Assessment Task 2: Data exploration and preparation

Name: Khoi Huynh

A. Initial Data Exploration

A1. Attribute types

Attribute Name	Description	Attribute Type
age	Age of person	Ratio
job	Type of job	Nominal
marital	Marital status	Nominal
education	Education level	Ordinal
default	Credit in default	Nominal
housing	Housing loan	Nominal
loan	Personal loan	Nominal
contact	Contact communication type	Nominal
month	Last contact month of year	Ordinal
day_of_week	Last contact day of the week	Ordinal
duration	Last contact duration	Ratio
campaign	Number of contacts performed during the campaign	Ratio
passed days	Number of days that passed by after the client was last contacted from a previous campaign	Ratio
previous	Number of contacts performed before the campaign	Ratio
poutcome	Outcome of the previous marketing campaign	Ordinal
variation rate	Employment variation rate – quarterly indicator	Interval
price index	Consumer price index – monthly indicator	Interval
confidence index	Consumer confidence index – monthly indicator	Interval
euribor3m	Euribor 3-month rate – daily indicator	Interval
no.employed	Number of employees – quarterly indicator	Ratio
subscribed	Subscribed a term deposit	Nominal
state	Name of the state	Nominal

A2 & A3. Summarized properties and exploration of attributes

Note: To avoid redundant data, the values in the **“Value (without missing values)”** column will be left blank or empty to indicate that it is the same as the corresponding values in the **“Value (with missing values)”** column. This signifies that no changes have occurred and NOT because it is actually null/empty/blank.

Attribute: age		
Statistics	**Value (with missing values)**	**Value (without missing values)**
Minimum	17	
Maximum	83	
Range	66	
25% Quantile	32	
50% Quantile (Median)	38	
75% Quantile	47	
Mean	39.869	
Mode	35	
Mean Absolute Deviation	8.353	
Standard Deviation	10.236	
Variance	104.776	
# Unique values	65	

The box plot in Figure 1 illustrates the age distribution, which is relatively diverse as the range covers most of the age groups from adolescents to the elderly. The youngest person to be observed is just under 20 years old whereas the oldest person is in the early 80s stage. Additionally, some observations are potentially outliers since they exceeded the threshold of $Q3 + 1.5 * IQR$.

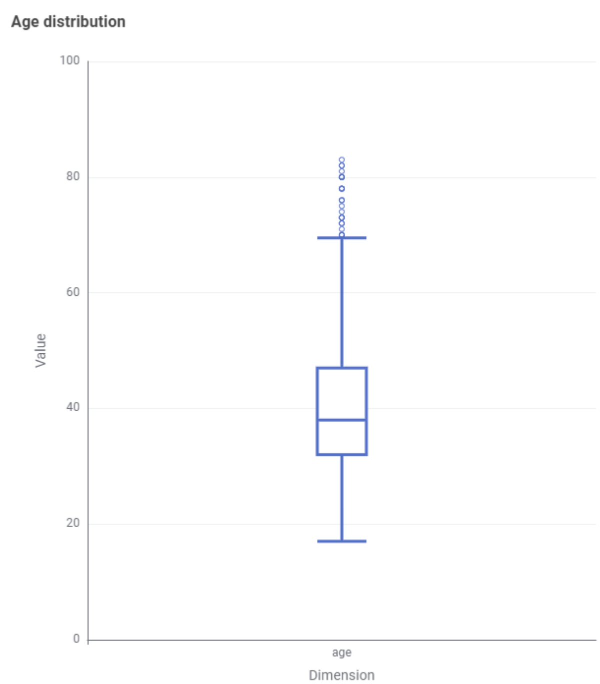


Figure 1

A closer inspection shows that the age distribution is slightly right skewed and the majority of the observations are in their 30s. This is evident from Figure 2a where the equal-width binning method has been carried out with $width = 10$. Furthermore, Figure 2b shows the average success index of different age groups. **The specific details on how an average success index can be found in the "poutcome" attribute section.** According to Figure 2b, the best-targeted age group for the marketing campaign will be people in their 60s.

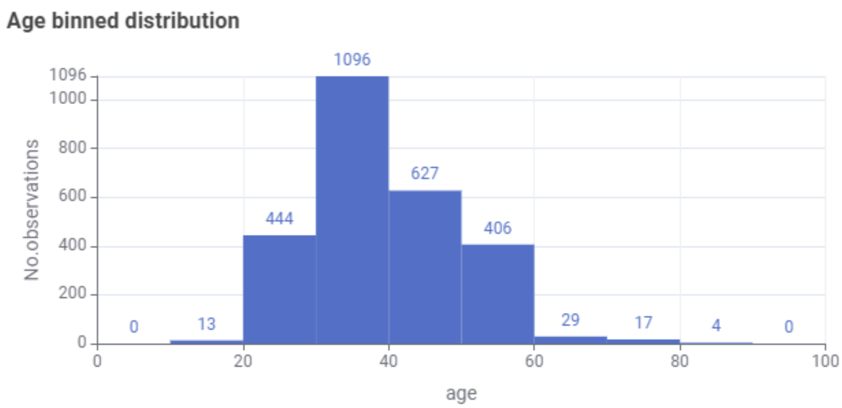


Figure 2a

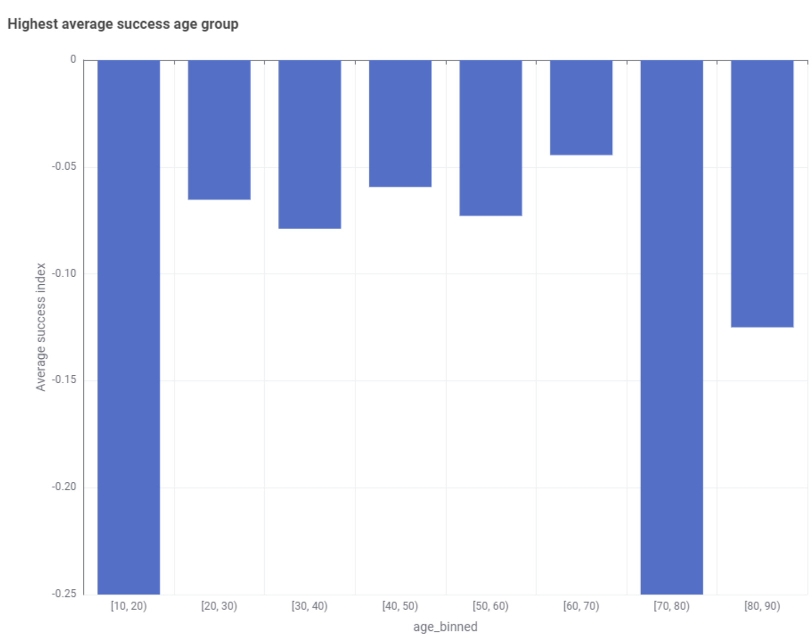


Figure 2b

Figure 3, where age has been discretized into age groups (boundaries shown in the table below), also conveys a similar finding where 80% of the observations fall into the middle-age group.

Age Group	Boundaries
Young	$(-\infty, 31)$
Middle Age	$[31, 61)$
Old	$[61, \infty)$

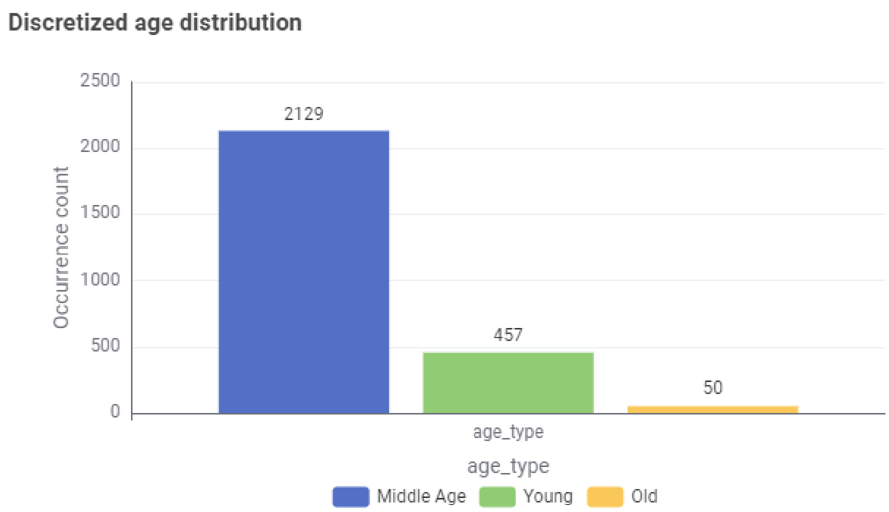


Figure 3

Attribute: job		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	12	11
Mode	admin	
10 most common values	1. admin (662; 25.11%) 2. blue-collar (580; 22.0%) 3. technician (464; 17.6%) 4. services (262; 9.94%) 5. management (178; 6.75%) 6. retired (95; 3.6%) 7. entrepreneur (91; 3.45%) 8. self-employed (87; 3.3%) 9. unemployed (65; 2.47%) 10. housemaid (63; 2.39%)	1. admin. (662; 25.41%) 2. blue-collar (580; 22.26%) 3. technician (464; 17.81%) 4. services (262; 10.06%) 5. management (178; 6.83%) 6. retired (95; 3.65%) 7. entrepreneur (91; 3.49%) 8. self-employed (87; 3.34%) 9. unemployed (65; 2.5%) 10. housemaid (63; 2.42%)

While there are missing values within the job attribute (Figure 4), it only accounts for approximately 1% of the entire dataset. Therefore, the chosen method for resolving missing values in this case is to remove them entirely (Figure 5). Since the proportion of missing values is minuscule, it is unsurprising that the impact caused by the missing values is insignificant. Both before and after the removal of missing values, the mode of the job attribute is still admin which takes up a quarter of the dataset followed closely behind by blue-collar jobs.

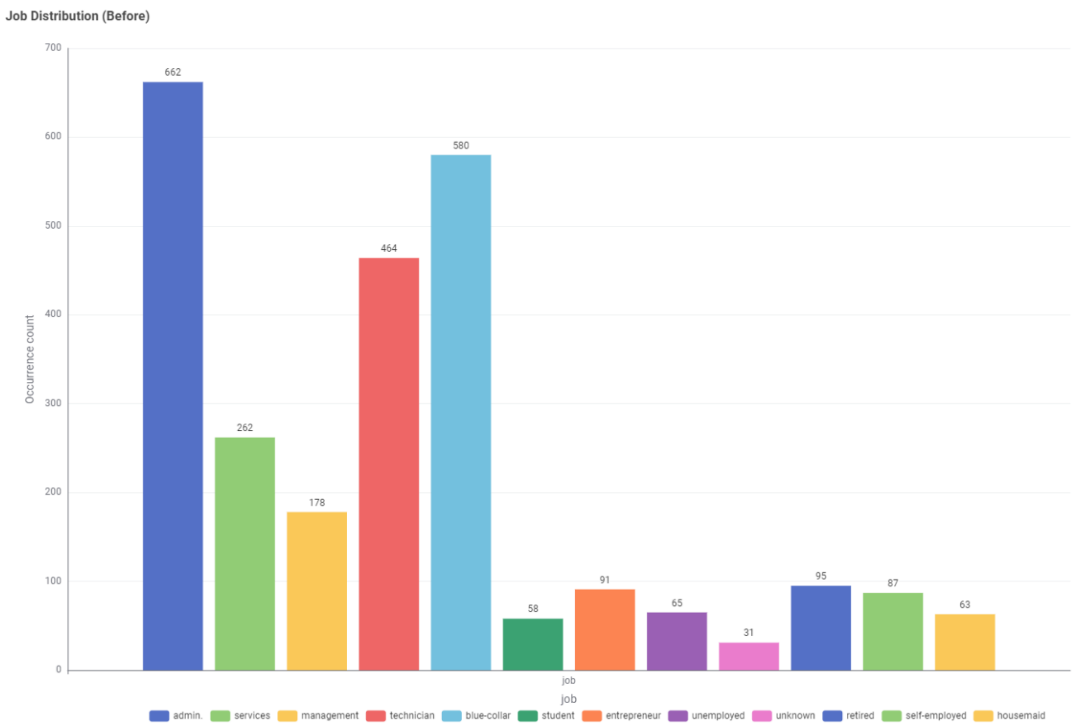


Figure 4

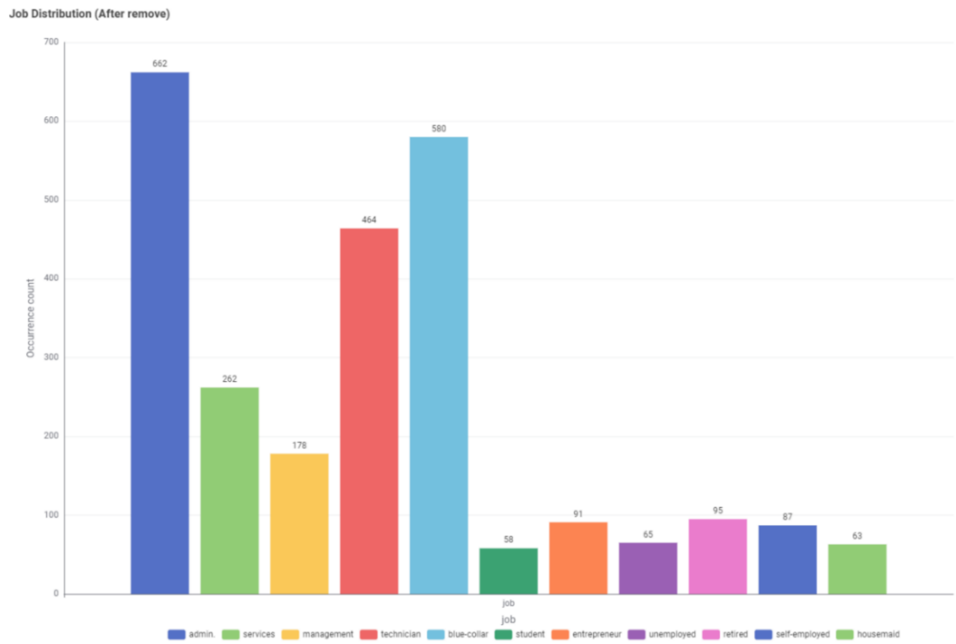


Figure 5

Furthermore, Figure 6 shows the education level of different jobs. From Figure 6, it can be seen that the majority of admins own a university degree which indicates that having a university degree is the standard. On the contrary, the education level required for technicians is more flexible as they can either choose to pursue a university degree or partake in a professional course. Finally, most people in the service industry stopped at high school.

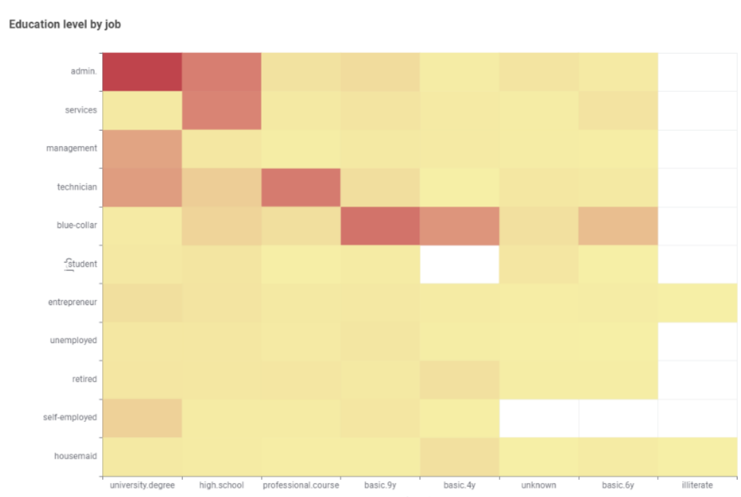


Figure 6

Attribute: marital		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	4	3
Mode	married	
10 most common values	1. married (1573; 59.67%) 2. single (736; 27.92%) 3. divorced (319; 12.1%) 4. unknown (8; 0.3%)	1. married (1573; 59.86%) 2. single (736; 28.01%) 3. divorced (319; 12.14%)

Similar to the "job" attribute, the missing values in the marital attribute are minuscule, taking up less than 1% of the entire dataset (Figure 7a). As a result, missing values are simply removed without having a serious impact on the dataset (Figure 7b). As evident from Figure 7 and Figure 8, the marital status of married is still the mode both before and after the removal of missing values. Roughly 60% of all observations indicated they are married, doubling that of the number of single people.

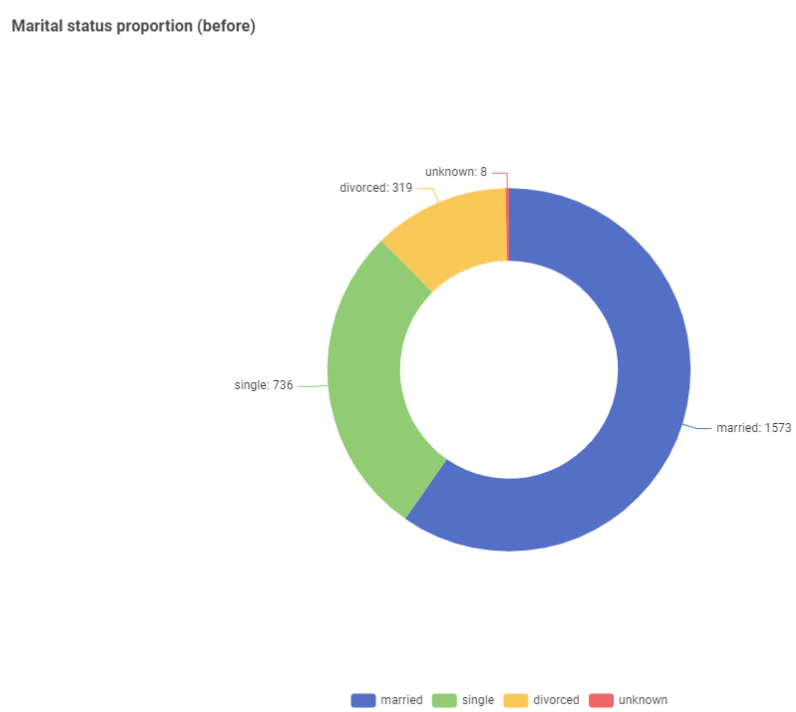


Figure 7a

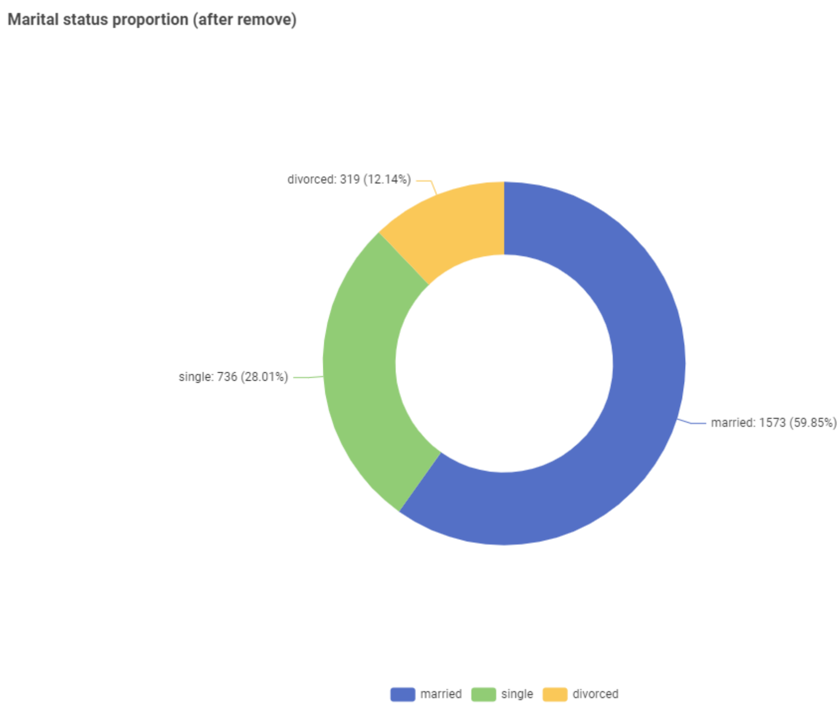


Figure 7b

Figure 8 below shows the average success index of each marital status. The marital status with the highest success index is “single”.

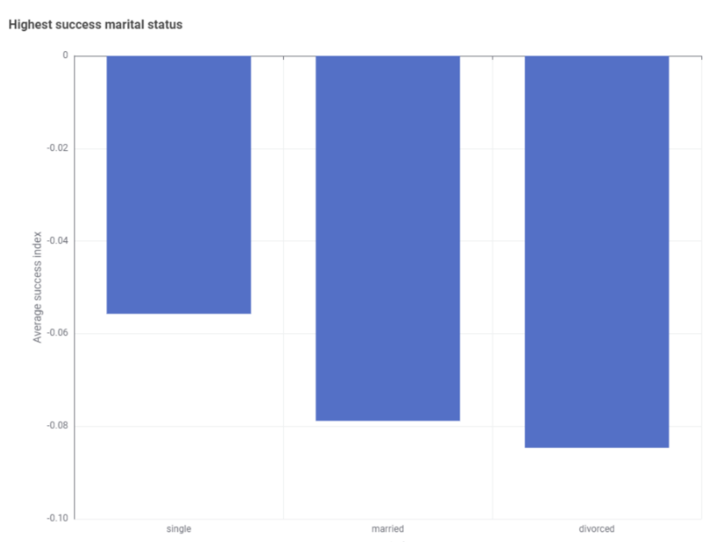


Figure 8

Attribute: education		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	8	7
Mode	university.degree	
10 most common values	1. university.degree (803; 30.46%) 2. high.school (587; 22.27%) 3. basic.9y (395; 14.98%) 4. professional.course (327; 12.41%) 5. basic.4y (256; 9.71%) 6. basic.6y (150; 5.69%) 7. unknown (116; 4.4%) 8. illiterate (2; 0.08%)	1. university.degree (919; 34.86%) 2. high.school (587; 22.27%) 3. basic.9y (395; 14.98%) 4. professional.course (327; 12.41%) 5. basic.4y (256; 9.71%) 6. basic.6y (150; 5.69%) 7. illiterate (2; 0.08%)

Unlike the job or marital attribute, the number of missing values in the “education” attribute can be considered (Figure 9) significant. As a result, the missing values cannot be discarded but they are imputed with the mode (Figure 10). Because the missing values are imputed with the mode, the mode itself does not change, however, its proportion will increase. After resolving the missing values, more than a third of the observations have a university degree, followed by a high school degree.

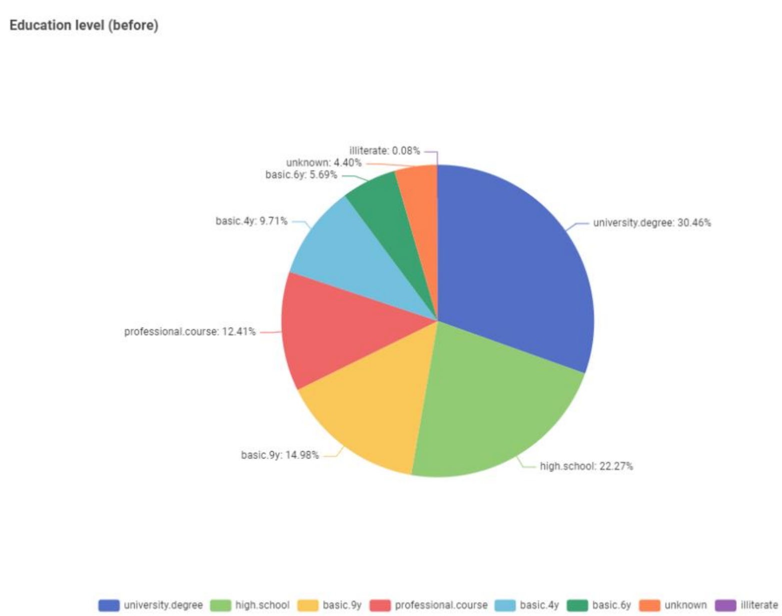


Figure 9

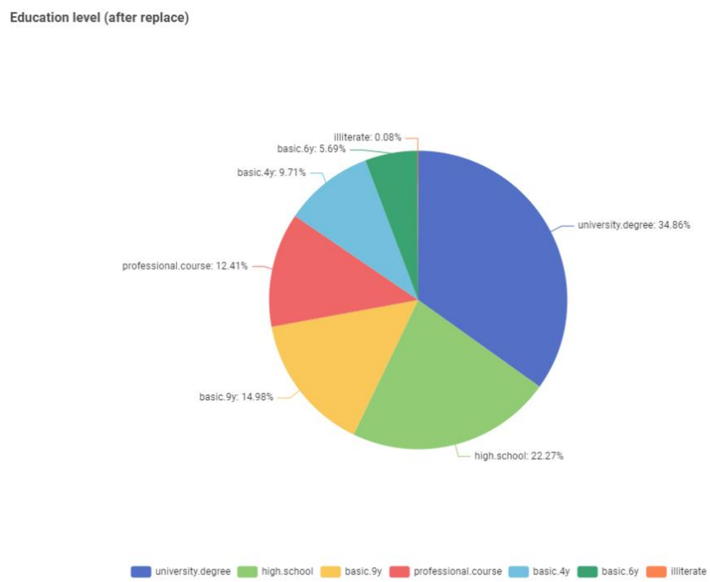


Figure 10

More importantly, since education is an ordinal datatype as it has an inherent ranking, we can assign numerical weights to each of the levels to extract a meaningful evaluation metric called the education index in order to gauge the average education level of each state. The details for the corresponding weights of each level are shown below:

Education level	Assigned value
illiterate	0
basic.4y	1
basic.6y	2
basic.9y	3
high.school	4
university.degree	5
professional.course	5

Figure 11 shows the average education level of each state, where there are only marginal differences between one another. This indicates that the education level of each state is quite similar to each other and a person on average will have completed high school regardless of which state they live in.

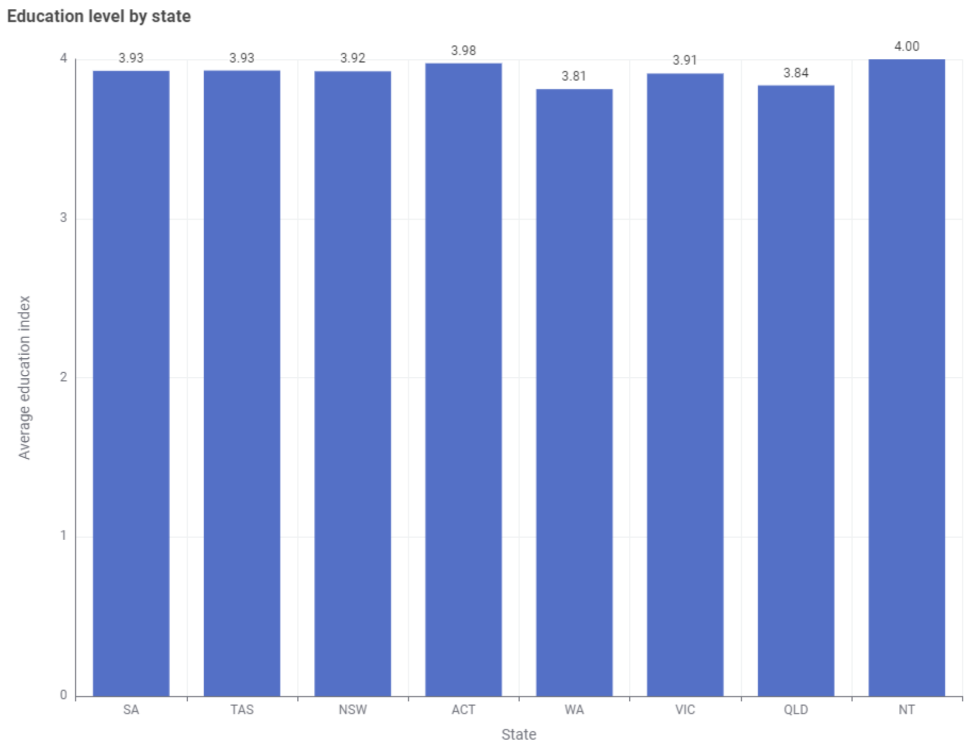
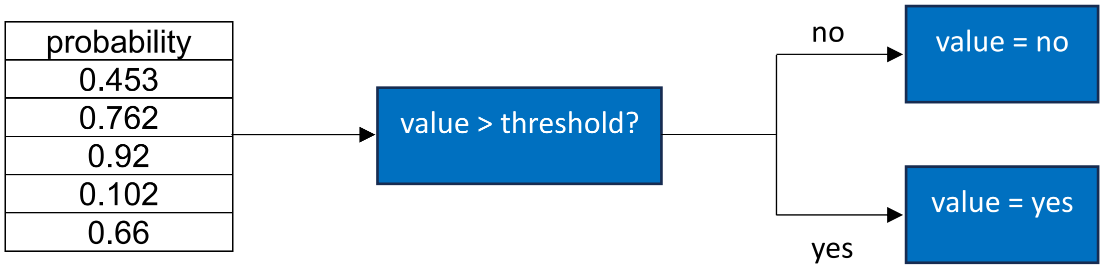


Figure 11

Attribute: default		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	2	
Mode	no	
10 most common values	1. no (2077; 78.79%) 2. unknown (559; 21.21%)	1. no (2495; 95.75%) 2. yes (112; 4.25%)

The default attribute is a binary attribute that takes a value of either yes or no. More importantly, a substantial proportion of the values are missing values (approximately 20%) as shown in Figure 12. Therefore, it can neither be removed completely nor imputed using the mode since it will just make the entire distribution to be 100% no which is very unlikely. To counteract this problem, we can do some feature engineering by creating a new attribute called probability. The attribute will subsequently be filled with randomly generated numbers within the range $[0, 1]$ to mimic the probability that its value is “no”. We will then use the randomly generated number and compare it with the probability threshold ($threshold = 0.7879$), to determine if the missing value is going to be “no” or “yes”.



Due to the numbers being randomly generated, the exact occurrence counts of “no” and “yes” will change each time we run the random number generator. Nevertheless, the overall proportion of “yes” and “no” should be very similar between each run since the values are determined based on the threshold. For example, there are currently 559 missing values and 78.79% of the observations are “no”, then we can assume that a similar proportion of the missing values will also be “no”.

$$Expected\ number\ of\ \text{"no" in missing values} = 559 * 0.7879 \approx 440$$

$$Expected\ number\ of\ \text{"yes" in missing values} = 559 - 440 = 119$$

Comparing the number of “yes” we got (112), we can see that it is very similar to the expected number (119). In fact, after testing multiple runs, the proportion of “yes” in the missing values is consistently in the range of $[4.0, 4.5]$. Figure 13 shows the distribution after imputing missing values where more than 95% of the values are “no”.

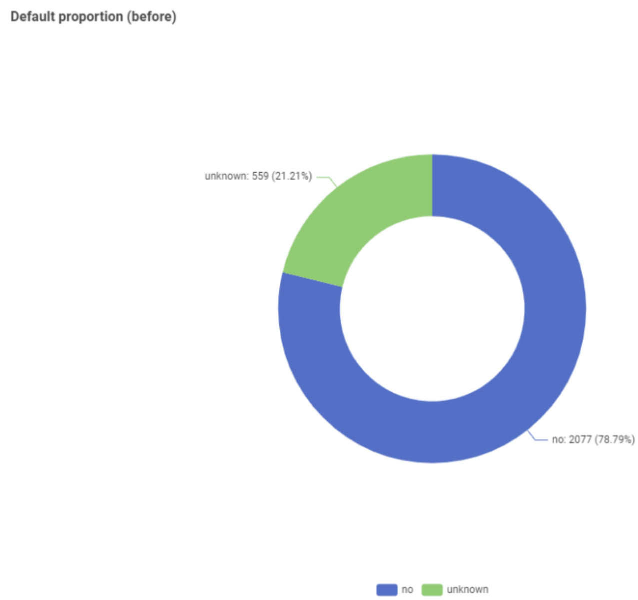


Figure 12

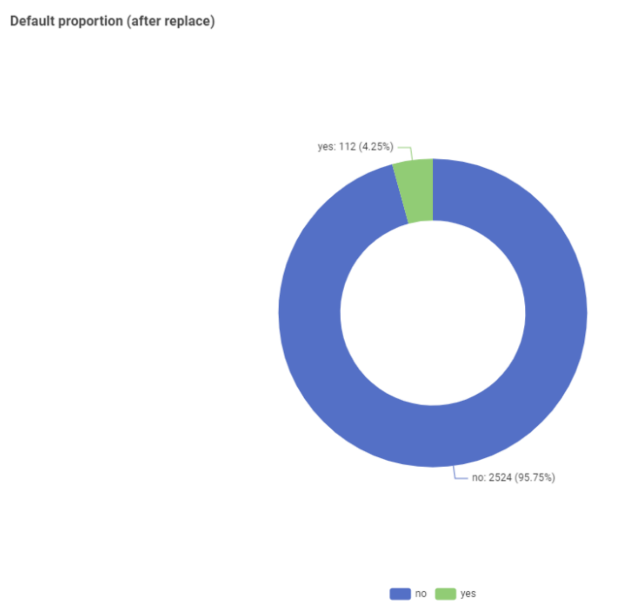


Figure 13

Attribute: housing		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	3	2
Mode	yes	
10 most common values	1. yes (1446; 54.86%) 2. no (1123; 42.6%) 3. unknown (67; 2.54%)	1. yes (1513; 57.40%) 2. no (1123; 42.6%)

Figure 14 shows the proportion of the binary attribute “housing”. It can be seen clearly that more than half of the observations answered “yes” as well as approximately 3% of the observations are missing values. To counteract the data quality issue of missing values, the missing values are imputed by replacing them with the mode as shown in Figure 15. The mode remains unchanged after imputation as the mode is still “yes”.

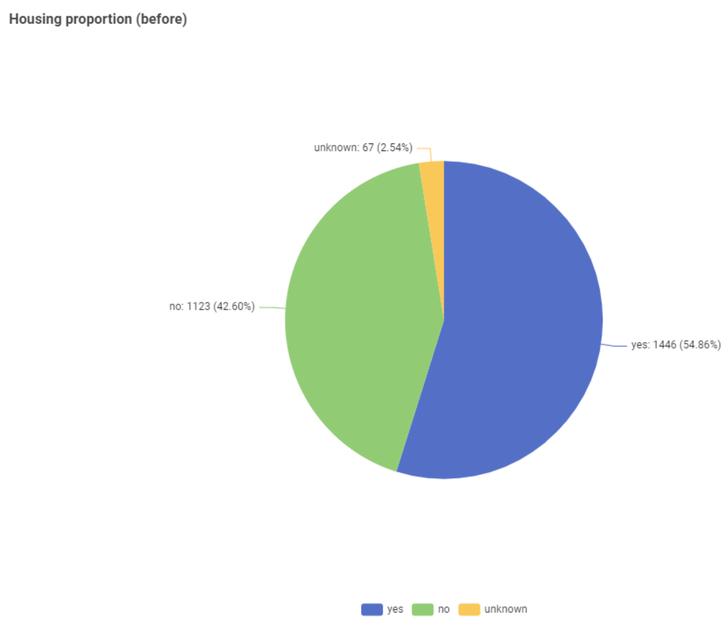


Figure 14

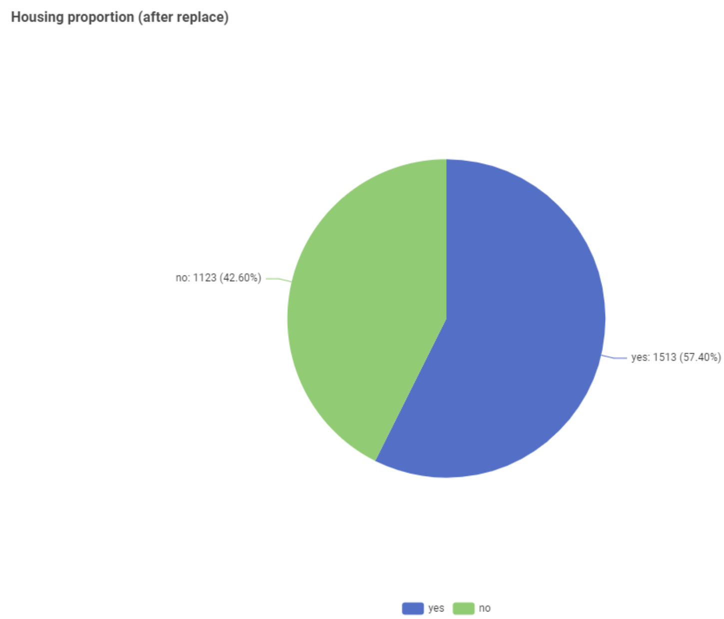


Figure 15

Attribute: loan		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	3	2
Mode	no	
10 most common values	1. no (2168; 82.25%) 2. yes (401; 15.21%) 3. unknown (67; 2.54%)	1. no (2235; 84.79%) 2. yes (401; 15.21%)

Figure 16 shows the proportion of the binary attribute “loan”. Evidently, an overwhelming number of observations answered “no” as well as approximately 3% of the observations are missing values. To counteract the data quality issue of missing values, the missing values are imputed by replacing them with the mode as shown in Figure 17. The mode remains unchanged after imputation as the mode is still “no”.

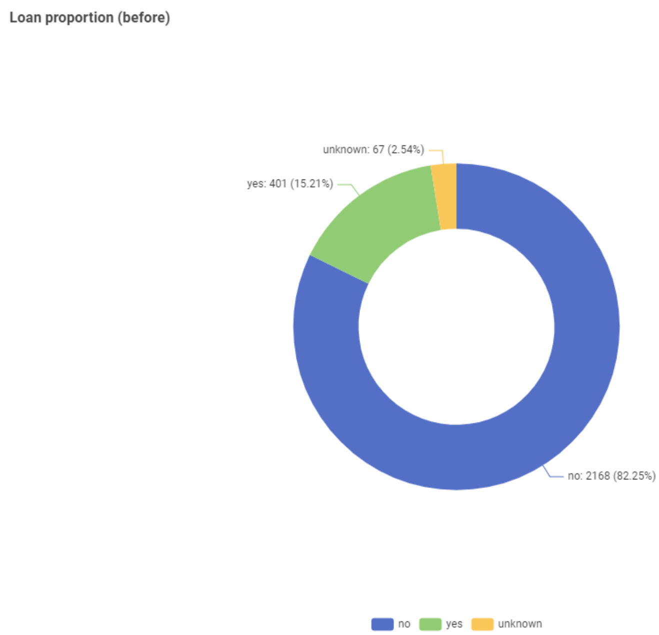


Figure 16

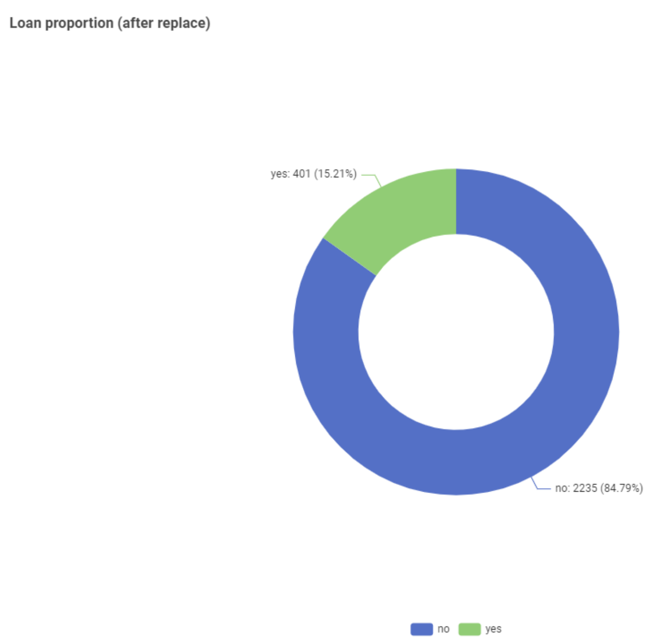


Figure 17

Attribute: contact		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	5	
Mode	Cellphone	
10 most common values	1. Cellphone (558; 21.17%) 2. Fax (523; 19.84%) 3. Email (522; 19.8%) 4. Telephone (517; 19.61%) 5. Mailing (516; 19.58%)	

Figure 18 shows the frequency of each contact communication type. Unsurprisingly, cellphone is the most popular contact method overall, followed by fax and email. On the contrary, mailing and telephone are the least popular contact methods. Nevertheless, when comparing the proportions of each contact method, the margin of difference between them is relatively small as each method equally takes up approximately 20% of the entire dataset. This indicates that all types of contact methods are roughly equally relevant to each other and none of them had gone out of date.

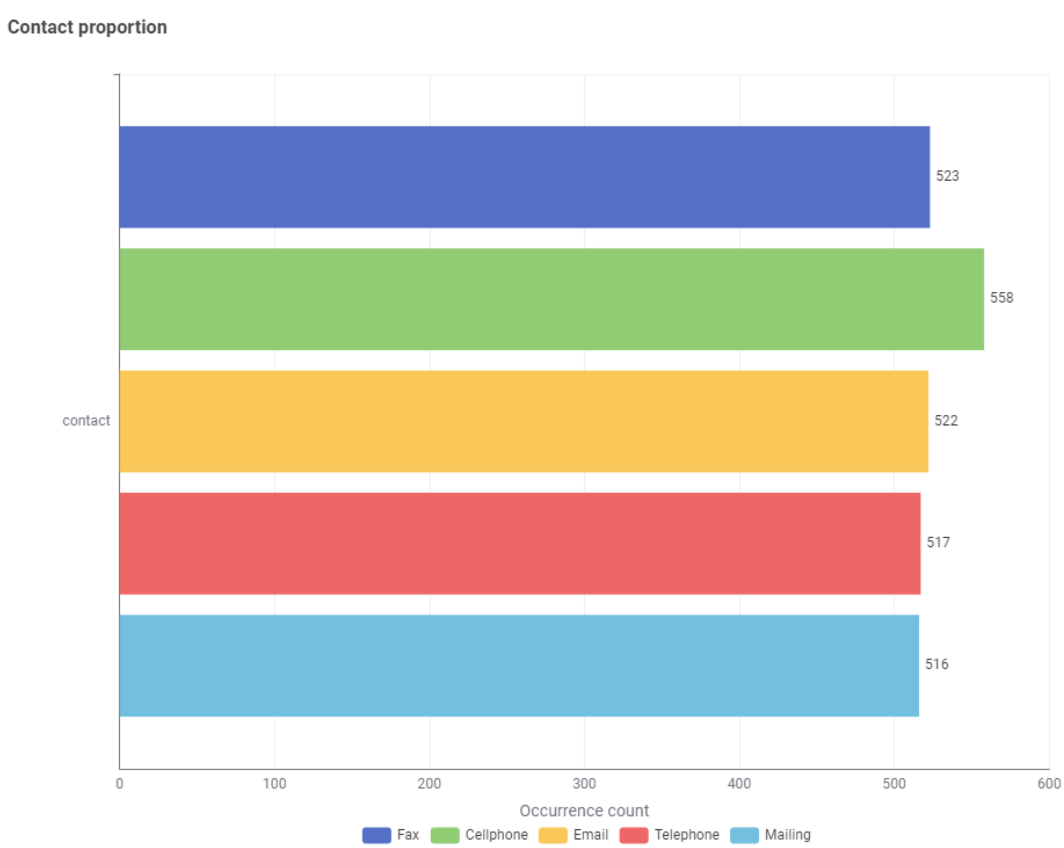


Figure 18

More interestingly, popularity does not always correspond to effectiveness as illustrated in Figure 19. Figure 19 shows the overall average success index of each contact method for the previous marketing campaign**.** Despite the average success index of all contact methods is in the negative, that does not mean they are all ineffective. In fact, it is because of the ineffectiveness of the previous marketing campaign that caused all contact methods to be negative. It can clearly be seen that email is the contact method that yields the highest success rate.

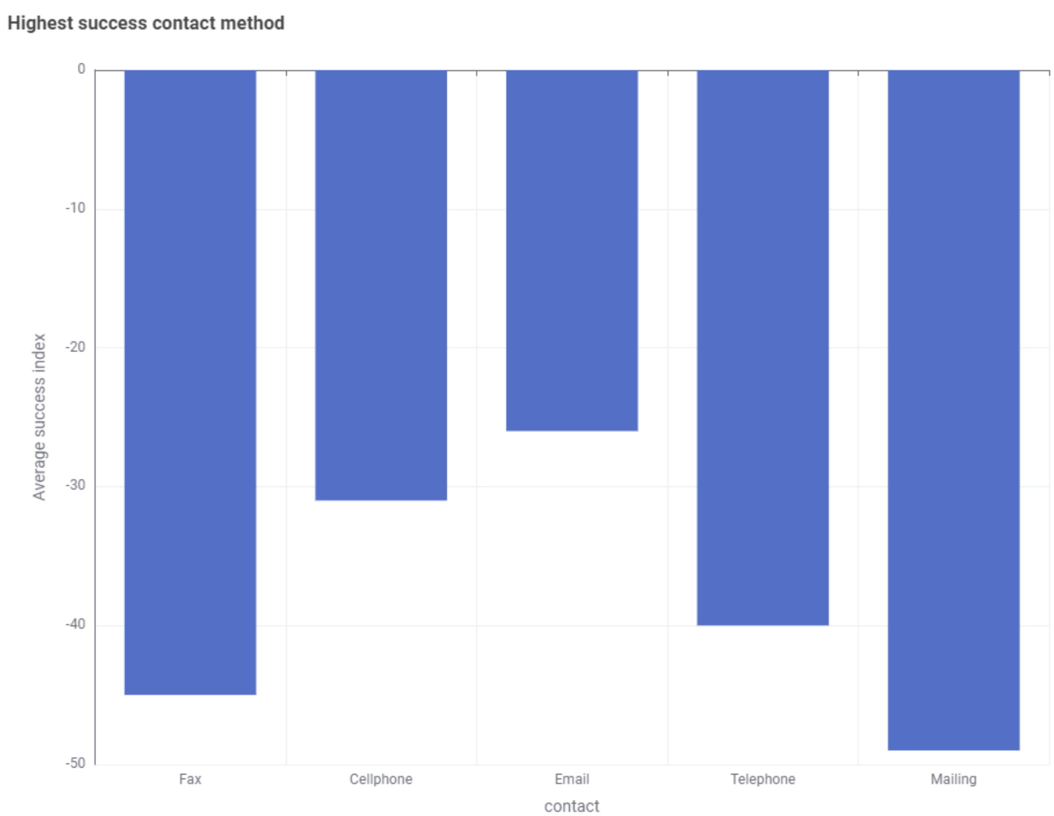


Figure 19

However, it is important to remember that Figure 19 only shows the average success index of each contact method **in isolation**; meaning without any other attributes such as “state” influencing it. This means if we want to contact a person without any relevant information, the best method to use will be email. But later in the report, other attributes might dictate or influence what contact method is the best. For example, it might be more effective to contact a person living in the ACT state via mailing rather than email.

Attribute: month		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	10	
Mode	1	
10 most common values	1. 1 (894; 33.92%) 2. 3 (435; 16.5%) 3. 4 (367; 13.92%) 4. 2 (352; 13.35%) 5. 6 (275; 10.43%) 6. 9 (181; 6.87%) 7. 5 (53; 2.01%) 8. 10 (41; 1.56%) 9. 8 (29; 1.1%) 10. 7 (9; 0.34%)	

Based on Figure 20, it is clear that most of the last contacts occurred in January. January alone takes up a third of all values in the attribute, doubling that of March, the second most frequent month. July, on the other hand, contributes to less than 1% of the total values. Also, it is noteworthy to point out that no contacts were made during November and December. There also seems to be a trend, while inconsistent, which shows that later in the year, the number of contacts decreases.

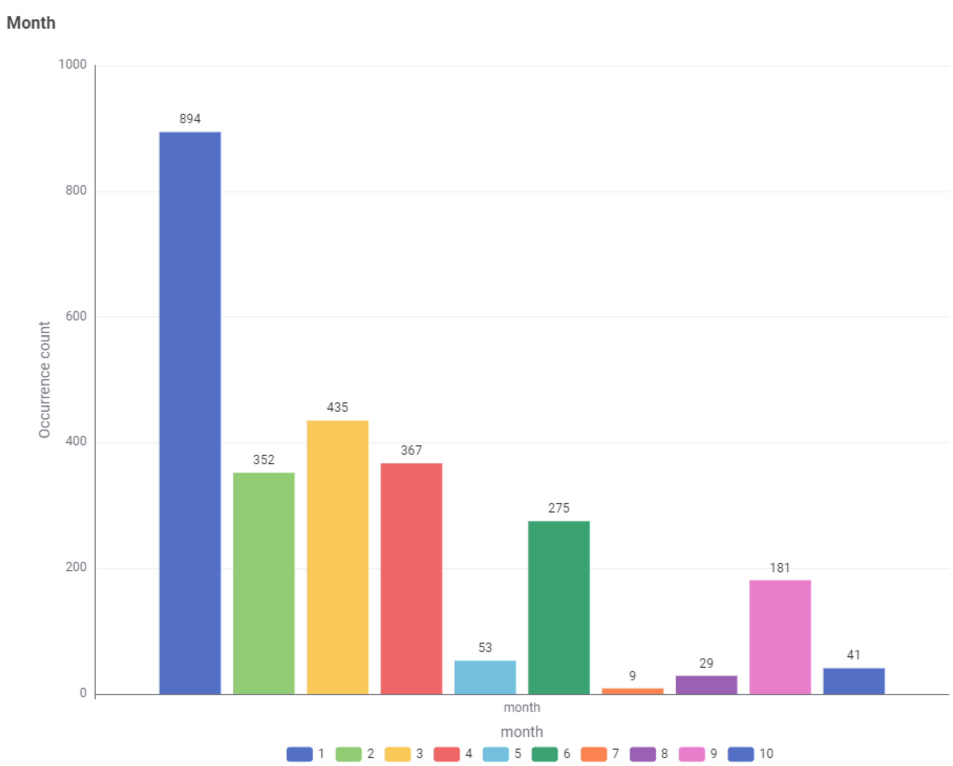


Figure 20

Out of the 10 months, only 2 months yield a positive average success index. Interestingly, the month with the highest average success index turns out to be the month with the lowest number of contacts, July, as shown in Figure 21. Furthermore, July is approximately 10 times higher than the second-best month, March, which despite having a positive average success index, is quite insignificant. Finally, the months February and August have an average success index of 0, not because there are no contacts made in these 2 months as seen in Figure 20 above, but because the number of successes and failures are the same.

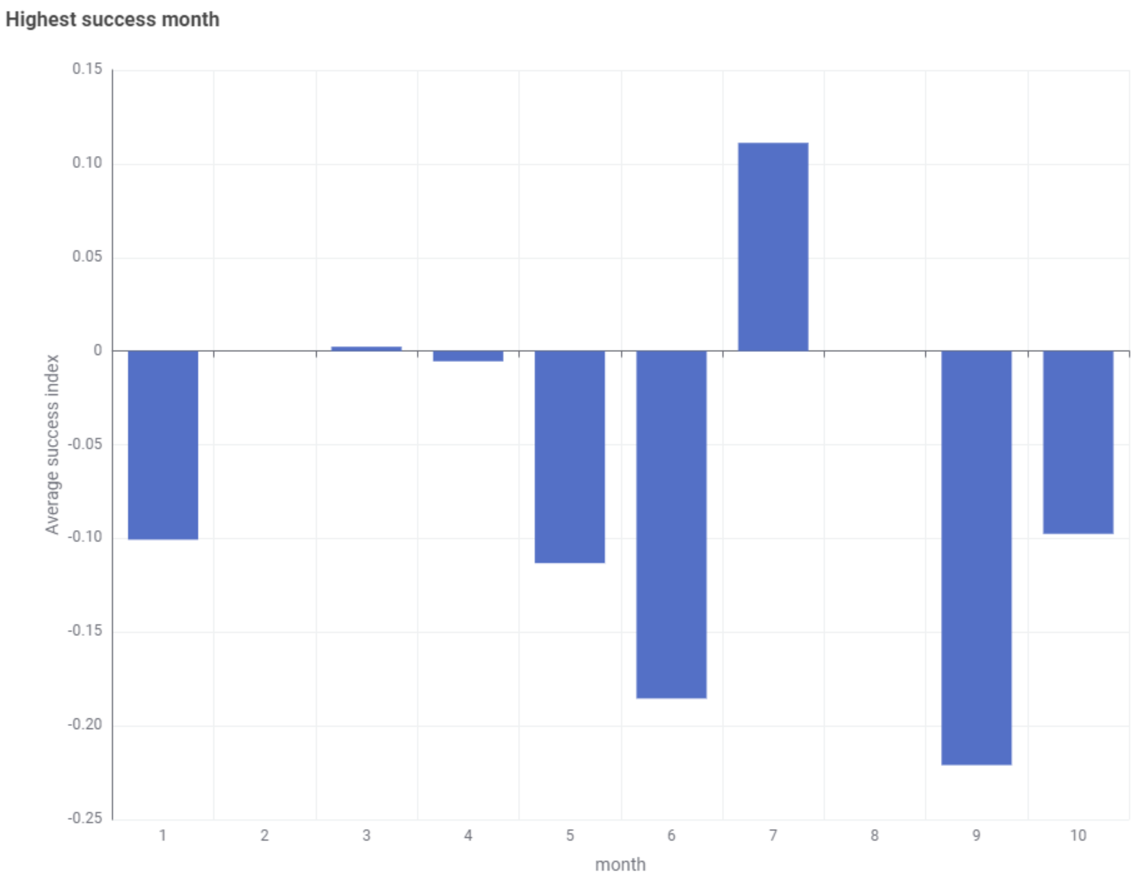


Figure 21

Attribute: day_of_week		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	5	
Mode	4	
10 most common values	1. 4 (561; 21.28%) 2. 3 (552; 20.94%) 3. 1 (540; 20.94%) 4. 5 (519; 19.69%) 5. 2 (464; 17.6%)	

Immediately from Figure 22, it can be seen that most of the last contacts occur on Thursday, closely followed by Wednesday. Tuesday has the lowest number of last contacts and is significantly lower than the other days. Interestingly, no last contacts are made on Saturday or Sunday as seen from the missing values of 6 and 7 from Figure 22.

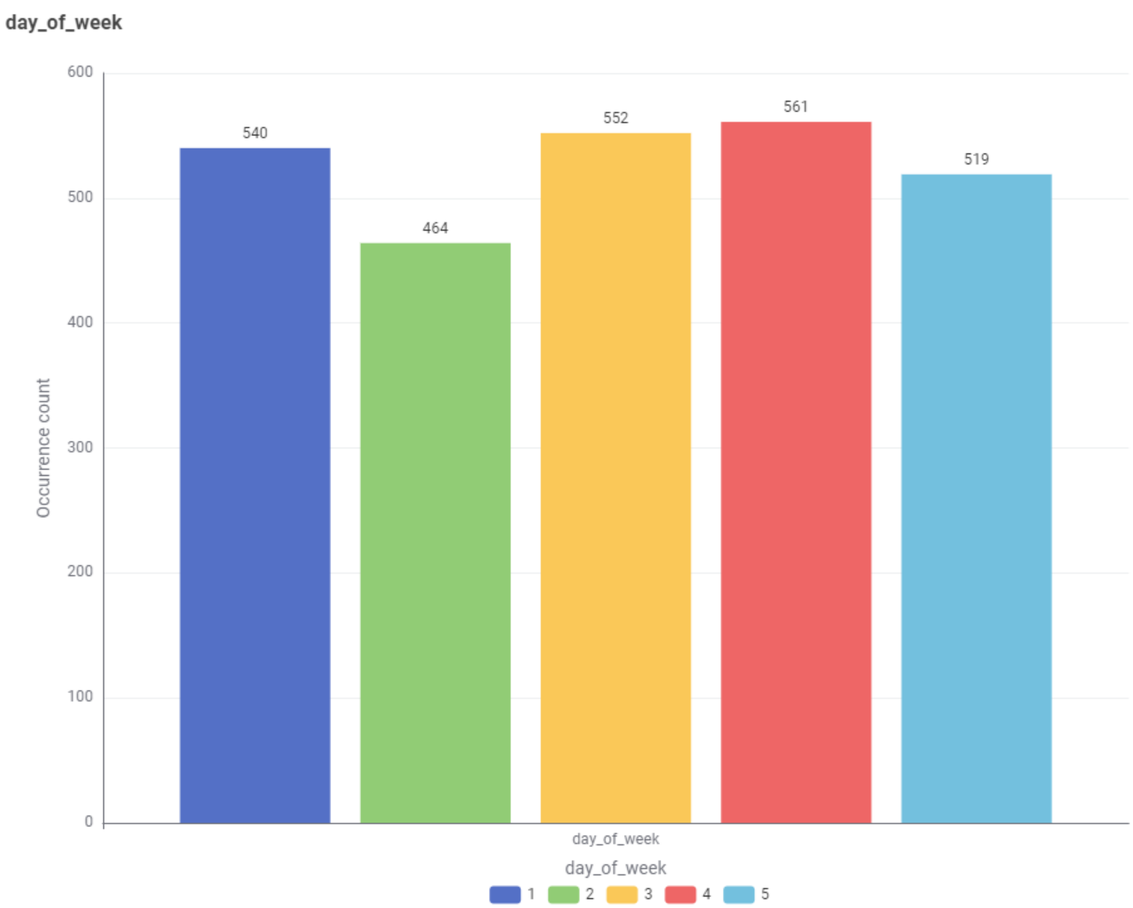


Figure 22

Additionally, Figure 23 shows the average success index for each day of the week. Evidently, Friday has the lowest average success index out of all the days in the week, approximately twice as bad as the 2 best days which are Tuesday and Wednesday, both sharing the same average success index. This indicates that when contacting a potential customer, it is best to avoid contacting on Friday and instead it is advised to contact on either Tuesday or Wednesday.

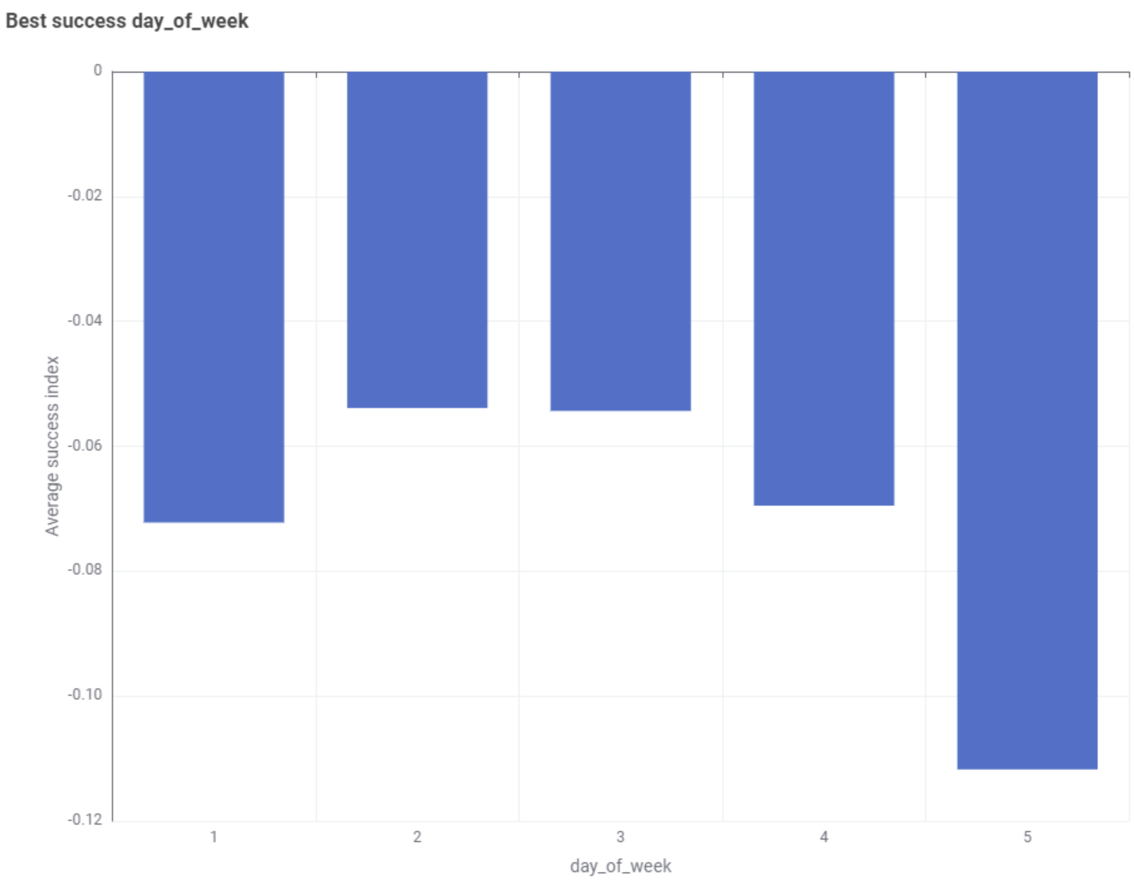


Figure 23

Attribute: duration		
Statistics	**Value (with missing values)**	**Value (without missing values)**
Minimum	1	
Maximum	4199	
Range	4198	
25% Quantile	101	
50% Quantile (Median)	177.5	
75% Quantile	318	
Mean	262.197	
Mode	135	
Mean Absolute Deviation	179.599	
Standard Deviation	299.263	
Variance	89558.343	
# Unique values	707	

The distribution of duration, measured in seconds, is displayed below in Figure 24. The longest contact duration recorded lasts up to 70 minutes whereas the shortest duration lasts only 1 second. Moreover, there is a significant number of outliers in the attribute as they have exceeded the threshold of $Q3 + 1.5 * IQR$.

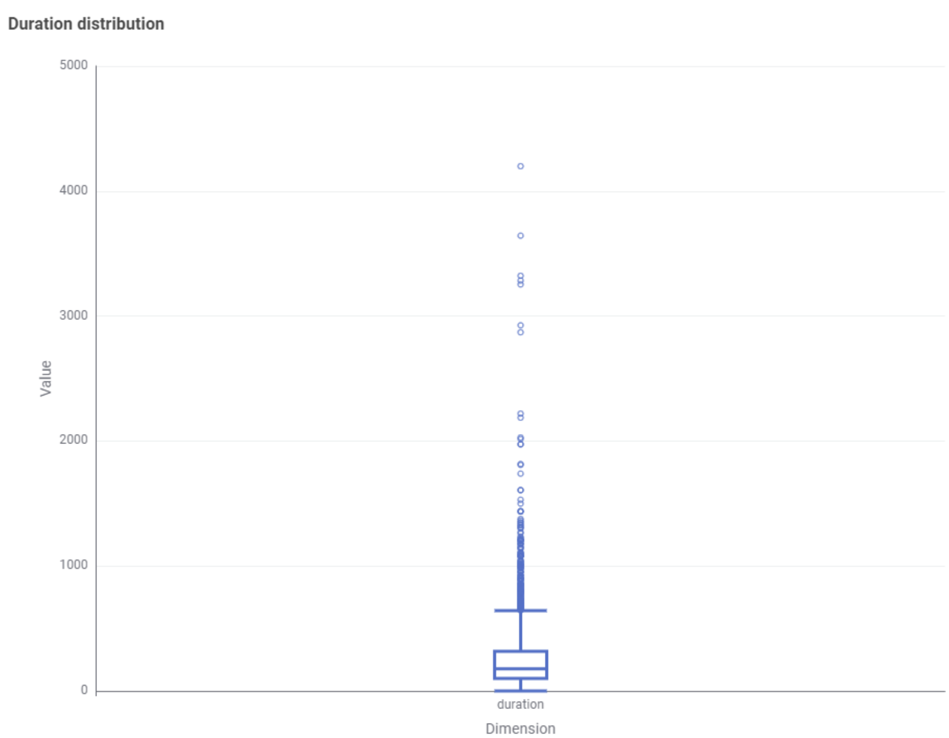


Figure 24

With such a wide range of duration values, the number of bins was ultimately determined using Sturges's rule shown below:

$$no.\,bins = 1 + 3.3 * \log(n) = 1 + 3.3 * \log(2636)$$

$$no.\,bins \approx 13$$

But with a range close to 4200 divided into 13 bins, that would result in each bin representing a 5 to 6 minutes difference which I found to be too wide to extract any useful findings. Additionally, since most outliers contribute to less than 0.01% of the values, I decided to restrict the range to make binning convey more useful information.

Figure 25 shows the proportion of duration after being binned using the equal-width method with $width = 120$, i.e. each bin represents a 2-minute difference. After binning, it is discovered that the majority of the contact durations (about 65%) only last up to 4 minutes.

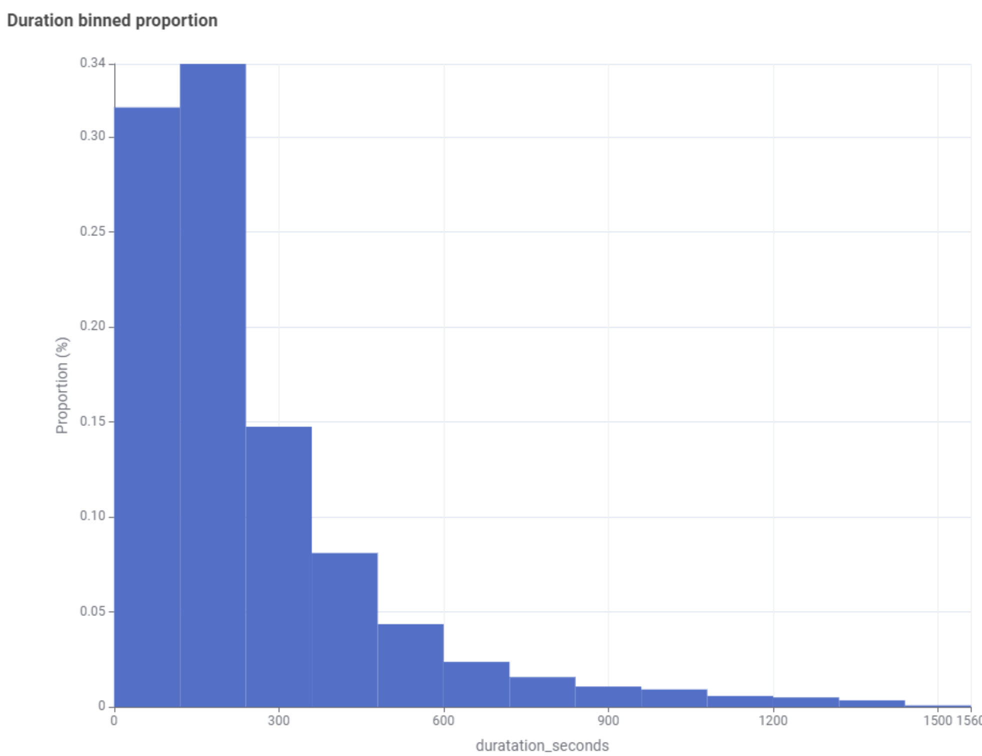


Figure 25

Attribute: campaign		
Statistics	**Value (with missing values)**	**Value (without missing values)**
Minimum	1	
Maximum	35	
Range	34	
25% Quantile	1	
50% Quantile (Median)	2	
75% Quantile	3	
Mean	2.552	
Mode	1	
Mean Absolute Deviation	1.635	
Standard Deviation	2.742	
Variance	7.519	
# Unique values	29	
Sum	6727	

The distribution of the campaign attribute is shown in Figure 26. The maximum and minimum number of contacts performed during the campaign recorded are 35 and 1 respectively. There are a total of 6727 contacts made during the campaign. Also, it can be seen that there are many outliers above the $Q3 + 1.5 * IQR$ threshold. Additionally, Figure 27 shows the proportion of the number of contacts made during the campaign where 92% of the time, less than 5 contacts will be made.

Figure 26

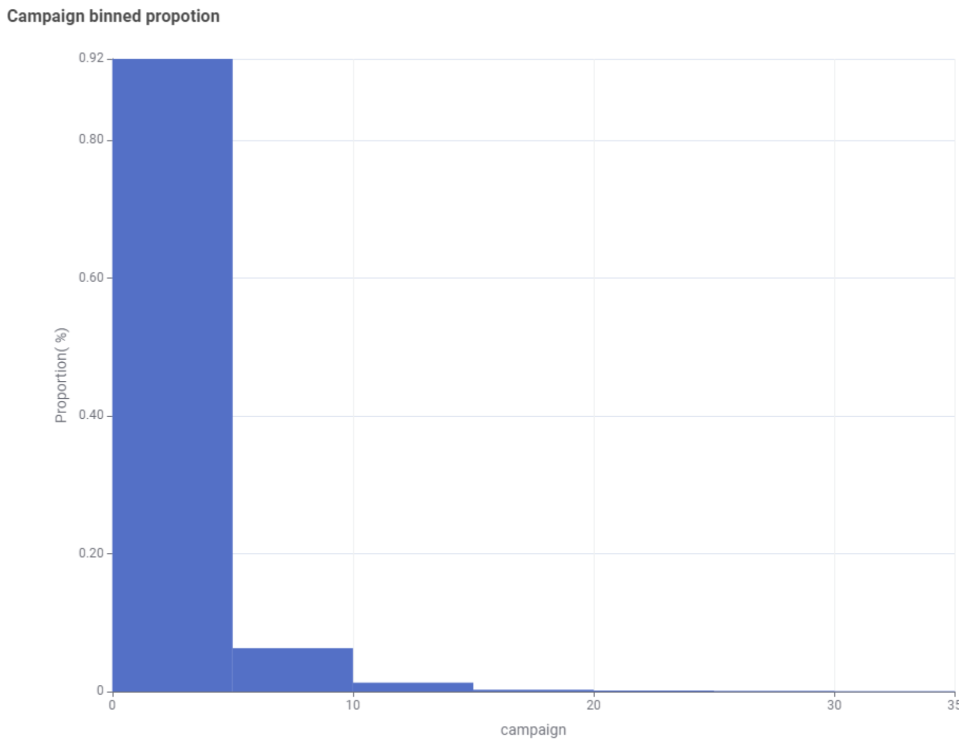


Figure 27

Attribute: passed days		
Statistics	**Value (with missing values)**	**Value (without missing values)**
Minimum	0	
Maximum	999	
Range	999	
25% Quantile	999	
50% Quantile (Median)	999	
75% Quantile	999	
Mean	962.851	
Mode	999	
Mean Absolute Deviation	69.665	
Standard Deviation	185.979	
Variance	34588.19	
# Unique values	18	

The distribution of the passed_days attribute can be seen below in Figure 28. An overwhelming majority of the observations (96%) have a value of 999, hence why the mode, Q1, Q2, and Q3 quantiles are all the same. The maximum and minimum number of passed_days recorded are 999 and 0 respectively. It can also be observed that there are some outliers below the $Q1 - 1.5 * IQR$ threshold.

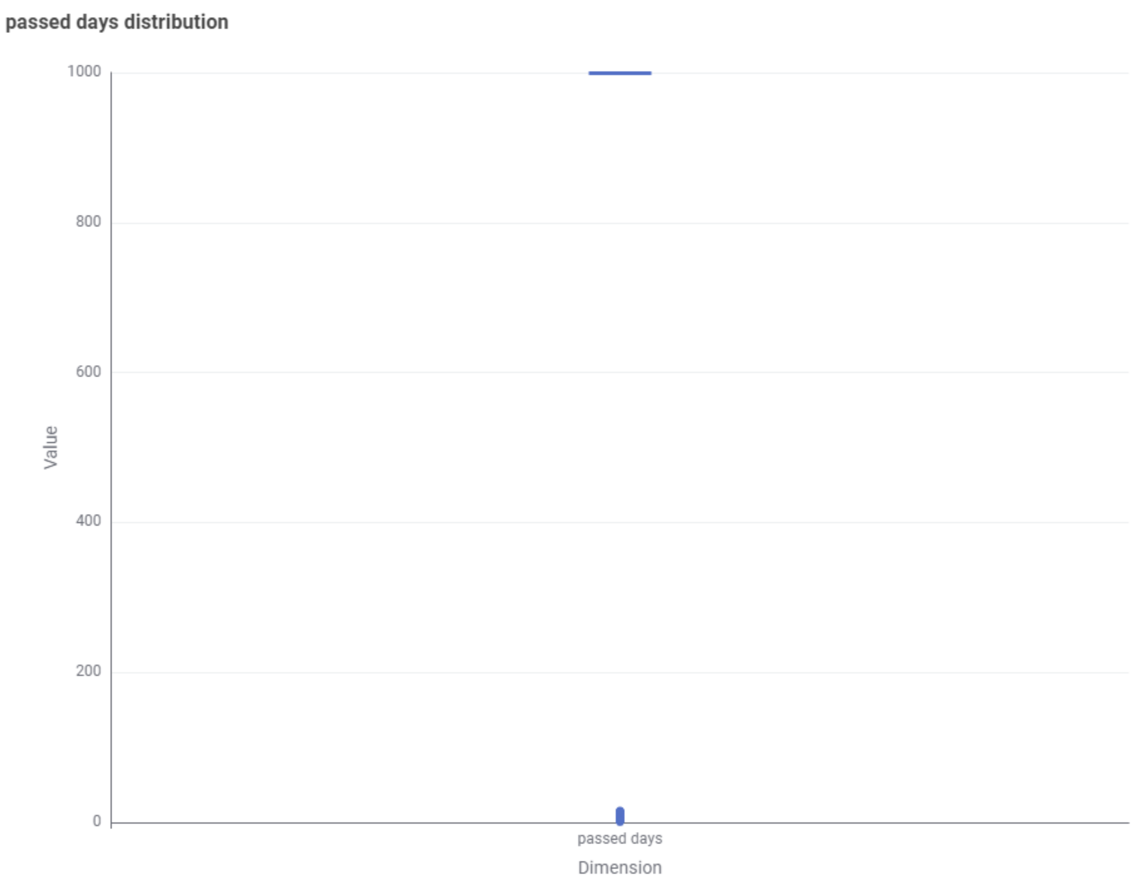


Figure 28

Attribute: previous		
Statistics	**Value (with missing values)**	**Value (without missing values)**
Minimum	0	
Maximum	4	
Range	4	
25% Quantile	0	
50% Quantile (Median)	0	
75% Quantile	0	
Mean	0.173	

Mode	0	
Mean Absolute Deviation	0.298	
Standard Deviation	0.481	
Variance	0.231	
# Unique values	5	
Sum	455	
10 most common values	1. 0 (2273; 86.23%) 2. 1 (293; 11.12%) 3. 2 (51; 1.93%) 4. 3 (16; 0.61%) 5. 4 (3; 0.11%)	

The distribution of the previous attribute can be seen below in Figure 29. Similar to the passed_days attribute, an overwhelming majority of the observations (86%) have a value of 0, hence why the mode, Q1, Q2, and Q3 quantiles are all the same. The maximum and minimum number of passed_days recorded are 4 and 0 respectively. It can also be observed that there are some outliers below the $Q3 + 1.5 * IQR$ threshold.

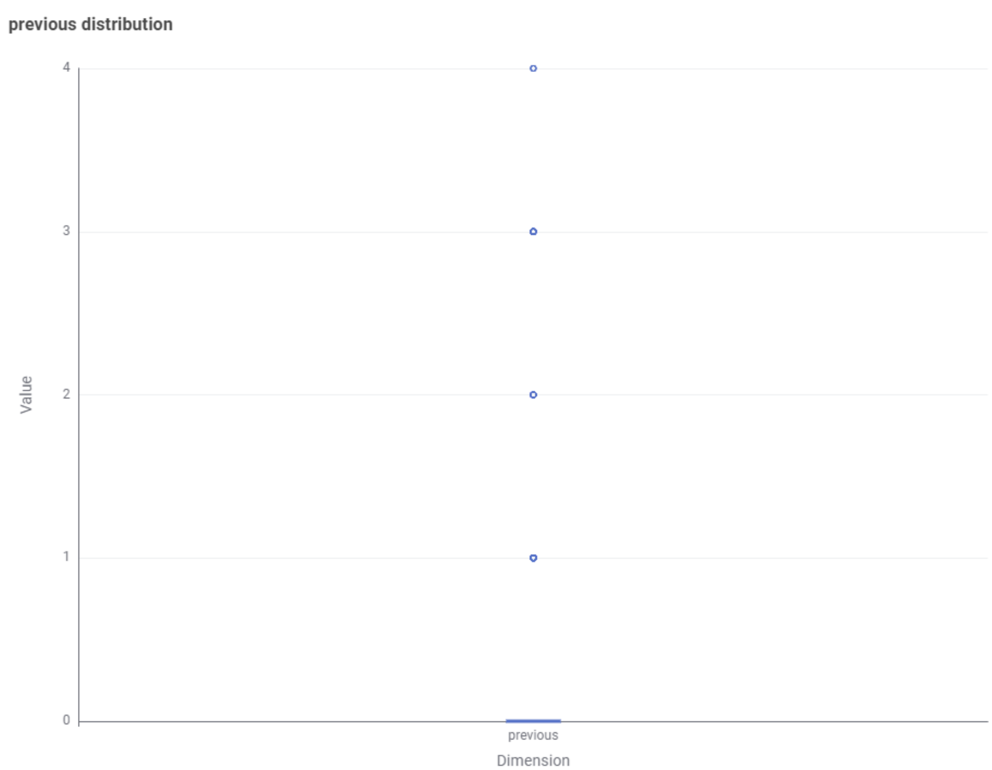


Figure 29

Based on the collected data, Figure 30 shows the ideal number of contacts to be made before the marketing campaign. Evidently, 3 contacts are the most effective amount with a considerable average success index of close to 0.40 when compared to other values. Additionally, it can be observed that more contacts being made previously does not translate to more success as seen when $previous = 4$. Therefore, any amount more than 3 is just a diminishing return. Finally, the average success index is the lowest when $previous = 1$, so we should avoid contacting only 1 time.

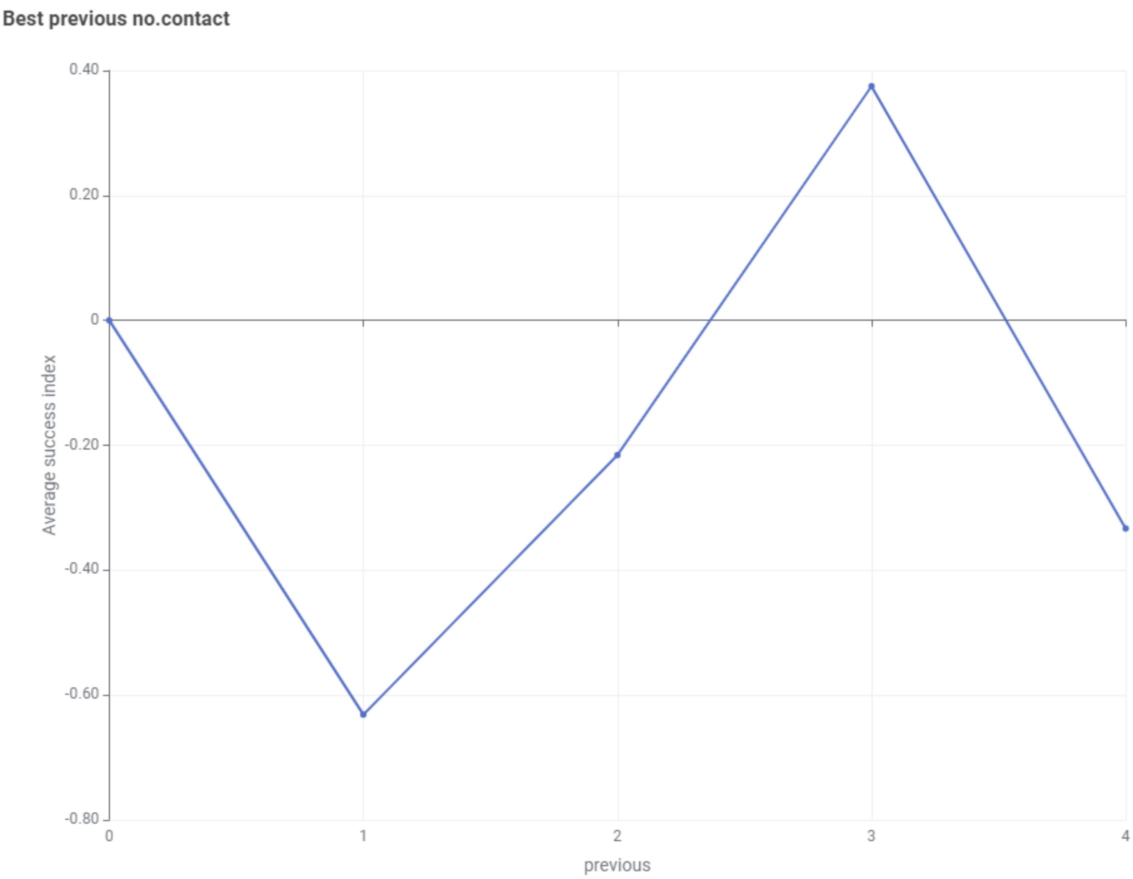


Figure 30

Attribute: poutcome		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	3	
Mode	nonexistent	
10 most common values	1. nonexistent (2273; 86.23%) 2. failure (277; 10.51%) 3. success (86; 3.26%)	

Figure 31 shows the proportion of the outcomes for the previous marketing campaign. Overall, it can be seen that the previous marketing campaign's effectiveness was mostly nonexistent with more than 80% of the observations indicating so. The next common previous outcome is "failure" followed by "success". The previous marketing campaign was very disappointing, not only due to the significant portion of the "nonexistent" outcome but also because the number of failures is triple that of the number of successes.

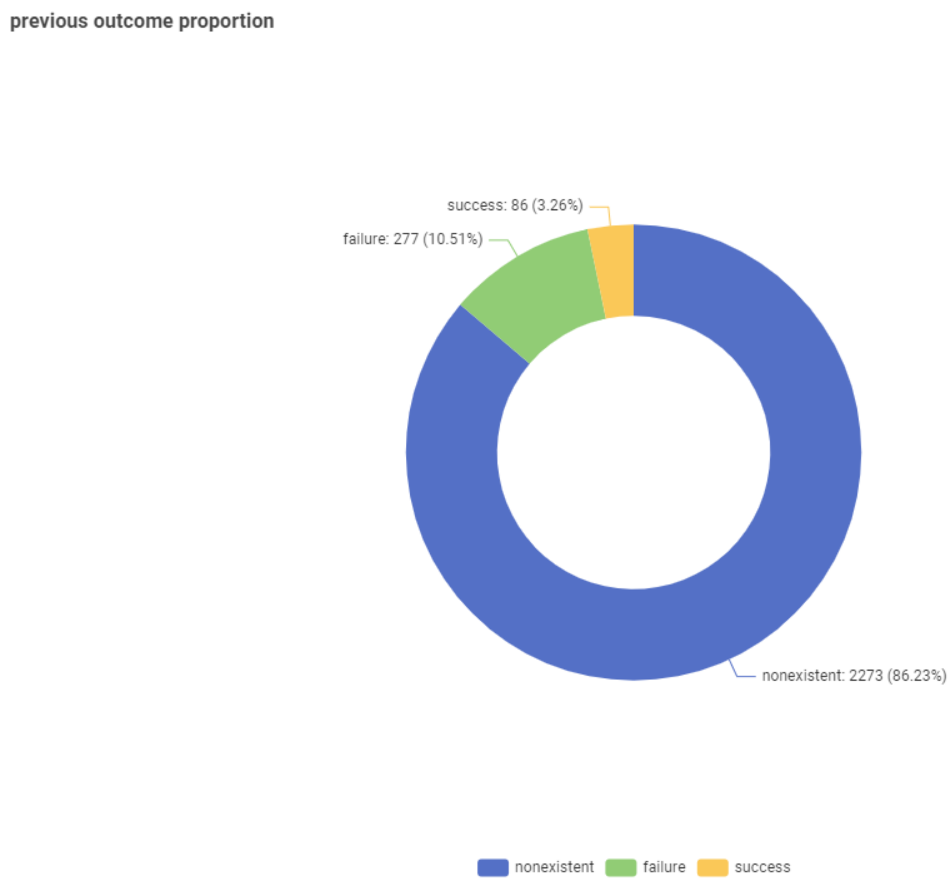


Figure 31

Similar to the attribute "education", different outcomes will be assigned a corresponding numeric value to extract a meaningful evaluation metric called the success index. The numeric values assigned to each outcome are shown in the table below.

Previous outcome	Assigned value
Failure	-1
Nonexistent	0
Success	1

Figure 32 shows the result after grouping and calculating the mean success index of each state. Out of all the states, the Southern Australia, Australian Capital Territory, and Northern Territory states have the highest average success index. Conversely, the state with the lowest average success index is Tasmania followed closely by the state of Western Australia.

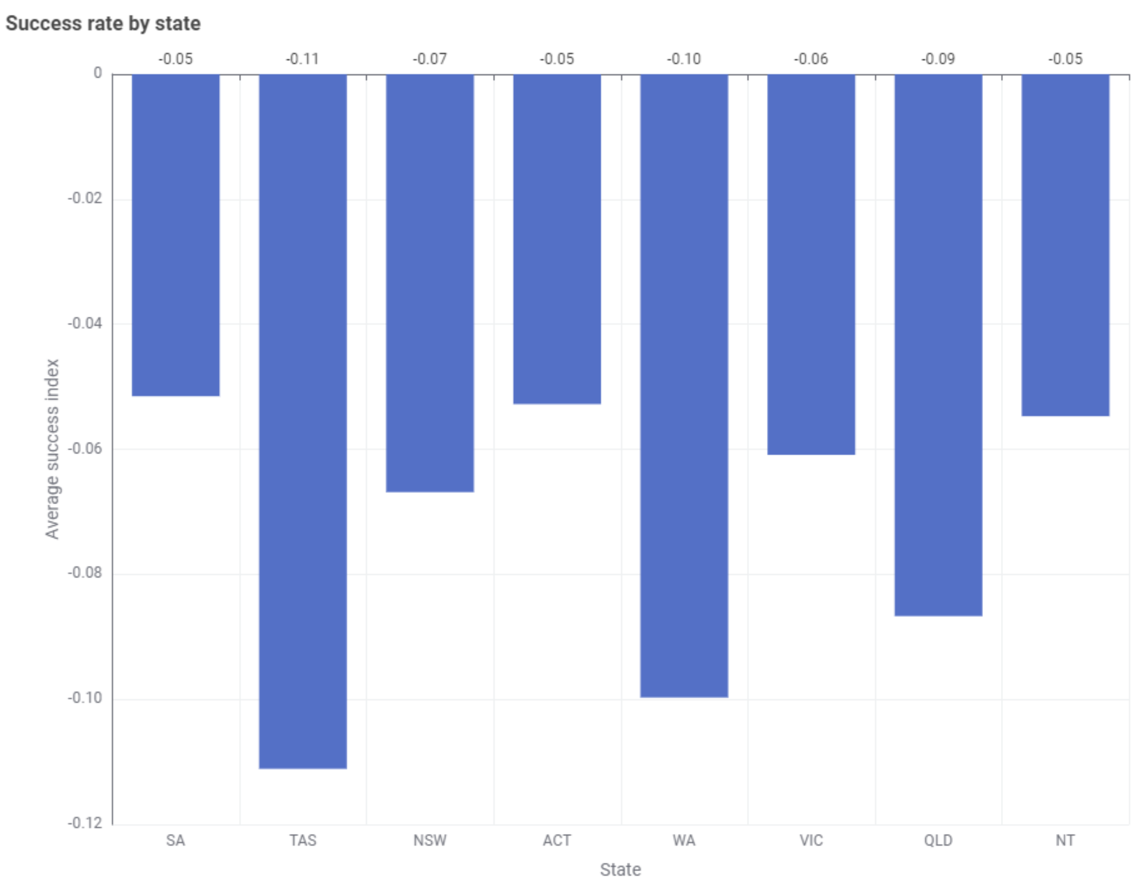


Figure 32

Attribute: variation rate		
Statistics	**Value (with missing values)**	**Value (without missing values)**
Minimum	-3.4	
Maximum	1.4	
Range	4.8	
25% Quantile	-1.8	
50% Quantile (Median)	1.1	
75% Quantile	1.4	
Mean	0.055	
Mode	1.4	
Mean Absolute Deviation	1.427	
Standard Deviation	1.58	
Variance	2.50	
# Unique values	9	

The distribution of the variation rate is shown below in Figure 33. The distribution is very left-skewed, however, there are no outliers spotted in the distribution. Additionally, the mean is very close to 0 indicating that on average, there is little to no change in the number of employees for each quarter.

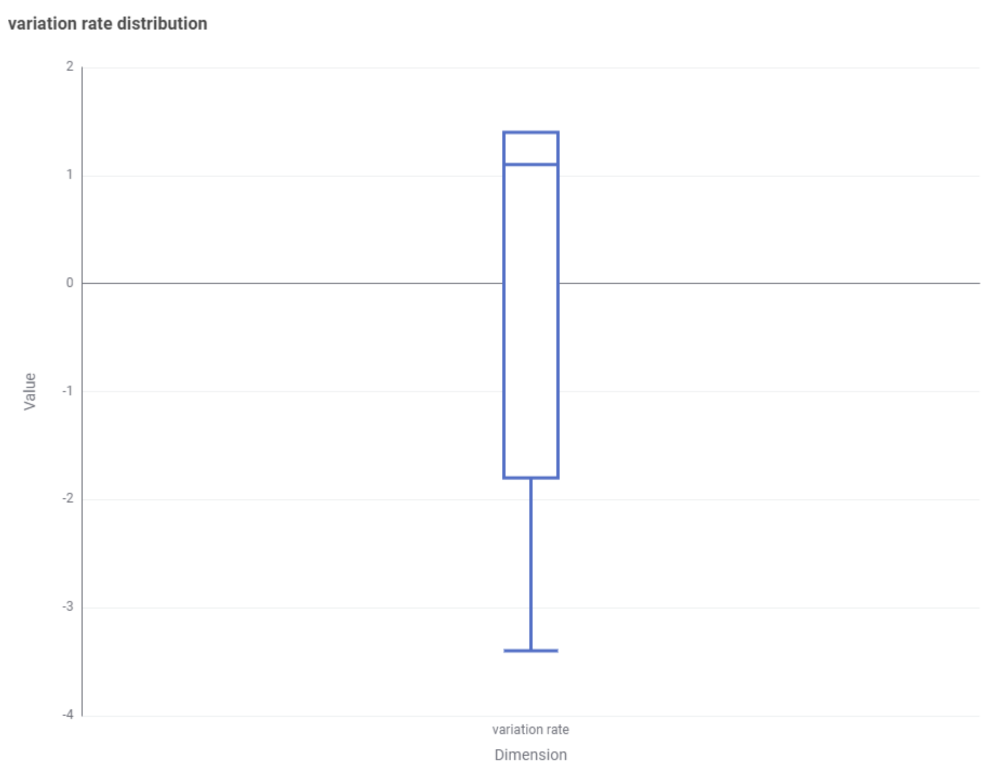


Figure 33

Most importantly, variation rate potentially has many interesting relationships with the attributes: price index, euribor3m, and Confidence Index that require further analysis. The supporting evidence can be seen from the matrices of Pearson's correlation coefficient and Spearman's rank correlation coefficient shown in Figure 34 and Figure 35 respectively. Pearson's correlation coefficient describes the strength and direction of a linear relationship between two attributes. On the other hand, Spearman's rank correlation coefficient describes how well can two attributes be modeled using a monotonic function.

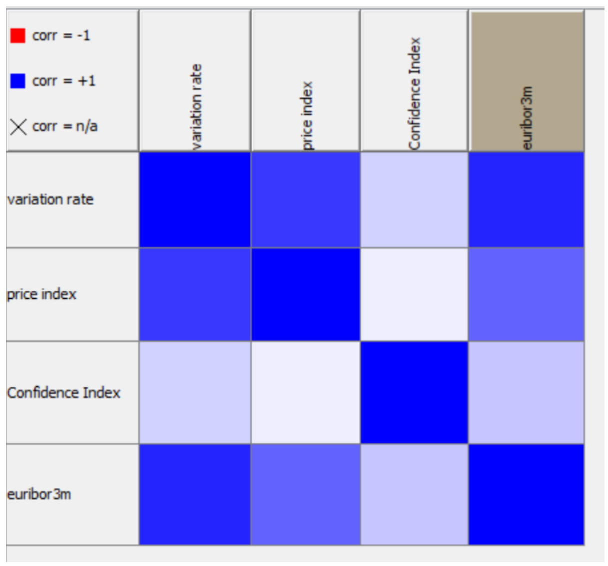


Figure 34. Pearson's correlation coefficient

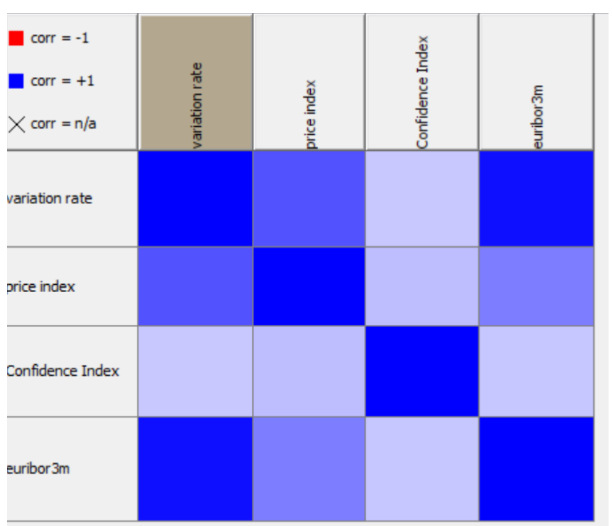


Figure 35. Spearman's rank correlation coefficient

As seen in Figure 36, there is a strong positive linear relationship between variation rate and price index with $pearson\ corr = 0.78$ and $spearman\ corr = 0.68$. Similarly in Figure 37, there is even a stronger positive linear relationship between variation rate and euribor3m with $pearson\ corr = 0.86$ and $spearman\ corr = 0.94$. Finally, while the relationship between variation rate and Confidence Index is non-linear as seen in Figure 38, a pattern can still be seen and should be analyzed in more detail.

This indicates that variation rate is a very good predictor of price index and euribor3m. However, if the objective is to predict the outcome of a marketing campaign, then including both variation rate and price index or euribor3m when training machine learning models might introduce the problem of multicollinearity.

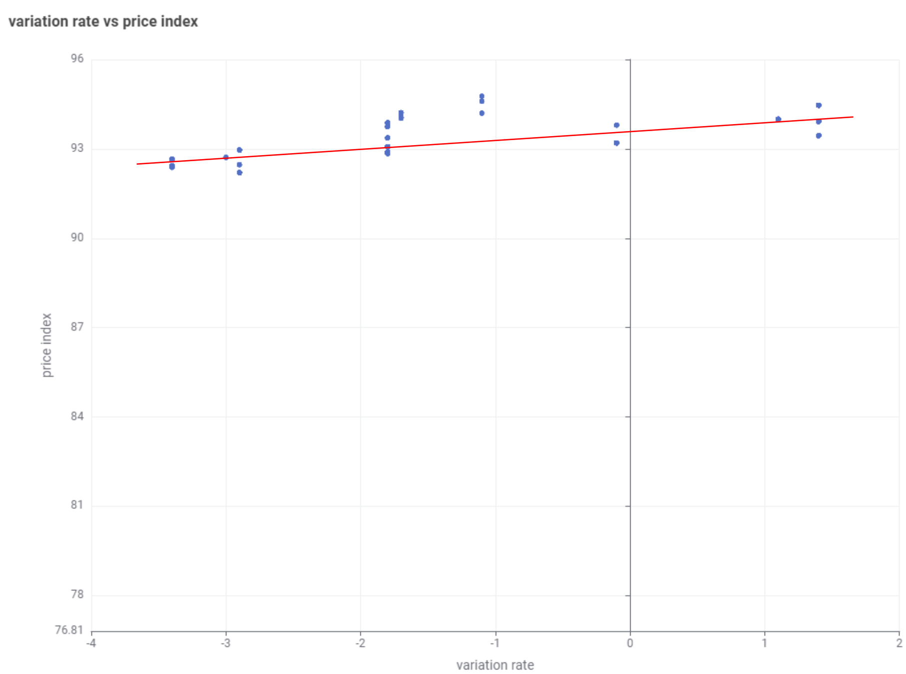


Figure 36

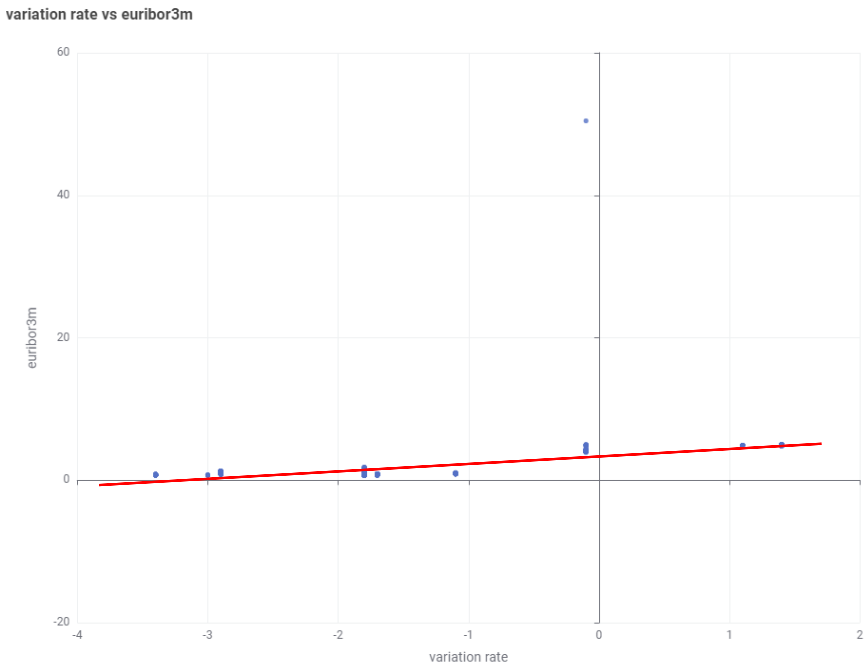


Figure 37

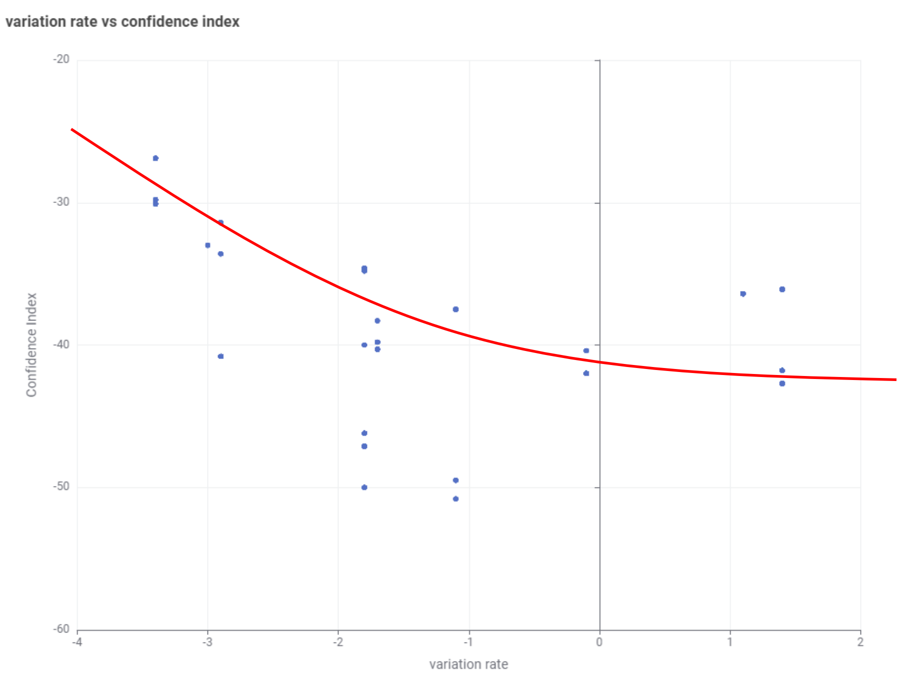


Figure 38

Attribute: price index		
Statistics	**Value (with missing values)**	**Value (without missing values)**
Minimum	92.201	
Maximum	94.767	
Range	2.566	
25% Quantile	93.075	
50% Quantile (Median)	93.596	
75% Quantile	93.994	
Mean	93.571	
Mode	93.994	
Mean Absolute Deviation	0.507	
Standard Deviation	0.575	
Variance	0.33	
# Unique values	25	

The distribution of the price index is shown below in Figure 39. The distribution is slightly left skewed but not significant and there are no outliers observed. Figure 40 shows the trend of the price index over the months. It seems like the price index will continuously drop starting from February to July and will start to stabilize in August.

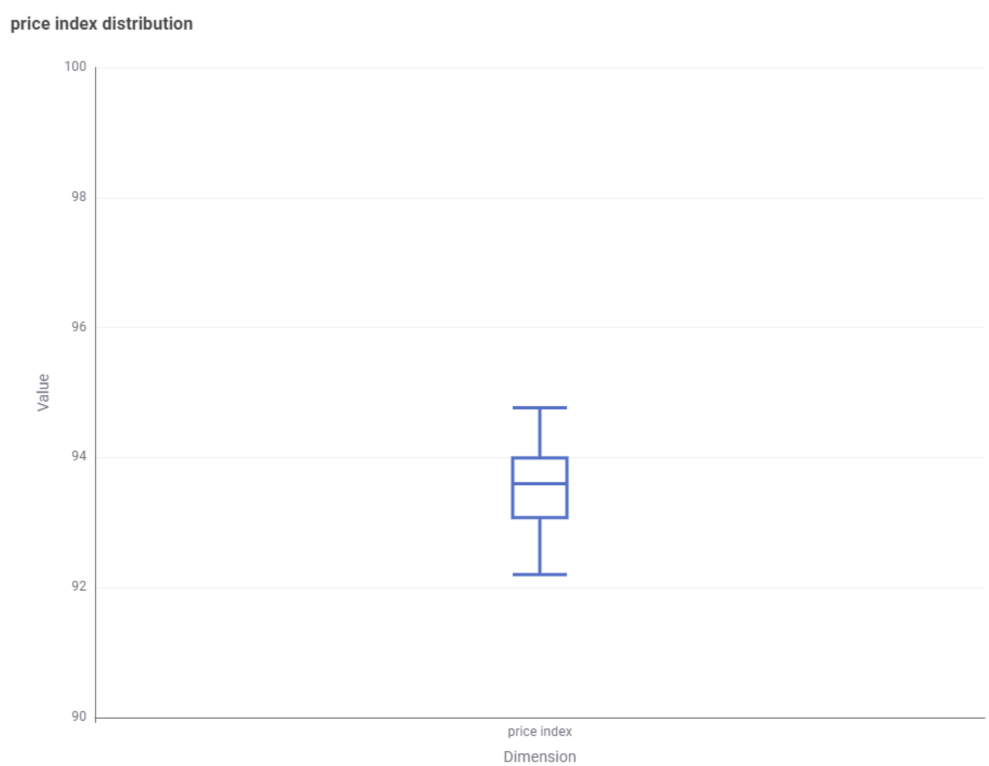


Figure 39

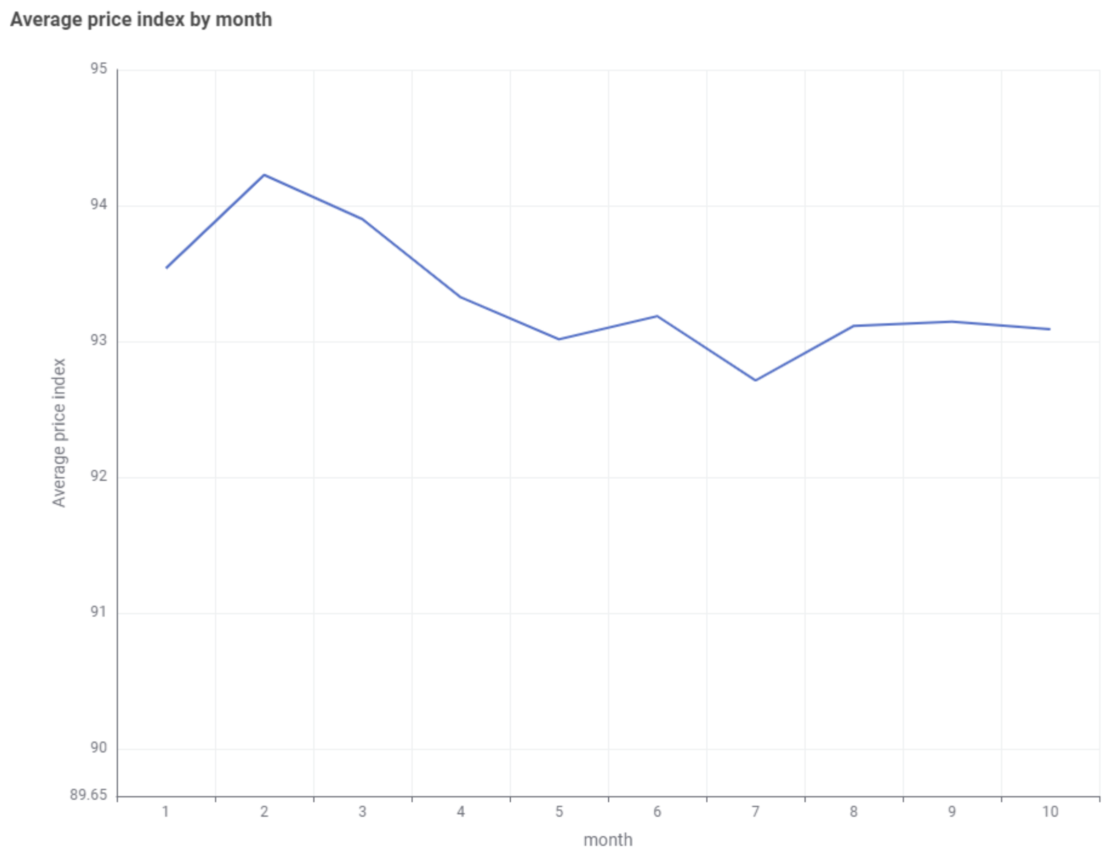


Figure 40

Attribute: Confidence Index		
Statistics	**Value (with missing values)**	**Value (without missing values)**
Minimum	-50.8	
Maximum	-26.9	
Range	23.9	
25% Quantile	-42.7	
50% Quantile (Median)	-41.8	
75% Quantile	-36.4	
Mean	-40.428	
Mode	-36.4	
Mean Absolute Deviation	3.928	
Standard Deviation	4.637	
Variance	21.50	
# Unique values	25	

Figure 41 shows the distribution of the Confidence Index which is extremely right-skewed with one point potentially being an outlier as it is slightly above the $Q3 + 1.5 * IQR$ threshold. The majority of the records reported that they are very pessimistic about the future financial outlook evidently from Figure 42. Furthermore, it is very concerning that there is not a singular instance indicating that they are even slightly optimistic. The binning technique used in Figure 42 is equal-width with $width = 20$. The table below gives the boundary of each category.

Confidence Outlook	**Boundary**
Very pessimistic	$(-\infty, -41)$
Moderately pessimistic	$[-41, -21)$
Slightly pessimistic	$[-21, 1)$
Optimistic	$[1, \infty)$

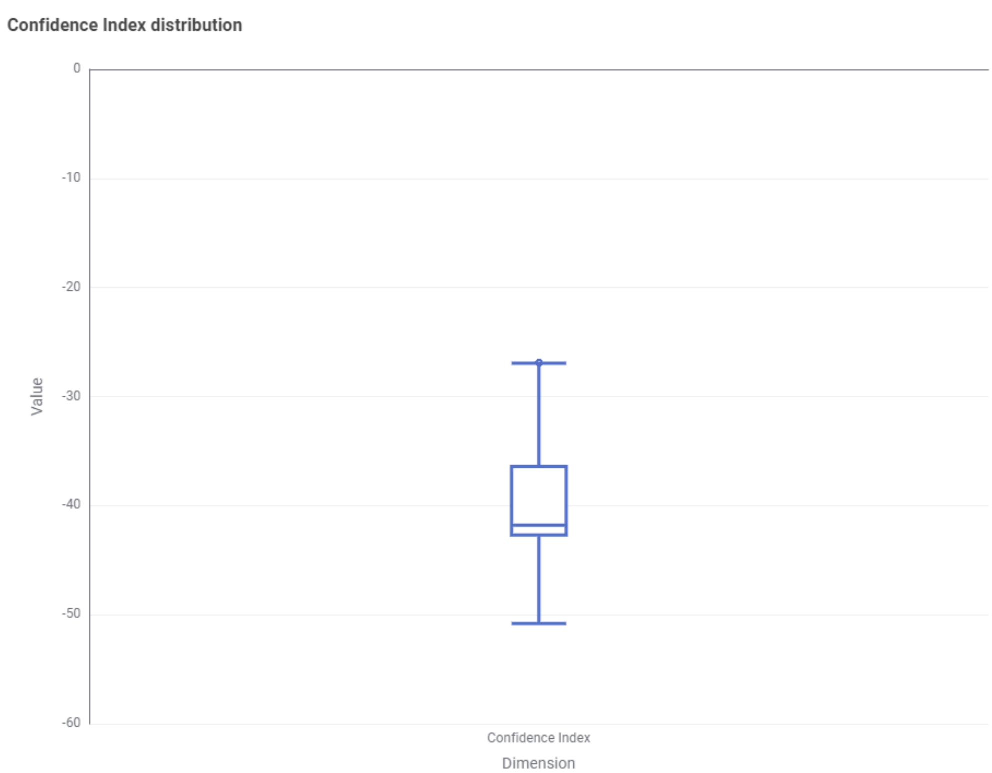


Figure 41

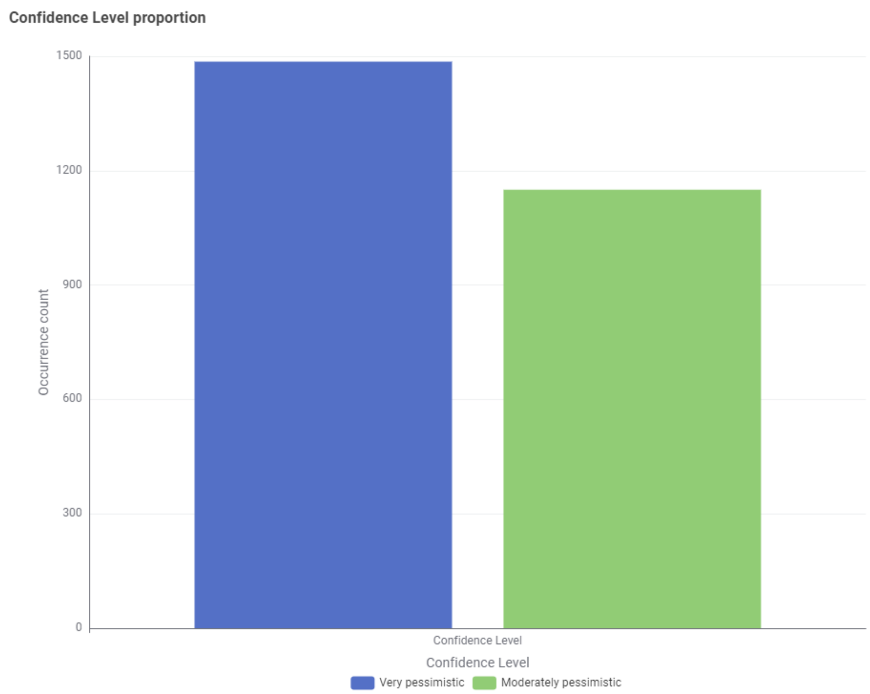


Figure 42

Initially, I wanted to test the hypothesis that the Confidence Index influences the outcome of the marketing campaign. Intuitively this makes sense since if the financial future outlook does not look good, customers would prefer to save money rather than spend it which makes marketing campaigns more unlikely to succeed. Surprisingly, there is no concrete evidence supporting the hypothesis since based on Figure 43, all the outcomes occur within roughly identical range. **The same happens when plotting poutcome against euribor3m and the price index. This concludes that the overall failure of the previous marketing was not due to economic factors.**

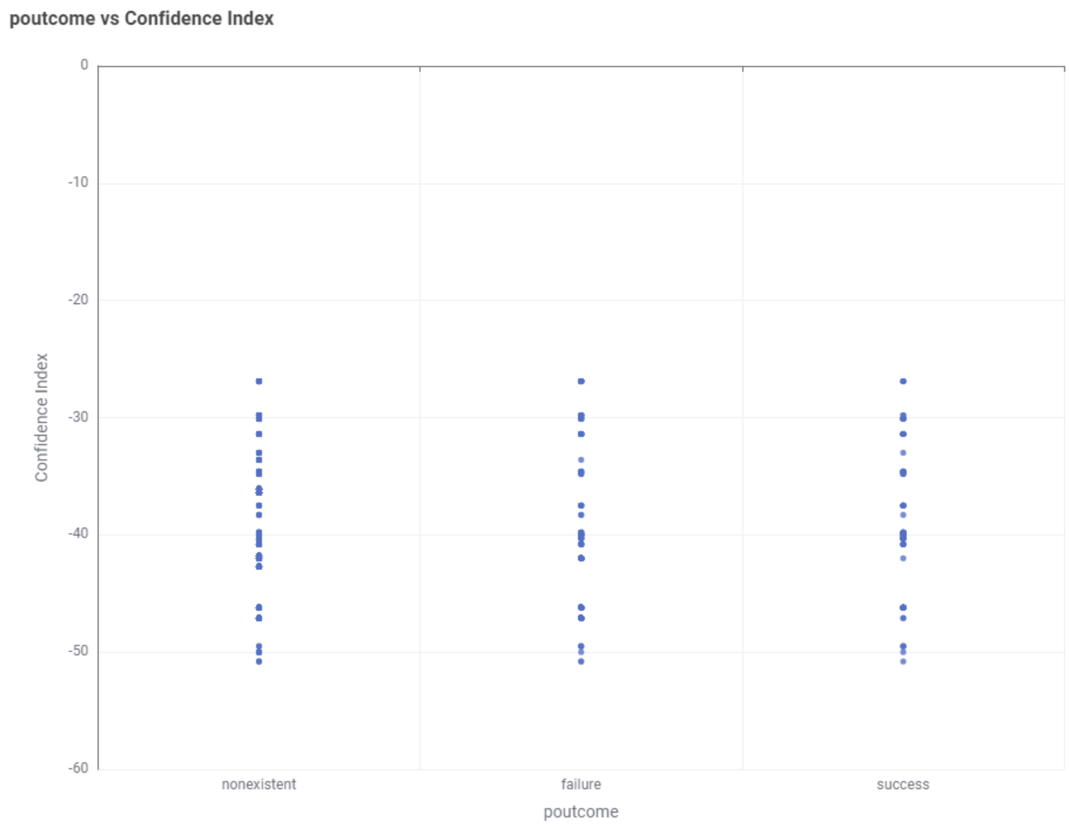


Figure 43

Attribute: euribor3m		
Statistics	**Value (with missing values)**	**Value (without missing values)**
Minimum	0.634	
Maximum	50.45	
Range	49.82	
25% Quantile	1.344	
50% Quantile (Median)	4.857	
75% Quantile	4.961	
Mean	3.622	
Mean Absolute Deviation	1.626	
Standard Deviation	1.964	
Variance	3.86	
# Unique values	199	

The distribution of euribor3m can be seen in Figure 44 and it is extremely left skewed since the median has the same value as the 75% quantile. There is a very significant outlier that is extremely far above the $Q3 + 1.5 * IQR$ threshold. Similar to the variation rate attribute, the attribute euribor3m potentially has a relationship with the price index as seen from the coefficient matrices in Figure 34 and Figure 35 above.

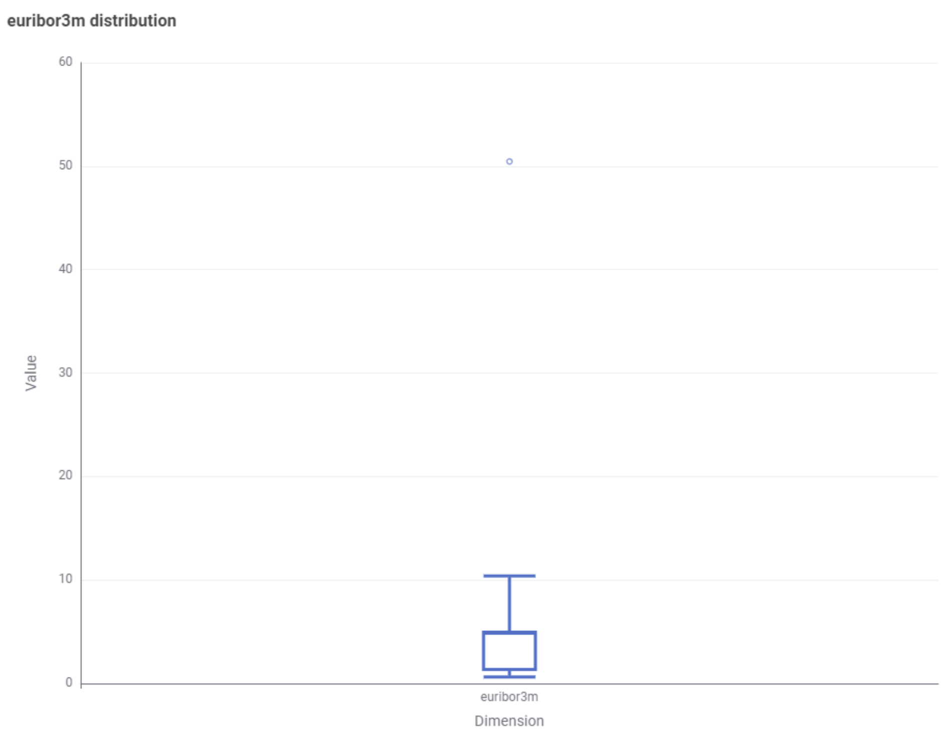


Figure 44

Based on Pearson's correlation coefficient matrix from Figure 34, euribor3m has a moderate positive linear relationship with the price index ($pearson\ corr = 0.62$). The estimated linear relationship is drawn below in Figure 45. Once again, while euribor3m is a decent predictor for the prince index, if the objective is to predict the marketing campaign outcome, then further analysis should be conducted to avoid the problem of multicollinearity.

Alternatively, the relationship between euribor3m and price index can also be potentially described by a non-decreasing monotonic function. This is evident as shown in Figure 46 below and the Spearman's rank correlation coefficient in Figure 35 above ($spearman\ corr = 0.51$).

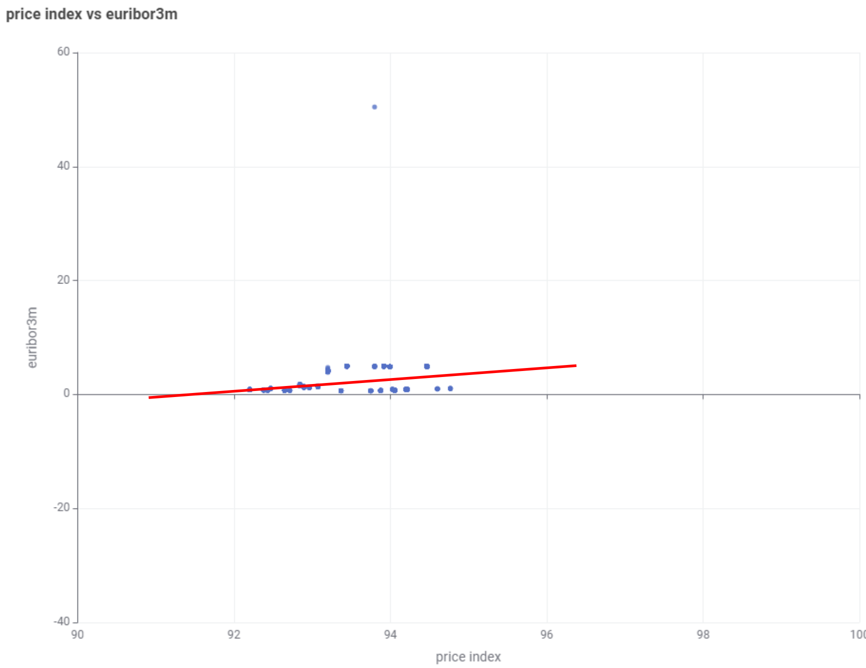


Figure 45

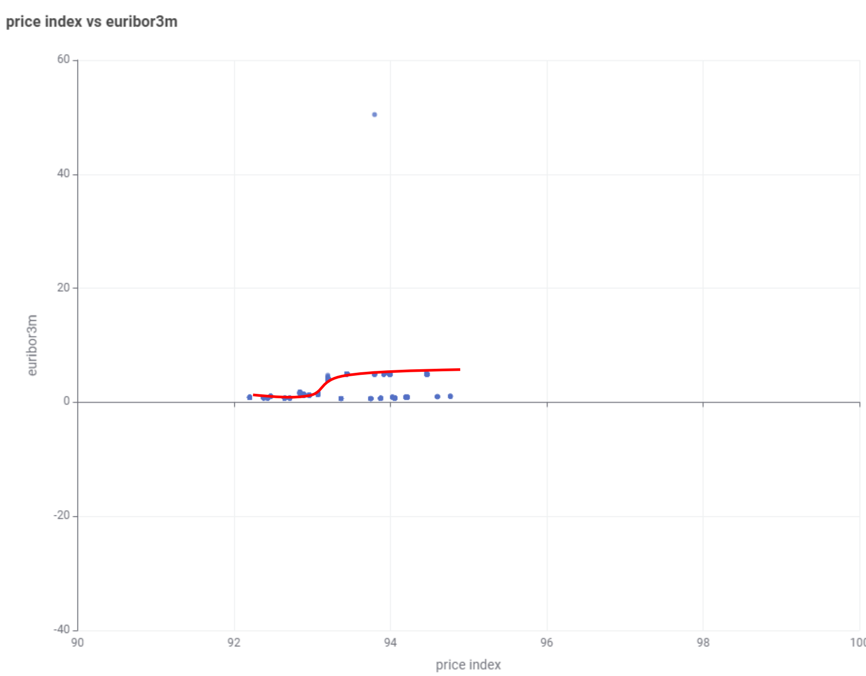


Figure 46

Attribute: No.employed		
Statistics	**Value (with missing values)**	**Value (without missing values)**
Minimum	4963.6	
Maximum	5228.1	
Range	264.5	
25% Quantile	5008.7	5017.5
50% Quantile (Median)	5076.2	5090.15
75% Quantile	5191	5191
Mean	5090.15	
Mean Absolute Deviation	81.977	75.446
Standard Deviation	90.565	86.881
Variance	8202.02	7548.31
# Unique values	11	12

The distribution of No.employed is relatively symmetric (refer to Figure 47), but there are some missing values that have been observed indicated with the symbol “?”. There are a total of 210 missing value records, which is approximately 8% of all the values. Since the proportion of missing values can be considered to be significant, they will be imputed with the mean.

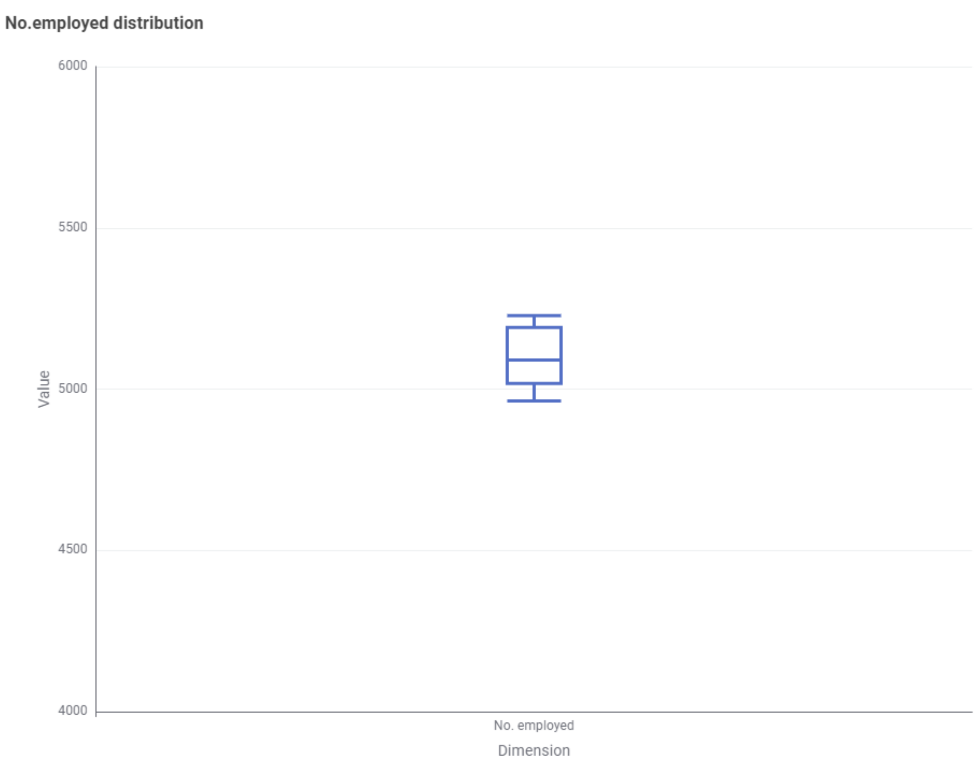


Figure 47

Attribute: Term Deposit		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	2	
Mode	0	
10 most common values	1. 0 (2343; 88.88%) 2. 1 (293; 11.12%)	

The pie chart below (refer to Figure 48) indicates the proportion of the binary values (0 and 1) for the attribute Term Deposit. By utilizing the logic that "Term Deposit" is a binary attribute, then "0" refers to customers who did not subscribe to a term deposit and vice versa. Evidently from the graph, the majority of the observations (approximately 90%) have reported that they did not subscribe to a term deposit.

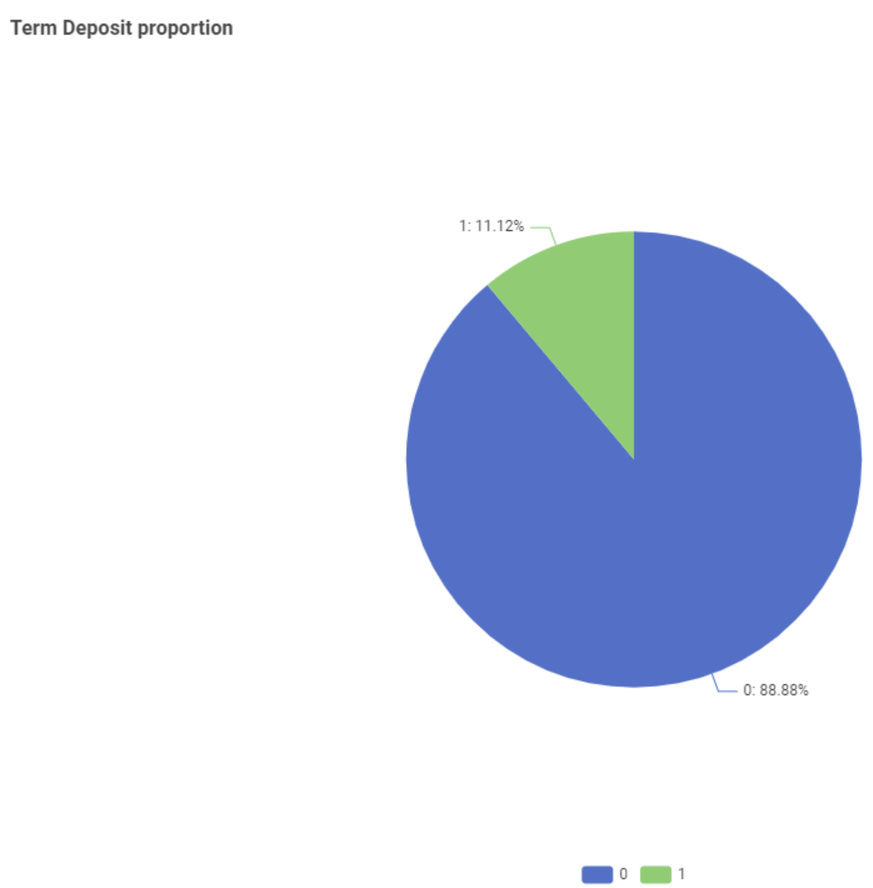


Figure 48

Attribute: State		
Statistics	**Value (with missing values)**	**Value (without missing values)**
# Unique values	8	
Mode	ACT	
10 most common values	1. ACT (360; 13.66%) 2. NSW (344; 13.05%) 3. WA (341; 12.94%) 4. SA (330; 12.52%) 5. NT (329; 12.48%) 6. QLD (323; 12.25%) 7. VIC (312; 11.84%) 8. TAS (297; 11.27%)	

The choropleth map below (refer to Figure 49) illustrates the proportion of observations belonging to each of the eight states listed in the table above. Most of the observations come from the state ACT, closely followed by the states NSW and WA respectively. On the contrary, the state TAS has the least number of records.

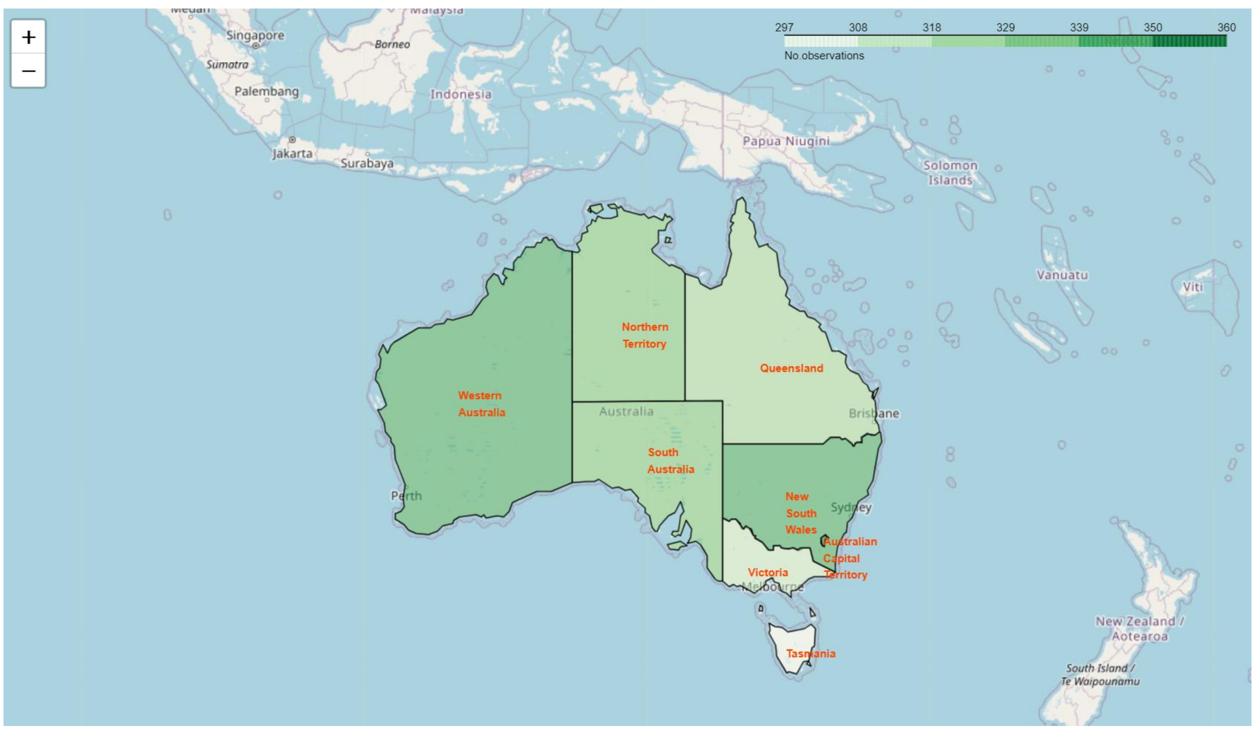


Figure 49

Figure 50 details the most popular contact method of each region. Interestingly, the state QLD mostly uses cellphones as their main contact method and does not use mailing as much in comparison. In a similar fashion, the state SA's main contact method is also the cellphone, and contacting via telephone or email is not popular amongst the state's residents. The state NSW on the other hand, mainly uses email whereas the states ACT, NT and WA are well-rounded overall.

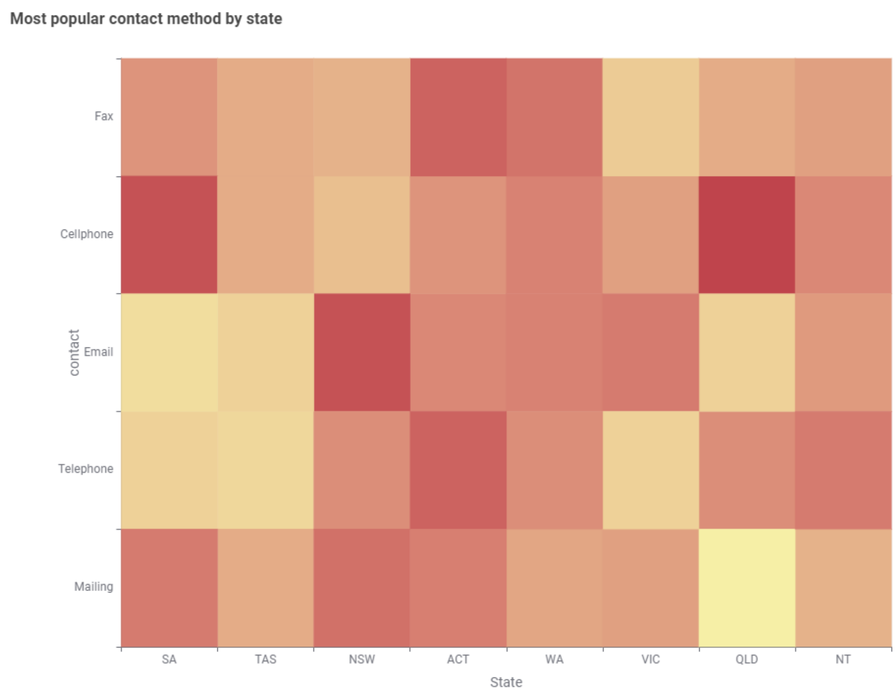


Figure 50

As mentioned in the "contact" attribute section above, the most popular contact method is not necessarily the most effective method. Furthermore, while the most effective contact method overall is email, the most effective contact method of each distinct state can be different as shown in Figure 51 below. The ribbon chart in Figure 51 shows the highest and lowest average success rates of different contact methods of each state. The states TAS, WA, and NT have the most success when contacted via email. Conversely, the states VIC and ACT have the most success when contacted via fax. Finally, the most successful contact method for the states of SA, NSW, and QLD is via mailing, telephone, and cellphone respectively.

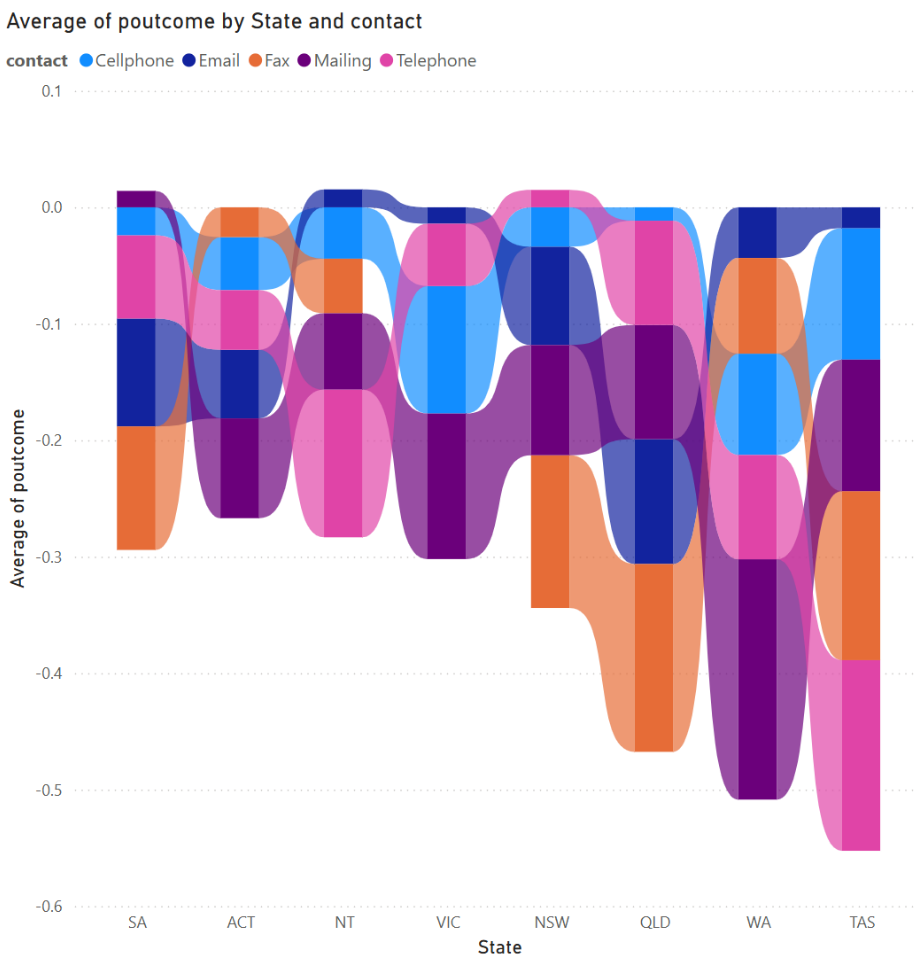


Figure 51

B. Data Preprocessing

B1. Binning techniques

- **Equi-width binning**

Equi-width is a binning technique where data points are categorized into bins of equal width. Examples of bins of equal width are: $(0, 10]$, $(10, 20]$, $(20, 30]$, etc. For the examples provided, the width of each bin is 10 since the first bin ranges from 0 (exclusive) to 10 (inclusive) and so on.

When binning the “age” attribute, the data points are classified into 6 distinct bins ($no.bins = 6$) to enforce the width of each bin to be 10 ($width = 10$) as shown in Figure 52, Figure 53, and Figure 54. The justification for this is because a decade is often used to measure the growth of a person. A person with age within the range [1, 11) is considered to be a child. However, a person within the age range [11, 21) will be considered as an adolescent. Additionally, an age too specific does not hold much significance which is why we often refer to age in a range of 10. For example, people often refer to each other’s age in a range such as “in the early-20s”, “in the mid-40s”, or “in the late-80s”.

Figure 52

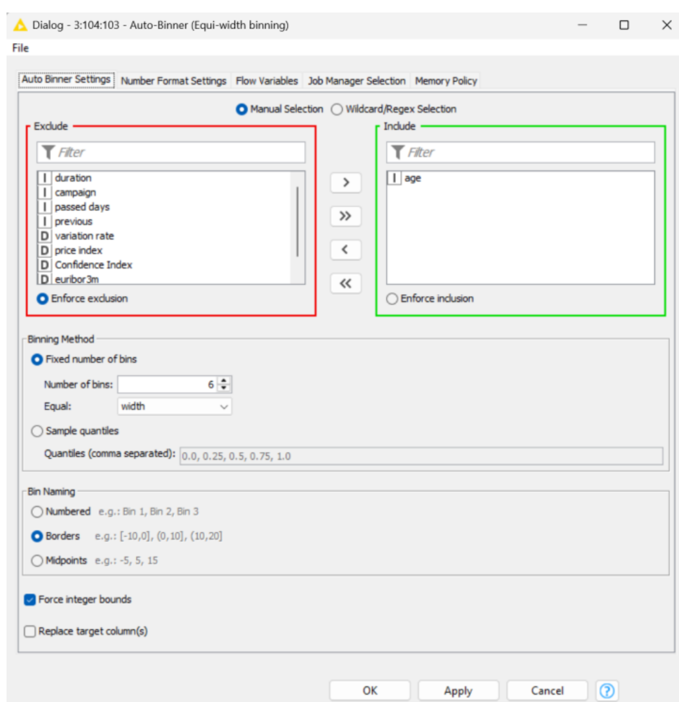
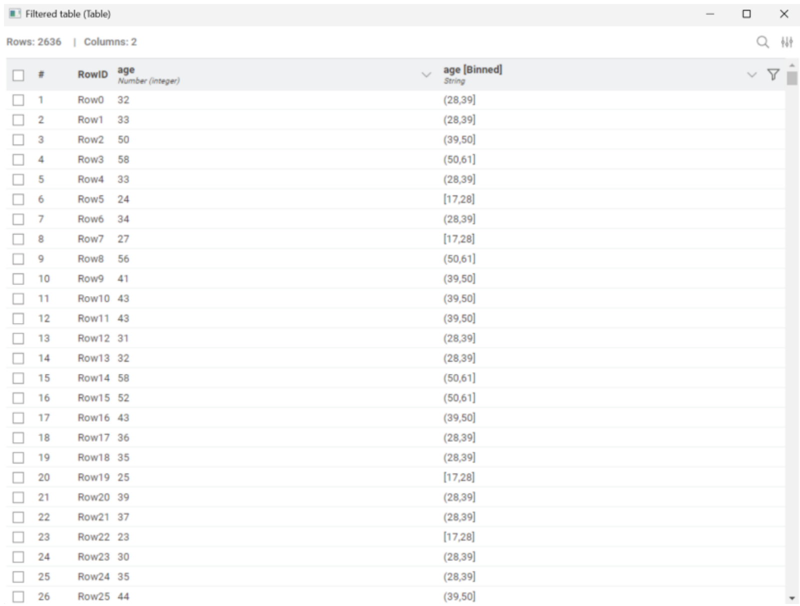


Figure 53



Filtered table (Table)

Rows: 2636 | Columns: 2

#	RowID	age (Number (integer))	age [Binned] (String)
1	Row0	32	(28,39]
2	Row1	33	(28,39]
3	Row2	50	(39,50]
4	Row3	58	(50,61]
5	Row4	33	(28,39]
6	Row5	24	[17,28]
7	Row6	34	(28,39]
8	Row7	27	[17,28]
9	Row8	56	(50,61]
10	Row9	41	(39,50]
11	Row10	43	(39,50]
12	Row11	43	(39,50]
13	Row12	31	(28,39]
14	Row13	32	(28,39]
15	Row14	58	(50,61]
16	Row15	52	(50,61]
17	Row16	43	(39,50]
18	Row17	36	(28,39]
19	Row18	35	(28,39]
20	Row19	25	[17,28]
21	Row20	39	(28,39]
22	Row21	37	(28,39]
23	Row22	23	[17,28]
24	Row23	30	(28,39]
25	Row24	35	(28,39]
26	Row25	44	(39,50]

Figure 54

- ## Equi-depth binning

Equi-depth is a binning technique where data points are categorized into bins of equal frequency. While it is not always possible to guarantee that every bin has an equal number of data points, the goal is to make them as close as possible to being equally distributed and still able to extract meaningful insights.

When binning the "age" attribute, the data points are classified into 7 distinct bins ($no.bins = 7$) as shown in Figure 55, Figure 56, and Figure 57. The justification for this is because $no.bins = 7$, gives the smoothest distribution (refer to Figure 58) while still retaining enough details and not too generalized when compared to other bins.

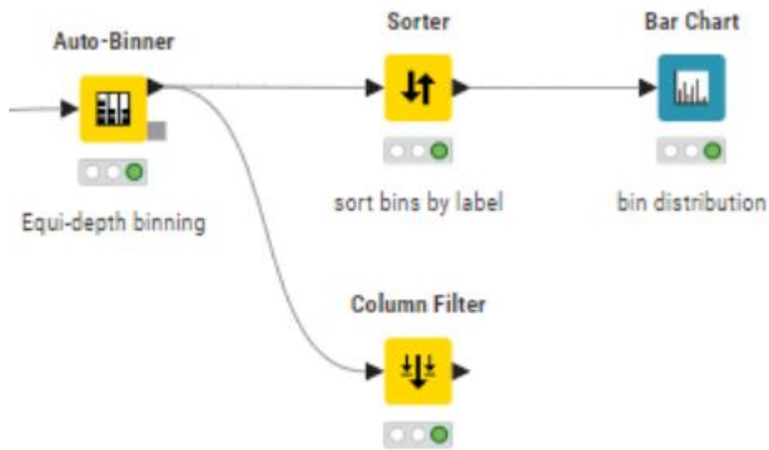


Figure 55

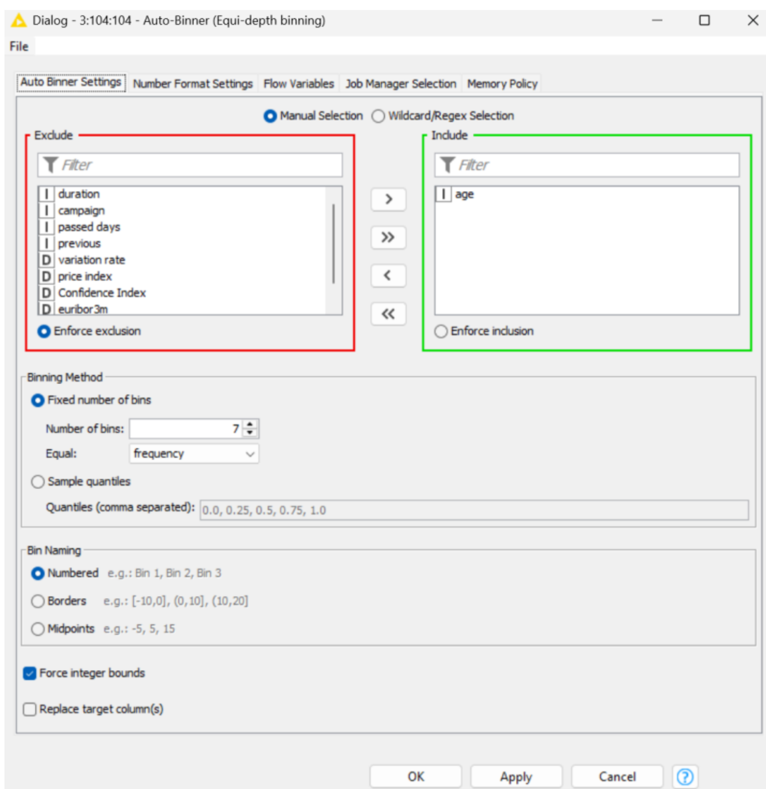


Figure 56

Filtered table (Table)

Rows: 2636 | Columns: 2

#	RowID	age (Number (integer))	age [Binned] (String)
1	Row0	32	Bin 2
2	Row1	33	Bin 3
3	Row2	50	Bin 6
4	Row3	58	Bin 7
5	Row4	33	Bin 3
6	Row5	24	Bin 1
7	Row6	34	Bin 3
8	Row7	27	Bin 1
9	Row8	56	Bin 7
10	Row9	41	Bin 5
11	Row10	43	Bin 5
12	Row11	43	Bin 5
13	Row12	31	Bin 2
14	Row13	32	Bin 2
15	Row14	58	Bin 7
16	Row15	52	Bin 7
17	Row16	43	Bin 5
18	Row17	36	Bin 4
19	Row18	35	Bin 3
20	Row19	25	Bin 1
21	Row20	39	Bin 4
22	Row21	37	Bin 4
23	Row22	23	Bin 1
24	Row23	30	Bin 2
25	Row24	35	Bin 3
26	Row25	44	Bin 5

Figure 57

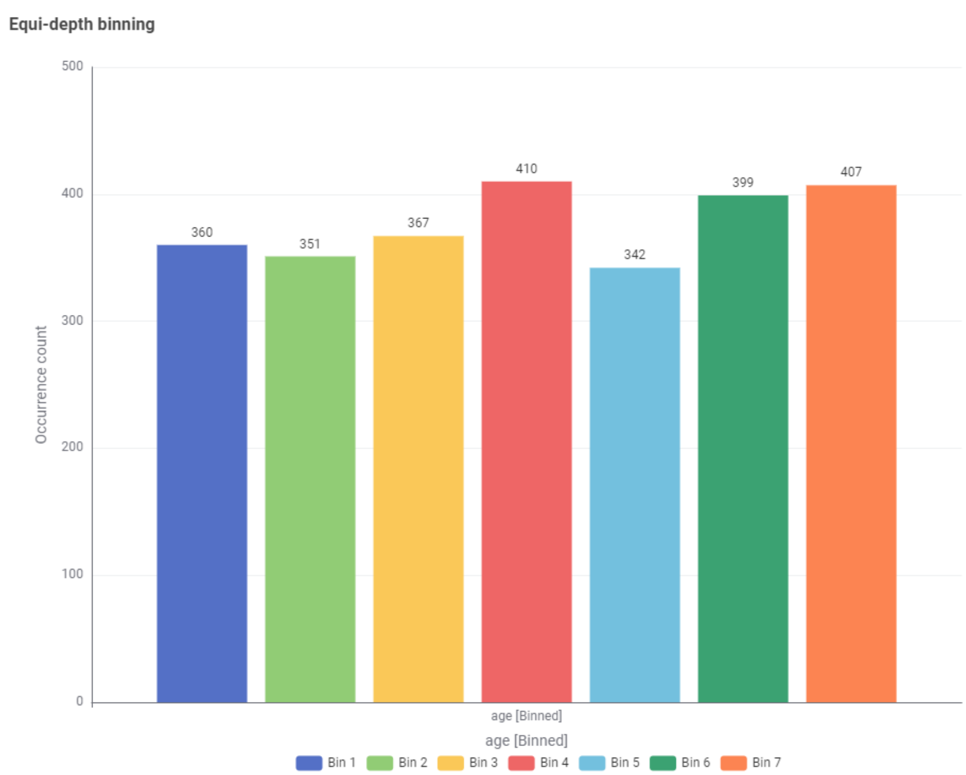


Figure 58

B2. Normalization

- **Min-Max normalization**

Min-Max is a normalization technique to limit the numeric values of an attribute onto the range [0, 1] for better interpretability and performance when training machine learning models. By using Min-Max normalization, it prevents the difference in the scale of values of different attributes from influencing the machine learning models as well as preventing the loss of information.

$$Min - Max: \quad x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

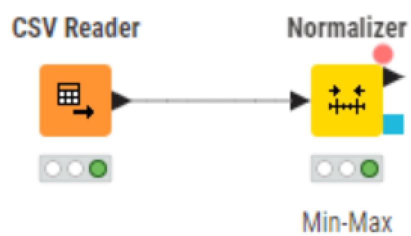


Figure 59

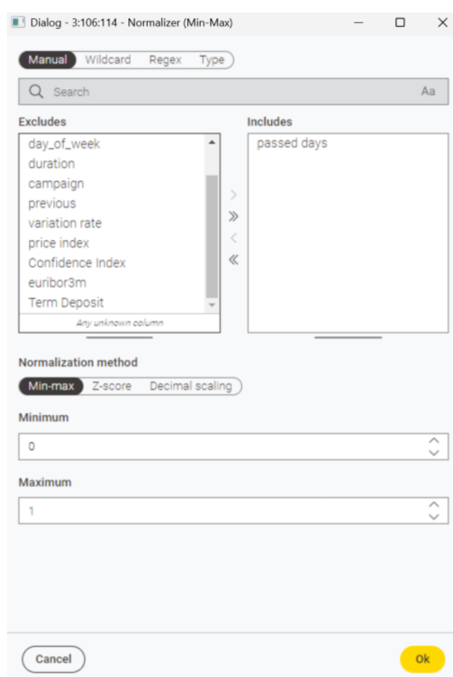
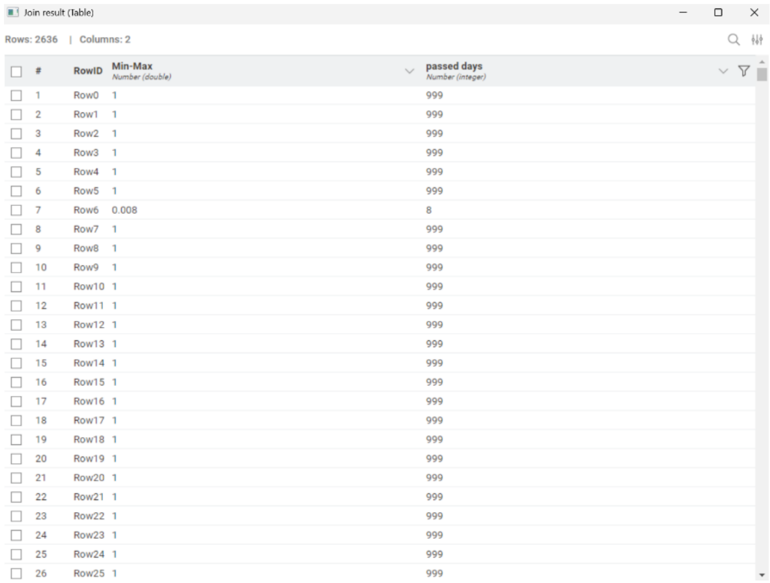


Figure 60



Join result (Table)

Rows: 2636 | Columns: 2

#	RowID	Min-Max Number (double)	passed days Number (integer)
1	Row0	1	999
2	Row1	1	999
3	Row2	1	999
4	Row3	1	999
5	Row4	1	999
6	Row5	1	999
7	Row6	0.008	8
8	Row7	1	999
9	Row8	1	999
10	Row9	1	999
11	Row10	1	999
12	Row11	1	999
13	Row12	1	999
14	Row13	1	999
15	Row14	1	999
16	Row15	1	999
17	Row16	1	999
18	Row17	1	999
19	Row18	1	999
20	Row19	1	999
21	Row20	1	999
22	Row21	1	999
23	Row22	1	999
24	Row23	1	999
25	Row24	1	999
26	Row25	1	999

Figure 61

- **Z-score normalization**

Z-score is a normalization technique to replace each corresponding value with their Z-score for better detection of outliers and performance when training machine learning models. Using Z-score normalization prevents the difference in the scale of values of different attributes from influencing the machine learning models as well as preventing the loss of information.

$$Z - score: \quad x_{normalized} = \frac{x - \mu}{\sigma}$$

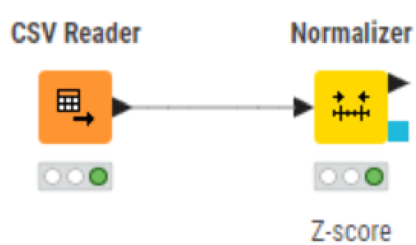


Figure 62

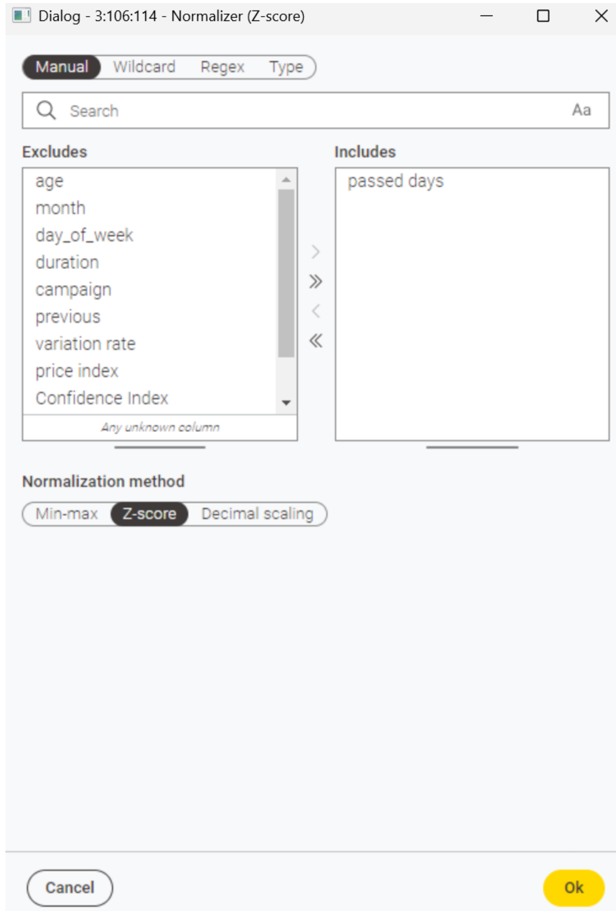
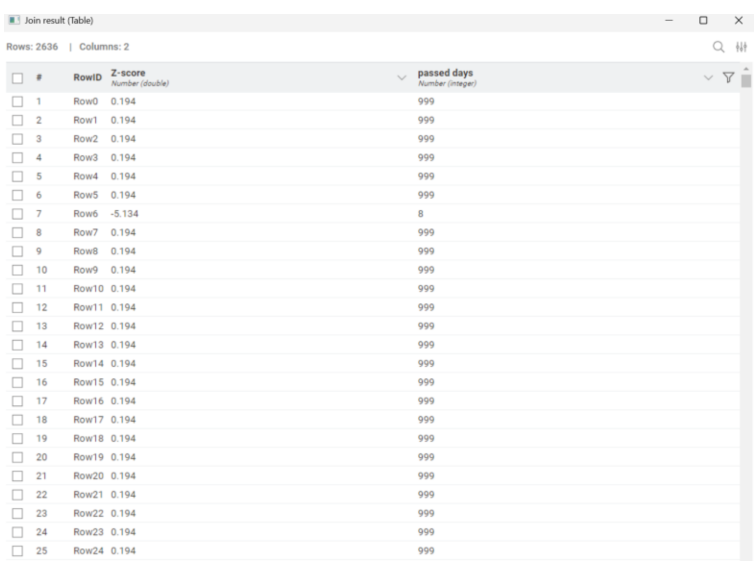


Figure 63


Join result (Table)

Rows: 2636 | Columns: 2

#	RowID	Z-score Number (double)	passed days Number (integer)
1	Row0	0.194	999
2	Row1	0.194	999
3	Row2	0.194	999
4	Row3	0.194	999
5	Row4	0.194	999
6	Row5	0.194	999
7	Row6	-5.134	8
8	Row7	0.194	999
9	Row8	0.194	999
10	Row9	0.194	999
11	Row10	0.194	999
12	Row11	0.194	999
13	Row12	0.194	999
14	Row13	0.194	999
15	Row14	0.194	999
16	Row15	0.194	999
17	Row16	0.194	999
18	Row17	0.194	999
19	Row18	0.194	999
20	Row19	0.194	999
21	Row20	0.194	999
22	Row21	0.194	999
23	Row22	0.194	999
24	Row23	0.194	999
25	Row24	0.194	999
26	Row25	0.194	999

Figure 64

B3. Discretization

Discretization is the process of mapping numerical values to an ordinal label for better interpretability. The attribute “variation rate” is discretized into three distinct categories: Low, Medium, and High as shown in the four figures below.

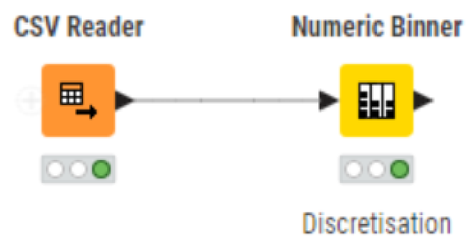


Figure 65

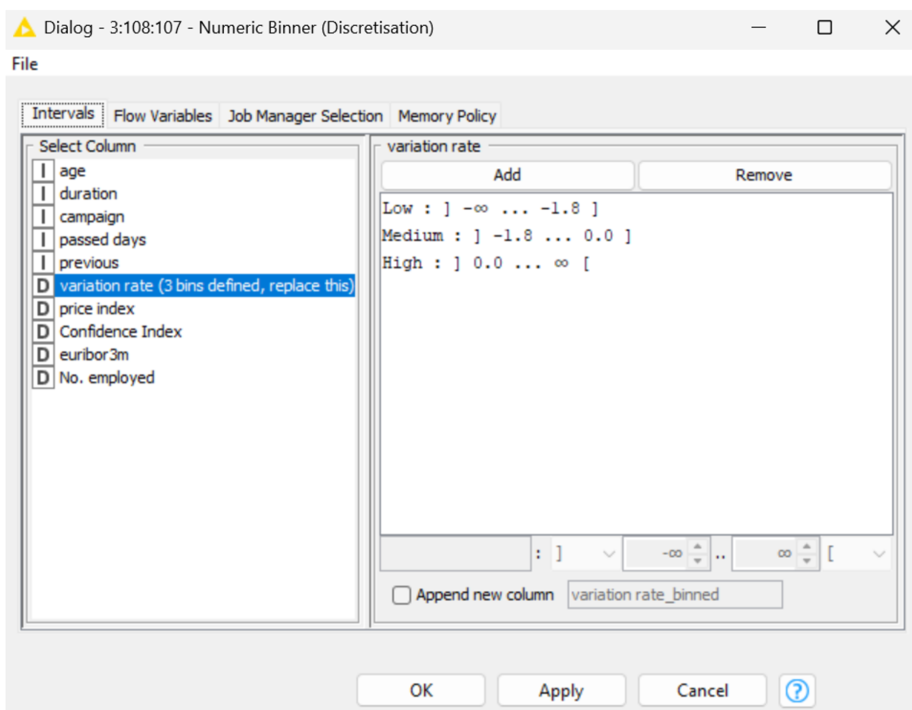


Figure 66

Filtered table (Table)

Rows: 2636 | Columns: 2

#	RowID	variation rate Number (double)	variation rate_binned String
1	Row0	-1.8	Low
2	Row1	-1.8	Low
3	Row2	1.4	High
4	Row3	1.4	High
5	Row4	-0.1	Medium
6	Row5	1.4	High
7	Row6	-1.8	Low
8	Row7	1.1	High
9	Row8	1.4	High
10	Row9	-0.1	Medium
11	Row10	-0.1	Medium
12	Row11	-1.8	Low
13	Row12	1.1	High
14	Row13	1.4	High
15	Row14	1.4	High
16	Row15	1.4	High
17	Row16	-3.4	Low
18	Row17	1.4	High
19	Row18	1.1	High
20	Row19	-1.8	Low
21	Row20	1.1	High
22	Row21	-2.9	Low
23	Row22	1.4	High
24	Row23	-1.8	Low
25	Row24	1.4	High
26	Row25	1.4	High

Figure 67

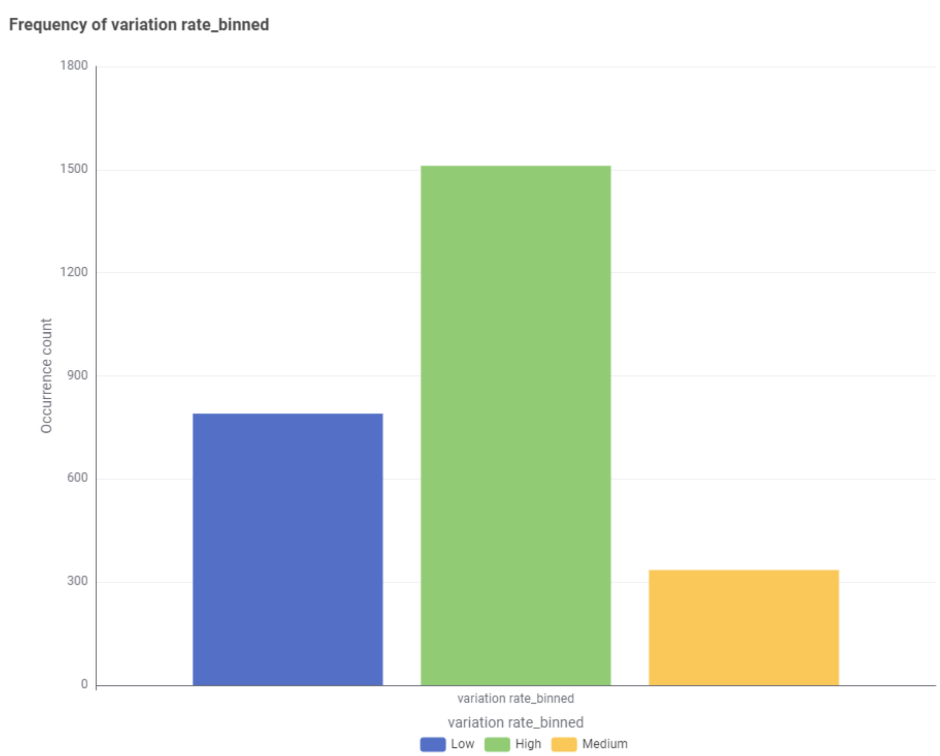


Figure 68

B4. Binarization

Binarization is the process of mapping categorical attributes into multiple binary variables. The process and result of performing binarization onto the attribute “contact” are shown in the three figures below.

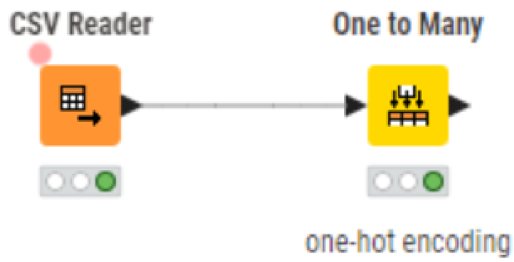


Figure 69

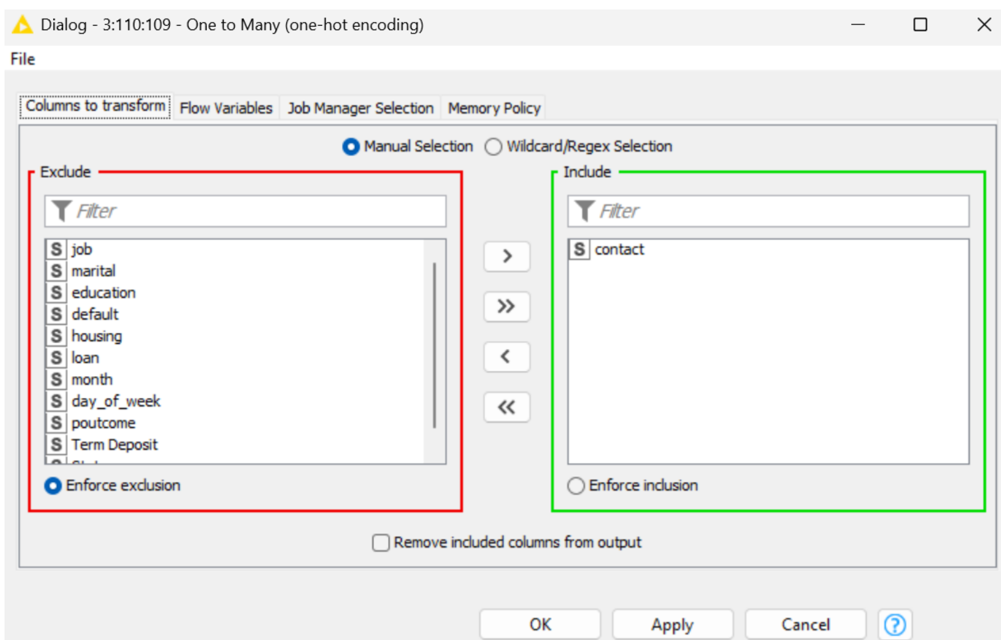


Figure 70

Filtered table (Table)

Rows: 2636 | Columns: 6

#	RowID	contact (String)	Fax (Number (integer))	Cellphone (Number (integer))	Email (Number (integer))	Telephone (Number (integer))	Mailing (Number (integer))
1	Row0	Fax	1	0	0	0	0
2	Row1	Fax	1	0	0	0	0
3	Row2	Cellphone	0	1	0	0	0
4	Row3	Email	0	0	1	0	0
5	Row4	Telephone	0	0	0	1	0
6	Row5	Mailing	0	0	0	0	1
7	Row6	Cellphone	0	1	0	0	0
8	Row7	Email	0	0	1	0	0
9	Row8	Cellphone	0	1	0	0	0
10	Row9	Mailing	0	0	0	0	1
11	Row10	Telephone	0	0	0	1	0
12	Row11	Email	0	0	1	0	0
13	Row12	Cellphone	0	1	0	0	0
14	Row13	Cellphone	0	1	0	0	0
15	Row14	Email	0	0	1	0	0
16	Row15	Telephone	0	0	0	1	0
17	Row16	Cellphone	0	1	0	0	0
18	Row17	Mailing	0	0	0	0	1
19	Row18	Mailing	0	0	0	0	1
20	Row19	Cellphone	0	1	0	0	0
21	Row20	Mailing	0	0	0	0	1
22	Row21	Email	0	0	1	0	0
23	Row22	Fax	1	0	0	0	0
24	Row23	Cellphone	0	1	0	0	0
25	Row24	Cellphone	0	1	0	0	0
26	Row25	Cellphone	0	1	0	0	0

Figure 71

C. Summary of Insights

The following details the most important insights extracted from the dataset:

- The majority of the observations are in their 30s, but the age group that yields the highest average success is people in their 60s. Overall, the target audience for the marketing campaign should be in the age range of $[20, 70)$. While there is not enough supporting evidence due to the small sample size to concretely say that people in the age group of $[10, 20)$, $[70, 80)$, and $[80, 90)$ are ineffective, they should be avoided for now.
- Out of the three distinct marital status categories: Single, Married, and Divorced, people who are single yield the highest average success.
- Overall, the cellphone is the most popular contact method, however, the most effective contact method overall is email. Nevertheless, the most popular/effective contact method will be influenced by which state the customers reside in. The table below summarizes the key contact methods per state.

State	Key points	Most effective
SA	• Mainly uses cellphone, followed by mailing and fax. • Email and telephone are not popular	Mailing
TAS	• Uses all contact methods relatively equally	Email
NSW	• Mainly uses email, followed by mailing and telephone	Telephone
ACT	• Uses all contact methods relatively equally	Fax
WA	• Uses all contact methods relatively equally	Email
VIC	• Telephone and fax are not popular	Fax
QLD	• Mainly uses cellphone followed by telephone • Mailing and email are not popular	Cellphone
NT	• Uses all contact methods relatively equally	Email

- The states SA, ACT, and NT have the highest average success index. It is also advised to avoid the states TAS, WA, and QLD since they have a significantly lower average success index compared to other states.
- Only the month of July has a significant positive average success index, therefore, it is advisable to focus and plan the marketing campaign around July and avoid the months of June and September.
- The price index follows a decreasing trend between February and July and will start to stabilize around August.
- It is not recommended to contact customers on Friday since that is the day with the lowest average success rate. The best days to contact is on Tuesday and Wednesday.
- Based on the dataset, for the marketing campaign to be as successful as possible, the optimal number of contacts performed before the campaign is 3. However, due to the limited range of the "previous" attribute, it is unsure whether 3 is truly the global optimal value. Additionally, it is inadvisable to contact less than 1 time.
- The relationships between variation rate, confidence index, euribor3m, and price index should be investigated more rigorously. Firstly, there is a strong positive linear relationship between variation rate and price index / euribor3m. Additionally, there is potentially a non-linear relationship between variation rate and confidence index. Finally, there is a moderate positive linear relationship between euribor3m and price index, but the relationship could also be non-linear.
- There is insufficient evidence to support the claim that the failure of the previous marketing campaign was due to external economic factors. The scatter plots between the poutcome and confidence index, price index, and euribor3m show no concrete correlation.