Contents

Introduction 3
1. **Initial Data Exploration** 3
1.1 Attributes 3
1.2 Summary statistics 5
1.2.1. Missing attribute values 5
1.2.2 Analysis for categorical attributes 5
1.2.3 Analysis for numerical attributes 15
1.3 Data exploration 23
1.3.1 Outliers 23
1.3.2 Interesting attributes 24
1.3.3 Clusters 24
1.3.4 Correlation 25
2. **Data Preprocessing** 27
2.1 Binning 27
2.1.1 Equi-width 27
2.1.2 Equi-depth 30
2.2 Normalisation 32
2.3 Discretisation 34
2.4 Binarisation 35
3. **Summary** 35

Introduction

This report includes a data exploration and preparation for the Analytics Unit. This report includes an initial data exploration discovering each attribute and statistical summaries, as well as, identifying outliers and clustering. Preprocessing is performed using binning techniques, normalisation, discretisation and binarisation.

The primary of these tasks is to develop a deep and comprehensive understanding of the given dataset and explore underlying relationships and nuances while cleaning and standardising data for analysis.

1. Initial Data Exploration

1.1 Attributes

Attribute Name	Attribute Type	Justification
Age	Ratio	Age is considered to have equal intervals between values and because it has a true-zero point, it is ratio. This means that at age zero it represents birth. In addition, it has a meaning ratio, for example, the age 10 is twice as old as age 5.
Job	Nominal	This distinguishes parties into categories or separate classes without any orders or ranking.
Marital	Nominal	The status of being “married”, “single” or “divorced” has no inherent order or ranking. The distinct categories are not labelled as “better” or “goes first” than another category.
Education	Ordinal	This is categorised in a hierarchy where higher education is considered “greater” than lower education. For example, university degree > high school.
Default	Nominal	This refers to having a state either a credit in default or a credit not in default, therefore, it labels the categories in no order.
Housing	Nominal	This refers to having “yes”, “no” or “unknown” to housing, therefore, it labels the categories in no order.

Loan	Nominal	This refers to having “yes”, “no” or “unknown” to housing, therefore, it labels the categories in no order.
Contact	Nominal	This is categorised by different means of communication and is not inherently ranked or ordered.
Month	Ordinal	This is categorised in a meaningful order as month of the year are in a natural sequence. After January is February then March then April, etc. It is also considered categorical as months are not quantitative, although they are represented by numerical data, they are labelled as categories.
Day of week	Ordinal	This is categorised in a meaningful order as days of the week are in a sequence or order. Although, they are represented by a numeric value, they are not quantitative.
Duration	Ratio	Duration is quantitative data and represents the time which can be measured. This also has a true zero point meaning that duration 0 means that there is no time passed.
Campaign	Ratio	Number of contacts is quantitative data and true zero point is meaningful (i.e. 0 means there are no contacts)
Passed days	Ratio	Number of days since client was last contacted is quantitative data and true zero point is meaningful (i.e. 0 means that no time has passed since contacted).
Previous	Ratio	Number of contacts is quantitative data and true zero point is meaningful.
Poutcome	Nominal	This variable is categorical as it represents classes like “non-existent” and “failure” with no order.
Variation rate	Ratio	This indicated a rate and can be ordered and measured. It also has a true zero point. This means that if variation rate is 0 it means that the rate is constant.
Price index	Interval	This is expressed as any number and has no true zero point meaning that prices can be below the base level (having the value 0 doesn’t mean that no prices exist)

Confidence index	Interval	This is expressed as any number and has no true zero point meaning that prices can be below the base level (having the value 0 doesn't mean that no consumer exist)
Euribor3m	Ratio	This indicated a rate and can be ordered and measured. It also has a true zero point.
No. employed	Ratio	This indicated several employed and can be ordered and measured. It also has a true zero point (i.e. zero employees mean employees do not exist).
Term deposit	Nominal	This is categorised by numeric values 0 and 1 meaning a true and false response respectively in binary.
State	Nominal	This is categorised by categories of different states and is not in inherent order.

1.2 Summary statistics

1.2.1. Missing attribute values

Attribute	Number of missing values
Job	28
Marital	6
Education	118
Default	584
Housing	76
Loan	76
No. employed	192

This summary statistic shows the count of missing values, represented by "unknown" or "?", for each attribute in the Excel spreadsheet, noting that not all attributes have missing values. Several impacts on summary statistics include inaccurate representations of measures such as mean and median leading to skewed data distribution.

1.2.2 Analysis for categorical attributes

1.2.2a Job

Job type	Frequency
Admin	645
Blue-collar	585
Technician	457

Services	249
Management	205
Retired	105
Self-employed	100
Entrepreneur	92
Housemaid	76
Unemployed	54
Student	40
Unknown	28

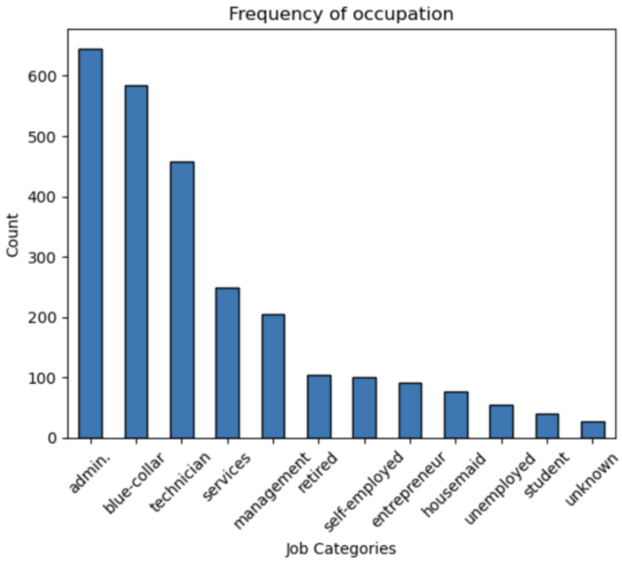


Figure 1. Bar graph for frequency of jobs

In this data, central tendency (mode) is discovered to be admin with 645 counts. The least common type was "unknown" which represents 28 entries of missing values. Although, it is a small proportion of the entire dataset, it is important to note that unclassified data might lead to biases in further analysis. However, it does not significantly skew the overall analysis as categories with larger counts and other relative sizes remain unchanged.

1.2.2b Marital

Marital	Count	Proportion (%)
Married	1611	61.1
Single	737	28.0
Divorced	282	10.7
Unknown	6	0.2
	2636	100

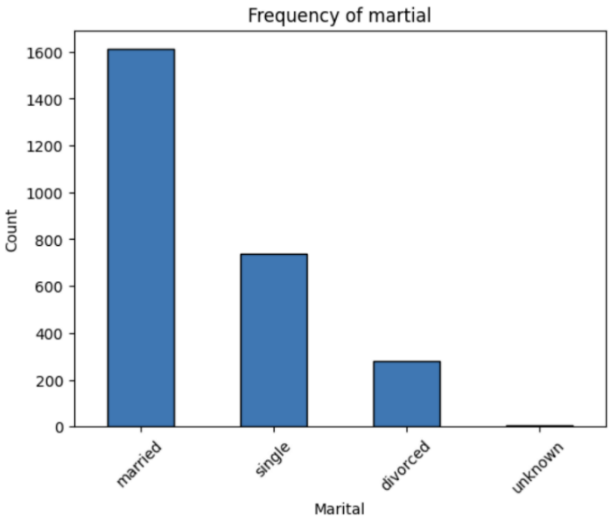


Figure 2. Bar graph of marital status count

The data above shows the marital status of 2636 entries. With the most common status, “married”, it dominates the others with a proportion of 61.1%. The distribution in figure 2, shows a heavy skew toward married individuals. This provides insight of the demographic makeup as the marketing campaign’s target population is dominantly married people. There could be financial priorities for married individuals such as investments and household savings that the company may consider to tailor offering.

1.2.2c Education

Education level	Count
University degree	799
High school	591
Basic 9Y	378
Professional course	326
Basic 4Y	261
Basic 6Y	162
Unknown	118
Illiterate	1

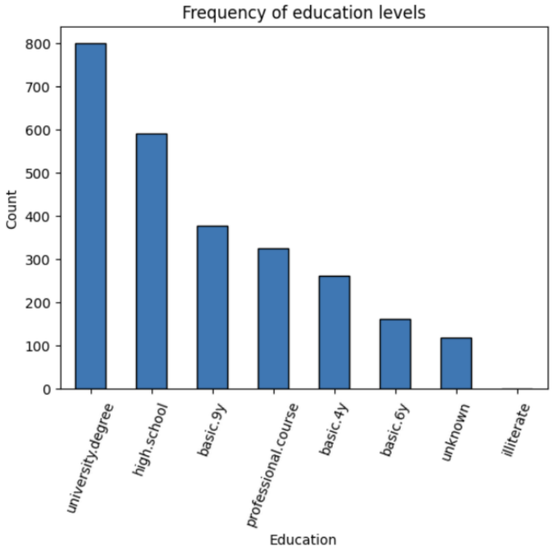


Figure 3. Graph of marital status count

In figure 3, the mode is university degree with 799 individuals and falling under is high school with 591 counts. University degree holders and individuals with certificates from a professional course are likely to have a more stable income and jobs. They are most likely to have interest in long-term financial planning and investments such a term deposits.

1.2.2d Default

Default	Count
No	2052
Unknown	584

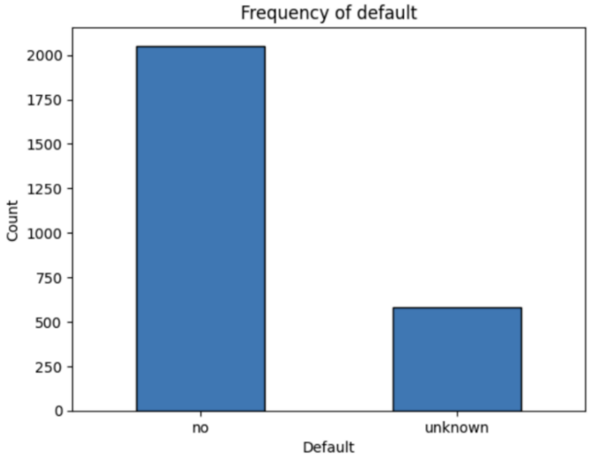


Figure 4. Bar graph of default

The column of “default” in the dataset represents the credit default of individuals. The frequency distribution is shown in figure 4, with the majority of individuals have said “no” to default and there are 582 instances of “unknown.” With a percentage of 22% of “unknown” accounts for many records missing or not recorded. Therefore, further investigation and data cleaning is required to minimise the impact of analysis, so that biased analysis is prevented that could distort data exploration.

1.2.2e Housing

Housing loan	Count
Yes	1348
No	1212
Unknown	76

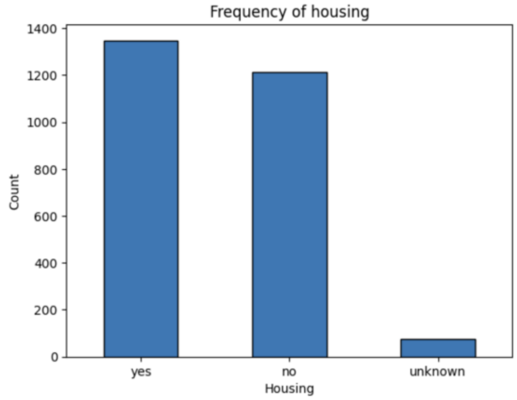


Figure 5a. Bar graph of housing loan count

Out of 2393 entries, a proportion of 1348 individuals have a housing loan versus 1212 that do not. The bar graph above (Figure 5a) shows that there is a relative split between having and not having a housing loan despite 76 unknown entries. For bank marketing campaign insights, the campaign may focus on financial advice or home equity loans as well as mortgaging and financial planning to tailor the need of both categories.

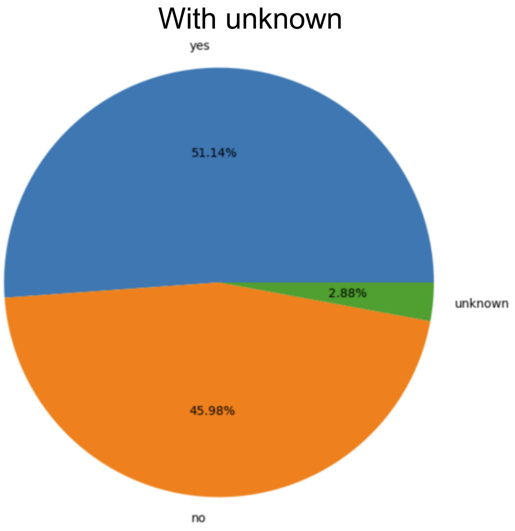


Figure 5b. Pie chart of housing with unknown

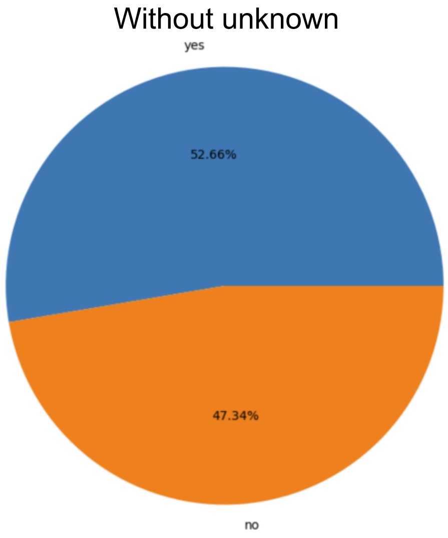


Figure 5c. Pie chart of housing without housing

The above pie charts show a comparison of results before and after data is cleaned. This enhanced data visualisation when comparing results of ‘yes’ and ‘no.’ Missing values in the data may result in misleading conclusions and enhances interpretation. Figure 5b has the original data set with the unknown values. After data cleaning, by filtering out the ‘unknown’ value, figure 5c shows adjustment of the proportions of results with answers of yes and no of 52.66% and 47.34% respectively.

1.2.2f Loan

Personal loan	Count
No	2174
Yes	386
Unknown	76

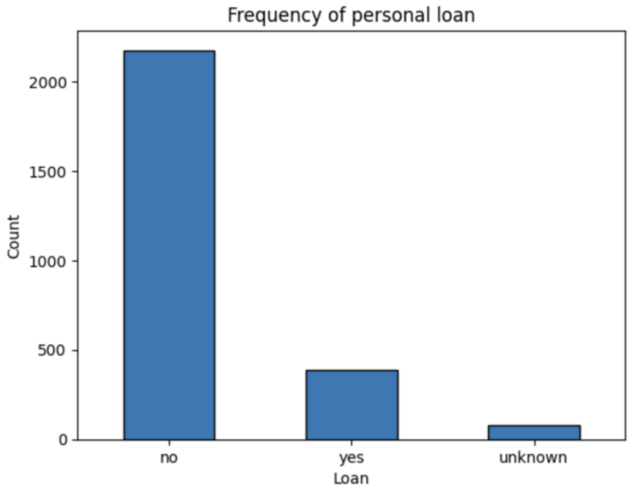


Figure 6a. Bar graph of personal loans

In figure 6a, a very large majority of individual does not have a personal loan with a proportion of 82%. This shows that most people did not opt for a loan or think a personal loan is needed. This observation provides insight for the bank to focus on offering personal loan products.

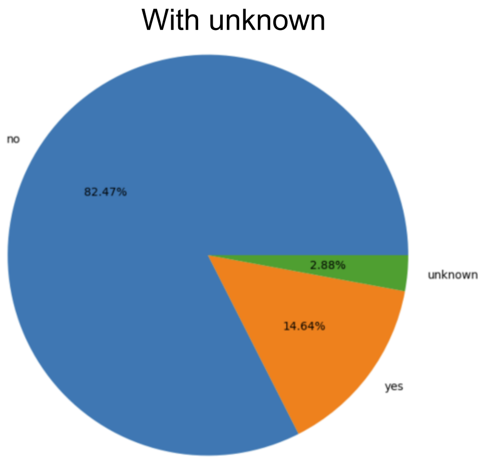


Figure 6b. Pie chart of housing with unknown

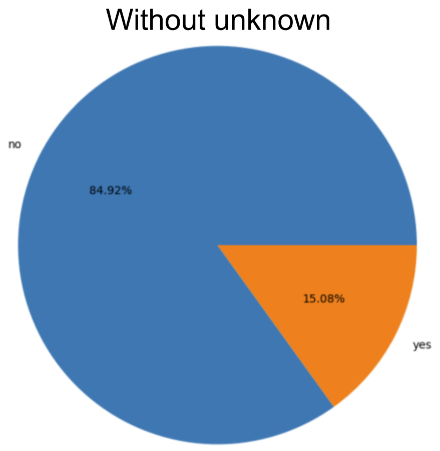


Figure 6c. Pie chart of housing without housing

Seen from the pie charts generated above, after data cleaning of filtering out the missing value labelled as ‘unknown,’ there is a small change of proportions. The effect of the ‘unknown’ value does not have a strong impact on the categorical sizes as results change from 82.47% to 84.92% for no and from 14.64% to 15.08% for yes. In figure 6c, the dominant category is still outstanding.

1.2.2g Contact

Mode of contact	Count
Fax	551
Email	551
Telephone	520
Mailing	509
Cell phone	505

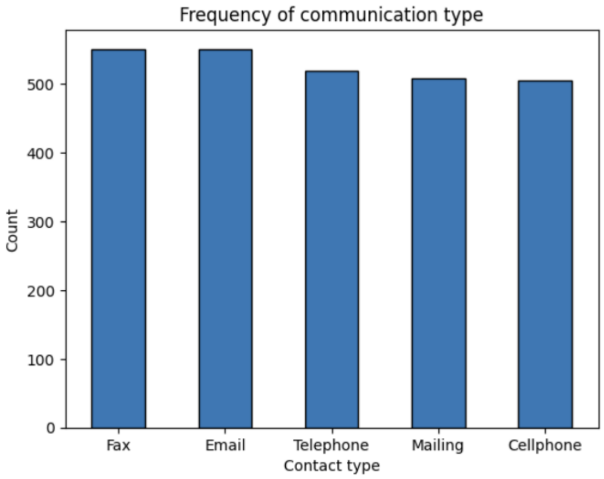


Figure 7. Bar graph of communication types

Overall, the mode of communication is equally distributed as no method does not dominate over the others. The central tendency (mode) of communication is both through fax and email which highlights digital/electronic communication. The dataset reveals an approach of multiple communication methods that demonstrated balanced distribution (Figure 7).

1.2.2h Month

Month	Count
1	867
3	492
4	366
2	348
6	272
9	162
5	47
8	34

10	30
7	18

Note: Each month is by number. (1- Jan, 2- Feb, 3- Mar, 4- Apr, 5- May, 6- June ,7, July, 8- Aug, 9- Sept ,10- Oct).

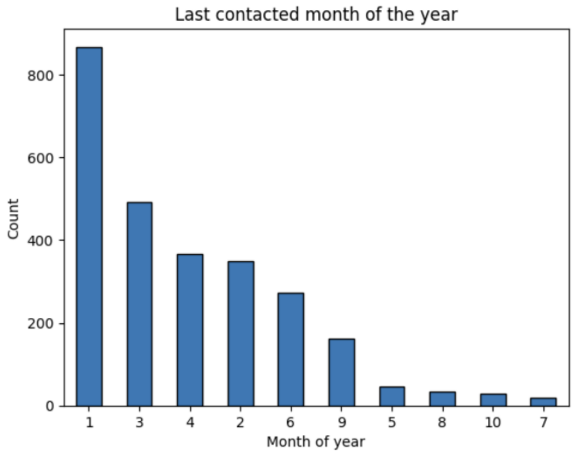


Figure 8. Bar graph of months when customer was last contacted

According to figure 8, most clients were contacted in January, followed by March and April. Off-peak months include July, August and October where customer engagement was low. This means that for the next campaign, a recommendation would be to push the campaign to capitalise customer interest leading to effective marketing strategies.

1.2.2i Day of week

Day	Count
4	575
3	556
1	543
5	489
2	473

Note: Each month is by number. (1- Mon, 2- Tues, 3- Wed, 4- Thurs, 5- Fri).

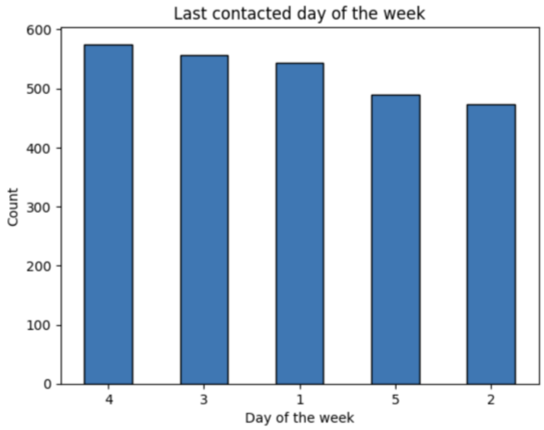


Figure 9. Bar graph of days of the week

In figure 9, it is seen that the mode is Thursday with a count of 575 when customers were last contacted. Closely behind is Tuesday with a frequency of 556. This implies high engagement where customers are most likely to engage with the marketing campaign.

1.2.2j Poutcome

Outcome	Count
Non-existent	2289
Failure	261
Success	86

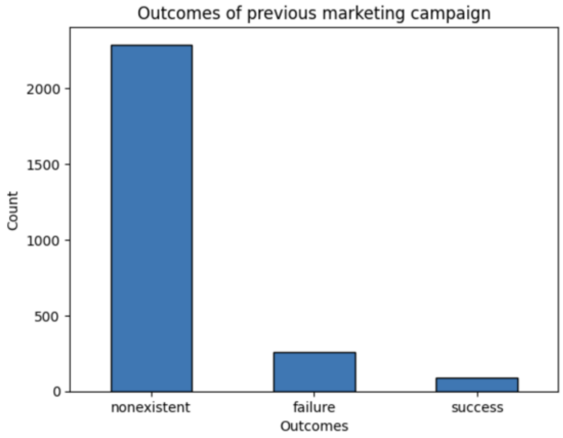


Figure 10a. Bar graph for outcomes of previous marketing campaign

Shown in figure 10a, there is a very small proportion of individuals in the dataset that had a successful campaign. Given the large volume of non-existent and failure results, it is confident to say that the overall outcome of the previous campaign was poor. This relates to the results of the engagement strategy as number of contacts performed before the campaign was dominantly zero (Figure 17). 2289 entries of non-existent outcomes implies that no outcome was record due to no response from the client and failure is due to low engagement and ineffective marketing strategies.

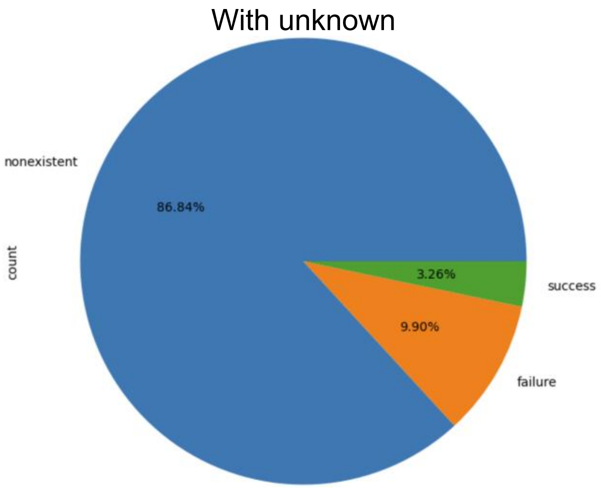


Figure 10b. Pie chart of housing with unknown

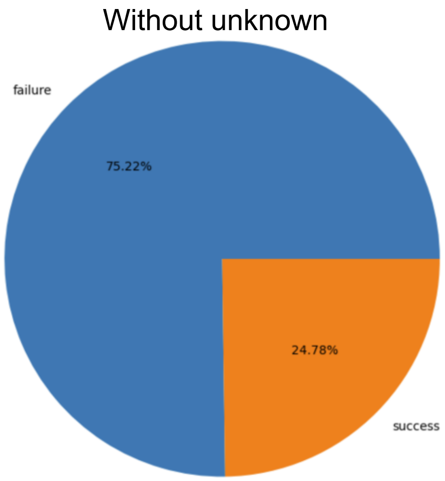


Figure 10c. Pie chart of housing without housing

It is crucial to consider unknown or missing data points to analyse data accurately and gain better understanding of a dataset. The outcome of the previous marketing campaign has a result majority that is non-existent, indicating that results may not be recorded or not captured. After this data it was cleaned, in figure 10c, it is seen that there is optimised visual enhancement. With a failure proportion of 75.22% and success percent of 24.78%. The overall impact of missing values is that it calculated inaccurate proportions which misrepresents the size of failure/success categories.

1.2.2k Term deposit

Subscribed to a term deposit	Count
0	2338
1	298

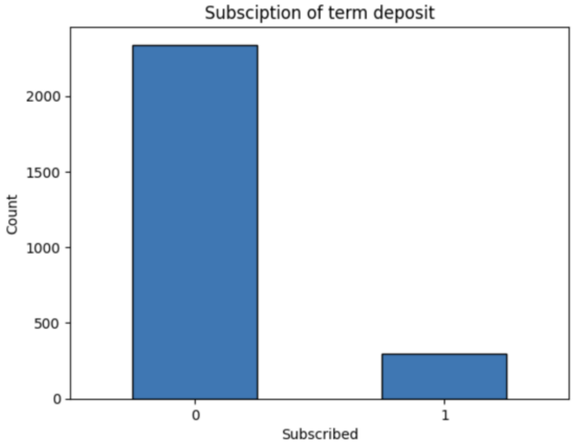


Figure 11. Bar graph of subscription to a term deposit

The values 0 and 1 is in a binary variable where 0 means not subscribed and 1 means to be subscribed to a term deposit. 2338 individuals account for 88.9% of the dataset of not subscribing to a term deposit. This indicated that outcome of the marketing campaign as not successful. The high proportion of non-subscribers suggests for more compelling and effective strategies for the campaign. With 298 entries of subscribers show that there is a need for improvements to convert customers.

1.2.2l State

State	Count
SA	356
VIC	341
NT	334
ACT	330
TAS	326
NSW	326
QLD	322

WA	301

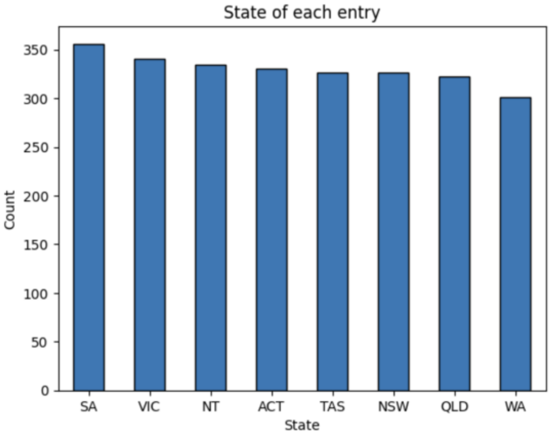


Figure 12. Bar graph of frequency of states in Australia

The statistical summary of states refers to the geographical location of each individual. The mode is SA which a count of 351 and the lowest count is WA with 301 entries. This indicates a even spread of different states and is supported by the bar graph shown in figure 13.

1.2.3 Analysis for numerical attributes

1.2.3a Age

Statistics	Value
Mean	40.0
Median (Q2)	38.0
Mode	31.0
Variance	104.1
Standard deviation	10.2
Q1 (25th percentile)	32.0
Q3 (75th percentile)	47.0
Interquartile range (IQR)	15.0
Maximum (higher whisker)	69.5
Minimum (lower whisker)	9.0

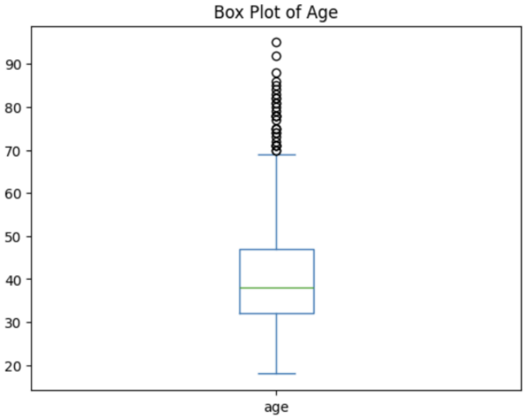


Figure 13. Box plot of age

The summary statistics from each individual's age show that the mean is 40 with a significant spread in data with a standard deviation of 10.2 and variance of 104.1. The age distribution indicates a broad range of target message for different age groups and the high standard deviation indicated diversity in age. There are outliers outside the maximum with the age of 69.5 years and no outliers are found below the minimum of 9 years.

1.2.3b Duration

Statistics	Value
Mean	263.9
Median (Q2)	179.0
Mode	96
Variation	73803.7
Standard deviation	271.7
Q1 (25th percentile)	101.0
Q3 (75th percentile)	323.3
Interquartile range (IQR)	15.0

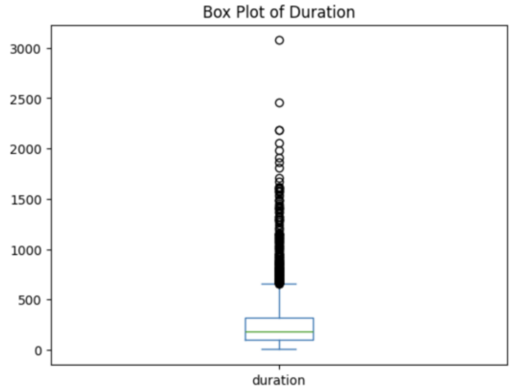


Figure 14a. Box plot of duration

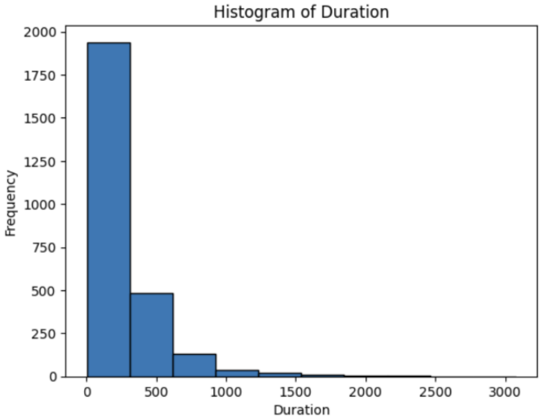


Figure 14b. Histogram of duration

The duration of last contacted is represented by two visuals, a box plot and histogram. Shown in figure 14a, many outliers are present and show a better visual for central tendency (mean, median, mode). The impact of these outliers results in a higher variance indicating a spread from the mean. The mean is greater than the median and the mode is less than the median, therefore, resulting in a positive or right-skewed distribution. Figure 14b supports the skewness as it shows a better visualisation that the data is skewed right.

1.2.3c Passed days

Statistics	Value
Mean	962.8
Median (Q2)	999.0
Mode	999

Variance	34628.1
Q1 (25th percentile)	999.0
Q3 (75th percentile)	999.0
Interquartile range (IQR)	0.0
Maximum	999.0
Minimum	999.0

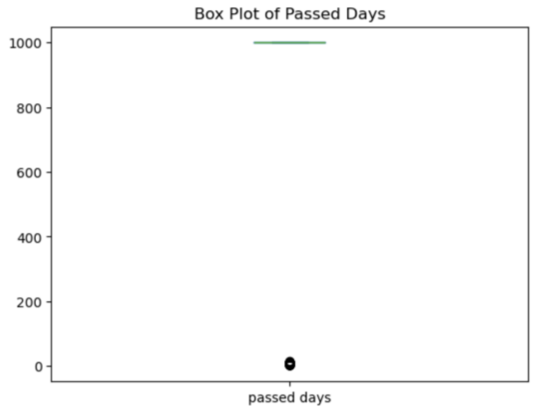


Figure 15a. Box plot of passed days

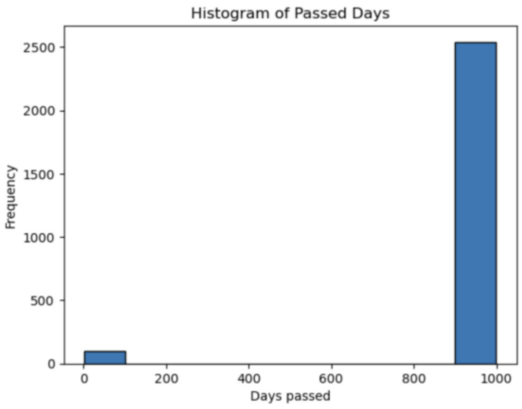


Figure 15b. Histogram of passed days

From the summary statistics, the distribution is significantly skewed as a large majority of values is 999. In the dataset, it shows low variability as the mode, median, Q1. Q3, maximum and minimum values is 999. This is supported by the boxplot with isolated points as outliers below the whisker (minimum-bound). Therefore, from the previous campaign, many of the clients was last contacted 999 days.

1.2.3d Campaign

Statistics	Value
Mean	2.5
Median (Q2)	2.0
Mode	1
Variance	6.7
Standard deviation	2.6
Q1 (25th percentile)	1.0
Q3 (75th percentile)	2.0
Interquartile range (IQR)	3.0
Maximum	6.0
Minimum	-2.0

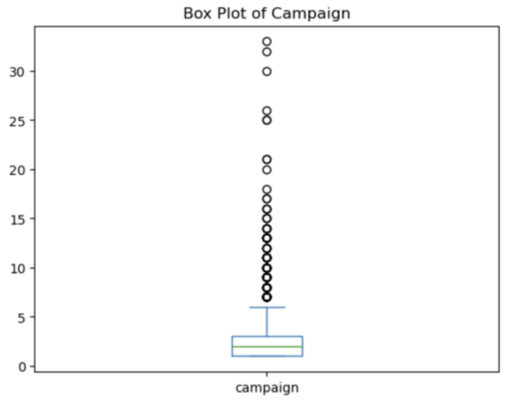


Figure 16a. Box plot of passed days

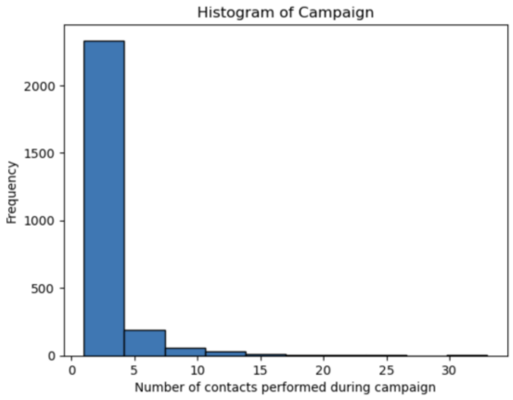


Figure 16b. Histogram of passed days

Shown in figure 15a, there are potential outliers as these points are above the maximum whisker of 6.0. This can also be seen in the histogram in figure 15b where the value count are mostly around 0-5 contacts. The dataset has a central tendency lower than the mean. The data has also variability as the values deviate from the mean by 2.6.

1.2.3e Previous

Number of contacts	Count
0	2289
1	295
2	38
3	10
4	3
7	1

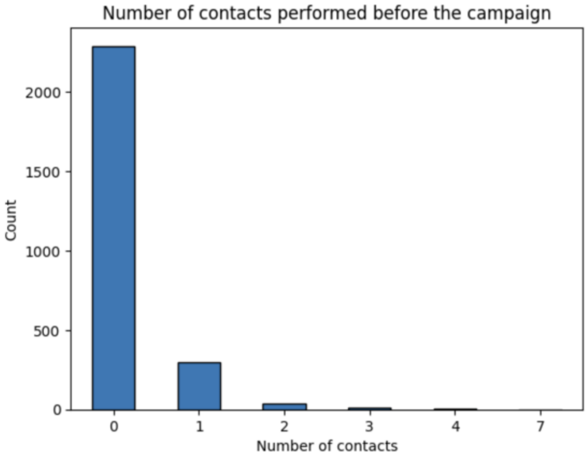


Figure 17. Bar graph for the number of contacts before the campaign.

A bar chart was generated as data is discrete and the values are whole numbers. This choice of graph shows each contact count into each bar. Most individuals had zero contacts with 2289 entries and making up 87% of the dataset. This suggests that there was no follow up contact or engagements after the initial interaction. Regarding the marketing campaign, it is important to focus on follow-up techniques to enhance customer engagement to result in a successful campaign.

1.2.3f Variation rate

Statistics	Value
Mean	0.1
Median (Q2)	1.1
Mode	1.4
Variance	2.5
Standard deviation	1.6
Q1 (25th percentile)	-1.8
Q3 (75th percentile)	1.4
Interquartile range (IQR)	3.2
Maximum	6.2
Minimum	-6.6

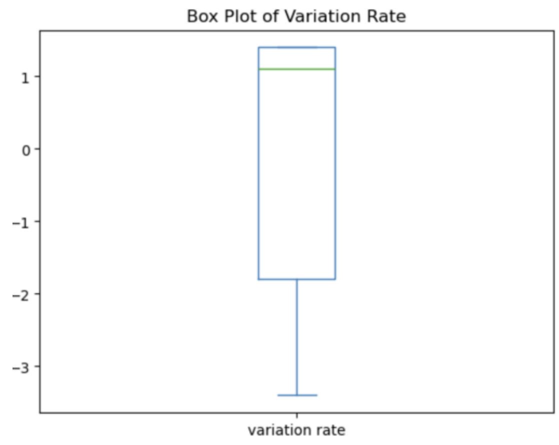


Figure 18a. Box plot of variation rate

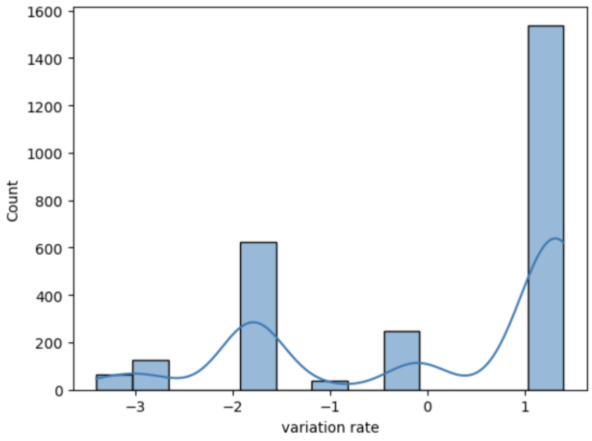


Figure 18b. Histogram of variation rate

The dataset shows the moderate distribution spread with a mean of 0.1 and 1.1 as the median, indicating slight left skewness. The spread is determined by the variance and standard deviation of 1.4. The mode identified is 1.4 and is shown as a peak in figure 18b. No outliers were identified as there was no isolated data points outside the minimum and maximum whiskers, with values -6.6 and 6.2 respectively, in the boxplot in figure 18a.

1.2.3g Price index

Statistics	Value
Mean	93.58
Median (Q2)	93.92
Mode	93.99

Variance	0.33
Standard deviation	0.58
Q1 (25th percentile)	93.08
Q3 (75th percentile)	94.00
Interquartile range (IQR)	0.92
Maximum	95.37
Minimum	91.70

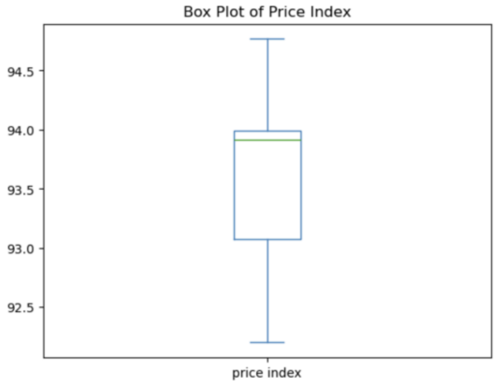


Figure 19a. Box plot of price index

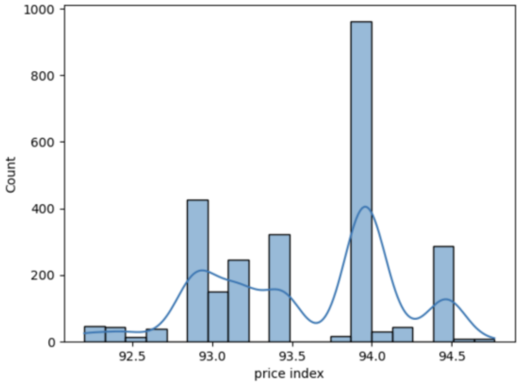


Figure 19b. Histogram and density plot of price index

The summary statistics show that the dataset is quite close to each other indicated by the low variance of 0.33 and standard deviation of 0.58. In figure 19a, it shows a relatively symmetric distribution as the mean, median and mode are proximity. From the histogram with density plot in figure 19b, the curves estimate the data points' distribution showing the tight clustering around the mean and median.

1.2.3h Confidence index

Statistics	Value
Mean	-40.52
Median (Q2)	-41.80
Mode	-36.40
Variance	20.96
Standard deviation	4.58
Q1 (25th percentile)	-42.70
Q3 (75th percentile)	-36.40
Interquartile range (IQR)	6.30
Maximum	-26.95
Minimum	-52.15

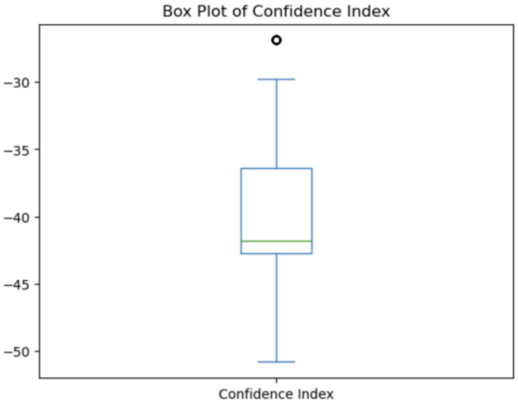


Figure 20a. Box plot of confidence index

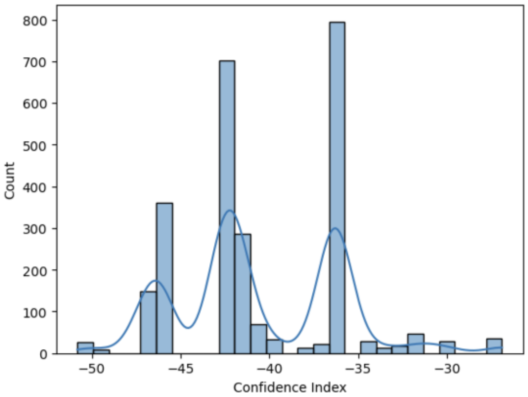


Figure 20b. Histogram and density plot of confidence index

The summary statistics show closeness between the mean and median indicating a nearly symmetric distribution. There is some spread in the dataset as the variance is 20.96 and points deviate from the mean of 4.58 on average. In figure 20a, an outlier is found as a data point is seen above the maximum whisker. The graph in figure 20b, shows a peak at approximately -47, -42 and -37 showing that in these values have the highest counts.

1.2.3i Euribor3m

Statistics	Value
Mean	2.5
Median (Q2)	2.0
Mode	1
Variance	6.7
Standard deviation	2.6
Q1 (25th percentile)	1.0
Q3 (75th percentile)	2.0
Interquartile range (IQR)	3.0
Maximum	6.0
Minimum	-2.0

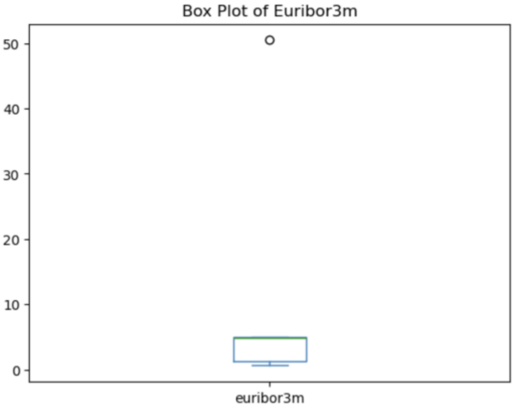


Figure 21a. Box plot of euribor3m

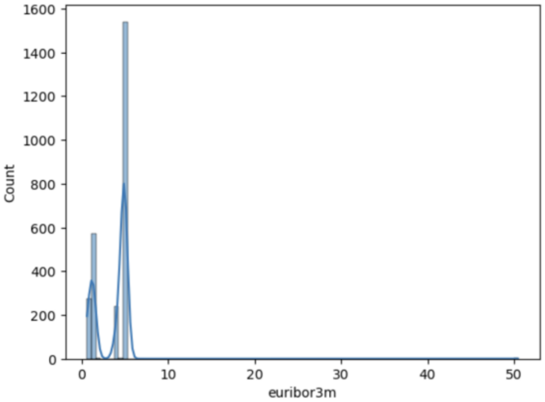


Figure 21b. Histogram and density plot of euribor3m

The statistics for euribor3m show a mean of 2.5 and median of 2.0 with a moderate spread around the mean with a variance of 6.7 and standard deviation of 2.6. Because the mean is slightly greater than the median, the distribution is skewed right. Figure 21a highlights an outlier at around 50 as the data point is isolated from the rest of the dataset, evident in the distance from the maximum whisker of 6.0. In figure 21b, the dataset is most frequent between 0 and 10 with some notable peaks at around 3 and 7.

1.2.3j No. employed

Statistics	Value
Mean	5089.13
Median (Q2)	5076.20
Mode	5191.00
Variance	8137.75
Standard deviation	90.21
Q1 (25th percentile)	5008.70
Q3 (75th percentile)	5191.00
Interquartile range (IQR)	182.30
Maximum	5464.45
Minimum	4735.25

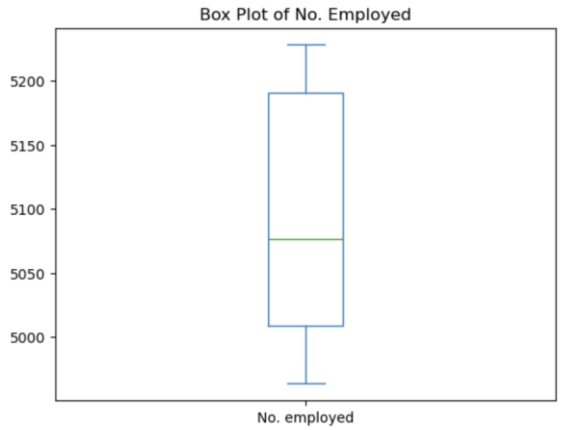


Figure 22a. Box plot of no. employed

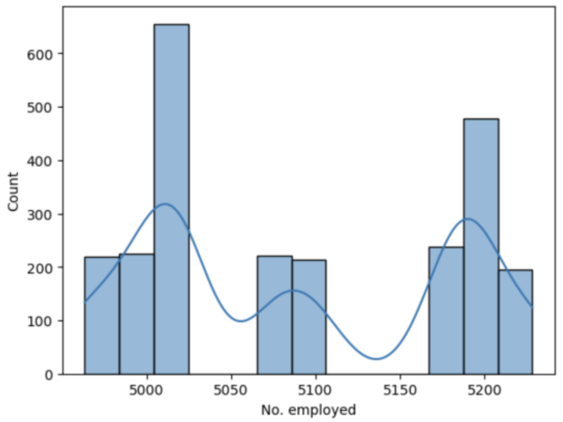


Figure 22b. Histogram and density plot of no. employed

The mean and median have a very close proximity, but the mean is slightly higher than the median suggesting a slight right-skew. The distribution is somewhat spread with a variance and standard wdeviation of 8137.75 and 90.21 respectively. The IQR (182.30) shows there is spread of the middle 50% and ranges from Q1 and Q3, suggesting variability, evident in figure 22a. Figure 22b shows distribution of the dataset with 3 peaks at around 5000, 5100 and 5200.

1.3 Data exploration

1.3.1 Outliers

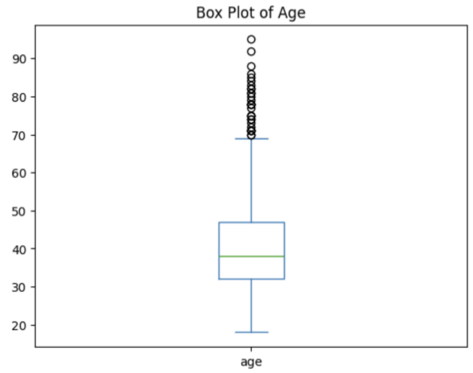


Figure 13. Box plot of age

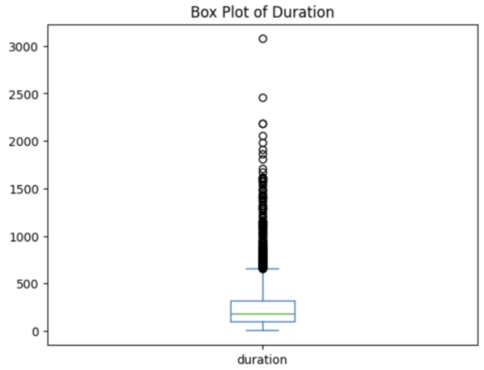


Figure 14a. Box plot of duration

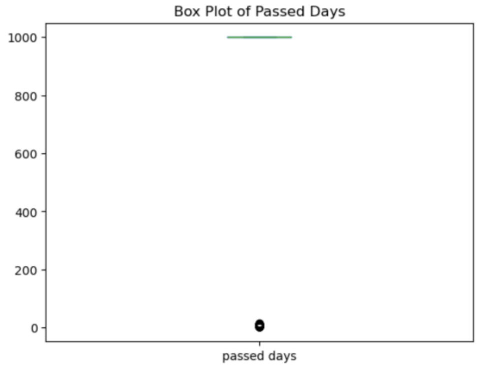


Figure 15a. Box plot of passed days

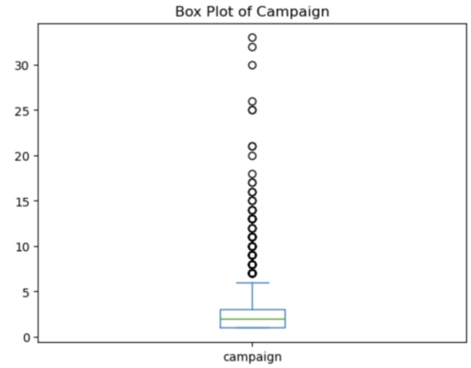


Figure 16a. Box plot of passed days

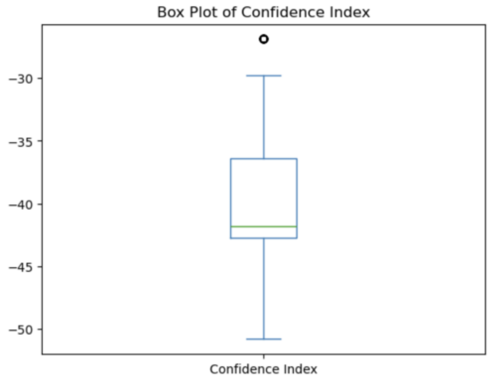


Figure 20a. Box plot of confidence index

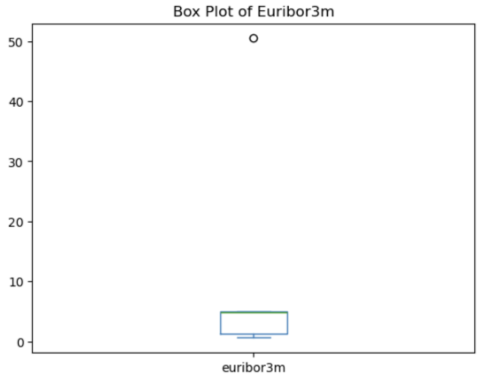


Figure 21a. Box plot of euribor3m

The graphs above shown is a compilation of box plots with outliers to identify. The outliers are identified by data points that are seen and are isolated from the majority of data points. Maximum and minimum whiskers is the threshold in the dataset that are within 1.5 times the IQR from the first and third quartile. Any data points that lie outside the range and considered as potential outliers. For example, in figure 14a, there is a significant amount of data points that lie outside the maximum whisker indicating outliers.

1.3.2 Interesting attributes

- In 1.2.2d default, there were two results for this attribute: ‘no’ and ‘unknown.’ There were 2052 instances for ‘no’ and 584 unknown instances were unknown. This conflicts with data quality as information are missing or not recorded. If that this result was ignored, it leads to biased results as analysis will only include ‘no.’ Therefore, leads to misleading conclusions.
- Summary statistics in 1.2.3c passed days, display meaningful results as there was a majority of 999 days instances, and very few outliers that deviate from the entire dataset, shown in the boxplot (figure 15a). This indicates that the users had no contact with the campaign for 999 days and a very small proportion had contact at most recent days.
- The euribor3m attribute has an abnormal data point as the outlier or maximum value is at 50.45. Euribor refers to a European interbank interest rate, so with a rate of 50.45 is extremely high. This can impact the central tendency of the mean as it skews the dataset, and results can represent misleading conclusions. It is important to confirm and analyse this abnormality to validate incorrect data entry.

1.3.3 Clusters

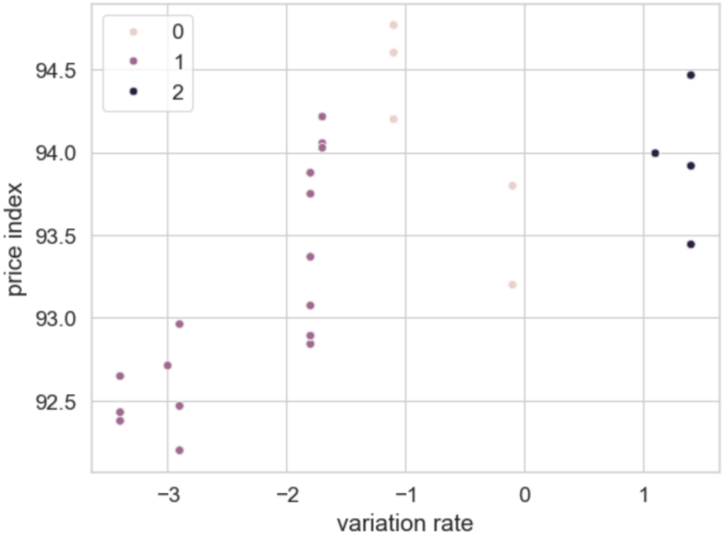


Figure 23. KMeans Clustering plot of variation rate and price index

A scatter plot with KMeans clusters is created through python. KMeans is applied to the dataset of columns (variation rate and price index) and the visual of the graph is customised by a grid to enhance visualisation. There are 3 clusters identified by three colours: peach, dark purple and purple. The peach colour is grouped by a negative variation rate ranging from approximately -3.5 to -1.8. Another group is clustered in dark purple that ranges from around -1 and 0. A variation rate greater of around 1 is grouped indicate by the colour purple. This clustering shows the relationship of the two attributes as it is seen to have a positive correlation (See 1.3.2).

1.3.4 Correlation

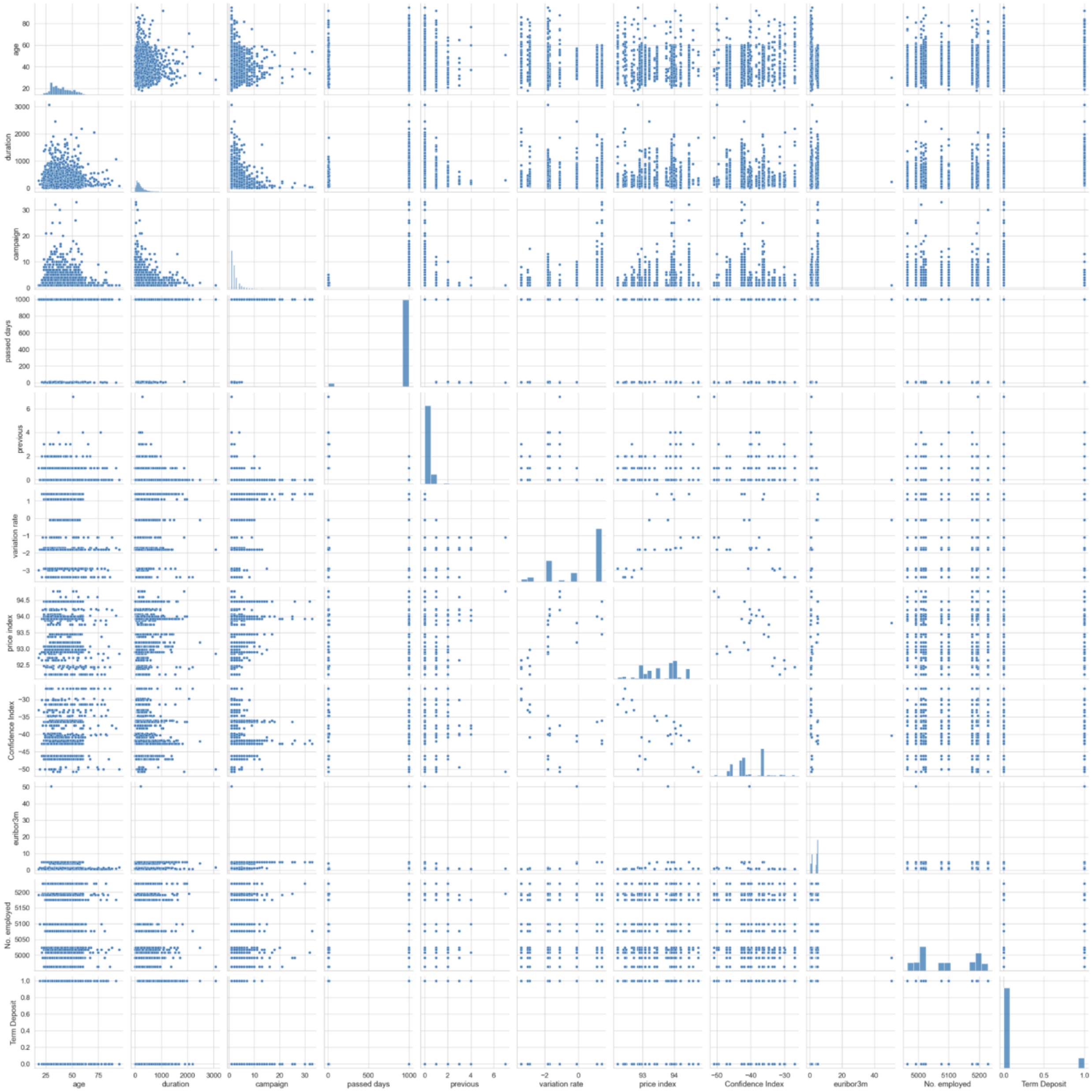


Figure 24. Pair plot of numerical attributes

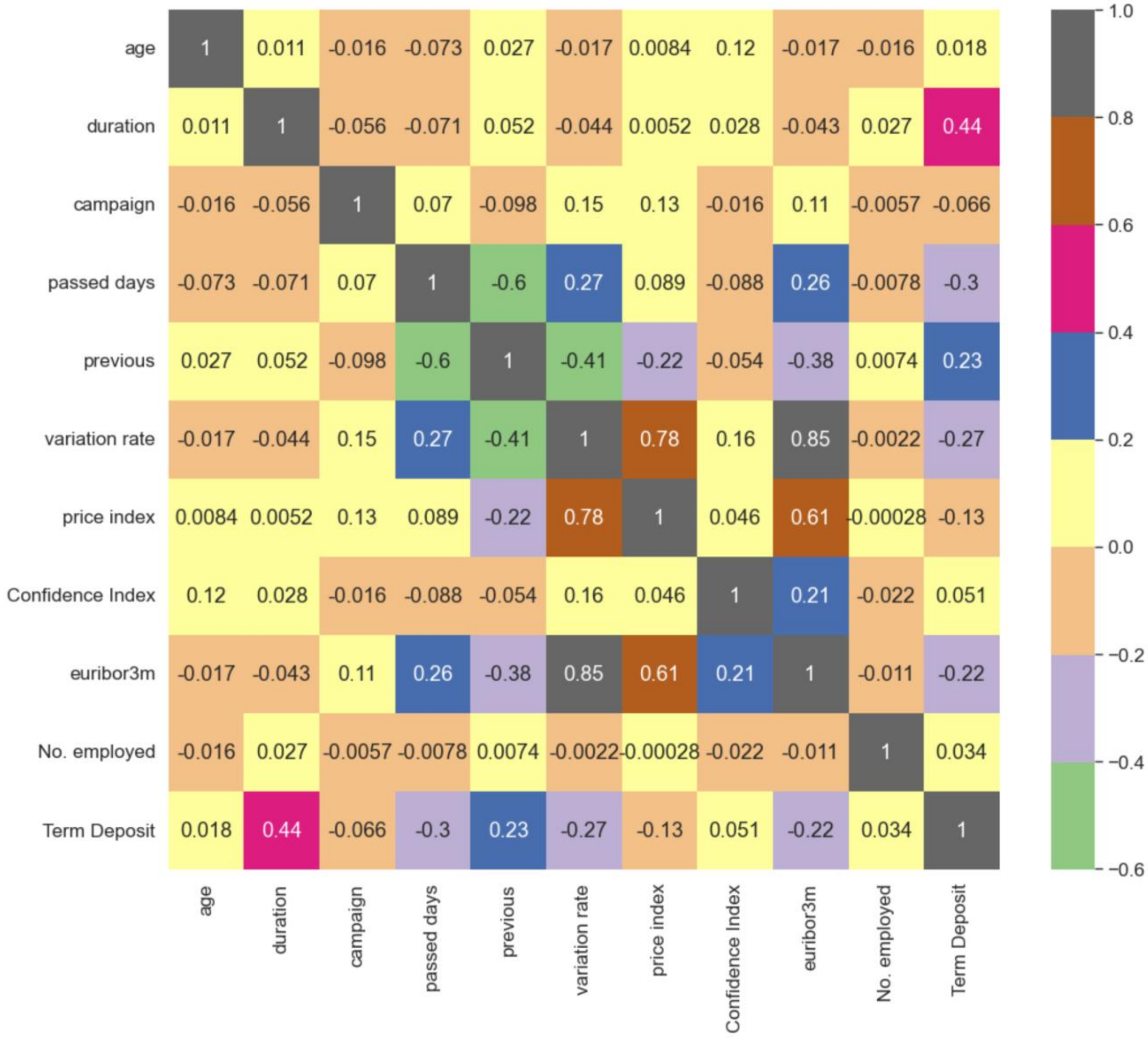


Figure 25. Heatmap of numerical attributes

The two graphs above demonstrate the correlation between numerical attributes. In figure 24, it is seen that a scatter plot is created with x and y axis as attributes. This helps show a pairwise relationship and enhance visualisation in determining correlations between the two variables. Figure 25 is a heatmap showing the correlation coefficient that measures the linear relationship between two variables. Values below 0 indicate a negative correlation, whereas values at 0 have no correlation (linear) and values above 0 have a positive correlation.

In 1.3.1, the correlation of price index and variation rate is plotted with a scatter plot with 3 clusters. The graph indicates a positive correlation, seen visually. Figure 25 supports this as the correlation value is at 0.78, representing a positive high correlation. A regression plot is created to outline that as variation rate increases, the price index proportionally increases.

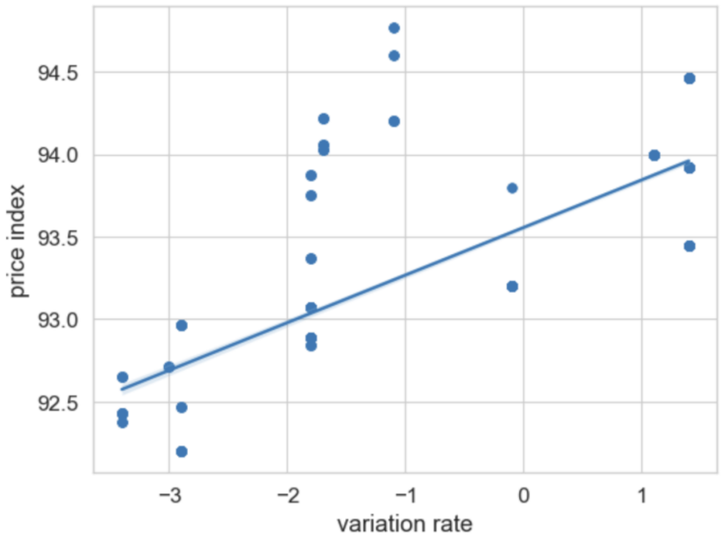


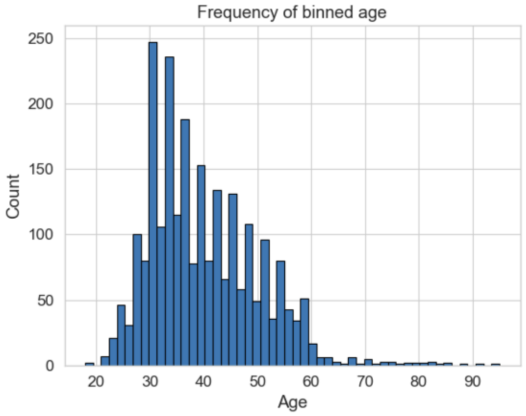
Figure 26. Regression plot for variation rate and price index

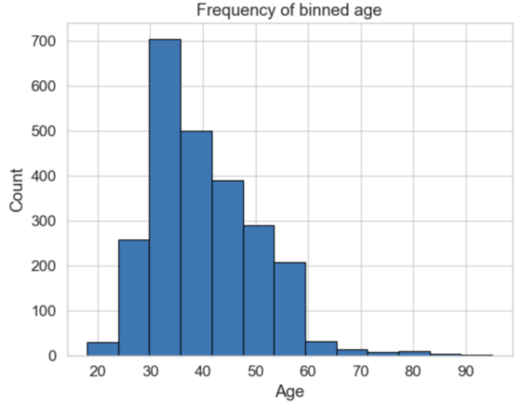
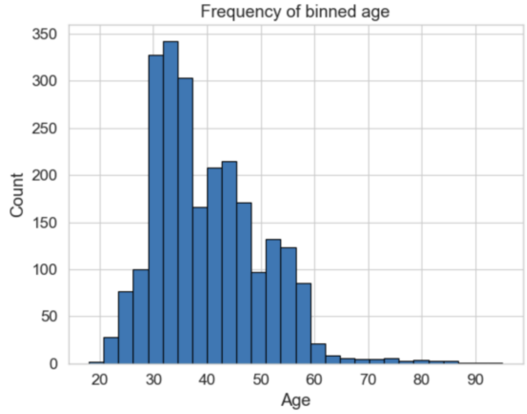
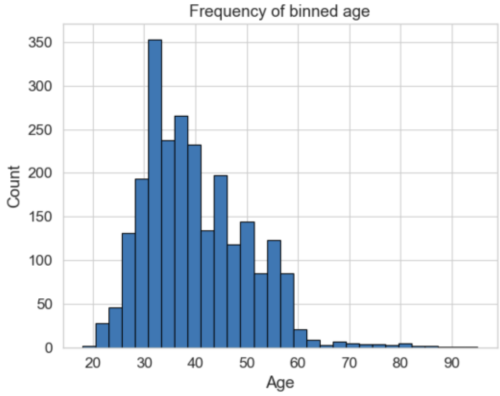
2. Data Preprocessing

2.1 Binning

2.1.1 Equi-width

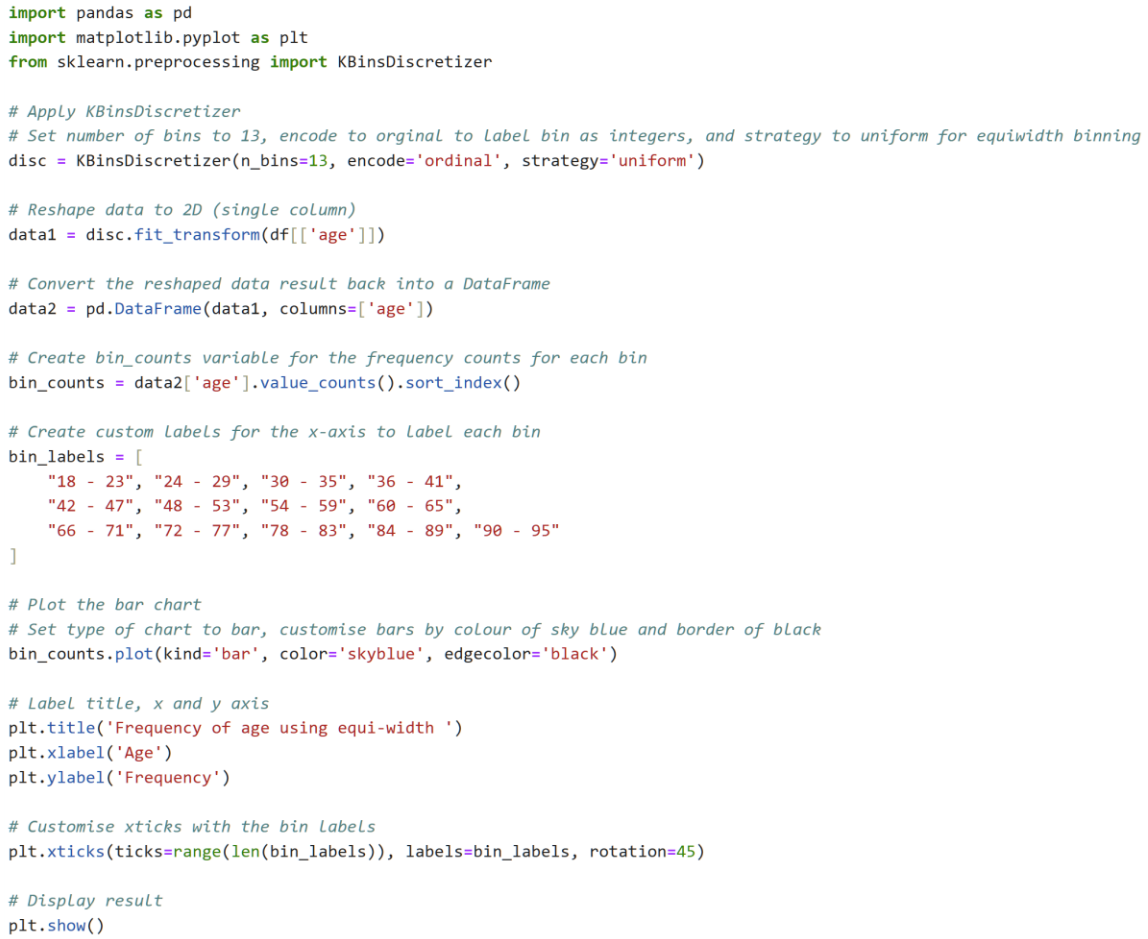
In this section, binning techniques are performed for the attribute ‘age.’ The number of bins is determined by 4 binning techniques in determining bins. The table below shows the number of bins after each calculation and are compared to other methods. After analysing each graph with corresponding bin numbers, judgement of appropriate number of bins is performed.

Method	Equation (where n = 2636)	Number of bins and graph
Square root	$\sqrt{n}$	52  Figure 27a. Binned age with square root method
Sturges	$1 + 3 \times 3 \log n$	13

		 Figure 27b. Binned age with Sturges method
Rice	$2 \times \sqrt[3]{n}$	28  Figure 27c. Binned age with Rice method
Scott	$\frac{\max - min}{3.5 \times \frac{stdev}{\sqrt[3]{n}}}$	30  Figure 27d. Binned age with Scott method

In the table above, figure 27a – figure 27d shows a visual representation of a binned age frequency graph with corresponding number of bins based on each method and equation. Figure 27a has 52 bins which shows an unclear distribution and thus, noisy data. The number of bins makes it difficult to analyse and understand patterns. However, using Sturges binning equation, creates a smooth and simplified pattern of the frequency as it does not include noisy data, enhancing overall visual clarity. However, since there are 2636 data points, it is important to consider that this oversimplifies the large volume of data. Figure 27c and figure 27d are both similar as the number of bins is 28 and 30 respectively. Because in figure 27c and 27d, it captures more detail, but this also means that it is more difficult to easily interpret data. Using Sturge's formula and giving 13 bins, thus shows a clearer distributing (data granularity) as it shows clarity especially in skewness and the overall trends.

Below are the techniques used using python in Jupyter notebook to create a binned age data.



```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import KBinsDiscretizer

# Apply KBinsDiscretizer
# Set number of bins to 13, encode to orginal to label bin as integers, and strategy to uniform for equiwidth binning
disc = KBinsDiscretizer(n_bins=13, encode='ordinal', strategy='uniform')

# Reshape data to 2D (single column)
data1 = disc.fit_transform(df[['age']])

# Convert the reshaped data result back into a DataFrame
data2 = pd.DataFrame(data1, columns=['age'])

# Create bin_counts variable for the frequency counts for each bin
bin_counts = data2['age'].value_counts().sort_index()

# Create custom labels for the x-axis to label each bin
bin_labels = [
    "18 - 23", "24 - 29", "30 - 35", "36 - 41",
    "42 - 47", "48 - 53", "54 - 59", "60 - 65",
    "66 - 71", "72 - 77", "78 - 83", "84 - 89", "90 - 95"
]

# Plot the bar chart
# Set type of chart to bar, customise bars by colour of sky blue and border of black
bin_counts.plot(kind='bar', color='skyblue', edgecolor='black')

# Label title, x and y axis
plt.title('Frequency of age using equi-width ')
plt.xlabel('Age')
plt.ylabel('Frequency')

# Customise xticks with the bin labels
plt.xticks(ticks=range(len(bin_labels)), labels=bin_labels, rotation=45)

# Display result
plt.show()
```

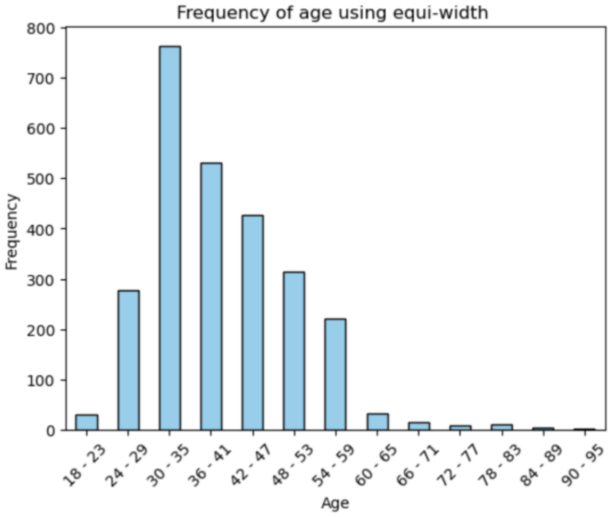


Figure 28. Equi-width binning of age

2.1.2 Equi-depth

It was found that the most appropriate number of bins is 7 by visualisation and comparing the standard deviation of various bin counts. A lower standard deviation is used for equi-depth as it equalises or divides each bin containing approximate frequency of data. Low standard deviation often indicates that there are less spread in the data, resulting in balanced bins.

Below is a result of the bins and its corresponding standard deviation.

Number of bins	Standard deviation
5	58.23
6	42.39
7	33.93
8	35.74
9	42.21
10	36.39
12	32.86
13	46.37
14	42.71
15	39.88
16	42.00
17	36.39
18	39.76
19	40.57
20	32.84

Three lowest standard deviations with the corresponding number of bins and bar chart:

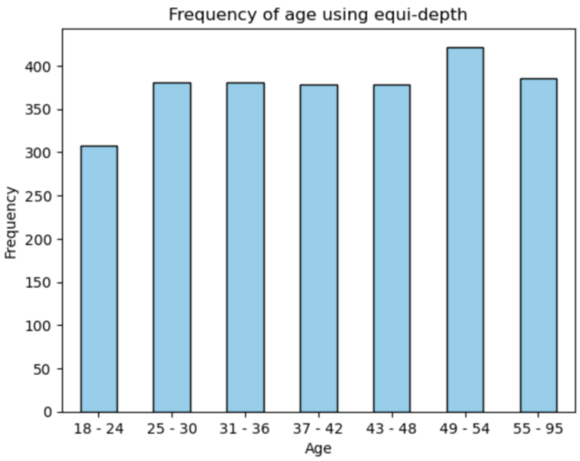


Figure 29a. Binned age data with 7 bins

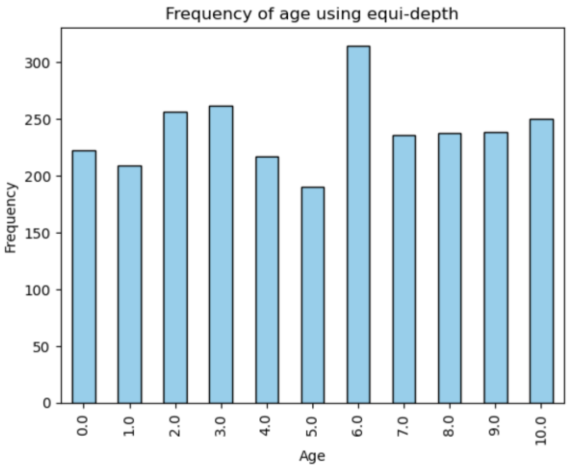


Figure 29b. Binned age data with 11 bins

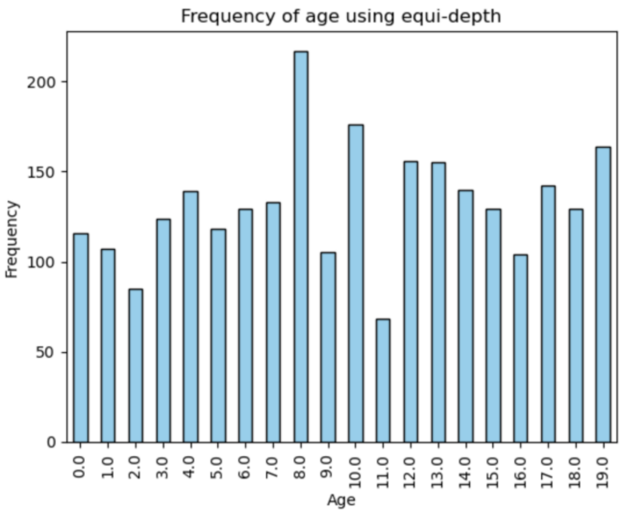
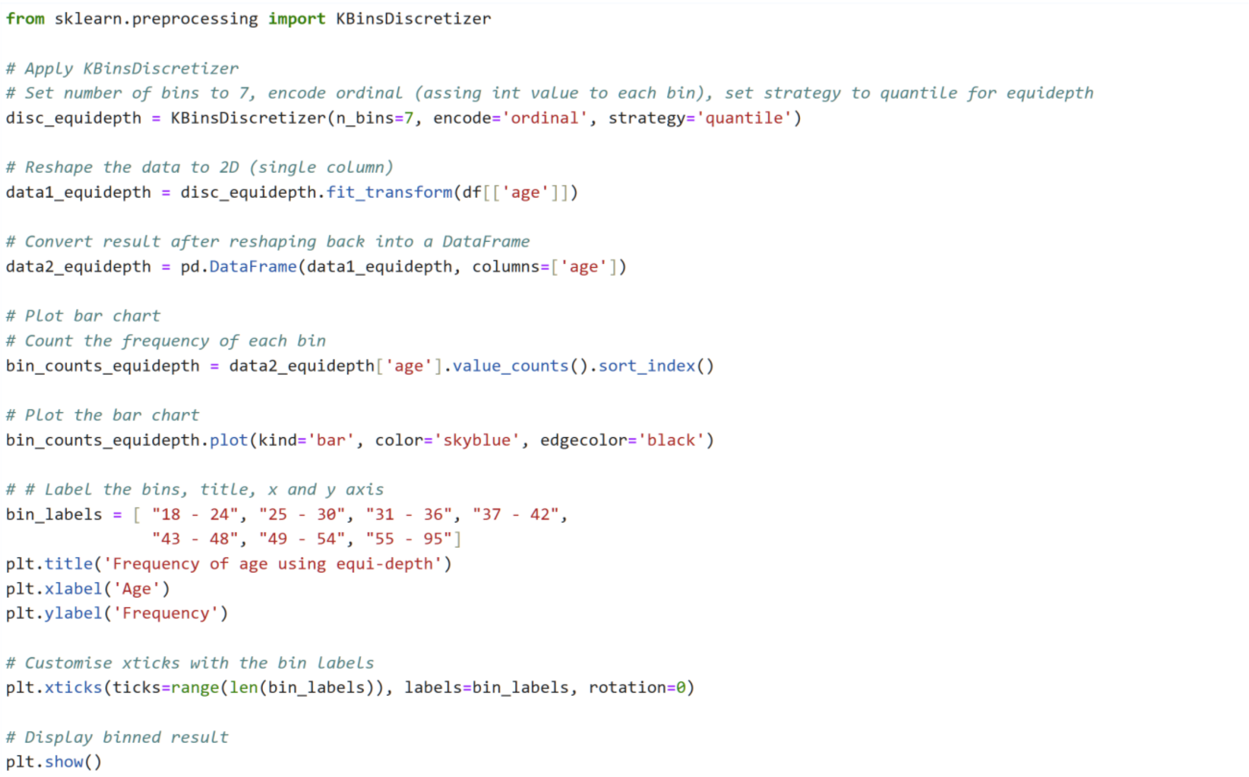


Figure 29c. Binned age data with 20 bins

From the results above, although 12 and 20 bins have a lower deviation than 12, the graph below is a comparison of the visual representation. From judgement, figure 29a has a good balance between interpretability as the bins are more evenly distributed visually. Figure 29b and figure 29c loses clarity and meaningful insight due to overcomplication.

Below is the technique used in python written in Jupyter notebook to create a binned age data bar chart using equi-depth with 7 bins.



```
from sklearn.preprocessing import KBinsDiscretizer

# Apply KBinsDiscretizer
# Set number of bins to 7, encode ordinal (assing int value to each bin), set strategy to quantile for equidepth
disc_equidepth = KBinsDiscretizer(n_bins=7, encode='ordinal', strategy='quantile')

# Reshape the data to 2D (single column)
data1_equidepth = disc_equidepth.fit_transform(df[['age']])

# Convert result after reshaping back into a DataFrame
data2_equidepth = pd.DataFrame(data1_equidepth, columns=['age'])

# Plot bar chart
# Count the frequency of each bin
bin_counts_equidepth = data2_equidepth['age'].value_counts().sort_index()

# Plot the bar chart
bin_counts_equidepth.plot(kind='bar', color='skyblue', edgecolor='black')

# # Label the bins, title, x and y axis
bin_labels = [ "18 - 24", "25 - 30", "31 - 36", "37 - 42",
               "43 - 48", "49 - 54", "55 - 95"]
plt.title('Frequency of age using equi-depth')
plt.xlabel('Age')
plt.ylabel('Frequency')

# Customise xticks with the bin labels
plt.xticks(ticks=range(len(bin_labels)), labels=bin_labels, rotation=0)

# Display binned result
plt.show()
```

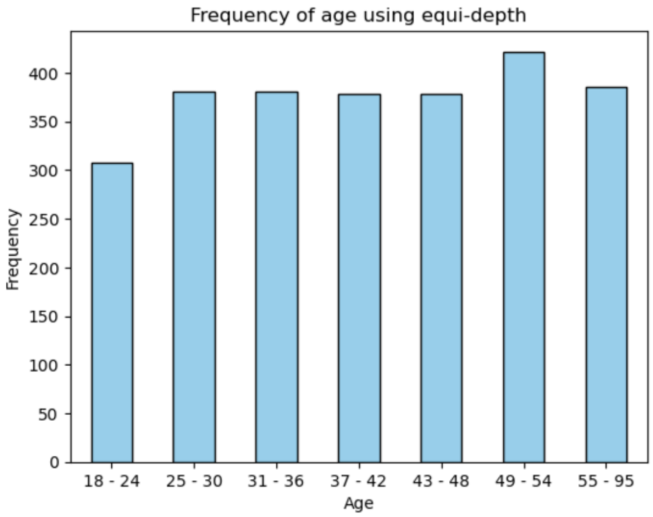
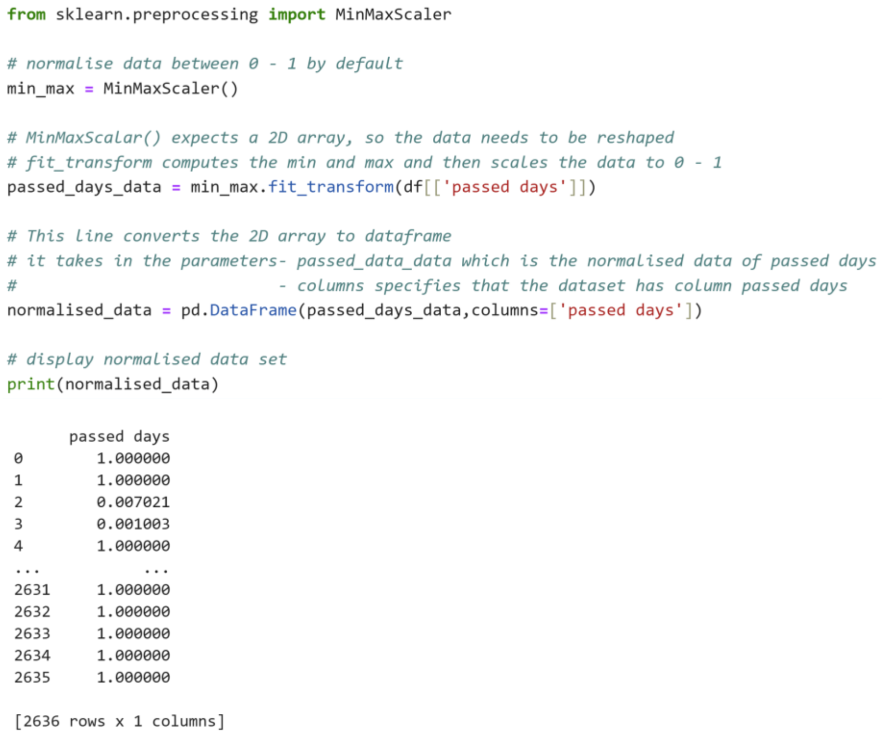


Figure 29. Equi-depth binning of age

2.2 Normalisation

Normalisation of ‘passed days’ attribute was conducted by using MinMaxScalar() and Z-scores. Min-max scaling rescaled the dataset to a range of [0,1].

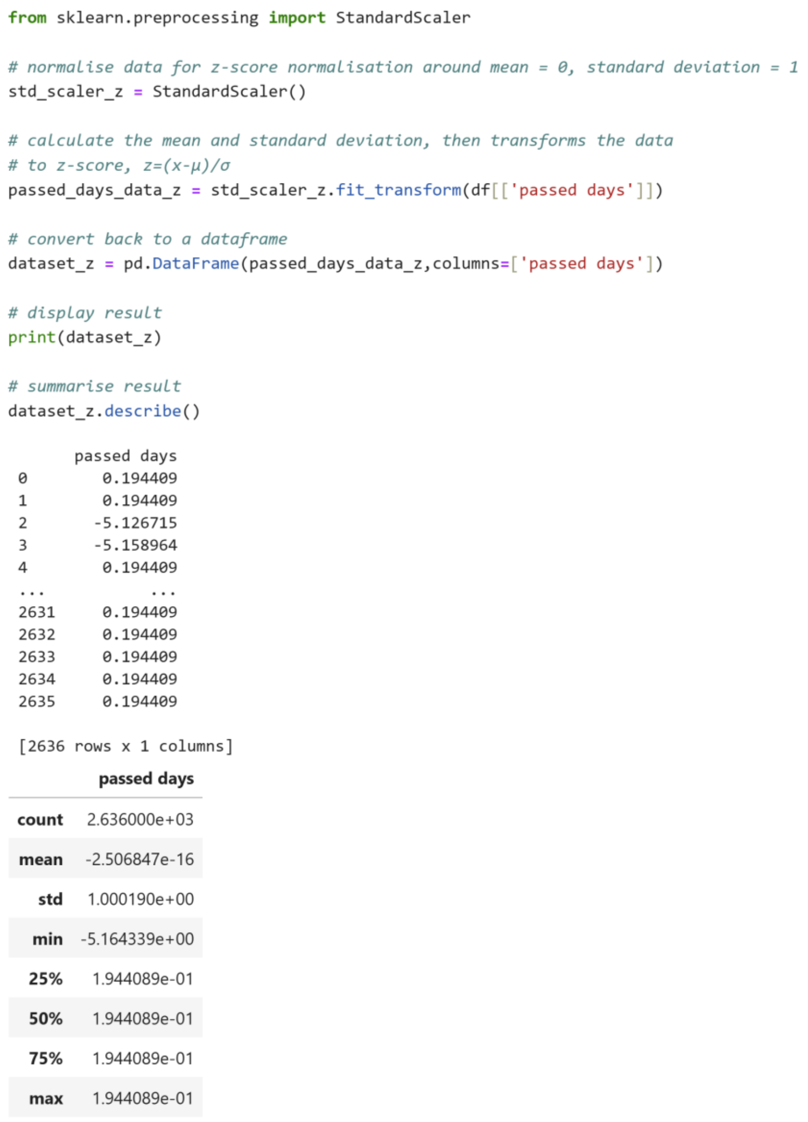
Below is the technique used to normalise the data using MinMaxScaler() from the sklearn library written in Jupyter notebook.



```
from sklearn.preprocessing import MinMaxScaler

# normalise data between 0 - 1 by default
min_max = MinMaxScaler()

# MinMaxScalar() expects a 2D array, so the data needs to be reshaped
# fit_transform computes the min and max and then scales the data to 0 - 1
passed_days_data = min_max.fit_transform(df[['passed days']])

# This line converts the 2D array to dataframe
# it takes in the parameters- passed_data_data which is the normalised data of passed days
#                           - columns specifies that the dataset has column passed days
normalised_data = pd.DataFrame(passed_days_data,columns=['passed days'])

# display normalised data set
print(normalised_data)
```

```
      passed days
0        1.000000
1        1.000000
2        0.007021
3        0.001003
4        1.000000
...           ...
2631     1.000000
2632     1.000000
2633     1.000000
2634     1.000000
2635     1.000000

[2636 rows x 1 columns]
```

Another method to normalise the ‘passed days’ variable was scaling by the z-scores. Z-score normalisation centered the data around the mean of 0 with a standard deviation of 1

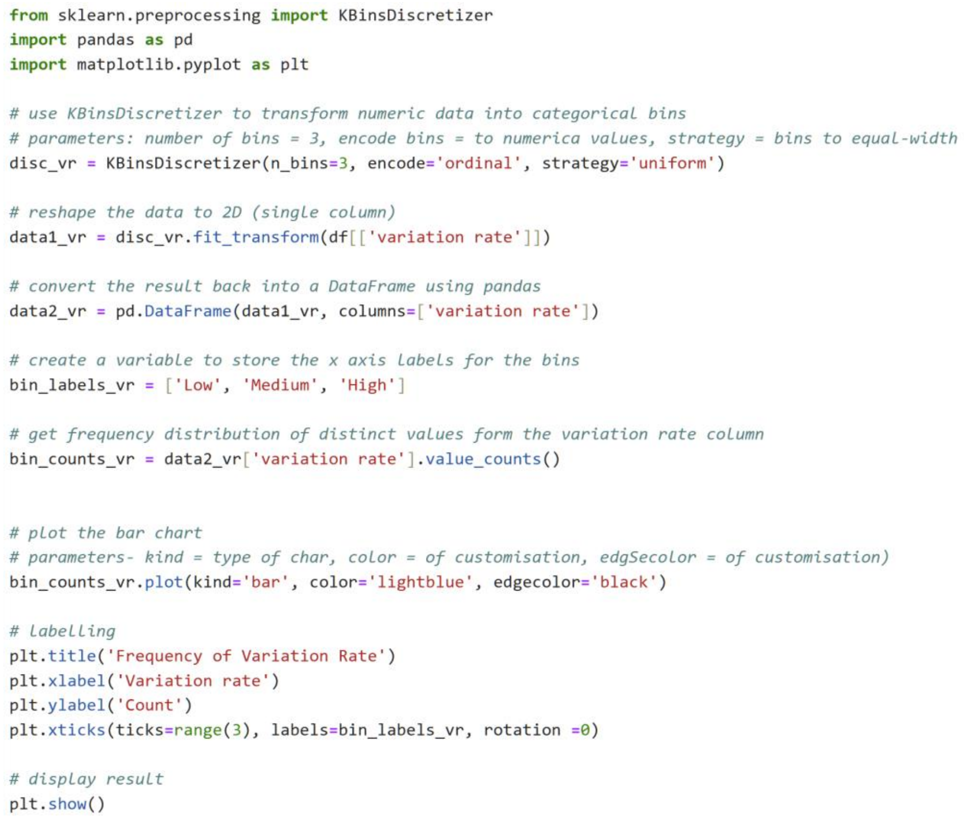


```
from sklearn.preprocessing import StandardScaler

# normalise data for z-score normalisation around mean = 0, standard deviation = 1
std_scaler_z = StandardScaler()

# calculate the mean and standard deviation, then transforms the data
# to z-score, z=(x-μ)/σ
passed_days_data_z = std_scaler_z.fit_transform(df[['passed days']])

# convert back to a dataframe
dataset_z = pd.DataFrame(passed_days_data_z,columns=['passed days'])

# display result
print(dataset_z)

# summarise result
dataset_z.describe()
```

```
      passed days
0        0.194409
1        0.194409
2       -5.126715
3       -5.158964
4        0.194409
...           ...
2631     0.194409
2632     0.194409
2633     0.194409
2634     0.194409
2635     0.194409

[2636 rows x 1 columns]
```

	passed days
count	2.636000e+03
mean	-2.506847e-16
std	1.000190e+00
min	-5.164339e+00
25%	1.944089e-01
50%	1.944089e-01
75%	1.944089e-01
max	1.944089e-01

2.3 Discretisation



```
from sklearn.preprocessing import KBinsDiscretizer
import pandas as pd
import matplotlib.pyplot as plt

# use KBinsDiscretizer to transform numeric data into categorical bins
# parameters: number of bins = 3, encode bins = to numerica values, strategy = bins to equal-width
disc_vr = KBinsDiscretizer(n_bins=3, encode='ordinal', strategy='uniform')

# reshape the data to 2D (single column)
data1_vr = disc_vr.fit_transform(df[['variation rate']])

# convert the result back into a DataFrame using pandas
data2_vr = pd.DataFrame(data1_vr, columns=['variation rate'])

# create a variable to store the x axis labels for the bins
bin_labels_vr = ['Low', 'Medium', 'High']

# get frequency distribution of distinct values form the variation rate column
bin_counts_vr = data2_vr['variation rate'].value_counts()


# plot the bar chart
# parameters- kind = type of char, color = of customisation, edgSecolor = of customisation)
bin_counts_vr.plot(kind='bar', color='lightblue', edgecolor='black')

# labelling
plt.title('Frequency of Variation Rate')
plt.xlabel('Variation rate')
plt.ylabel('Count')
plt.xticks(ticks=range(3), labels=bin_labels_vr, rotation =0)

# display result
plt.show()
```

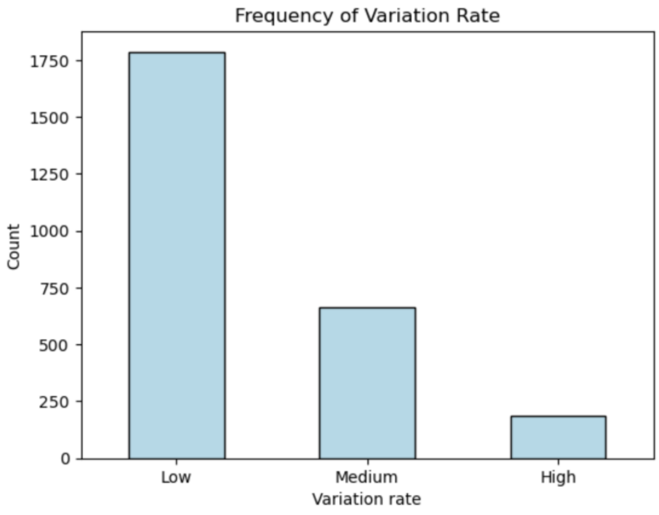
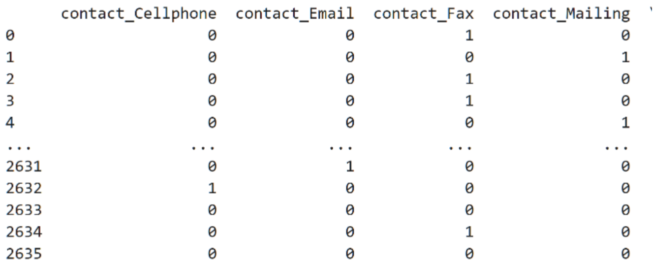


Figure 30. Discretised variation rate

Bins	Count
Low variation	1786
Medium variation	662
High variation	188

2.4 Binarisation

```
from sklearn.preprocessing import OneHotEncoder

# instantiate OneHotEncoder
oneHot = OneHotEncoder(dtype=int, sparse_output=False)

# apply one-hot encoding technique on the 'contact' variable
contact = oneHot.fit_transform(df[['contact']])

# take the categories (feature names) of 'contact'
contact_categories = oneHot.get_feature_names_out(['contact'])

# create a new DataFrame with the one-hot encoded data with contact and assigns each category as a separate labelled column
contact_encoded = pd.DataFrame(contact, columns=contact_categories)

# display one-hot encoded data
print(contact_encoded)
```



```
      contact_Cellphone  contact_Email  contact_Fax  contact_Mailing
0                     0              0            1                0
1                     0              0            0                1
2                     0              0            1                0
3                     0              0            1                0
4                     0              0            0                1
...                 ...            ...          ...              ...
2631                  0              1            0                0
2632                  1              0            0                0
2633                  0              0            0                0
2634                  0              0            1                0
2635                  0              0            0                0
```

3. Summary

Summary statistics:

- The categories were split into 2 types of data: categorical and numerical.
  - Categorical data had bar charts and/or pie charts.
  - Numerical data had table of statistics calculating mean, median, mode, Q1, Q3, IQR etc., a box plot and density/histogram chart.
- The presence of missing values on categorical attributes had a range of effects on the corresponding dataset. A pie chart and bar chart were implemented for visualisation to compare the frequency or count of each category.
  - For attributes with missing values, a pie chart was generated to explore its impact on the dataset and proportions of results. It was found that some categories had a relatively small proportion of missing values which did not have a greatly change the sizes of the categories.
  - However, there were larger percentages of missing value of a dataset which significantly gain insight on false representations without cleaning the data.
  - For example, for the 'housing' attribute, the majority of results was 'nonexistent.' This particular category is vague and does not reveal a clear explanation of what it means as non-existent. This indicates that data may have not been recorded, data is incomplete or missing. This required investigation to recover the dataset with 'yes' and 'no' results only, and discover what this category means.

Data exploration:

- Outliers were present in some numerical attributes which was identified when a datapoint deviates or exceeds from the minimum and maximum whisker in the box plot.
- Some interesting attributes were notes including the ‘contact’ and ‘passed days attribute’ which had missing data that can result in misleading information and affecting central tendency of data, respectively.
- There was extreme instance of the ‘euribor3m’ attribute with a maximum value of 50.45, which was quite unusual for an interest rate. So, further investigation for this datapoint is required as it skews the dataset.
- The variation rate and price index attributes were plotted in a scatter plot using KMeans. One cluster included lower price index and variation rate; the second cluster was grouped in the middle of the x-axis and price index was higher; the third cluster had datapoints to the positive or more to the right of the x-axis and price index was moderately high.
- A heatmap was created to display the correlation coefficient of numerical attributes. The correlation coefficient for variation rate and price index was at 0.78, indicating a positive correlation.

Data preprocessing:

- Binning techniques (equi-width and equi-depth) were used on the ‘age’ attribute to smoothen out the data for granularity and ease of interpretation.
    - For equi-width, the number of bins was selected by using binning method formulas and compared the results to judge the most appropriate bin count. It was found that 13 bins were sufficient to gain insight on the skewness of age without displaying noisy data.
    - For equi-depth, the bar chart with 7 bins showed the most balanced out bars for each bin.
- Normalisation of ‘passed days’ attribute was conducted. After analysing the z-score normalisation, some values indicated that they were close to the mean (like 0.194...). However, there were some that deviated very far from the mean with a value of -5.142...
- Discretising the ‘variation rate’ converted continuous data into three categories with three bins of equal width. The dataset was group by low, medium and high categories to group similar values together, allowing easy analysis.
- The ‘contact’ variable was binarised using the one-hot encoding technique resulting in a 2D vector. This allowed the categories to have unique values with 0 or 1 values.