 13486927
Ye Min Oo | 13506858

Table of Contents

1. Business Problem

With the increasing popularity of passive income strategies implemented to assist with long term secured savings, many banks have opportunities to launch direct marketing campaigns targeted at potential clientele, who prefer regularity in their returns, due to fixed interest rates. Generally the target customer for these campaigns is someone planning for long term returns, instead of short-term cash flow injection, as specified within a standard fixed term deposit.

Unique records within the *'BANK_DIRECT_MARKETING'* dataset explicitly documents real-world examples of these direct marketing campaigns, and is a subset of a larger dataset extracted from a Portuguese bank, over a time period of two years, from May 2008 to November 2010 (Moro et al., 2011). The total number of unique records amounts to 10,578 entries [VERIFY]. During this time, the bank documented data relating to 17 marketing campaigns, targeted at convincing potential customers to enter a term deposit contract, via in-house telemarketing (Moro et al., 2011). Using this dataset, our goal is to provide meaningful insights based on 18 input attributes such as job type, education, and marital status, mentioned in section 3.1 – data dictionary. Ultimately, upon completion of initial data exploration of attribute data, we aim to highlight key insights or trends within the data, and describe every attribute in respect to their distribution within the dataset. These steps will create the basis for further modelling and prediction steps within the data mining process; to eventually predict which customers would be more likely to agree to a term deposit, and help streamline future marketing campaigns.

2. Report Structure

This report aims to present key insights and observations derived from a comprehensive examination of the selected *'BANK_DIRECT_MARKETING'* data set (UCI Machine Learning Repository, 2012), which will be documented below. The report is segregated into three major sections: data exploration, comparative analysis, and data pre-processing. Within the data exploration section, attributes will be individually categorised, highlighting their distribution, along with key metrics. Building upon this, comparative analysis will delve deeper into the data by providing bivariate and multivariate relationship insights for certain attributes. These comparisons will provide a basis for any required pre-processing that will occur in the following section. Lastly, a summary of findings will be collated to document useful trends identified in the first three sections of the report; as well as future objectives and challenges faced during exploratory data analysis. All observations catalogued below were obtained with the use of the *SAS Viya for Learners* cloud platform.

3. Initial data exploration

In this section, attributes will be categorised into data types, and will have their summary statistics defined. Where relevant, summary statistics may be visualised to display interesting results.

3.1. Data dictionary

In preparation for further data analysis, attributes within the provided dataset will be classified into four types. These being:

Qualitative:

- **Nominal (categoric) –** attributes that have no significant meaning to them, other than being referred to as labels. Therefore, they cannot be ordered (e.g. Colours – 'Red', 'Blue', 'Green')
- **Ordinal –** attribute labels that can be ordered to show hierarchical meaning; however numerical operations such as addition and subtraction are not rational (e.g. Height – 'Tall', 'Average', 'Short')

Quantitative:

- **Interval –** attributes that can be ordered and are measured in fixed units. Subtraction is rational as the distance can be derived from two different values. Absolute zero values do not exist (e.g. Temperature in °F – '100°F', '50°F', '32°F')
- **Ratio –** attributes that can be ordered and are treated as real numbers, with an absolute zero value. All numerical operations can be conducted on these attributes. (e.g. Yearly income – '$155678.45', '$30000.00', '$22445.34')

The data dictionary below represents data types of input attributes, with possible modifications made, to better adhere to the business problem. Any changes to an attribute's type will be outlined in the pre-processing section of the report, following individual attribute exploration.

Attribute (in order of occurrence)	Data Type	Description	Example Value
Education	Nominal	Highest education level of client	Tertiary
Marital	Nominal	Marital status	Divorced
Job	Nominal	Occupation type	Student
Age	Interval	Age of potential client	25
Customer_id	Nominal	Unique row identifier	5122
Default	Nominal	Has the client ever defaulted on credit debt?	Yes
Balance	Ratio	Average yearly account balance (in Euros)	2500
Housing	Nominal	Does the client have an active housing loan?	No
Loan	Nominal	Does the client have an active personal loan?	Yes
Contact	Nominal	Communication type	Telephone
Day	Interval	Last contact day in the month (of the current/most recent campaign)	31
Month	Ordinal	Last contact month in the year (of the current/most recent campaign)	Jan
Duration	Interval	Duration of the last call with the client in seconds	245
Campaign	Interval	Number of contacts/calls conducted during this campaign (inclusive of the last contact)	6
Pdays	Ordinal	Number of days passed since last contact with client from a previous campaign	-1: target client was not previously contacted 1: target client was contacted in the last 1-273 days 2: target client was contacted more than 273 days ago
Previous	Interval	Number of contacts/calls for this client in the previous campaign	5
Poutcome	Nominal	Outcome of previous campaign for the client	Success
Target Variable: Y	Nominal	Has the client subscribed to a term deposit?	Yes

Table 1: Data Dictionary

3.2. Attribute information, summary statistics & data distribution

3.2.1. Education

Education is classified as a nominal attribute as it does not have an order, and is solely symbolic of the customer's highest education level. While it is possible to rank levels within this attribute ordinally from highest to lowest education, the order should not impact predictions for this business problem, and is therefore classed as nominal.

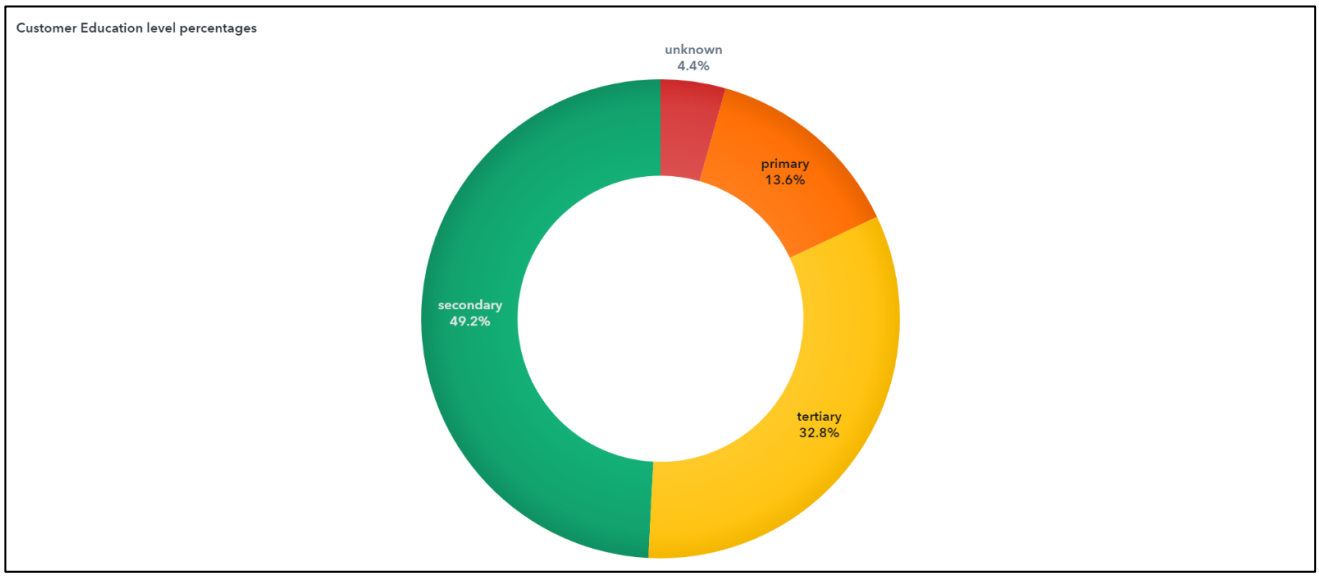

Figure 1: Education – Pie Chart

Prior to any pre-processing, it is interesting to note that the majority of clients in the bank's dataset have a secondary education as their highest academic achievement, amounting to almost 50%. It should also be noted that 4.4% of education values are marked as unknown, which could be indicative of a lack of reporting during the campaign. These are shown in red in Figure 1. To lower the number of dimensions in this attribute, the unknown category *may* be subsumed into the mode (most occurring) level for this segment, being secondary education.

3.2.2. Marital

Marital is classed as a nominal attribute, as ordinally ranking single, divorced, or married categories, does not provide additional value to the raw data. This attribute is representative of the marital status of each potential client, as reported to the Portuguese banking institution, during their telemarketing campaigns that ran from 2008-2010.

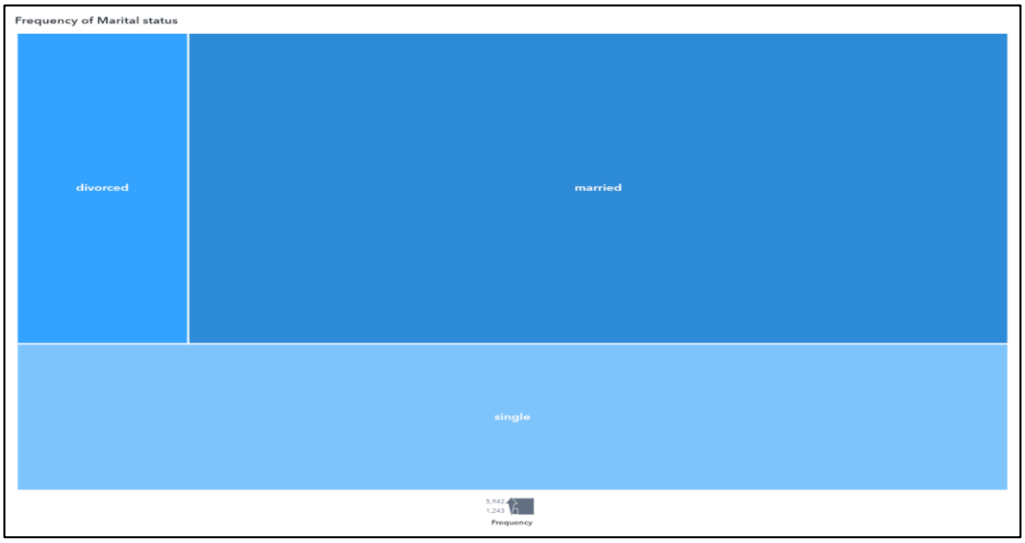

Figure 2: Marital – Pie Chart

Figure 2 above, visualises these 3 categories into a tree map. As shown, 'married' has the largest count amongst all potential customers, totalling to 5942. Contrastingly, potential clients who reported a 'divorced' marital status, had the lowest frequency, equating to 1243.This 378% increase from divorced to married, could be strategic within the bank's telemarketing campaign. It could be assumed that married customers have more disposable income to invest in a term deposit, in comparison to divorced individuals, and therefore were targeted more often. The latter

category may have had to deal with expensive legal fees, wealth sharing between both parties, and may not be in the right mindset to open a term deposit. This notion will be further explored in section 4, comparing relevant attributes together such as, balance, age, and housing. Potential external factors such as the global financial crisis at the time of these campaigns, may have also influenced the subset of targeted clients sought out by the Portuguese banking institute.

3.2.3. Job

Job is a nominal attribute as it designates recorded occupation types during the bank's telemarketing campaigns. The tree map below illustrates occupation type, using frequency as its size, and gradient factor.

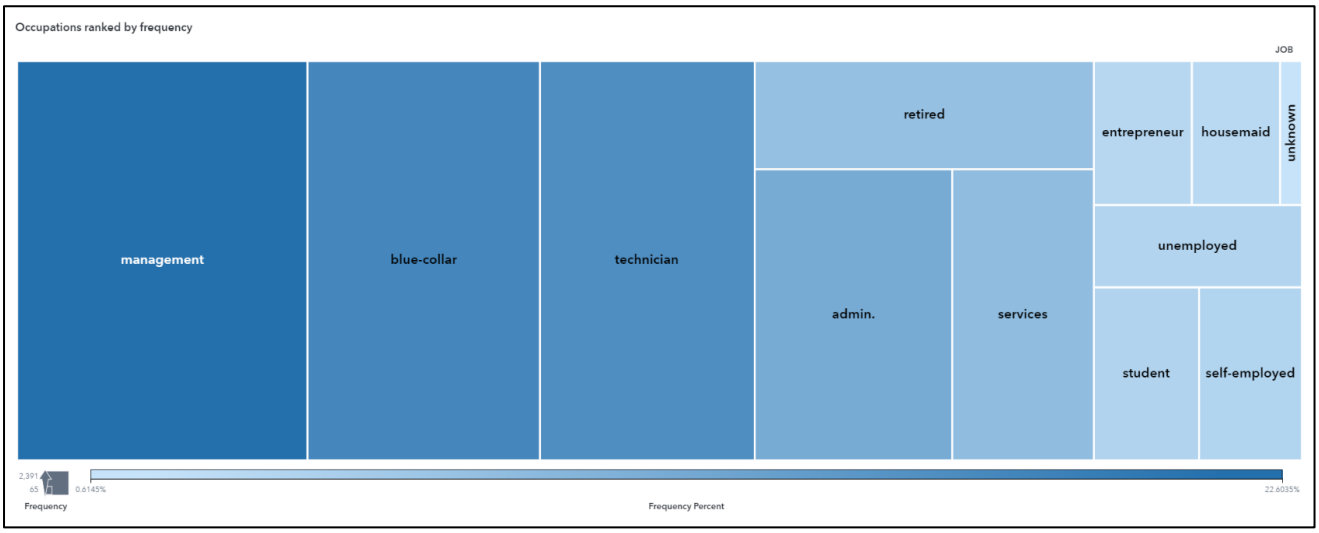

Figure 3: Job – Tree Map

JOB	Frequency	Frequency Percent ▾
management	2,391	22.60%
blue-collar	1,914	18.09%
technician	1,768	16.71%
admin.	1,185	11.20%
services	850	8.04%
retired	757	7.16%
student	375	3.55%
self-employed	367	3.47%
unemployed	353	3.34%
entrepreneur	291	2.75%
housemaid	262	2.48%
unknown	65	0.61%

Figure 4: Job – Tabulated Job Frequencies

The Job attribute contains 12 categoric levels, ranging from the largest being management at 22.6%, to housemaid being 2.48%. As depicted in the tree map in Figure 3, emphasis was placed on contacting potential clients with secure occupations such as managers, technicians, and administrators. Figure 4 reinforces this, with the majority of occupations below double-digit percentages (10% or lower), being mainly sole trader-based occupations. These may be assumed to be less financially secure, as retirees, students, housemaids and self-employed categories, may not have a steady stream of income. An exception to this is the 'service' category, which may have not been of great importance to the bank at the time. Given the global financial crisis that took place during these campaigns; service worker income may have become unstable. It can be assumed that the demand for service-based businesses dwindled during this period due to many people losing their own disposable income streams. Aside from this exception, the 'unknown' job category is representative of only 0.61% of the dataset, and could be due to incorrect recording during the collection process.

3.2.4. Age

Age is classified as an interval attribute, as it can be deduced that the difference between the ages of 20 and 30, is 10 years, which is the same difference between 30 and 40. While age can also be classified as a ratio, in this dataset there are no absolute 0 age values, as the minimum is 18.

Statistic:	Mean	Median	Range	Mode	Standard Deviation	Relative Variance	Skewness	Kurtosis
Value:	41.2268	39	95 (max) – 18 (min) = 77	31	11.9977	0.2910	0.8442	0.5614

Table 2: Age - Summary statistics

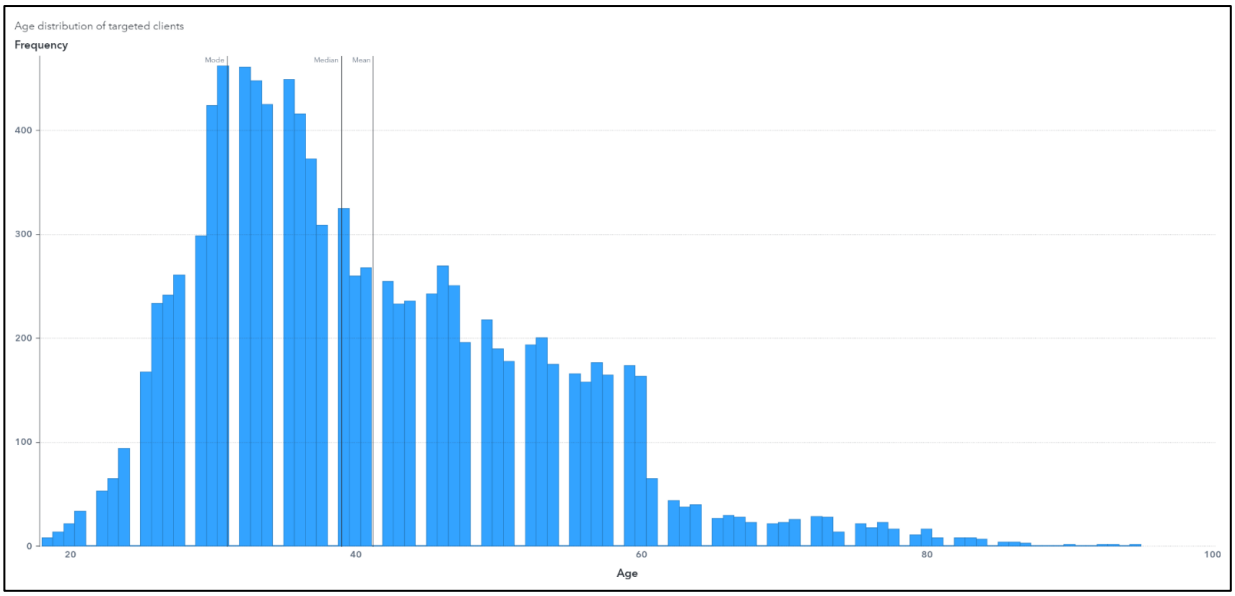

Figure 5: Age – Binned Histogram

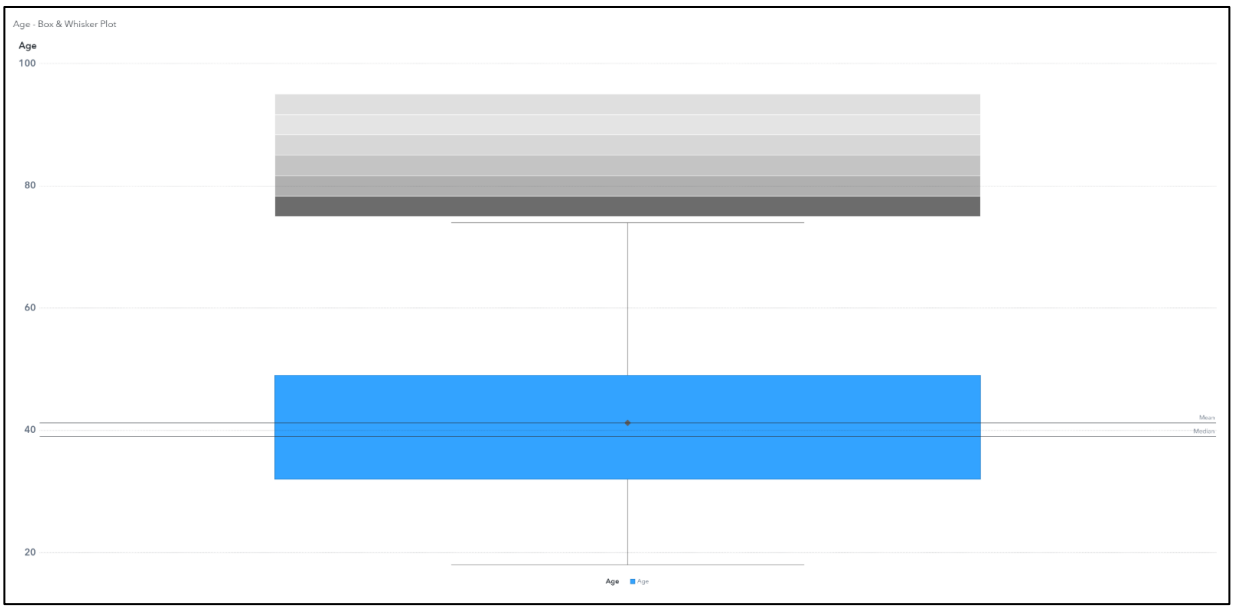

Figure 6: Age – Box & Whisker Plot

This attribute is representative of the age of potential clients targeted by the telemarketing campaigns, ran by the Portuguese banking institution. As shown in Table 2, the most occurring age amongst potential clients is 31, indicating that the distribution of ages is slightly positively skewed, with a skewness value of 0.8. This coincides with the big drop off in ages from 60 and older as shown in Figure 5, aligning with the positively skewed characteristic, of trailing outliers towards the right of the distribution (Klima, 2021, para. 2). Additionally, the kurtosis of 0.56, is indicative of relatively heavier tails, reinforcing the outliers shown visually in the positively skewed histogram in Figure 5 (Kenton, 2023; SAS, 2015). The box and whisker plot shown in Figure 6 also corroborates this narrative, and depicts a multitude of outliers shown in a grayscale gradient, above the third quartile of the chart. The least occurring outlier range shown is in the 92-95 bin, with a frequency of 6. The most occurring outlier range is the 75-80 bin, with a frequency of 80.

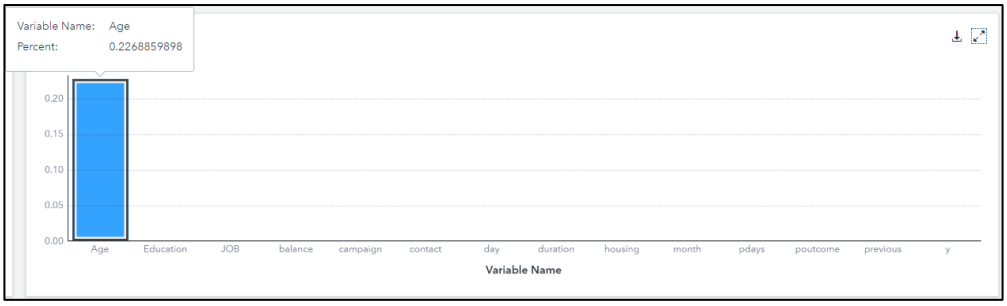

Figure 7: Age – Missing Values

It is important to note that Age is the only attribute in the entire dataset that has missing values. Using the data exploration node in SAS Model Studio, it is evident that this equates to 24 missing values, documented by the percentage in Figure 7. These values will be addressed in section 5.1 – data cleaning. It should also be noted that both Figures 5 & 6 have been illustrated with binned age values, with the histogram being capped at 100 bins due to SAS Viya limitations.

3.2.5. Customer_ID

'Customer_id' is classified as a nominal attribute as it does not have any meaning aside from labelling each row in the dataset. The number of rows within the dataset equates to 10,578. This attribute will be omitted from training, validation, and test sets during pipeline creation. Its removal will be addressed further in section 5.3 – dimensionality reduction.

3.2.6. Default

Default is classified as a nominal attribute which indicates whether a client has defaulted on credit debt.

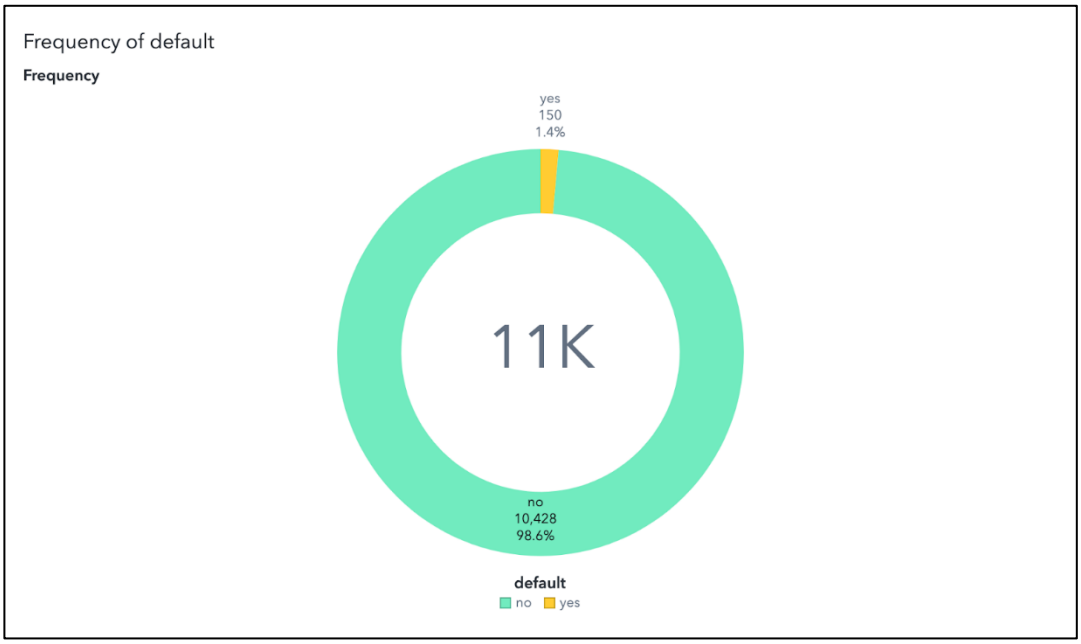

Figure 8: Default – Pie Chart

According to Figure 8, 10,428 people (98.6%), have not missed payments on credit debt, contrasted with 150 people that have. The incredibly low default rate of 1.4% is significant, showing that the vast majority of Portuguese clients in this sample are financially responsible, and are effective at managing their credit debt. This is in the best interest of the Portuguese institution, so it is logical that this distribution of target clients was chosen for the marketing campaign.

3.2.7. Balance

Balance is a ratio attribute which indicates the potential client's average yearly balance in euros. It is identified as a ratio attribute as it has a logical zero value in the dataset. Moreover, the attribute has multiplicative and divisive properties that enable balances such as 600, to represent a value twice the size of 300 - this can be useful in explaining differences between data points.

Statistic:	Mean	Median	Range	Mode	Standard Deviation	Relative Variance	Skewness	Kurtosis
Value:	1,548.5298	566	81,204 (max), -3058 (min)	Not Applicable (unique value per row)	3,130.5653	2.0216	7.7168	119.6499

Table 3: Balance – Summary statistics

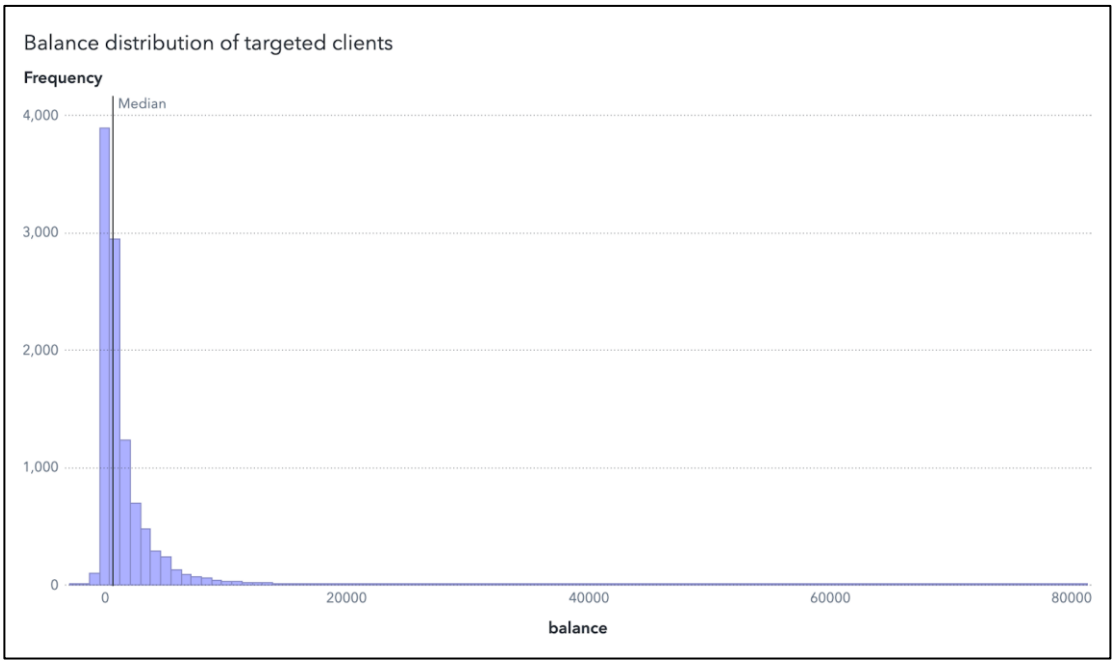

Figure 9: Balance – Binned Histogram

As taken from Table 3, the kurtosis value of 119.65 is significantly higher than a normal distribution, characterising the distribution as platykurtic (Frost, 2022). In contrast to a normal distribution, platykurtic distributions contain heavy tails and a sharp peak, which show that there are more outliers present (Glen, n.d., para. 1). This characteristic is clearly evident in the distribution shown in Figure 9. When referring to bank balances, this indicates that a disproportionately large percentage of customers have very high average yearly balances. This is corroborated by the significant positive skew of the distribution, equating to 7.72. Positive skewness means the tail is longer on the right side than the left. This indicates that the majority of targeted clients have balances that are lower than the mean, with a small number of them having extremely high balances that are pushing the mean upward. For relative variance, a value greater than 1 (in this case, 2.02) denotes a high degree of dispersion within the data.

3.2.8. Housing

Housing is a nominal attribute and is indicative of the client's housing loan status, and is therefore a binary variable.

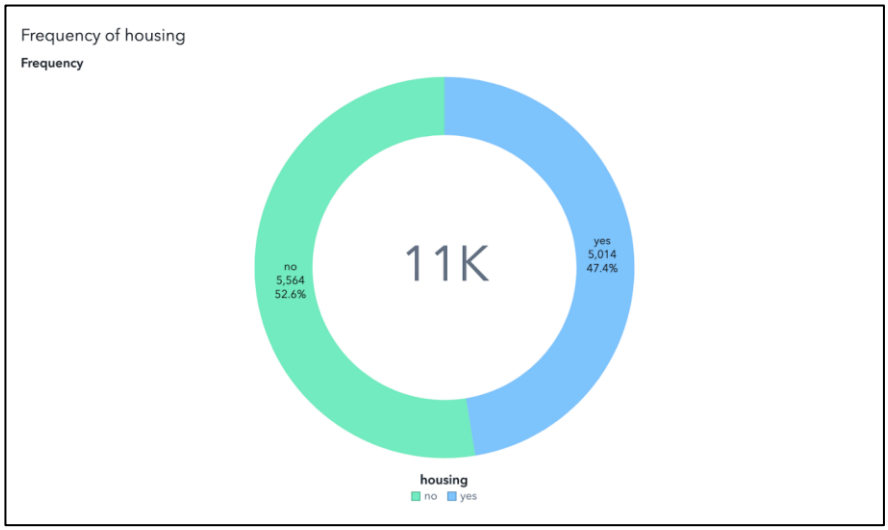

Figure 10: Housing – Pie Chart

As shown in Figure 10, the results indicate that 5,564 people (52.6%), do not have an active housing loan and 5,014 people (47.4%), hold an active one. The data shows a nearly equal distribution of clients with and without active mortgages. During comparative analysis, it would be interesting to look at how old these two groups are. For example, while older age groups may have paid off their mortgage loans, younger people may have not.

3.2.9. Loan

Loan is characterised as a nominal type, capturing whether a client has an active personal loan or not. Upon analysing the dataset, an overwhelming majority of clients, precisely 87.1% or 9,210 individuals, have reported not having an active personal loan, as visualised in Figure 11. Contrarily, 1,368 clients, representing 12.9% of the dataset, confirmed the presence of a personal loan. This significant skew towards the "no" category indicates that the majority of potential clients approached by the bank's telemarketing campaigns do not have other personal loan commitments, potentially making them more financially flexible.

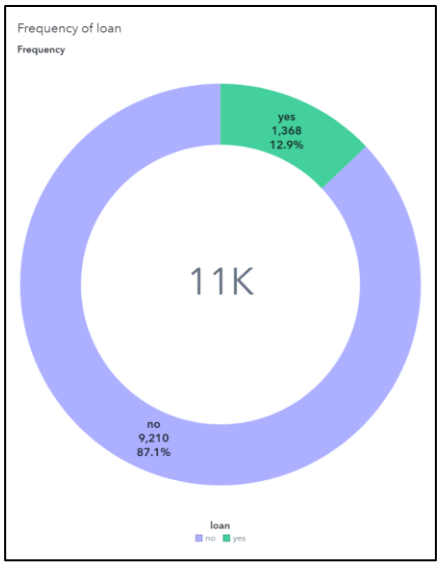

Figure 11: Loan – Pie Chart

3.2.10. Contact

'Contact' is delineated as a nominal type, indicating the mode of communication through which clients communicate with the bank. This attribute is comprised of three distinct categories: cellular, unknown, and telephone. Within the dataset, a dominant majority of 7,682 clients, equivalent to 72.6%, have opted for cellular communication. In contrast, a notably smaller fraction, 712 clients or approximately 6.7%, have chosen the telephone as their mode of contact. Interestingly, there's a sizable segment, about 20.6% or 2,184 clients, where the mode of communication remains 'unknown'. This category can be attributed to several factors such as data entry errors, or perhaps the possibility of utilising alternate contact methods, not specified in the template used during initial data collection. To potentially improve model predictions, considering strategies to handle this 'unknown' category, such as imputation or grouping may be employed, if it's deemed important in swaying the outcome of the dependent variable, '*y*'.

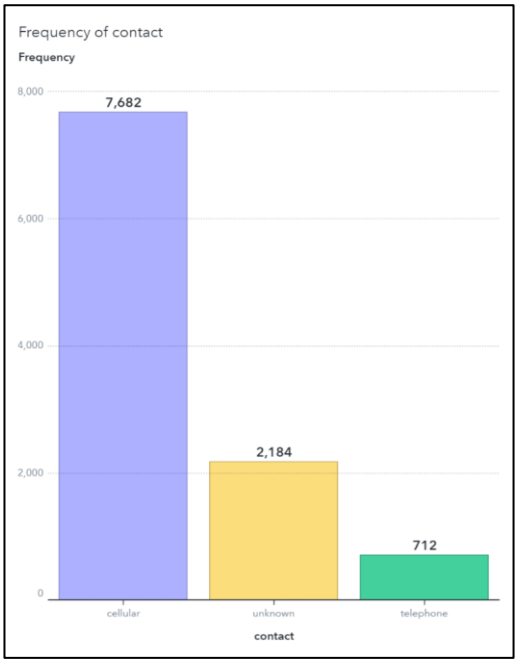

Figure 12: Contact – Histogram

3.2.11. Day

The 'Day' attribute is representative of the last contact day in the month of the most recent campaign. Day is an ordinal data type, as it is meaningful to rank them categorically, in accordance with the Gregorian calendar. As shown in Figure 13, the call frequency ranges from 110 to over 500. On most days, the frequency is in the range of 300 to slightly over 400. There are a few days with abnormally low frequency, such as the 1st, 10th, 24th and the 31st, where the frequency drops below 200.

Statistic:	Mean	Median	Range	Mode	Standard Deviation	Skewness	Kurtosis
Value:	14.4759	15	31 (max), 1 (min)	20	8.4138	0.1294	-1.0608

Table 4: Day – Summary statistics

As shown in the summary statistics table above, a skewness value of 0.1294 indicates that the 'Day' attribute has a very negligible positive skew. This slight skew is more characteristic of a relatively symmetrical distribution, rather than a strong peak. This is also validated by the negative kurtosis value of approximately -1, stipulating relatively even tails, with no explicit outliers, as shown in Figure 13.

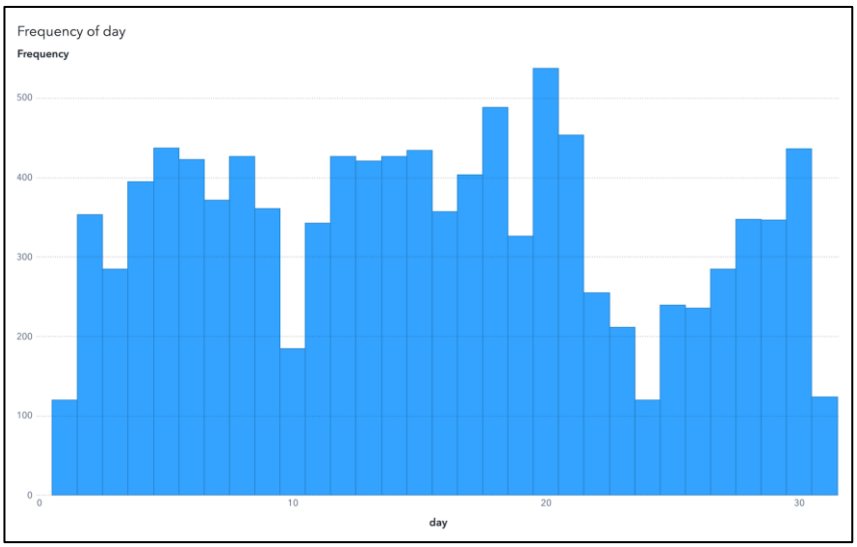

Figure 13: Day – Histogram

Moreover, the frequency of contacts drops relatively low at the start and end day of the month. This reduction could indicate that the customers refused to interact with the firm, or the bank was running on skeleton staff during these periods. Additionally, the contact frequency spikes on day 20, with approximately 537 contacts, followed by day 18 with 488 contacts. The increase in contacts on these days may be a result of potential clients receiving their monthly wage, based on industry trends; and therefore have more resources allocated towards contacting them. These spikes are followed by an off-peak period from the 22nd to the 26th, symbolising the only period of relatively low-frequency contact.

3.2.12. Month

The 'Month' attribute is indicative of the last contact period within the year. This attribute is classified as ordinal, as it follows the standard Gregorian calendar.

Statistic	**Value**
Mode	May (Frequency: 2,603)

Table 5: Month – Summary statistics

Timing is crucial when it comes to campaign deployment and data reconciliation. The selected contact period may significantly enhance the engagement of the potential consumer; and attract a greater response rate, thereby increasing the probability of signing up for a term deposit. A few factors that the Portuguese banking institution may have considered when planning telemarketing campaigns include: the seasonal trends in local markets, consistency of campaigns from 2008-2010 and the global financial crisis.

May accounted for the most contacts, with a value of 24.6%, as shown in Figure 14 below. Subsequent months, June, July, and August, also have high contact frequencies in the year. These four months contribute to nearly two-thirds of the bank's total calls in a year. In contrast, there were a few months when the communication between both parties reduced. These months were September, October, December, January, and March, and only contributed to approximately 15% of the total call frequency.

Considering the ordinality of the data, and the fact that these real-world campaigns were conducted during the global financial crisis, it is important to also consider external factors that may have influenced the monthly contact distribution. While the contact year is not provided within the dataset, it may be assumed that the high contact frequencies in May, could have been due to upward-trending deposit interest rates, as shown in Figure 15. This trend marks the beginning of the recovering Portuguese economy towards the start of 2010, in which the last marketing campaigns were conducted. It may be assumed that the majority of contacts made in May, June, July and August took place in 2010; where potential clients could be motivated by higher interest rates (TheGlobalEconomy.com, 2023), and therefore justifies the bank's motivation to increase call frequencies.

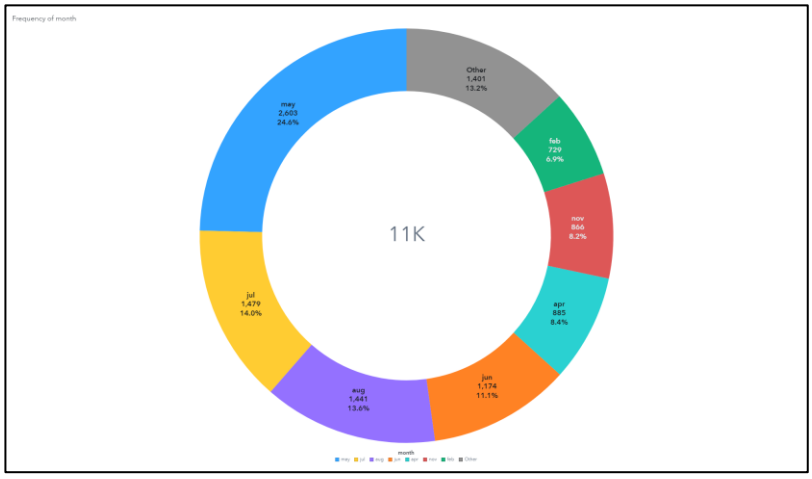

Figure 14: Month – Pie Chart

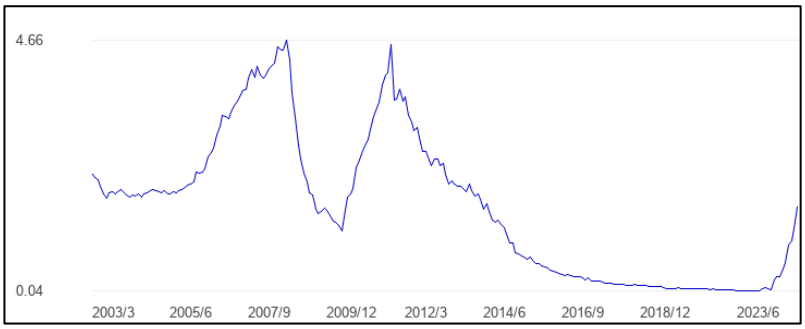

Figure 15: Historical deposit interest rates in Portugal – Line Graph (TheGlobalEconomy.com, 2023)

3.2.13. Duration

Duration is designated as an interval attribute, due to it representing the duration (in seconds) of the previous call with the potential client. This allows differences between two or more call durations to be made, aiding in comparative data point analysis.

Statistic:	Mean	Median	Range	Mode	Standard Deviation	Relative Variance	Skewness	Kurtosis
Value:	379.2437	260	3,881 (max) – 4 (min) = 3877	1,086 calls in bin: 120.31 – 159.08 seconds	350.9718	0.9255	2.1468	7.3608

Table 6: Duration – Summary statistics

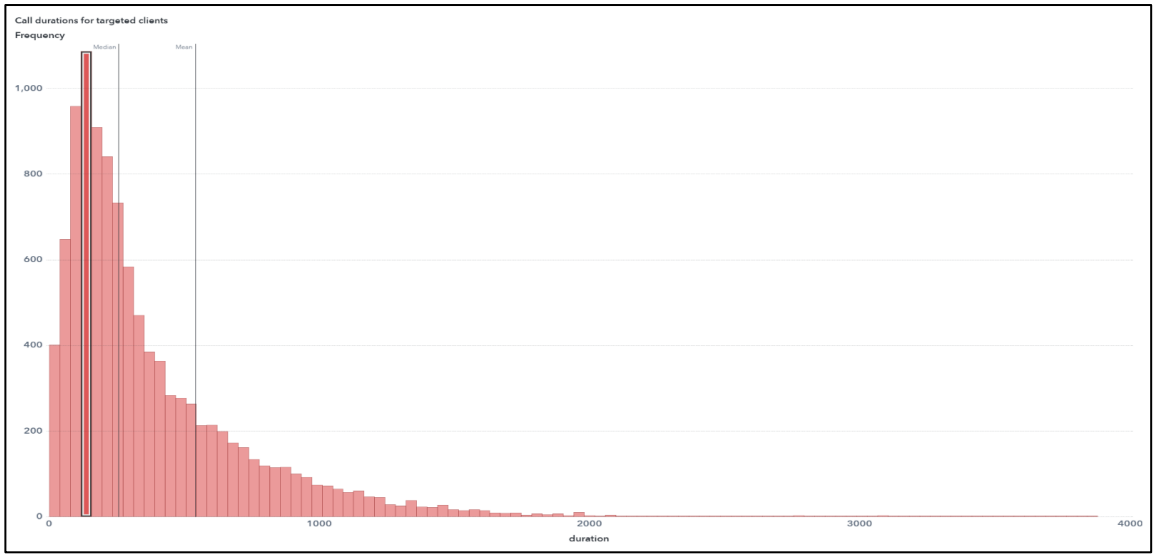

Figure 16: Call durations for targeted clients – Binned Histogram

Figure 17: Duration – Box & Whisker Plot

Call duration may be valuable in determining if targeted clients sign up for a term deposit, as greater call times may provide insight into client fatigue. Figure 16 depicts the distribution for contacted clients, with a noticeable peak towards the left of the chart. This is validated by the skewness value of approximately 2.15, indicating a strong positive skew. Contributing to this skew is the binned mode of approximately 120 – 159 seconds, representing a frequency of 1086. This highlights the fact that most calls to potential clients only last between 2 - 3 minutes, perhaps establishing a premise that most people do not need a long amount of time to decide on a term deposit. This notion will be further examined in section 4 – comparative analysis. Additionally, the histogram illustrates the decreasing graduation of bins towards the tail end of the graph. This consistently dropping frequency is also representative of the relatively high kurtosis value of approximately 7.36. This abnormally high concentration of outliers, when compared to a normal distribution, is also evident in Figure 17. The box plot depicts these outliers in a grayscale gradient above the third quartile, with the darkest section being indicative of the most occurring outlier bin (1052 – 1523.5 seconds), with a frequency of 450. It could be assumed that these duration outliers represent clients with lower average bank balances, and therefore need more time to decide on signing up for a term deposit. This supposition will also be considered in section 4.

3.2.14. Campaign

Campaign is an attribute representing the number of contacts performed during this campaign for the potential client. It is defined as an interval attribute, as there are no 0-values present, or these records would simply not exist in the dataset, as they were never contacted. This is confirmed by the minimum value of 1.

Statistic:	Mean	Median	Range	Mode	Standard Deviation	Relative Variance	Skewness	Kurtosis
Value:	2.4748	2	50 (max) – 1 (min) = 49	1	2.6152	6.8392	5.0976	44.6295

Table 7: Campaign – Summary statistics

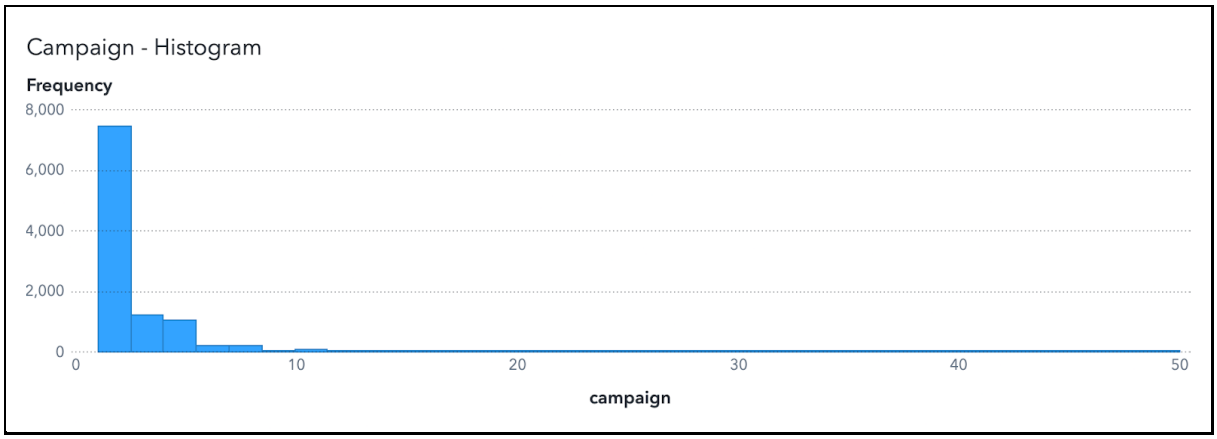

Figure 18: Campaign – Histogram

Table 7 shows the summary statistics regarding the attribute. The values range from 1 to 50, with the median being 2. Figure 18 depicts that the majority of contacts within this campaign are under 10, and is indicative of a positive skew.

3.2.15. Pdays

PDays refers to the number of days that have passed after a potential client was last contacted from a previous campaign. The Pdays attribute can be regarded as ordinal, as it includes distinct categorical values that can be ordered logically. These are tabulated below, and given their tiered hierarchy of durations, there is a clear logical order that can be followed.

PDays	Frequency	Frequency Percentage
-1 (Not Contacted)	7866	74.36%
1 (contacted within the last 1 - 273 Days)	1109	10.48%
2 (contacted more than 273 Days ago)	1603	15.15%

Table 8: Pdays – Summary statistics

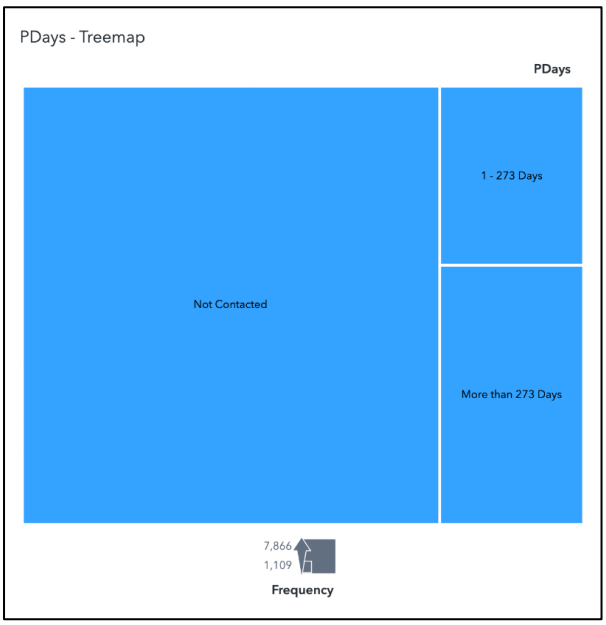

Figure 19: Pdays – Treemap

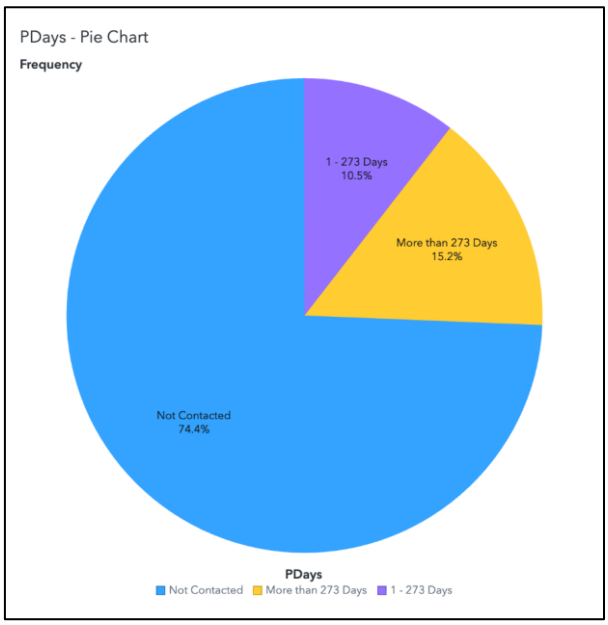

Figure 20: Pdays – Pie Chart

The tree-map presented in Figure 19 highlights the dominance of potential clients that were never contacted in a prior campaign to this one. The pie chart above also reinforces the idea, as this category accounts for 74.4% of the entire dataset. This provides insight into the bank's marketing strategy, suggesting that potentially untapped clientele, may be more willing to register for a term deposit, as they were never previously contacted before the current campaign.

It appears that those customers who were contacted a long time ago (more than 273 days), make up 15.2% of the set, which is bigger than the 10.5%, which is made up of customers who had more recent contacts. Customers who had been contacted previously may have biases towards the bank, which may influence their decision on whether to subscribe a term deposit.

3.2.16. Previous

This attribute refers to the number of contacts made before the current marketing campaign. It has values ranging from 0 to 275. It is considered as interval data, as a true zero point in this case provides no further statistical meaning, aside from representing that call frequency in a previous campaign was 0 for a particular target client.

Statistic:	Mean	Median	Range	Mode	Standard Deviation	Relative Variance	Skewness	Kurtosis
Value:	0.8525	0	275 (max), 0 (min)	0	3.4721	12.0556	48.5081	3694.2714

Table 9: Previous – Summary statistics

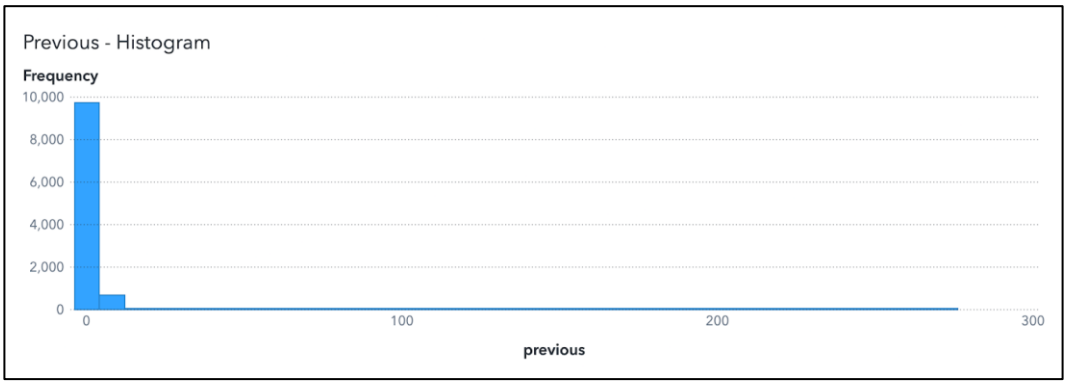

Figure 21: Previous – Binned histogram

Figure 22: Previous – Box & Whisker Plot

In Table 9, the summary statistics show that the median and mode equate to 0, which proposes that a significant portion of target clients had no interactions before the current campaign. Figure 21 demonstrates the high positive skewness of data, with a value of 48.5, and an extremely high kurtosis of 3694, indicating the occurrence of

significant outliers. This is corroborated by the box plot, showing that most target clients were contacted between 1 and 50 times. It is also noteworthy to look at the one outlier who was contacted 275 times prior to this campaign, and resultantly, played a major role in the high kurtosis value.

3.2.17. Poutcome

Poutcome is an attribute that represents the outcome of the previous marketing campaign. It is categorical in nature and contains values tabulated below. Given the nature of these categories, they don't have a clear order which can be meaningful. While it could be argued that 'failure' is before 'success' in terms of measuring campaign success, 'other' and 'unknown' values don't have a clear position. Therefore, *poutcome* is a nominal attribute.

Poutcome	Frequency	Frequency Percentage
Success	1055	10%
Failure	1153	10.9%
Other	502	4.7%
Unknown	7868	74.4%

Table 10: Poutcome – Summary statistics

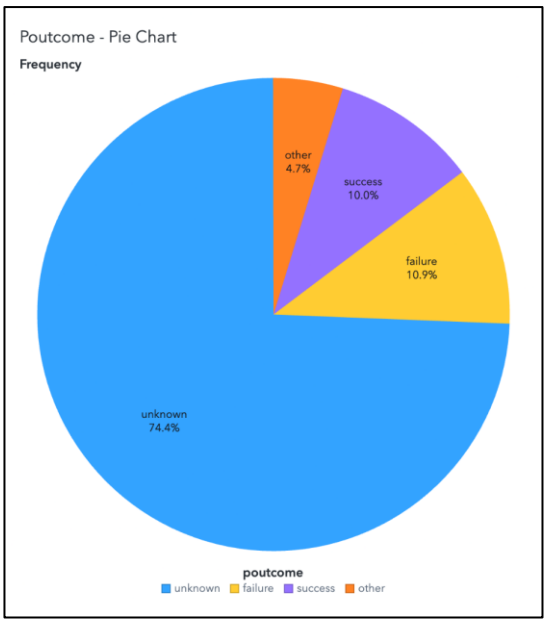

Figure 23: Poutcome – Pie Chart

It can be seen that 74.4% of the values are categorised as 'unknown', indicating that the majority of the outcomes from the previous marketing campaign are unknown. Both 'success' and 'failure' outcomes have similar frequencies, with 10.0% and 10.9% respectively. This suggests that when the outcome of the previous marketing campaign is known, there is a an approximate even split.

4. Comparative Analysis

Comparative analysis is a vital step within the exploratory data analysis process, as it can unravel key insights into relationships between chosen variables. This can improve dataset literacy, and assist in making informed pre-processing decisions, as outlined in section 5. By examining key relationships between attributes, potential trends and correlations can be derived, painting a narrative on the factors surrounding the telemarketing dataset, even considering external factors at the time of data collection.

4.1. Bivariate comparisons

4.1.1. Age and Housing

Age and Housing are assumed to be two important attributes when it comes to having free cashflow. This is especially important when attempting to persuade target clients to sign up for a term deposit, essentially locking their funds away for a specified period of time. Figure 24 below combines these two attributes into a single scatter plot to assist in correlation interpretation.

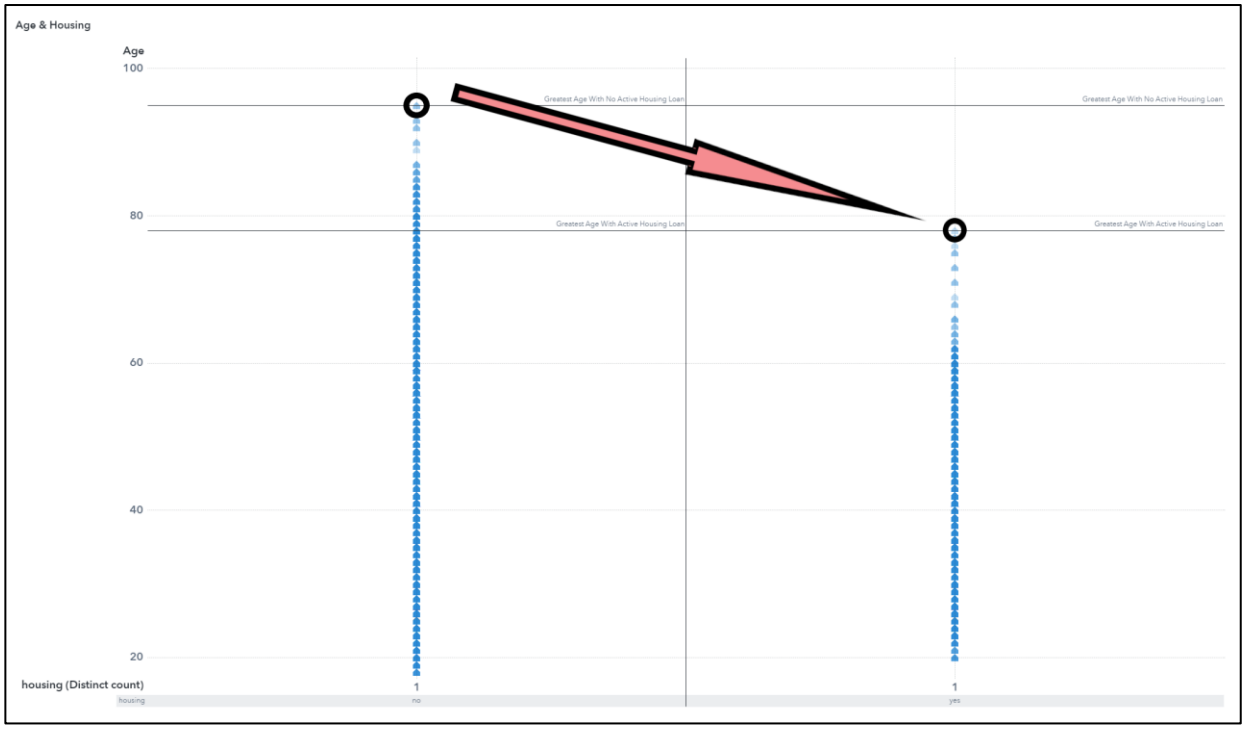

Figure 24: Age & Housing – Scatter Plot

As shown by the descending arrow, it is evident that there is a weak trend portraying that older clients have no active housing loans, whilst younger clients do. The blue colour gradient displayed on the house icons, indicate the concentration of ages amongst the overall distribution. As shown, there is a higher concentration of ages 20 – 60 for those that have an active housing loan, and 20 – 80 for those that do not. Two specific points have also been highlighted by black circles. The leftmost circle represents the greatest recorded client age who does not have an active housing loan, whilst the rightmost circle shows the greatest cage of a client who does. These points equate to a value of 95 and 78 respectively. This again corroborates the narrative that older targeted clients have paid off their mortgages and younger clients have yet to do so.

Additional external factors at the time of data collection may have also played a role in this relationship. The BANK_DIRECT_MARKETING dataset was assembled during the global financial crisis, by collating 17 telemarketing campaigns from May 2008 to November 2010. During this time, it was likely that free cashflow plummeted for residents and those with active home mortgages had to defer payments or overdraft credit accounts to continue making payments. This financial instability could've influenced this data for those with active loans, showcased by the 24 clients who were between the ages of 63 and 78 (inclusive of both). The pension age in Portugal in 2008 was 61.5 years, allowing clients above or at this age to receive their payments (Campos & Pereira, 2008, p. 44). This external data assists in classifying these 24 clients as outliers, as even with an assumed pension, they still have an active mortgage. Furthermore, it is highly unlikely that these clients within the 63-78 age range were purchasing new houses during the global financial crisis.

During this period, Portugal was also going through pension reform, by setting a transitionary period to slowly increase the retirement age, and therefore the pension payouts (Campos & Pereira, 2008). During the telemarketing campaign, the retirement age increased from 61.5 to 62.5, in half-year increments. This may have also impacted those that were close to retirement age, but still had an active mortgage, making them less enthusiastic to sign up for a term deposit. This is corroborated in Figure 25 below, where the people with active mortgages who agreed to a term deposit, are proportionally less than those who did not have one. This characteristic has been highlighted by a black oval encapsulating the column's data points.

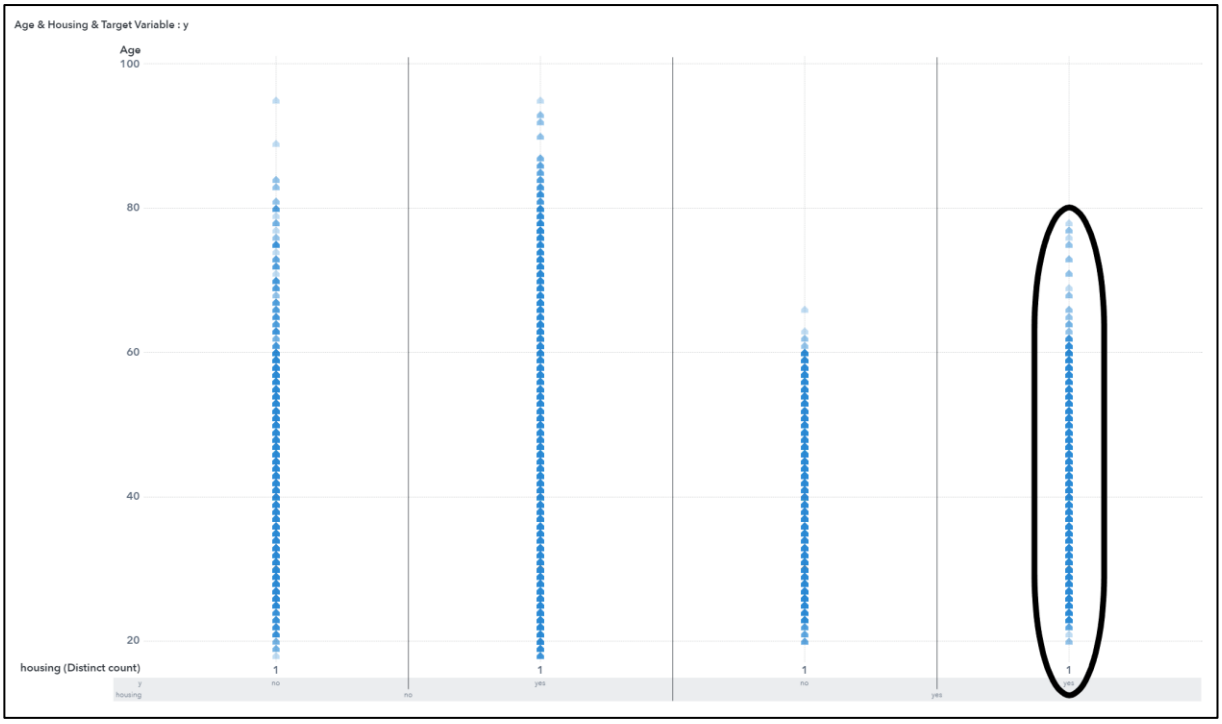

Figure 25: Age, Housing & Target Variable: Y – Scatter Plot

4.1.2. Balance and Job

The scatter plot depicts the association between the average annual account balance (in Euros), and the kind of occupation. The plot moderately suggests that those with higher-paying jobs such as managers and technicians tend to have higher average yearly account balances. Interestingly, retirees also tend to have relatively high average bank balances, which may represent their access to an influx of cash flow from their retirement pensions. Holistically, the average yearly account balance and occupation type do not have a generalised linear relationship. The average yearly account balance does not increase linearly with occupation type. For example, the average yearly account balance for management is not twice the average salary of a housemaid, and therefore the relationship is largely comparative, rather than direct.

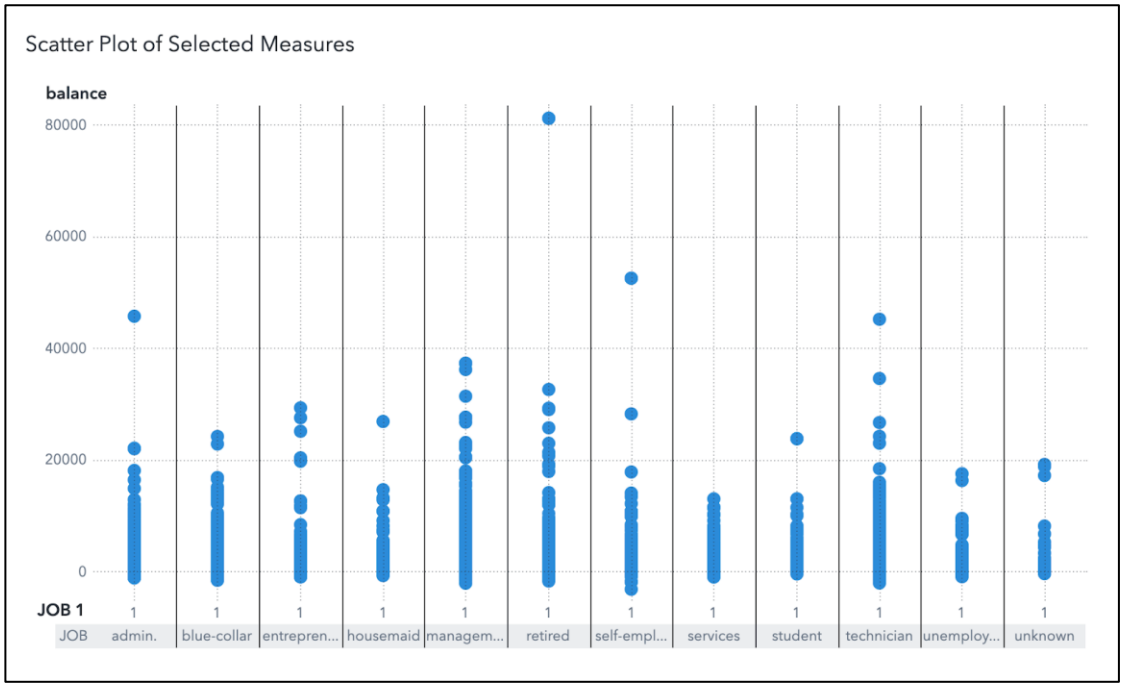

Figure 26: Balance & Job – Scatter Plot

4.1.3. Balance and Duration

The scatter plot provides insight into the relationship between a client's average yearly account balance (measured in Euros) and the duration of their last call with the bank (in seconds). The x-axis of the plot represents the duration, while the y-axis captures the balance. From the plot, it's evident that there's a positive correlation between the two

variables, suggesting that clients with higher balances generally spend less time on calls. This tendency might be due to their financial flexibility, enabling them to make quick decisions whether to sign up for a term deposit, as they may have a plethora of other investment strategies. Contrastingly, there are of several outliers that indicate an inverse relationship. Notably, there are points that represent target clients with low balances but with lengthy call durations, as depicted by the tail towards the bottom right of the plot. Such patterns may indicate clients facing financial difficulties and seeking assistance on how effective a term deposit may be for them.

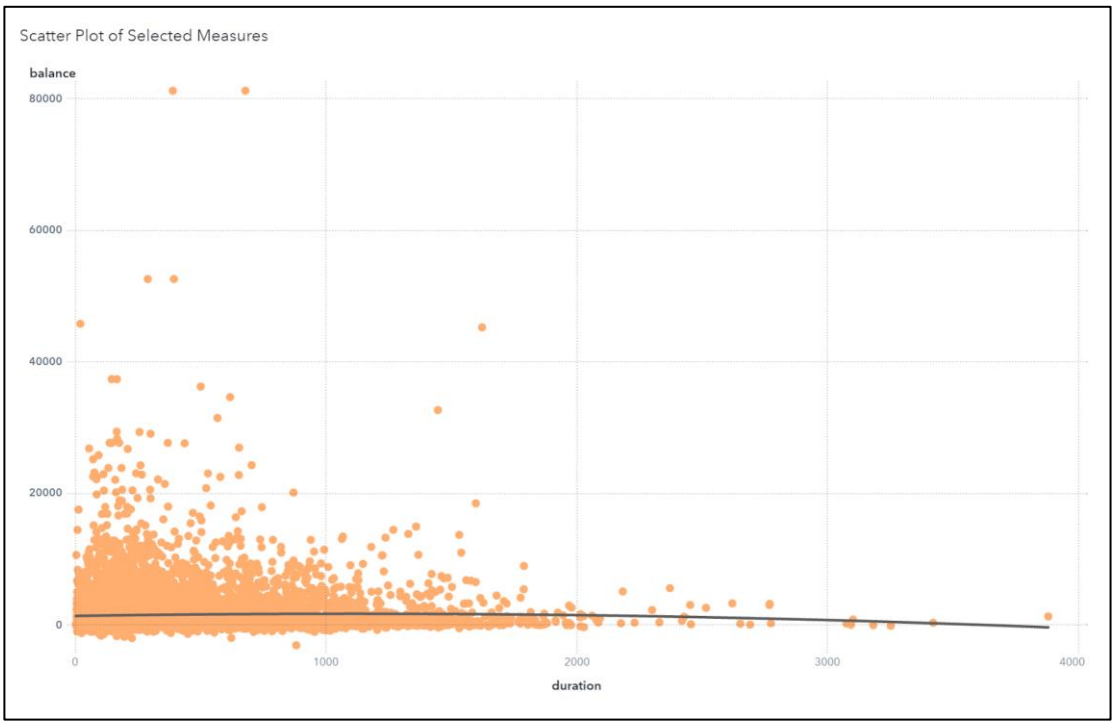

Figure 27: Balance & Duration – Scatter Plot

4.1.4. Campaign and Poutcome

To plot two qualitative attributes together, the parallel coordinates chart was used. In the visualisation, it can be seen that most unknown values of *poutcome* are related to having a value less than 5 in *campaign*. It is also evident that most campaign values fall under 15, and all values higher than 15 result in an unknown value for the poutcome variable. Therefore, it can be deduced that the previous marketing outcome for most target clients who were contacted 5 times or less, is unknown. The same is true for all clients who were contacted more than 15 times. Additionally, as shown below, it is clear that success, other or failure outcomes, do not have an explicit distribution abnormality, and instead are relatively equal. This indicates that other factors identified in previous relationships have a greater impact on the success of previous campaigns.

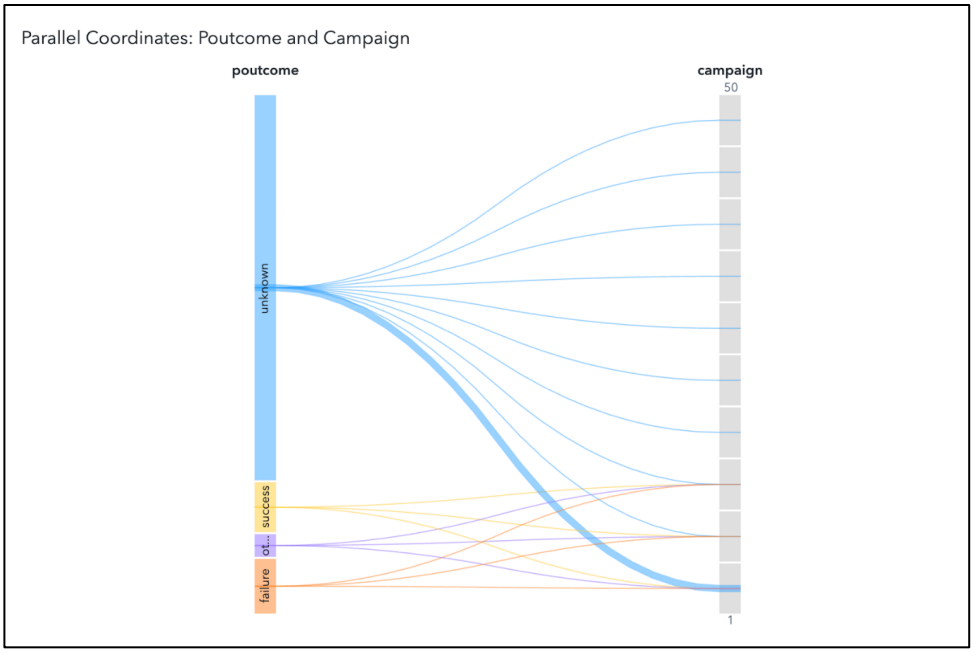

Figure 28: Campaign & Poutcome – Parallel Coordinates Chart

4.1.5. Age and Education

The scatter plot shown in Figure 29 below, indicates that most age groups are from 20 - 80 years old, with a few outliers above 90. Figure 29 also indicated a weak negative correlation between age and education level. A possible thesis for this statement is that the younger generation must undertake mandatory education up to the secondary level in Portugal. The scatter plot reveals that the maximum age for tertiary education is 84. Moreover, only a minority of people in this education group are aged over 80, again reinforcing the potentially new education standards set out for the younger generations.

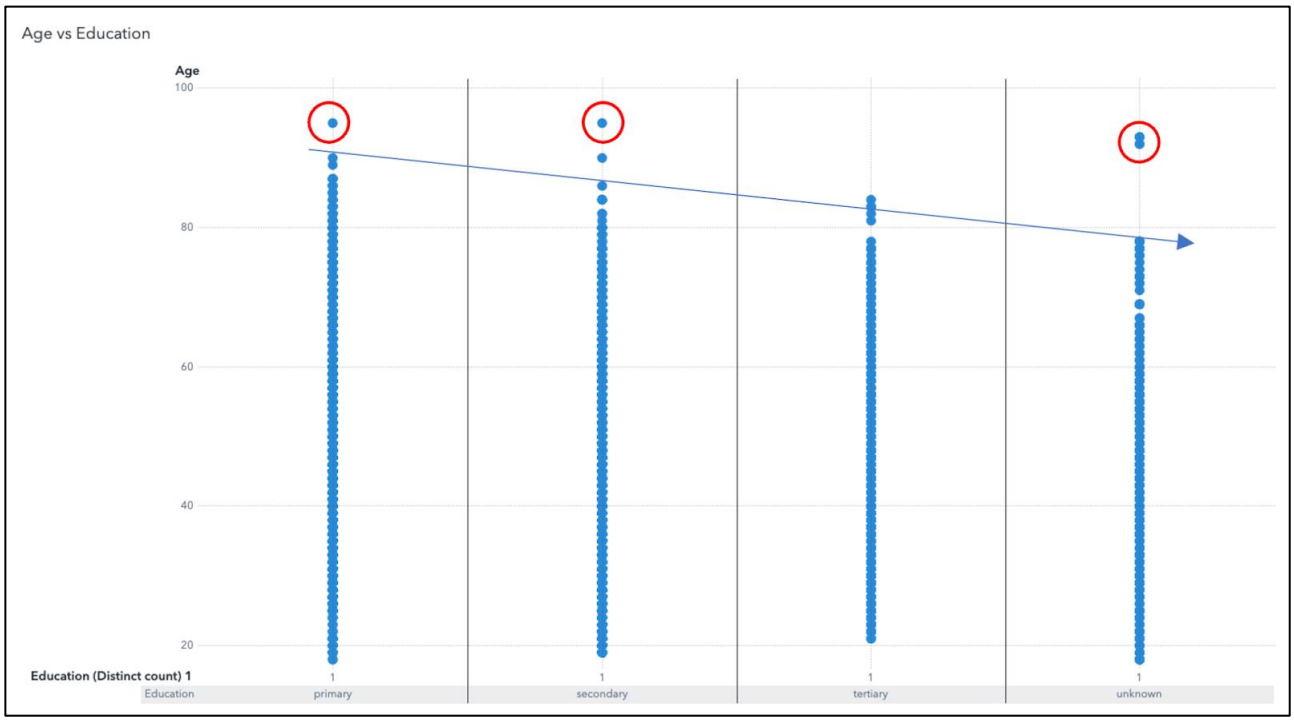

Figure 29: Age & Education – Scatter Plot

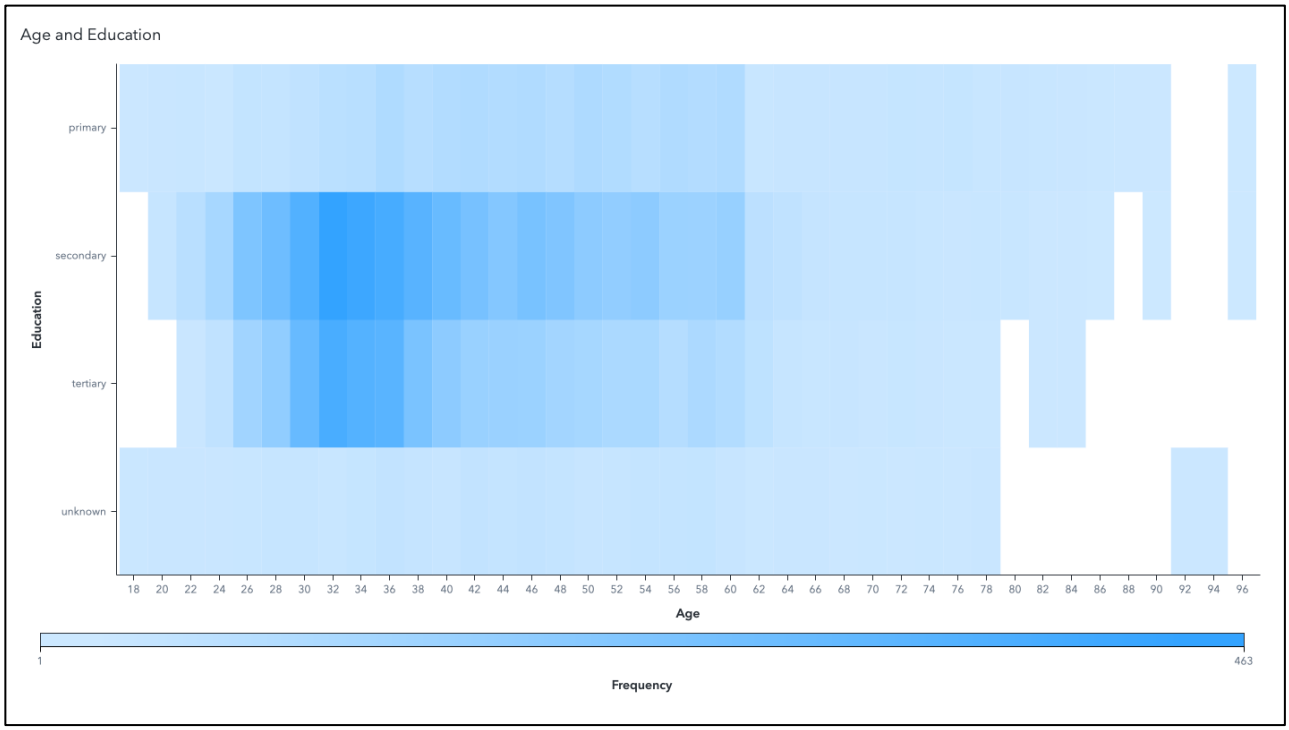

Figure 30: Age & Education – Heat Map

The heat map in Figure 30 also indicates that most people with tertiary education levels are from 26 to 40 years old, with a frequency of approximately 200 - 400 individuals. With a similar frequency, secondary education starts at the same age of 26, but spread out to nearly 60. The distribution of the other two categories is dispersed among all ages with no apparent pattern.

4.1.6. Age and Balance

According to the scatter plot in Figure 31, there is a clear curve, that tapers towards the right due to outliers. Most account balances are below 20,000 euros across all ages, and most ages are in the range of 25 to around 60. This age range is the universal working age range, as well as the approximate working range in Portugal. Target clients below the age of 24, do not have bank balances of greater than 10,000 euros. This is understandable, as most of these individuals may be students or have recently graduated. Potential clients aged 60 or older, have significantly reduced average yearly balances, by up to approximately 50%, to around 10,000 euros. This loose downwards trend creates a weak-moderate inverse relationship, with the increase of age and the decrease of balance. This observation may be due to retirees not earning regular salaries, and instead using their pensions to travel or live a better lifestyle.
The next group of bank balances are from 20,000 - 40,000 euros, potentially representing a group with more stable jobs and higher salaries. Most of this group also is between 25 and 60 years old. The distribution highlights one extreme outlier circled in red, with a balance of 81,204 euros, at the age of 84. This could mean that the individual has a very strong financial situation, or in an unlikely case, there may be a chance the data was incorrectly recorded.

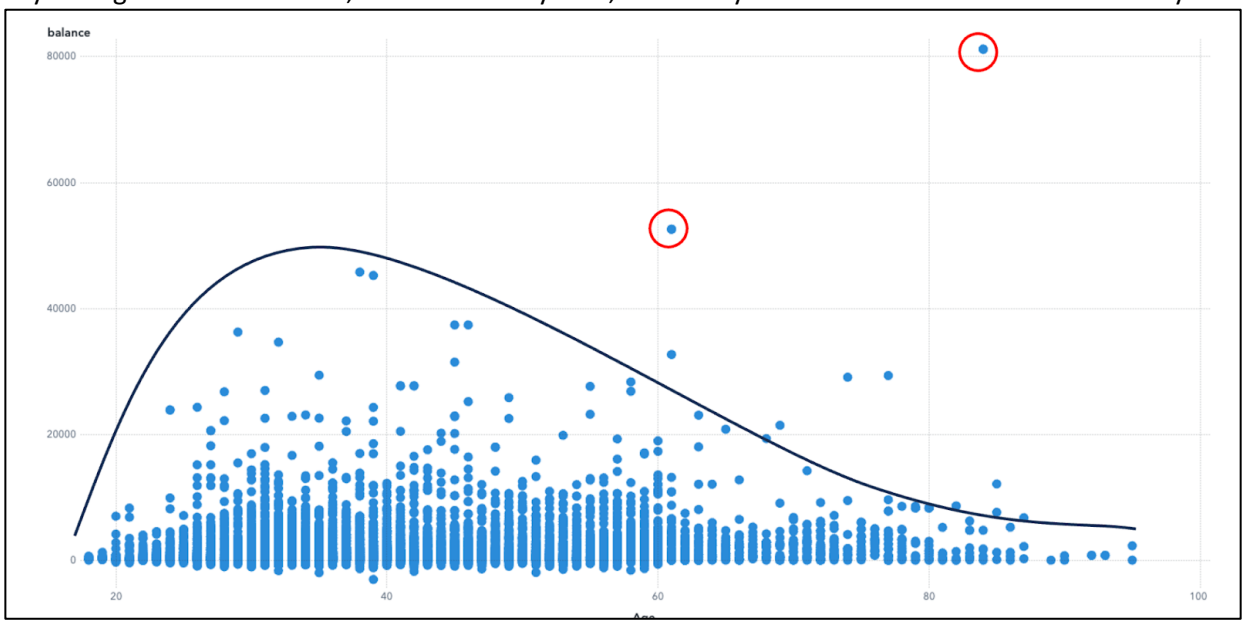

Figure 31: Age & Balance – Scatter Plot

4.2. Multivariate comparisons

4.2.1. Balance, Education and Job

Figure 32 illustrates a multivariate relationship, with job on the x-axis, balance on the y-axis, and education type as the colour of each data point.

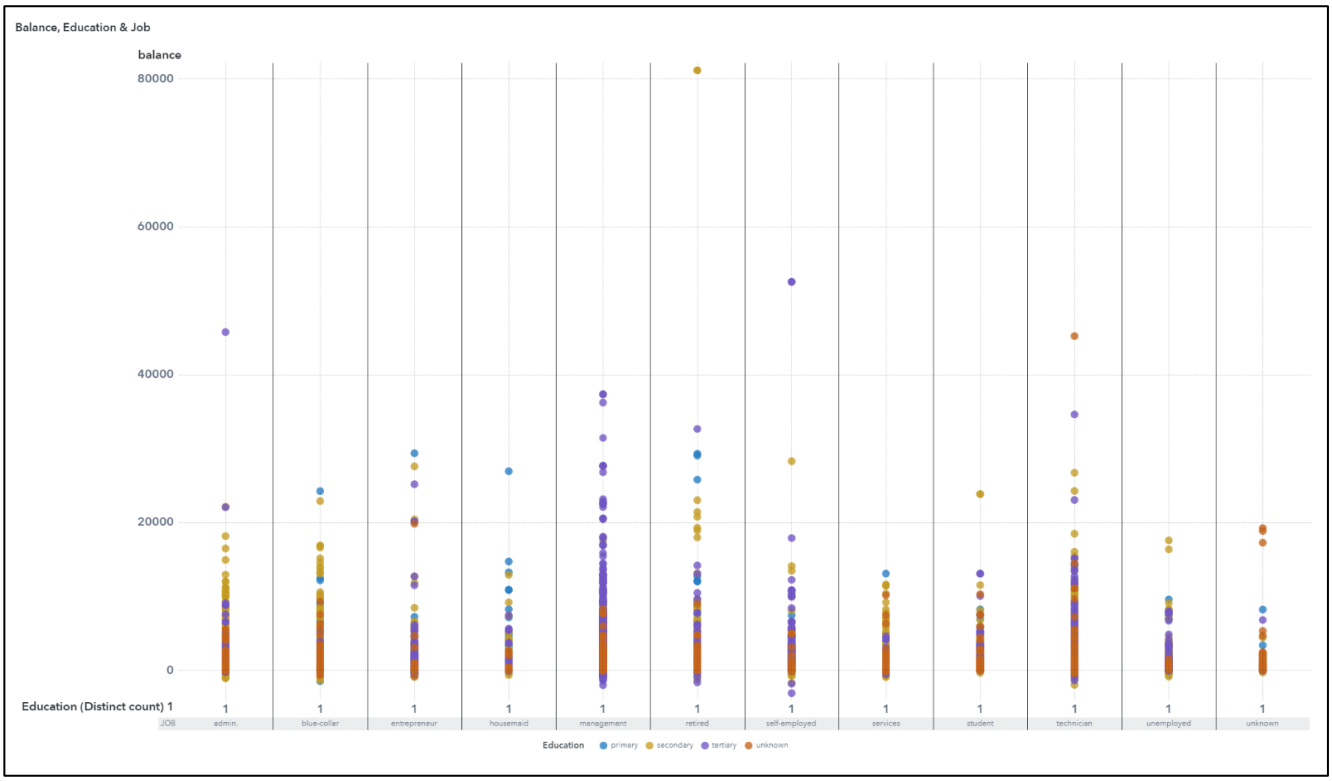

Figure 32: Balance, Education & Job – Scatter Plot

The scatter plot indicates that the majority of people in management roles tend to have the highest bank balances, followed by technicians and blue-collar jobs. This also corroborates the narrative suggested by Figures 3 & 4, where these three roles were the most common in the dataset. This again alludes to the possibility that the bank would prefer clients with stable, high-income jobs, instead of more flexible, but potentially insecure sole-trader roles, such as student or entrepreneur.

Another interesting find is that management roles seem to have higher education levels, with clients completing their tertiary studies. A similar trend is seen in technician roles, albeit to a lesser degree. This suggests that these fields prefer clients with higher education statuses, and therefore attract more competition, and therefore a higher salary. This notion could also justify the bank's disproportionate distribution of data points, in favour of these roles. Contrastingly, whilst blue-collar jobs are also in the top three job types of potential clients, they have minimal tertiary education points, and instead illustrate either unknown or secondary. This coincides with the fact that many blue-collar roles such as electrician, plumber and landscaper do not require a formal tertiary education to provide certification.

4.2.2. Balance, Default and Education

To understand if there is a general trend between education type, personal loan defaults and balance, a scatter plot is used to plot categorical variables education and default on the x-axis, and balance on the y-axis. This multivariate relationship has been depicted in Figure 33 below.

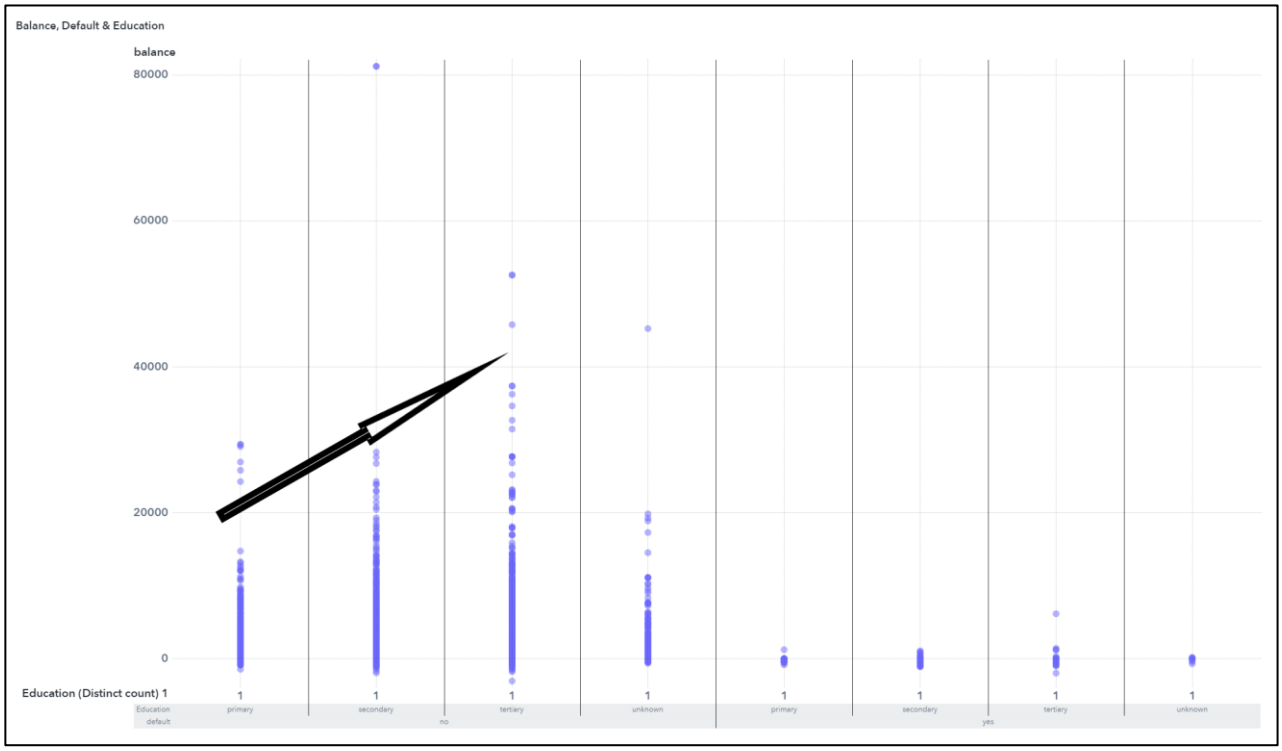

Figure 33: Balance, Default & Education – Scatter Plot

The plot highlights the blatant dichotomy between no and yes default categories, with no being the majority of data points. Within this category, the highest average yearly balance is more closely related to potential clients with tertiary education levels, followed by secondary. This suggests that higher education levels attract higher average salaries. The big difference in 'no' and 'yes' default values proposes that the bank prefers contacting clients that have not defaulted on their personal home loans, and may be more financially flexible. This intentional dichotomy may be strategic on the bank's plan of targeting financially literate clients, who do not like delaying their loan repayments, likely making them loyal customers. Unknown education types that exist within the dataset, may be a resultant of data input error or a lack of recent communication with the potential client.

4.2.3. Age, Balance, Job and Marital

Plotting these attributes in a multivariate scatter plot helps uncover a few weak correlations and trends. In this case, it is evident that the majority of job types do not have their marital status skewed towards a certain type, and instead highlight a relatively proportionate distribution. Two categories diverge from this notion, with 'retired' consisting of predominantly married clients, and 'student' being largely single. This intuitively matches reality, in

which retirees are more likely to be married, whilst students are not. Similarly, the age distribution of retirees is mainly over 60, whilst students are in the 18-50 range. Another interesting finding was that the preponderance of clients who had an average yearly balance of above 20,000 euros, were married. This is not indicative of a strong correlation; however, it could be symbolic of a very weak one. Resultantly, it cannot be said that the bank puts significant effort into telemarketing their term deposit plans, to a higher proportion of married people.

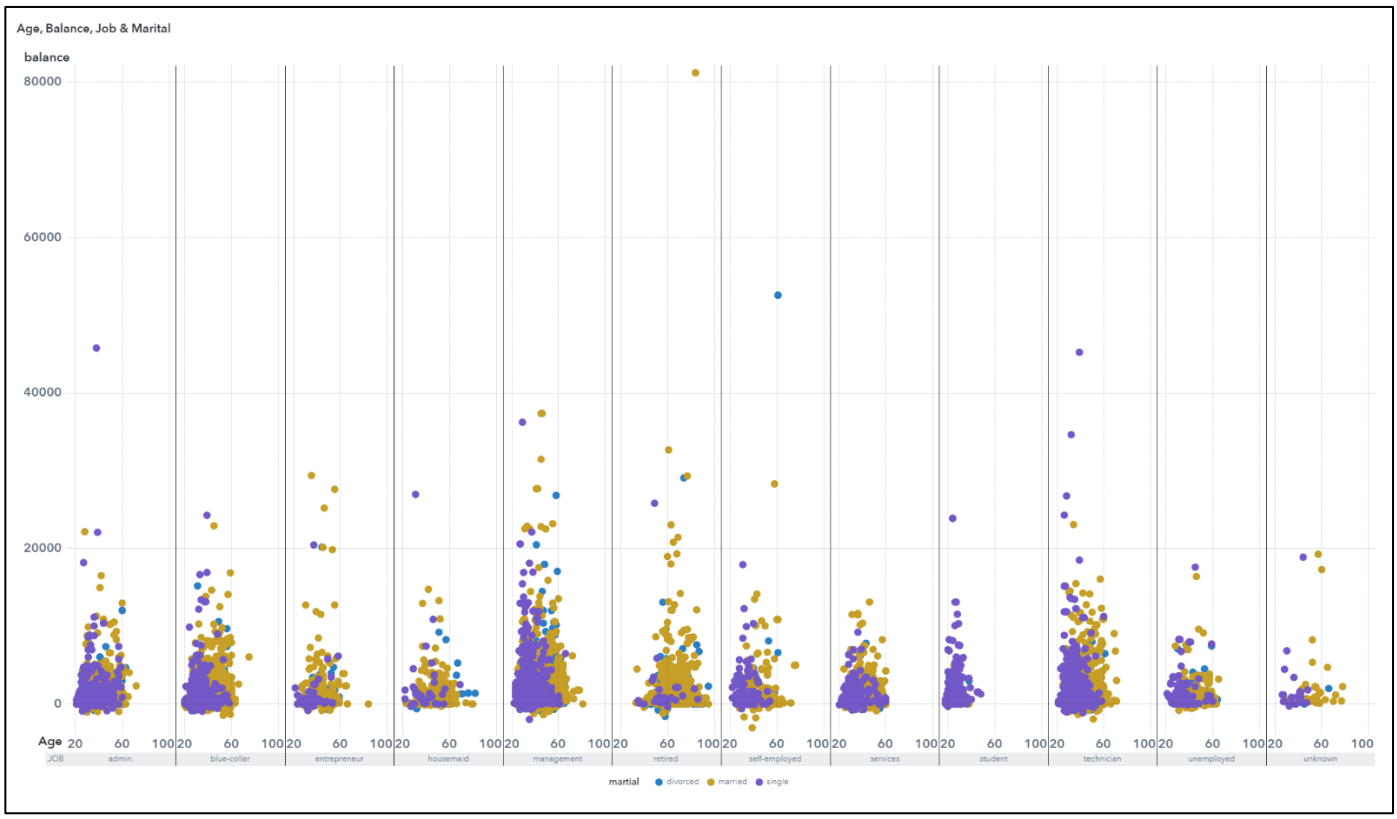

Figure 34: Age, Balance, Job & Marital – Scatter Plot

4.2.4. Balance, Housing and Loan

Figure 35 below, highlights the correlation between balance, personal loans, and housing loans. As shown by the blue dots, potential clients who have no personal loan, also tend to not have any active mortgages. This suggests a weak relationship between both variables. It is also evident that target clients that don't have either type of loan, tend to have higher balances, when compared to those who have an active mortgage. Whilst this an explicit observation, it does not have a strong correlation. This may also be indicative of the fact that mortgages may be more financially troublesome than personal loans, therefore, those with mortgages, have slightly lower average bank balances. The distributions of personal loan categories across those with and without active mortgages is too similar to derive any further insight.

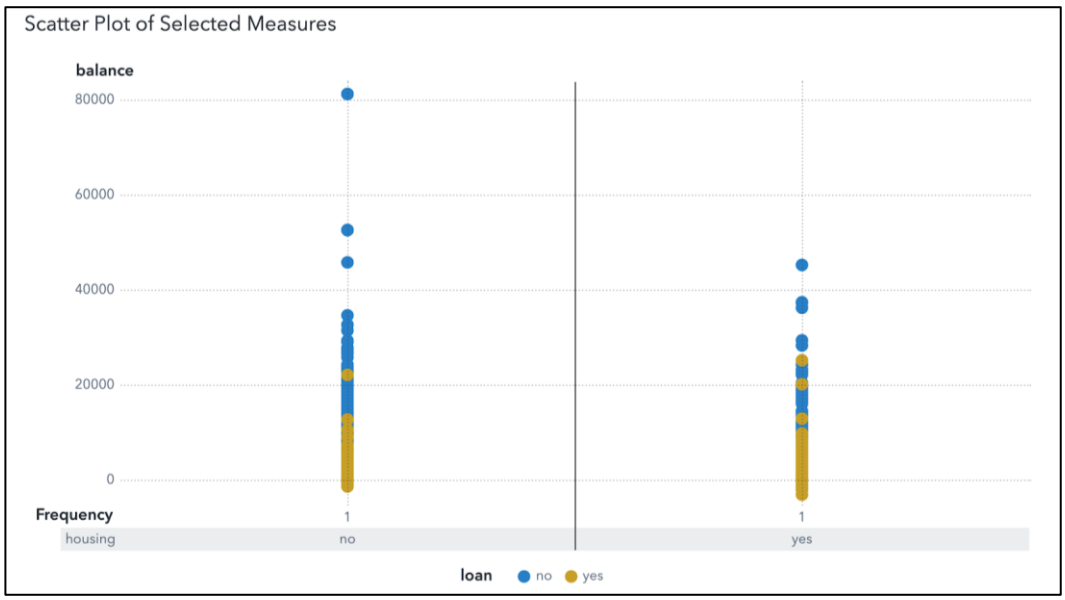

Figure 35: Balance, Housing & Loan – Scatter Plot

4.2.5. Campaign, Contact, Duration and Poutcome

In Figure 36 below, subplots containing '*poutcome*' values of failure, success and other; are grouped towards the left, indicating longer call durations, but fewer contacts during the campaign. The blue points are predominant, suggesting that most contacts during failed, other and success categories, were via cell phones. A few yellow dots can be observed, indicating a lesser reliance on landline communication during these interactions. The purple (unknown) dots, although present, are quite sporadic, showing infrequent instances of unidentified contact methods in these outcome categories.

The unknown subplot stands out both in its size and colour distribution. The most striking feature here is the dominance of purple (unknown) dots, suggesting that campaigns with unidentified outcomes also had a significant portion of interactions with unknown contact methods. This may refer to poor data input templates, not having specific contact options, or data collection errors. Similar with other plots, blue (cellular) dots are still prevalent, indicating that even in cases with uncertain outcomes, cellular communication was prominent. There is a general downward trend, indicating that the higher number of contacts made during the current campaign, the lower the duration of the call.

Across all scatterplots, there's a discernible pattern: data points are majorly concentrated towards the left side, suggesting that irrespective of the campaign's outcome, most interactions were of relatively longer durations, but involved fewer contacts during the campaign (y-axis). Cellular communication remains the preferred mode across all outcomes, highlighting its significance in the bank's communication strategy. Cellular calls also exhibited a generalisation across all plots, with this contact method consisting of the longest durations. This may have been intentional, so that potential clients have a greater likelihood of answering the call. Contrarily, the only exception is the 'unknown' outcome, where unidentified contact methods overshadow other modes. This hints at a potential discrepancy caused by input issues that transcended data collection.

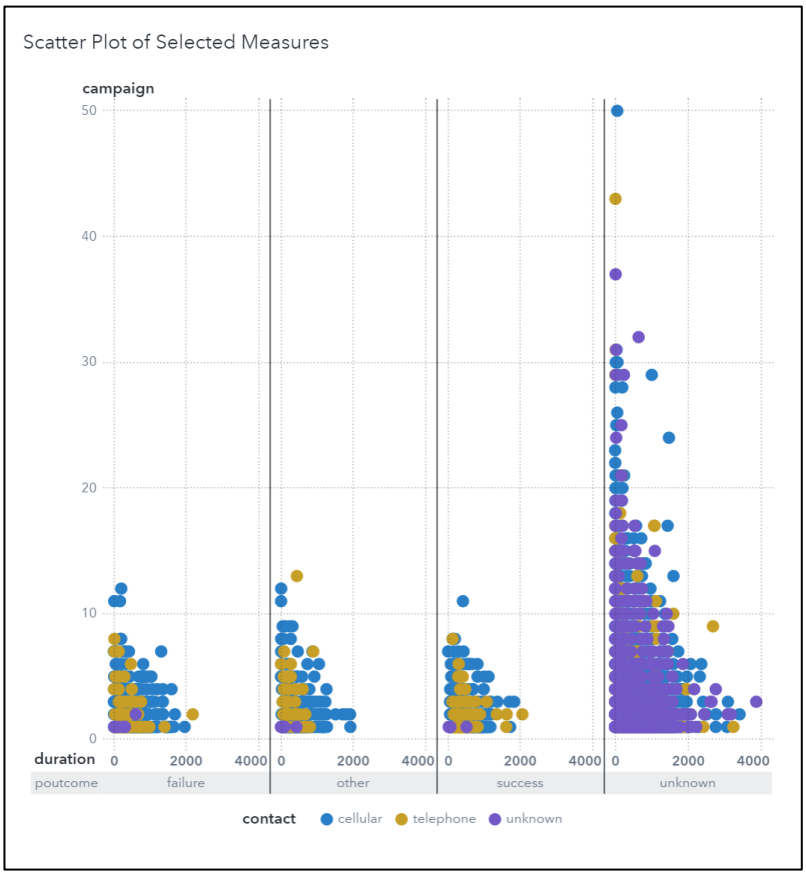

Figure 36: Campaign, Contact, Duration & Poutcome – Scatter Plot

4.2.6. Balance, Pdays and Poutcome

The scatter plot below visualises the relationship between balance, pdays and poutcome. One interesting finding is that all unknown 'poutcome' values are directly connected to PDays, as they all represent target clients which were never contacted in a previous campaign. This indicates that the outcome of the previous campaign is only unknown when the targeted client in that campaign was never contacted. This is a very intriguing discovery, as it provides an explanation for over 70% of the unknown 'poutcome' values present in the dataset.

Another finding is that the clients with higher balances are contacted more recently, particularly between 1-273 days prior to the current campaign. This is dissimilar to the majority of target clients who possessed an average yearly balance of less than 20,000 euros, and were contacted more than 273 days prior to the current campaign. This solidifies the narrative that the bank is more likely to recontact potential clients with higher balances, in the hopes of persuading them to sign up for a term deposit. There are also two main outliers present who have more than 80,000 euros as their average yearly balance. One of these target clients was contacted between 1 – 273 days prior to the current campaign, with the other being contacted more than 273 days prior. Another outlier exists with a 50,000 balance value, albeit significantly smaller.

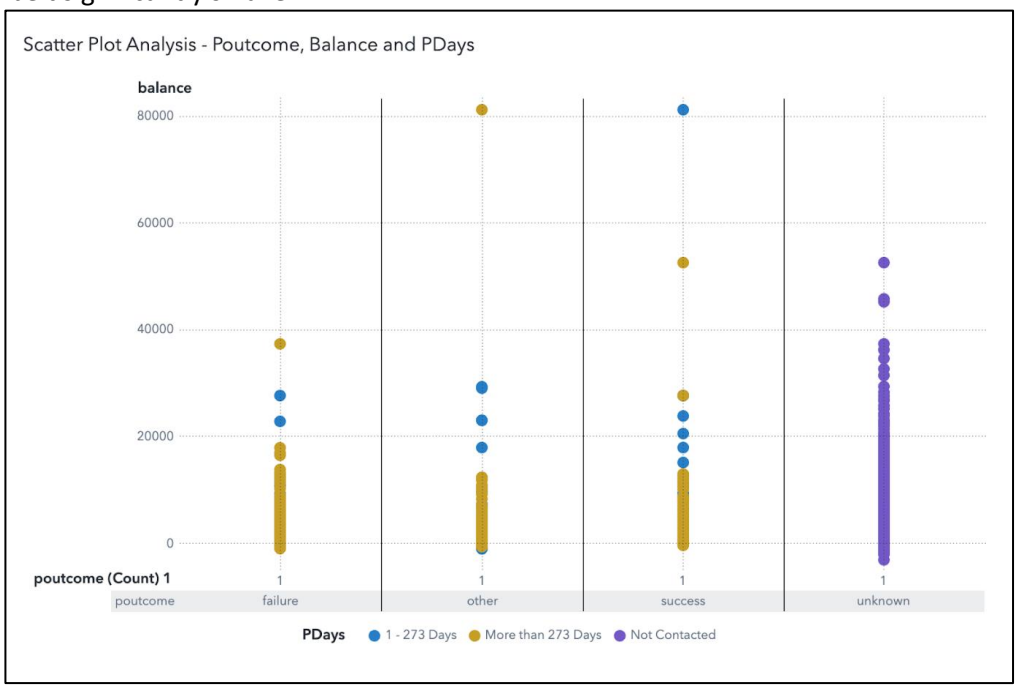

Figure 37: Balance, Pdays & Poutcome – Scatter Plot

4.2.7. Default, Poutcome and Previous

In the scatter plot, it can be seen the almost all unknown values have a 'yes' value for the default attribute. It's also evident that a number of failure values for 'poutcome', also have the default value of yes. A small portion of people of have defaulted on their credit loans, are observed in the 'other', *poutcome* category. In contrast, it is observed that a significant portion of people with a successful 'poutcome' value, have not defaulted on their credit loans. This suggests that those who have not defaulted on credit debt are more likely to have been successful candidates for the previous marketing campaign, indicating that they may have more available cash flow for a term deposit. It can also be seen that all clients who were deemed successful in the previous marketing campaign were contacted less than 25 times. Therefore, it is apparent that too many contacts led to failure or other outcomes for the previous marketing campaign. Interestingly, there is an outlier who was contacted more than 250 times, with an outcome of 'other'.

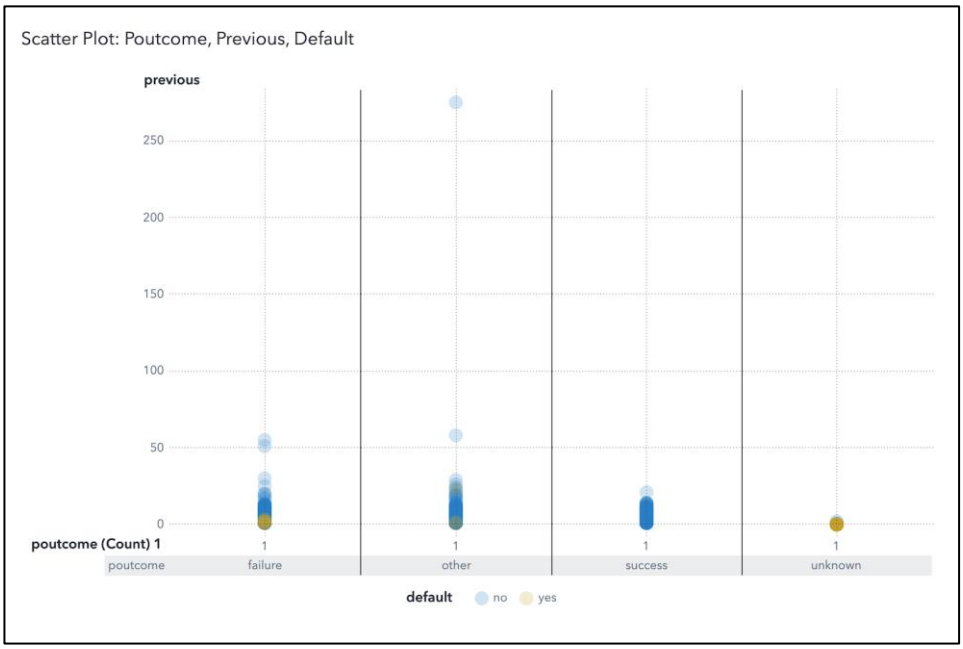

Figure 38: Default, Poutcome & Previous – Scatter Plot

4.2.8. Age, Balance and Loan

Figure 39 indicates that older adults above 60 are less likely to have an existing loan. Only 10 out of more than 10,000 individuals over 60 have a personal loan, as shown in red. This demographic may have a lower risk tolerance, meaning that they are avoiding debt in case unable to repay it from their pensions.

Having a lower bank balance is symbolic of having a personal loan, particularly for the younger age demographic, as explicitly depicted in Figure 40. Making monthly interest payments towards these loans may minimise their bank balance, and motivation to sign up for a term deposit. A subset of the older demographic also has low bank balances, which may indicate their ongoing payments for a personal item, as retirees may not be earning income, and instead are relying on pension payments. Additionally, those with no personal loan commitments are more equally distributed amongst the balance range, and resultantly, no explicit trend can be derived.

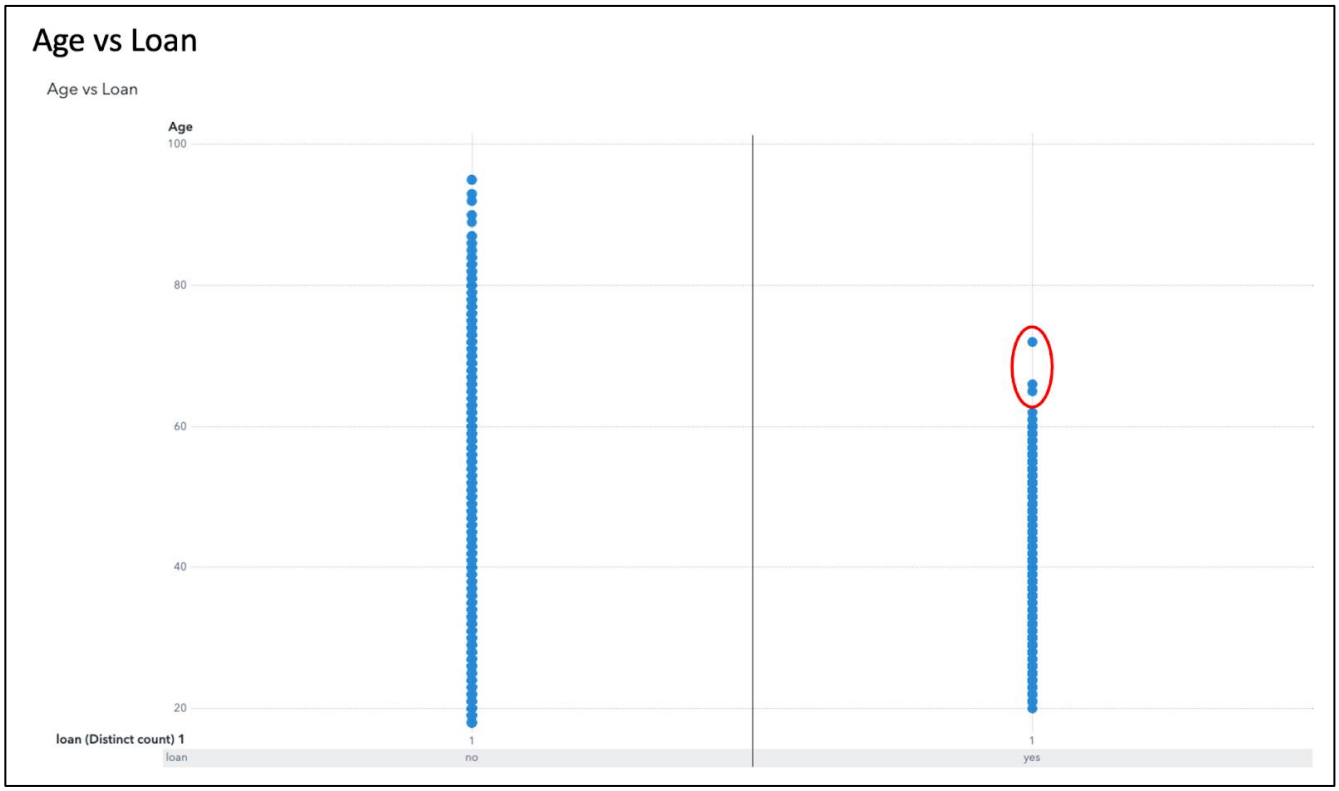

Figure 39: Age & Loan – Scatter Plot

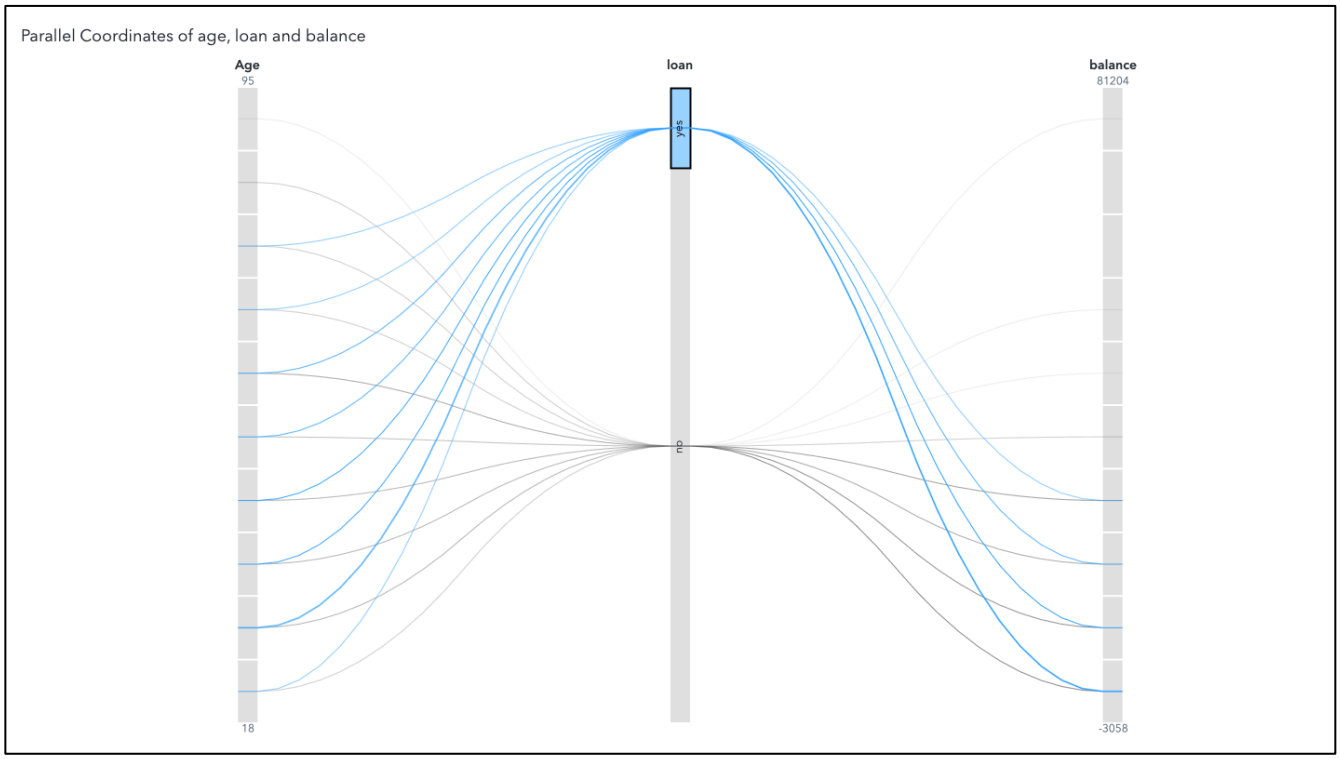

Figure 40: Age, Balance & Loan – Parallel Coordinates Chart

5. Data pre-processing

Data pre-processing involves performing modifications on the input dataset, 'BANK_DIRECT_MARKETING'; so that the data can be more accurately represented based on initial data exploration, and comparative analyses. This process also aims to streamline the dataset by conducting data cleaning, transformation, and dimensionality reduction, so that abnormalities are dealt with prior to predictive modelling.

5.1. Data cleaning

Data cleaning involves ensuring that all records within the BANK_DIRECT_MARKETING dataset are free from corruption, duplication, or omission. As shown in Figure 41 below, there are 24 Age data points missing. Aside from this, the dataset contains no duplicates or corrupt data that requires cleaning.

Obs	Variable Name	Role	Measurement Level	Order	Label	Count	Number of Missing Values
1	Age	INPUT	INTERVAL			75	24
2	Education	INPUT	NOMINAL			4	0
3	JOB	INPUT	NOMINAL			12	0
4	_PartInd_	PARTITION	NOMINAL		Partition Indicator	3	0
5	_dmIndex_	KEY	NOMINAL			254	0
6	balance	INPUT	INTERVAL			254	0
7	campaign	INPUT	INTERVAL			34	0
8	contact	INPUT	NOMINAL			3	0
9	customer_id	REJECTED	INTERVAL			254	0
10	day	INPUT	INTERVAL			31	0
11	default	INPUT	BINARY			2	0
12	duration	INPUT	INTERVAL			254	0
13	housing	INPUT	BINARY			2	0
14	loan	INPUT	BINARY			2	0
15	martial	INPUT	NOMINAL			3	0
16	month	INPUT	ORDINAL			12	0
17	pdays	INPUT	ORDINAL			3	0
18	poutcome	INPUT	NOMINAL			4	0
19	previous	INPUT	INTERVAL			32	0
20	y	TARGET	BINARY			2	0

Figure 41: Missing Value & Attribute register

These 24 Age values have been imputed with the mean for the attribute. This was completed with the imputation node, with a mean value of 41, which was rounded down from 41.2268. This was done to conform with the Age attribute's integer numbering scheme, and can be seen below in Figure 42.

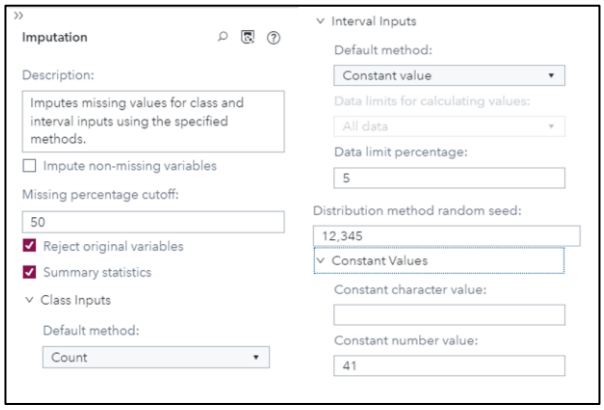

Figure 42: Imputation Node Parameters

5.2. Data transformation

Data transformation involves directly modifying suitable attributes after they have gone through any cleaning processes. Given the lack of data quality abnormalities within the BANK_DIRECT_MARKETING dataset, these transformations will be minimal, and are segregated into binning & data type conversions (SAS, 2021).

5.2.1. Data type conversion

To create a modicum of normalisation across all binary attributes, they were level encoded using a *SAS Code* node to assign numerical and nominal properties to them. These attributes include default, housing, and loan, as shown in Figure 43 below.

```
/* Transform all binary attributes to numeric + nominal */
/* Transformation Method = LEVELENCODE */
Length 'LEVENC_default'n 8;
Label 'LEVENC_default'n = 'Transformed_default_level_encoded';
Length _val_80926280 $5;
Drop _val_80926280;
_val_80926280 = ktrim(put('default'n,$CHAR5.));
if _val_80926280 in ( 'no' ) then
    'LEVENC_default'n = 0 ;
else if _val_80926280 in ( 'yes' ) then
    'LEVENC_default'n = 1 ;

/* Transformation Method = LEVELENCODE */
Length 'LEVENC_housing'n 8;
Label 'LEVENC_housing'n = 'Transformed_housing_level_encoded';
Length _val_80926280 $5;
Drop _val_80926280;
_val_80926280 = ktrim(put('housing'n,$CHAR5.));
if _val_80926280 in ( 'no' ) then
    'LEVENC_housing'n = 0 ;
else if _val_80926280 in ( 'yes' ) then
    'LEVENC_housing'n = 1 ;

/* Transformation Method = LEVELENCODE */
Length 'LEVENC_loan'n 8;
Label 'LEVENC_loan'n = 'Transformed_loan_level_encoded';
Length _val_80926280 $5;
Drop _val_80926280;
_val_80926280 = ktrim(put('loan'n,$CHAR5.));
if _val_80926280 in ( 'no' ) then
    'LEVENC_loan'n = 0 ;
else if _val_80926280 in ( 'yes' ) then
    'LEVENC_loan'n = 1 ;
```

Figure 43: Binary level encoding conversion code

Additionally, both 'month' and 'Pdays' attributes were changed to an ordinal data type, to better reflect the business problem and data dictionary shown in sections 1 and 3.1, respectively. *Month* was modified to ordinal, to better reflect the standard Gregorian calendar. *Pdays* was changed to ordinal to demonstrate the tiered hierarchy of its durations, as mentioned in section 3.2.15. These modifications may assist in providing more meaningful data during predictions made with neural networks. This may highlight external factors such as the end of financial year, influencing the likelihood of subscribing to a term deposit.

☑	month	Character	Input	Ordinal	Default
☑	pdays	Numeric	Input	Ordinal	Default

Figure 44: Categorical data type modification register

5.2.2. Binning

To perform discretisation of suitable interval attributes in the dataset, it is important to consider binning to reduce absolute values. This has been employed for two main attributes within the dataset: duration and balance. These were chosen specifically to reduce their extremely high-level count of greater than 254, as shown below. To resolve this, equal-width binning was selected, with a total bin count of 15 each. Generally, when using equal-width binning for attributes with high kurtosis values, and platykurtic distributions; outliers may not be accurately represented (LinkedIn, 2023, para. 4). As '*balance*' falls within this description, it is important to combat this without the use of standard bucket binning. Alternatively, tree-based binning was used to optimally segregate values in respect to the target attribute, '*y*'; perhaps providing better accuracy, whilst reducing the number of records (SAS, 2018, para. 3; SAS, 2021). As the analysis process can be cyclic in nature, permutations of this transformation method may be employed when determining the best model. A snippet of this binning code is shown in Figure 46 below.

☐	Variable... ↑	Type	Role	Level	Order	Number of Levels
☐	balance	Numeric	Input	Interval	Default	>254
☐	duration	Numeric	Input	Interval	Default	>254

Figure 45: Attributes with high level counts

```
* Transformation Method = TREEBIN ;
Label 'BIN_balance'n = 'Transformed_balance_tree_binned';
Length 'BIN_balance'n $22;
if missing('balance'n) then
'BIN_balance'n= '00:_MISSING_';
else
if 'balance'n <= -888.66 then
'BIN_balance'n = '01:low--888.66';
else
if 'balance'n <= 202.68 then
'BIN_balance'n = '02:-888.66-202.68';
else
if 'balance'n <= 748.35 then
'BIN_balance'n = '03:202.68-748.35';
else
if 'balance'n <= 1839.69 then
'BIN_balance'n = '04:748.35-1839.69';
else
if 'balance'n <= 5113.71 then
'BIN_balance'n = '05:1839.69-5113.71';
else
if 'balance'n <= 5659.38 then
'BIN_balance'n = '06:5113.71-5659.38';
```

Figure 46: Binning code snippet

5.3. Dimensionality reduction

To assist in reducing training time and to remove unnecessary attributes that have no meaning to the value or are highly correlated to each other, it is important to undergo the process of dimensionality reduction. In this dataset, 'customer_id' has been assigned a rejected role (see Figure 47), as it does not provide meaningful data to future predictions, given its sole purpose of record labelling. This assignment will exclude the attribute from any future analysis in the model pipeline.

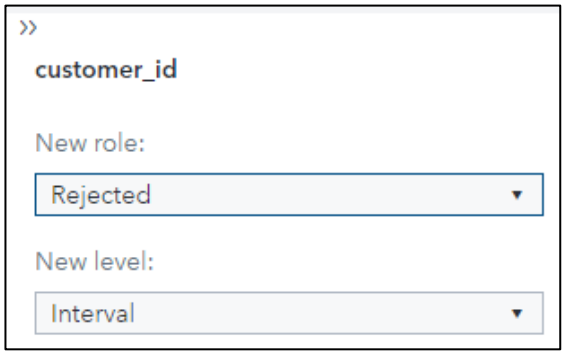

Figure 47: Customer_ID updated role assignment

It is important to note that the dataset has a significant amount of 'unknown' or 'never contacted' (-1) values in attributes such as Poutcome and Pdays. It was originally thought that these unknown values were due to data collection or data input templating errors, as mentioned in section 3.2. Following comparative analysis, it was uncovered that these values were in fact accurate, given the context of the business problem. Unknown and never contacted (-1) values in these variables are related to each other, as they indicate no prior communication, and therefore no prior term deposit registration. A similar notion is exhibited by the 'previous' attribute, where a '0' value denotes that these same target clients were contacted 0 times prior to the current campaign. As is evident, attributes such as these are related, even though the correlation score between them is not extremely high. For example 'previous' and 'Pdays' has a moderate correlation of approximately 41%, as shown in Figure 48. Prior to training, one of these attributes will be omitted to measure the potential dichotomy in model accuracy, when removed.

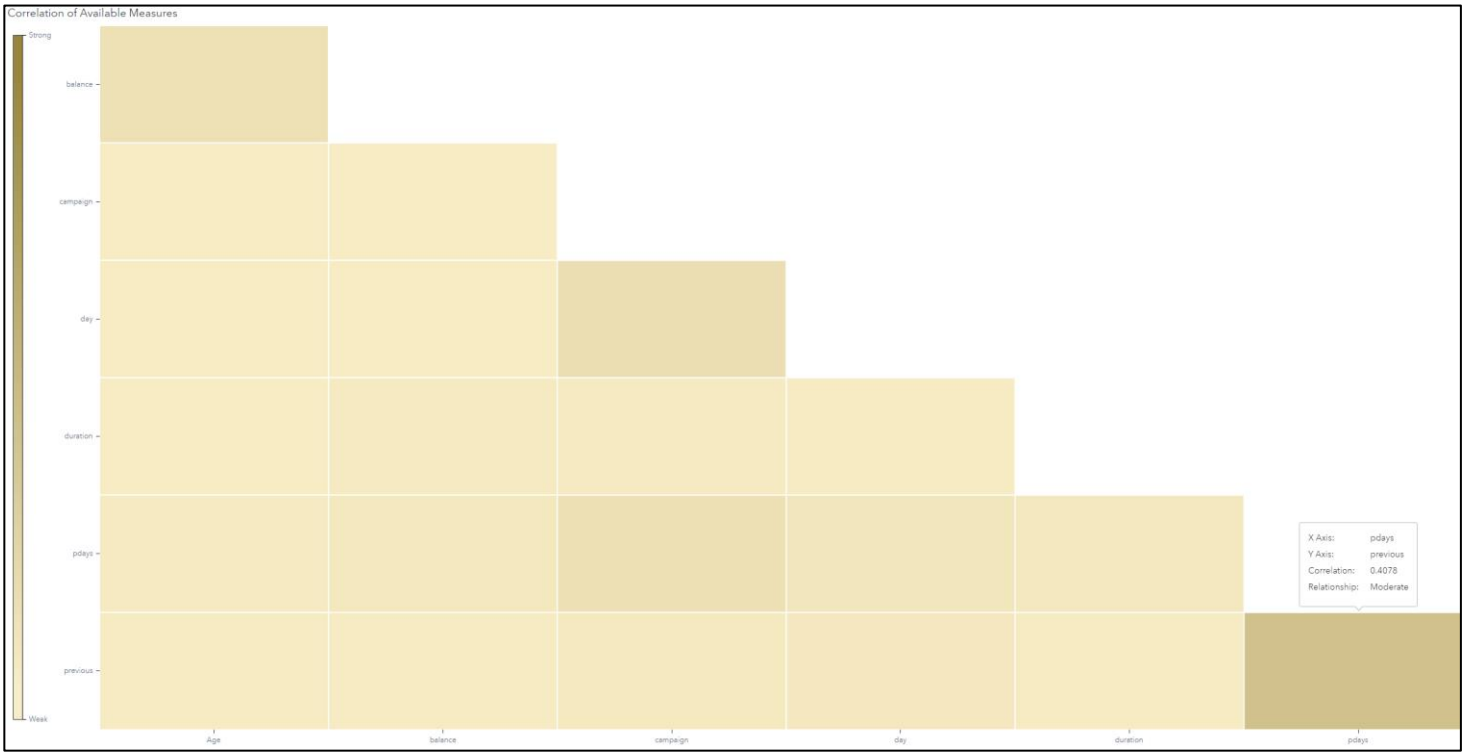

Figure 48: Correlation matrix

5.4. General cluster analysis

To verify that dimensionality reduction steps proposed in the previous section are appropriate for this dataset; clusters, features and correlations, will be examined after pre-processing, to determine if any additional modifications need to occur. Similar to randomised parameter selection for modelling, it is important to note that certain attributes may be omitted to ascertain any accuracy differences in results. Given the nature of this trial-and-error approach, it may act upon decisions made by educated guesses based on the business problem, or randomisation.

5.4.1. Correlation Matrix – SAS Visual Analytics

The correlation matrix shown above in Figure 48, plots compatible numeric attributes against each other in accordance with how closely related they are. Darker shades indicate higher correlations, with the highest being between 'pdays' and 'previous' at 41%. This suggests that both attributes have a relatively higher chance of providing the same insight, when determining if the client subscribes to a term deposit. As mentioned earlier these attributes will be used together and separately, so that any observations in accuracy differences between each combination can be examined. As Figure 48 does not have any highly correlated attributes, no further actions will be taken to omit specific attributes based solely on this matrix.

5.4.2. Relative Importance – Data Exploration Node

The data exploration node in SAS Model Studio utilises a tree-split algorithm to discover the importance variations between attributes in the data set, in respect to the target variable. More precisely, the data exploration node can calculate the relative importance, which denotes the predictive power of each compatible attribute. As shown below, the highest value is 1, belonging to 'duration'. This denotes that *duration* has the greatest influence on whether the customer will agree to a term deposit (*y*). Subsequent attributes such as the outcome of the previous campaign (*poutcome*), and the last contact *month,* both have weak scores of approximately 0.3. The remaining attributes have negligible relative importance scores, and resultantly, these may be used during the trial-and-error attribute omission approach, mentioned in section 5.4.

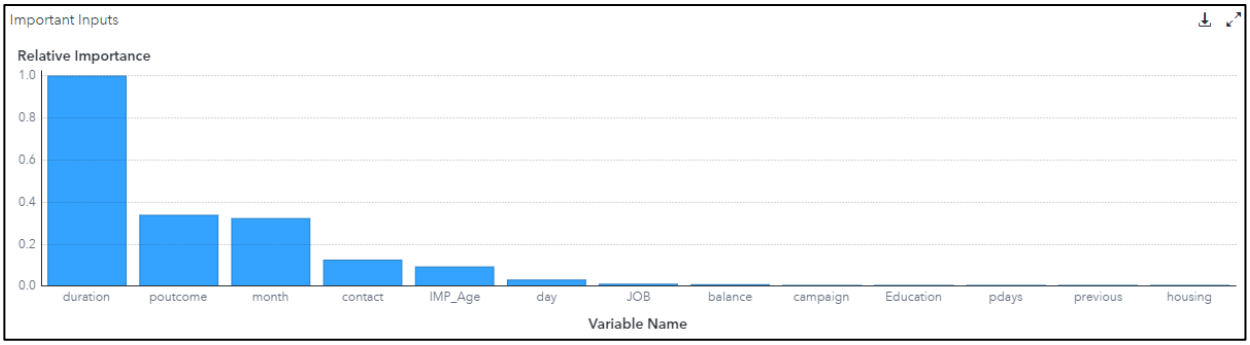

Figure 49: Relative Importance Histogram

5.4.3. Variable Clustering Node

The variable clustering node generates weighted graphs based on computed relation probabilities between various attributes and their subsidiary categories. Figure 50 below illustrates these clusters across all compatible attributes, configured with the use of 25 clustering steps, and an intra-cluster correlation coefficient (RHO value) of 0.8. To allow enough iterations to develop a more insightful cluster diagram, 25 steps was employed, as it sits in the middle of the 50-step limit. Selecting the RHO value is also important, and in this case, 0.8 was chosen. This creates a weighted balance, tilted towards ensuring that two data points randomly selected from within a cluster, are at least 80% more likely to have corresponding values, than two random points selected from the entire set (Sturgis, 2004, para. 7). Evidently, this is a measure of intra-cluster homogeneity, assisting in determining insightful connections (Search.r-project.org, n.d.).

Whilst convoluted, the graph below demonstrates that the majority of all inter-cluster relations are connected to the same attribute, albeit different category values. The greatest weighted connection is between *housing_yes* and *LEVENC_housing*. This is to be expected, as the former indicates if a potential client has confirmed their use of an active housing loan, whilst the latter is a numerical binarisation (completed in section 5.2.1) of the same attribute, i.e. yes = 1 and no = 0. This known relation explains its heavily weighted connection, with a value of 0.517. All other connections do not offer any further insight into inter-attribute relationships, but instead reinforces the general variance between each attribute, potentially aiding in accuracy. This justifies the dimensionality reduction steps carried out earlier, with no explicit modifications derived from this graph.

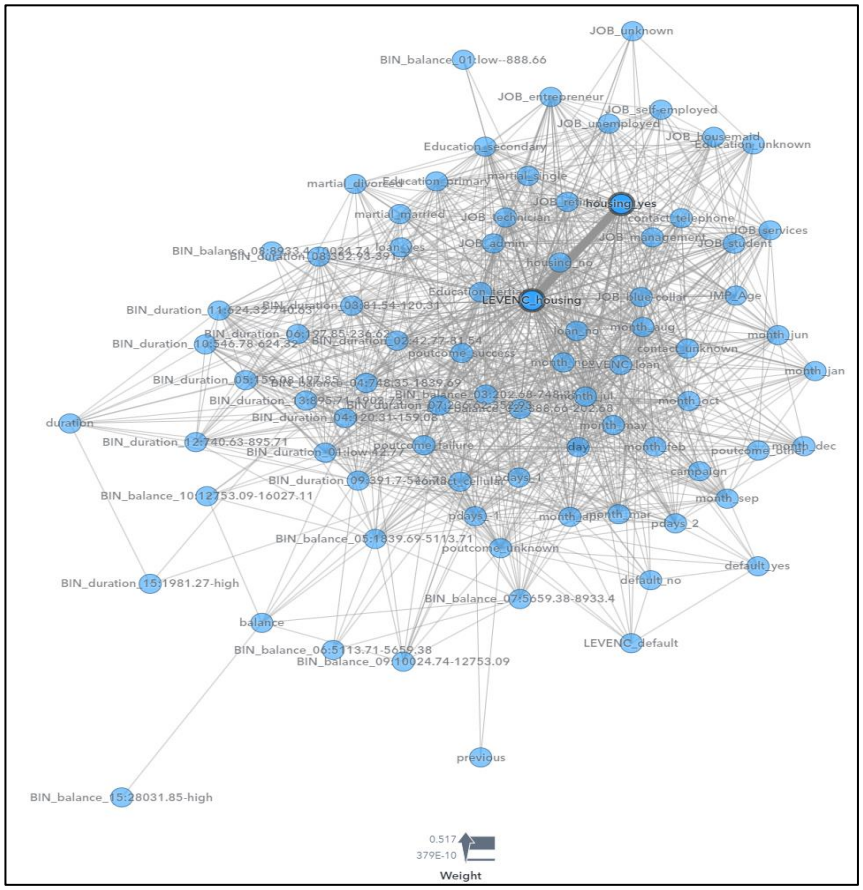

Figure 50: Variable cluster graph with all compatible attributes

To simplify the cluster graph in an attempt to uncover potentially hidden relationships, the penalised log-likelihood cluster selection method was used, as shown in Figure 51 below.

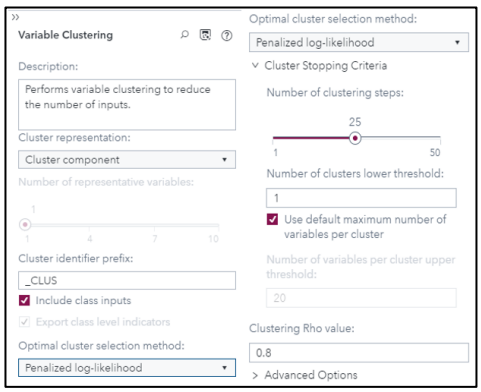

Figure 51: Variable clustering node configuration

The graph below is the output of this new configuration, and it highlights a few key relationships:

- Unknown 'poutcome' values have a strong relation to -1 (never contacted) 'pdays' values.
    - This connection reinforces the thorough examination conducted in section 5.3 (dimensionality reduction), regarding the same findings.

- Tertiary education has a relatively strong connection with the management job type, and a weak one with blue collar jobs.
    - This relationship also corroborates findings analysed in section 4.2.1 (balance, education, and job), where it was noted that people with management positions were more likely to have completed tertiary education, when compared to blue collar roles.

Other connections in the cluster graph are either negligible and do not provide further insight into the dataset, or are heavily weighted due to being intra-attribute categories.

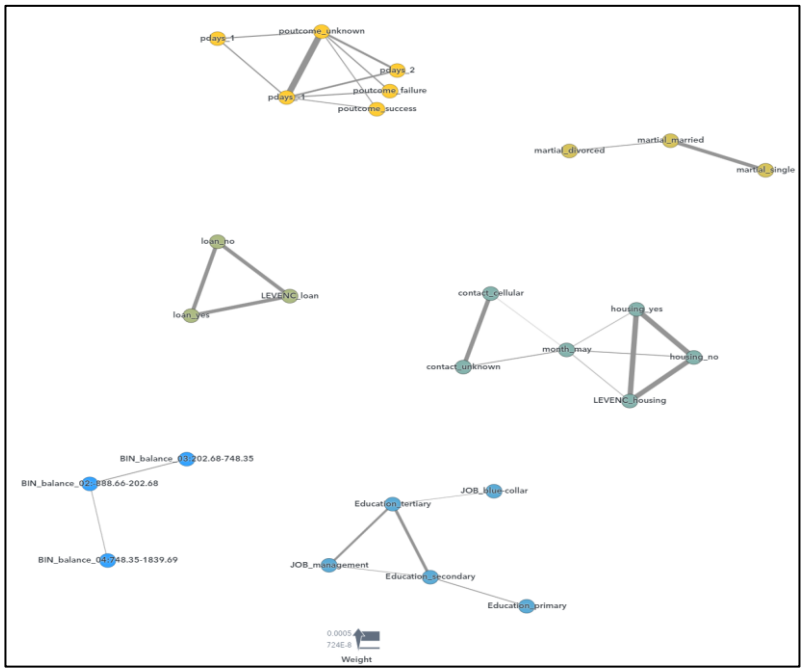

Figure 52: Modified variable cluster graph

5.4.4. Feature Machine Node

The feature machine node is an automated method to attempt at improving model accuracy. It is able to perform complex transformations based to address data quality issues such as high kurtosis, skewness, missing values, and outliers. Examining the details of the results table in Figure 53, it becomes apparent that a myriad of complex transformations have been executed on individual variables, delineating the depth of automated preprocessing

implemented. Noteworthy variables such as *'Transformed_duration_tree_binned'* and *'Transformed_default_level_encoded'* possess relatively high-ranking criteria values, implying their potential prowess in predictive modelling. The high-ranking value verifies that transformations done in section 5.2, prove to be comparatively beneficial as input variables. Moreover, the recommendation of several other transformations such as Box-Cox, median imputation and target / weight-of-evidence encoding may be used in predictive modelling, if accuracy is an issue during modelling.

Generated Features

Obs	Feature	Description	Level	Input Variable	Input Label	Ranking Criterion	Feature Rank
1	cpy_nom_mode_imp_lab_BIN_balance	BIN_balance: Low missing rate - mode imputation + label transformation	NOMINAL	BIN_balance	Transformed_balance_tree_binned	0.01316	1
2	cpy_nom_mode_imp_lab_var_1_	BIN_duration: Low missing rate - mode imputation + label transformation	NOMINAL	BIN_duration	Transformed_duration_tree_binned	0.10258	1

Figure 53: Generated Features – Feature Machine Node

5.4.5. Feature Extraction Node

Feature extraction is a technique that converts high-dimensional data into a smaller collection of features, that maintain the most significant information in the form of principal components. Principal component analysis (PCA) is a technique for generating groups of uncorrelated variables known as principal components, to understand holistic trends across the dataset (Walker & Rogers, 2020). The principal component coefficient represents the weight or the contribution of each original variable to the principal component (PennState Eberly College of Science, n.d.). For example, as shown in Figure 54, principal component 1 correlates most strongly with housing (0.52), and age has the most negative correlation of (-0.49). While this cluster does not exhibit signs of inter-correlation with other generated principal components, this may be indicative of the relatively low dimensionality of the dataset. This can be seen by the drastic differences between components 1 and 2 below. Typically, PCA is conducted to reduce dimensionality with an inordinate number of variables; however, as there is no holistic trend across components, this makes PCA less useful for this dataset. If accuracy issues occur during model prediction stages, this node may be re-examined to determine any overlooked biases.

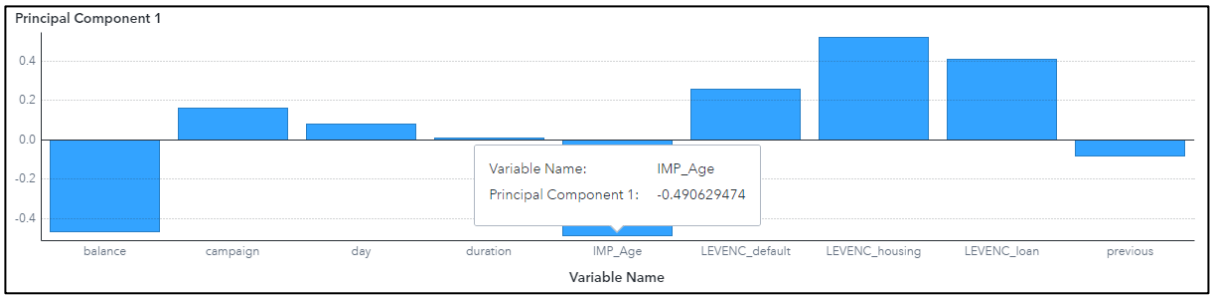

Figure 54: Principal Component 1 – Feature Extraction Node

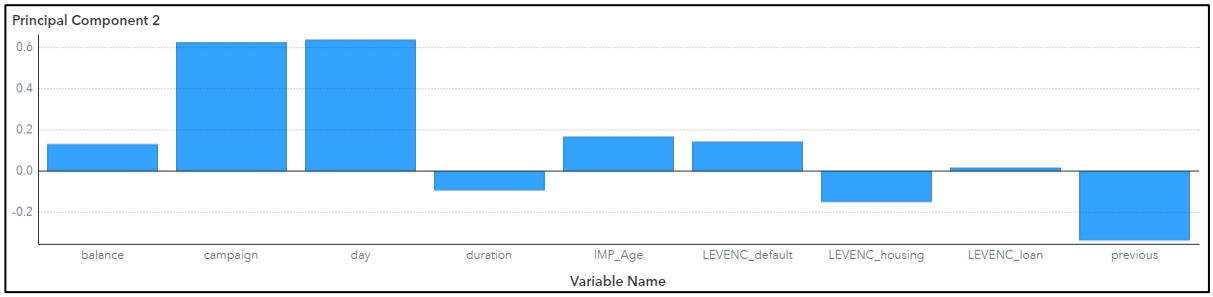

Figure 55: Principal Component 2 – Feature Extraction Node

5.4.6. Anomaly Detection Node

The anomaly detection node is used to detect if outliers are present in the dataset, and attempts to generalise them by dropping observations that may cause any over or under fitting issues during predictive modelling. This node may be used to inform modelling decisions in the next stage of the data analytics cycle.

5.4.7. Clustering Node

The clustering node applies definable automatic pre-processing steps such as imputation, encoding and standardisation, to better segregate data into identifiable clusters. This is more useful with a large dataset with holistic trends identified by principal component analysis; so that smaller training, test, and validation sets can be generated if a subset of the data needs to be examined separately. As this is not characteristic of the BANK_DIRECT_MARKETING dataset, it will not be explicitly used unless specified during predictive modelling.

6. Summary of findings

The BANK_DIRECT_MARKETING dataset was examined and analysed methodically to identify attribute types, distributions, and correlations, so that a baseline can be provided for further investigations. The business problem was clearly established, aiding in providing insight for certain relationships and observations that were derived from precise, reproducible results. These showcased various interesting trends between related attributes, such as the discovery of management roles intrinsically preferring tertiary education. This demographic contributed to a major portion of the dataset (approximately 22%), alluding to the narrative that these stable roles attract higher average yearly balances; and therefore may be more willing to subscribe to a term deposit.

Overall, after investigating the dataset, it was found that exploratory data analysis involving individual variable examination, helped find clusters and formed precise viewpoints which led to the formation of relationship narratives. Another example of this is the connection between Pdays and Poutcome, where unknown outcomes indicate that the target client was never contacted before. This finding helped explain the enormous disproportionate distribution of 74.4% of all records being marked as 'unknown' for the Poutcome attribute. It was uncovered that all unknown values were directly linked to clients who were never contacted before ('-1'), as described by the Pdays attribute. This creates a strong holistic narrative that the Portuguese banking institution was skewed towards targeting a majority of new clients, instead of previously contacted clients. This decision could have been motivated by the global financial crisis at the time of data collection; and executed as a bid to increase their term deposit signup rate by introducing new targets.

To utilise these valuable insights, predictive models such as neural networks, random forests and decision trees will be added to the end of the pre-processing pipeline shown in Figure 56 below. As specified by the business problem, their purpose will be to accurately determine if a potential client would signup for a term deposit, by predicting the dependent variable, *'y'*.

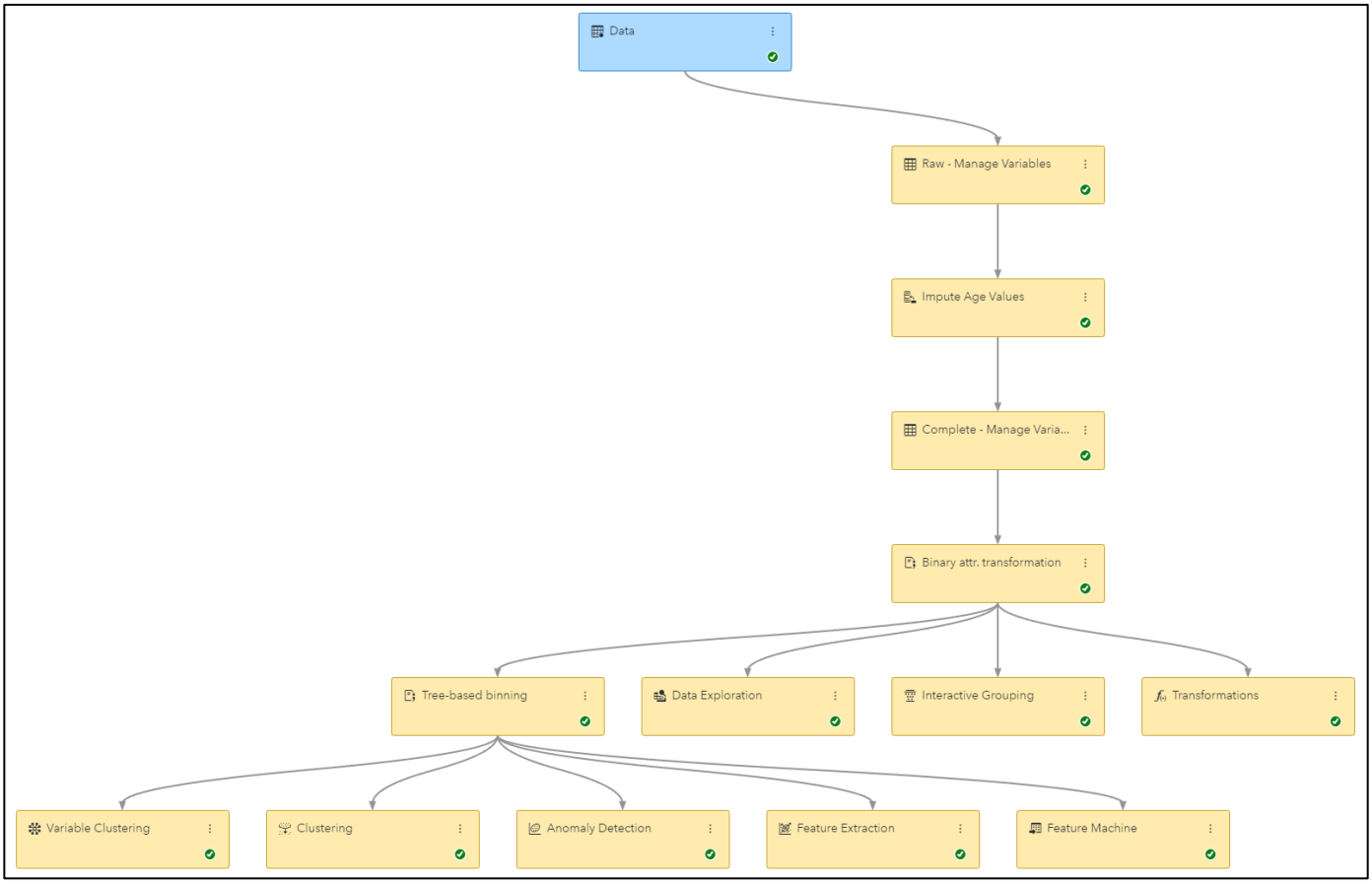

Figure 56: Data exploration and pre-processing pipeline

6.1. Challenges encountered during exploratory data analysis

Throughout the exploration stage of the analytics cycle, multiple challenges were faced, relating to the SAS environment itself, as well as the dataset. These are listed below with brief descriptions:

- Lack of clarity when using certain black box nodes such as anomaly detection
    - Specific nodes such as clustering and anomaly detection are not well documented, as they are majorly automated in nature. This made it difficult to gauge its results after running the pipeline.

- Learner account features are restricted, and projects could not be shared
    - This created many issues for the team, essentially putting the final pipeline design on a single account, instead of being able to collaborate freely. Pipelines were also unable to be shared via the exchange to other students for manual download.

- Attribute binarization
    - To alter the default binarization from 1 and 2 to the more standard 0 and 1, a SAS code node was needed as it was not straightforward to complete this via the model studio user interface.

- Scatter plot attribute type conversions
    - To display nominal / ordinal attributes on scatter plots, they had to be manually converted to their distinct count representations to plot against numerical values, which required significant research as this was not obvious at the time.

- Auto binning within SAS Visual Analytics
    - When viewing a continuous attribute with many levels, such as Age or Balance, SAS visual analytics auto bins to a maximum of 100. This made it hard to determine which values were either included or excluded within each bin, as these properties were not explicit.

7. References

In-text citations and referencing present within this report utilise APA 7th reference styling.

Campos, M. M., & Pereira, M. C. (2008). Impact of the recent reform of the Portuguese public employees' pension system. Research Papers in Economics. https://www.bportugal.pt/sites/default/files/anexos/papers/wp200812.pdf

Frost, J. (2022). Kurtosis: Definition, leptokurtic & platykurtic. Statistics By Jim. https://statisticsbyjim.com/basics/kurtosis/

Glen, S. (n.d.). Kurtosis: Definition, leptokurtic, platykurtic. Statistics How To. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/kurtosis-leptokurtic-platykurtic/

Kenton, W. (2023, April 12). Kurtosis definition, types, and importance. Investopedia. Retrieved September 7, 2023, from https://www.investopedia.com/terms/k/kurtosis.asp

Klima, K. (2021, March 21). Normality testing - Skewness and kurtosis. Connect, Learn, and Share | The GoodData Community. Retrieved September 7, 2023, from https://community.gooddata.com/metrics-and-maql-kb-articles-43/normality-testing-skewness-and-kurtosis-241

LinkedIn. (2023, August 31). What are the advantages and disadvantages of equal-width and equal-frequency Binning methods? Retrieved September 10, 2023, from https://www.linkedin.com/advice/1/what-advantages-disadvantages-equal-width#:~:text=Both%20are%20unsupervised%20binning%20methods,the%20data%20is%20uniformly%20distributed

Moro, S., Cortez, P., & Laureano, R. (2011, October). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology [Paper presentation]. European Simulation and Modelling Conference. https://www.researchgate.net/publication/236231158_Using_Data_Mining_for_Bank_Direct_Marketing_An_Application_of_the_CRISP-DM_Methodology

PennState Eberly College of Science. (n.d.). 11.4 - Interpretation of the principal components. Retrieved September 11, 2023, from https://online.stat.psu.edu/stat505/lesson/11/11.4#:~:text=Interpretation%20of%20the%20principal%20components%20is%20based%20on%20finding%20which,of%20course%20a%20subjective%20decision

SAS. (2015, March 13). Descriptive statistics — The more, the merrier in SAS visual analytics. SAS Blogs. https://blogs.sas.com/content/sgf/2015/03/13/descriptive-statistics-the-more-the-merrier-in-sas-visual-analytics/

SAS. (2018, June 25). Interactive binning node vs. transformation node binning. SAS Communities. Retrieved September 10, 2023, from https://communities.sas.com/t5/SAS-Data-Science/Interactive-Binning-Node-vs-Transformation-Node-Binning/td-p/393634

SAS. (2021, April 29). Transformations node in SAS model studio on SAS viya. SAS Communities. Retrieved September 10, 2023, from https://communities.sas.com/t5/SAS-Communities-Library/Transformations-Node-in-SAS-Model-Studio-on-SAS-Viya/ta-p/737959

SAS. (2021, March 4). SAS Viya: Optimal binning. SAS Communities. Retrieved September 11, 2023, from https://communities.sas.com/t5/New-SAS-User/SAS-Viya-Optimal-Binning/td-p/539999

Search.r-project.org. (n.d.). Intracluster correlation coefficients for clustered data. https://search.r-project.org/CRAN/refmans/fishmethods/html/clus.rho.html

Sturgis, P. (2004). Analysing complex survey data: Clustering, stratification and weights. Social Research Update. Retrieved September 12, 2023, from https://sru.soc.surrey.ac.uk/SRU43.html

TheGlobalEconomy.com. (2023, June). Portugal bank deposit interest rate, percent, June, 2023 - Data, chart. Retrieved September 8, 2023, from https://www.theglobaleconomy.com/Portugal/deposit_interest_rate/

UCI Machine Learning Repository. (2012, February 13). Bank Marketing. https://archive.ics.uci.edu/dataset/222/bank+marketing

Walker, C., & Rogers, W. (2020). Principal component analysis demystified. SAS Customer Support Site | SAS Support. https://support.sas.com/resources/papers/proceedings20/5110-2020.pdf

8. Appendix

8.1. Work distribution table

Team Member	Completed Sections	Extra Contributions
Manjyot Joher - 12897981 ***(Team lead)***	• 1 – Business Problem • 2 – Report Structure • 3.2.1 – Education • 3.2.2. – Marital • 3.2.3 – Job • 3.2.4 – Age • 3.2.5 – Customer_ID • 3.2.13 – Duration • 4.1.1 – Age & Housing • 4.2.1 – Balance, Education & Job • 4.2.2 – Balance, Default & Education • 4.2.3 – Age, Balance, Job & Marital • 5.2 – Data transformation (binning/conversion) • 5.4.1 – Correlation Matrix • 5.4.3 – Variable Clustering • 6 - Summary	• 3.2.12 – Month • 4.2.6 – Balance, Pdays & Poutcome • 5.4.5 – Feature Extraction • Document collation/editing
Vi Nguyen - 13592629	• 3.2.11 – Day • 3.2.12 – Month • 4.1.5 – Age & Education • 4.1.6 – Age & Balance • 4.2.8 – Age, Balance & Loan • 5.1 - Data cleaning • 5.4.2 – Relative Importance • 5.4.6 – Anomaly detection • 6.1 – Challenges encountered	• 1 – Business problem • 5.4.1 – Correlation Matrix • 5.2.2 – Binning • 6 - Summary
Aiden Ye Yint Hlyan - 14017432	• 3.2.14 – Campaign • 3.2.15 – Pdays • 3.2.16 – Previous • 3.2.17 – Poutcome • 4.1.4 – Campaign & Poutcome • 4.2.6 – Balance, Pdays & Poutcome • 4.2.7 – Default, Poutcome & Previous • 5.3 – Dimensionality reduction • 6.1 – Challenges encountered	• 1 – Business problem • 5.2 – Data transformation • 5.1 - Data cleaning • 6 - Summary
Su Myat Than Cin - 13486927	• 3.1 – Data dictionary • 3.2.6 – Default • 3.2.7 – Balance • 3.2.8 – Housing • 4.1.2 – Balance & Job • 4.2.4 – Balance, Housing & Loan • 5.4.5 – Feature Extraction • 6.1 – Challenges encountered	• 4.2.5 – Campaign, Contact, Duration & Poutcome • 1 – Business problem • 5.3 – Dimensionality reduction • 6 - Summary
Ye Min Oo - 13506858	• 3.2.9 – Loan • 3.2.10 – Contact • 4.1.3 – Balance & Duration • 4.2.5 – Campaign, Contact, Duration & Poutcome • 5.4.4 – Feature Machine • 5.4.7 – Clustering • 6.1 – Challenges encountered • Summary stat verification for all attributes	• 4.2.6 – Balance, Pdays & Poutcome • 1 – Business problem • 5.4.3 – Variable Clustering • 6 - Summary