Assignment : Data Exploration And Preparation

Vraj Mehta (13488642)
Spring 2019
University of Technology Sydney

1A. Initial Data Exploration

Section 1

The following is a table with Attributes and their types(Nominal, Ordinal, Interval, Ratio).

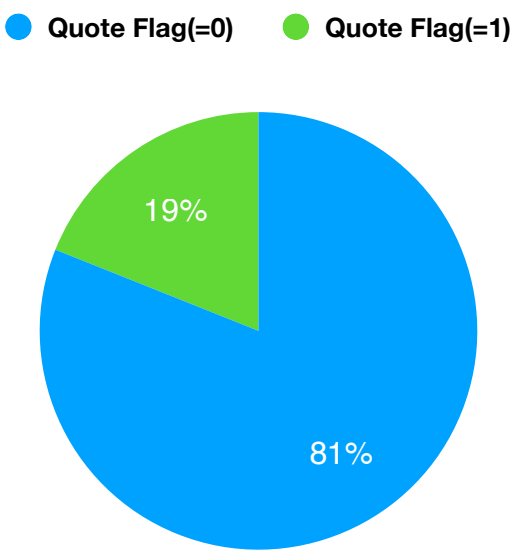
Attribute Name	Attribute Type	Reason
Quote Attributes		
Quote_ID	Nominal	Binary values of '0' & '1' as labels
Quote_Date	Ordinal	Date can be ordered (no fixed distance in the data given- **not interval**)
Quote_Flag	Nominal	Binary values of '0' & '1' as labels
Field Attributes		
Field_Info1	Nominal	Fixed set of values, 'B', 'C','D', 'E', 'F', 'J', 'K' as labels
Field_Info2	Ordinal	Normalised values between 0.8746 to 1.0101(no fixed interval)
Field_Info3	Nominal	String values (no fixed interval)
Field_Info4	Nominal	Binary values of 'Y' & 'N' as labels
Coverage Attributes		
Coverage_Info1	Ordinal	Set of values from -1 & 1 to 25 (0 excluded - not interval)
Coverage_Info2	Ordinal	Numeric values of 1,2,22,5 (no fixed interval)
Coverage_Info3	Nominal	Fixed set of values, from 'A' to 'L' as labels
Sales Attributes		
Sales_Info1	Nominal	Binary values of '0' & '1' as labels
Sales_Info2	Interval	Numeric values of 2,3,4,5 (fixed interval)
Sales_Info3	Ordinal	Numeric values from 1 to 24 except 2 (Not interval)
Sales_Info4	Nominal	Fixed set of values, 'K', 'M','P', 'Q', 'R', 'T', 'V' as labels
Sales_Info5	Ordinal	Many values ranging from 82 to 67153

Attribute Name	Attribute Type	Reason
Personal Attributes		
Personal_Info1	Nominal	Binary values of 'Y' & 'N' as labels
Personal_Info2	Interval	Numeric values from 1 to 25 included (fixed interval)
Personal_Info3	Nominal	Fixed set of values, 'XB', ' XC',….. 'ZW' as labels
Personal_Info4	Nominal	Binary values of '0' & '1' as labels
Personal_Info5	Nominal	Only single value '2' as a label
Property Attributes		
Property_Info1	Nominal	Binary values of 'Y' & 'N' as labels
Property_Info2	Nominal	Only single value '0' as a label
Property_Info3	Nominal	Fixed set of values, 'A', ' D', … 'S' as labels
Property_Info4	Nominal	Binary values of '0' & '1' as labels
Property_Info5	Interval	Numeric values from 1 to 25 included (fixed interval)
Geographic Attributes		
Geographic_Info1	Ordinal	Set of values from -1 & 1 to 25 (0 excluded - not interval)
Geographic_Info2	Ordinal	Set of values from -1 & 4 to 25 (0,1,2,3 excluded - not interval)
Geographic_Info3	Nominal	Fixed set of values, '-1' & '25' as labels
Geographic_Info4	Nominal	Binary values of 'Y' & 'N' as labels
Geographic_Info5	Nominal	Fixed set of values (US state names), 'CA', 'IL', 'NJ', 'TX' as labels

Note - I have used Tableau for visualisation and taken screenshot of that, please zoom if not visible clearly.

1A. Section 2

(i) Quote Flag : Indicates whether a person bought the insurance policy or not.



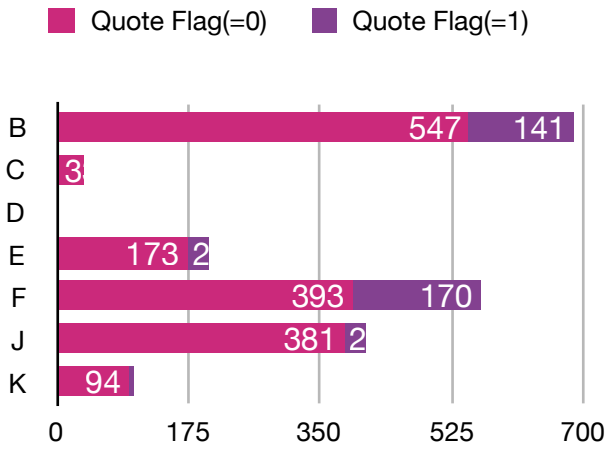
This pie chart depicts the **81%** *(1621/2000)* people did not bought the insurance policy.(Quote Flag = 0)
While, **19%** *(379/2000)* people did buy the policy. (Quote Flag = 1)

The most important attribute as ultimately we want to check people who bought the policy, and important to note how other attributes affect **Quote_Flag** attribute.

(ii) Quote Date : The date when the insurance policy was purchased or not. Following is a *Treemaps* showing *policy purchases* with respect to *month.*

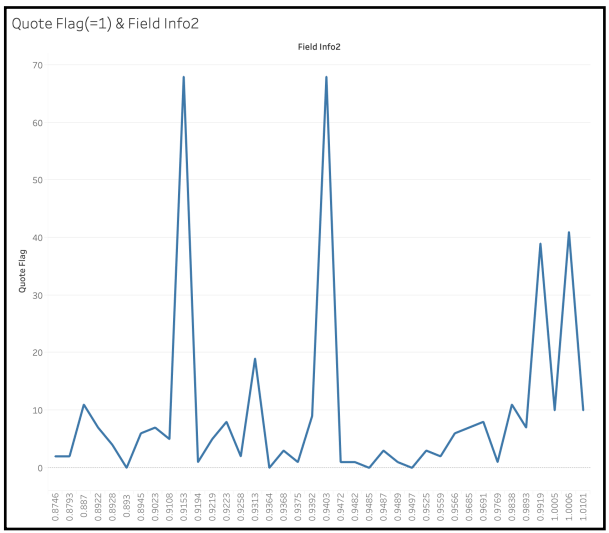
April(46 occurrences) & May(44 occurrences) are the months when the most people bought the policy. Whereas in **July(16 occurrences)** least people bought the policy. This is calculated over a period of 3 years from 2013-2015.

(iii) Field Info1 : Set of Seven labels, 'B', 'C', 'D', 'E', 'F', 'J' & 'K'.



B has the highest occurrence with total of **688 times.** But **F** having occurrence **563 times, 170** people bought the policy while only **141** people bought policy incase of **B.**
D has only **1** occurrence, least important field.

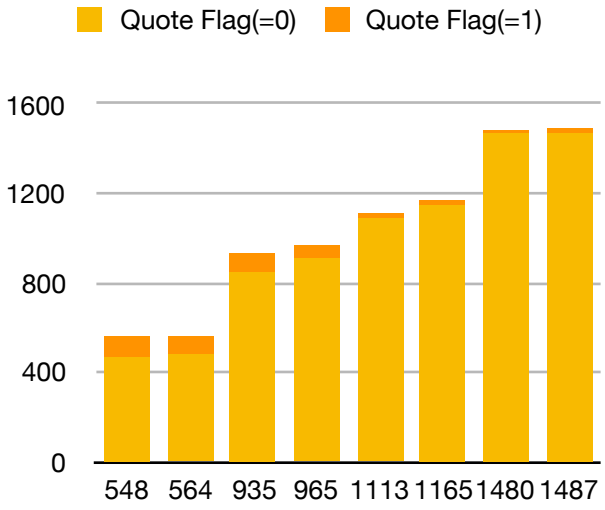
(iv) Field Info2 : Seems to be **Normalised values** ranging from **0.8746 to 1.0101 .**



No of Quote Flag(=1) for **Field Info2=0.9153 & 0.9403 is 68.** While majority of others are having *Quote Flag(=1) < 10.*
These are the most important values in Field Info2.

Mean = *0.9391*
Standard Deviation = *0.0377*

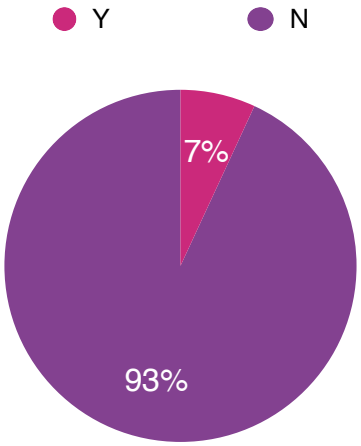
(v) Field Info3 : Set of values eight values 548, 564, 935, 965, 1113, 1165, 1480, 1487.



The histogram clearly depicts that *frequency* is **increasing** with increasing values from *548 to 1487.* *Conversely,* the frequency of that covering to buying policy is opposite to that.
As the Value in Field Info3 increases Quote Flag(=1) decreases. **Quote Flag(=1) is the highest for 548 and the least for 1487.**

Hence, Quote Flag(=1) & Field Info3 are *inversely proportional* .

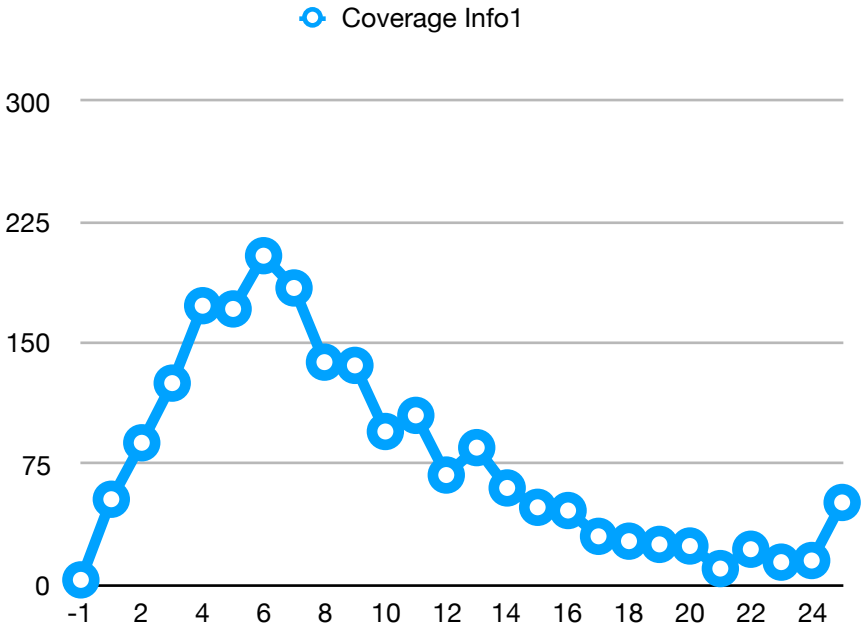
(vi) Field Info4 : Binary values of 'Y' & 'N' of Nominal type.



The pie chart depicts that majority of the entries around **93% are 'N'** and only **7% are 'Y'**.

Field Info4	Frequency
N	1862
Y	138

(vii) Coverage Info1 : Range of Ordinal values from 1 to 25 including (-1).

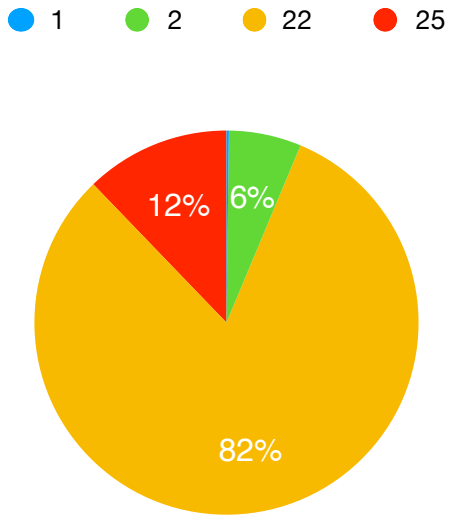


The line chart shows the occurrences of values from 1 to 25 including (-1).

'6' having the *highest (204 occurrence)*
'-1' having the *least (3 occurrence)*

Range = 26
Mean = 8.956
Standard Deviation = 5.662.

(vi) Coverage Info2 : Four values of Ordinal type, '1', '2', '22', '25'.

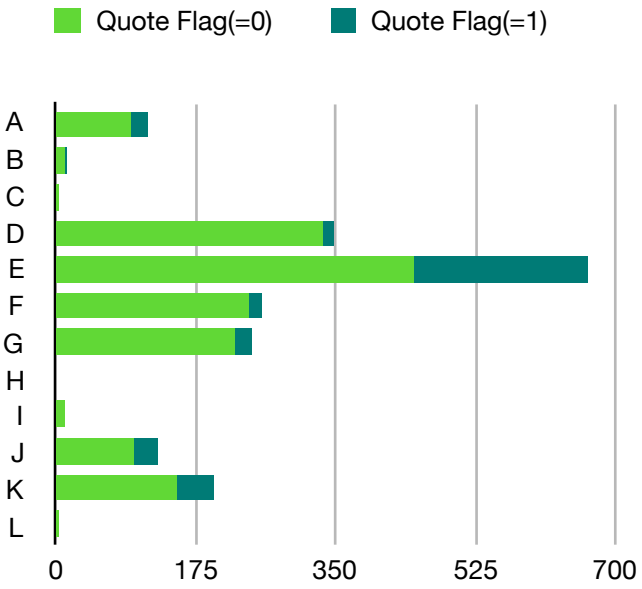


Coverage Info2	Frequency
1	4
2	122
22	1630
25	244

The pie chart depicts the frequency of values occurring. '**22'** is the *most frequently* occurring value with around **82%**(1630/2000) and **'1'** being the l*east occurring* value with less than **<1%.**

Mean = 21.104
Standard Deviation = 5.059

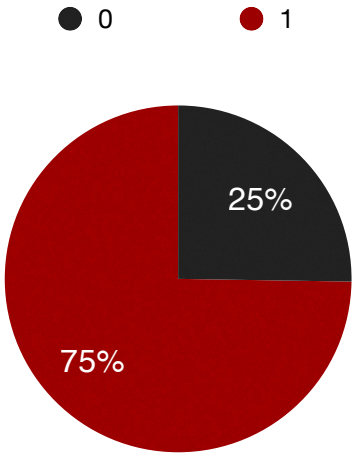
(vii) Coverage Info3 : Nominal variables from 'A', 'B', 'C', …. , 'L'.



The bar graph represents that variable **'E'** has the *highest occurrence* and the *highest quote flag(=1).* **'H'** has the least occurrence.

Coverage Info3	Frequency
A	116
B	15
C	6
D	356
E	663
F	258
G	243
H	1
I	12
J	129
K	197
L	4

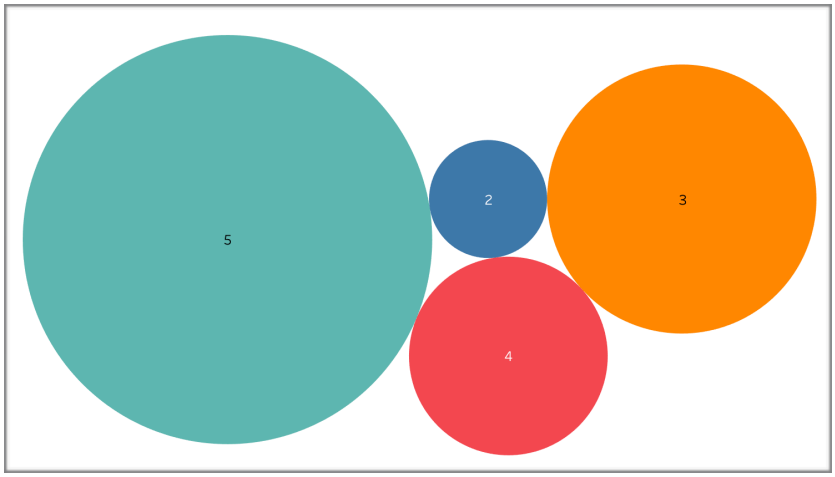
(viii) Sales Info1 : Nominal variable with '0' and '1' values.



The pie chart depicts that majority of the entries around 75**% are 'Y'** and the rest **25% are 'N'**.

Sales Info1	Frequency
0	505
1	1495

(ix) Sales Info2 : Interval variable with four values, 2, 3, 4, 5 .

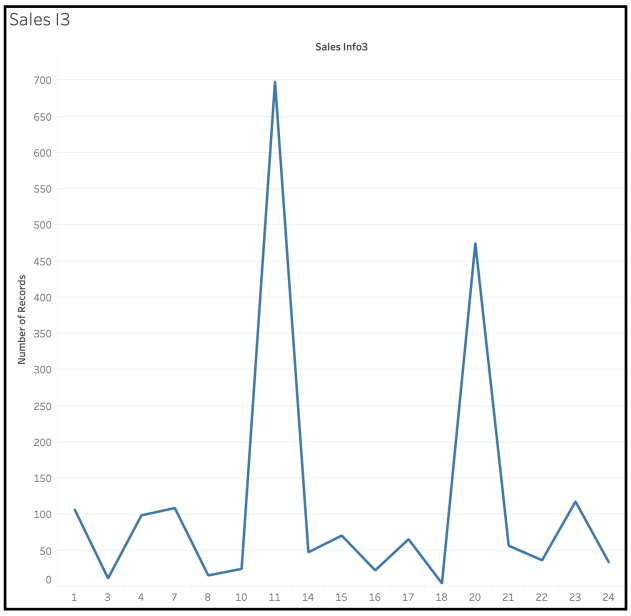


Packed Bubbles depict the frequency of the variables.
The larger the bubble, more is the frequency.
'5' has the *highest* and **'2'** has the *lowest* frequency.

Mean = 4.229
Std Dev=0.978

Sales Info2	Frequency
2	95
3	494
4	269
5	1142

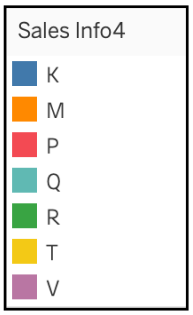
(x) Sales Info3 : Ordinal set of values from 1 to 25 excluding 2.

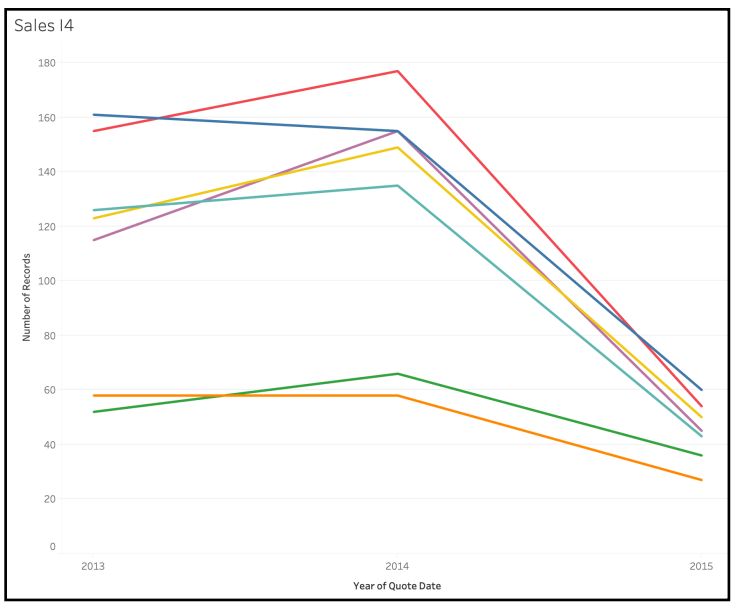


Sales Info3 =11 is having the *highest occurrence* (*698 times*) & **Sales Info3=20** having occurrence of *475 times* as shown in fig.

Range = 24 (Max-Min = 25-1)
Mean = 13.858
Standard Deviation = 6.2485

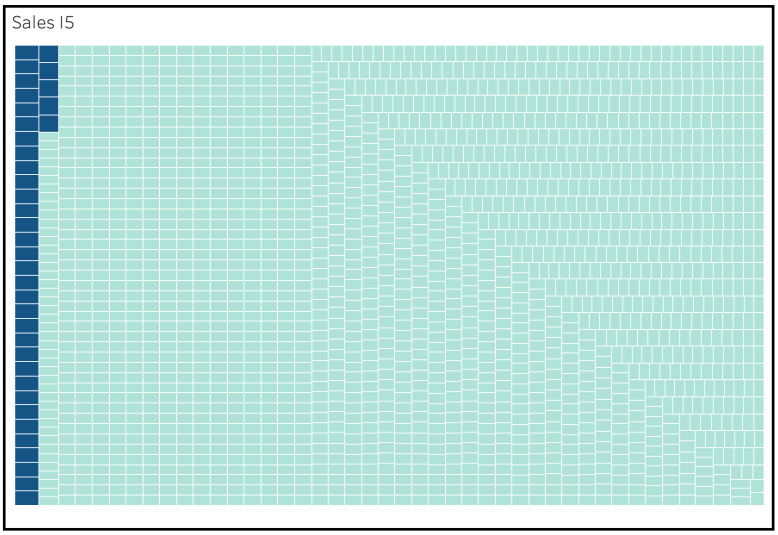
(xi) Sales Info4 : Set of Nominal values, 'K', 'M', 'P', 'Q', 'R', 'T', 'V'.

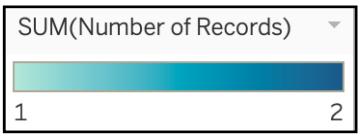




This line chart depicts the **Sales Info4** with respect to **year** from 2013 to 2015.
We can see that overall sales reached a peak in 2014 but declined significantly from there in 2015.
P is having the *highest sales*, while **M** having the *lowest sales.*

(xii) Sales Info5 : Ordinal values ranging from 82 to 67153.

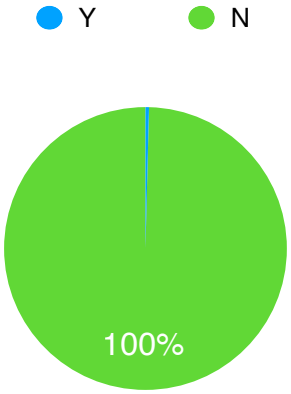




This Treemaps depicts that most of the values are having *frequency of 1.*
There are only **35 values** in the range specified that have *frequency of 2.*
(Dark Blue Colour- frequency 2)
(Light Green Colour - frequency 1)

Range = 67065
Mean = 34016
Standard Deviation = 19067

(xiii) Personal Info1 : Binary values of 'Y' & 'N' of Nominal type.

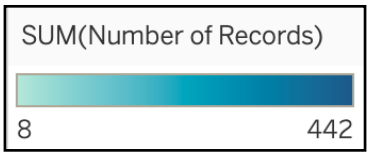


This pie chart depicts that around **99.6 %** values are *'N' (1992/2000)*
And only **0.4%** values are *'Y' (8/2000)*

There is **one missing value** which I have replaced by the most *frequently occurring number.*

(xiv) Personal Info2 : Ordinal values in the range of 1 to 25 including (-1).

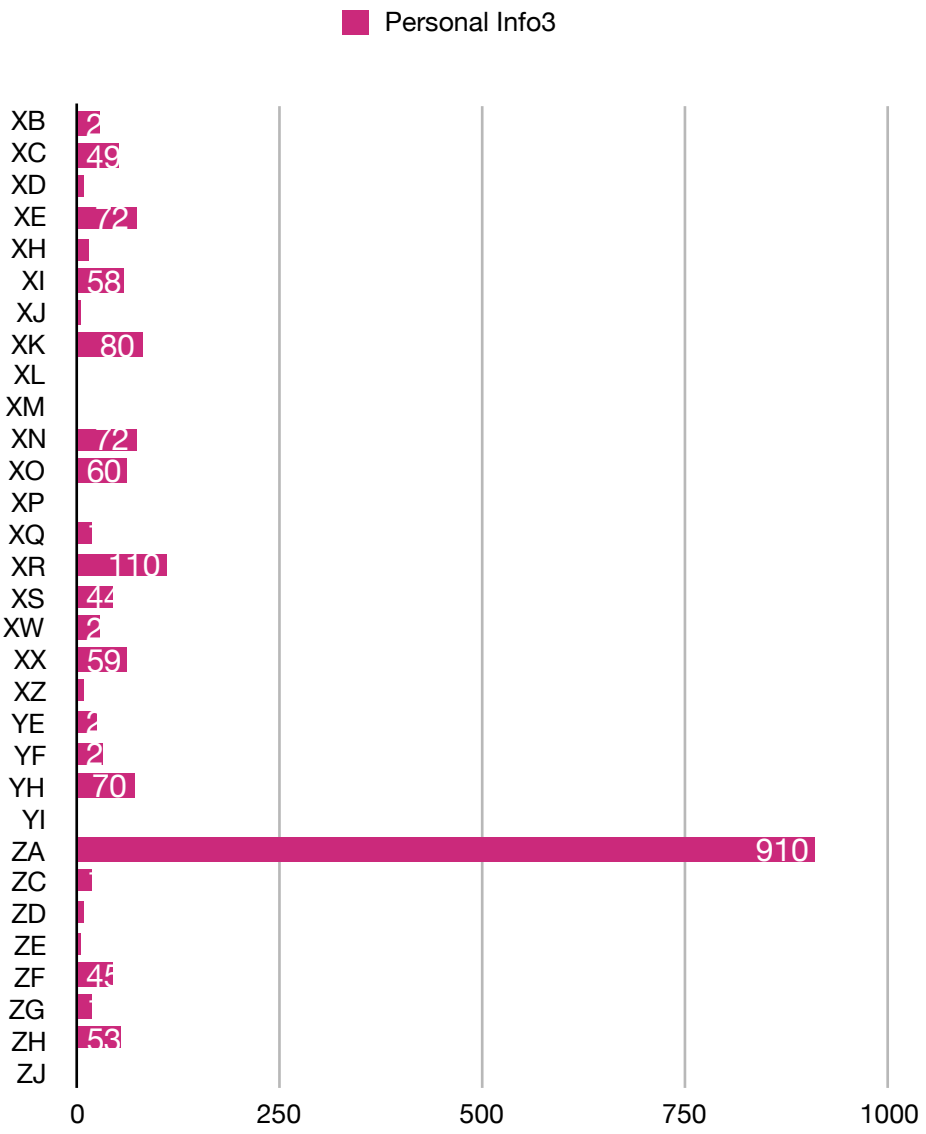
Personal Info2	
-1	442
1	43
2	8
3	21
4	115
5	335
6	277
7	189
8	77
9	33
10	34
11	36
12	40
13	35
14	22
15	31
16	34
17	16
18	35
19	28
20	20
21	24
22	14
23	24
24	28
25	39



This table represents the frequency of values in Personal Info2.
'-1' is having the *highest frequency* followed by **'5' & '6'.**
'2' is *lowest* occurring value.

Range = 26
Mean = 6.773
Standard Deviation = 6.712

(xv) Personal Info3 : Nominal values.



'ZA' has the highest occurrence *910 times*.

Rest of the labels are having occurrence of *<110 times.*

'XK', 'YI', 'ZJ' having the least occurrence *1 time only.*

(xvi) Personal Info4 : Nominal variable with '0' and '1' values.

Personal Info4	
0	1,998
1	2

99.9 % of values are **'0',** while only **0.01%** of values are **'1'.**

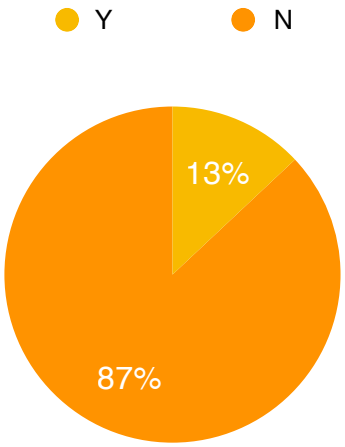
(xvii) Personal Info5 : Nominal variable with '1' and '2' values.

Personal Info5	
Null	983
1	12
2	1,005

Almost **50%** of data is ***missing(null)***. So won't be ideal to predict or replace these values with most frequent, or any other method.

50% of values are **'2'** and only **0.6%** of values are b.

(xviii) Property Info1 : Nominal values having **'Y' & 'N'.**

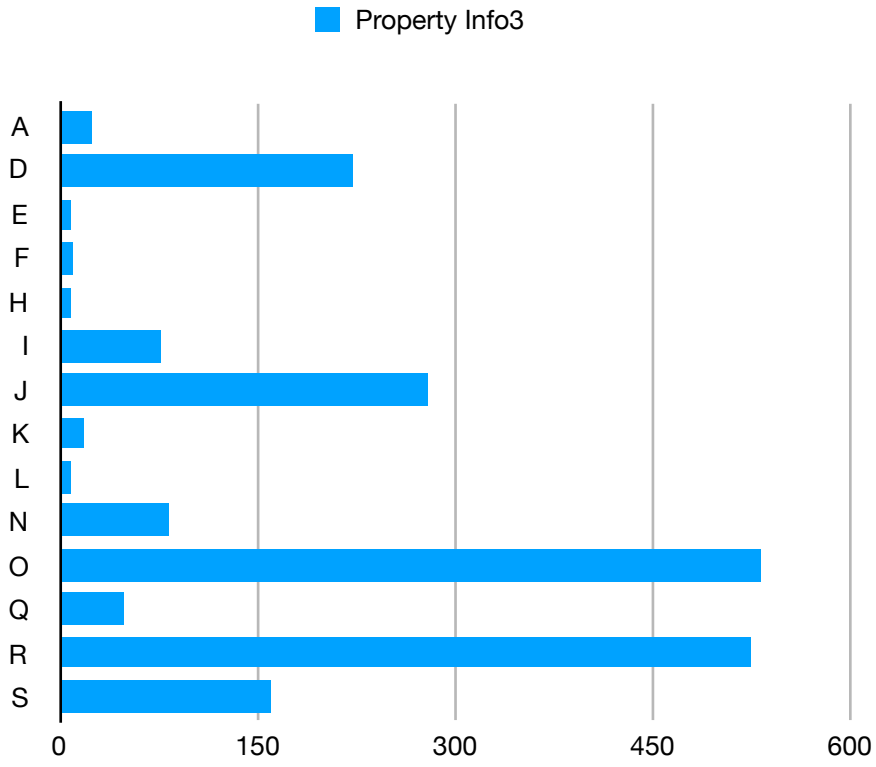


The pie chart depicts that **87%** values are **'N'** & **13 %** values are **'Y'.**

Property Info1	Frquency
Y	260
N	1740

(xix) Property Info2 : ***All*** *the values are* ***zero***. No other values are present in the data.

(xx) Property Info3 : Nominal values, **'A', 'D', … , 'S'.**



Property Info3	
A	24
D	223
E	8
F	9
H	8
I	77
J	280
K	17
L	8
N	82
O	532
Q	48
R	525
S	159

'O' is having the *highest occurrences* followed by **'R'.**
'H' & 'L' has the *lowest occurrences.*

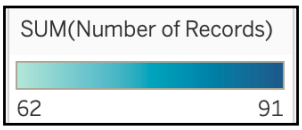
(xxi) Property Info4 : Nominal variable having **'0' and '1' .**

Property Info4	
0	673
1	1,327

66.35 % of values are having label as **'1',** while **33.65**% of values are having label as **'0'.**

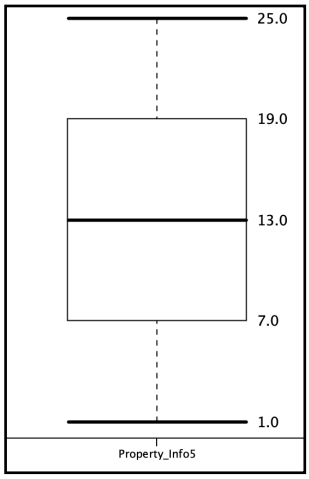
(xxii) Property Info5 : Interval type ranging from, 1 to 25 (both included).

Property Info5	
1	77
2	85
3	82
4	64
5	82
6	72
7	78
8	69
9	88
10	91
11	79
12	90
13	91
14	86
15	82
16	87
17	91
18	70
19	85
20	62
21	68
22	90
23	81
24	72
25	78

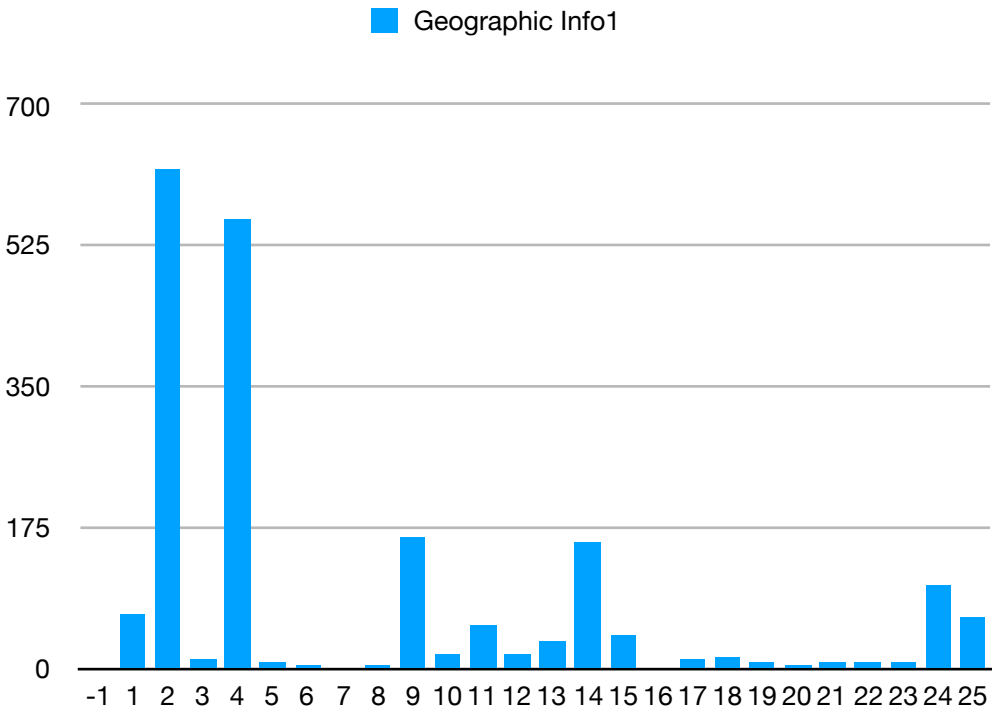


This table represents the *frequency* of values in *Property Info5*.
Almost all the values are having frequency that is **equally distributed** between **62 to 91.**

Range = 24 *(25-1)*
Mean = 12.98
Standard Deviation = 7.1041



(xxiii) Geographic Info1 : Ordinal variable having values from 1 to 25 including (-1)

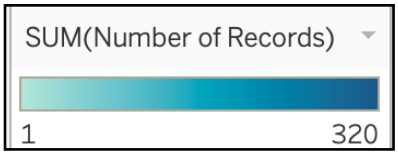


Geographic ..	
-1	1
1	68
2	620
3	12
4	557
5	9
6	5
7	2
8	5
9	164
10	19
11	55
12	18
13	33
14	157
15	42
16	1
17	12
18	16
19	9
20	4
21	7
22	8
23	8
24	104
25	64

'2' & '4' have the highest occurrence with 620 and 557 times respectively.
'-1' has the least occurrence with only 1 time.
Range = 26
Mean = 7.403
Standard Deviation = 6.9969

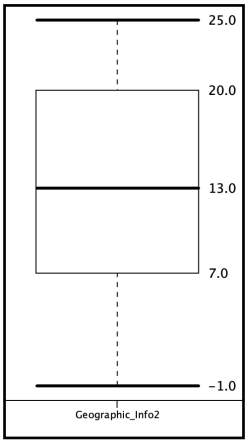
(xxiv) Geographic Info2 : Ordinal set of values from (-1) & 4 to 25 (0,1,2,3 excluded)

Geographic Info2	
-1	1
4	320
5	82
6	82
7	82
8	50
9	98
10	60
11	77
12	79
13	96
14	79
15	70
16	81
17	65
18	78
19	94
20	101
21	82
22	89
23	75
24	77
25	82



The tables shows that **'4**' has *highest frequency.*
All the other values of Geographic Info2 except -1 lie in the range of **50 to 101** in frequency.
'-1' has the least frequency of 1.

Range = 26
Mean = 13.375
Standard Deviation = 6.904



(xxv) Geographic Info3 : Nominal variable with '-1' and '25' values.

Geographic Info3	
-1	1,937
25	63

96.85% values are **'-1'**, while only **3.15%** values are **'25'.**

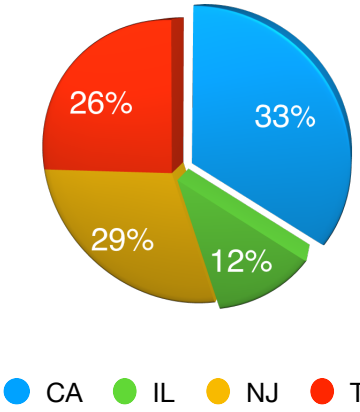
(xxvi) Geographic Info4 : Nominal variable with 'Y' and 'N' values.

Geographic Info4	
Null	1
N	1,949
Y	50

97.5% values are **'N'** , while **2.5%** values are **'Y'.**

There is **one missing value** which I have replaced by the most *frequently occurring value.*

(xxvii) Geographic Info5 : Nominal set of values having US names, **'CA', 'IL', 'NJ' & 'TX'.**

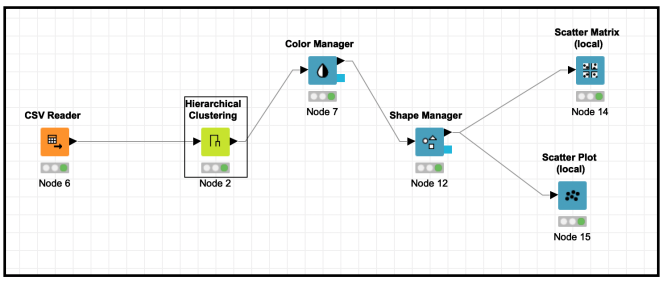


Geographic Info5	
CA	688
IL	236
NJ	564
TX	512

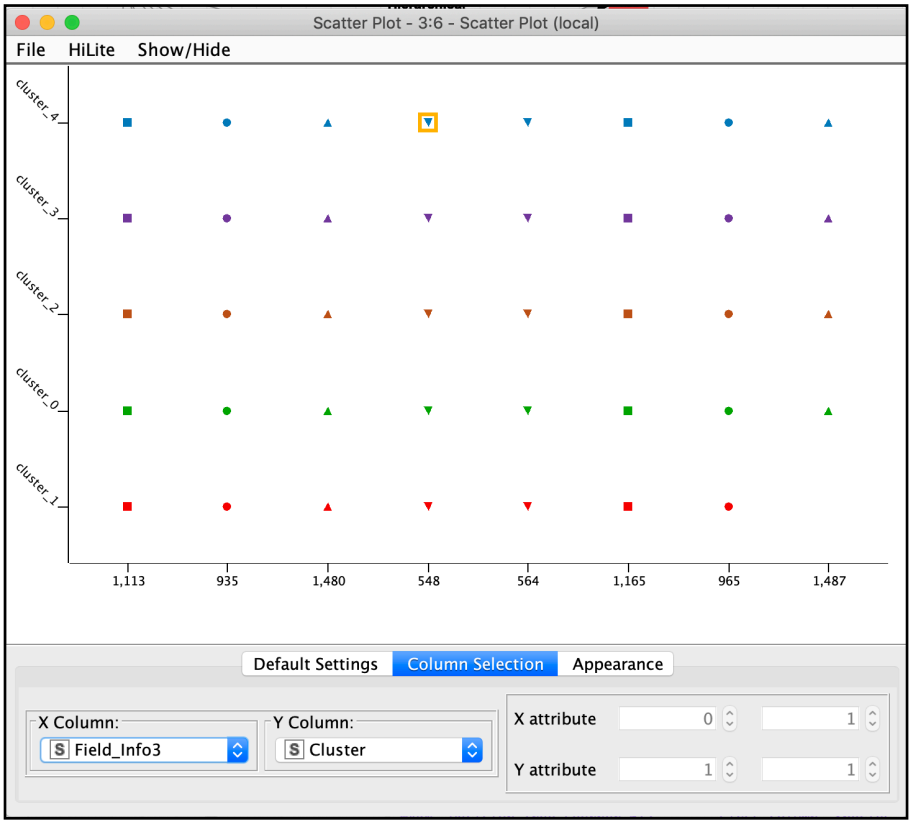
'CA' is having the most occurrences(33%) as shown the pie chart and table.

Whereas, **'IL'** is having the least occurrences (12%).

1A. Section 3



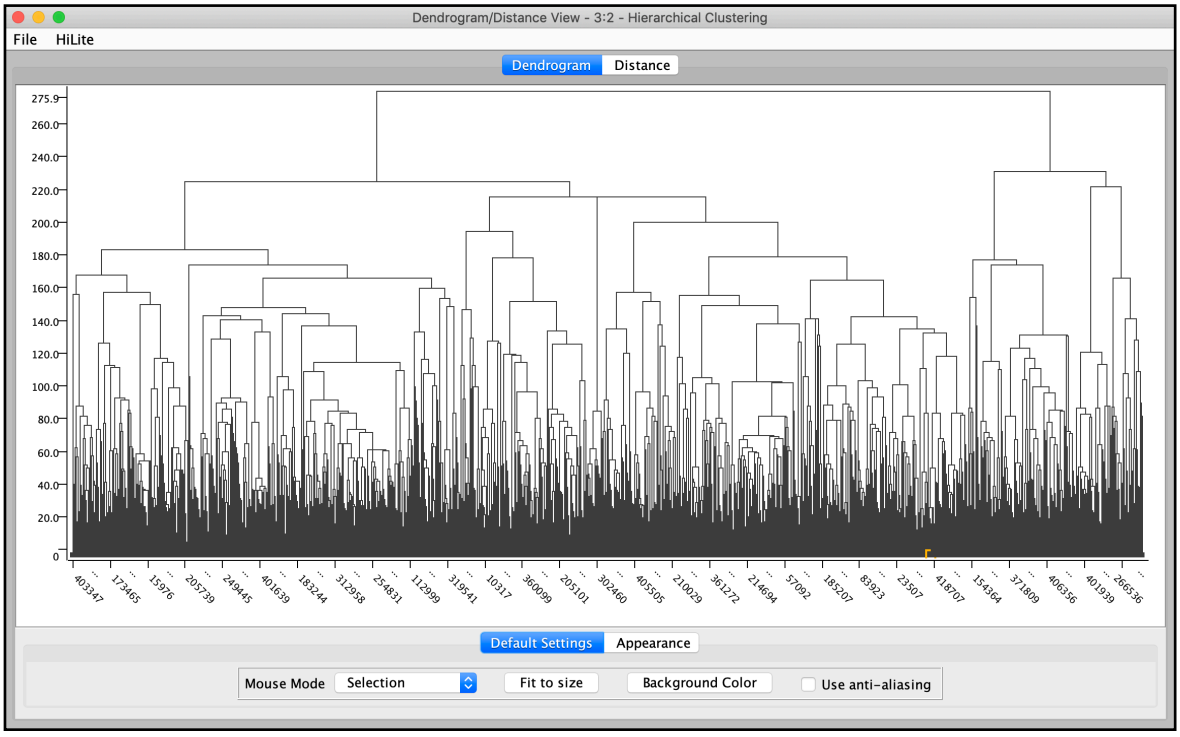
Clustering is done using **Hierarchical Clustering** Node.

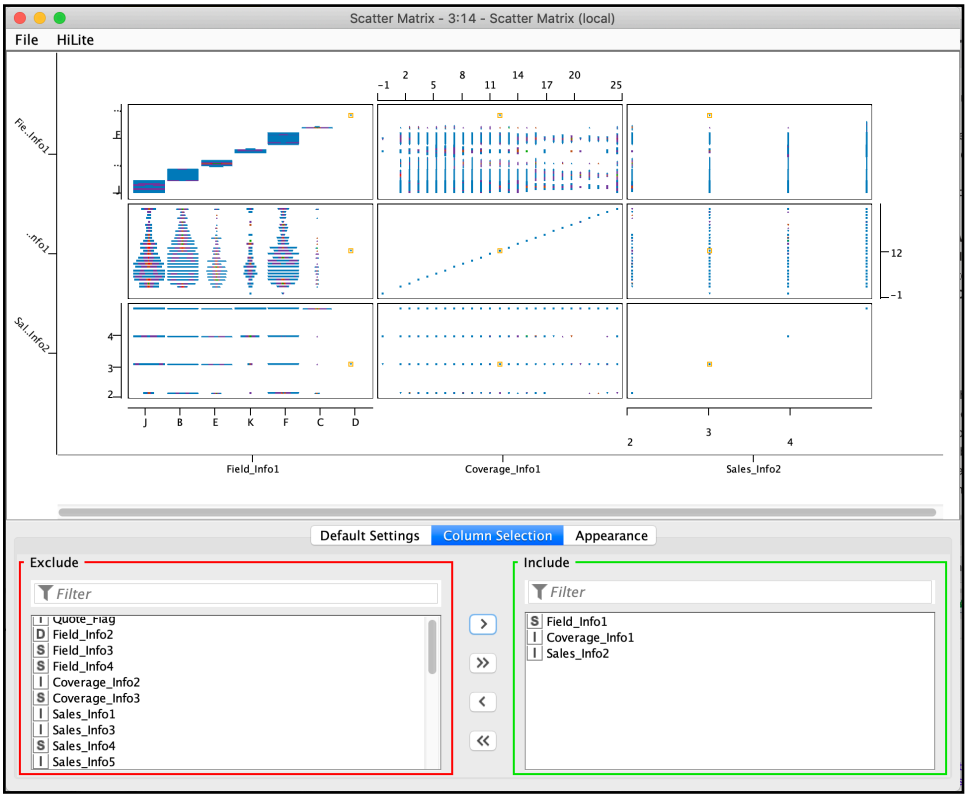


I have chosen **5** clusters. When plotted using a scatter plot, the **same colour shapes** belong the **same cluster**.
The histogram represents 5 clusters named **cluster_0, cluster_1,…, cluster_4** for **Field_Info3.**

*Check the *excel sheet* for **clusters** corresponding to **each data point** name **1A_Clusters.**

From the **Dendrogram** we can clearly see that there are 5 major clusters.



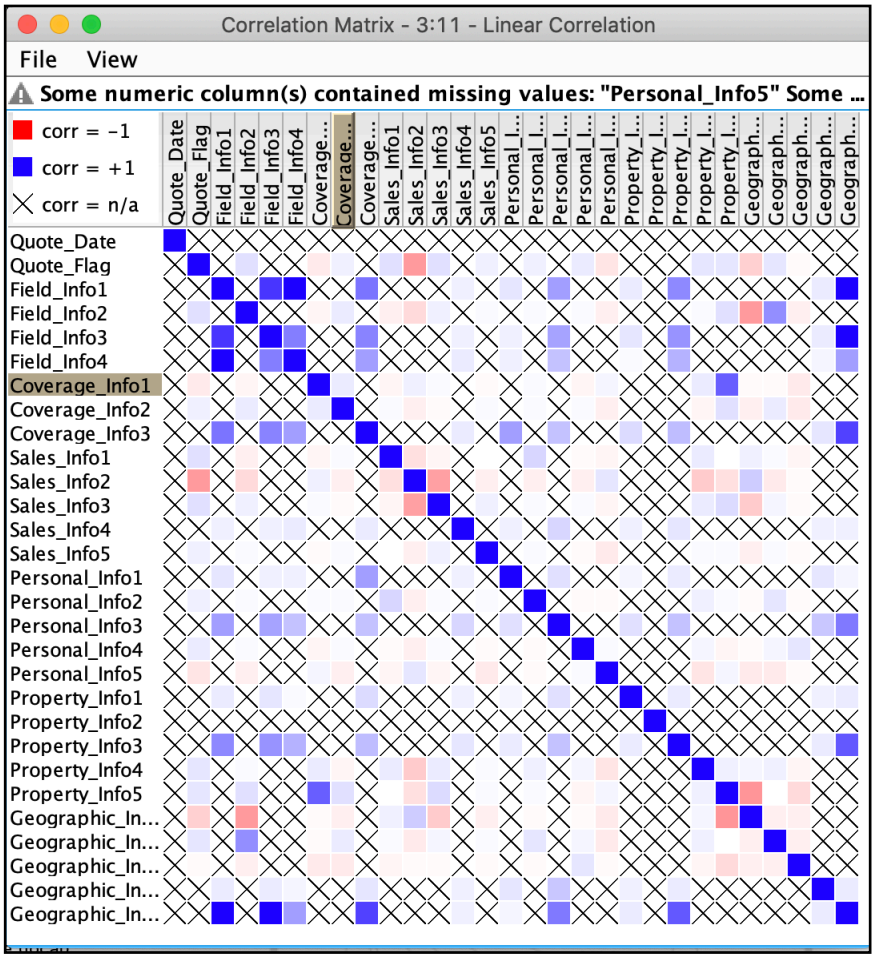


For finding outliers I have used **Scatter Matrix**, plotted a matrix using three attributes. **Field_Info1, Coverage_Info1 & Sales_Info2** as seen in the figure.

I have *highlighted* the outlier. It basically distinguish itself in each of the matrix expect the diagonal.

***Check the highlighted point in the figure (Scatter Matrix), that is the outlier.**

Row ID	Quote...	Quote...	Field_I...	Field_I...	Field_I...	Field_I...	Cover...	Cover...	Cover...	Sales_...	Sales_...
257566	13/10/13	0	B	0.94	965	N	10	22	D	1	5
359574	6/2/14	1	D	0.969	548	N	12	22	B	1	3
133397	13/3/15	1	B	0.915	935	N	6	22	E	1	2
223107	27/12/14	0	B	0.915	935	N	4	22	D	1	5



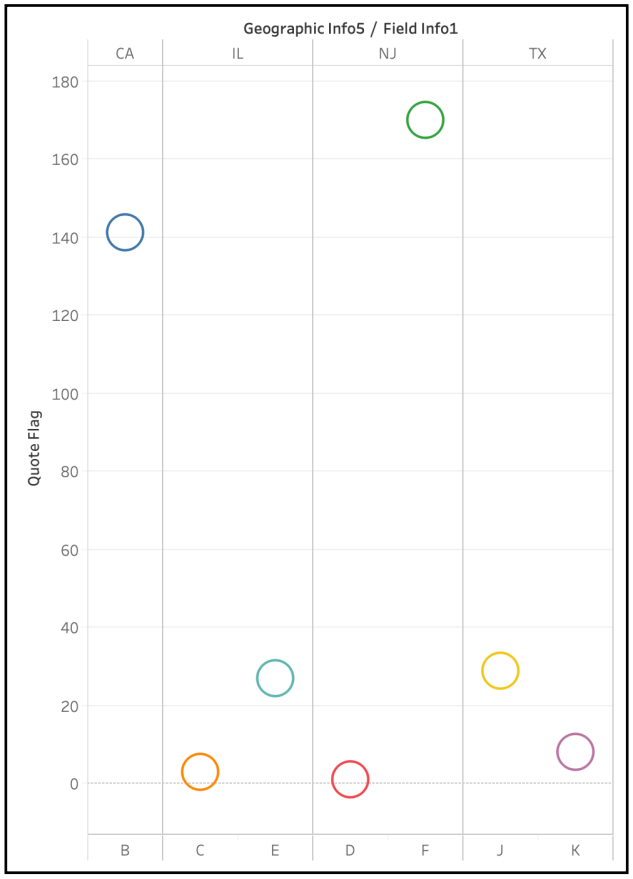
***Quote_ID** (Row ID) = **359574** is the **outlier** in the dataset.

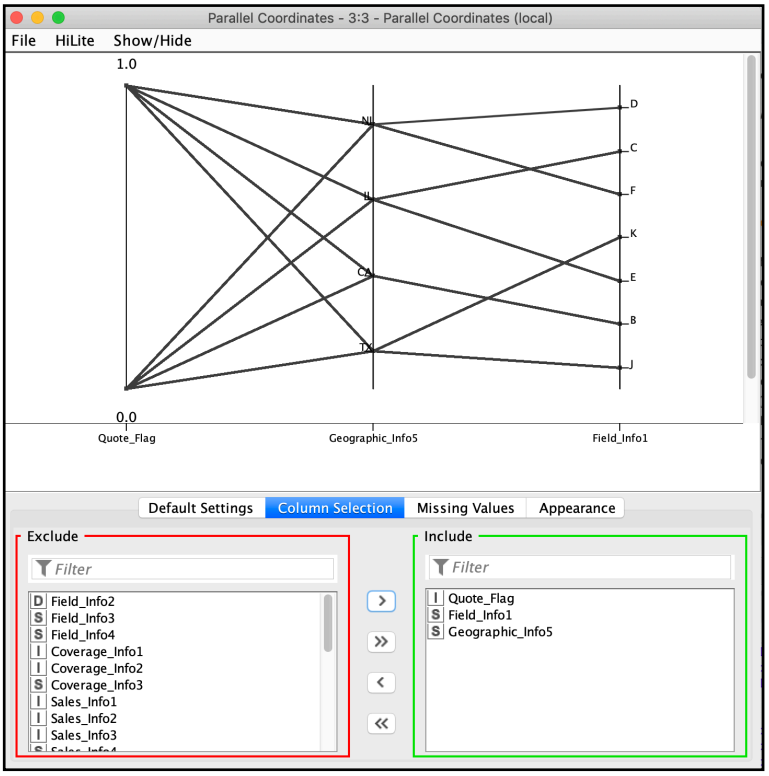
The **correlation matrix** node shows the pair of attributes that are correlated to each other as shown in the figure below.

The pairs that are **highly correlated** according to correlation matrix:

1)Geographic_Info5 & Field_Info1
2)Geographic_Info5 & Field_Info3
3)Field_Info1 & Field_Info4

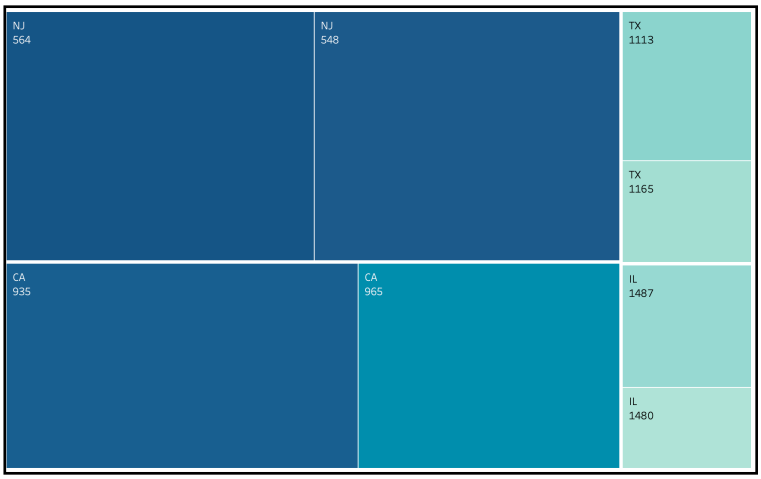
The above figure is plotted using





Parallel Coordinates node. I have included three attributes Quote_Flag, Field_Info1 & Geographic_Info5. All the four values in Geographic_Info5, **'CA', 'IL', 'NJ', 'TX'** have Quote_Flag as 0 & 1. But when corresponded to **Field_Info1,** Each value in *Geographic_Info5* corresponds to **distinct values** of *Field_Info1*. **'B'** belongs to only **'CA'**, **'C' & 'E'** belongs to only **'IL'**, **'D' & 'F'** belongs to only **'NJ'** and **'J' & 'K'** belongs only to **'TX'.**

*Check the bar-circle graph for proper visualisation along with Quote_Flag.
It show that **Geographic_Info5 = 'NJ' & Field_Info1='F'** has the **highest** people buying the insurance policy.



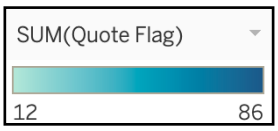
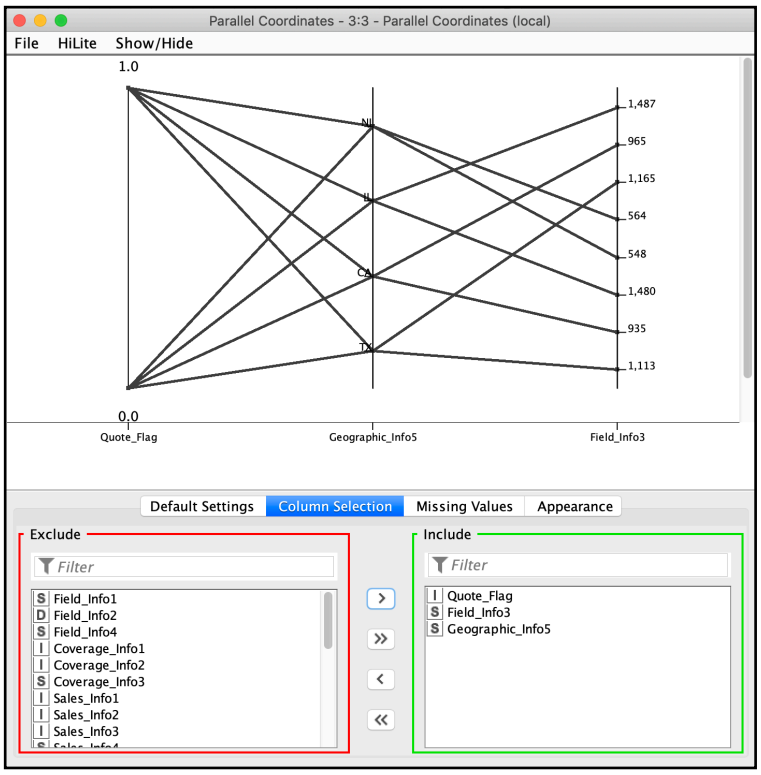
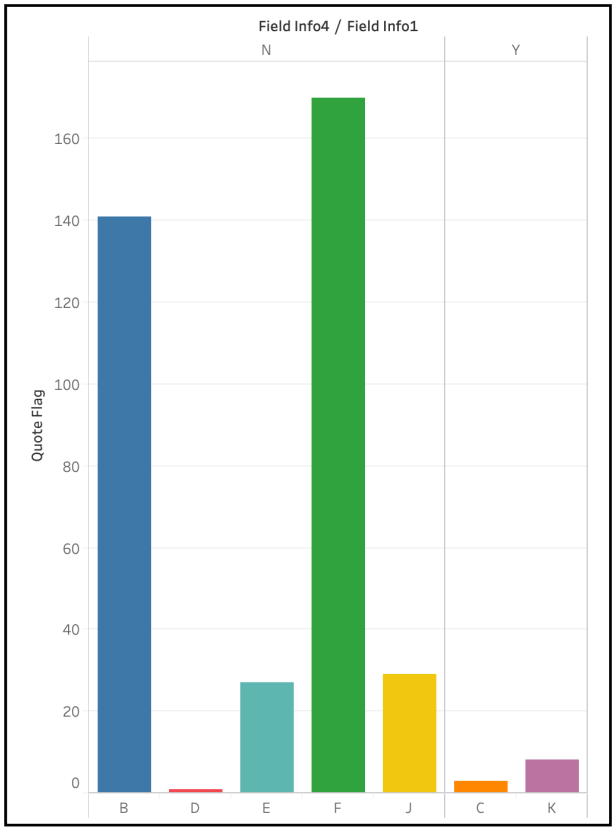


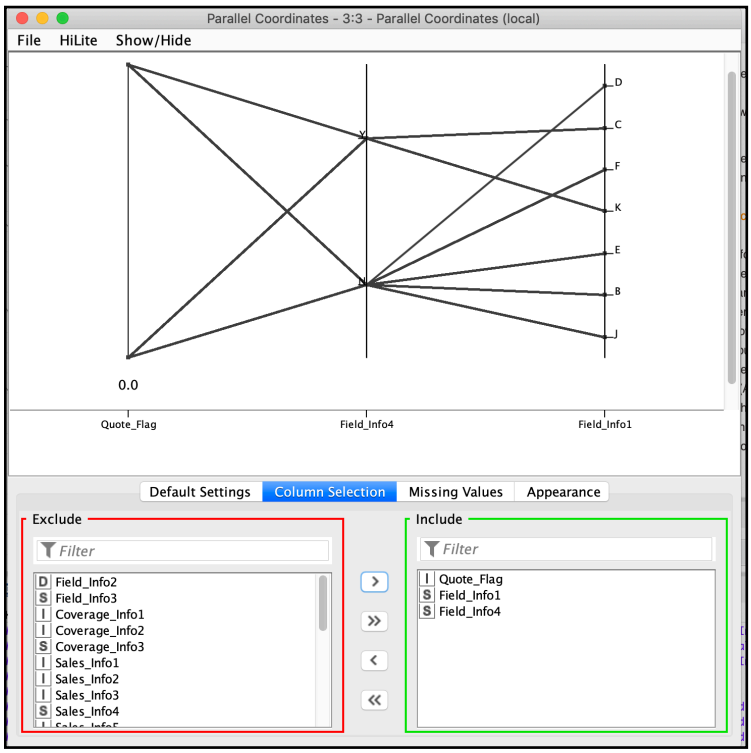
Fig. Treemaps



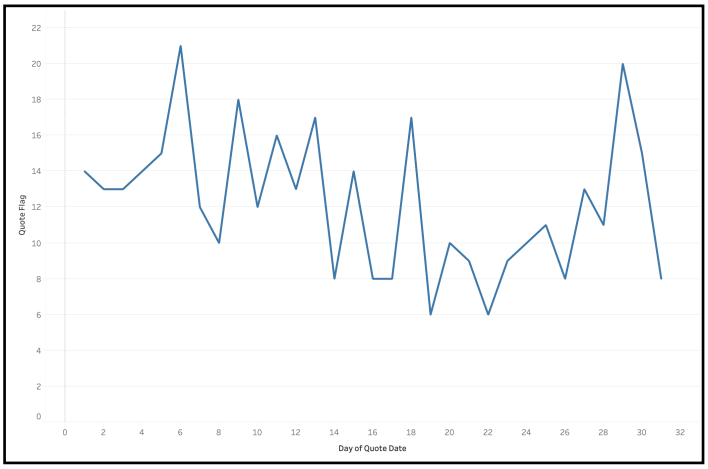
The above figure is plotted using **Parallel Coordinates** node. I have included three attributes Quote_Flag, Field_Info3 & Geographic_Info5. All the four values in Geographic_Info5, **'CA', 'IL', 'NJ', 'TX'** have Quote_Flag as 0 & 1. But when corresponded to **Field_Info3**, each value in *Geographic_Info5* corresponds to **distinct values** of *Field_Info3*. **'965' & '935'** belongs to only **'CA'**, **'1487' & '1480'** belongs to only **'IL'**, **'564' & '548'** belongs to only **'NJ'** and **'1165' & '1113'** belongs only to **'TX'.**

*Check the **Treemaps figure** for proper visualisation along with Quote_Flag.
It shows that **Geographic_Info5 = 'NJ' & Field_Info3='564'** has the **highest** people buying the insurance policy.

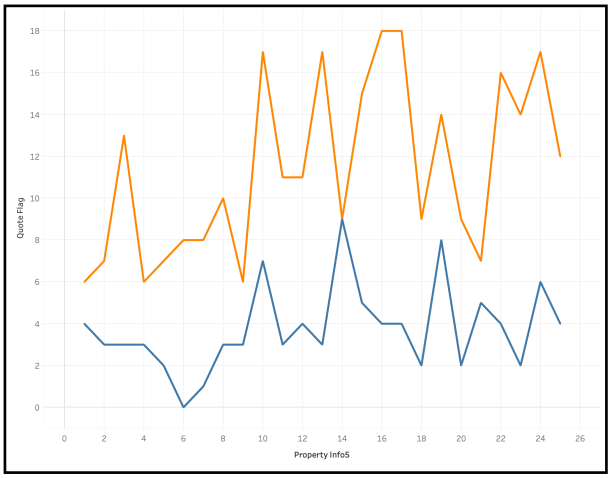




The above figure is plotted using **Parallel Coordinates** node. I have included three attributes Quote_Flag, Field_Info4 & Field_Info1. Both the values in Field_Info4, **'Y' & 'N'** have Quote_Flag as 0 & 1. But when corresponded to **Field_Info1**, each value in *Field_Info4* corresponds to **distinct values** of *Field_Info1*. **'C' & 'K'** belongs to only **'Y'**, **'B', 'D', 'E', 'F', 'J'** belongs to only **'N'.**
*Check the **Bar Graph** for proper visualisation along with Quote_Flag.
It shows that **Field_Info4= 'N' & Field_Info1='F'** has the **highest** people buying the insurance policy.

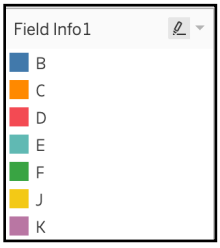


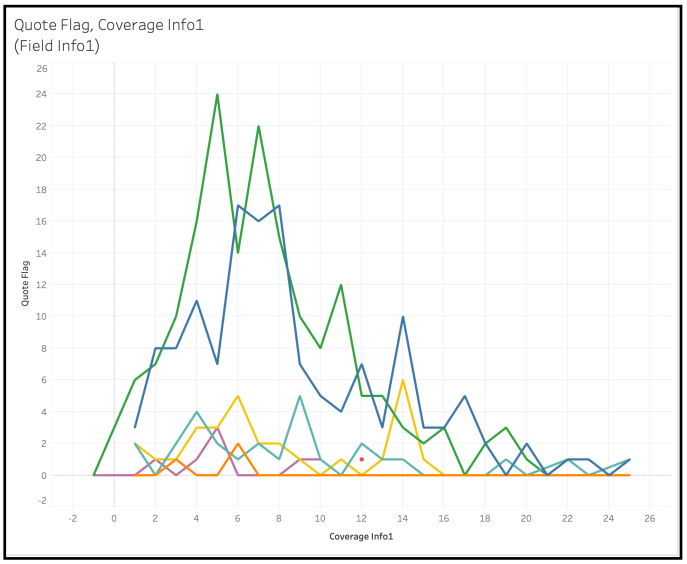
This line graph shows the relationship between Quote_Flag and Day of the month. During the **first week** Quote_Flag attains the **peak**, that gradually drops . **Last week** sales are also **good.**



This line chart corresponds to Property_Info4 and Property_Info5. We can see that for **Property_Info4=1** (irrespective of Property_Info5) has **higher Quote_Flag** than **Property_Info4=0.**

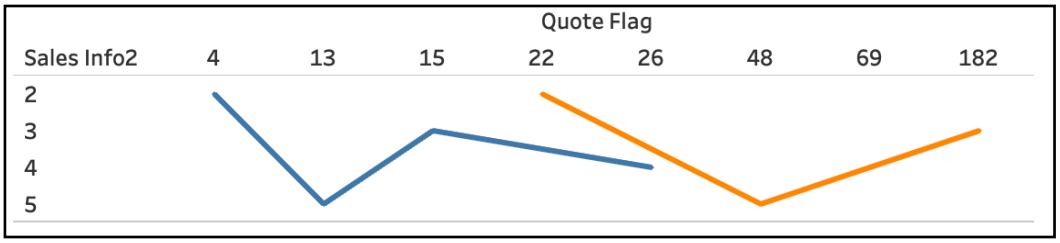
More people buy the policy if **Property_Info4 is 1.**





This line chart corresponds to **Coverage_Info1** and **Field_Info1**. For **Field_Info1 = 'F' & 'B' & Coverage_Info1=5 & 7** attains a **highest Quote_Flag.**

More people buy policy if **Field_Info1 ='F' & Coverage_Info1='5'.**



This line chart above corresponds to **Sales_Info1** & **Sales_Info2.**
We see that for **Sales_Info1 = 1** (irrespective of Sales_Info2) has **higher** Quote_Flag than **Sales_Info1 = 0.**
More people buy the policy if **Sales_Info1 is 1.**

Also **Quote_Flag** is high when **Sales_Info2 = 5.**

1B. Data Preprocessing

Based on the knowledge gained during data exploration, the data set can be cleansed and massaged to eliminate and compensate for bad data and missing data. Some data transformation might be done as well in order to analyse the dataset.

a. Binning

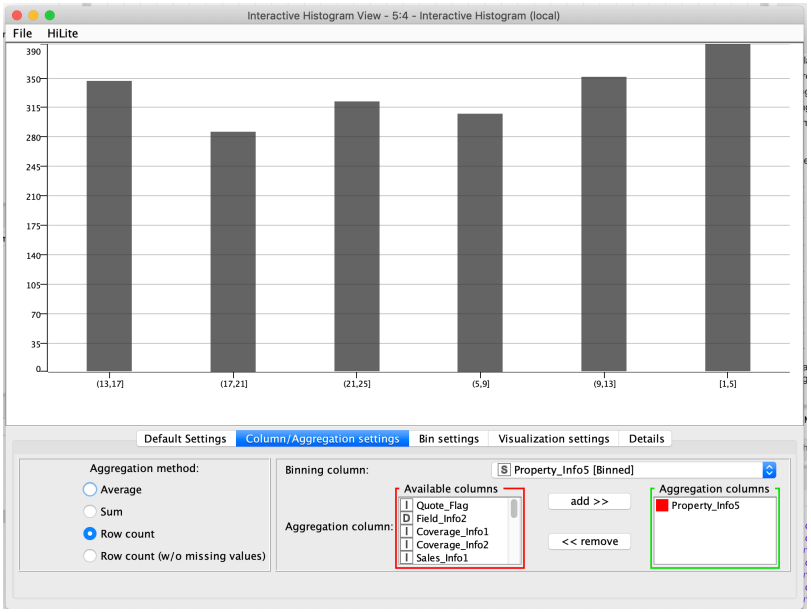
Binning is used to *categorise variables*. Numerical variables may be converted to categorical in order to perform binning. This analysis shows binning of **Property_Info5** column. Binning has two unsupervised method which are **Equi-width** and **Equi-depth**.

Equi-width: (6 bins)

Equi-width binning is dividing data into equal intervals.
W = (max – min)/k
Where 'W' is the width of interval.
The interval boundaries are (*min+W, min+2W.....,min+(k-1)*)

For Property_Info5, max value = 25 and min value =1.

W = (25-1)/k = 24/k
'k' should give the equal size of interval. So, k can be either **4, 6, 8, 12.**
I plotted histograms for each of these values.
Choosing **6 bins** is appropriate as this provides more granularity and shows the distribution clearly.



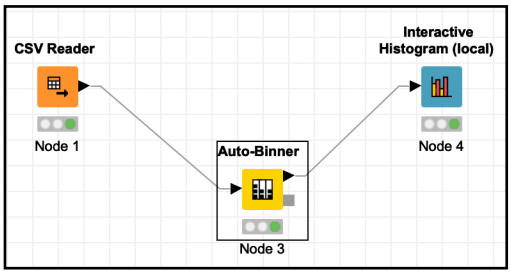
This histogram represents **6 bins.**

Bin 1 = [1,5]
Bin 2 = (5,9]
Bin 3 = (9,13]
Bin 4 = (13,17]
Bin 5 = (17,21]
Bin 6 = (21,25]

*Check the *excel sheet* with name **1B_EquiWidth** for results.

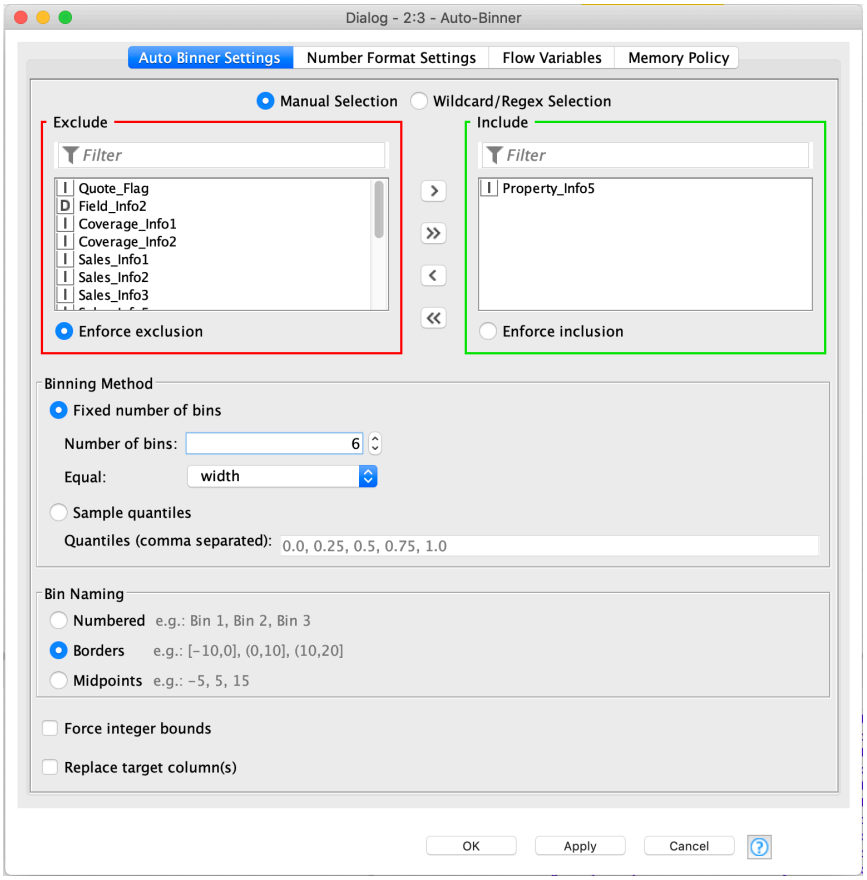
The steps to perform **Equi-width** in KNIME are:

1) Open KNIME Workflow.
2) Add **CSV Reader** node, configure it to the input location of the excel file.
3) Add **Auto-Binner** node.



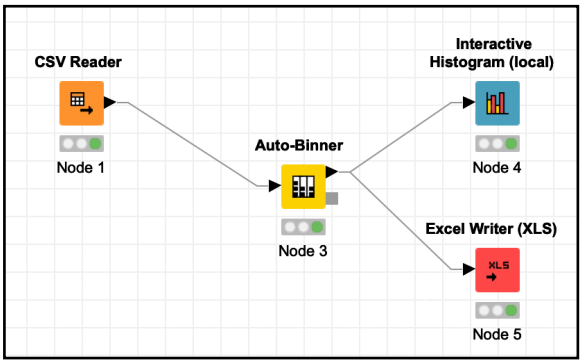
4) Configure **Auto-Binner**, Include only **Property_Info5 .**
5) In Binning select the N*umber of bins* as **6** and choose **width.**

6) Select **Borders** in *Bin Naming*
7) Click on *Apply* then *OK*



8) Add **Interactive-Histogram** to check whether is bins are smoothed or not.
9) Connect *Auto-Binner* and I*nteractive Histogram.*
10) If the numbers of bins are perfect, then write the binned data in the excel sheet.

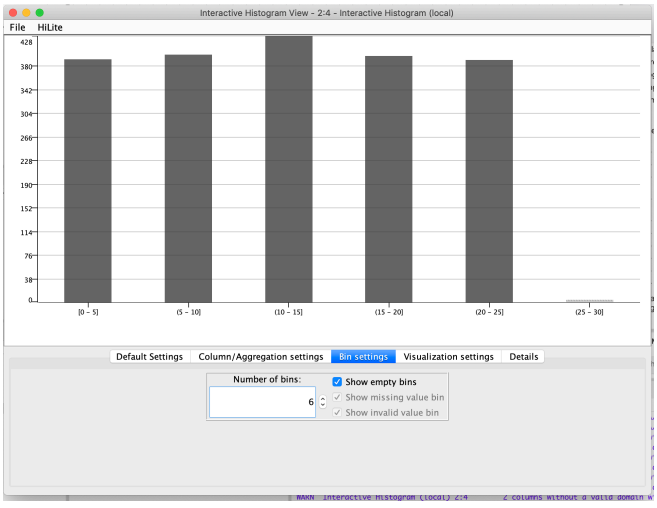
11) Add **Excel Writer** Node.
12) Configure it to output location, add the columns you want to write.



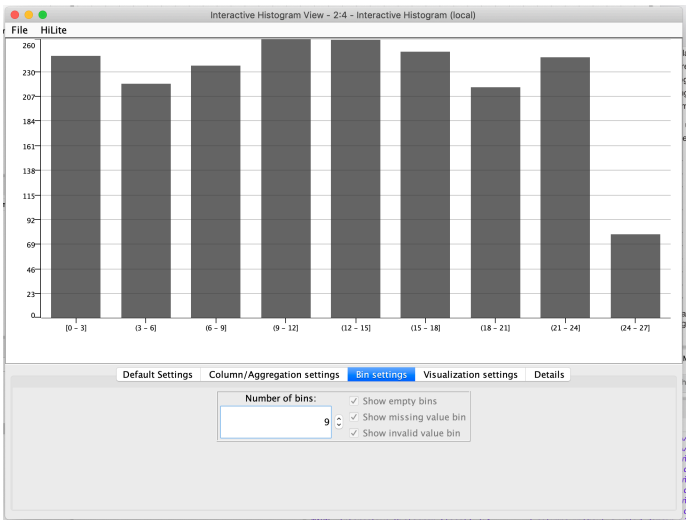
Equi-depth: (5 bins)

Equi-depth is achieved by distributing almost **equal number of elements in each bin.**

I chosed numbers from 4 to 10 and plotted a histogram to check which has almost equal number of elements in each bin.



This histogram represents **6 bins** formed by equi-dept. As clearly seen **Bin 6 - (25,30]** is empty which results in uneven distribution of data.
Hence, **not** the ideal bin size.



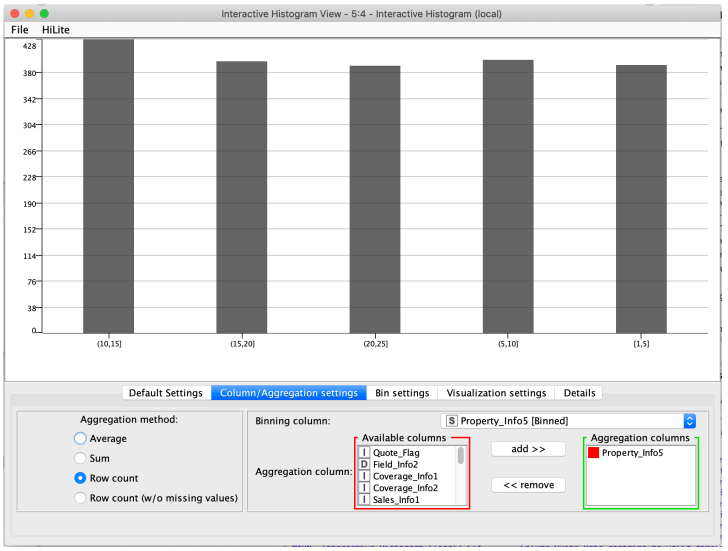
This histogram represents **9 bins** formed by equi-dept. As clearly seen **Bin 9 - (24,27]** is has very less number of elements compared to other Bins.
Hence, **not** the ideal bin size.

Choosing **5 bins** is appropriate as this provides more granularity and shows the distribution clearly.

Bin 1 = [1,5]
Bin 2 = (5,10]
Bin 3 = (10,15]
Bin 4 = (15,20]
Bin 5 = (20,25]

*Check the *excel sheet* with name **1B_EquiDept** for results.

The histogram represents **5 bins** formed by equi-dept.

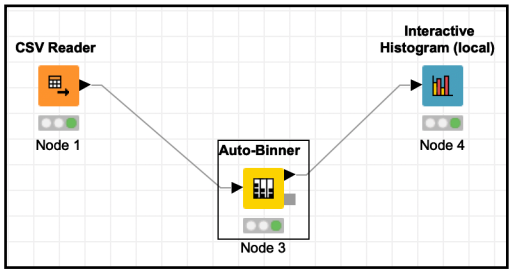


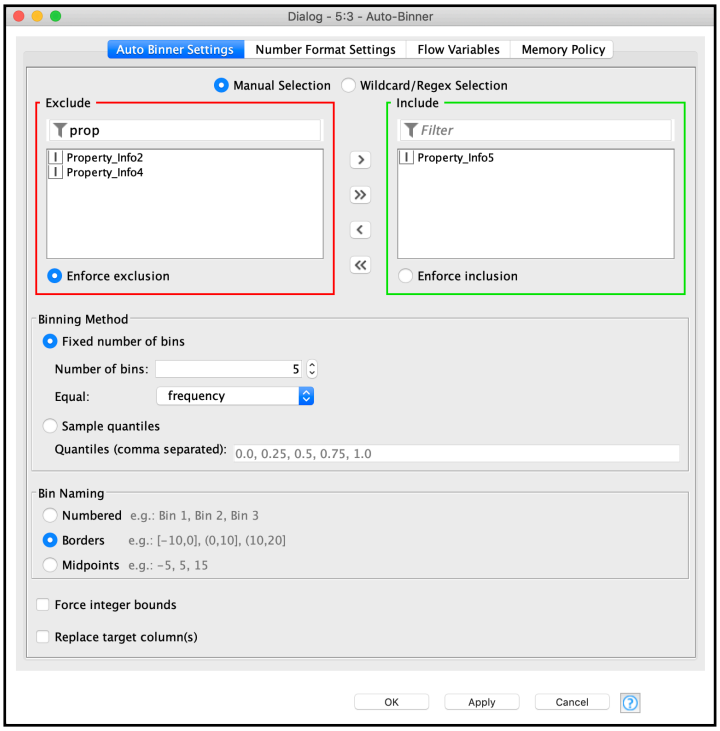
The steps to perform Equi-dept in KNIME are:

1) Open KNIME Workflow.
2) Add **CSV Reader** node, configure it to the input location of the excel file.
3) Add **Auto-Binner** node.

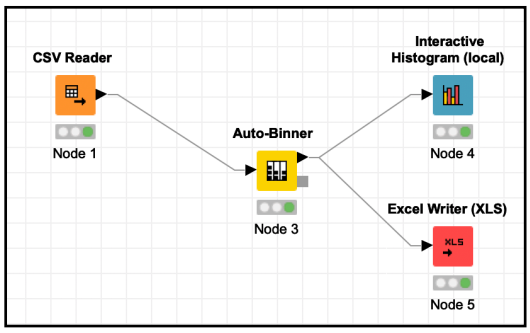
4) Configure **Auto-Binner**, Include only **Property_Info5 .**
5) In Binning select the N*umber of bins* as **5** and choose **frequency.**

6) Select **Borders** in *Bin Naming*
7) Click on *Apply* then *OK*





8) Add **Interactive-Histogram** to check whether is bins are smoothed or not.
9) Connect *Auto-Binner* and I*nteractive Histogram.*
10) If the numbers of bins are perfect, then write the binned data in the excel sheet.
11) Add **Excel Writer** Node.
12) Configure it to output location, add the columns you want to write



b. Normalize

Min-Max Normalization

Min-Max Normalization is used to normalize the data and scale the data between 0 and 1.
From the given dataset, **Sales_Info5** is normalized to transform the data between min = 0.0 and max = 1.0 .

The formula for **min-max normalized:**

$$v' = \frac{v - min}{max - min}(h - l) + l,$$

Sales_Info5 range from 82 to 67153. Thus, **h = 67153 and l = 82.**
v is each value in Sales_Info5 and **v'** is the normalized value.

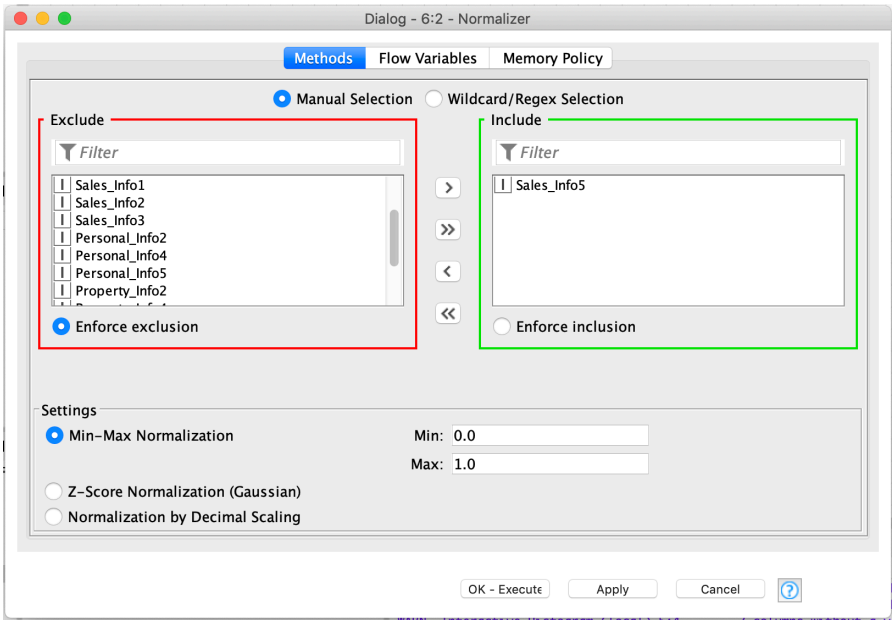
To calculate Min-Max Normalization using KNIME:
1) Open KNIME Workflow.
2) Add **CSV Reader** node, configure it to the input location of the excel file.
3) Add **Normalizer** node.

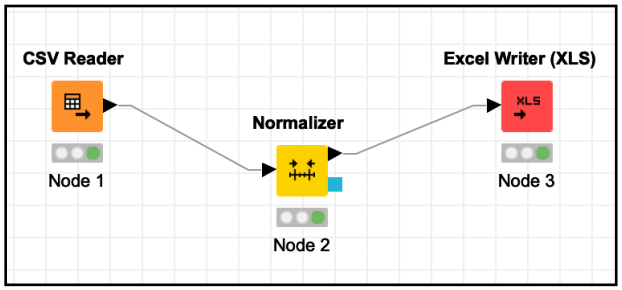
4) Configure **Normalizer**, Include only **Sales_Info5 .**

5) In Setting section, choose *Min-Max Normaliztion* and set **Min = 0.0** and **Max = 1.0**

6) Click Apply then Ok

7) Add **Excel Writer** Node.
8) Configure it to output location, add the columns you want to write.





*Check the *Excel Sheet* for Normalized values for **1B_MinMax**

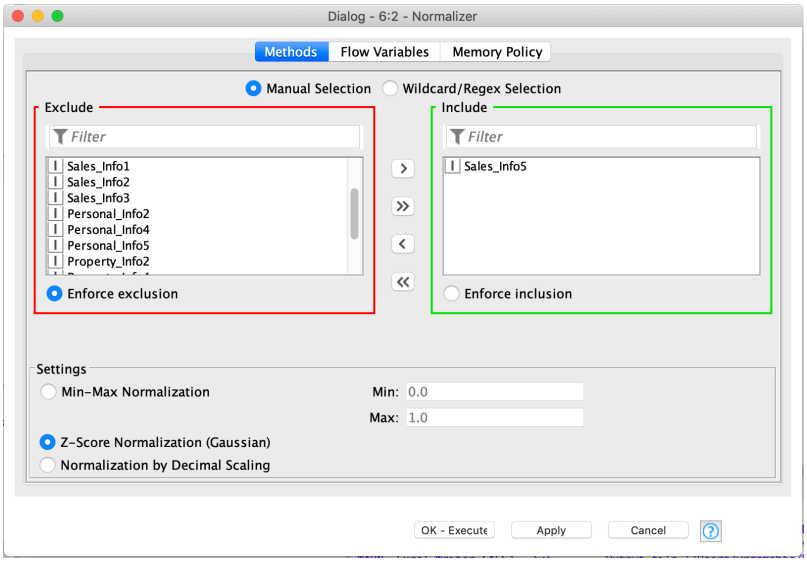
Z Score Normalization

The standard **score** (more commonly referred to as a **z-score**) is a very useful statistic because it (a) allows us to calculate the probability of a **score** occurring within our normal distribution and (b) enables us to compare two **scores** that are from different normal distributions.

Row ID	Min	Max	Mean	Std. deviation	Skewn...	Variance	Kurtosis	Overal...	No. m...	
Sales_Info5	82	67,153	34,016.55	19,066.928	0.002	363,547,739.062	-1.184	68,033,100	0	0

The equation for z-score normalization is:

Z score = (x – Mean)/standard_deviation
Mean = 34016 & Standard Deviation = 19067
x are the values in **Sales_Info5**



To calculate Z Score Normalization using KNIME:

Same method as *Min-Max Normalization,*
Except In **Normalizer Node** change Setting from *Min-Max Normalization* to **Z-Score Normalization (Guassian).**

*Check the Excel Sheet for Normalized values for **1B_ZScore.**

c. Discretise

The task is to discretise the Column **Coverage_Info1** into four categories: **Basic, Low, Medium and High.**

Row ID	Column	Min	Max	Mean	Std. deviation	Skewn...	Variance	Kurtosis	Overal
Coverage_Info1	Coverage_Info1	-1	25	8.956	5.662	1.045	32.058	0.644	17,912

Since we don't what **Coverage_Info1** refers to. I have used *Gaussian Distribution* for dividing into categories.

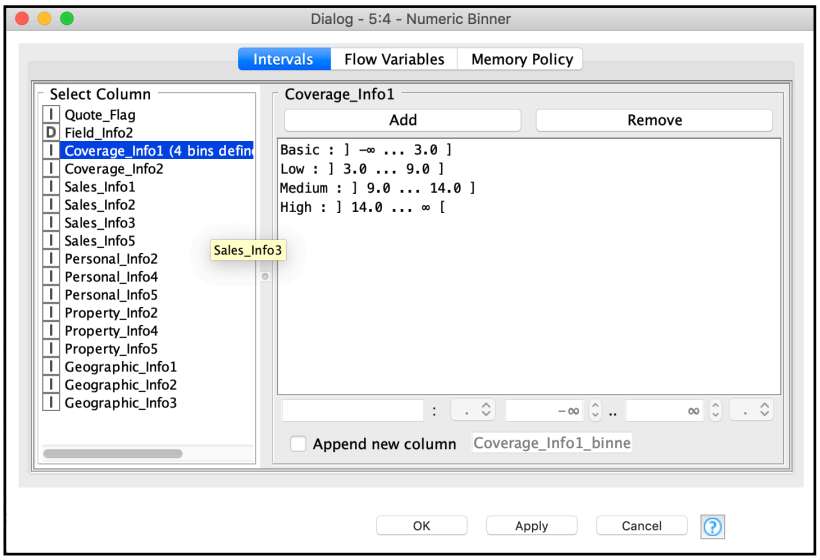
Mean (M) = 8.956
Standard Deviation (SD)= 5.662

Basic : M - SD = 8.956 - 5.662 = 3.294
Range : **[Min, M-SD] = [-1, 3.294] —-> [-1, 3]**

Low : M - SD = 8.956 - 5.662 = 3.294
Range : **[M-SD, M] = [3.294, 8.956] —> (3, 9]**

Medium : M + SD = 8.956 + 5.662 = 14.618
Range : **[M, M+SD] = [8.956, 14.618] —-> (9, 14]**

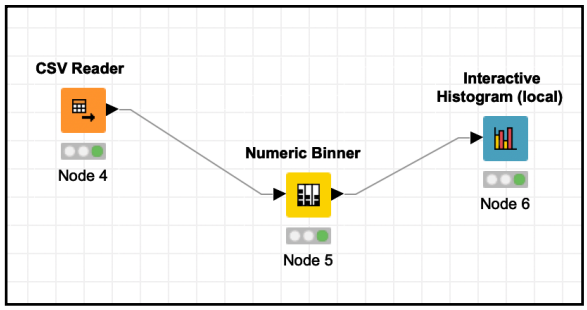
High : M + SD = 8.956 + 5.662 = 14.618
Range : **[M+SD, Max] = [14.618, 25] —> (14, 25]**



To calculate **Discretisation** using KNIME:

1)Open KNIME workflow
2)Add **CSV Reader** node, configure it to the input location of the file.
3)Add **Numeric Binner** Node, Configure it using the values calculated above for **Basic, Low, Medium, High.**

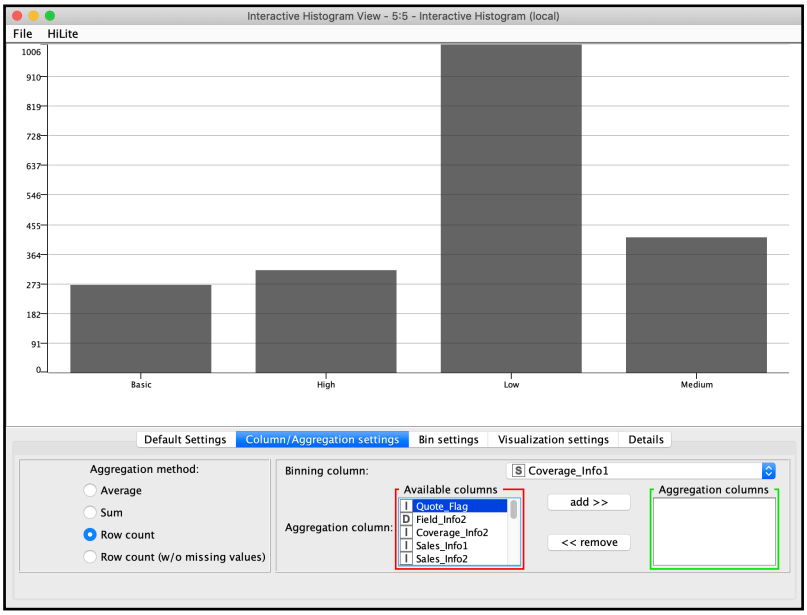
4)Add **Excel Writer** node
5)Configure it to output location, add columns you want to write.



	A	B
1	Coverage_Info1	Coverage_Info1
2	12	Medium
3	5	Low
4	9	Low
5	1	Basic
6	11	Medium
7	4	Low
8	13	Medium
9	18	High
10	1	Basic
11	2	Basic
12	18	High

Histogram of the Discretised values.

*Check the Excel Sheet for Discretised values for **1B_Discrete**



d. Binarisation

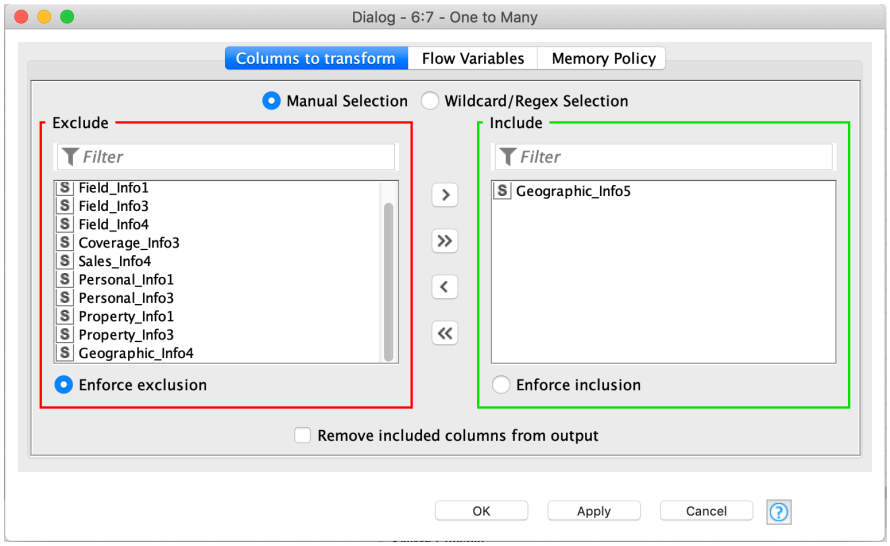
The binarisation is the method of converting nominal to binary values(0's and 1's). from the given dataset, the binarisation operation is performing on Geographic_Info5 to binarise, **'CA', 'IL', 'NJ' & 'TX' .**

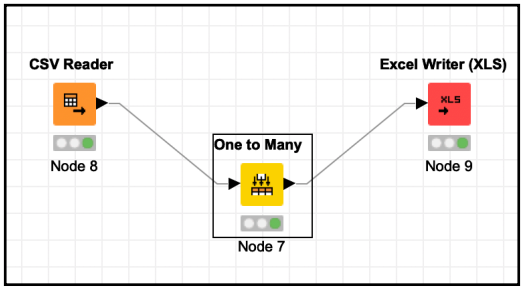
To compute Binarisation Normalization using KNIME:

1) Open KNIME Workflow.
2) Add **CSV Reader** node, configure it to the input location of the excel file.
3) Add **One to Many** node.

4) Configure **One to Many**, Include only **Geographic_Info5 .**

5) Click Apply then Ok.
7) Add **Excel Writer** Node.
8) Configure it to output location, add the columns you want to write.





*Check the *Excel Sheet* for Binarised values **1B_Binarise.**

Sample output is shown.

	A	B	C	D	E
1	Geographic_Info5	CA	NJ	IL	TX
2	CA	1	0	0	0
3	NJ	0	1	0	0
4	NJ	0	1	0	0
5	IL	0	0	1	0
6	NJ	0	1	0	0
7	NJ	0	1	0	0
8	TX	0	0	0	1
9	CA	1	0	0	0
10	IL	0	0	1	0
11	NJ	0	1	0	0
12	IL	0	0	1	0
13	CA	1	0	0	0
14	CA	1	0	0	0
15	TX	0	0	0	1
16	CA	1	0	0	0
17	TX	0	0	0	1

'1' represent true if the value is there, '0' represents false.

Number of binarised column = Number of distinct values .

1C. Summary

The main attributes of interest in the data set Quote_Flag(indicates whether a person bought the policy), Quote_Date(for trends), Geographic_Info5 (name of states US), atleast one attribute from each category.

As the head of Analytics Unit, the main goal is to increase the sale of insurance policy. To achieve that, check what all attributes are related to having Quote_Flag= 1 or having high sales.

1) **Geographic_Info5 = 'NJ'** has the highest sales for Field_Info1 = 'F'. The company can target other states with **Field_Info1='F'**, a key factor in increasing sales for Geographic_Info5 = 'NJ'.

2) Similarly, Geographic_Info5 = 'NJ' has the highest sales for Field_Info3 = '564'. The company can target other states with **Field_Info1='564**', a key factor in increasing sales for Geographic_Info5 = 'NJ'.

3) **Field_Info1='N'** has significantly high sales(high Quote_Flag) comparing to Field_Info1='Y'. The insurance company might remove Field_Info1 characteristics from the policy, if Field_Info1 'Y' & 'N' means having a particular characteristic in the policy.

4) **Field_Info4 ='B' & 'F'** has much higher Quote_Flag compared to other in Field_Info4. The company may include Field_Info4 = 'B' or 'F' in the characteristic of their policy.

5) **Quote_Date** also has a relation with Quote_Flag. For **April and May** the sales of policy are highest while July records for the lowest sales. Day also affects the sale. For a particular month, the highest sales occur in first week and last week of the month.

6) **Property_Info4** having binary values, '0' & '1'. The company has significantly high sales (high Quote_Flag) for **Property_Info4=1** then Property_Info4=0. The company must include Property_Info4=1 characteristic in their policy.

7) **Coverage_Info1 = 5** corresponds to highest sales registered for Field_Info1='F'. The company should target other coverage areas with Field_Info1 = 'F' for higher sales of the policy.

8) **Sales_Info1** having binary values, '0' & '1'. The company has significantly high sales (high Quote_Flag) for Sales_Info1=1 then Sales_Info1=0. The company must include Sales_Info1=1 characteristic in their policy.

9) **Sales_Info2 = 5** or higher the value of Sales_Info2 better is the sale of the policy. Sales_Info2=5 attains the highest sales.