Name: Vishesh Kalsi

SID: 13731652

1A. Initial data exploration

A)

S No.	Attribute	Type	Reason
1	BATHRM	Interval	We can make meaningful comparisons between different values of number of bathrooms and perform arithmetic functions such as '+' and '-' which give us results that make sense
2	HEAT_D	Nominal	It is only making sense when we distinguish between different types of heating description such as warm cool and evp cool but nothing more
3	AC	Nominal	Air conditioning is a binary attribute which makes it a nominal attribute
4	NUM_UNITS	Interval	We can make meaningful comparisons between different values of number of units and perform arithmetic functions such as '+' and '-' which give us results that make sense
5	ROOMS	Interval	We can make meaningful comparisons between different values of number of rooms and perform arithmetic functions such as '+' and '-' which give us results that make sense
6	BEDRM	Interval	We can make meaningful comparisons between different values of number of bedrooms and perform arithmetic functions such as '+' and '-' which give us results that make sense
7	AYB	Interval	We can make meaningful comparisons between different values of the earliest time the main portion of the building was built and perform arithmetic functions such as '+' and '-' which give us results that make sense
8	YR_RMDL	Interval	We can make meaningful comparisons between different values of last year when residence was remodelled and perform arithmetic functions such as '+' and '-' which give us results that make sense
9	SALEDATE	Interval	We can make meaningful comparisons between different values of date of most recent sale and perform arithmetic functions such as '+' and '-' which give us results that make sense
10	PRICE	Ratio	All possible mathematical operations when applied turn out to be meaningful and provide us with data worthy of capturing
11	QUALIFIED	Nominal	Air conditioning is a binary attribute which makes it a nominal attribute

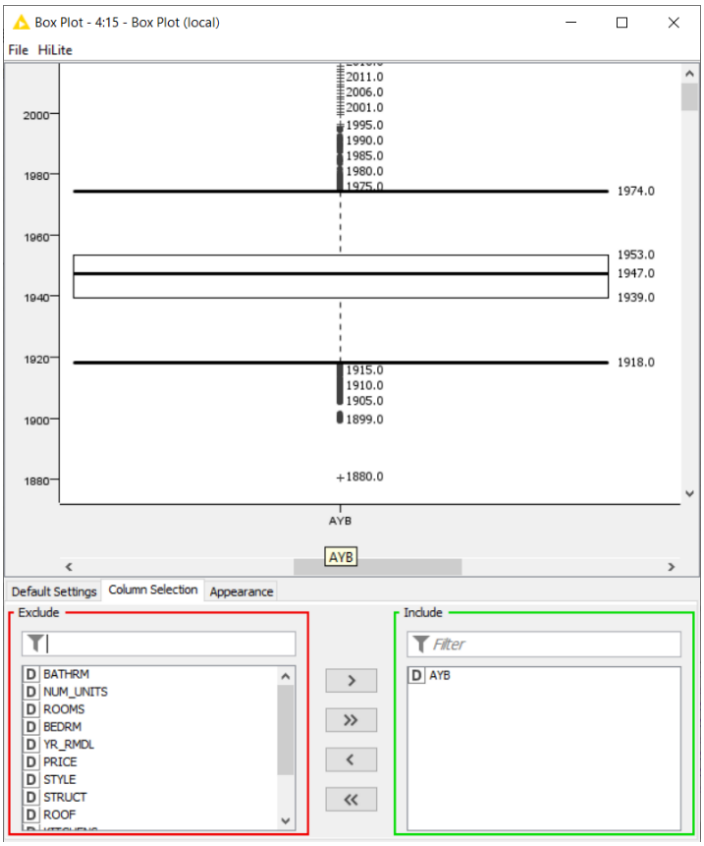
12	STYLE	Nominal	It is only making sense when we distinguish between different types of style codes but nothing more
13	STRUCT	Nominal	It is only making sense when we distinguish between different types of structure codes but nothing more
14	GRADE_D	Nominal	It is only making sense when we distinguish between different types of grade descriptions but nothing more
15	CNDTN_D	Nominal	It is only making sense when we distinguish between different types of condition descriptions but nothing more
16	EXTWALL_D	Nominal	It is only making sense when we distinguish between different types of exterior wall descriptions but nothing more
17	ROOF	Nominal	It is only making sense when we distinguish between different types of codes but nothing more
18	INTWALL_D	Nominal	It is only making sense when we distinguish between different types of interior descriptions but nothing more
19	KITCHENS	Ratio	All possible mathematical operations when applied turn out to be meaningful and provide us with data worthy of capturing
20	FIREPLACES	Ratio	All possible mathematical operations when applied turn out to be meaningful and provide us with data worthy of capturing
21	LANDAREA	Ratio	All possible mathematical operations when applied turn out to be meaningful and provide us with data worthy of capturing
22	GIS_LAST_MOD_DTTM	Interval	We can make meaningful comparisons between different values of last modified datetime and perform arithmetic functions such as '+' and '-' which give us results that make sense

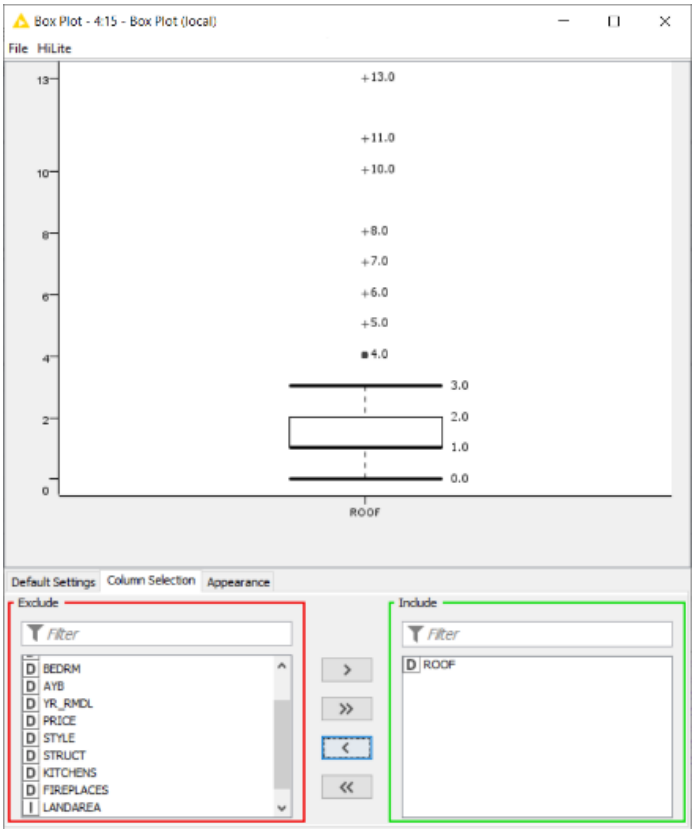
B)

S No.	Attribute name	Range	Mean	Median	Mode	Variance	Standard Deviation	Min	Max
1	BATHRM	8	1.635	1	1	0.759	0.871206061	0	8
2	HEAT_D				Hot Water Rad				
3	AC				Y				
4	NUM_UNITS	4	1.192	1	1	0.491	0.700713922	0	4
5	ROOMS	24	6.862	6	6	5.895	2.427962108	0	24
6	BEDRM	12	3.022	3	3	0.979	0.989444288	0	12
7	AYB	2018	1950.201	1947	1950			0	2018
8	YR_RMDL	101	1999.438	2006	2017			1917	2018
9	SALEDATE				1900-01-01T00:00:00.000Z				
10	PRICE	25100000	161624.748	130000				0	25100000
11	QUALIFIED				N				
12	STYLE	15	4.009	4	4			0	15
13	STRUCT	8	4.857	6	1			0	8
14	GRADE_D				Average				
15	CNDTN_D				Average				
16	EXTWALL_D				Common Brick				
17	ROOF	13	2.063	1	1			0	13
18	INTWALL_D				Hardwood				
19	KITCHENS	5	1.2	1	1			0	5
20	FIREPLACES	4	0.241	0	0			0	4
21	LANDAREA	62349	3497.832	2847.5	4000	0.267	0.516720427	760	63109
22	GIS_LAST_MOD_DTTM				2018-07-22T18:01:43.000Z				

In the following snapshots we come across various relations for different kinds of attributes of the dataset.

We observe that a box plot of AYB tells us various things about the variable. It tells us about the different quantiles, outliers, median, interquartile range etc. It is a very useful

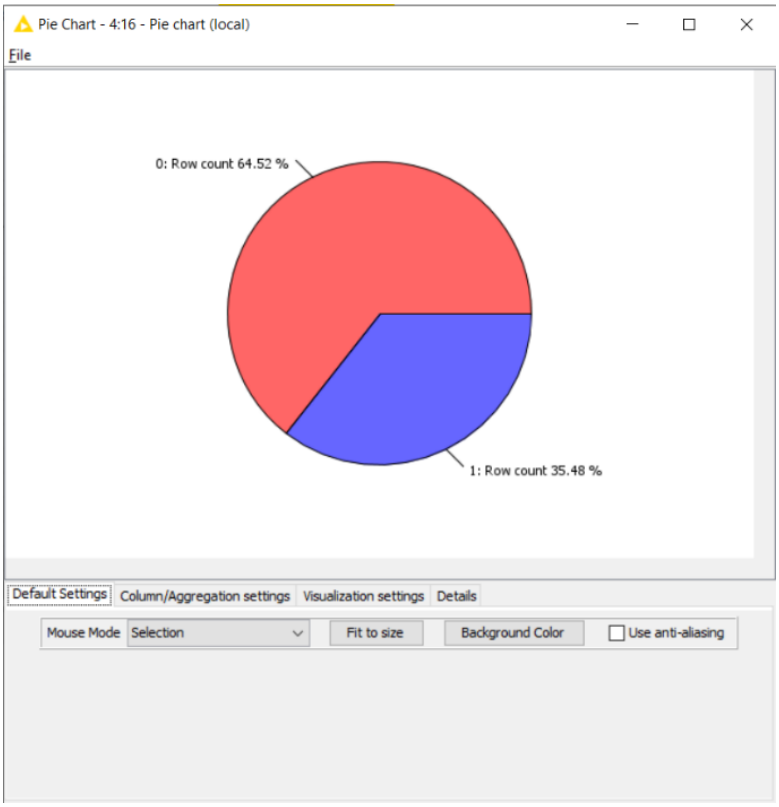


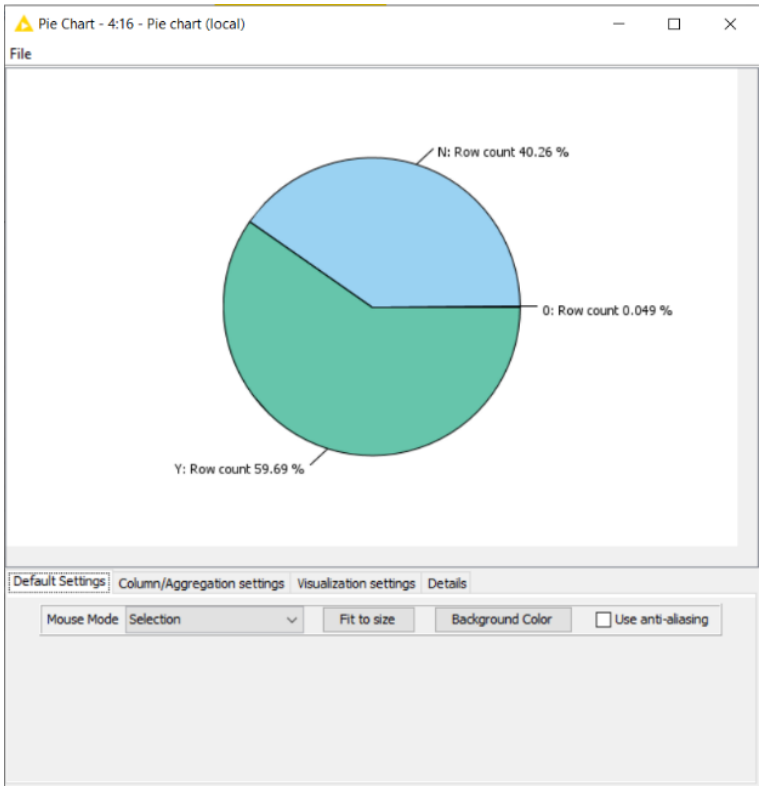


When we make a box plot of another variable available to us such as roof then we get to know a lot about this too. We get to know that it has a median of 1 and we gain a handful of knowledge on its other aspects which include the quantiles, outliers etc.

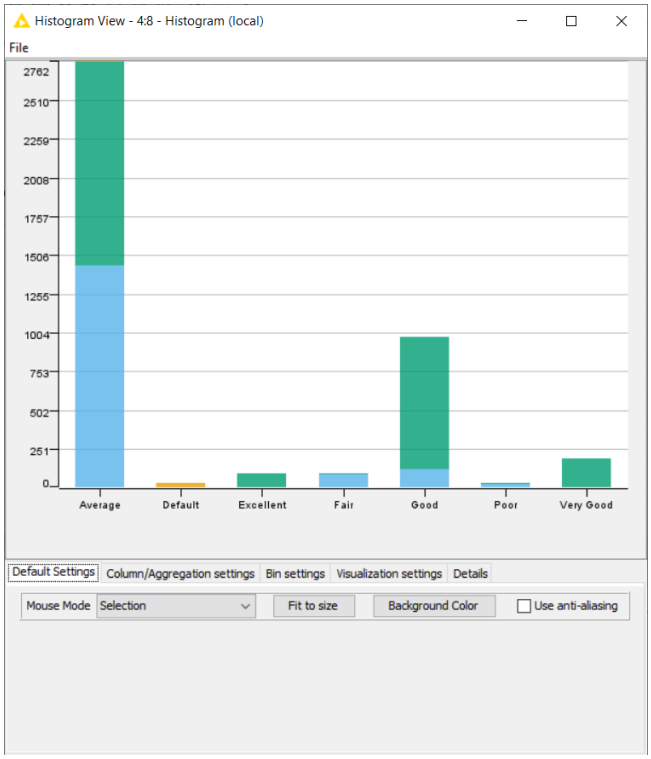
When we make a histogram for the variable named ‘BEDRM’, we get to know various things about it. We infer a knowledge about how the data in the variable in spread out and how is it made up.

If we observe a pie chart of a binary variable such as Qualified then we get to know about the distribution of the data that resides in it. We get to know that majority of the data in the variable is made up of 0 that is most of the houses are unqualified. This proves to be very useful for us.



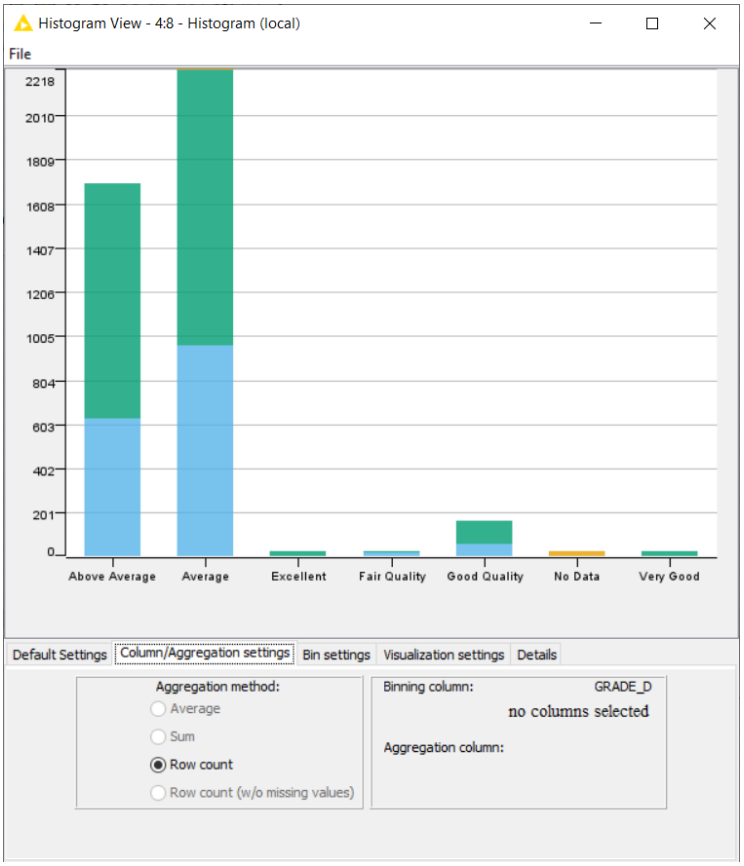


We also have another binary variable in our data set that is AC. When we make a pie chart for thi variable, it also turns out to be pretty useful for us as it gives us a detailed insight into the variable. We get to know that a major portion of the variable consists of the values Yes which in turn tell us that a major number of houses do have AC fitted.



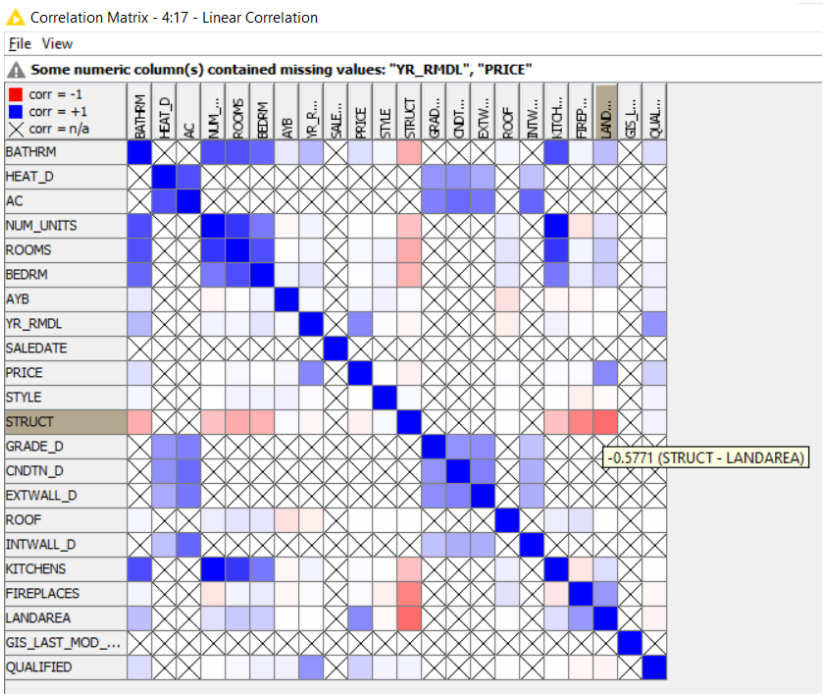
When we make a histogram of a variable such as CNDTN_D then we get to know a lot about it too. We get to know about the spread of the data among the variable. It tells us that majority of the values in the variable are of Average followed by Good and so on. This proves out to be very useful for us.

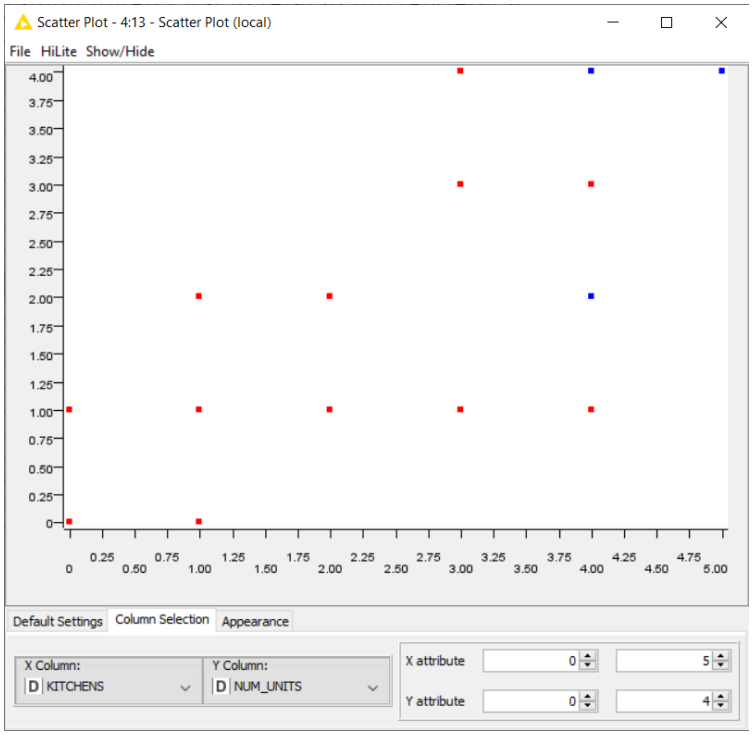
When we make a histogram of a variable such as GRADE_D then we get to know a lot about it too. We get to know about the spread of the data among the variable. It tells us that majority of the values in the variable are of Average followed by Above Average and so on. This proves out to be very useful for us.



C)

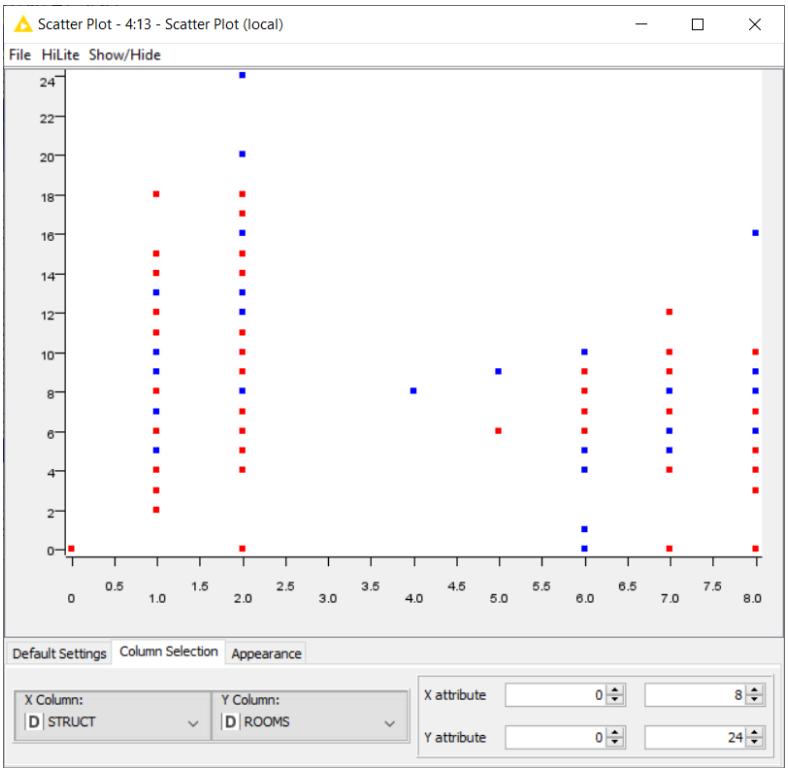
The most useful node to help us find relations between all present variables is a linear correlation node. It helps us to identify the relations in a very interactive and easy method. For ex the linear correlation between Struct and Landare is -0.5771

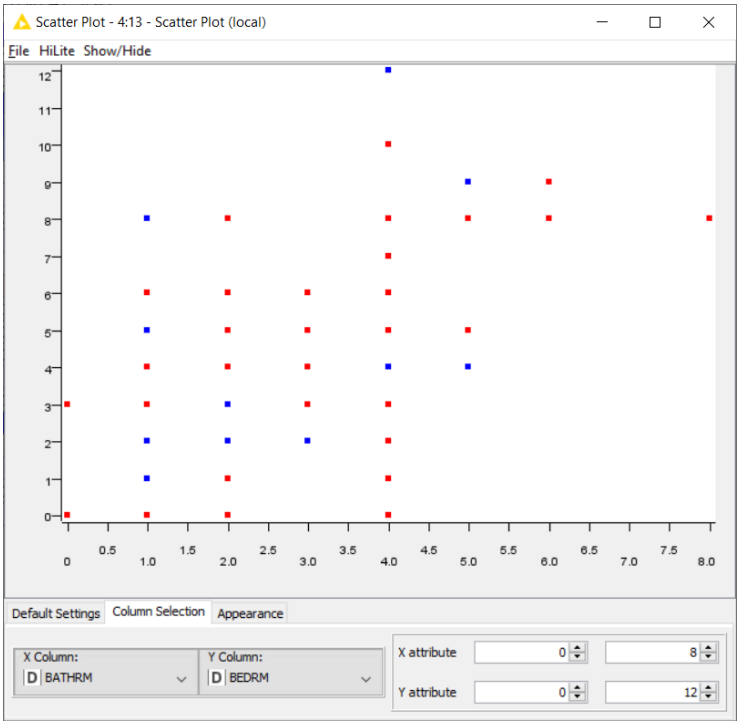




In this we have found a weak positive relation between the variables Kitchens and Num_units by using a scatter plot. It proves out to be very useful for us as it provides us with great insights of the variables and relations between them.

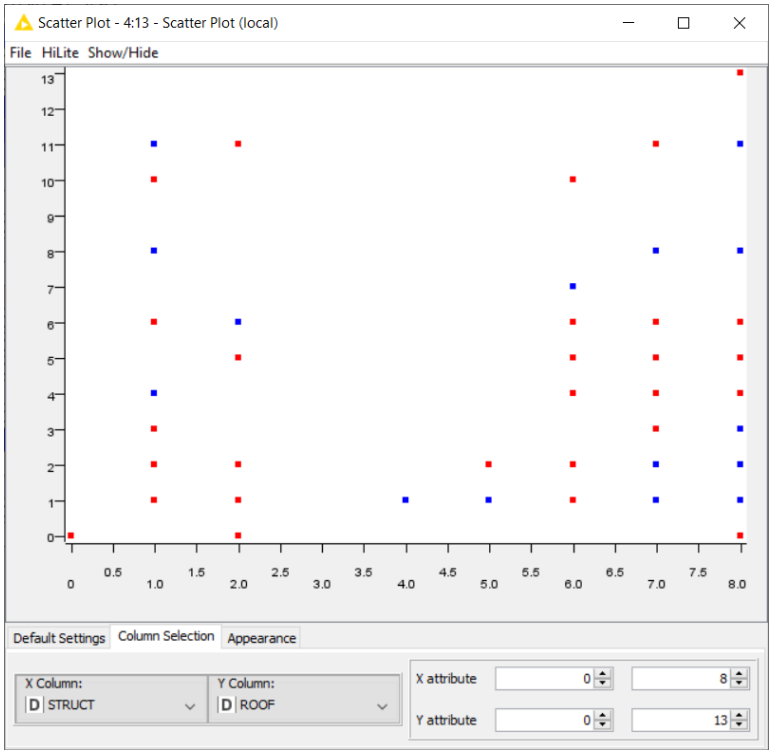
In this we have found a strong negative relation between the variables Struct and Rooms by using a scatter plot. It proves out to be very useful for us as it provides us with great insights of the variables and relations between them

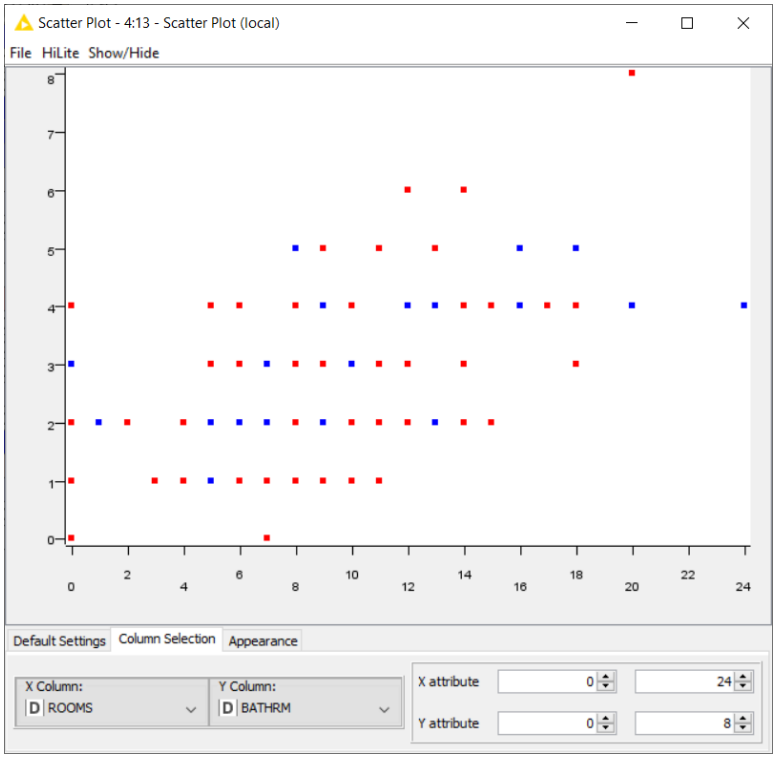




In this we have found a strong positive relation between the variables Bathrm and Bedrm by using a scatter plot. It proves out to be very useful for us as it provides us with great insights of the variables and relations between them

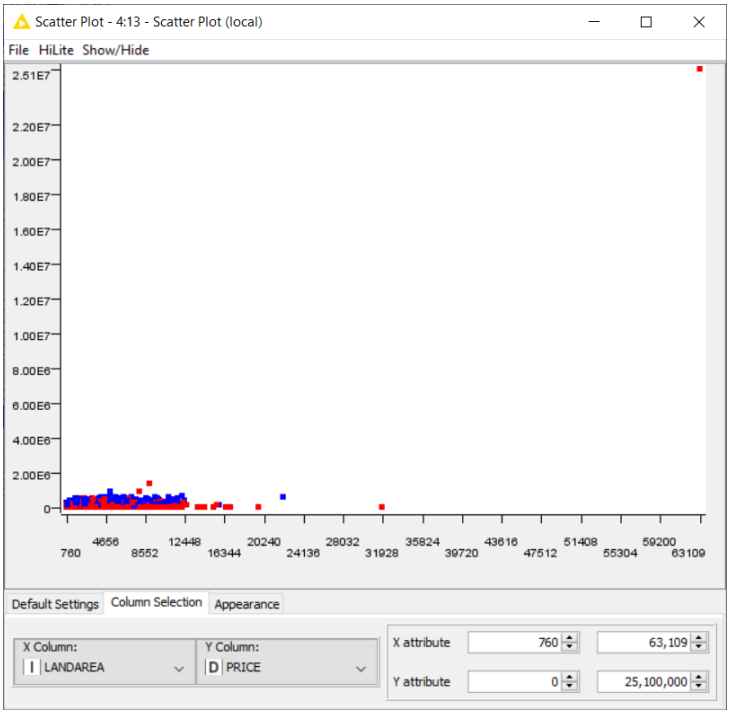
In this we have found a weak positive relation between the variables Struct and Roof by using a scatter plot. It proves out to be very useful for us as it provides us with great insights of the variables and relations between them





In this we have found a weak positive relation between the variables Rooms and Bathrm by using a scatter plot. It proves out to be very useful for us as it provides us with great insights of the variables and relations between them

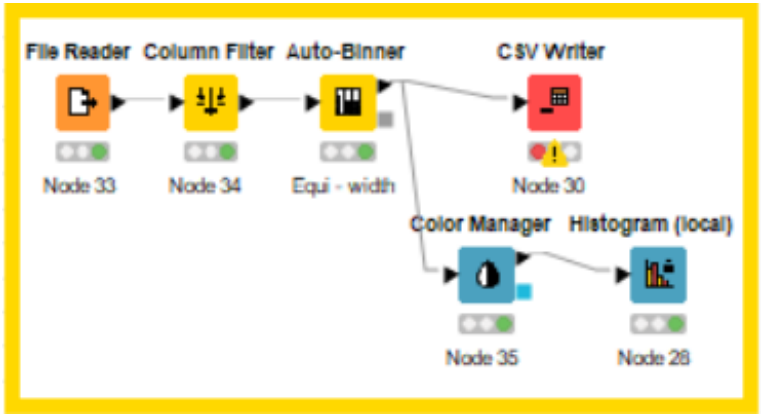
In this we have found a weak positive relation between the variables Landarea and Price by using a scatter plot. It proves out to be very useful for us as it provides us with great insights of the variables and relations between them



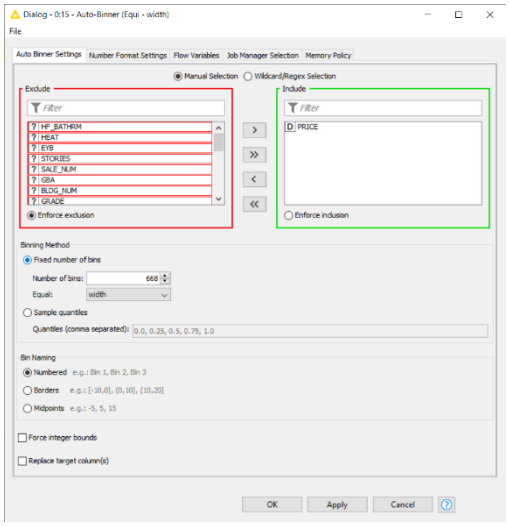
1B. Data preprocessing

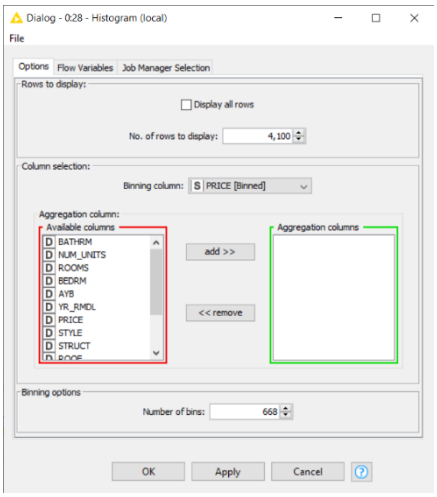
A). In the first part of the pre processing of the data set we had to do binning to smooth the values of the PRICE attribute by 2 methods:

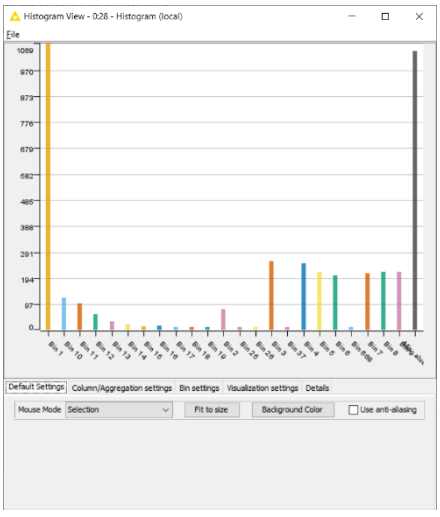
Equi-width binning



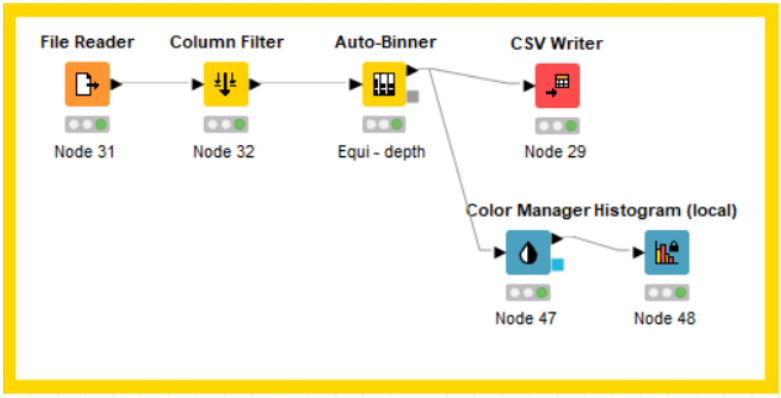
I read the data into a csv reader and then applied a column filter node to keep just the required columns. Then I chose the auto binner node and configured it by just including the required column, Price. Then I chose the number of bins as 668 using the Sturgess formula and saw the data set and found out that these number of bins did justice to the data set and set the binning method to equal width. Hence, I was satisfied with them. I then executed the node and saw the required data set using a histogram. The data set thus obtained can be found in the spreadsheet named: Equi-Width Binning in the attached workbook to provide a better reference on the results found.



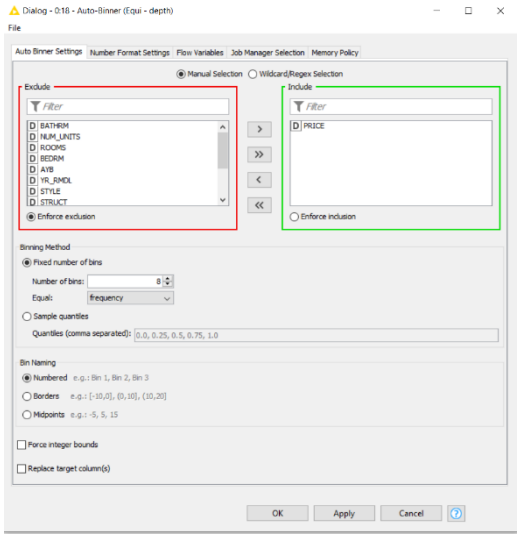


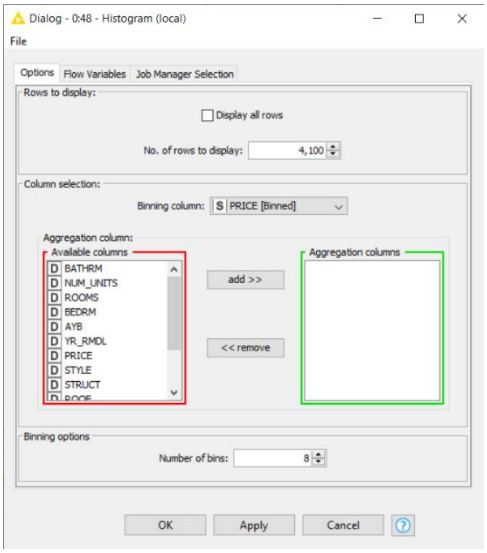


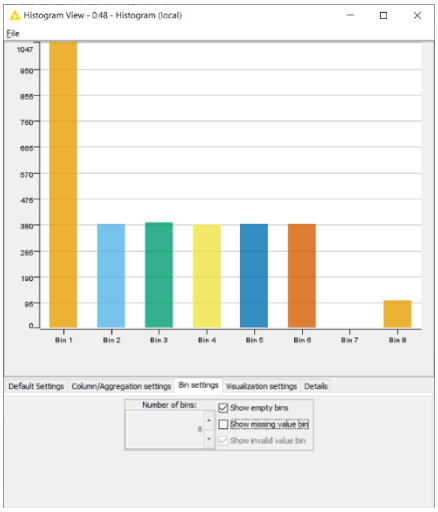
Equi-depth binning



I read the data into a csv reader and then applied a column filter node to keep just the required columns. Then I chose the auto binner node and configured it by just including the required column, Price. Then I chose the number of bins as 8 using saw the data set and found out that these number of bins did justice to the data set and set the binning method to equal frequency. Hence, I was satisfied with them. I then executed the node and saw the required data set using a histogram to which color was added using a color manager. The data set thus obtained can be found in the spreadsheet: Equi-Depth Binning in the attached workbook to provide a better reference on the results found.

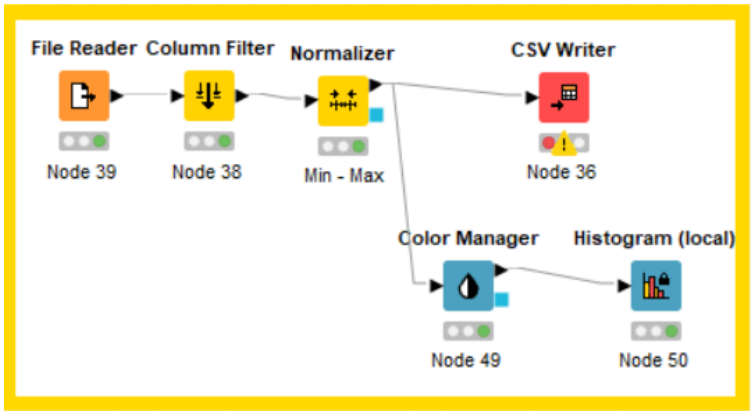




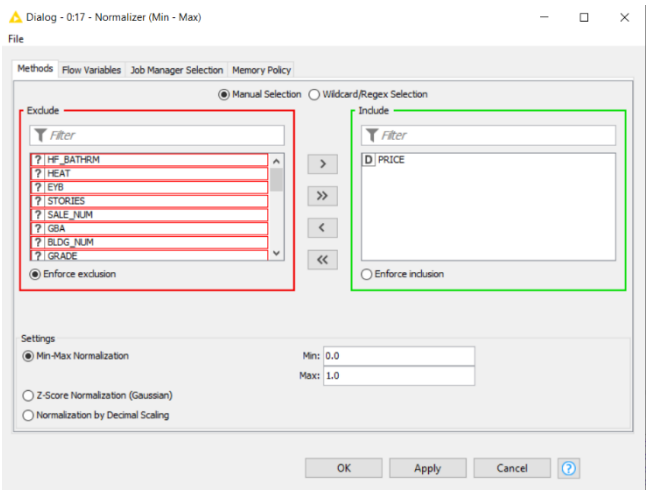


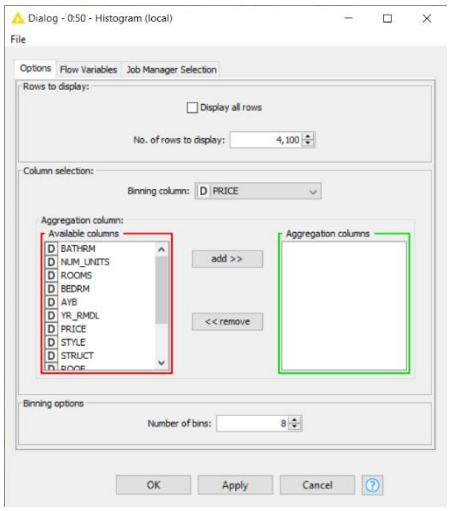
B). In the second part of the pre processing of the data set we had to normalise the attribute PRICE attribute by 2 methods:

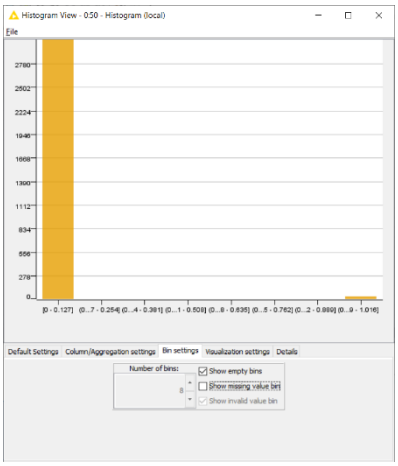
Min-Max normalization to transform the values onto the range [0.0-1.0]



I read the data into a csv reader and then applied a column filter node to keep just the required columns. For this I chose the normalizer node. I then configured it by just including the required column, Price. Then I changed the settings to Min-Max Normalization and set the Min and Max to 0.0 and 1.0, respectively. I then executed the node and saw the column of Price get normalized using the histogram to which color was added using the color manager. The data set can be found in the spreadsheet named: Min-Max Normalization in the attached workbook to provide a better reference on the results found. The data set was written uing the csv writer.

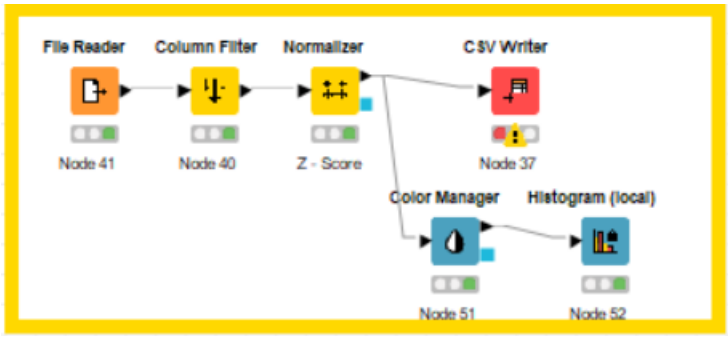


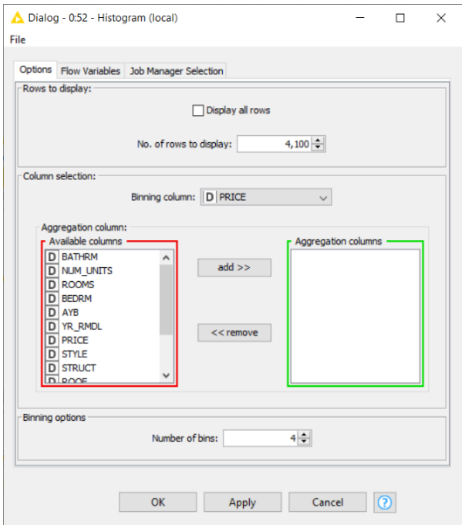


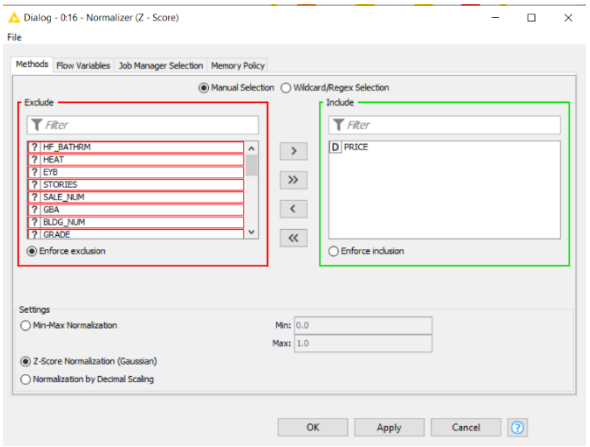


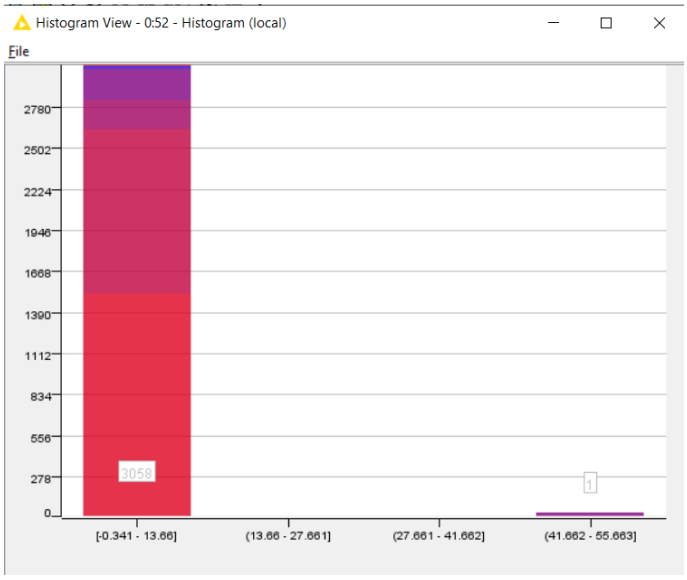
z-score normalization to transform the values

I read the data into a csv reader and then applied a column filter node to keep just the required columns. For this I chose the normalizer node. I then configured it by just including the required column, Price. Then I changed the settings to Z - Score (Guassian). I then executed the node and saw the column of Price get normalized according to the Guassian method which can be found in the spreadsheet named: Z-Score (Guassian) Normalization in the attached workbook to provide a better reference on the results found. The file was written using a csv writer.





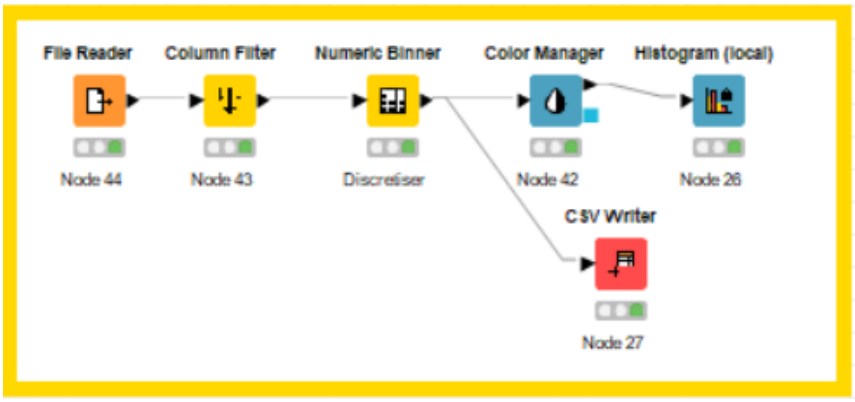


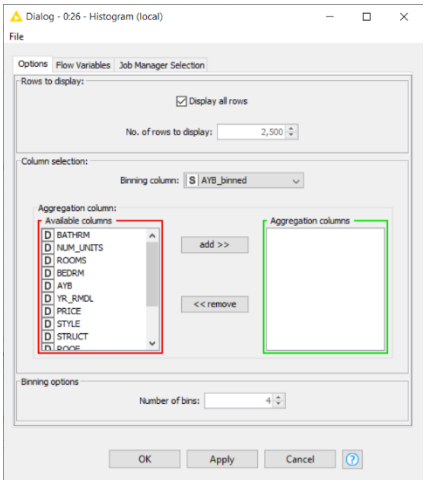


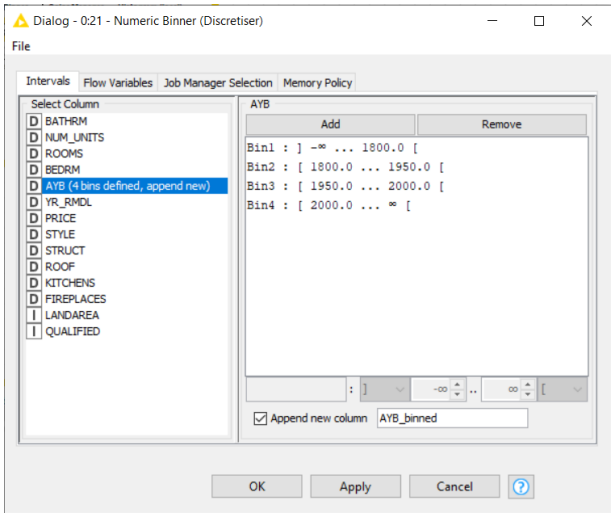
C).

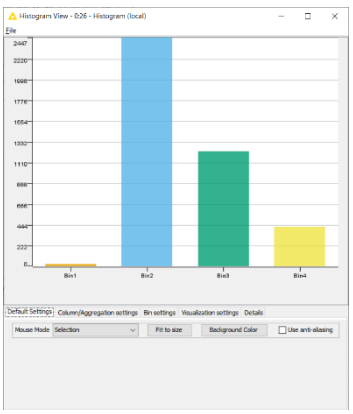
In the third part of the pre processing of the data set we had to discretize the AYB attribute into the following categories: Very Old=0-1800; Old=1801-1950; New=1951-2000; Very New=2001 + and then provide the frequency of each category in the data set

For this I chose the numeric binner and histogram node. I then configured it by adding 4 bins of with the expected details and executed it. After this I configured the histogram as by setting the binning column to AYB_binned and then executing it. The histogram along with the discretised data can be found in the spreadsheet named: Discretise in the attached workbook to provide a better reference on the results found.





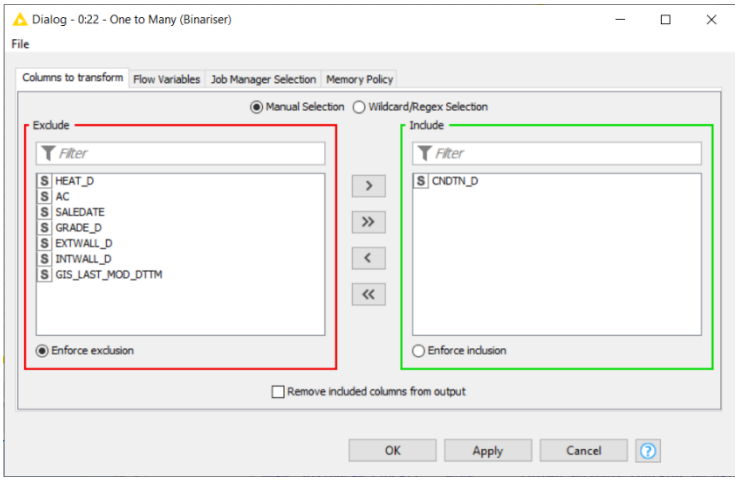


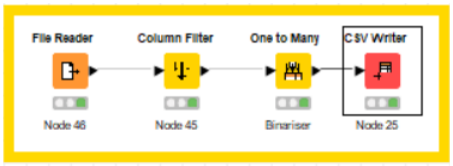


D).

In the fourth part of the pre processing of the data set we had to Binarise CNDTN_D variable [with values "0" or "1"].

For this I chose the one-to-many node. I binarized the required variable by including it only using the filter. It was then executed, and the data can be found in the spreadsheet named: Binarise in the attached workbook to provide a better reference on the results found.





Summary

The relation between the attribute landare and price is positive as landarea increases the price also increases which makes it very intuitive. Another relation of landarea is that when the landarea increases the number of rooms also increase. We can also observe that when the variable struct decreases the variable landarea increases as they have a negative correlation. After completing this assignment it can be reported that the variable Price has a few outliers that skew the whole data and make it difficult to examine. Recent remodeling of the property indicates a better condition of the few residences. It is a wonderful dataset and I would love to work on such a dataset n the future for sure