

Assignment Name:

Student Name:

Student Number:

1a. Initial Data Exploration

1. Attribute types

First way:

Attribute Name: Quote_Id

Attribute Type: Nominal

Justification: The Attribute type of Quote_Id is 'Nominal' as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

Attribute Name: Quote_Date

Attribute Type: Interval

Justification: The Attribute type of Quote_Date is 'Interval' as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated.

Attribute Name: Quote_Flag

Attribute Type: Ordinal

Justification: The Attribute type of Quote_Flag is 'Ordinal' as it represents discrete and ordered units and is used to label variables which can be ordered.

.....

Second way:

Attribute Name	Attribute Type	Justification
Quote_Id		
Quote_Date		
Quote_Flag		
.....		
.....		

2. The summarising properties for the attributes

Frequency & Distribution Data Visualisations

Statistics	Value
Mean	
Median	
Minimum Value	
Maximum Value	
Standard deviation	
Variance	
.....	
.....	

The pie chart below (refer to figure 1) indicates the binary values (0 and 1) that support the Quote_Flag attribute. In the given information it was outlined that the Quote_Flag refers to whether a customer has purchased a policy or not. Utilising that logic in this data, '0' refers to customers who have not purchased the policy and '1' refers to the customer who have purchased the policy. Based on the row count 18.43% of the recorded people have purchased the insurance policy and 81.57% have not purchased the insurance policy. Evident from the Frequency graph (refer to figure 2), the frequency for '0' is nearly quadruple than that of '1' harbouring 2447 selections as opposed to 553.

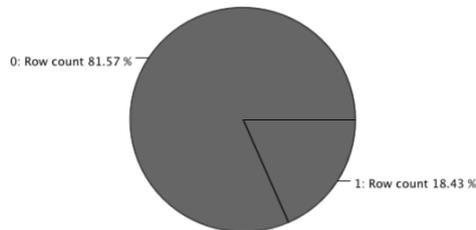


Figure 1 Pie Chart for Quote_Flag

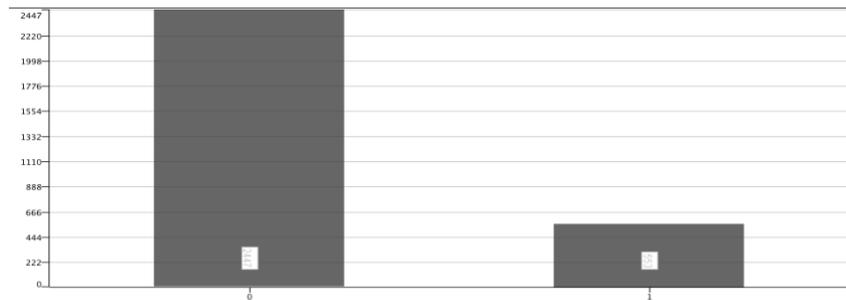


Figure 2 Frequency Graph for Quote_Flag

3 exploration

By utilising hierarchical clustering and calculating the year difference to today, this scatter plot outlines that the most personal_info3 nominal data was provided 6 and 7 years ago. The most recent data collection in this dataset record was 5 years ago – however, it had lower personal_info3 nominal data than 6 & 7 years ago. The lowest collection of personal_info3 nominal data was 8 years ago. Not only does this represent a decline in the recording of personal_info3 nominal data but also represents a decline in Quote_IDs essentially that less quotes were given out 5 years ago.

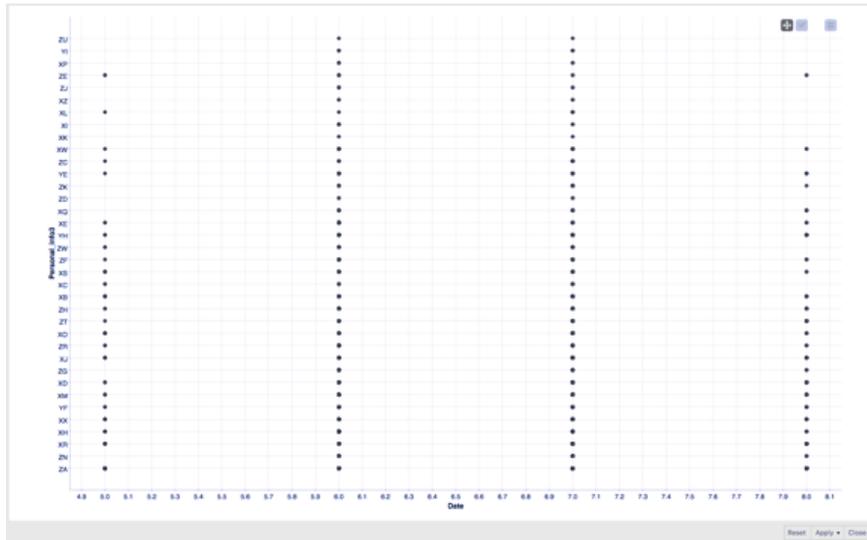


Figure 3 Scatter Plot Date vs Personal_info3

In the figure below, there is a linear correlation between Field_info1 and Geographic_info5. Primarily focusing on the 4 main clusters, the potential customer who has chosen their geographical location as IL correlates to the Field_info1 value E, similarly customers who have selected geographical location asw CA heavily correlates to B, NI correlates to F and TX correlates to J.

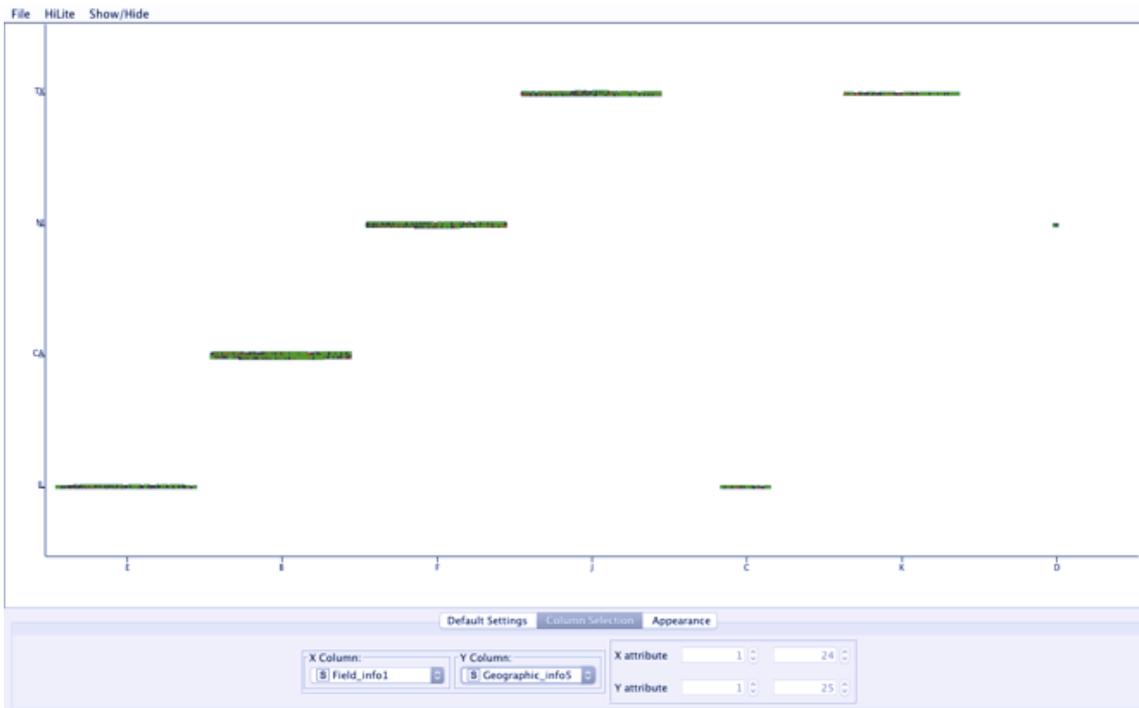


Figure 4 Field_info1 vs Geographic_info5

The rank correlation and linear correlation chart below identifies the correlations between the different attributes that have been used in the discussions within this report.

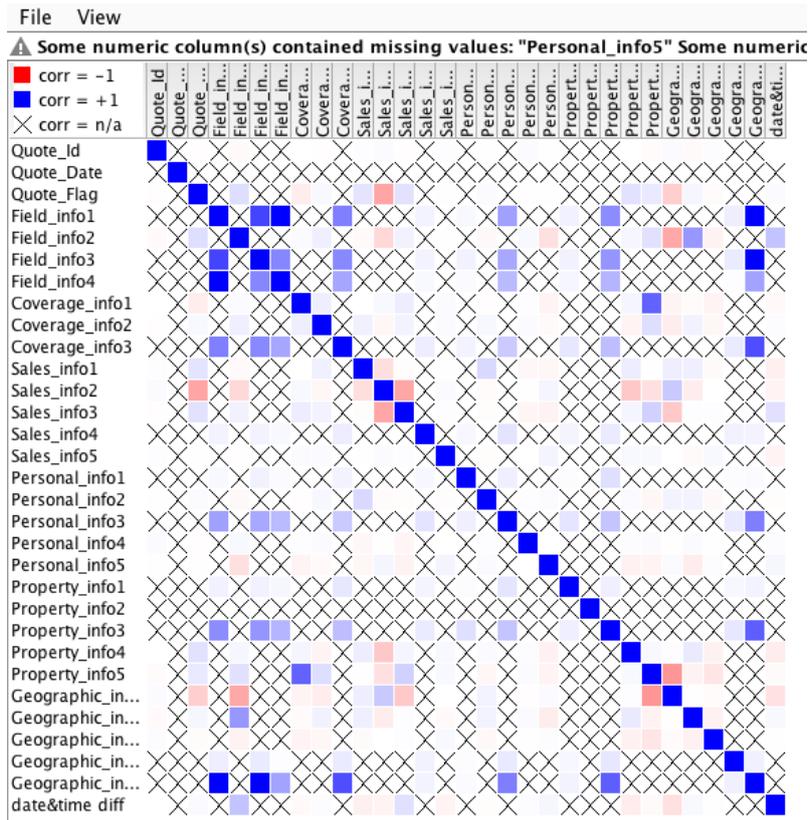
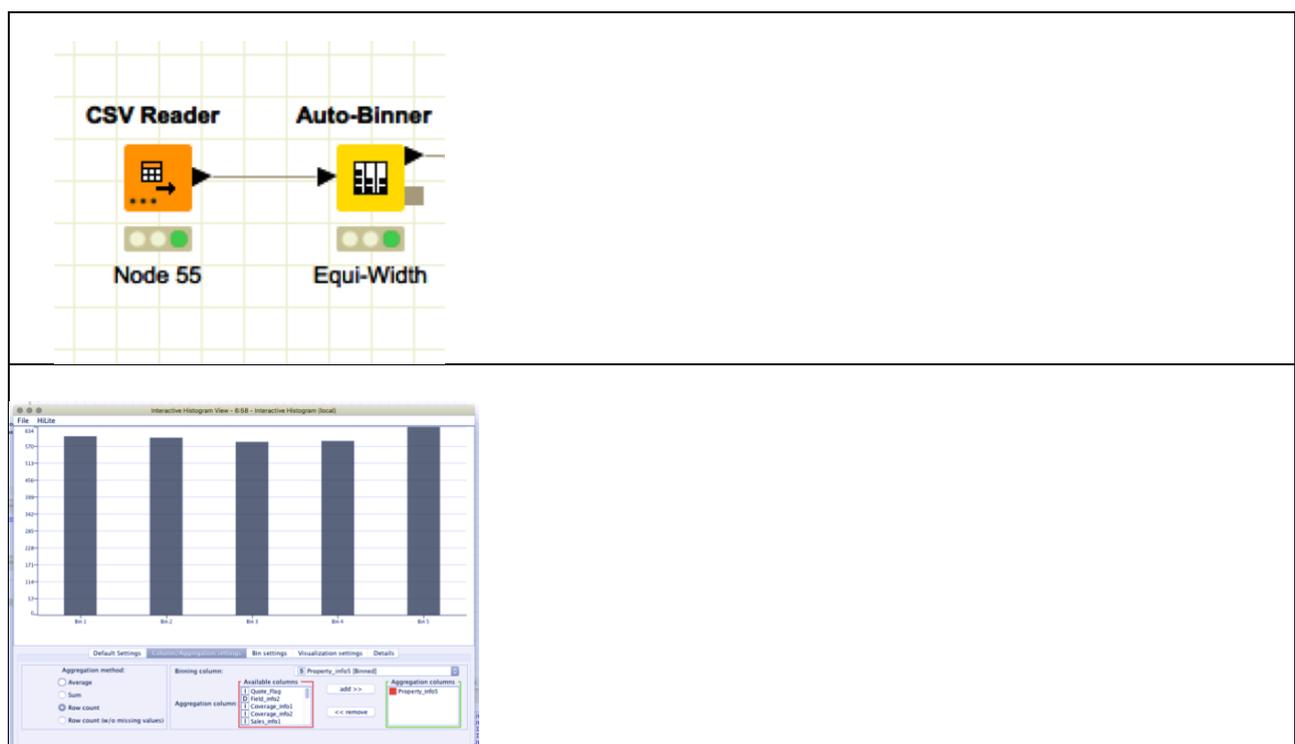


Figure 5 Linear Correlation and Rank Correlation Graphs

1B. Data Pre-processing

1. Binning techniques

Equi-width binning.



Processed data - 6:69 - One to Many (Binarisation)

Table "default" - Rows: 3000 Spec - Columns: 34 Propertie

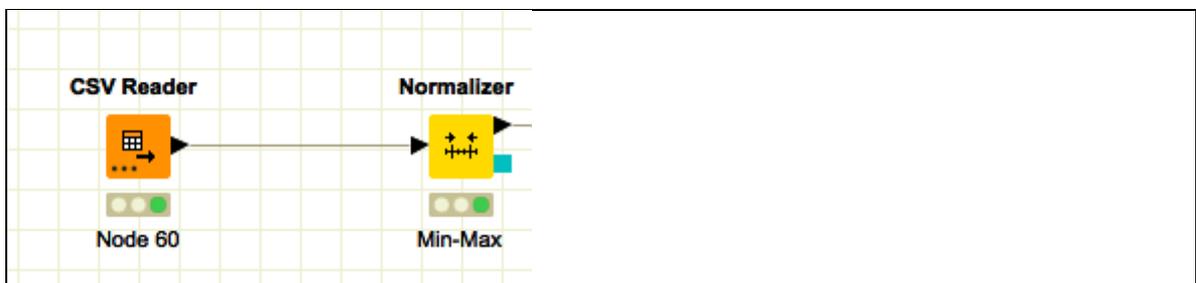
Row ID	Prope...	Prope...	Geogr...	Geogr...	Geogr...	Geogr...	Geogr...
Row0	7	9	25	25	N	IL	
Row1	25	2	4	-1	N	CA	
Row2	15	4	22	-1	N	NJ	
Row3	4	9	23	-1	N	IL	
Row4	25	4	20	-1	N	NJ	
Row5	10	24	13	-1	N	TX	
Row6	24	2	13	-1	N	CA	
Row7	21	4	16	-1	N	NJ	
Row8	20	1	15	-1	N	CA	
Row9	12	9	24	-1	N	IL	
Row10	13	4	19	-1	N	NJ	
Row11	7	2	10	-1	N	CA	
Row12	3	4	18	-1	N	NJ	
Row13	19	14	5	-1	Y	TX	
Row14	17	4	21	-1	N	NJ	
Row15	24	4	22	-1	N	NJ	
Row16	13	2	10	-1	N	CA	
Row17	6	9	23	-1	N	IL	
Row18	10	24	15	-1	N	TX	
Row19	5	24	11	-1	N	TX	
Row20	21	2	4	-1	N	CA	
Row21	5	2	9	-1	N	CA	
Row22	8	4	16	-1	N	NJ	
Row23	10	2	5	-1	N	CA	
Row24	21	2	10	-1	N	CA	
Row25	18	24	14	-1	Y	TX	
Row26	15	23	4	-1	N	TX	
Row27	9	14	8	-1	N	TX	

- **Equi-depth binning:** do the same as equi-width

2. Normalisation

In a Min-max normalisation

The purpose of normalisation is to change the values of numeric columns in the dataset to use a common scale but preventing differences in the ranges of values or losing information.



Processed data - 6:69 - One to Many (Binarisation)

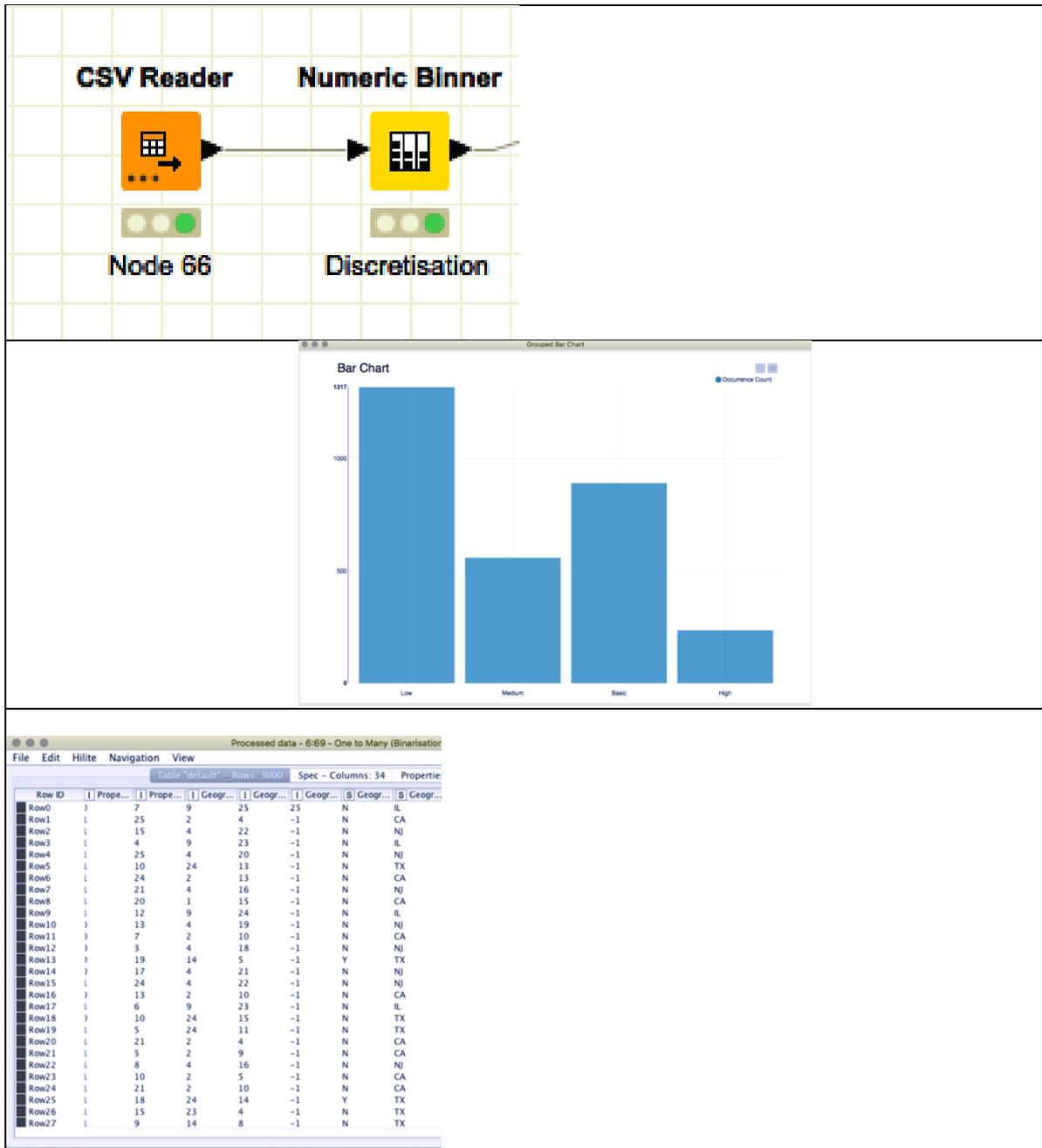
Table "default" - Rows: 3000 Spec - Columns: 34 Propertie

Row ID	Prope...	Prope...	Geogr...	Geogr...	Geogr...	Geogr...	Geogr...
Row0	7	9	25	25	N	IL	
Row1	25	2	4	-1	N	CA	
Row2	15	4	22	-1	N	NJ	
Row3	4	9	23	-1	N	IL	
Row4	25	4	20	-1	N	NJ	
Row5	10	24	13	-1	N	TX	
Row6	24	2	13	-1	N	CA	
Row7	21	4	16	-1	N	NJ	
Row8	20	1	15	-1	N	CA	
Row9	12	9	24	-1	N	IL	
Row10	13	4	19	-1	N	NJ	
Row11	7	2	10	-1	N	CA	
Row12	3	4	18	-1	N	NJ	
Row13	19	14	5	-1	Y	TX	
Row14	17	4	21	-1	N	NJ	
Row15	24	4	22	-1	N	NJ	
Row16	13	2	10	-1	N	CA	
Row17	6	9	23	-1	N	IL	
Row18	10	24	15	-1	N	TX	
Row19	5	24	11	-1	N	TX	
Row20	21	2	4	-1	N	CA	
Row21	5	2	9	-1	N	CA	
Row22	8	4	16	-1	N	NJ	
Row23	10	2	5	-1	N	CA	
Row24	21	2	10	-1	N	CA	
Row25	18	24	14	-1	Y	TX	
Row26	15	23	4	-1	N	TX	
Row27	9	14	8	-1	N	TX	

z-score normalisation: similar to min-max

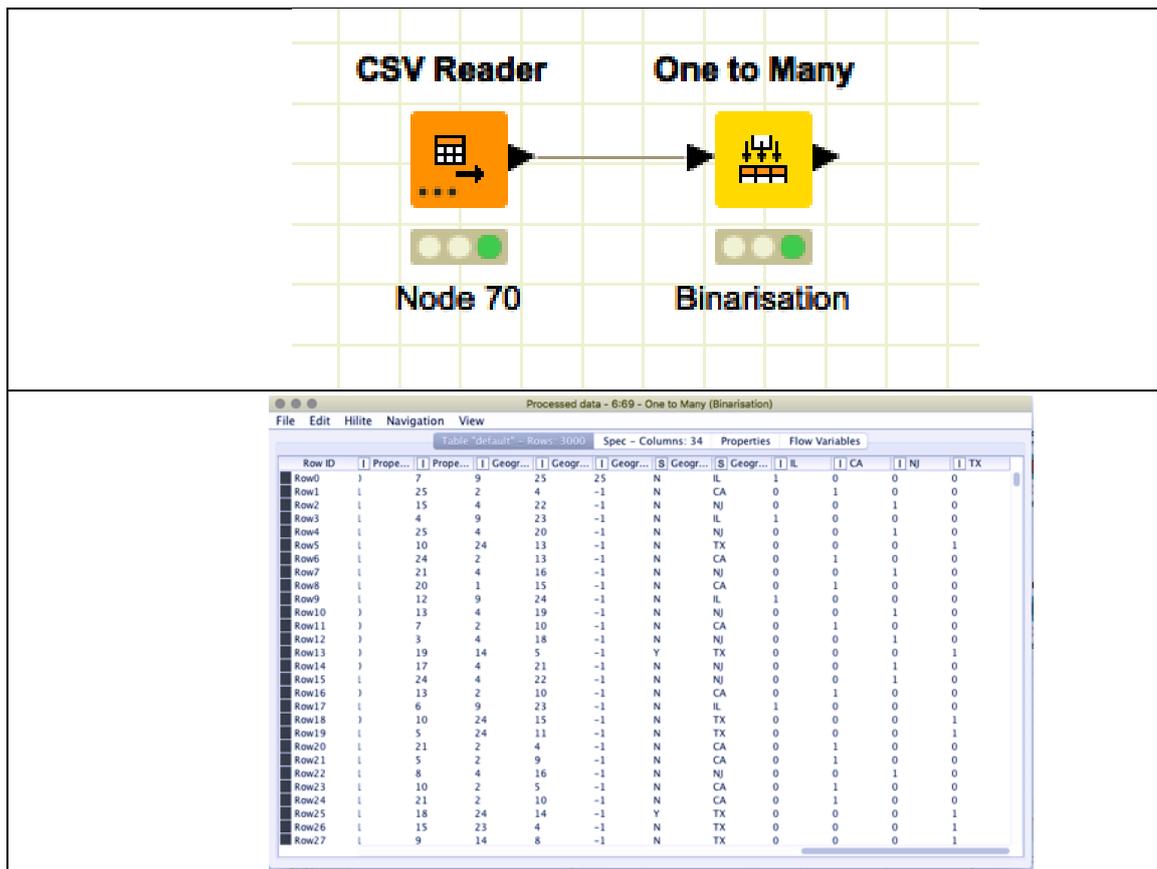
3. Discretisation

Discretisation is the process of converting a continuous attribute into an ordinal attribute to map the number of values into a small number of categories. In this case, discretisation of the Coverage_Info1 attribute is to be done into the following categories: Basic, Low, Medium and High.



1. Binarisation

The purpose of binarisation is to map a continuous or categorical attribute into one or multiple binary variables. To perform the binarisation in the Geographic_Info5 variable [with values "0" or "1"] the following steps were followed in KNIME:



1C. Summary

The most important findings of this report include the following:

- Coverage_info1 attribute: the data ranges from -1 to 25 however, there are only 4 out of 3000 instances of '-1' which is lower than all other values when comparing their frequencies thus very uneven distribution of data. This could indicate that '-1' represents a data collection error rendering it as an outlier as it is possible that the data quality can be affected by human error when recording data.
- unanswered and confidential – this could potentially explain the 45.87% of blank data/missing values in regard to their personal information. This is worth investigating further as to if the data is missing values that need to be recovered.
- By utilising hierarchical clustering and calculating the year difference to today, the scatter plot outlines that the most sales_info5 data that could presumably represent monetary value were provided 6 and 7 years ago. The trend over the latest year represents a decline in the sales_info5 data and thus, possibly lower monetary value incoming from potential customers. This association should be investigated and visually examined more rigorously in the future.
- Personal_info3 is plotted against Geographic_info5, the main clustered data with high similarity is around 'CA' which could potentially represent a geographical location as mentioned in its attribute name and how the people from that state have only selected 'ZA' as their data in personal_info3, this represents a strong correlation between the two attributes. This association should be investigated and visually examined more rigorously in the future.