Assessment 2

Data exploration and preparation

1A. Initial data exploration

1. Types of each attribute in the dataset:

The attribute type of the attributes in the given dataset is as follows:

Heading	Attribute Type	Explanation
SK_ID_CURR	Nominal	This attribute is nominal because it represents unique identifiers for each customer. It is categorical in nature and doesn't have a natural order or hierarchy. Each customer ID is distinct and serves as a label or identifier without any inherent numeric or ordinal meaning.
TARGET	Nominal	TARGET is nominal as it represents a categorical label for loan applicants. It has two distinct values: 0 (indicating the applicant is unable to repay the loan) and 1 (indicating the applicant can repay the loan). These values are nominal labels without an inherent numeric order.
NAME_CONTRACT_TYPE	Nominal	This attribute is nominal because it categorizes loan contracts into two distinct types: "Cash loans" and "Revolving loans." There is no inherent order or hierarchy between these contract types, making them nominal categories.
CODE_GENDER	Nominal	CODE_GENDER is nominal as it categorizes loan applicants into gender categories,"M" and "F". Like other nominal attributes, it doesn't imply any order or ranking between these categories; they

		are distinct labels for gender identification.
FLAG_OWN_CAR	Nominal	FLAG_OWN_CAR is nominal as it categorizes applicants into two distinct groups: "Y" (owning a car) and "N" (not owning a car). These categories have n no natural order and represent a nominal classification based on car ownership status.
FLAG_OWN_REALTY	Nominal	(Same reason as FLAG_OWN_CAR)
CNT_CHILDREN	Ordinal	CNT_CHILDREN represent a meaningful order or ranking based on the number of children an applicant has. The order is essential to understand the data contect in family size and allows comparison between different count of children.
AMT_INCOME_TOTAL	Ratio	AMT_INCOME_TOTAL is classified as a ratio variable because it represents a continuous, quantifiable measurement of total income with a true zero point, allowing for meaningful mathematical operations and statistical analysis.
AMT_CREDIT	Ratio	(Same reason as AMT_INCOME_TOTAL)
AMT_ANNUITY	Ratio	(Same reason as AMT_INCOME_TOTAL)
AMT_GOODS_PRICE	Ratio	(Same reason as AMT_INCOME_TOTAL)
NAME_TYPE_SUITE	Nominal	
NAME_INCOME_TYPE	Nominal	NAME_INCOME_TYPE is classified as nominal because it represents different income source categories for loan applicants. It includes categories such as "Pensioner", "Working," “Commercial associate" and more. These categories serve as distinct

		labels for characterizing the source of income for each applicant. There is no inherent order or ranking between these categories, making it a nominal attribute.
NAME_EDUCATION_TYPE	Nominal	
NAME_FAMILY_STATUS	Nominal	
NAME_HOUSING_TYPE	Nominal	
REGION_POPULATION_RELATIVE	Nominal	
DAYS_BIRTH	Ratio	DAYS_BIRTH is a ratio variable, as it represents a continuous and quantifiable measurement of age in days, with a true zero point, allowing for mathematical and statistical analysis.
DAYS_EMPLOYED	Nominal	(Same reason as DAYS_BIRTH)
DAYS_REGISTRATION	Nominal	(Same reason as DAYS_BIRTH)
DAYS_ID_PUBLISH	Nominal	(Same reason as DAYS_BIRTH)
FLAG_MOBIL	Nominal	
FLAG_EMP_PHONE	Nominal	
FLAG_WORK_PHONE	Nominal	
FLAG_CONT_MOBILE	Nominal	
FLAG_PHONE	Nominal	
FLAG_EMAIL	Nominal	
CNT_FAM_MEMBERS	Ordinal	(Same reason as CNT_CHILDREN)
REGION_RATING_CLIENT	Ordinal	REGION_RATING is considered an ordinal variable because it consists of ordered categories with a meaningful ranking based on the rating levels of regions.
REGION_RATING_CLIENT_W_CITY	Ordinal	
WEEKDAY_APPR_PROCESS_START	Nominal	
HOUR_APPR_PROCESS_START	Nominal	

REG_REGION_NOT_LIVE_REGION	Nominal	REG_REGION_NOT_LIVE_REGION is classified as a nominal variable as it represents distinct categories without a meaningful order, and it serves to categorically differentiate between two conditions regarding the client's address information.
REG_REGION_NOT_WORK_REGION	Nominal	(Same reason as REG_REGION_NOT_LIVE_REGION)
LIVE_REGION_NOT_WORK_REGION	Nominal	(Same reason as REG_REGION_NOT_LIVE_REGION)
REG_CITY_NOT_LIVE_CITY	Nominal	(Same reason as REG_REGION_NOT_LIVE_REGION)
REG_CITY_NOT_WORK_CITY	Nominal	(Same reason as REG_REGION_NOT_LIVE_REGION)
LIVE_CITY_NOT_WORK_CITY	Nominal	(Same reason as REG_REGION_NOT_LIVE_REGION)
ORGANIZATION_TYPE	Nominal	
EXT_SOURCE_2	Ratio	EXT_SOURCE attributes are classified as ratio variables because they are continuous, quantifiable measurements with a true zero point, allowing for mathematical and statistical analysis.
EXT_SOURCE_3	Ratio	
OBS_30_CNT_SOCIAL_CIRCLE	Ordinal	OBS_30_CNT_SOCIAL_CIRCLE is classified as an ordinal variable as it represents ordered categories with a meaningful ranking based on the count of observations, and the analysis of such variables often involves methods suited for ordinal data.
DEF_30_CNT_SOCIAL_CIRCLE	Ordinal	
OBS_60_CNT_SOCIAL_CIR	Ordinal	

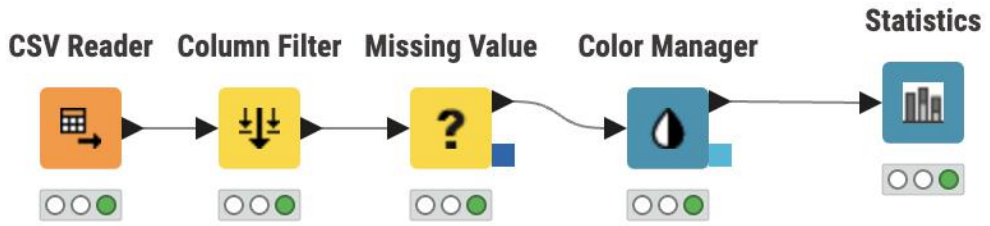
CLE		
DEF_60_CNT_SOCIAL_CIRCLE	Ordinal	
DAYS_LAST_PHONE_CHANGE	Nominal	
FLAG_DOCUMENT_2	Nominal	
FLAG_DOCUMENT_3	Nominal	
FLAG_DOCUMENT_4	Nominal	
FLAG_DOCUMENT_5	Nominal	
FLAG_DOCUMENT_6	Nominal	
FLAG_DOCUMENT_7	Nominal	
FLAG_DOCUMENT_8	Nominal	
FLAG_DOCUMENT_9	Nominal	
FLAG_DOCUMENT_10	Nominal	
FLAG_DOCUMENT_11	Nominal	
FLAG_DOCUMENT_12	Nominal	
FLAG_DOCUMENT_13	Nominal	
FLAG_DOCUMENT_14	Nominal	
FLAG_DOCUMENT_15	Nominal	
FLAG_DOCUMENT_16	Nominal	
FLAG_DOCUMENT_17	Nominal	
FLAG_DOCUMENT_18	Nominal	
FLAG_DOCUMENT_19	Nominal	
FLAG_DOCUMENT_20	Nominal	
FLAG_DOCUMENT_21	Nominal	
AMT_REQ_CREDIT_BUREAU_HOUR	Ordinal	AMT_REQ_CREDIT_BUREAU_HOUR is classified as an ordinal variable AS it represents ordered categories with a meaningful ranking based on the count of inquiries, and the analysis of such variables often involves methods suited for ordinal data.
AMT_REQ_CREDIT_BUREAU_DAY	Ordinal	
AMT_REQ_CREDIT_BUREAU_WEEK	Ordinal	
AMT_REQ_CREDIT_BUREAU_MON	Ordinal	
AMT_REQ_CREDIT_BUREAU_QRT	Ordinal	

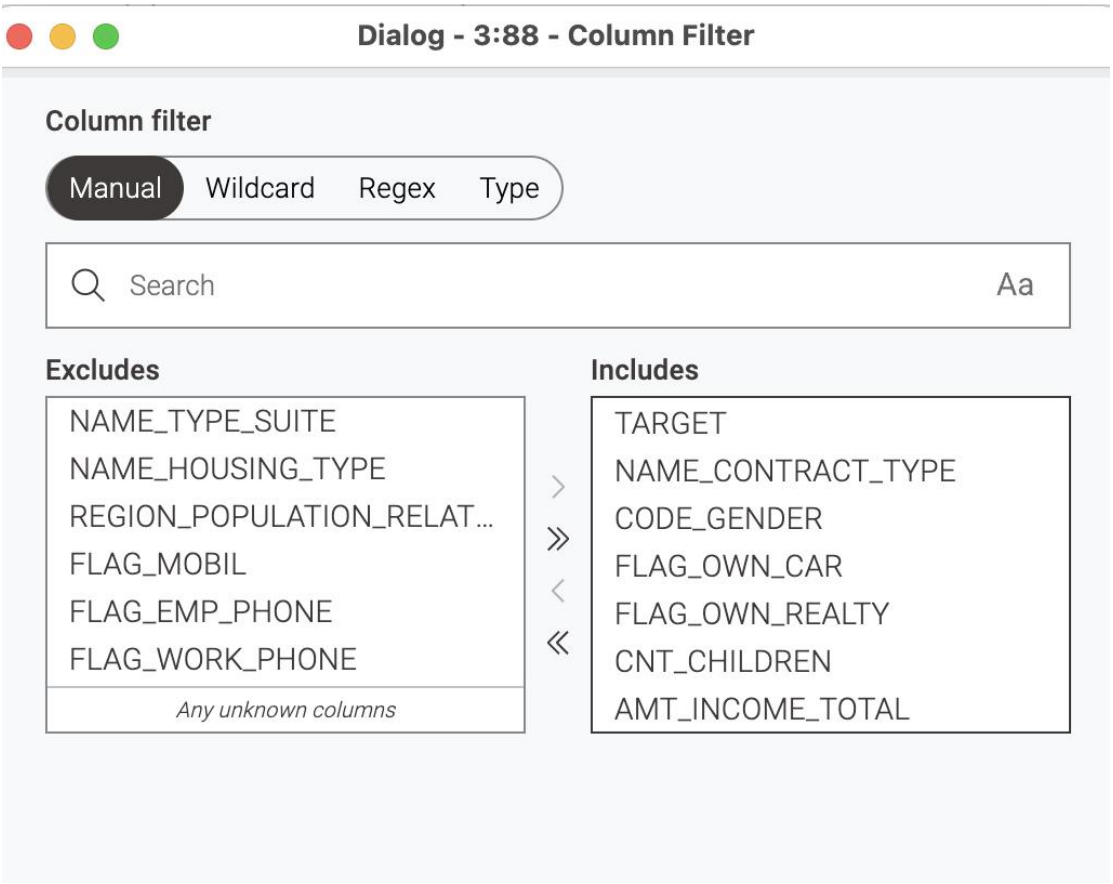
AMT_REQ_CREDIT_BUREAU_YEAR	Ordinal	

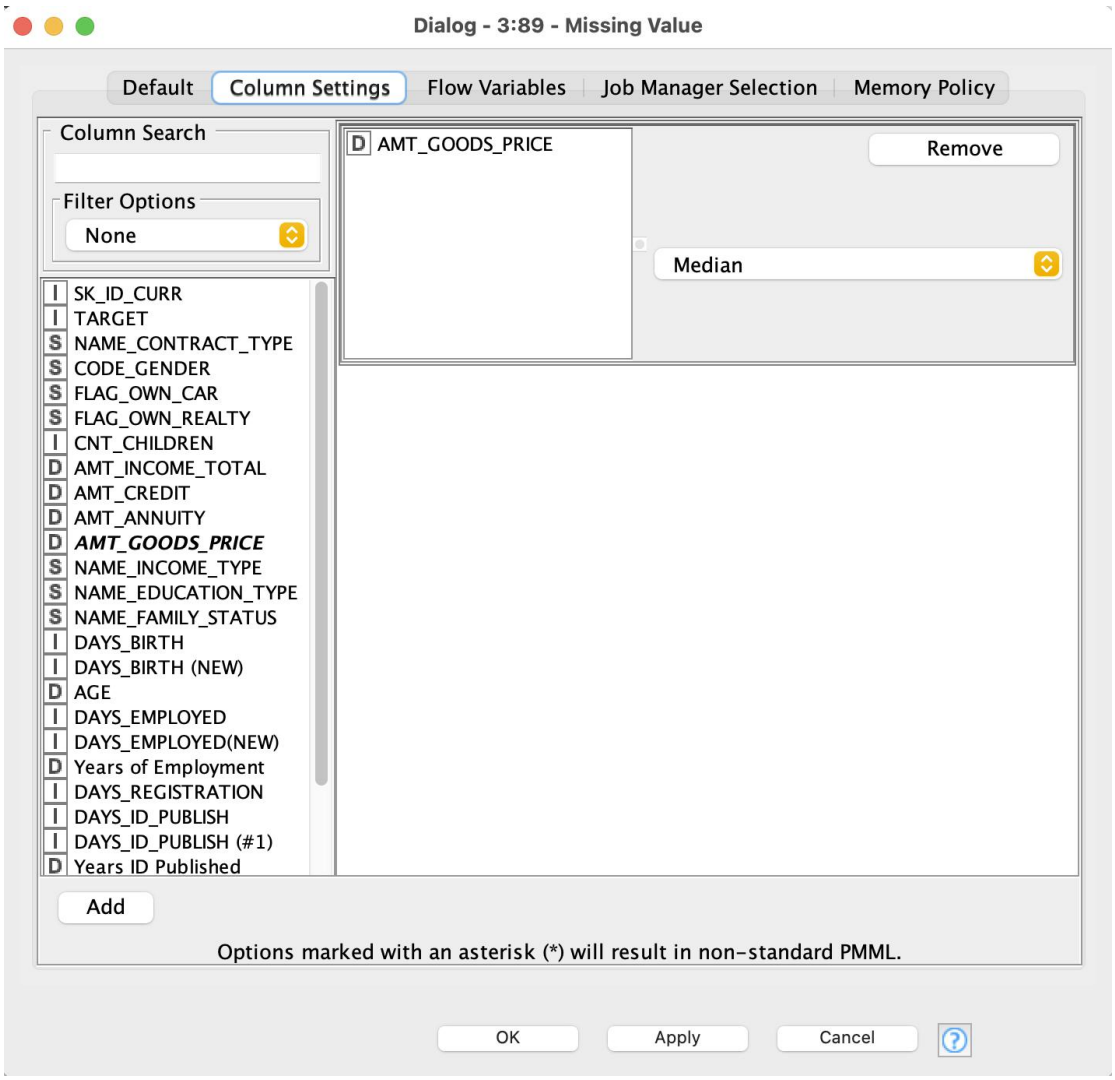
*The **bolded** attributes are the attributes that I have chosen for the analytic process.*

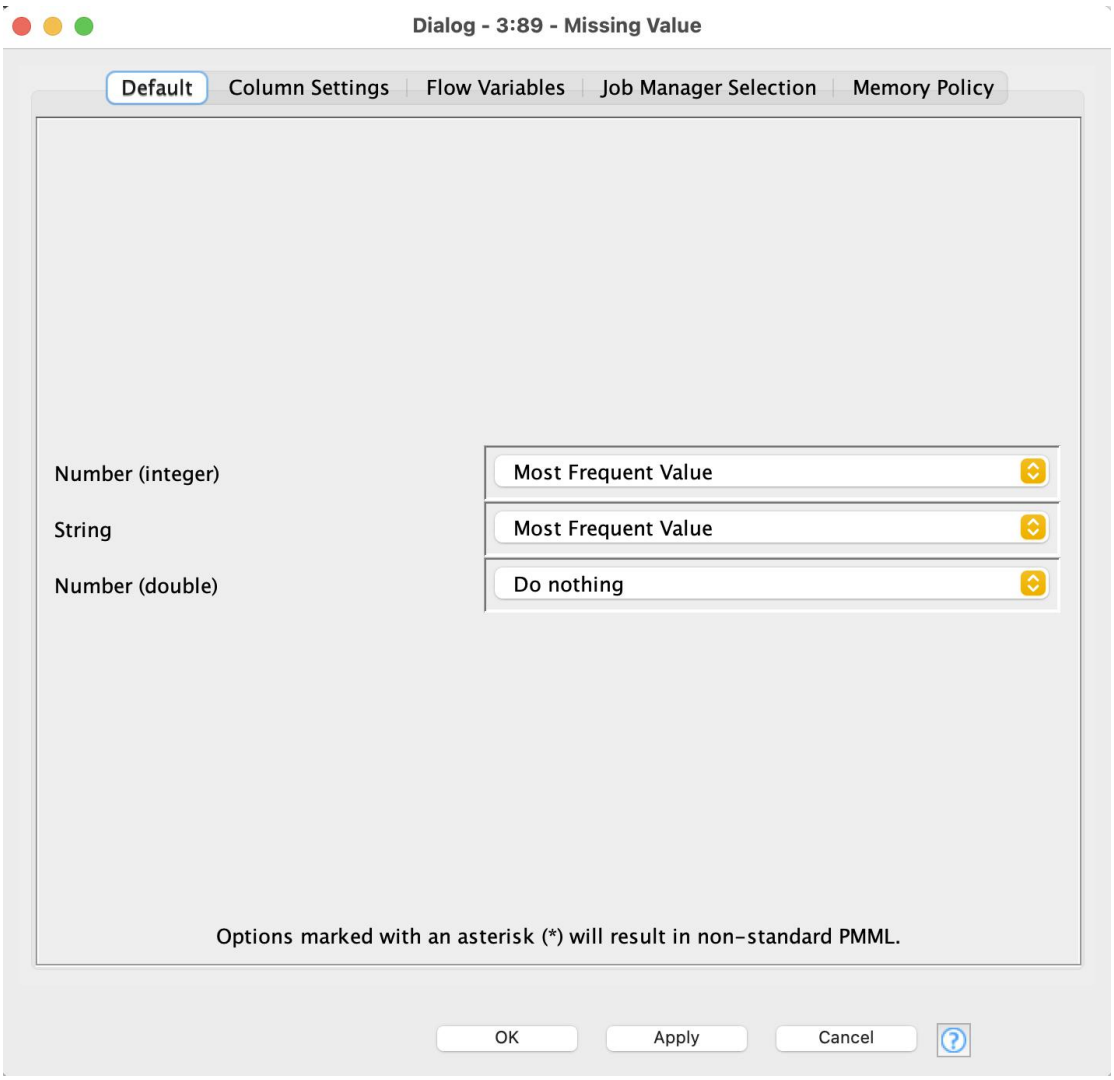
2. Identifying Summary Statistics and Visualizations for Attributes:

2.1. Steps for analysing the quantitative data






Dialog - 3:89 - Missing Value
Default
Column Settings
Flow Variables
Job Manager Selection
Memory Policy
Column Search
Filter Options
None
I SK_ID_CURR
I TARGET
S NAME_CONTRACT_TYPE
S CODE_GENDER
S FLAG_OWN_CAR
S FLAG_OWN_REALTY
I CNT_CHILDREN
D AMT_INCOME_TOTAL
D AMT_CREDIT
D AMT_ANNUITY
D AMT_GOODS_PRICE
S NAME_INCOME_TYPE
S NAME_EDUCATION_TYPE
S NAME_FAMILY_STATUS
I DAYS_BIRTH
I DAYS_BIRTH (NEW)
D AGE
I DAYS_EMPLOYED
I DAYS_EMPLOYED(NEW)
D Years of Employment
I DAYS_REGISTRATION
I DAYS_ID_PUBLISH
I DAYS_ID_PUBLISH (#1)
D Years ID Published
D AMT_GOODS_PRICE
Remove
Median
Add
Options marked with an asterisk (*) will result in non-standard PMML.
OK
Apply
Cancel



Statistics

Rows: 8 | Columns: 7

Name	Type	Minimum	Maximum	50% Quantile (M...	Mean	Variance
AMT_INCOME_TO...	Number (double)	27,000	3,150,000	139,500	165,264.437	11,101,354,120.424
AMT_CREDIT	Number (double)	45,000	3,020,760	512,338.5	587,956.262	147,467,722,821.0...
AMT_ANNUITY	Number (double)	2,173.5	225,000	25,447.5	27,335.25	206,895,248.725
AMT_GOODS_PRICE	Number (double)	45,000	2,700,000	450,000	523,220.821	122,991,049,108.6...
DAYS_BIRTH	Number (integer)	-25,132	-7,783	-15,014.5	-15,527.477	19,217,263.849
DAYS_EMPLOYED	Number (integer)	-16,348	365,243	-1,136.5	56,462.063	18,126,271,088.847
DAYS_REGISTRATI...	Number (integer)	-17,039	0	-4,368	-4,700.426	11,799,750.532
DAYS_ID_PUBLISH	Number (integer)	-6,155	-4	-3,107	-2,918.524	2,259,665.311

Figure 1. Steps for analysing the statistical summary of quantitative data

1. I utilised the CSV Reader to import the data file, followed by the Column Filter to extract the selected 20 attribute attributes.
2. To address missing values, I focused on the 'Total Amount of Goods' attribute, which had two missing values. These gaps were filled with the median value, providing a robust way to impute the missing data.
3. Additionally, I addressed missing values in 'OBS_60_CNT_SOCIAL_CIRCLE,' 'DEF_60_CNT_SOCIAL_CIRCLE,' 'OBS_30_CNT_SOCIAL_CIRCLE,' and 'DEF_30_CNT_SOCIAL_CIRCLE' by replacing them with the most frequent value, which in this case was '0.' Notably, the proportion of missing values in these attributes was minimal, accounting for less than 1% of the database. This imputation strategy ensures that missing values do not significantly impact the analysis.
4. I have selected the ratio attributes to analyse the Types of the attributes, Minimum, Maximum, Median, Mean and Variance of the details.

2.1.1 Identifying Summary Statistics and Visualizations for Attributes:

1. Amount Income

Statistics

Rows: 8 | Columns: 7

Name	Type	Minimum	Maximum	50% Quantile (M...	Mean	Variance
AMT_INCOME_TO...	Number (double)	27,000	3,150,000	139,500	165,264.437	11,101,354,120.424

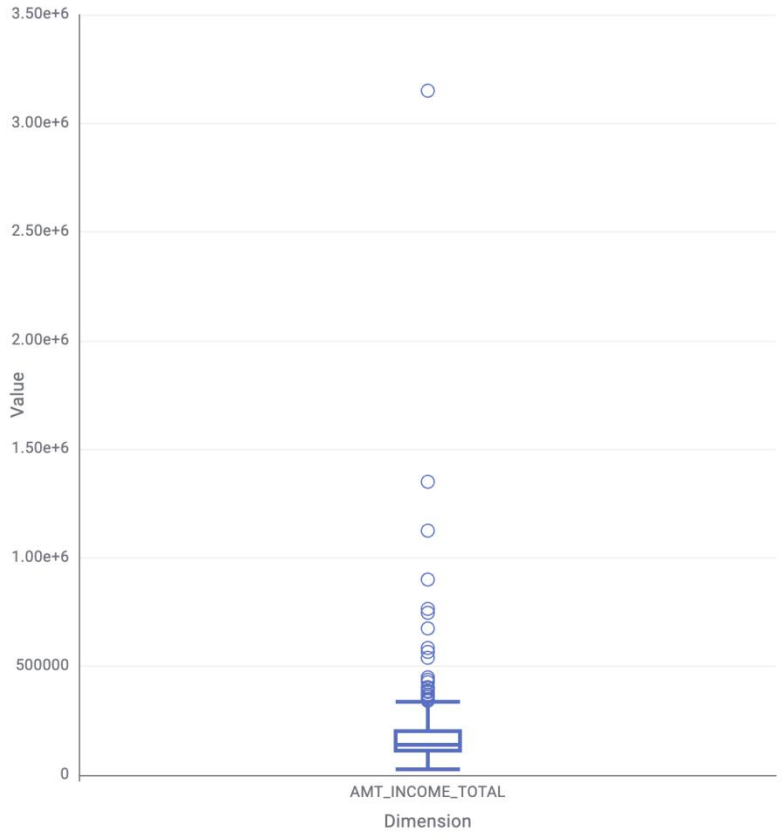


Figure 2 Box Plot for Total Income Amount

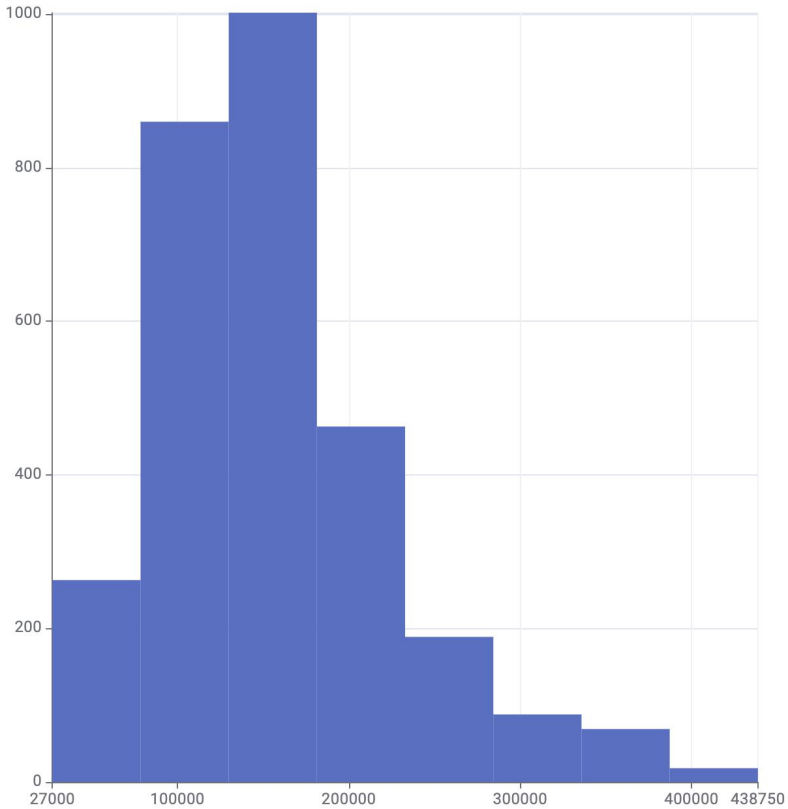


Figure 3 Histogram for Total Income Amount after removing the outlier

- The dataset's "Total Income Amount" attribute exhibits exciting characteristics. The mean income is approximately 165,264.44, while the median income is slightly lower at 139,500. This suggests the presence of right-skewness in the income distribution.
- Multiple potential outliers are evident in the "Total Income Amount." These outliers, represented as high-value data points, are noticeable in the box plot visualisation. Removing these outliers could significantly impact the distribution and statistics related to income, making them intriguing facts for further analysis.
- Based on the histogram, approximately one-third of the individuals in the dataset fall within the income range of 130,000 to 180,000, indicating a concentration of income in this range.

2. Amount Credit

Statistics

Rows: 1 | Columns: 7

Name	Type	Minimum	Maximum	50% Quantile (M...	Mean	Variance
AMT_CREDIT	Number (double)	45,000	3,020,760	512,338.5	587,956.262	147,467,722,821.0...

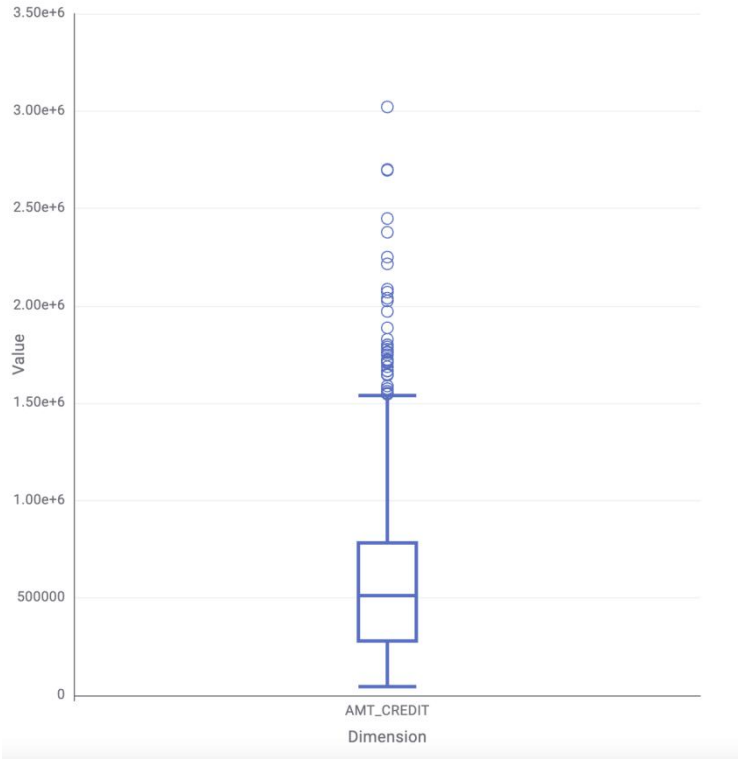


Figure 4 Box Plot for Amount of Credit

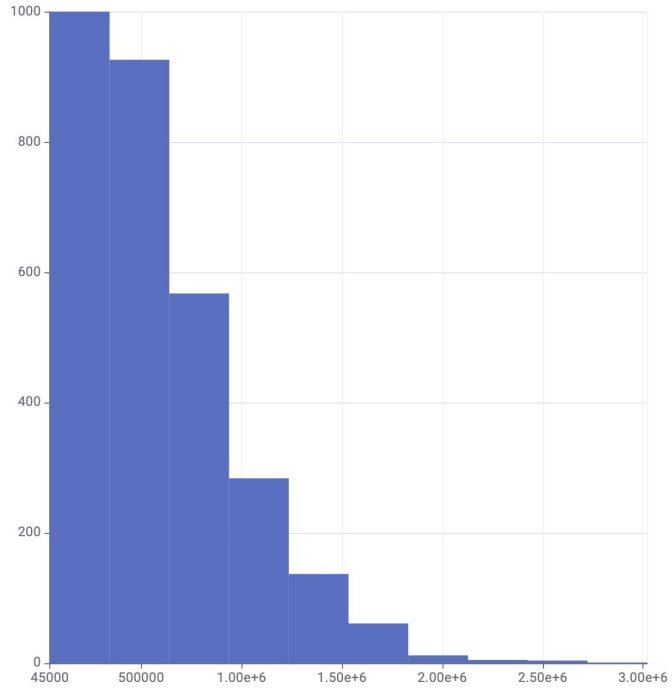


Figure 5 Histogram for Amount of Credit

- The "Total Amount of Credit" attribute spans a substantial range of 45,000 to 3,020,760.26.
- On average, loan applicants have a total amount of credit of approximately

587,956.26, as indicated by the mean. However, the median value, 512,338.5, falls below the standard.

- Based on the histogram, the "Total Amount of Credit" distribution appears right-skewed, with most loans concentrated in the lower to middle credit amount range.
- The difference between the mean and median suggests that some loans with notably high credit amounts pull the mean upwards, contributing to the right-skewness.
- The box plot visualisation reveals the presence of potential outliers with highly high credit amounts. These outliers could significantly impact the distribution and statistics associated with credit amounts and warrant further investigation.

3. Amount of annuity

Statistics

Rows: 1 | Columns: 7

Name	Type	Minimum	Maximum	50% Quantile (M...	Mean	Variance
AMT_ANNUITY	Number (double)	2,173.5	225,000	25,447.5	27,335.25	206,895,248.725

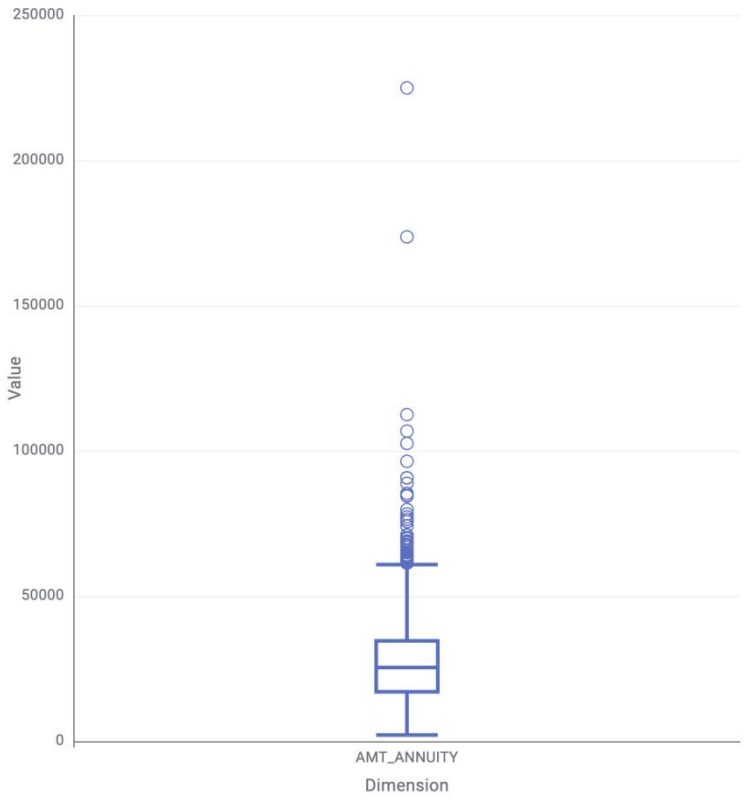


Figure 6 Box Plot of Amount of Annuity

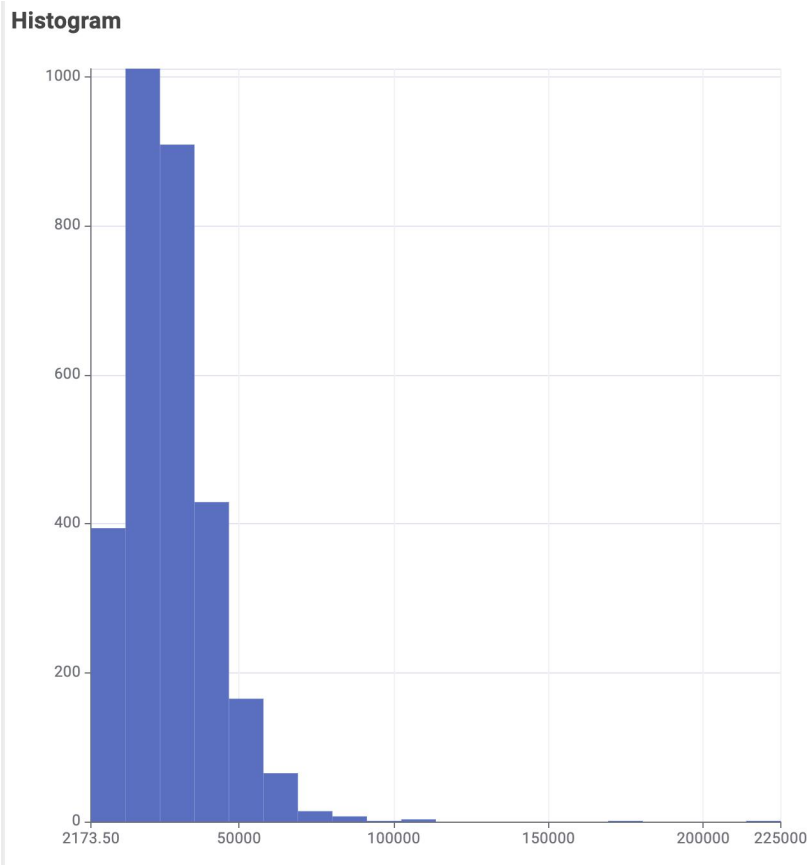


Figure 7 Histogram for Total Income Annuity with 20 bins

- The "Annuity Amount" attribute exhibits a wide range of values, with the value range spanning from 2,173.50 to 225,000.
- On average, loan applicants have an annuity of approximately 27,335.25, as indicated by the mean. However, the median value, 25,447.5, falls slightly below the standard.
- The distribution of "Annuity Amount" is right-skewed, with the majority of annuity amounts concentrated in the lower to middle range.
- There are potential outliers with exceptionally high annuity amounts, as evident from the box plot and histogram visualisation.
- Based on the histogram, it's evident that around 50% of applicants have annuity amounts ranging from 13,000 to 35,000.
- Furthermore, the "Annuity Amount" attribute displays significant variability among applicants, as indicated by the high variance. This variability suggests that some applicants have substantially higher or lower annuity amounts than the mean.

4. Amount of Good Price

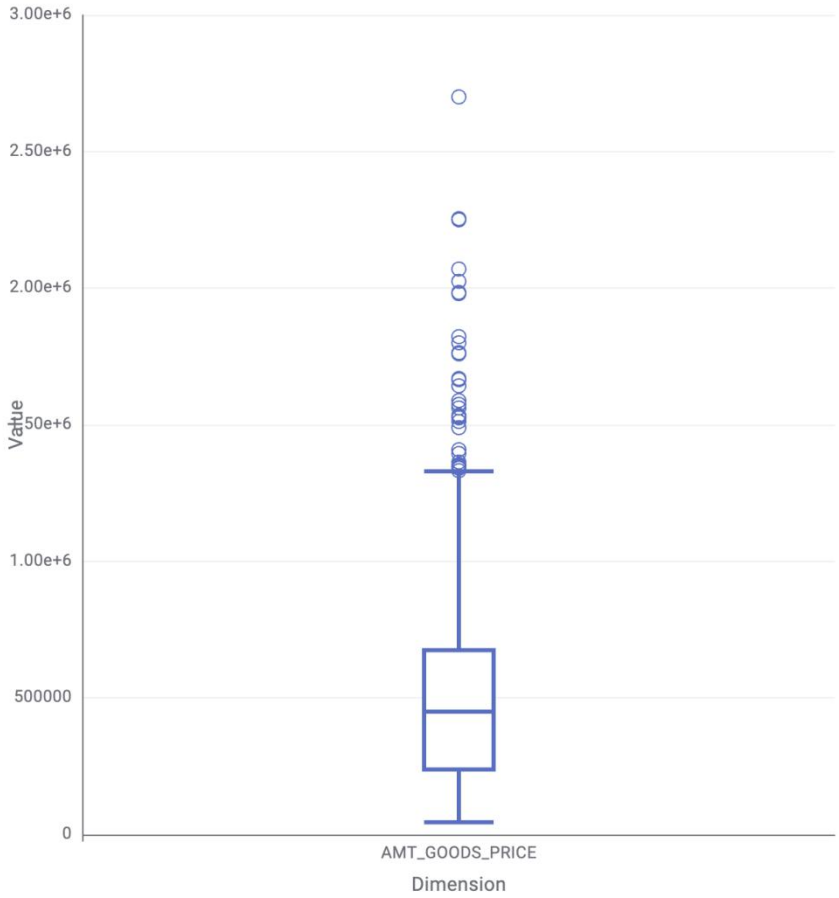


Figure 8 Box Plot of Amount of Good Price

- The "Goods Price Amount" attribute encompasses a broad range of values from 45,000 to 2,700,000.00.
- On average, loan applicants have a goods price of approximately 523,172.007, represented by the mean. However, the median value of 450,000 is notably lower, indicating a right-skewed distribution.
- The "Goods Price Amount" distribution is right-skewed, with most prices concentrated in the lower to middle range.
- Potential outliers with exceptionally high goods prices exist, clearly discernible in the box plot visualisation.
- Additionally, the "Goods Price Amount" attribute exhibits significant variability among applicants, signified by the substantial variance. This variance underscores that some applicants possess substantially higher or lower goods price amounts than the mean.

5. Days_Birth

Name	Minimum	Maximum	50% Quantile (M...	Mean
DAYS_BIRTH	-25,132	-7,783	-15,014.5	-15,527.477
AGE	21.32	68.85	41.14	42.541

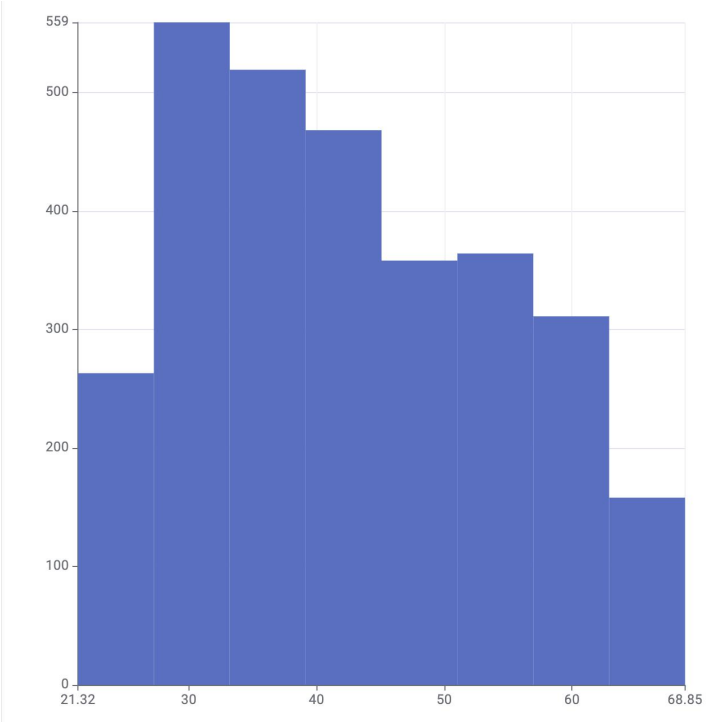


Figure 9 Histogram Days of Birth

- The "Days of Birth" attribute has been transformed into ages in years for improved interpretability. This conversion involved dividing the number of days of birth by 365 and representing it as a positive number.
- The age of applicants in the dataset spans from around 21.32 years to 68.85 years.
- On average, the age of applicants is approximately 42.541 years, as indicated by the mean. Meanwhile, the median age stands at about 41.14 years.
- Based on the histogram, a significant portion of clients, approximately 50%, falls within the age range of 28 years to 50 years.
- Moreover, within this broader range, approximately one-third of the dataset is concentrated in the narrower age range from 28 to 40.

6. Days Employment

Name	Minimum	Maximum	50% Quantile (M…	Mean
DAYS_EMPLOYED	-16,348	365,243	-1,136.5	56,462.063
Years of Employm…	0	46.68	11.97	12.878

Figure 10 Statistic Table of Days Employment

The above table shows the potential of error data in DAYS_EMPLOYED. The maximum number of “365,243” appears in 1/7 of the row with a positive number; meanwhile, other data in the same column is in a negative number.

Name	Minimum	Maximum	50% Quantile (M…	Mean
DAYS_EMPLOYED	-16,348	-15	-1,465	-2,207.539
Years of Employm…	0.04	44.79	4.01	6.048

Figure 11 Statistic Table of Days Employment after modify

The above table shows the result after removing the row that contains “365,245” in the “DAYS_EMPLOYED” column with the row filter node in KNIME.

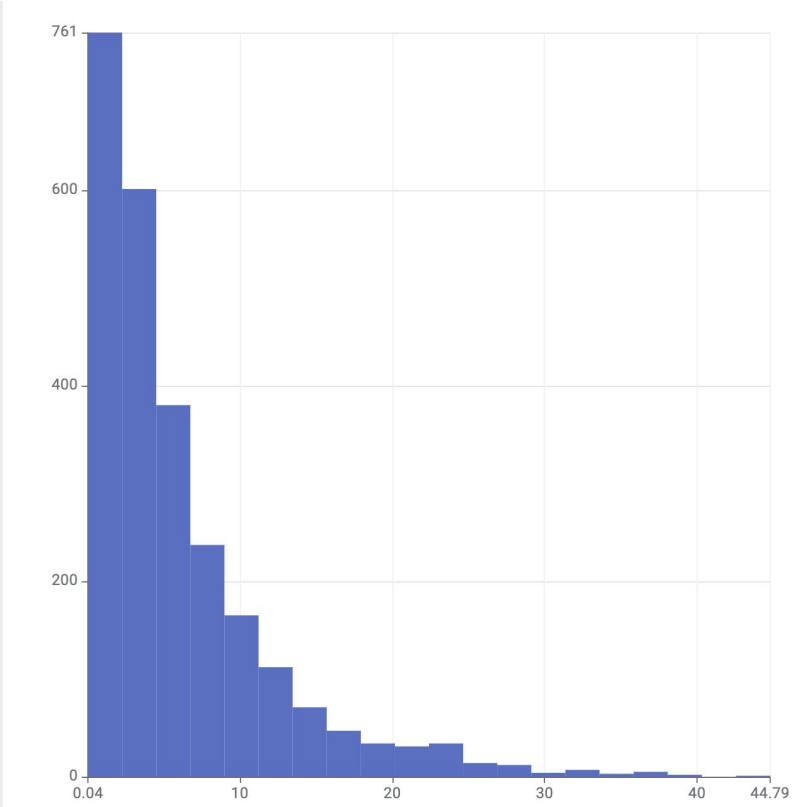


Figure 12 Histogram of Days Employed

- The "Days Employed" attribute has been transformed into employment durations in years for improved interpretability, a transformation carried out using Excel.
- The employment duration of applicants in the dataset spans a wide range, from approximately 0.04 years to 44.79 years.
- On average, applicants have an employment duration of approximately 6,048 years, as represented by the mean. Meanwhile, the median employment duration is about 4.01 years.
- Notably, one-third of the applicants have fewer than five years of employment.

7. Days ID Published

Name	Minimum	Maximum	50% Quantile (M...	Mean
DAYS_ID_PUBLISH	-6,155	-4	-3,107	-2,918.524
Years ID Published	0.01	16.86	8.51	7.996

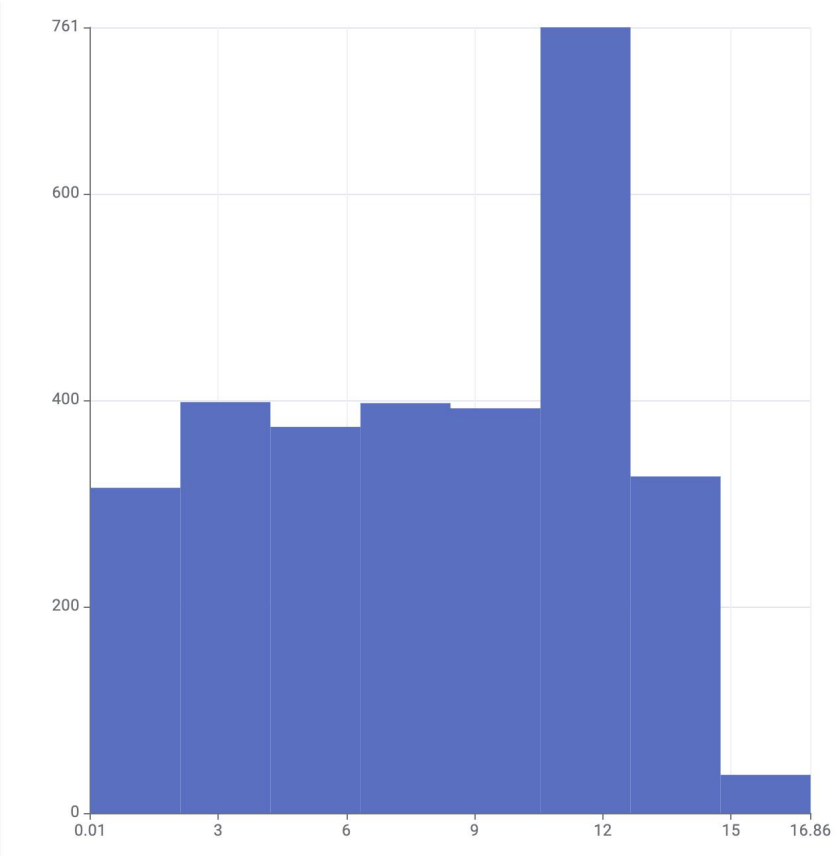


Figure 13 Histogram of Days ID Published

- The "Days ID Published" attribute has been transformed into ID document publishing durations in years to enhance interpretability, a transformation using Excel.
- ID document publishing in the dataset ranges from approximately 0.01 to 16.86 years.
- On average, the duration is approximately 7.996 years, as indicated by the mean. Meanwhile, the median time stands at about 8.51 years.
- The distribution of ID document publishing durations exhibits a left-skewed pattern, with most applicants having durations falling between 10 to 13 years, as evident from the histogram.

2.2. Frequency count and visualisation of the qualitative attribute (Nominal and Ordinary)

1. Target

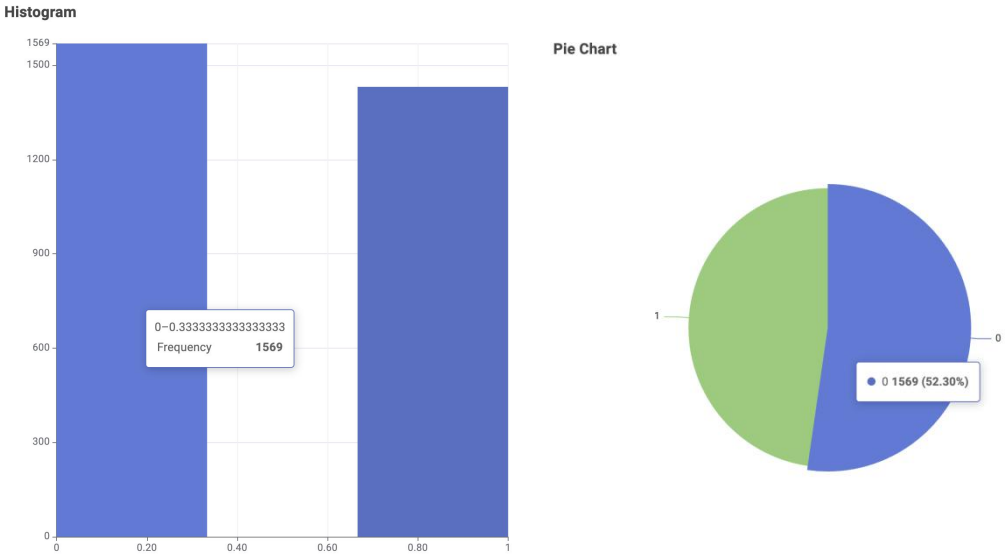


Figure 14 Histogram and Pie Chart of Target

- To facilitate the calculation of frequency and proportion, I performed a transformation on the 'TARGET' attribute. Specifically, I converted numeric values to string labels using the 'Number to String' node in KNIME. This transformation streamlined counting occurrences of 'TARGET' values 1 and 0.
- Subsequently, I employed a pie chart to represent the proportions of loan repayment within the dataset visually:
- Proportion of Loan Repayment (TARGET=1): Approximately 47.7%, 1431 Applicants.
- Proportion of Non-Repayment (TARGET=0): Approximately 52.3%, 1569 Applicants.

2. Count of Children

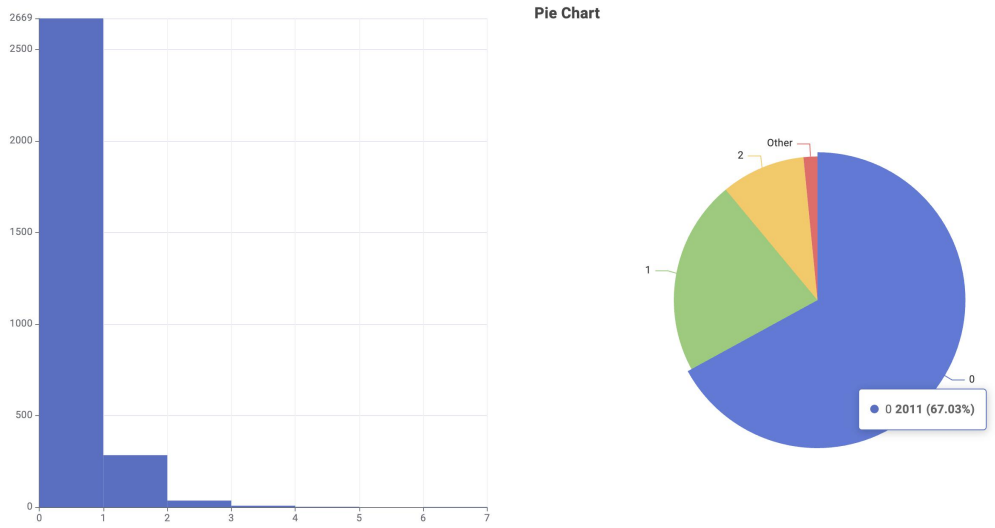


Figure 15 Histogram and Pie Chart of Count of Children

- I transformed the "Count of Children" attribute from numeric values to string labels for frequency analysis.
- The range of the children count varies from 0 to 7, reflecting the number of children each applicant has.
- The histogram shows that a substantial portion (1/6) of the applicants fall into having 0 or 1 child.
- Approximately 67.03% of applicants have no children, underscoring that many applicants do not have dependents or a limited number of dependents.

3. Name Contract Type

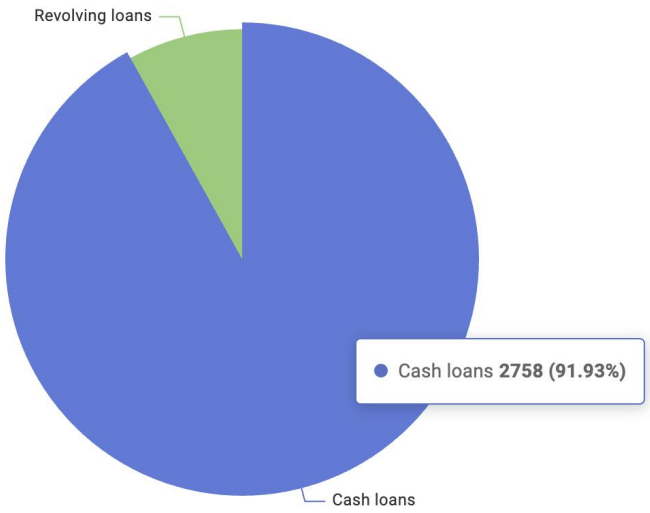


Figure 16 Pie Chart for Contract Type

- The primary loan contract types are "Cash loans" and "Revolving loans."
- "Cash loans" constitute the majority, accounting for approximately 91.93% of the loan types in the dataset.

4. Code Gender

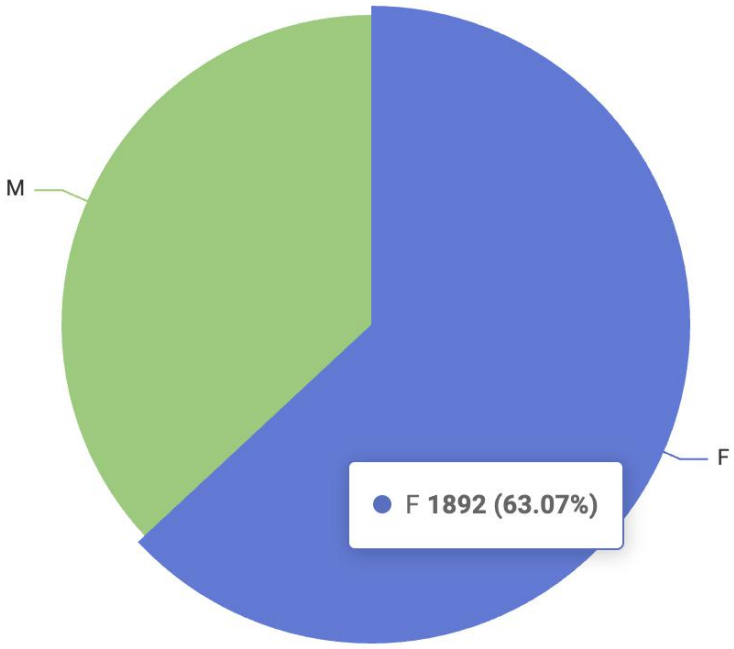


Figure 17 Pie Chart for Gender Code

- The dataset includes primarily two gender categories, "Female" and "Male."
- "Female" applicants (63.07%) are more common than "Male" applicants (36.93%).

5. Flag Own Car

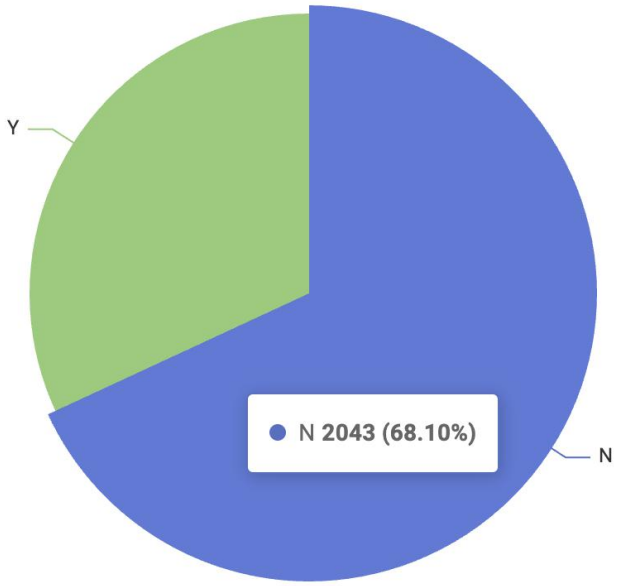


Figure 18 Pie Chart for Flag Own Car

- The dataset contains applicants who both own and do not own cars, with "No Car" being slightly more common (68.10%) with an occurrence frequency of 2043.

6. Flag Own Reality

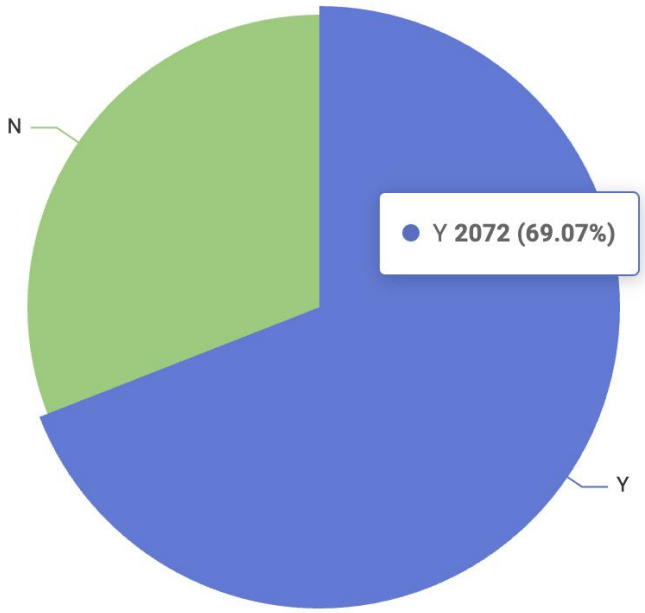


Figure 19 Pie Chart for Flag Own Reality

- The dataset contains applicants who both own realty and do not own realty, with "Owns Realty" being more common (69.07%) with an occurrence frequency of 2072.

7. Name Income Type

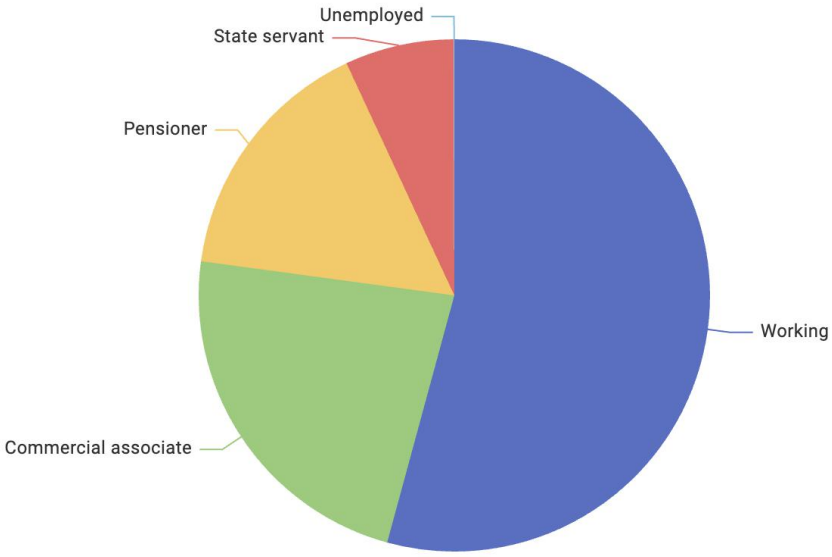


Figure 20 Pie Chart for Income Type

- The most common income type among applicants is "Working," accounting for approximately 54.23% of the total.
- "Commercial associate" and "Pensioner" are the most common income types, making up around 24.90% and 15.93%, respectively.
- Other income types, such as "State servant" and "Unemployed are less common, each representing a small percentage of the total.

8. Name Education type

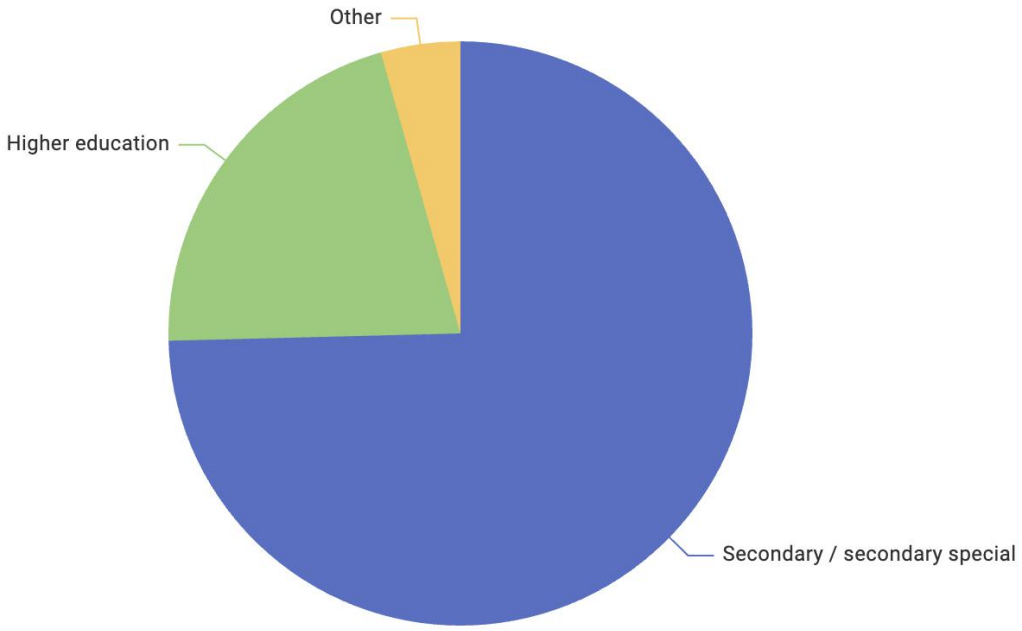


Figure 21 Pie Chart for Education Type

- The most common education type among applicants is "Secondary/secondary special," accounting for approximately 74.60% of the total.
- Higher education" is the next most common type, making up around 21% of the total.
- Other education types, such as "Incomplete higher," "Lower secondary," and "Academic degree," are less common, each representing a smaller percentage of the total.

9. Name Family Status

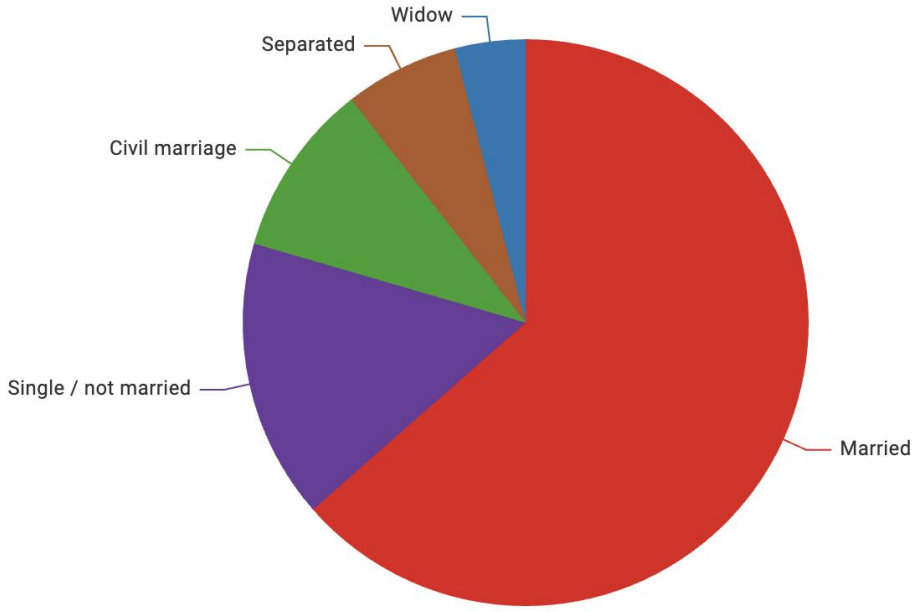


Figure 21 Pie Chart for Family Status

- The most common family status among applicants is "Married," accounting for approximately 63.53%.
- "Single / not married" is the most common family status, making up around 15.97%.
- Other family statuses, such as "Civil marriage," "Separated," "Widow," and "Unknown," are less common, each representing a smaller percentage of the total.

10. OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE

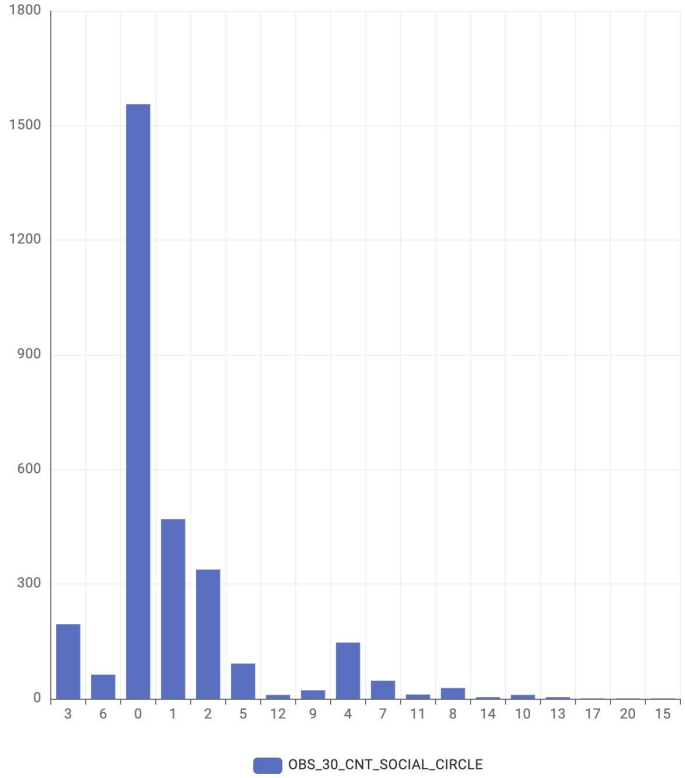


Figure 22 Histogram for OBS_30_CNT_SOCIAL_CIRCLE

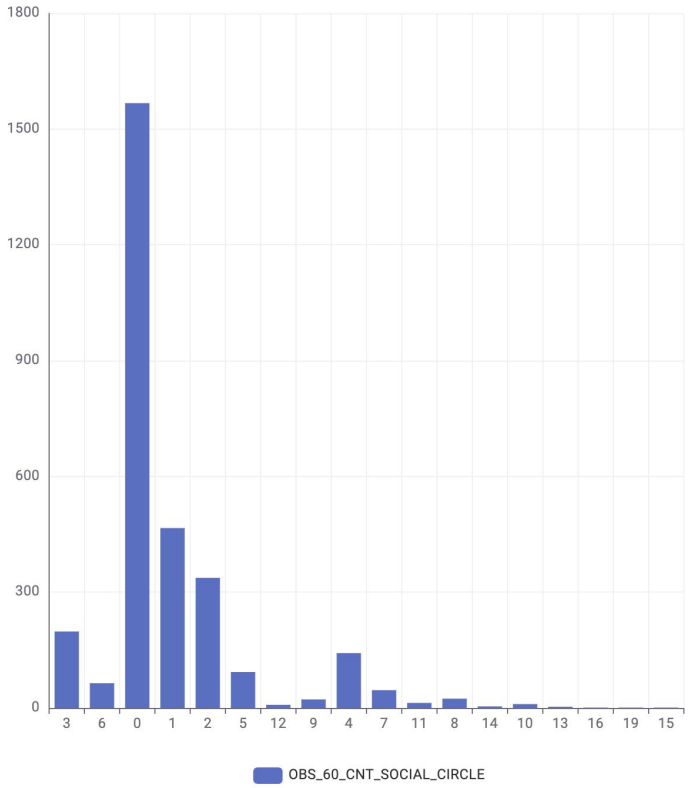


Figure 23 Histogram for OBS_60_CNT_SOCIAL_CIRCLE

- The majority of "OBS_30_CNT_SOCIAL_CIRCLE" and "OBS_60_CNT_SOCIAL_CIRCLE" attributes are equal to 0, which indicates that a significant portion of the applicants in the dataset have no observable instances of payment delays of 30/60 days or more among their social circle members.

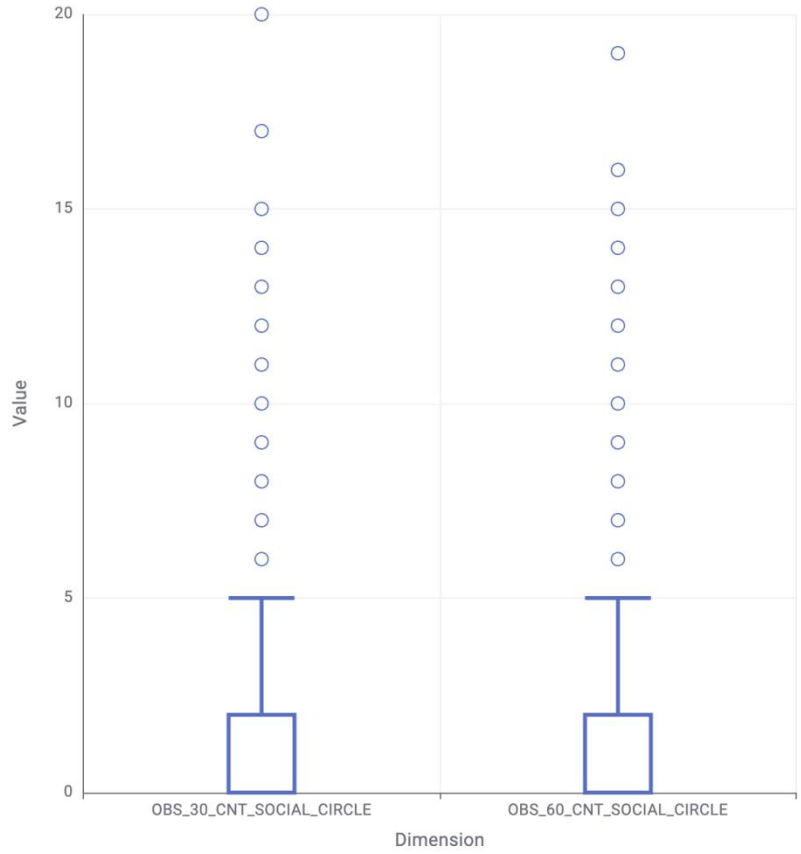


Figure 24 Box Plot for OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE

- The box plot shows the outliers of OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE. These outliers could be flagged for further investigation, as they may represent unique cases with potentially higher risk.

11. DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE

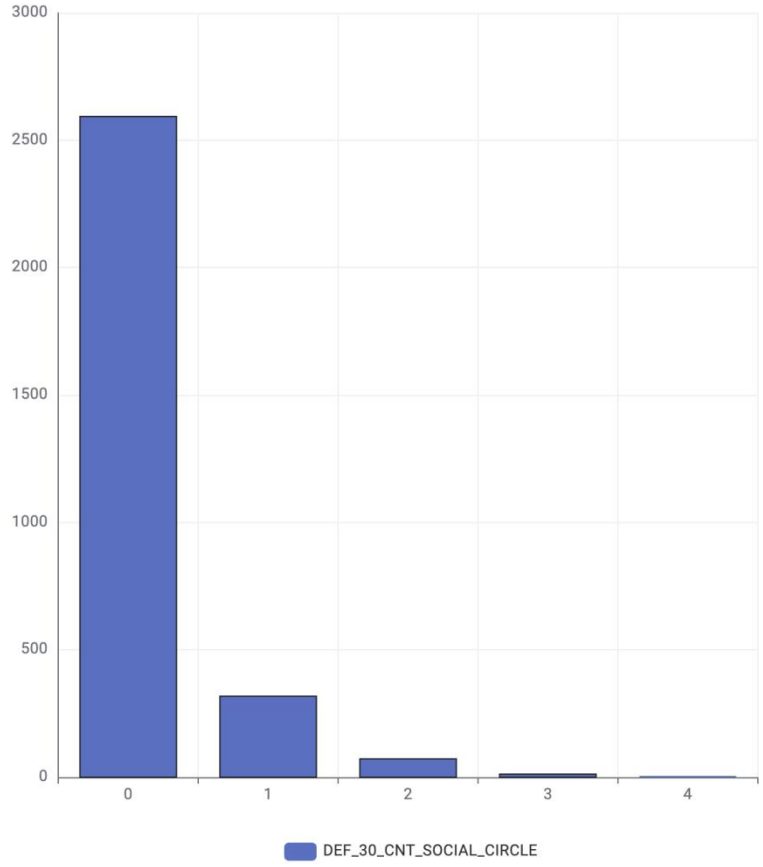


Figure 25 Histogram for DEF_30_CNT_SOCIAL_CIRCLE

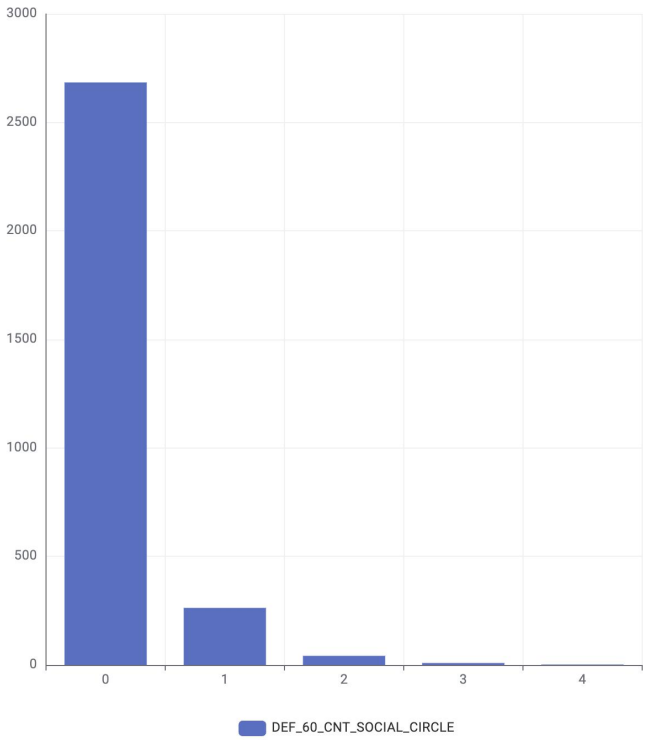


Figure 26 Histogram for DEF_60_CNT_SOCIAL_CIRCLE

- The majority of values in the "DEF_60_CNT_SOCIAL_CIRCLE" attribute are equal to 0, which indicates that a significant portion of the applicants in the dataset has no observable instances of severe payment delays (30/60 days or more) among their social circle members.

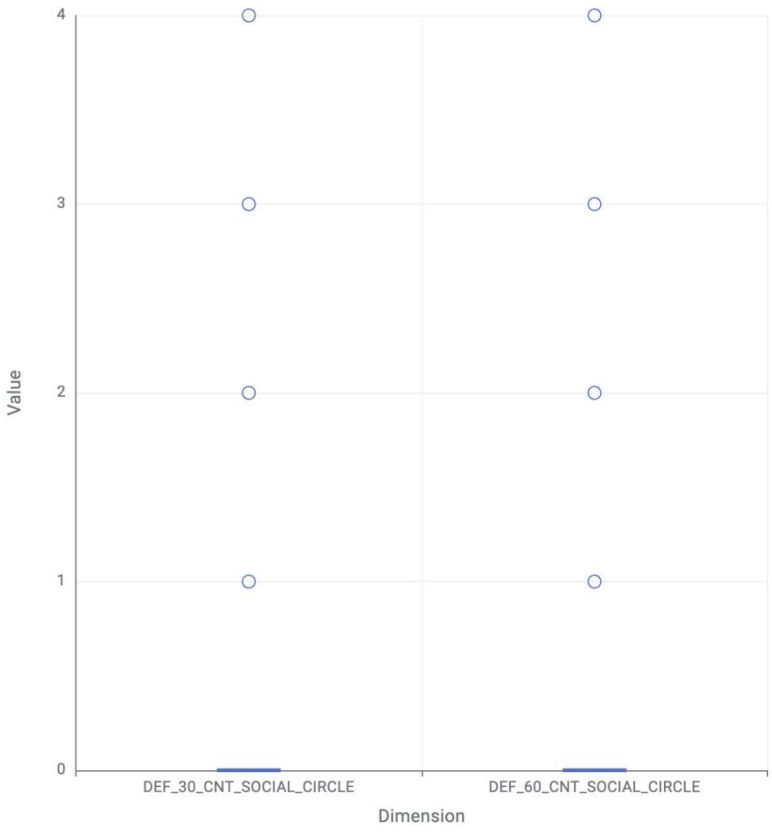


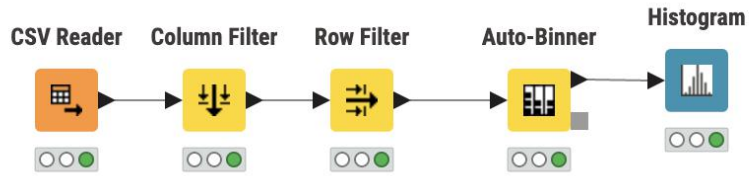
Figure 27 Box Plot for DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE

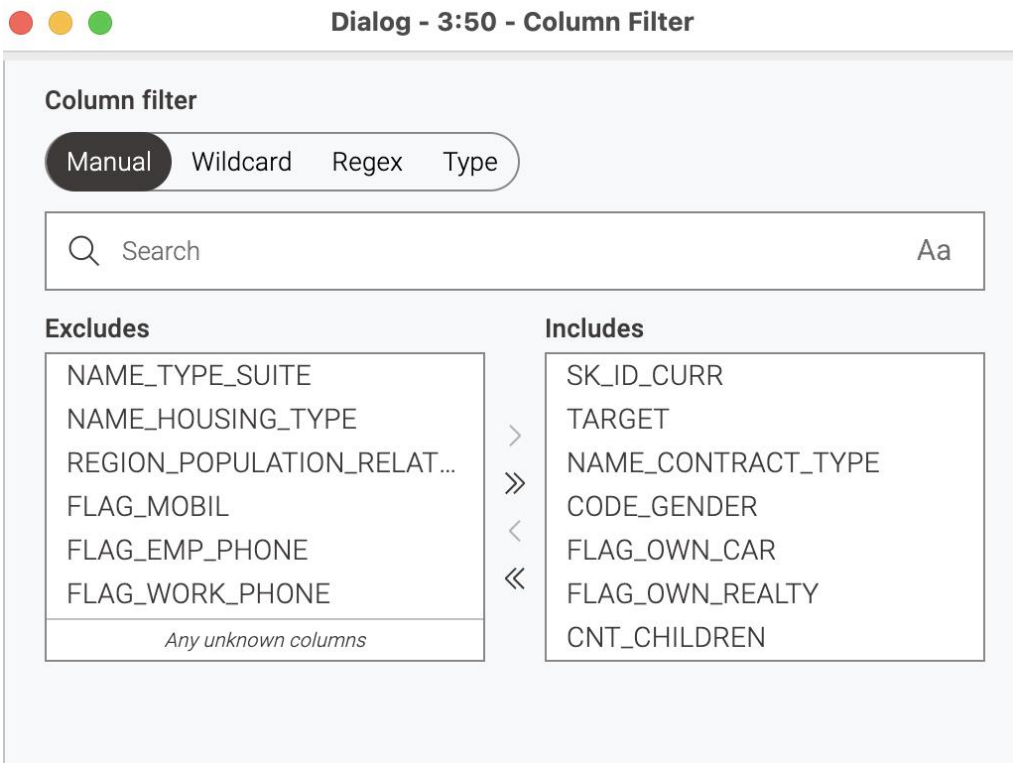
- The box plot shows the outliners of DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE. These outliers could be flagged for further investigation, as they may represent unique cases with potentially higher risk.

1B Data Preprocessing

1. Equi-depth and Equi-width Binning:

1.1.1. DAYS_EMPLOYED preprocess the data





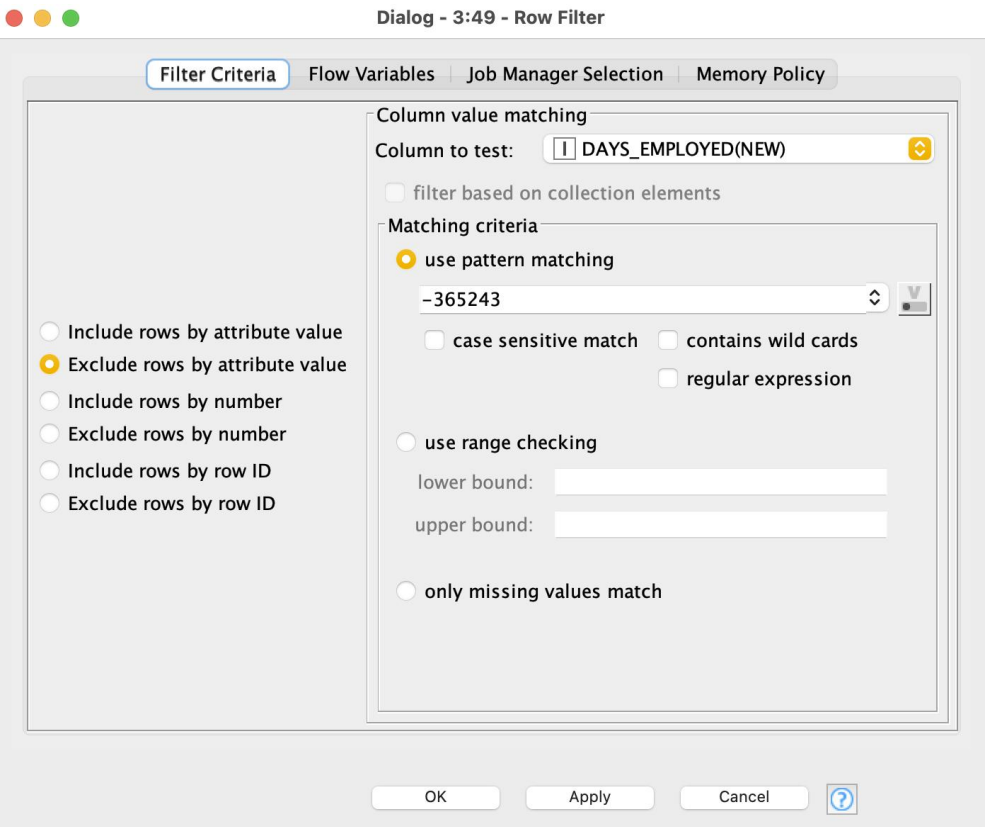


Figure 28 Steps to preprocess the Days of Employed with KNIME techniques

- The "DAYS_EMPLOYED" data numbers were preprocessed in Excel before being imported into the CSV Reader to enhance interpretability. The data is converted into positive numbers before applying binning.
- A column filter was applied to select the most critical 20 attributes, as determined by the analyst, to avoid overwhelming the table with excessive data.
- A row filter excluded rows with potentially erroneous data in the "DAYS_EMPLOYED" column. Specifically, rows containing the maximum value of "365,243" appeared in approximately 489 rows, and these values were associated with positive numbers. In contrast, other data in the same column were predominantly represented as negative numbers.

1.1.2 Equi-width binning

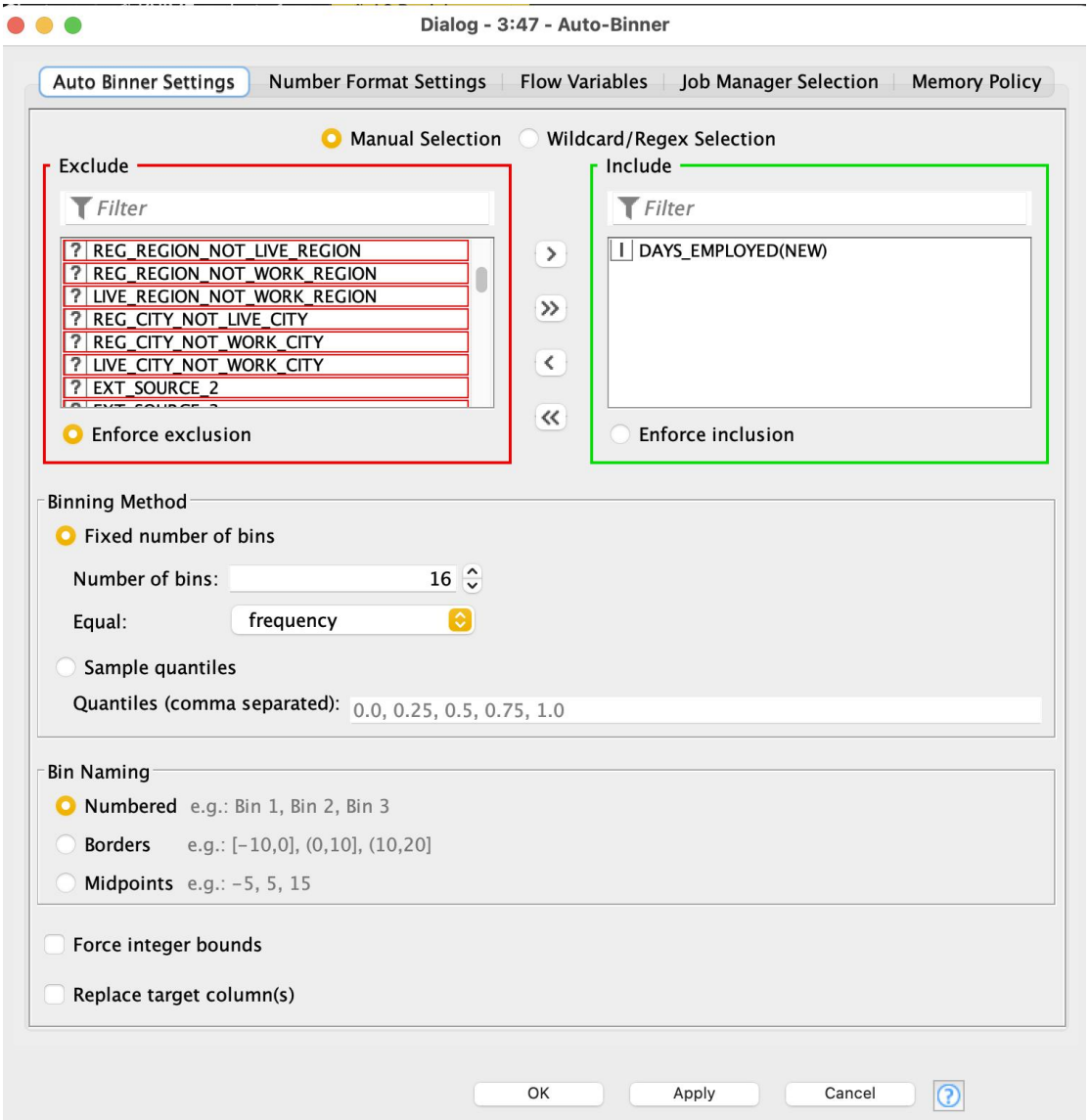


Figure 29 Equi-width binning with 16 bins

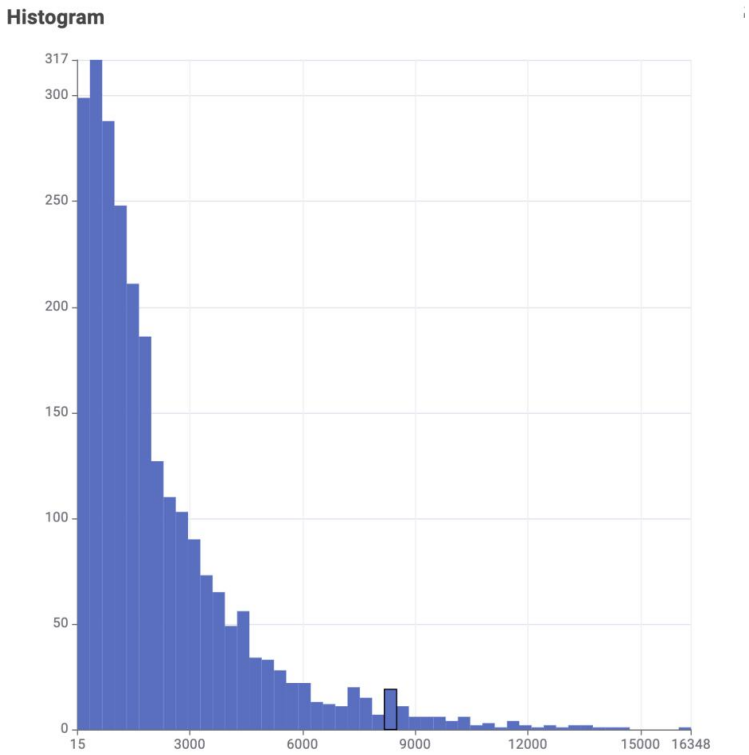


Figure 30 Histogram for 50 bins equi-width binning for DAYS_EMPLOYED

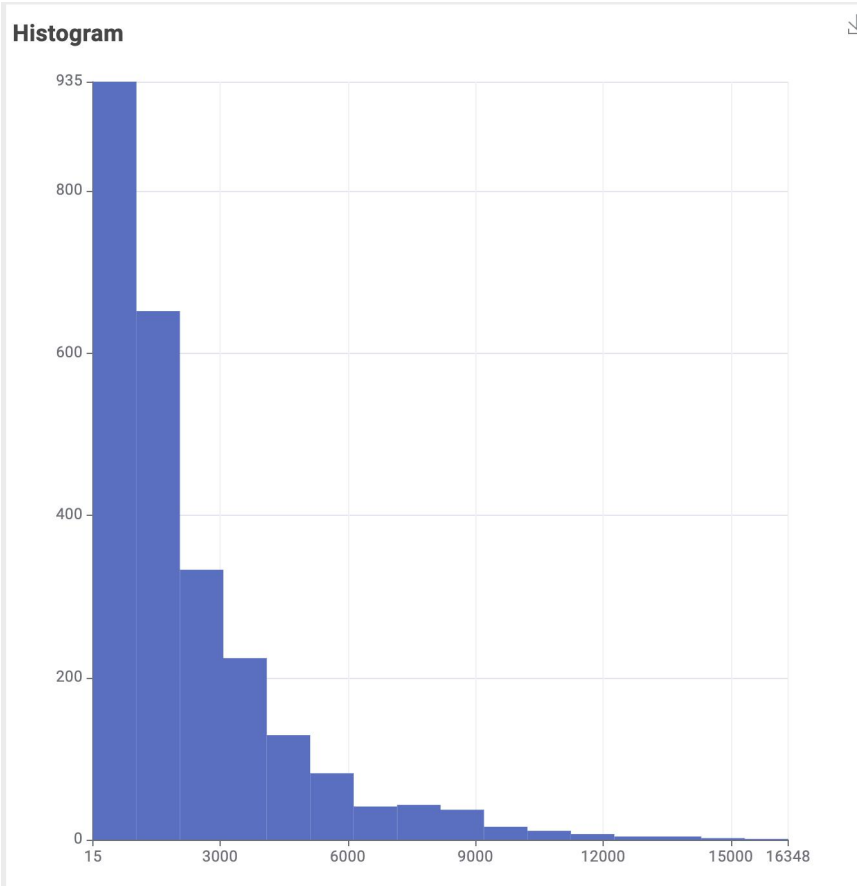


Figure 31 Histogram for 16 bins equi-width binning for DAYS_EMPLOYED

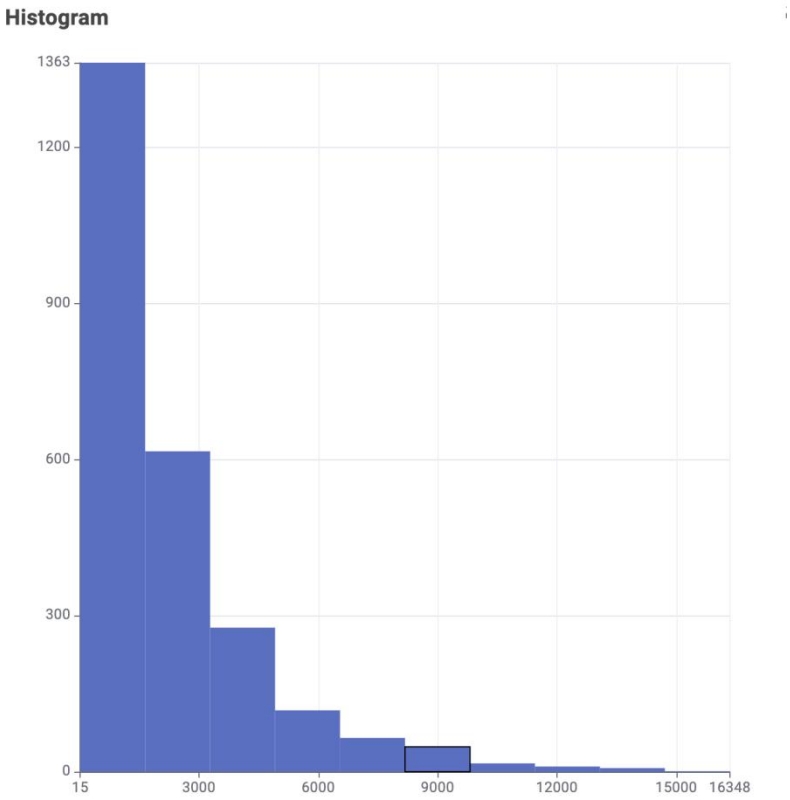


Figure 32 Histogram for ten bins equi-width binning for DAYS_EMPLOYED

- Initially, 50 bins were chosen for the histogram of the "DAYS_EMPLOYED" attribute based on the square root rule (sqrt(1,521)).
- After visualising the data with different bin counts, it was decided to use 16 bins. This choice aimed to strike a balance: it preserved the underlying patterns observed with 50 bins while reducing noise in the visualisation. Using 16 bins provided a more detailed view than a lower bin count, such as ten bins, without overcomplicating the representation.
- Additionally, selecting 16 bins per bin effectively translates to approximately 2.83 years of employment per bin. Which can be about look at three years. The choice of 16 bins was deemed suitable for the dataset because it matches the timeframes of employment durations in the data.

1.1.3. Equa-depth Binning

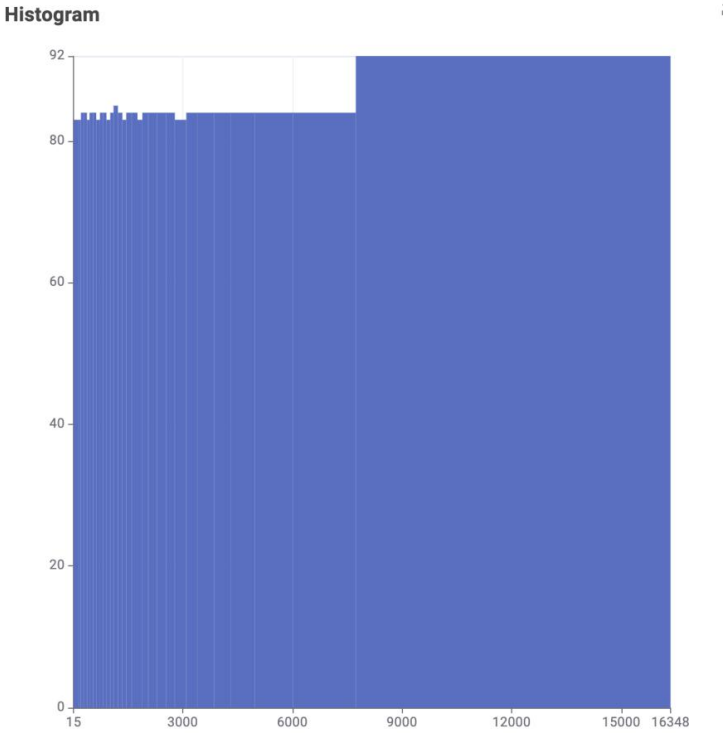


Figure 33 Equi-depth binning with 30 bins

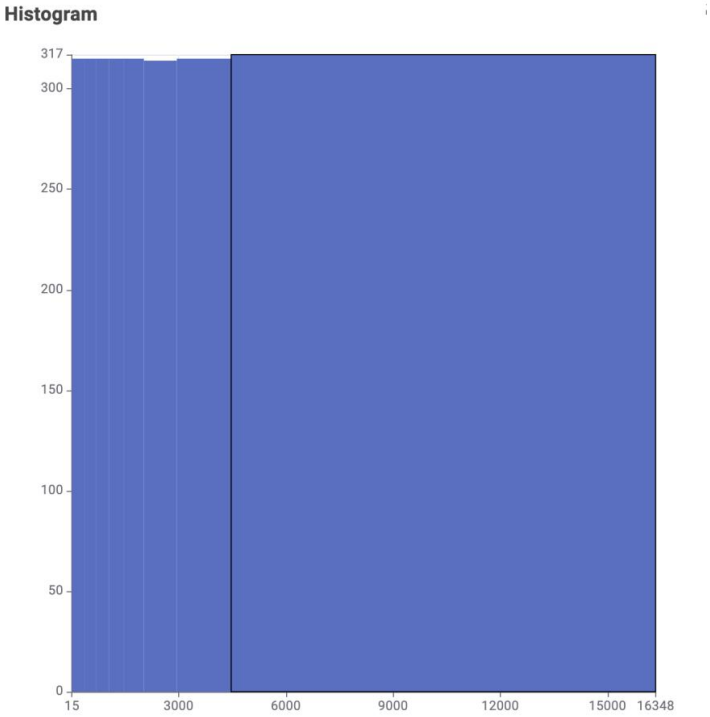


Figure 34 Equi-depth binning with ten bins

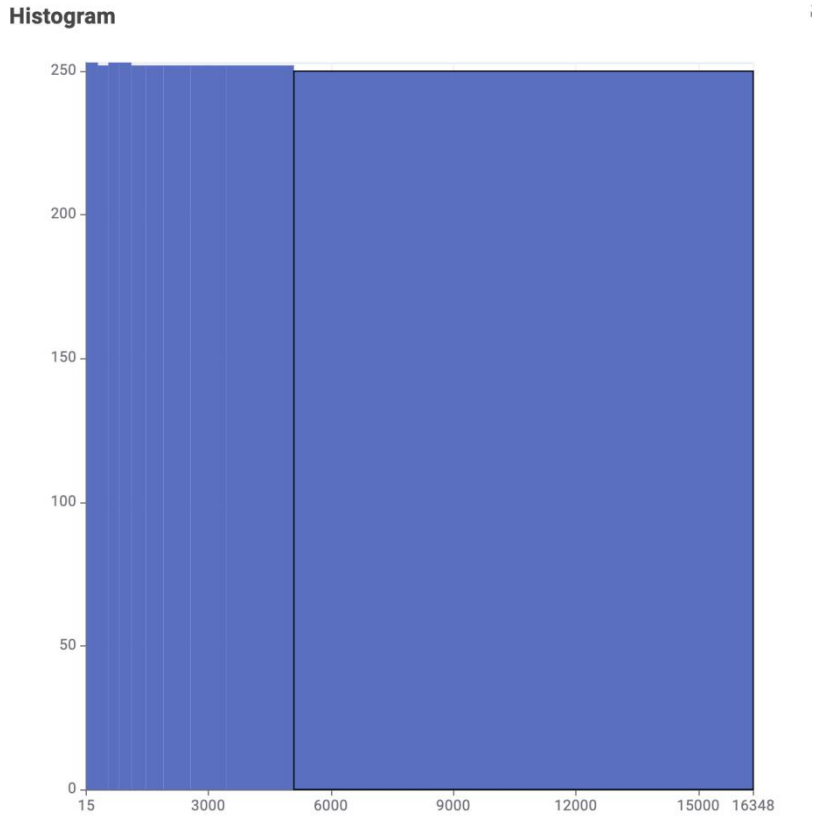


Figure 35 Equi-depth binning with eight bins

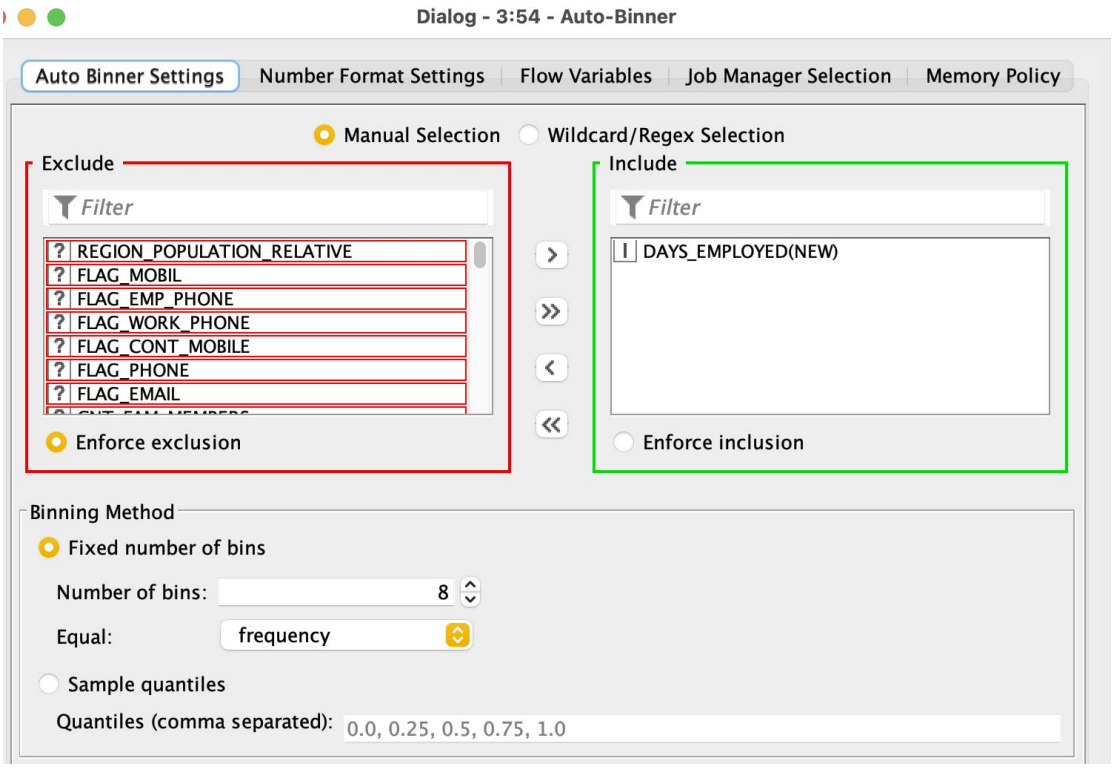


Figure 36 Equi-depth binning with eight bins

- Based on the histogram, using 30 bins for bin depth resulted in a cluttered graph where it was challenging to discern clear patterns. Reducing the bin depth to 10 bins improved visibility, but the visualisation was still difficult to interpret. Ultimately, a choice of 8 bins was made because it allowed for a more distinct depiction of varying bin depths.
- The histogram revealed that most applicants had less than three years of employment, which gradually decreased as time passed.

1.2. DAYS_ID_PUBLISHED

1.2.1. Equi-width Binning

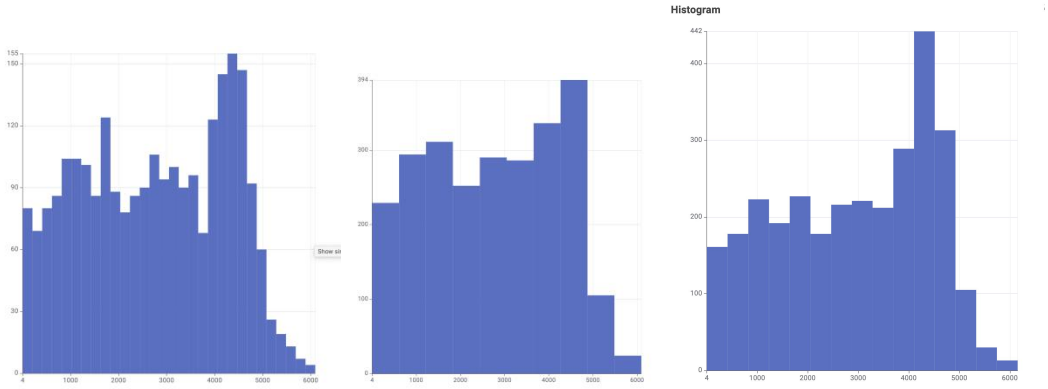


Figure 37 Histogram for Equi-width binning with 20, 10 and 15 bins.

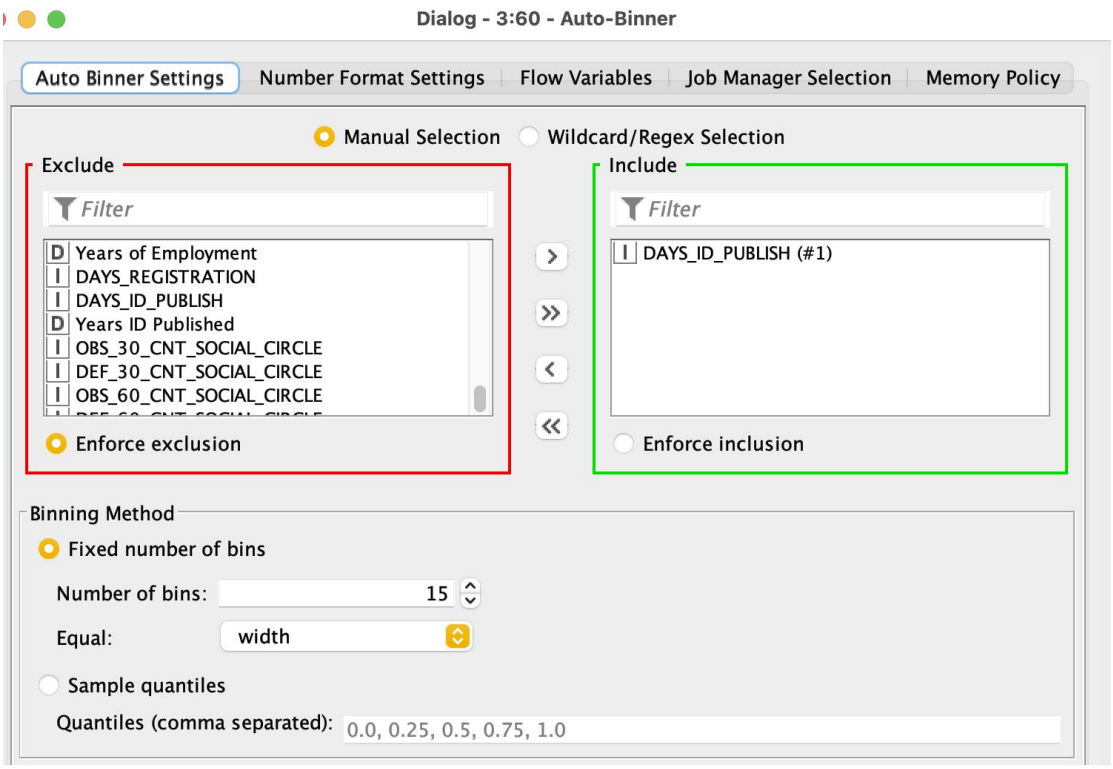


Figure 38 Equi-width binning with 15 bins

- The 'Days ID Published' data was preprocessed in Excel to convert it into positive numbers before applying binning.
- Histograms were created using different bin widths, including 30, 10, and 15 bins.
- In the 30-bin histogram, it's evident that most ID publishing durations fall within the range of 4000 to 5000 days.
- However, the data becomes overly simplified when using ten bins, and the distinctions between containers are less clear.
- Therefore, a decision was made to opt for 15 bins, as it effectively maintains the underlying pattern observed in the 30-bin histogram while reducing visual noise and avoiding excessive simplification.

1.2.2. Equi-depth binning

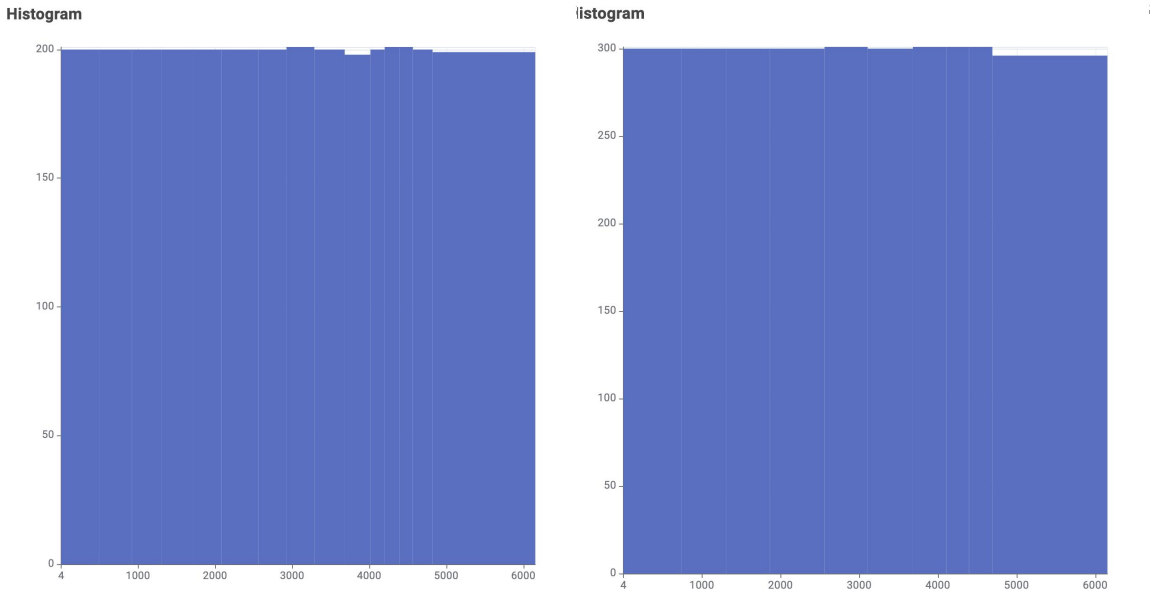


Figure 39 Histogram for Equi-depth binning with respectively 10 and 15 bins.

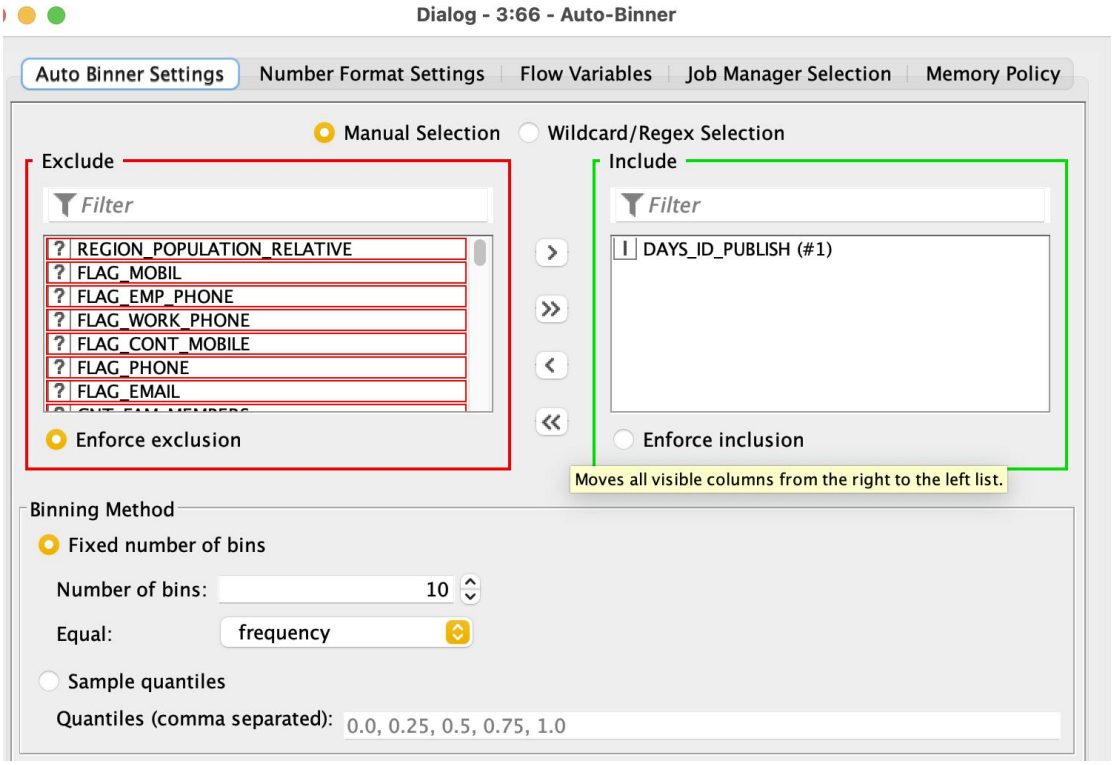


Figure 40 Equi-depth binning with ten bins

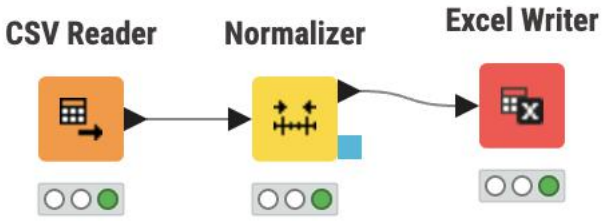
- Histograms were created using ten bins and 15 bins.
- The range of 4000 to 5000 days exhibited the highest frequency.
- The decision was made to use ten bins as it simplified the visualisation,

making it easier to interpret and extract meaningful information.

- The choice of 10 bins was practical, as it balances simplifying the visualisation and retaining essential insights.

2. Min-max and Z-score normalisation:

2.1 Normalise



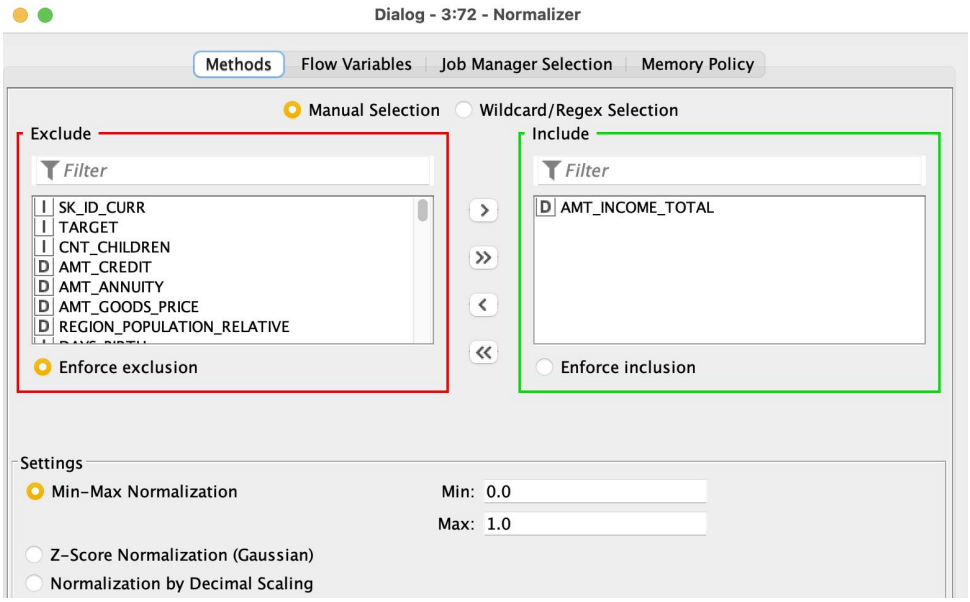


Figure 41 Steps to do Min-max Normalization

- I imported the data file using the CSV Reader.
- To normalise the 'AMT_INCOME_TOTAL' attribute, I scaled it to a range between 0 and 1.
- Normalized_value = (original_value - min_value) / (max_value - min_value) is the formula for min-max normalisation.

- The normalised data was then exported into an Excel file.

#	Row...	SK_ID_C... Number (int...	TARGET Number (int...	NAME_... String	CODE_G... String	FLAG_O... String	FLAG_O... String	C. N...	AMT_INCO... Number (double)
1227	Row...	338222	0	Revolving loa...	F	Y	Y	0	0
631	Row...	268319	1	Cash loans	F	N	Y	0	0.001
1214	Row...	384035	0	Cash loans	F	N	N	0	0.001
1029	Row...	211914	0	Cash loans	F	N	N	0	0.002
375	Row...	274230	1	Cash loans	F	N	N	1	0.003
583	Row...	152467	1	Cash loans	F	N	Y	0	0.003
2916	Row...	329680	1	Cash loans	F	N	Y	0	0.003
2951	Row...	199692	1	Cash loans	M	N	Y	0	0.003
1981	Row...	225046	1	Cash loans	F	N	N	2	0.004
1660	Row...	127310	1	Cash loans	F	N	N	0	0.004
1724	Row...	390957	0	Cash loans	F	N	N	0	0.004
2058	Row...	194412	1	Cash loans	F	N	N	0	0.004
2713	Row...	216915	1	Cash loans	M	Y	N	0	0.004

Figure 42 Ascending order of normalise Amount Total Income with a minimum of 0.

#	Row...	SK_ID_C... Number (int...	TARGET Number (int...	NAME_... String	CODE_G... String	FLAG_O... String	FLAG_O... String	C. N...	AMT_INCO... Number (double)
1616	Row...	387126	1	Cash loans	F	Y	Y	1	1
151	Row...	134013	0	Cash loans	F	N	N	0	0.424
2370	Row...	144317	1	Cash loans	F	N	N	0	0.352
1707	Row...	133753	0	Revolving loa...	F	N	Y	0	0.28
339	Row...	291492	0	Revolving loa...	M	Y	Y	4	0.236
1267	Row...	324099	0	Cash loans	F	Y	Y	0	0.231
577	Row...	114321	0	Cash loans	M	Y	N	0	0.207
627	Row...	180915	1	Cash loans	M	N	Y	2	0.207
1102	Row...	183027	1	Cash loans	M	N	Y	2	0.207
1490	Row...	117605	1	Cash loans	F	N	Y	0	0.207
1625	Row...	391254	0	Cash loans	M	Y	Y	0	0.207
1962	Row...	287933	0	Revolving loa...	M	N	N	1	0.207
2206	Row...	303156	0	Cash loans	M	Y	Y	0	0.207

Figure 43 Descending order of normalise Amount Total Income with a maximum of 1.

- The normalisation of 'AMT_INCOME_TOTAL,' with a maximum value of 1, reveals a significant gap between the highest income level and the second highest, which stands at 0.424. This disparity indicates a substantial income difference between the top earners and the next tier.
- This notable income gap requires extra attention to avoid potential bias.

2.2 Z-score Normalization

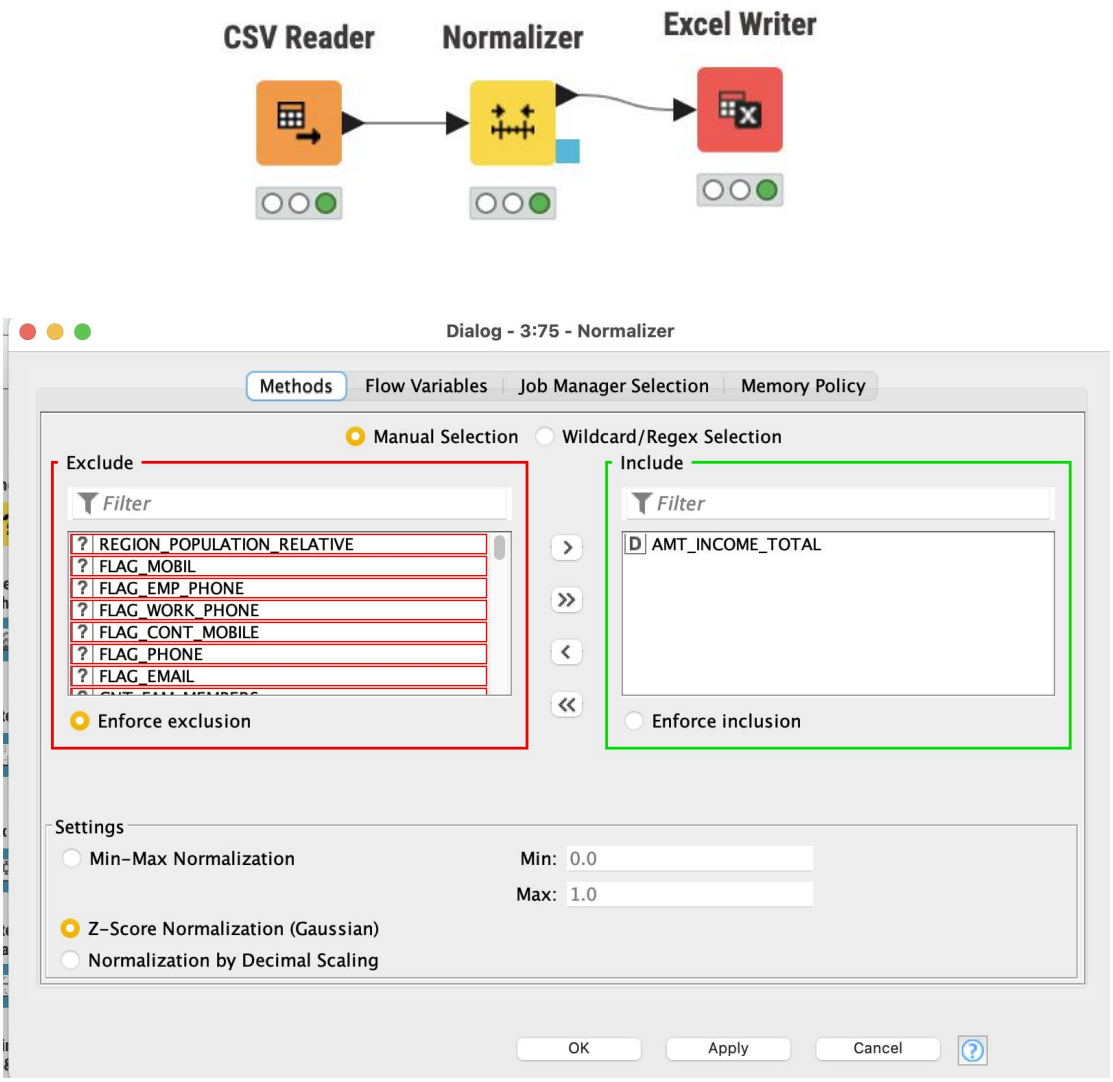


Figure 44 Steps to do Z-score normalisation

- I imported the data file using the CSV Reader.
- To z-score normalise the 'AMT_INCOME_TOTAL' attribute, I selected the Z-score normalisation option to configure a normalisation node window. (Figure 44)
- The Z-Score normalised data was then exported into an Excel file.

#	Row...	SK_ID_C... Number (int...	TARGET Number (int...	NAME_... String	CODE_G... String	FLAG_O... String	FLAG_O... String	CNT_C... Number (int...	AMT_INCOM... Number (double)
1227	Row...	338222	0	Revolving loa...	F	Y	Y	0	-1.312
631	Row...	268319	1	Cash loans	F	N	Y	0	-1.282
1214	Row...	384035	0	Cash loans	F	N	N	0	-1.27
1029	Row...	211914	0	Cash loans	F	N	N	0	-1.248
375	Row...	274230	1	Cash loans	F	N	N	1	-1.227
583	Row...	152467	1	Cash loans	F	N	Y	0	-1.227
2916	Row...	329680	1	Cash loans	F	N	Y	0	-1.223
2951	Row...	199692	1	Cash loans	M	N	Y	0	-1.21
1981	Row...	225046	1	Cash loans	F	N	N	2	-1.205
1660	Row...	127310	1	Cash loans	F	N	N	0	-1.184
1724	Row...	390957	0	Cash loans	F	N	N	0	-1.184
2058	Row...	194412	1	Cash loans	F	N	N	0	-1.184
2713	Row...	216915	1	Cash loans	M	Y	N	0	-1.184

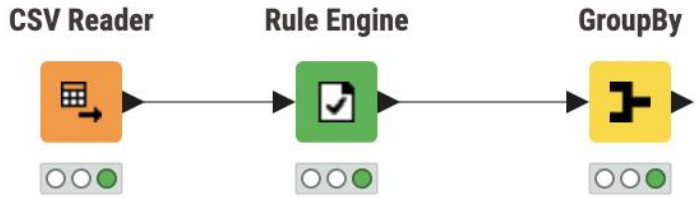
Figure 45 Ascending order of Z-Score Normalisation

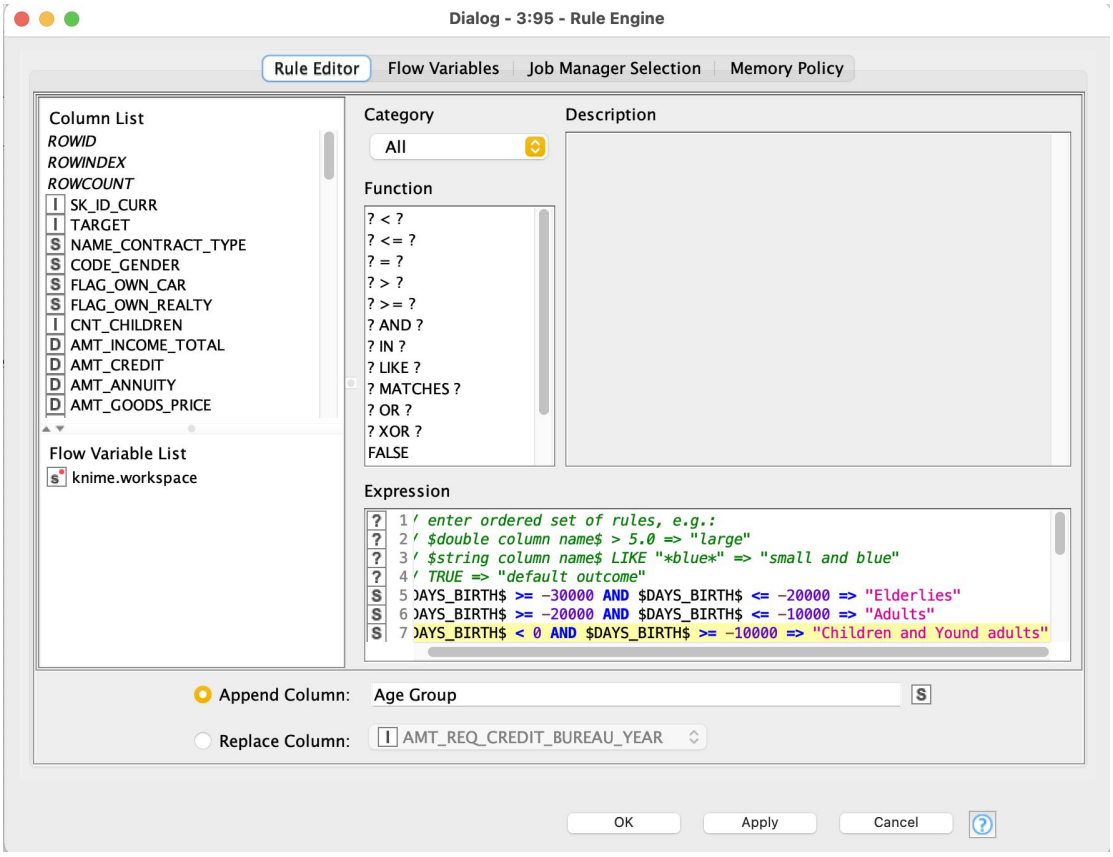
#	Row...	SK_ID_C... Number (int...	TARGET Number (int...	NAME_... String	CODE_G... String	FLAG_O... String	FLAG_O... String	CNT_C... Number (int...	AM... Number (do...	AMT_C... Number (do...
606	Row...	403497	0	Cash loans	F	N	N	0	-0.01	263,686.5
1758	Row...	410614	1	Cash loans	M	N	N	0	-0.01	521,280
1374	Row...	258069	1	Cash loans	F	N	N	0	-0.014	1,329,579
131	Row...	271257	0	Cash loans	F	N	Y	0	-0.031	922,666.5
485	Row...	356028	0	Cash loans	F	Y	Y	1	-0.031	528,633
486	Row...	311894	1	Cash loans	F	N	N	0	-0.031	1,293,502.5
833	Row...	150832	1	Cash loans	F	N	N	0	-0.031	900,000
1269	Row...	119793	1	Cash loans	M	Y	Y	0	-0.031	541,323
1466	Row...	381629	0	Cash loans	M	N	Y	0	-0.031	540,000
1558	Row...	277730	0	Cash loans	F	N	Y	0	-0.031	792,346.5
1564	Row...	141912	0	Cash loans	M	Y	N	1	-0.031	900,000
1624	Row...	100190	0	Cash loans	M	Y	N	0	-0.031	263,686.5
1704	Row...	355295	0	Revolving loa...	F	Y	Y	0	-0.031	135,000
1765	Row...	273082	0	Cash loans	M	N	Y	0	-0.031	463,284
2114	Row...	384836	0	Cash loans	F	N	Y	1	-0.031	753,840
2236	Row...	232507	1	Cash loans	F	N	Y	0	-0.031	640,080
2340	Row...	100627	0	Cash loans	F	N	Y	0	-0.031	874,152

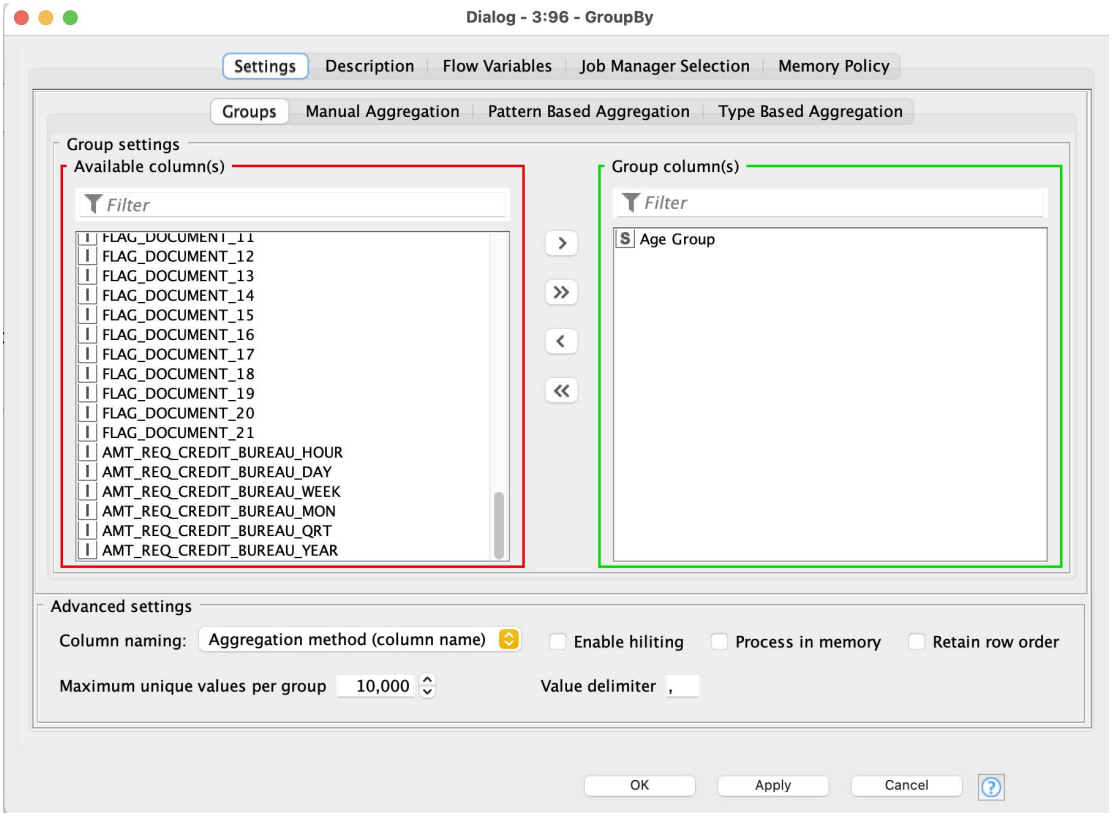
Figure 46 The nearest point to the mean

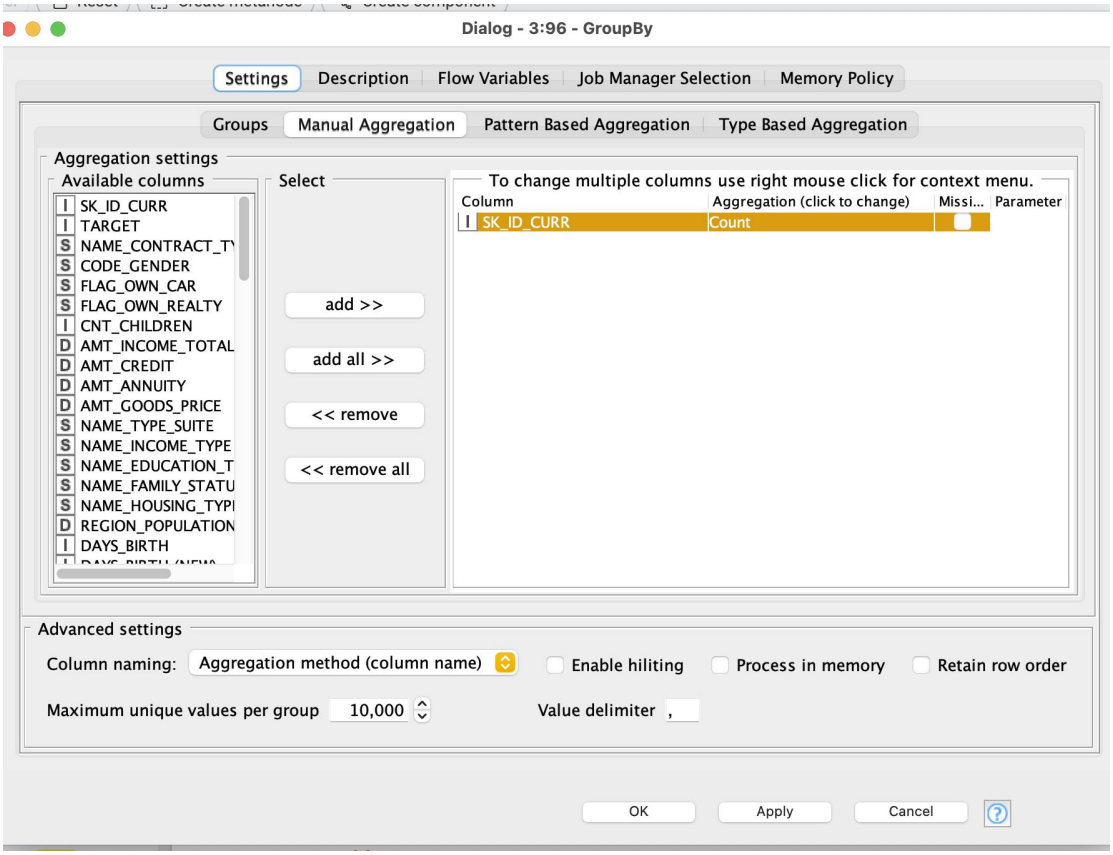
- After applying the Z-score transformation, the lowest value corresponds to a Z-score of -1.312, indicating its position below the mean. (Figure 45)
- The Z-score values are centred around 0, with 0 being the exact point representing the mean income.
- The nearest data point to this mean, representing a Z-score close to 0, is approximately -0.01, indicating its proximity to the crest of the distribution. (Figure 46)

3. Discretization






Dialog - 3:96 - GroupBy
Settings
Description
Flow Variables
Job Manager Selection
Memory Policy
Groups
Manual Aggregation
Pattern Based Aggregation
Type Based Aggregation
Group settings
Available column(s)
Filter
FLAG_DOCUMENT_12
FLAG_DOCUMENT_13
FLAG_DOCUMENT_14
FLAG_DOCUMENT_15
FLAG_DOCUMENT_16
FLAG_DOCUMENT_17
FLAG_DOCUMENT_18
FLAG_DOCUMENT_19
FLAG_DOCUMENT_20
FLAG_DOCUMENT_21
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
Group column(s)
Filter
Age Group
Advanced settings
Column naming:
Aggregation method (column name)
Enable hiliting
Process in memory
Retain row order
Maximum unique values per group
10,000
Value delimiter
,
OK
Apply
Cancel


Dialog - 3:96 - GroupBy
Settings
Description
Flow Variables
Job Manager Selection
Memory Policy
Groups
Manual Aggregation
Pattern Based Aggregation
Type Based Aggregation
Aggregation settings
Available columns
SK_ID_CURR
TARGET
NAME_CONTRACT_T
CODE_GENDER
FLAG_OWN_CAR
FLAG_OWN_REALTY
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
NAME_TYPE_SUITE
NAME_INCOME_TYPE
NAME_EDUCATION_T
NAME_FAMILY_STATU
NAME_HOUSING_TYPI
REGION_POPULATION
DAYS_BIRTH
Select
add >>
add all >>
<< remove
<< remove all
To change multiple columns use right mouse click for context menu.
Column
Aggregation (click to change)
Missi...
Parameter
SK_ID_CURR
Count
Advanced settings
Column naming:
Aggregation method (column name)
Enable hiliting
Process in memory
Retain row order
Maximum unique values per group
10,000
Value delimiter
,
OK
Apply
Cancel

Rows: 3 | Columns: 2 Table Statistics

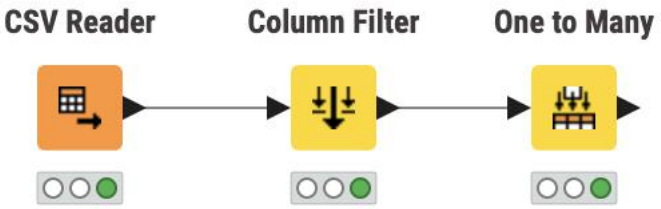
#	Row...	Age Group String	Count*(SK_ID_CURR) Number (integer)
1	Row0	Adults	2134
2	Row1	Children and Yound adults	276
3	Row2	Elderlies	590

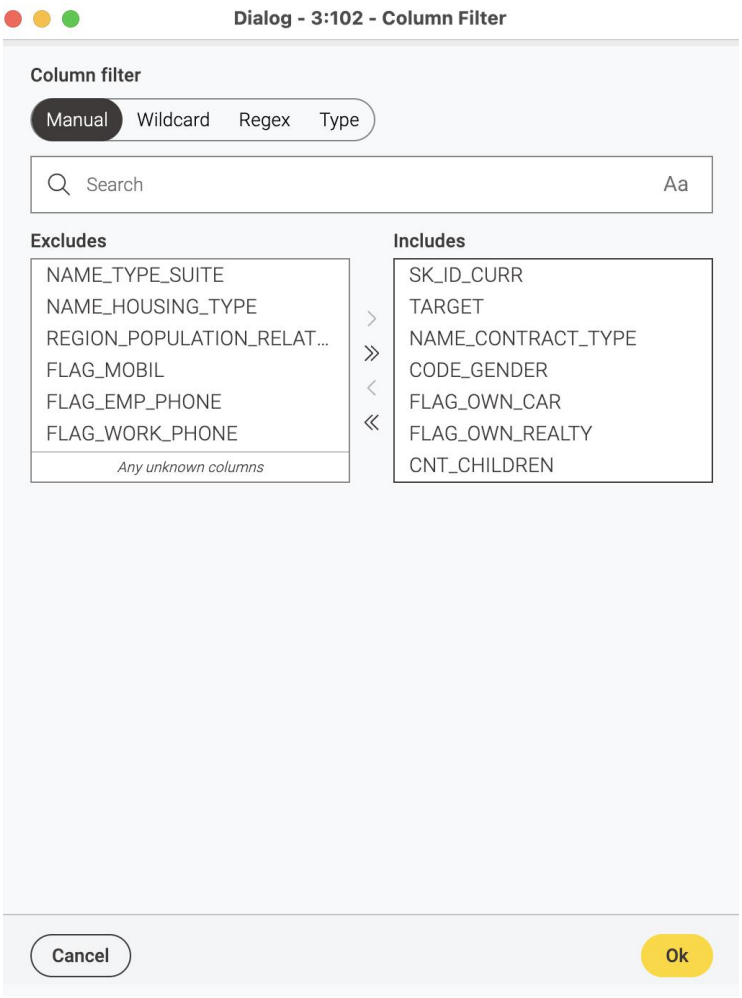
Figure 47 Discretise age and look for the frequency

The process involved in discretising the birthdays into categories and obtaining the frequency count for each Day of Birth is illustrated in the above images. The following steps were followed:

1. I connected the imported file CSV Reader node to the "Rule Engine" node in KNIME.
2. Configured expressions for age groups:
   $DAYS_BIRTH$ >= -30000 AND $DAYS_BIRTH$ <= -20000 => "Elderlies"
   $DAYS_BIRTH$ >= -20000 AND $DAYS_BIRTH$ <= -10000 => "Adults"
   $DAYS_BIRTH$ < 0 AND $DAYS_BIRTH$ >= -10000 => "Children and Young adults"
3. Labeled the column as "Age Group."
4. I connected the "Group By" node to find the frequency count.
5. I selected the age group as the group column.
6. Used the manual aggregation method with customer ID as the column and count as the aggregation method.
7. This process displayed the number of times each age group appeared in the dataset.

4. Binarise the CODE GENDER attribute:





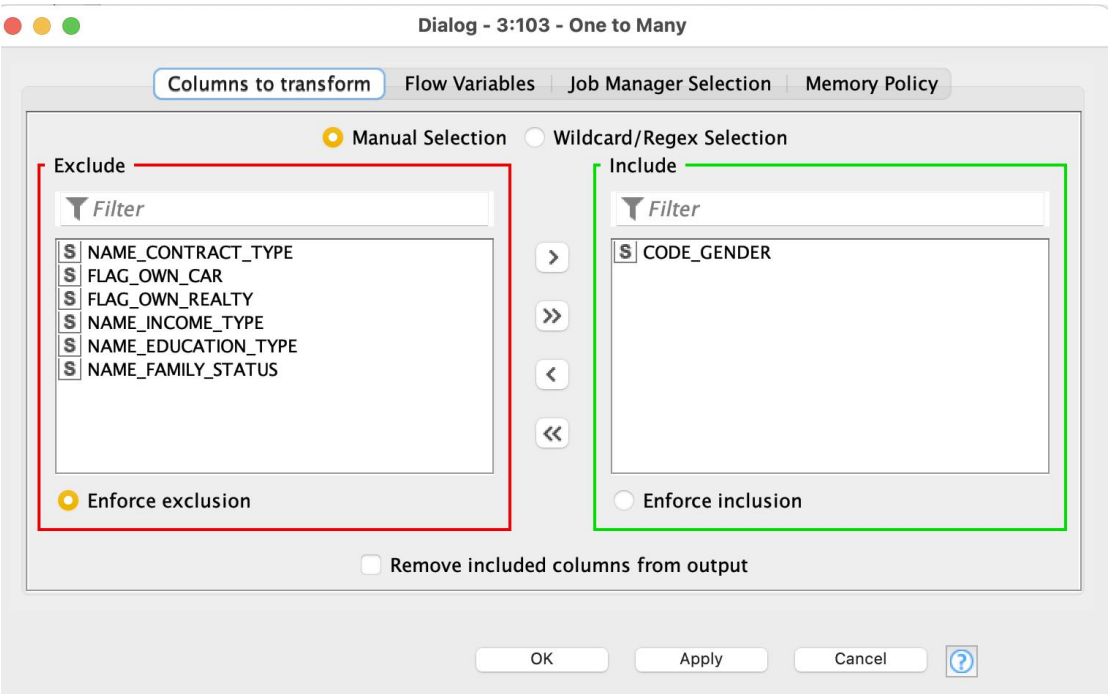


Figure 48 Steps of binarise CODE GENDER

To binarise the CODE GENDER, the above image illustrates the following steps:

1. Connected the imported file CSV Reader node to the "Column Filter" node in KNIME to filter out the selected 20 attributes.
2. Connected the node to the “One to Many” node.
3. Selected CODE GENDER for inclusion in the process.
4. Two columns were created, Column F and Column M.
5. In Column F, 1 represents Female, and 0 illustrates Male. Vice versa for Column M.

1C. Summary

During the dataset analysis process, I discovered some critical insights that the Analytics Unit should pay attention to. Firstly, from the applicant's total income ('AMT_INCOME_TOTAL') attribute, I have found out that there is a big gap, especially among those who earn the most. The substantial income gap among applicants might raise concern for potential financial stability and repayment capacity disparities, as significant income gaps may indicate varying financial risk. The income gap in the dataset should be further investigated for how the income gap affects the decision on who to give loans to.

Furthermore, the amounts of credit applicants apply for ('AMT_CREADIT'), most applicants are used within the specific credit range. This should be further investigated, and the reason should be found so that the lending strategies can be customised accordingly.
Moreover, applicants' social circle (attributes like OBS_30_CNT_SOCIAL_CIRCLE) have shown some unusual cases. This needs to be investigated further to understand how the social process of applicants affects the ability of the applicant to repay the loan.

Next, the age and employment duration ('BAYS_BIRTH' and 'DAYS_EMPLOYED') have indicated that many applicants have been employed for less than five years, and more than half of the applicants are between 28 and 50. Breaking down the employment duration and period into categories might give us valuable insights while analysing the data.

From the gender (CODE_GENDER) attribute, we have discovered that female applicants have a slightly higher percentage than male applicants. This might lead to a possibility that we might need to consider addressing the risk in a way specific to gender.

In conclusion, throughout my findings, it is suggested that we should look closely at the income gaps, understand the reason applicants apply for a certain amount of credit, investigate the outliers of applicants' social circles, consider the category of duration of employment and age, and think about gender-specific risk assessment. Doing this might help us make better decisions about loans and risks in the future.

Acknowledgment:

The authors would like to express their gratitude to OpenAI for providing access to ChatGPT version 3.5, a powerful language model that significantly contributed to the preparation of this paper. ChatGPT 3.5 played a crucial role in various aspects of the article, including answering factual questions by leveraging information available on the internet, aiding in the drafting and structuring of ideas, generating suggestions for graphics and visuals, critically analysing written content for validity, refining grammar and writing structure, experimenting with diverse writing styles, and overcoming writer's block.

The collaboration with ChatGPT 3.5 proved instrumental in enhancing the overall quality and efficiency of the research and writing process. The sections of the paper that benefited from AI assistance include the data exploration and preprocessing steps, statistical analysis interpretations, and the generation of concise summaries. The ability of ChatGPT 3.5 to comprehend complex information and provide coherent and contextually relevant responses greatly facilitated the creation of a more robust and insightful paper.

The authors acknowledge the AI tool's valuable contributions, recognising it as a supportive and innovative resource in pursuing academic excellence.