UTS: ENGINEERING & INFORMATION TECHNOLOGY

SUBJECT NUMBER & NAME	**NAME OF STUDENT(s)** (PRINT CLEARLY)	**STUDENT ID(s)**
32130 Fundamentals of Data Analytics - Spring 2023	Aishwarya Rajesh Panhale	24587976

STUDENT EMAIL	**STUDENT CONTACT NUMBER**
aishwaryarajesh.panhale@student.uts.edu.au	0466394417

ASSESSMENT ITEM NUMBER & TITLE

Assignment 2
Data exploration and preparation

☐ I confirm that I have read, understood and followed the guidelines for assignment submission and presentation on page 2 of this cover sheet.
☐ I confirm that I have read, understood and followed the advice in the Subject Outline about assessment requirements.
☐ I understand that if this assignment is submitted after the due date it may incur a penalty for lateness unless I have previously had an extension of time approved and have attached the written confirmation of this extension.

Declaration of originality: The work contained in this assignment, other than that specifically attributed to another source, is that of the author(s) and has not been previously submitted for assessment. I understand that, should this declaration be found to be false, disciplinary action could be taken and penalties imposed in accordance with University policy and rules. In the statement below, I have indicated the extent to which I have collaborated with others, whom I have named.

Statement of collaboration:

Signature of student(s)_____ARP_____ Date _____29/09/2023_____

Table of Contents

Table of Figures

Introduction

This report includes the practical experience in data visualisation, exploration, and preparation conducted on the dataset assigned. The dataset consists of 3000 instances and 72 attributes.

Task : Data Exploration

Attribute Types

- Nominal : Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values.
- Ordinal: Attributes consist of value that can be ordered but cannot be differentiated between values.
- Interval: Attributes consist of numeric value that can be differentiated to obtain outcome but cannot have true zero value.
- Ratio: Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .

Identifying Attributes

N NO.	**Attribute**	**Type**	**Description**	**Justification**
1	SK_ID_CURR	Nominal	Customer ID for clients application for loan	It consists of customer ID which is unique.
2	TARGET	Nominal	It labels client application of being able to repay the loan where 0=No and 1=Yes	Variable for loan approval for clients application .
3	NAME_CONTRACT_TYPE	Nominal	Contract loan type	Consist of category for loan application
4	CODE_GENDER	Nominal	Gender of client applicating for loan where F=Female and M=Male	Consist of category of client gender

5	FLAG_OWN_CAR	Nominal	It includes categorical value for client owning car or not where 0=No and 1=Yes	Consist of 0 or 1 value for client owning car (1) or not (0)
6	FLAG_OWN_REALTY	Nominal	It includes categorical value for client owning Realty or not where 0=No and 1=Yes	Consist of 0 or 1 value for client owning Realty (1) or not (0)
7	CNT_CHILDREN	Ordinal	Count of children an applicant has at the time of application	Attributes consist of value that can be ordered but cannot be differentiated between values.
8	AMT_INCOME_TOTAL	Ratio	Income total has value that can ordered based on income of client higher or lower during application and it consist of numeric value which can include zero value	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value ..
9	AMT_CREDIT	Ratio	Credit amount includes value for client with higher credit which can be ordered for predicting risk of loan approval	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
10	AMT_ANNUITY	Ratio	Annuity includes value of client annuity based on their income, credit	Attributes includes value that can be ordered, differentiated to obtain outcome

			and income type.	and also it can be multiplied and divided which can include true zero value .
11	AMT_GOODS_PRICE	Ratio	Good price values can be differentiated based on application of client and can consist true zero value.	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
12	NAME_TYPE_SUITE	Nominal	It consists distinct value if someone accompanied client during application	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values.
13	NAME_INCOME_TYPE	Nominal	Clients income category includes businessman, working, maternity leave	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values. .
14	NAME_EDUCATION_TYPE	Nominal	Category of education client has finished.	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values

15	NAME_FAMILY_STATUS	Nominal	Marital status of an applicant	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values.
16	NAME_HOUSING_TYPE	Nominal	Housing situation of the client in categories.	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values.
17	REGION_POPULATION_RELATIVE	Ratio	It consists of population based on region	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
18	DAYS_BIRTH	Interval	Client age in days at the time of application	Attributes consist of numeric value that can be differentiated to obtain outcome but cannot have true zero value.
19	DAYS_EMPLOYED	Interval	Client employment in days at the time of application	Attributes consist of numeric value that can be differentiated to obtain outcome but cannot have true zero value.

20	DAYS_REGISTRATION	Interval	Number of days for registration of client	Attributes consist of numeric value that can be differentiated to obtain outcome but cannot have true zero value.
21	DAYS_ID_PUBLISH	Interval	Number of days client published any ID since the time of application	Attributes consist of numeric value that can be differentiated to obtain outcome but cannot have true zero value.
22	FLAG_MOBIL	Nominal	Did the client provide a mobile phone (0=No, 1=Yes)	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
23	FLAG_EMP_PHONE	Nominal	Value includes if the client provide an employer phone (0=No, 1=Yes)	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
24	FLAG_WORK_PHONE	Nominal	Value includes if the client provide a work phone (0=No, 1=Yes)	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
25	FLAG_CONT_MOBILE	Nominal	Values include if the client mobile phone was reachable (0=No, 1=Yes)	Attributes consist of categorical data which are distinct and cannot be ordered or

				differentiated between values
26	FLAG_PHONE	Nominal	Value includes if client provide a home phone (0=No, 1=Yes)	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
27	FLAG_EMAIL	Nominal	Value consist if the client provide an email (0=No, 1=Yes)	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
28	CNT_FAM_MEMBERS	Ratio	Consist of number of family members of client	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
29	REGION_RATING_CLIENT	Ordinal	This attribute consists of region rating as 1,2, 3	Attributes consist of value that can be ordered but cannot be differentiated between values.
30	REGION_RATING_CLIENT_W_CITY	Ordinal	This attribute consists of region rating for city as 1,2, 3	Attributes consist of value that can be ordered but cannot be differentiated between values.

31	WEEKDAY_APPR_PROCESS_START	Nominal	Category based on days of the week	Categorical variable representing application day.
32	HOUR_APPR_PROCESS_START	Interval	Hour when client started the application process	Attributes consist of numeric value that can be differentiated to obtain outcome but cannot have true zero value.
33	REG_REGION_NOT_LIVE_REGION	Nominal	Values contains client's permanent address does not match contact address at region level 0=Same, 1=Different	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
34	REG_REGION_NOT_WORK_REGION	Nominal	Values contains client's permanent address does not match work address at region level (0=Same, 1=Different)	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
35	LIVE_REGION_NOT_WORK_REGION	Nominal	Values contains client's contact address does not match work address at region level (0=Same, 1=Different)	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
36	REG_CITY_NOT_LIVE_CITY	Nominal	Values contains client's permanent address does not match contact address at city level (0=Same, 1=Different)	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values

37	REG_CITY_NOT_WORK_CITY	Nominal	Values contains client's permanent address does not match work address at city level (0=Same, 1=Different)	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
38	LIVE_CITY_NOT_WORK_CITY	Nominal	Values contains client's contact address does not match work address at city level (0=Same, 1=Different)	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
39	ORGANIZATION_TYPE	Nominal	Category of client employment organization	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values
40	EXT_SOURCE_2	Ratio	External data source 2 values	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
41	EXT_SOURCE_3	Ratio	External data source 3 values	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can

				include true zero value .
42	OBS_30_CNT_SOCIAL_CIRCLE	Ratio	Observation in client social circle	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
43	DEF_30_CNT_SOCIAL_CIRCLE	Ratio	Defaulters in client social circle	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
44	OBS_60_CNT_SOCIAL_CIRCLE	Ratio	Observation in client social circle	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
45	DEF_60_CNT_SOCIAL_CIRCLE	Ratio	Defaulters in client social circle	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can

				include true zero value .
46	DAYS_LAST_PHONE_CHANGE	Interval	Client changed phone since the application.	Attributes consist of numeric value that can be differentiated to obtain outcome but cannot have true zero value.
47	FLAG_DOCUMENT_2, FLAG_DOCUMENT_3, FLAG_DOCUMENT_4, FLAG_DOCUMENT_5, FLAG_DOCUMENT_6, FLAG_DOCUMENT_7, FLAG_DOCUMENT_8, FLAG_DOCUMENT_9, FLAG_DOCUMENT_10, FLAG_DOCUMENT_11, 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,	Nominal	Values contain if client gave document or not	Attributes consist of categorical data which are distinct and cannot be ordered or differentiated between values

	FLAG_DOCUMENT_21, FLAG_DOCUMENT_22			
48	AMT_REQ_CREDIT_BUREAU_HOUR	Ratio	Number of queries in credit bureau for client based on hour	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
49	AMT_REQ_CREDIT_BUREAU_DAY	Ratio	Number of queries in credit bureau for client based on days	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .

50	AMT_REQ_CREDIT_BUREAU_WEEK	Ratio	Number of queries in credit bureau for client based on week	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
51	AMT_REQ_CREDIT_BUREAU_MON	Ratio	Number of queries in credit bureau for client based on month	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
52	AMT_REQ_CREDIT_BUREAU_QRT	Ratio	Number of queries in credit bureau for client based on quartile	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .
53	AMT_REQ_CREDIT_BUREAU_YEAR	Ratio	Number of queries in credit bureau for client based on year	Attributes includes value that can be ordered, differentiated to obtain outcome and also it can be multiplied and divided which can include true zero value .

Graphical and Statistical Representation

1. TARGET

The pie chart shows the number of clients having difficulties with payment due to late payments and other cases. From the pie chart we understand that 47.17% clients have difficulties in payment and 52.83% clients wither have some other issues or they face no difficulties in payments

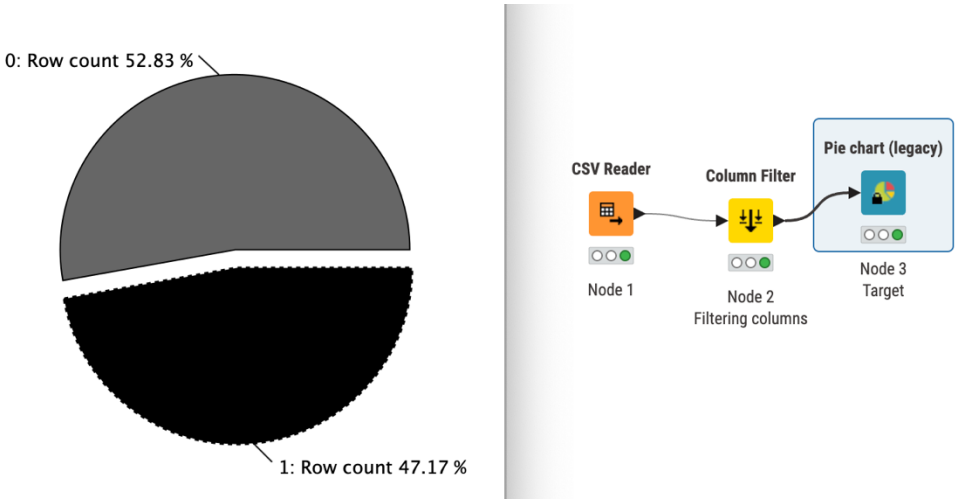


Figure 1 Pie chart and KNIME workflow for TARGET attribute

2. NAME_CONTRACT_TYPE

The pie chart shows the number of type of loans such as cash and revolving loans. The chart depicts the cash loan is higher than revolving loans as cash loans are 91.7% whereas revolving loans are just 8.3%. It is evident that clients prefer cash loans over revolving loans through this dataset.

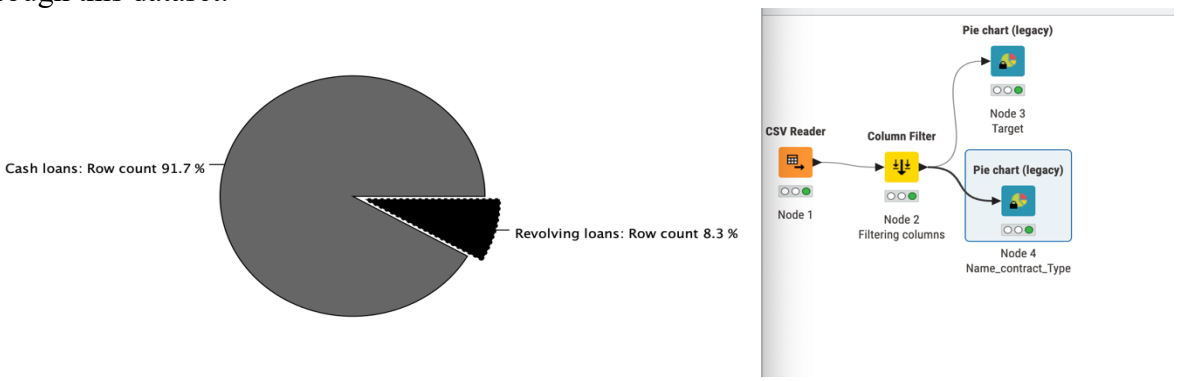


Figure 2 Pie chart and KNIME workflow for NAME_CONTRACT TYPE attribute

3. CODE_GENDER

The pie chart highlights the number of Female and Male clients. Based on the above pie chart this pie chart includes color manager which provides color to the pie chart and is helpful to identify the average value of clients who are male and female. The outcome states that female clients are 61.83% which is higher than male client that are 38.17%.

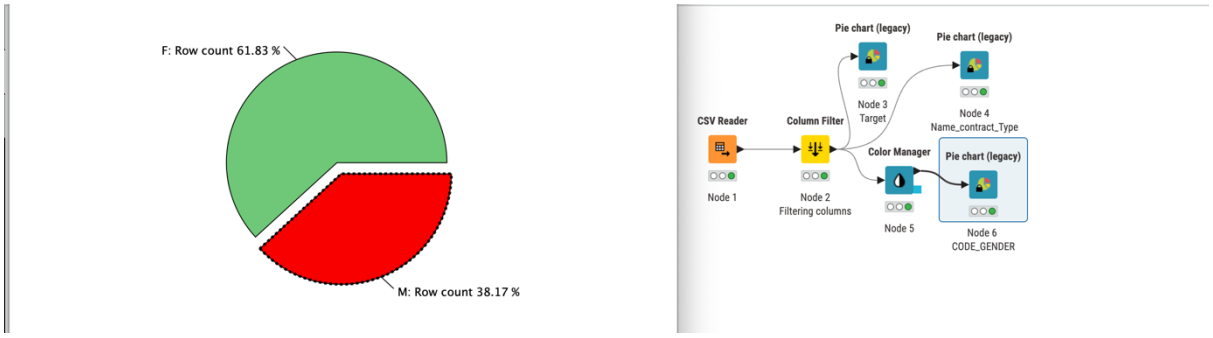


Figure 3 Pie chart and KNIME workflow for GENDER attribute

4. FLAG_OWN_CAR

The histogram of client owning a car or not. Where 0 represent no car is owned and 1 represent client that own a car. The histogram highlights that client not owning car is higher than client owning car.

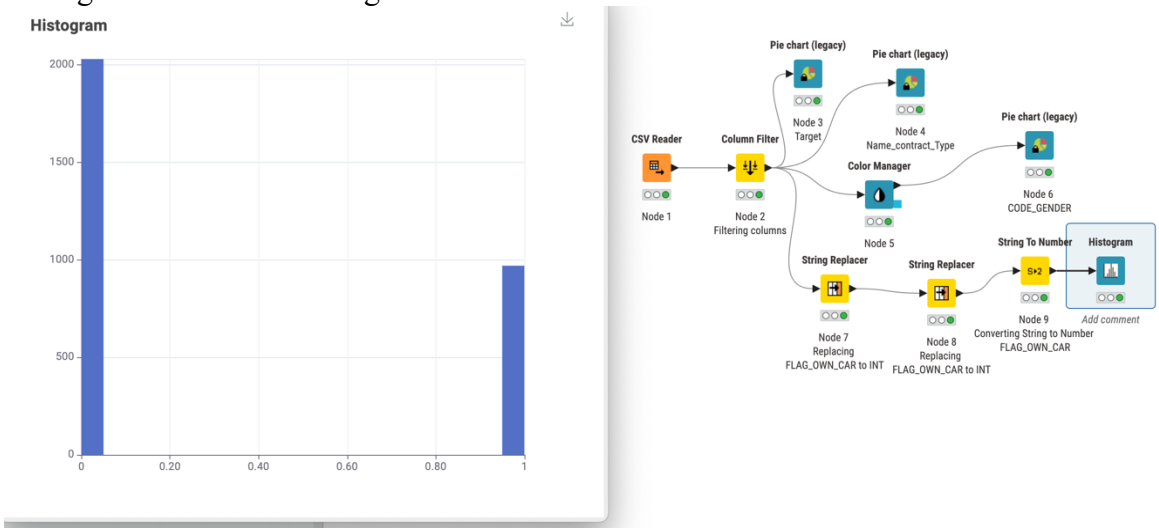


Figure 4 Histogram and KNIME workflow for FLAG_OWN_CAR attribute

5. FLAG_OWN_REALTY

The histogram of client owns realty or not. Where 0 represent no realty is owned and 1 represent client that owns realty . The histogram highlights that client owning realty is higher than client not owning realty. In this histogram I have enabled ‘show bar values’ to get the count of client that owns realty or not.

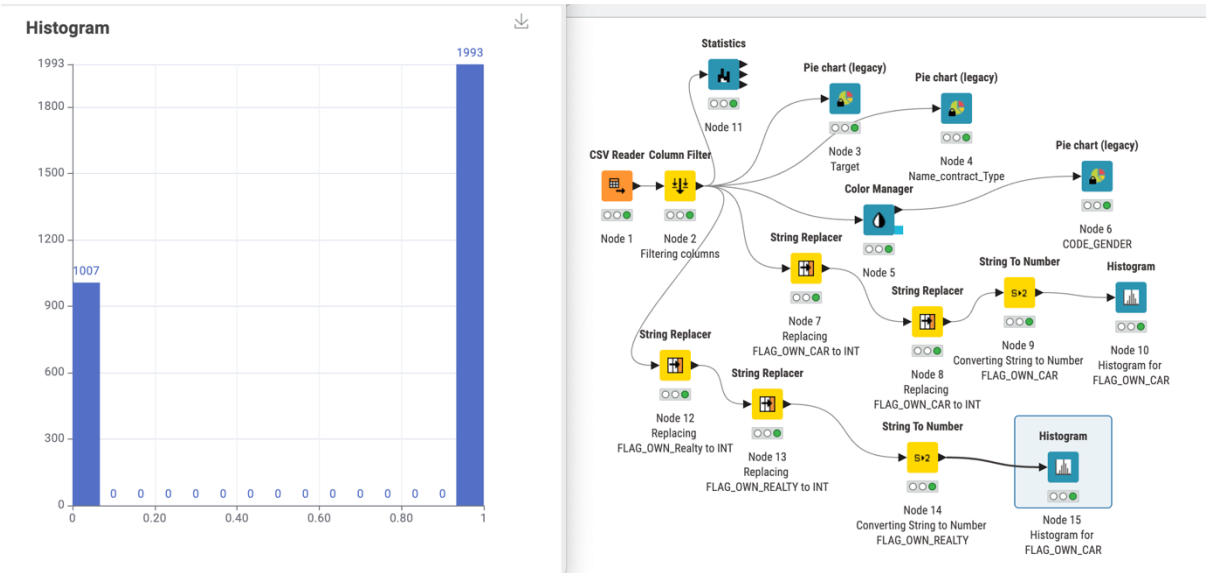


Figure 5 Histogram and KNIME workflow for FLAG_OWN_REALTY attribute

6. CNT_CHILDREN

The value indicates the count of children client has ranging from 0 to 9. Below table shows statistics and histogram for the attribute.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
CNT_CHILDREN	0.0	0.4367	0.0	9	0.7476	2.0596	7.654	0	0	0	0 9

Figure 6 Statistics and histogram of CNT_CHILDREN

7. AMT_INCOME_TOTAL

The value indicates the total income of the client starting from 27000. Figure x shows statistics and histogram for the attribute.

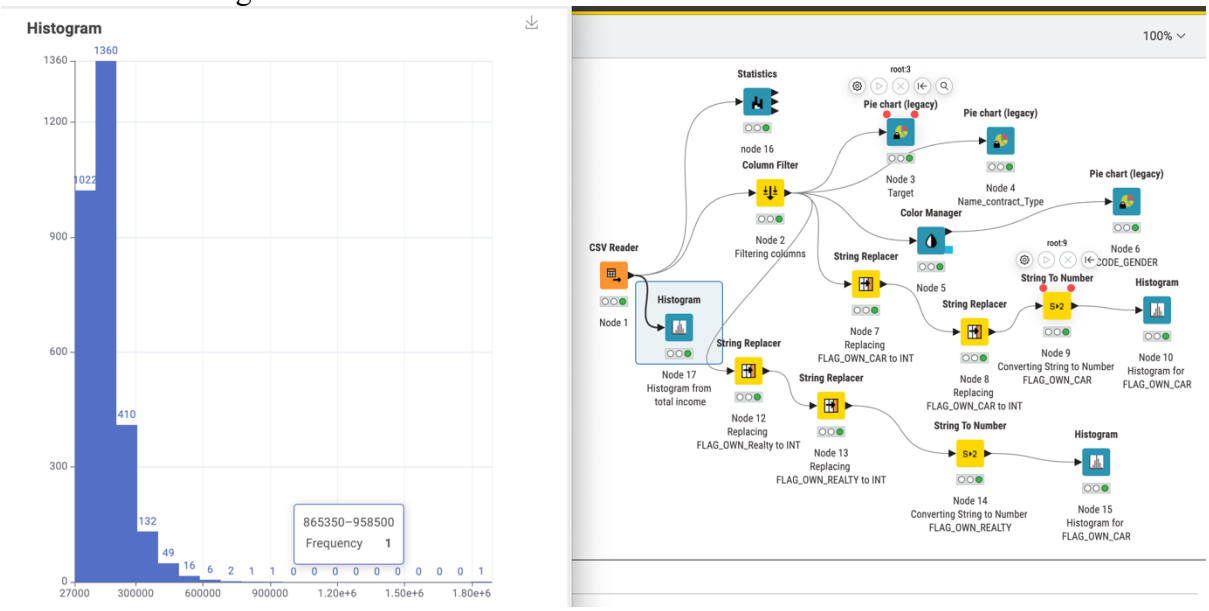


Figure 7 Histogram and Knime workflow for AMT_INCOME_TOTAL

8. AMT_CREDIT

The attribute represents the credit value client can apply for loan. It has minimum value as 45,000 and maximum value as 2,250,000. With the help of Statistics node, the statistics and histogram for the attribute are shown.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
AMT_CREDIT	45,000	581,979.2955	508,495.5	2,250,000	379,952.5137	1.1555	1.3272	0	0	0	45.000 2.250.000

Figure 8 Statistics and histogram of AMT_CREDIT

9. AMT_ANNUITY

The attribute depicts the client annuity. The histogram of annuity ranges from 2174 to 145485 With the help of Statistics node, the statistics and histogram for the attribute are shown.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
AMT_ANNUITY	2,173.5	26,895.0105	25,321.5	145,485	13,633.2805	1.2034	3.4834	0	0	0	2.174 145.485

Figure 9 Statistics and histogram of AMT_ANNUITY

10. NAME_INCOME_TYPE

The attribute Income type states category in which client get their income from. The pie chart depicts the income category. Unemployment and student has lower count as 2 and 1 respectively whereas maximum income type is from working with 54.27% of total.

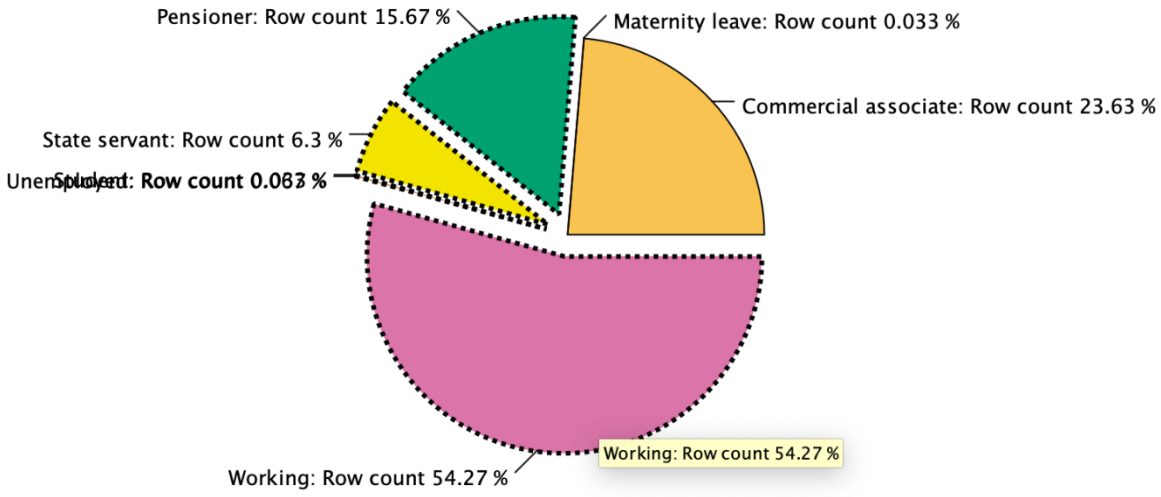


Figure 10 Pie chart of NAME_INCOME_TYPE attribute

11. NAME_EDUCATION_TYPE

The bar graph shows the category of education received by client where maximum client has completed Secondary education, one academic degree, 90 client with incomplete higher education and 30 lower secondary education.

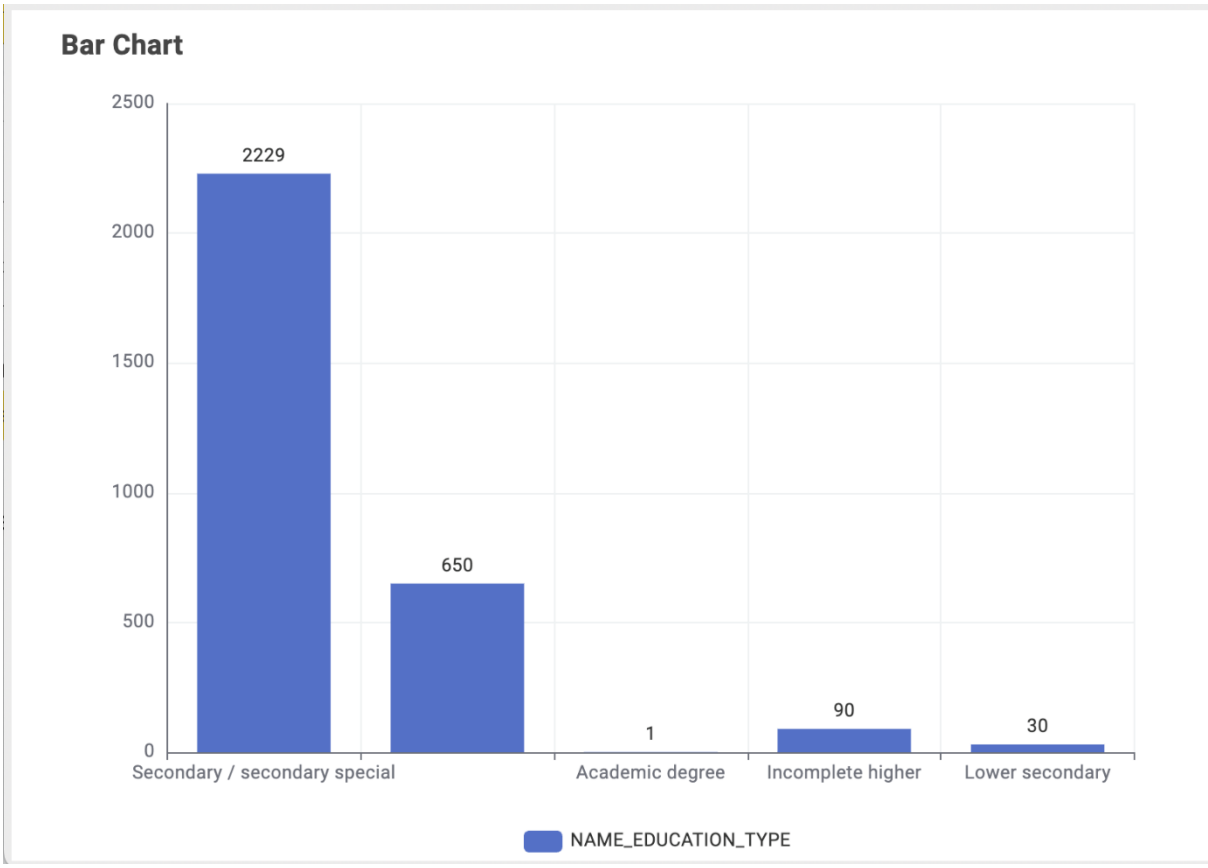


Figure 11 Histogram of NAME_EDUCATION_TYPE

12. NAME_FAMILY_STATUS

The pie chart shows client with marriage status of the client where 63.4% client are married,15.2% client is single, 10.57% client had civil marriage, 6.6% client are separated and 4.23% client are widow.

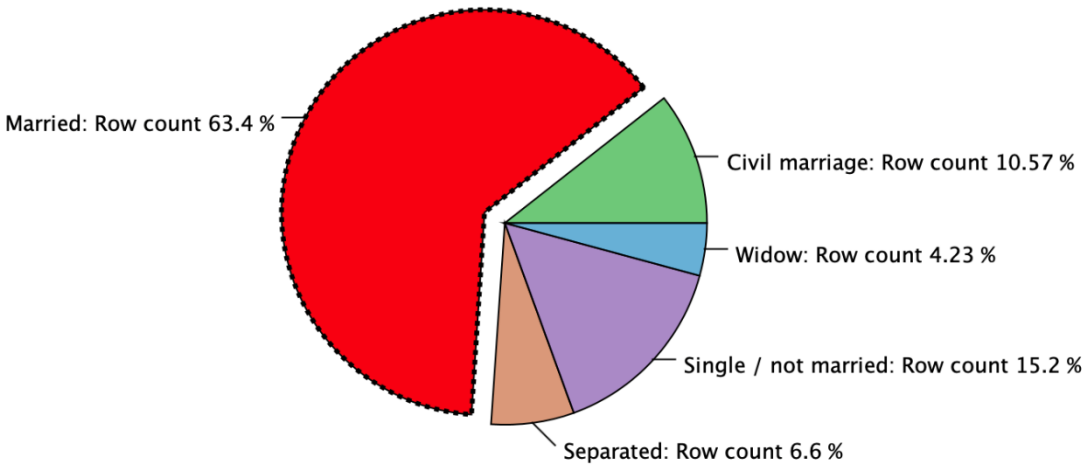


Figure 12 Pie chart of NAME_FAMILY_STATUS

13. NAME_HOUSING_TYPE

The pie chart shows the housing type of client where maximum client 87.13% belong to house/apartment housing type. Lowest is co-op apartment with 0.3% of client.

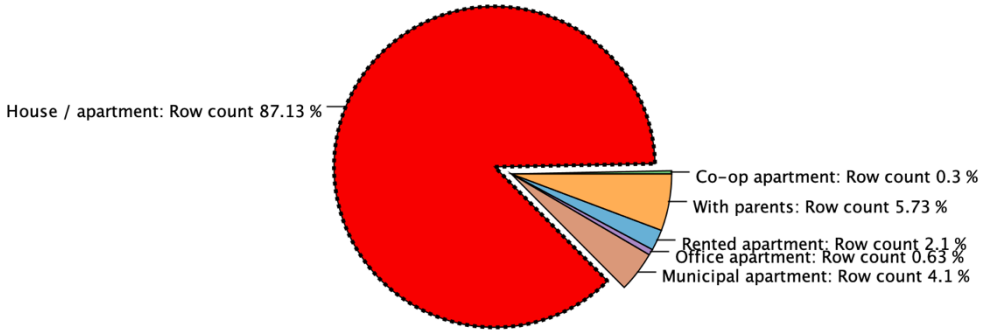


Figure 13 Pie chart of NAME_HOUSING_TYPE

14. DAYS_BIRTH

The attribute represents the number of days of client at the time of loan application. This numeric data can be used for discretization. The value for this attribute is in same range as seen in histogram which includes Adults . With the help of Statistics node, the statistics and histogram for the attribute are shown.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
DAYS_BIRTH	-24,982	-15,561.9397	-15,111	-7,717	4,350.403	-0.2517	-0.9819	0	0	0	-24.982 -7.717

Figure 14 Statistics and histogram of DAYS_BIRTH

15. DAYS_EMPLOYED

The attribute has values in days for client to be employed it also has negative range highlighting the days client is not employed since the application of loan.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
DAYS_EMPLOYED	-16,360	55,637.8587	-1,147	365,243	133,816.6185	1.8822	1.5455	0	0	0	-16.360 365.243

Figure 15 Statistics and histogram of DAYS_EMPLOYED

16. DAYS_REGISTRATION

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
DAYS_REGISTRATION	-16,366	-4,686.3373	-4,256	0.0	3,372.81	-0.6318	-0.3104	0	0	0	-16.366 0

Figure 16 Statistics and histogram of DAYS_REGISTRATION

17. DAYS_ID_PUBLISH

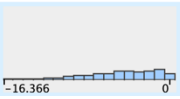
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
DAYS_REGISTRATION	-16,366	-4,686.3373	-4,256	0.0	3,372.81	-0.6318	-0.3104	0	0	0	

Figure 17 Statistics and histogram of DAYS_ID_PUBLISH

18. EXT_SOURCE_2

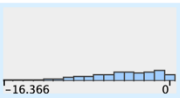
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
DAYS_REGISTRATION	-16,366	-4,686.3373	-4,256	0.0	3,372.81	-0.6318	-0.3104	0	0	0	

Figure 18Statistics and histogram of EXIT_SOURCE_2

19. OBS_60_CNT_SOCIAL_CIRCLE

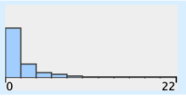
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
OBS_60_CNT_SOCIAL_CIRCLE	0.0	1.4769	0.0	22	2.3631	2.4641	8.2613	8	0	0	

Figure 19 Statistics and histogram of OBS_60_CNT_SOCIAL_CICRLE

20. DEF_60_CNT_SOCIAL_CIRCLE

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
DEF_60_CNT_SOCIAL_CIRCLE	0.0	0.1133	0.0	4	0.3812	3.8837	17.6201	8	0	0	

Figure 20 Statistics and histogram of DEF_60_CNT_SOCIAL_CICRLE

21. AMT_REQ_CREDIT_BUREAU_YEAR

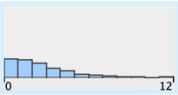
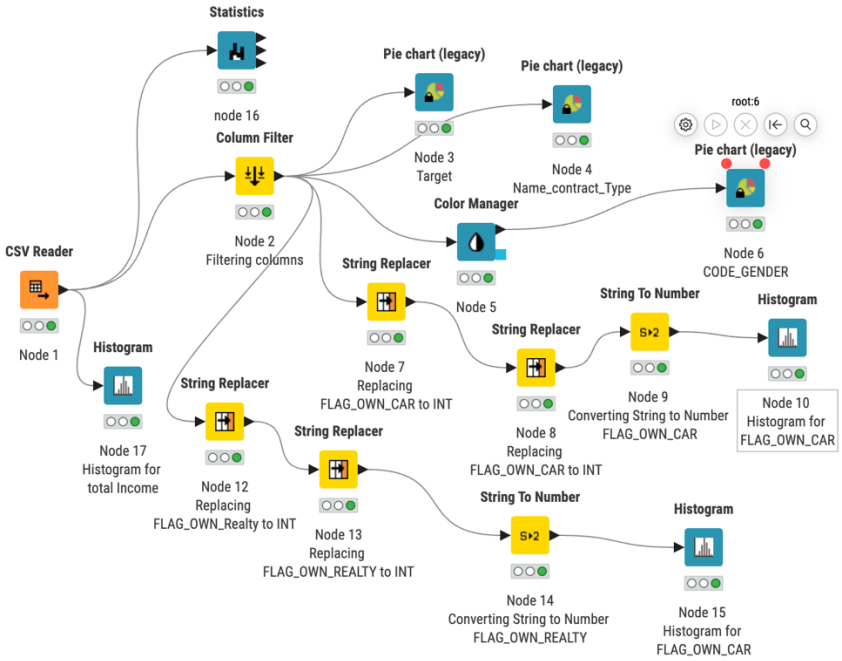
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
AMT_REQ_CREDIT_BUREAU_YEAR	0.0	2.007	2	12	1.9367	1.1929	1.3481	423	0	0	

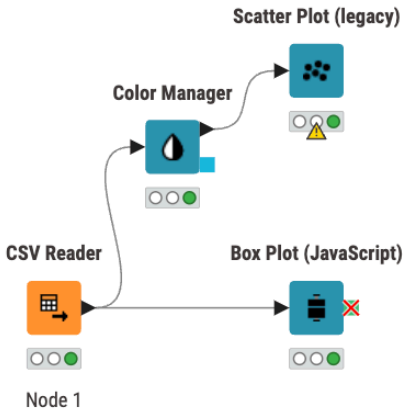
Figure 21 Statistics and histogram of AMT_REQ_CREDIT_BUREA

Knime Workflow



Outliers and clusters

To Identify outliers and clusters, box plot and scatter plot help identifying it. The analysis of client data allows to note several insights about the dataset and information gathered.



The first box plot of Amt-Annuity highlights the noticeable number of outliers. These outliers can be used for skew statistics and predictions for future analysis. The outlier in AMT_ANNUITY is one client with 150,000 annuity.

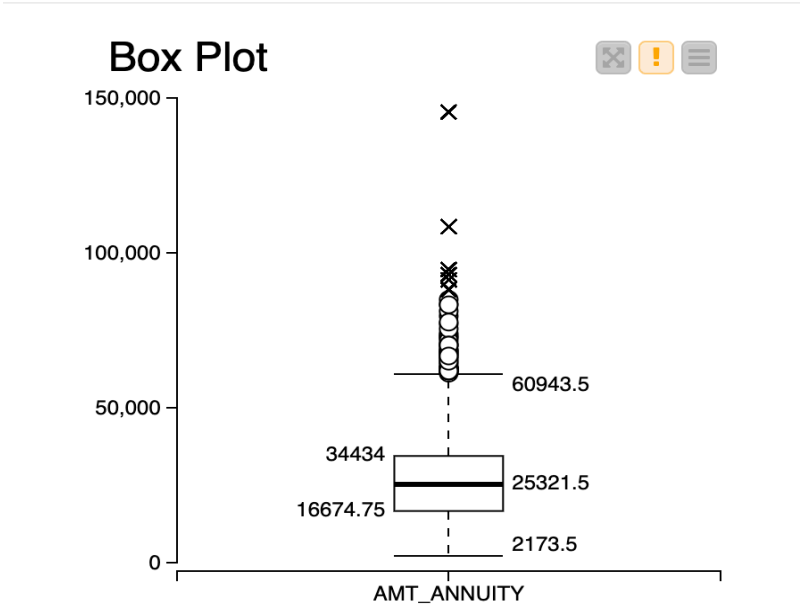


Figure 22 Box Plot of AMT_ANNUITY

The scatter plot highlights the category of client based on their education and combines the amount client can borrow.
X column has NAME_EDUCATION_TYPE and y- column includes scatter chart of AMT_CREDIT. IT highlights that client with Secondary and higher education has higher chance of credit amount to borrow whereas Academic degree client get less credit to borrow.

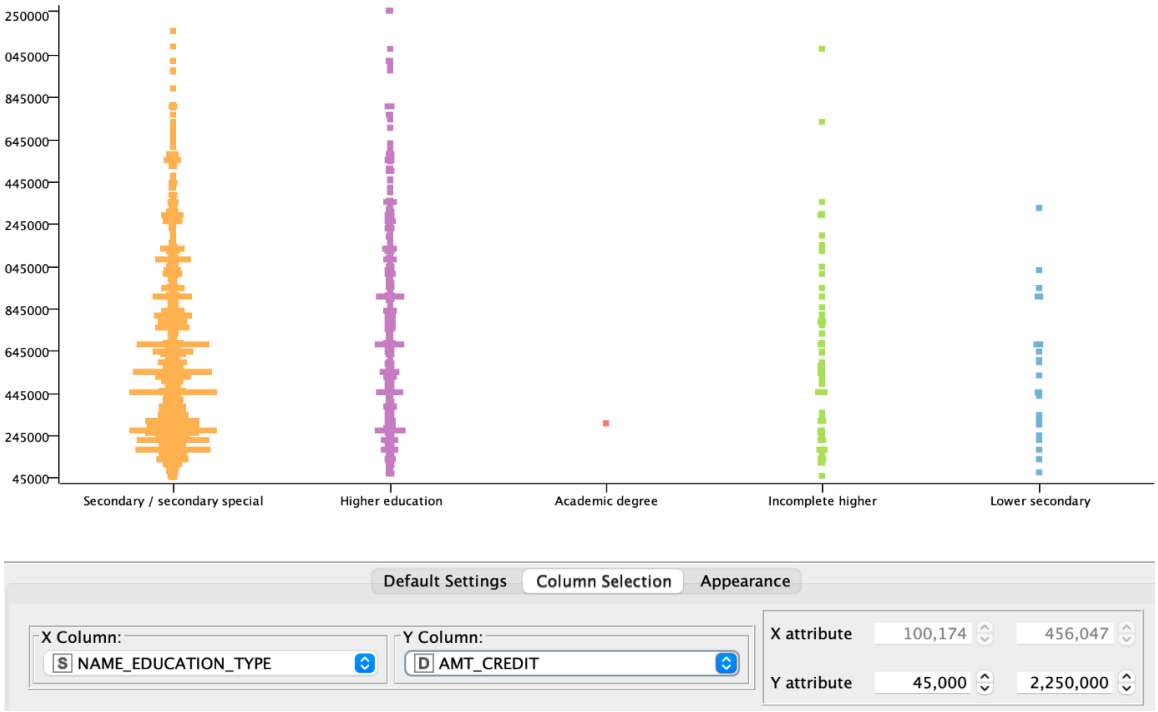


Figure 23 Scatter plot of NAME_EDUCATION_TYPE and AMT_CREDIT

The attribute AMT_INCOME_TOTAL has enough outliers to stand out. It indicates the income level of client who are loan applicants, and this attribute plays important part for

identifying defaulters. Box plot highlights client with high income of about more than 1,500,000 amount.

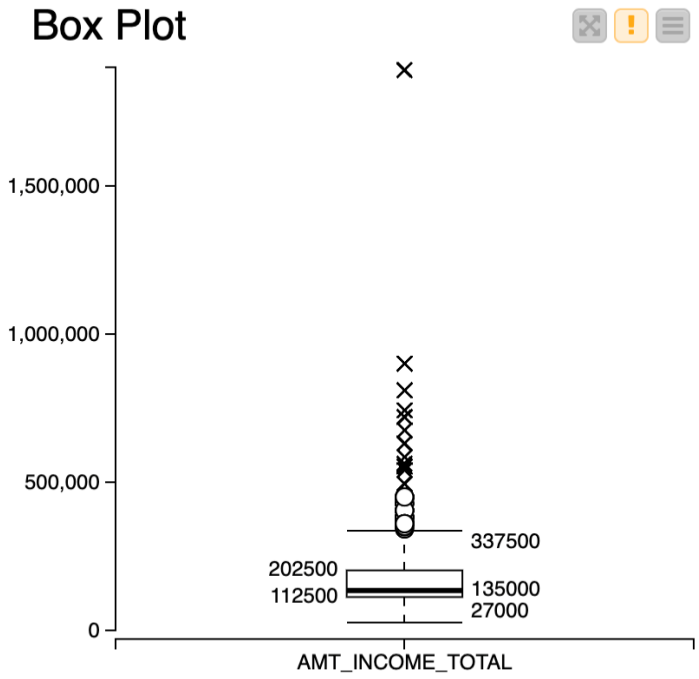


Figure 24 box plot of AMT_INCOME_TOTAL

The attribute DAYS_BIRTH has no outliers as all instances has no positive value which implies the client age is correct and can be relied for processing. With the help of box-plot identifying these elements are easier.

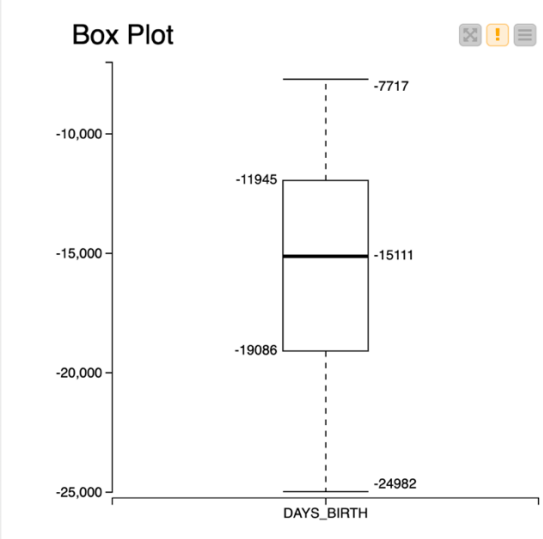


Figure 25 box plot DAYS_BIRTH

Attribute DAYS_EMPLOYED includes an outlier which has value that exceeds 350,000 days of employment at the time of application which is invalid if counted in years Through this box plot we could be able to overview this information and flag this instance for further

verification. This outlier value is impossible for client to be employed and results incorrect data.

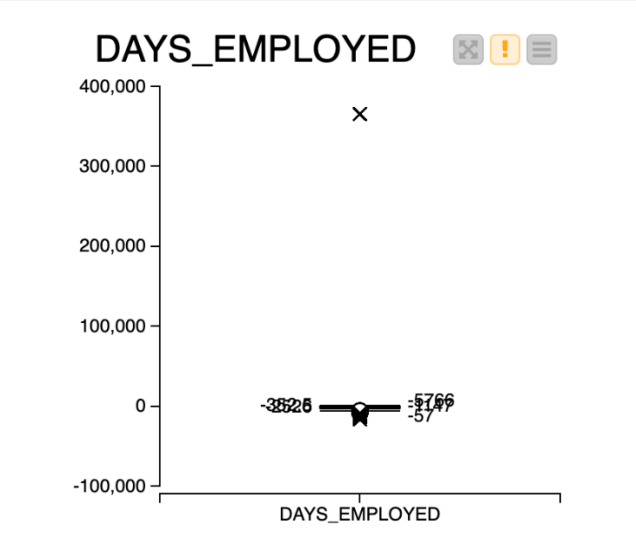


Figure 26 box plot of DAYS_EMPLOYED

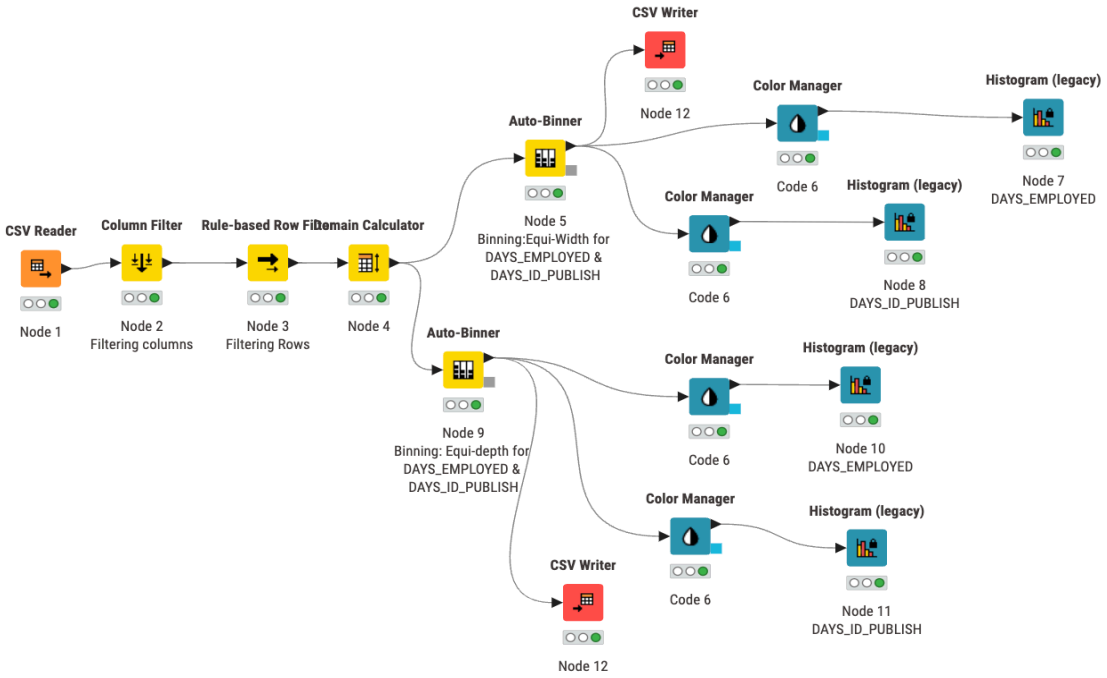
Task : Data Preprocessing

This task includes multiple pre-processing such as
Binning
Normalisation
Discretisation
Binarization

Binning

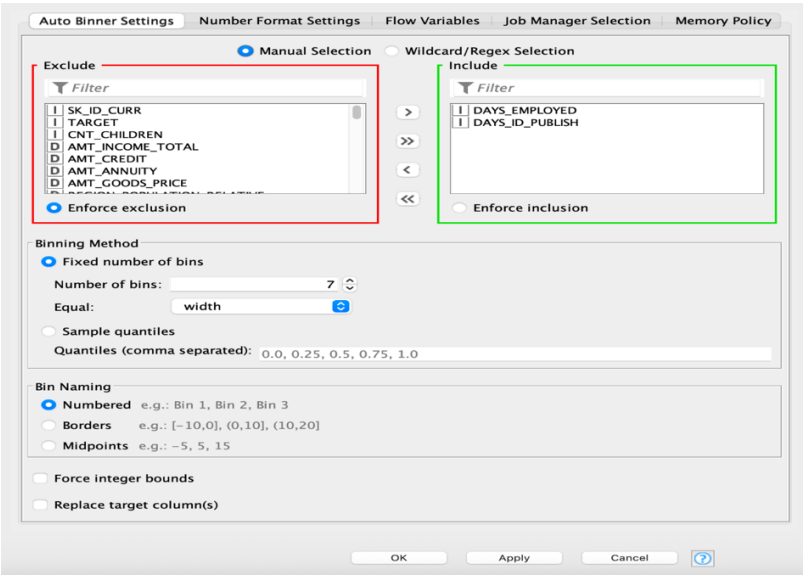
Binning includes process in which data is grouped based on criteria. The task includes binning two attributes DAYS_EMPLOYED and DAYS_ID_PUBLISH by techniques like equi-width and equi-depth binning.

Knime Workflow for Binning



To perform Equi-Width Binning following are the steps performed:

1. Connect Column and row Filter to CSV reader node to filter out the dataset.
2. Connect Auto-Binner for equi-width binning
3. To calculate the equi-width for range (h-1)/n equal width.
4. Setting the bin size to bin = 7
5. Save the configuration
6. Connect color manager for differentiating values
7. Connect Histogram to get data visualization



The histogram below shows the visualization of Equi-Width Binning on DAYS_EMPLOYED attribute

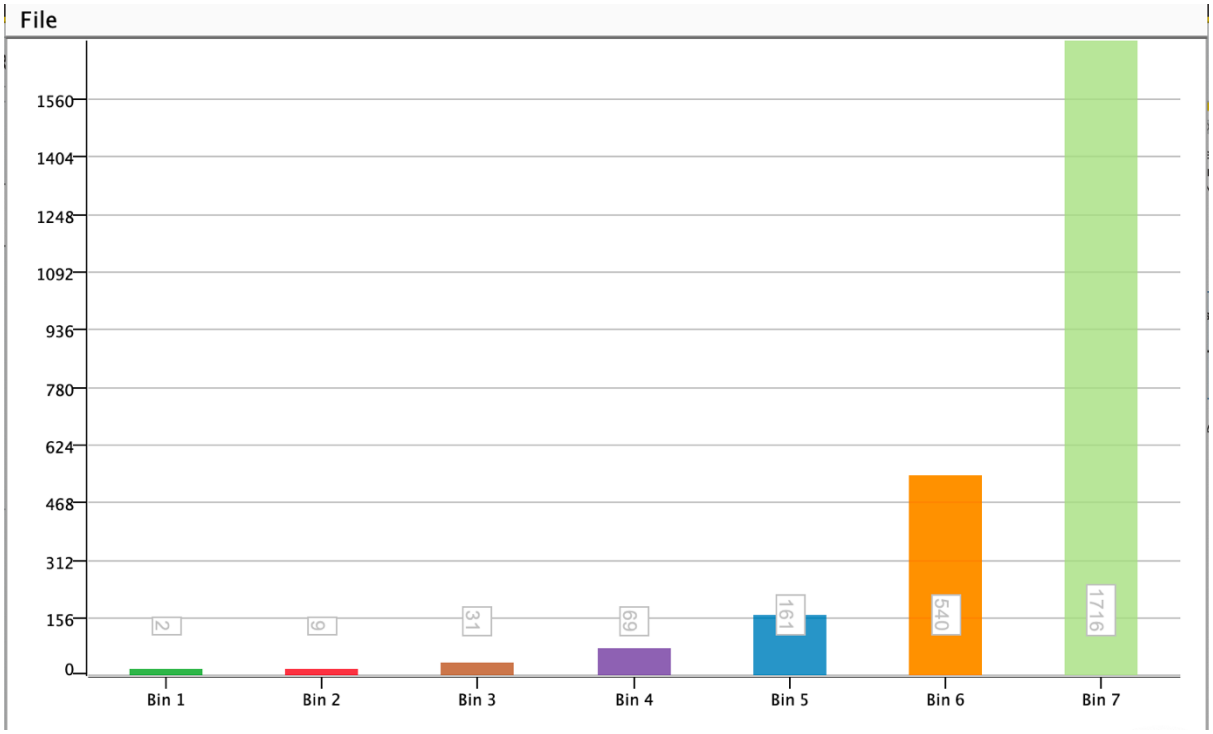


Figure 27 Histogram for equi-width binning for DAYS_EMPLOYED

The histogram below shows the visualization of Equi-Width Binning on DAYS_ID_PUBLISH attribute

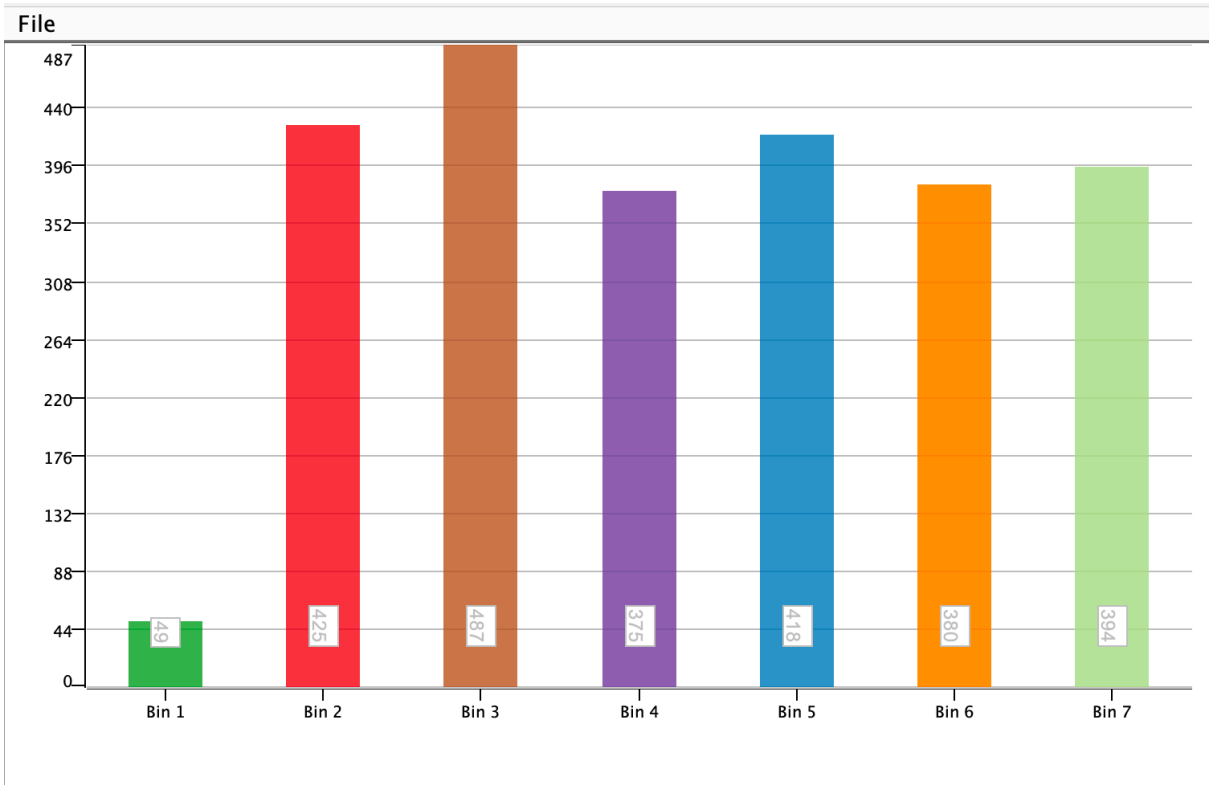
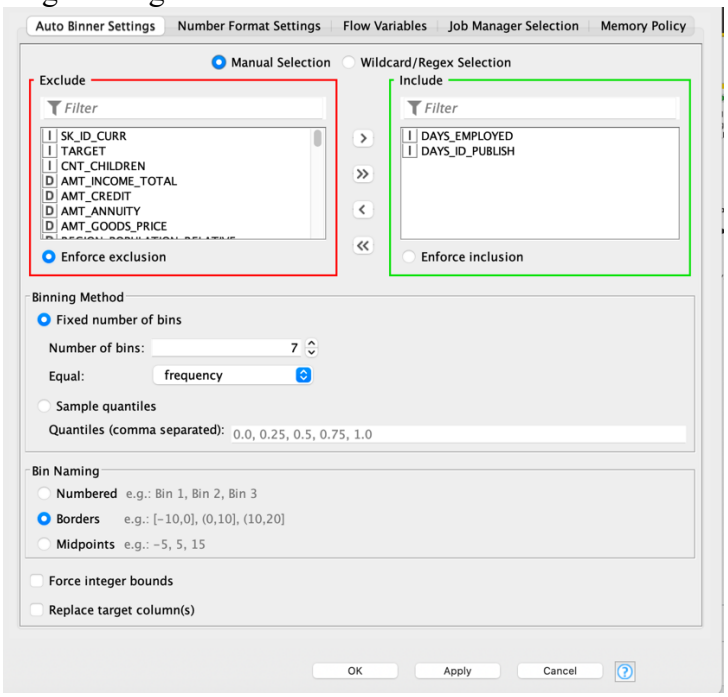


Figure 28 Histogram for equi-width binning for DAYS_ID_PUBLISH

To perform Equi-Depth Binning following are the steps performed:

1. Connect Column to CSV reader node to filter out the dataset.
2. Added row filter to filter out the positive values and zero to identify outliers by adding the code $DAYS_EMPLOYED$ <= 0 => TRUE
3. Connect Auto-Binner for equi-depth binning
4. To calculate the equi-width for range (h-1)/n equal depth.
5. Setting the bin size to bin = 7
6. Save the configuration
7. Connect color manager for differentiating values
8. Connect Histogram to get data visualization



The histogram below shows the visualization of Equi-depth Binning on DAYS_EMPLOYED attribute

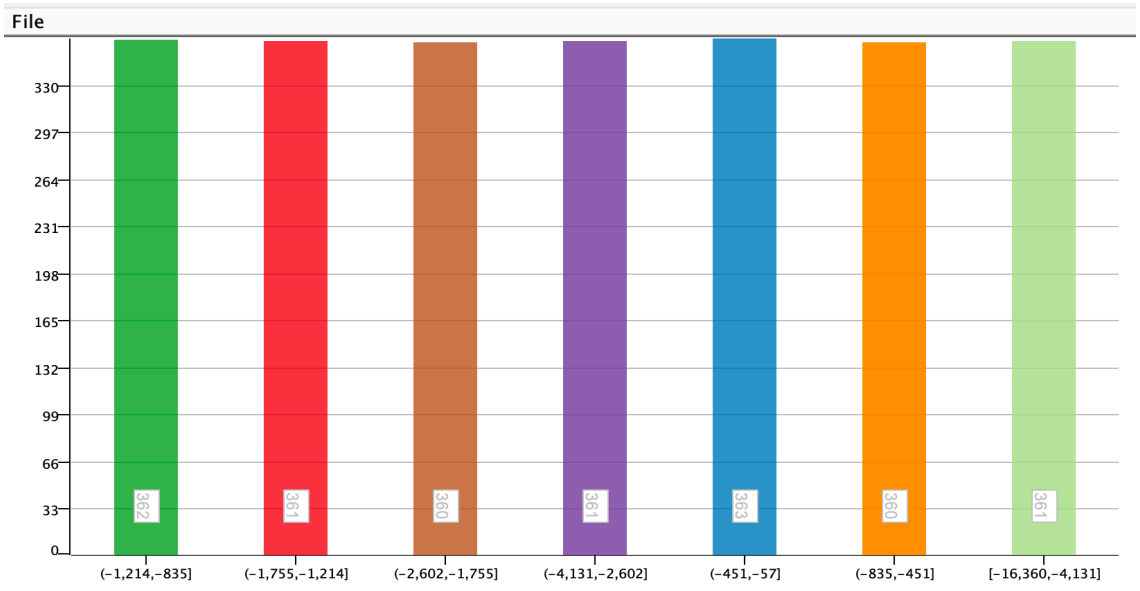


Figure 29 Histogram for equi-depth binning for DAYS_ EMPLOYED

The histogram below shows the visualization of Equi-Depth Binning on DAYS_ID_PUBLISH attribute

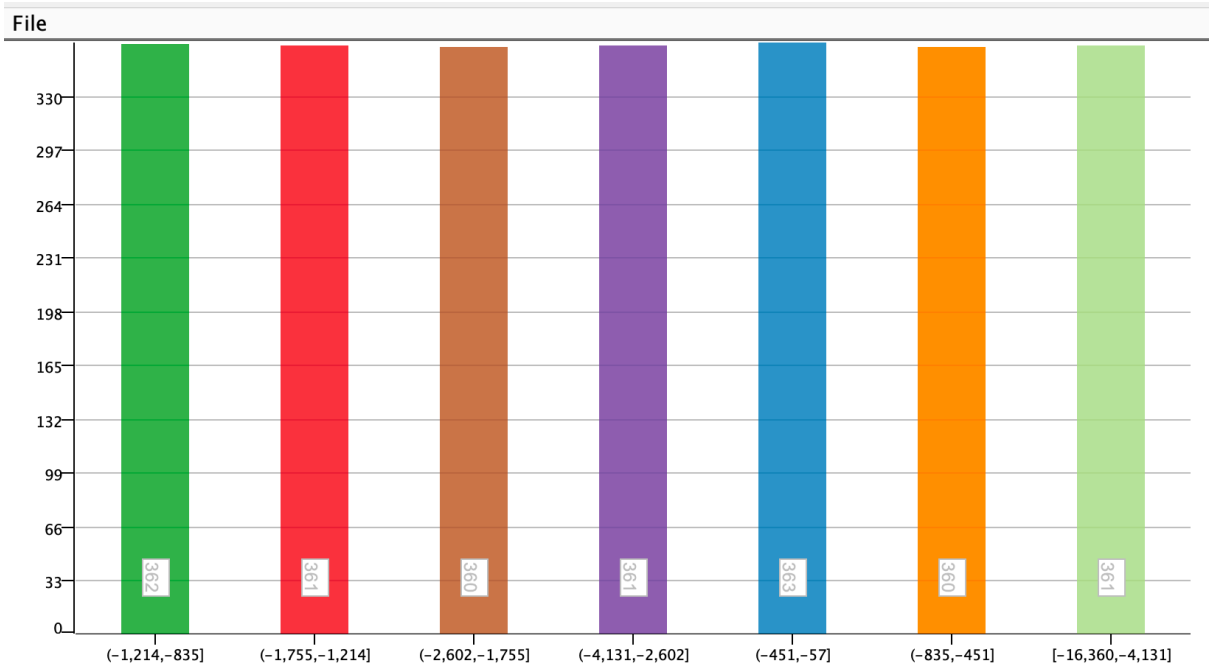
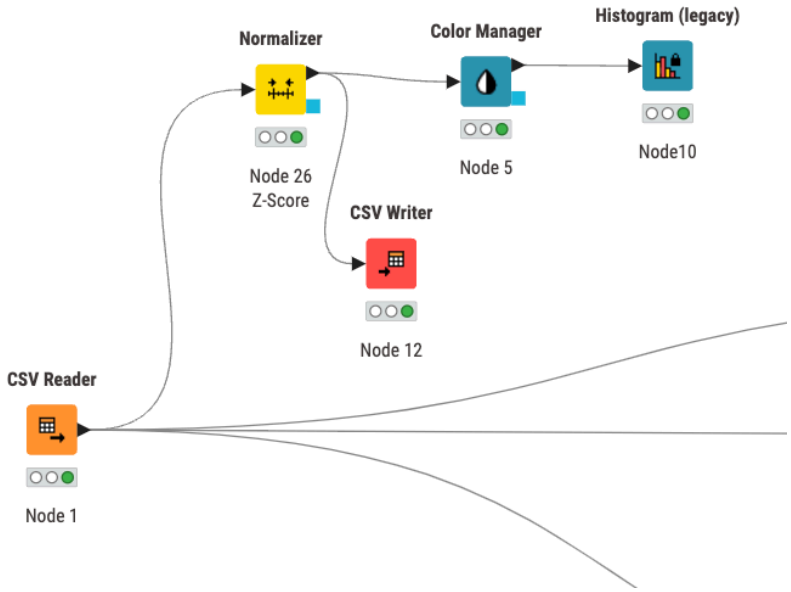


Figure 30 Histogram for equi-depth binning for DAYS_ ID_PUBLISH

Normalization

Normalization helps in changing the range of data to be organized in distributed values of attributes. Application of Min-max normalization and Z-score is performed.

Knime workflow for normalization



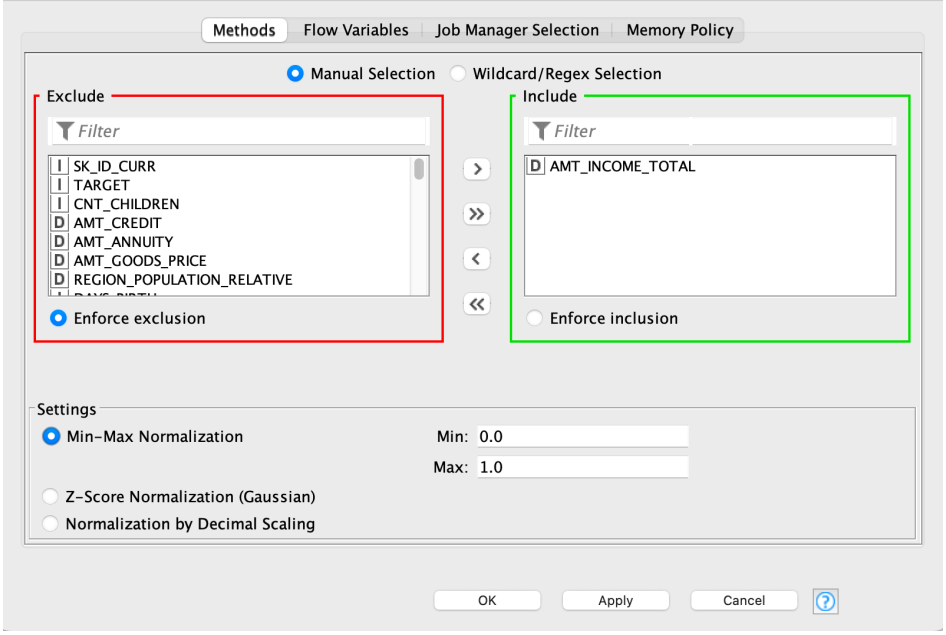
Steps performed

1. CSV reader is connected to Normalizer
2. Configuring normalizer node has settings such as Min-Max normalization. It includes value that has boundaries, in this scenario 0.0 min and 1.0 max value to include data

that are in that range. Any value beyond that boundary will be considered are out of boundary.

3. Connect color manager for differentiating values
4. Connect Histogram to get data visualization

Settings for Min-Max Normalization :



The histogram below shows the visualization of Min-Max Normalization on AMT_INCOME_TOTAL attribute

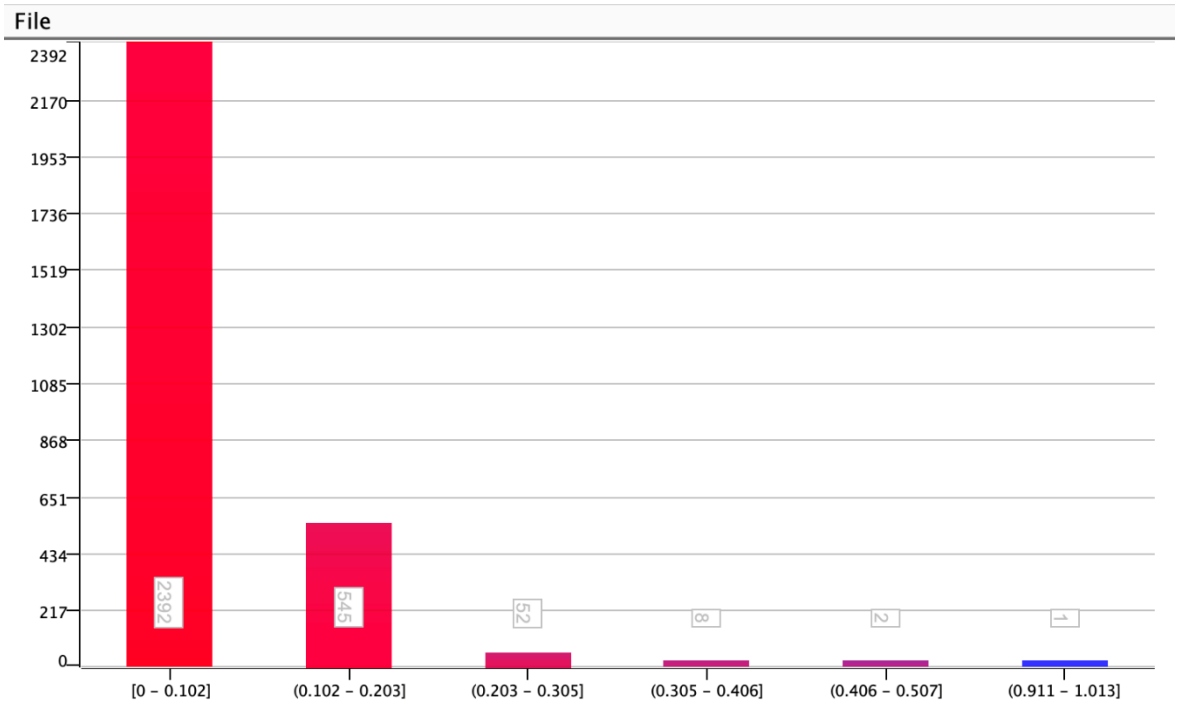
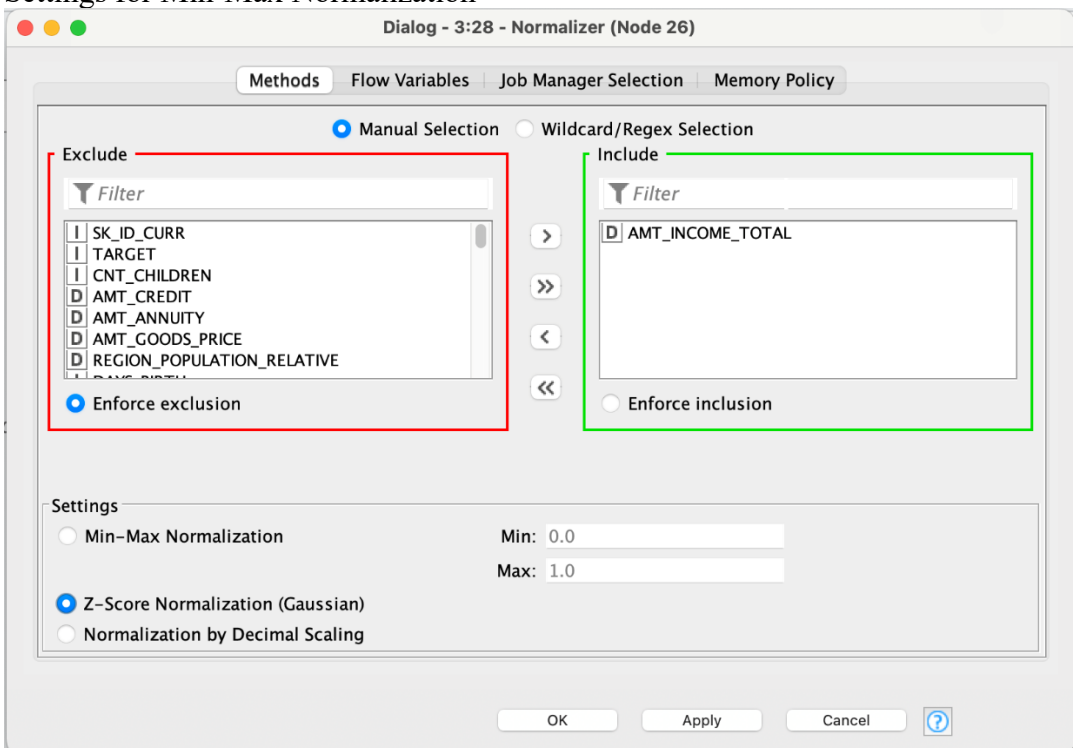


Figure 31 Histogram of AMT_INCOME_TOTAL using Min-Max Normalization

Settings for Min-Max Normalization



The histogram below shows the visualization of Z-Score Normalization on AMT_INCOME_TOTAL attribute

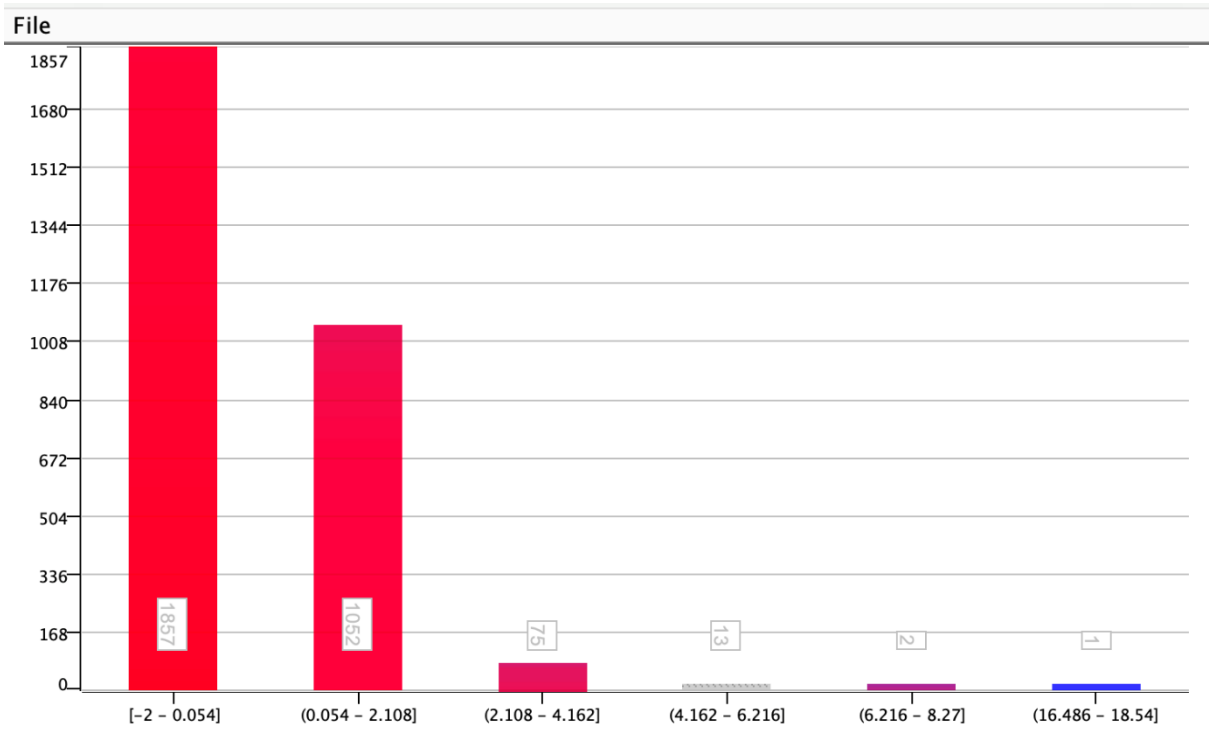
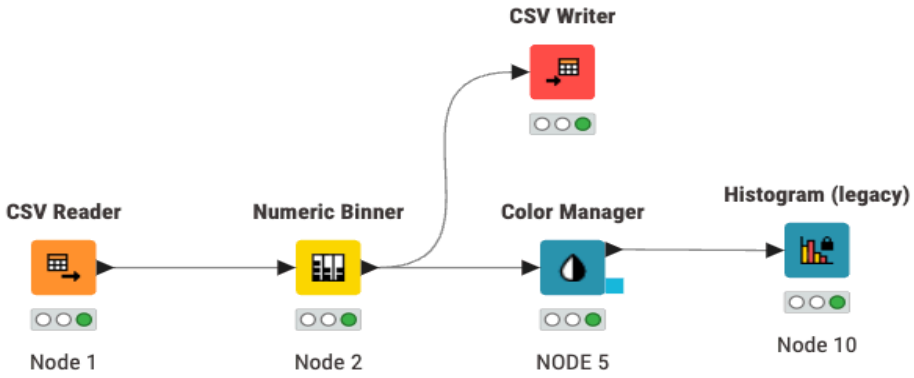


Figure 32 Histogram of AMT_INCOME_TOTAL using Z-Score Normalization

Discretization

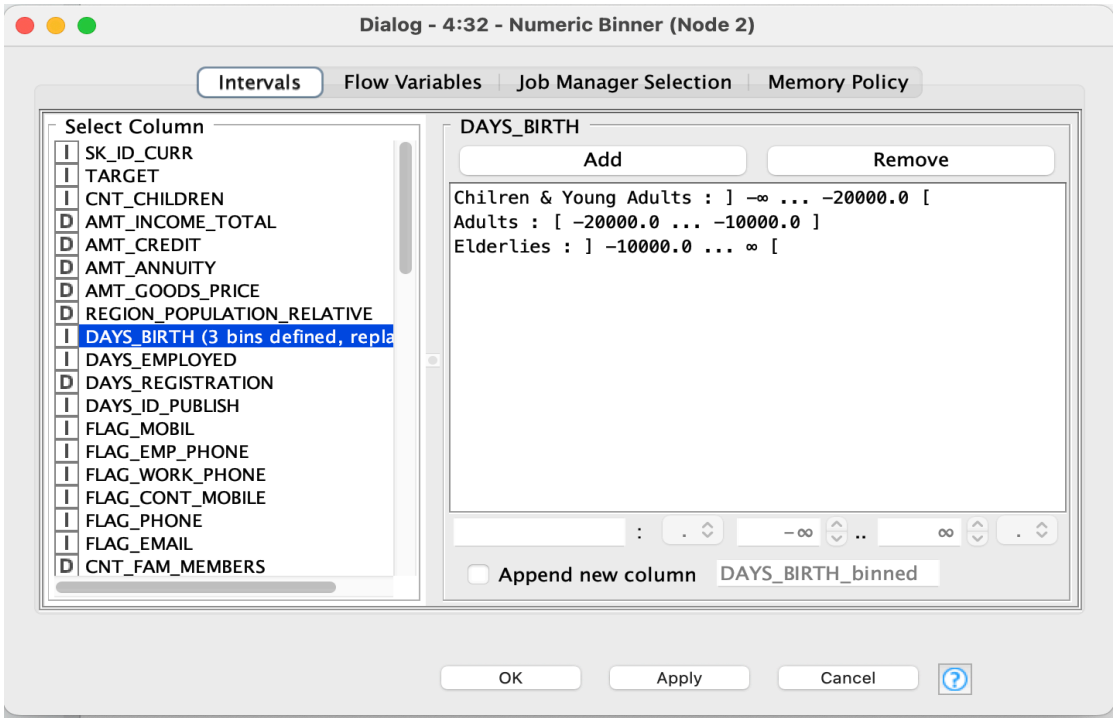
Discretise the column DAYS_BIRTH attribute is to reduce the occurrence of values by mapping the numeric value to categorical term.

Knime Workflow for Discretization



To Categorize the numeric value following steps were performed

1. Numeric binner node was connected to CSV Reader
2. DAYS_BIRTH attribute is selected
3. Based on conditions 3 bins are created such as Children and Young adults (-10,000 – 0), Adults (-20,000 — - 10,000), and Elderlies (-30,000 — -20,000).
4. Configurations are saved and executed.
5. With the help of color manager histogram has color-coded for different categories.



Visual representation of Discretised outcome for DAYS_BIRTH attribute
Frequency of categories are:
Adults – 2128
Children & Young Adults – 592
Elderlies - 280

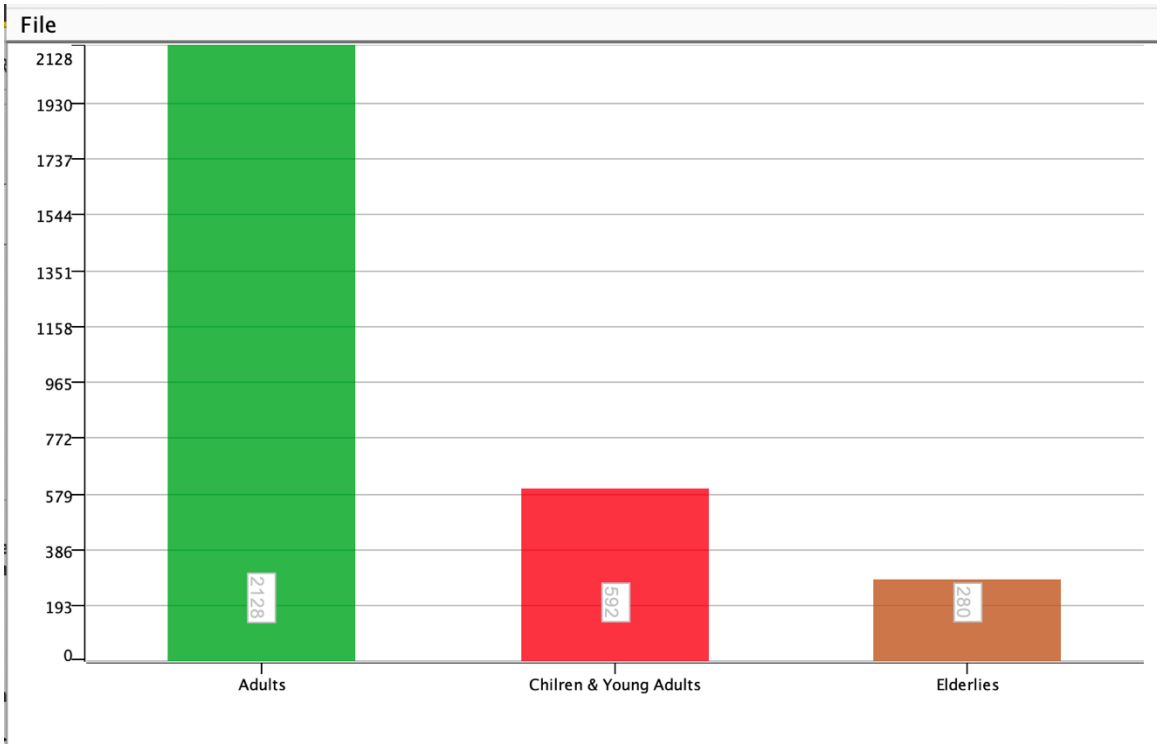
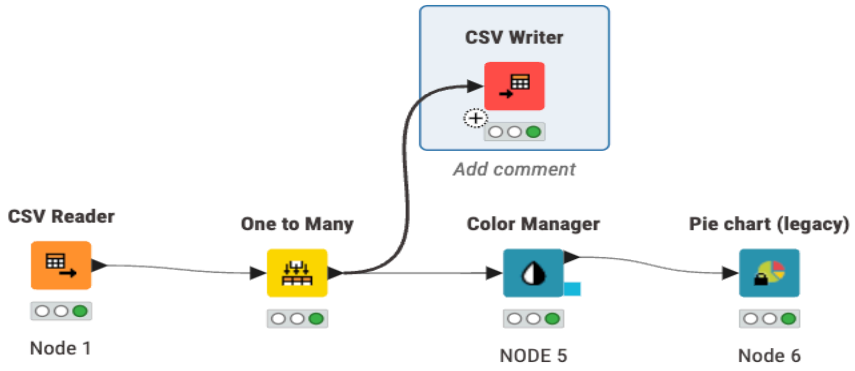


Figure 33 Histogram of DAYS_BIRTH using Discretisation

Binarization

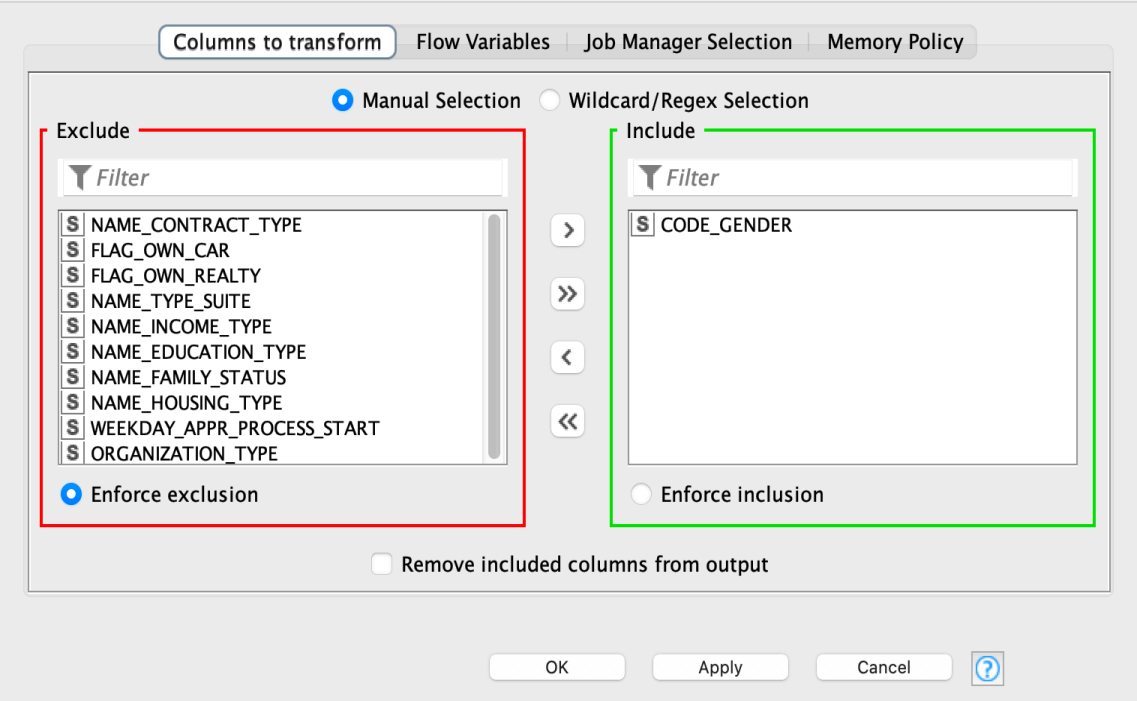
Binarizing CODE_GENDER includes technique to convert distinct value to binary value as 0 or 1. In this condition, CODE_GENDER attribute value, if the condition is satisfied value is true or else it is false.

Knime workflow for Binarization



To Binarize the CODE_GENDER attribute following steps were performed

1. Connect CSV Reader to One-to-Many node
2. Configure One-to-Many Node add the CODE_GENDER and save the configuration.



Visualization of Binarized data for CODE_GNDER attribute

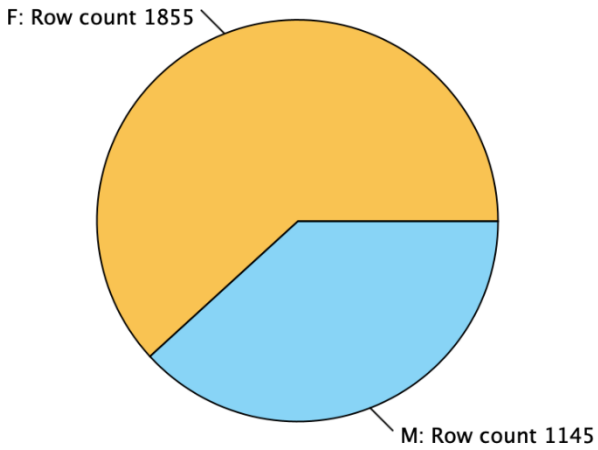


Figure 34 Pie chart of CODE_GENDER using Binarization

Summary

The dataset was analysed well to create this report and all the attributes along with their type is described in detailed in this report.

Using Knime framework visualizing the data and for in-depth understanding of data I have selected 21 attributes from the dataset which provided analysis and summary of information through data. Using charts and graphs to get valuable insights use of attributes were well analysed based on their compatibility to provide information.
This dataset helped in predicting the pattern for loan defaulters with the use of outliers and clustering potential standout were identified. Using scatterplot, the pattern indicates that credit amount is higher for secondary and higher education perceived client. Other attributes for clustering were AMT_ANNUITY to check client annuity, AMT_INCOME_TOTAL to check client's income at the time of loan application to ensure the client credit score is good and reduce risk of defaulters. Data quality assessment was an underlined task as executing box plot I encountered DAYS_EMPLOYED data has unrealistic information about number of days client was employed and the box plot highlighted the input with more than 300,000 days of employment which is incorrect and can issue. Validating such information helps to process the information well and reduce the risk of defaulters.

This assignment included data understanding, data quality check, data processing and providing knowledge based on the numeric and categorical data. It included investigation of clients who can be potential defaulters based on factors leaning into specific relationship nd patterns in dataset.