Assignment Name: Assessment Task 2: Data exploration and preparation
Student Name: Yongyan Liu
Student Number: 14214338

Content

1A. Initial Data Exploration

1. Attribute types

Sl. No.	Attribute Name	Attribute Type	Justification
1	BATHRM	Ratio	The type of BATHRM is 'Ratio' because it has a value that defines the zero point, which is the proportional origin. And it can sort based on the exact difference between the values. Zero means no bathroom.
2	HF_BATHRM	Ratio	The type of HF_BATHRM is 'Ratio' because it has a value that defines the zero point, which is the proportional origin. And it can sort based on the exact difference between the values. Zero means no complete half bathroom.
3	HEAT	Nominal	The type of HEAT is 'Nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups
4	HEAT_D	Ordinal	The type of HEAT_D is 'Ordinal' because can categorize and rank data, but cannot tell for the interval value between rankings. It can be classified according to different values, but it does not mean that one method is more advantageous than another, and it does not mean that all differences are the same.
5	AC	Nominal	The type of AC is 'Nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups
6	NUM_UNITS	Ratio	The type of NUM_UNITS is 'Ratio' because it has a value that defines the zero point, which is the proportional origin. And it can sort based on the exact difference between the values. Zero means no unit.
7	ROOMS	Ratio	The type of ROOMS is 'Ratio' because it has a value that defines the zero point,

			which is the proportional origin. And it can sort based on the exact difference between the values. Zero means no complete room.
8	BEDRM	Ratio	The type of BATHRM is 'Ratio' because it has a value that defines the zero point, which is the proportional origin. And it can sort based on the exact difference between the values. Zero means no bedroom.
9	AYB	Interval	The type of AYB is 'Interval' because variable values can be compared in size, and the difference between the two values is meaningful, and the values will be evenly distributed, but there will be no true zero.
10	YR_RMDL	Nominal	The type of YR_RMDL is 'Nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
11	EYB	Interval	The type of EYB is 'Interval' because variable values can be compared in size, and the difference between the two values is meaningful, and the values will be evenly distributed, but there will be no true zero.
12	STORIES	Ordinal	The type of STORIES is 'Ordinal' because can categorize and rank data, but cannot tell for the interval value between rankings. It can be classified according to different values, but this does not mean that higher floors or lower floors have more advantages.
13	SALEDATE	Interval	The type of SALEDATE is 'Interval' because variable values can be compared in size, and the difference between the two values is meaningful, and the values will be evenly distributed, but there will be no true zero. Because it could not have a date of sale that showed it was AD 0.
14	PRICE	Ratio	The type of PRICE is 'Ratio' because it has a value that defines the zero point, which is the proportional origin. And it can sort based on the exact difference between the values.

			Zero means free. Higher prices mean more expensive.
15	SALE_NUM	Nominal	The type of SALE_NUM is 'Nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups. There is no fixed logical order for sale number.
16	GBA	Ratio	The type of GBA is 'Ratio' because it has a value that defines the zero point, which is the proportional origin. And it can sort based on the exact difference between the values. Zero represents no floor area.
17	BLDG_NUM	Nominal	The type of BLDG_NUM is 'Nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
18	STYLE	Nominal	The type of STYLE is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
19	STYLE_D	Nominal	The type of STYLE_D is 'Nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups. There is no fixed logical order for sale number.
20	STRUCT	Nominal	The type of STRUCT is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to

			label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
21	STRUCT_D	Nominal	The type of STRUCT_D is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
22	GRADE	Interval	The type of GRADE is 'Interval' because variable values can be compared in size, and the difference between the two values is meaningful, and the values will be evenly distributed, but there will be no true zero. The higher the code, the better the level.
23	GRADE_D	Nominal	The type of GRADE_D is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
24	CNDTN	Interval	The type of CNDTN is 'Interval' because variable values can be compared in size, and the difference between the two values is meaningful, and the values will be evenly distributed, but there will be no true zero. The higher the code, the better the level.
25	CNDTN_D	Nominal	The type of CNDTN_D is 'Nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
26	EXTWALL	Nominal	The type of EXTWALL is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually

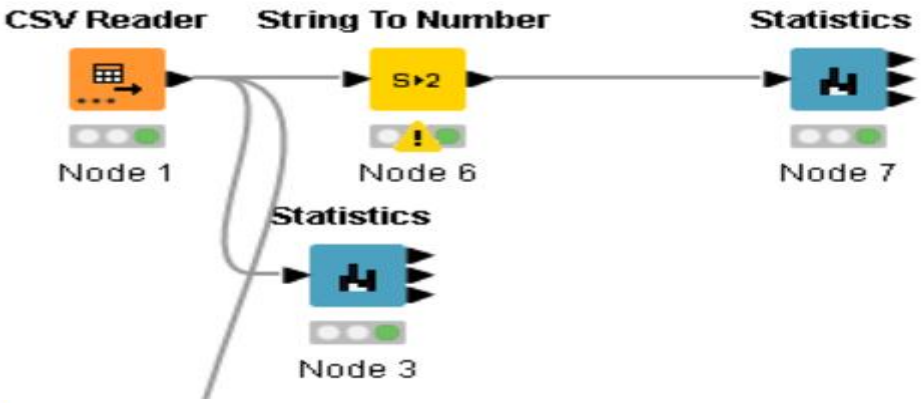
			exclusive categories or groups.
27	EXTWALL_D	Nominal	The type of EXTWALL_D is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
28	ROOF	Nominal	The type of ROOF is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
29	ROOF_D	Nominal	The type of ROOF_D is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
30	INTWALL	Nominal	The type of INTWALL is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
31	INTWALL_D	Nominal	The type of INTWALL_D is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
32	KITCHENS	Ratio	The type of KITCHENS is 'Ratio' because it has a value that defines the zero point, which is the proportional origin. And it can sort based on the exact difference between the values. Zero represents no kitchen.
33	FIREPLACES	Ratio	The type of FIREPLACES is 'Ratio' because it has a value that defines the zero

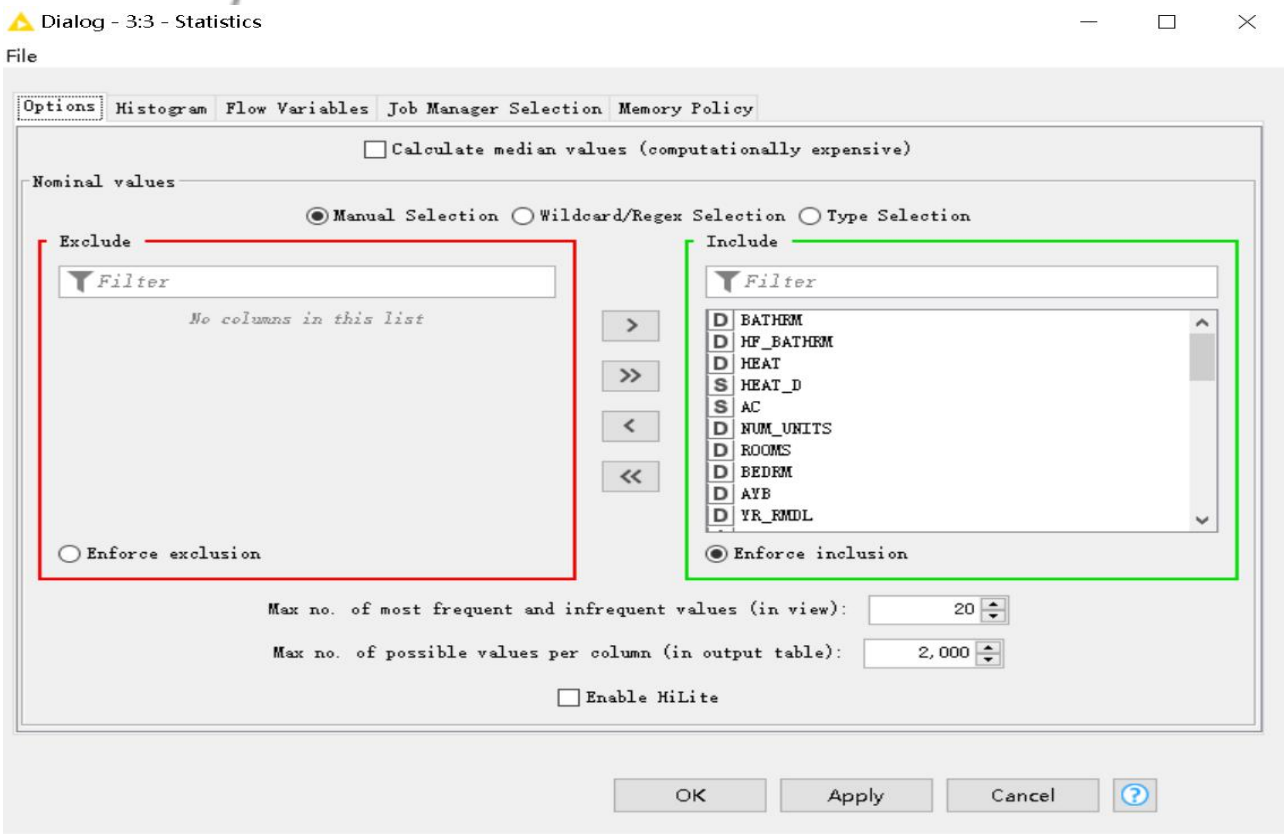
			point, which is the proportional origin. And it can sort based on the exact difference between the values. Zero represents no fireplaces.
34	USECODE	Nominal	The type of USECODE is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
35	LANDAREA	Ratio	The type of LANDAREA is 'Ratio' because it has a value that defines the zero point, which is the proportional origin. And it can sort based on the exact difference between the values. Zero represents no floor area.
36	GIS_LAST_MOD_DTTM	Nominal	The type of GIS_LAST_MOD_DTTM is 'nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups.
37	QUALIFIED	Nominal	The type of QUALIFIED is 'Nominal' because it is categorical variables without ranking and is not sortable. The data is a discrete unit that can be represented and is used to label variables but has no quantitative value. The purpose of this data is to classify the data into mutually exclusive categories or groups. There is no fixed logical order for qualified and unqualified types.

2. Summarising properties for the attributes

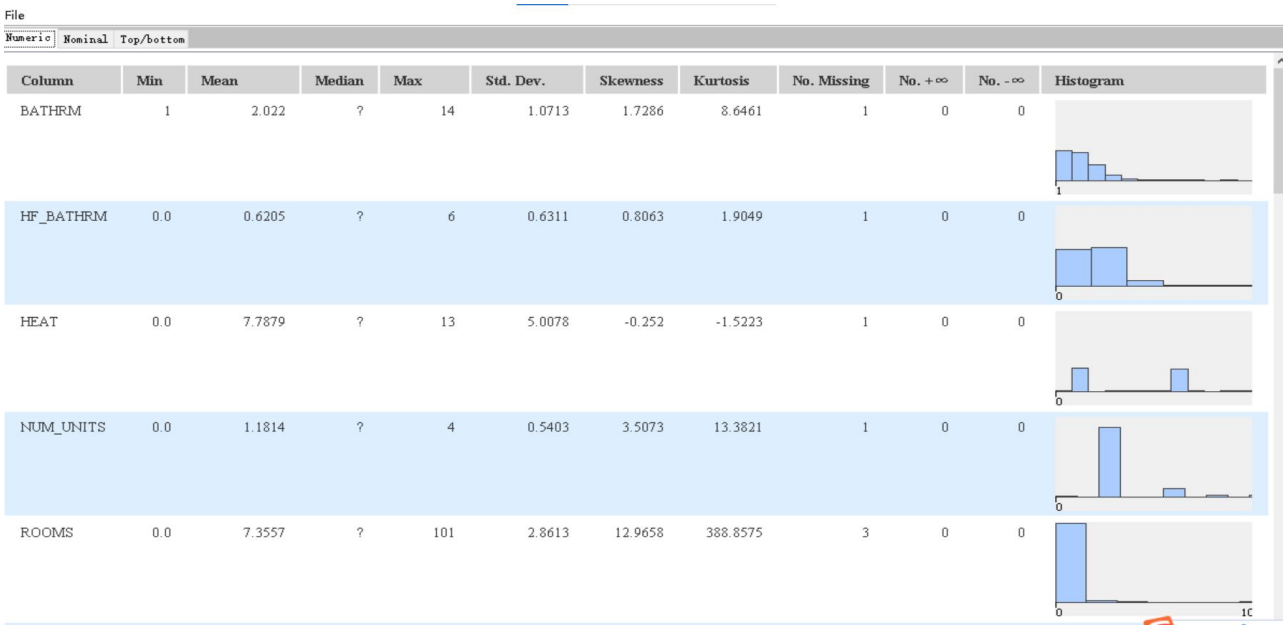
2.1 Data Visualisations

2.1.1 Utilizing "statistics" to view data

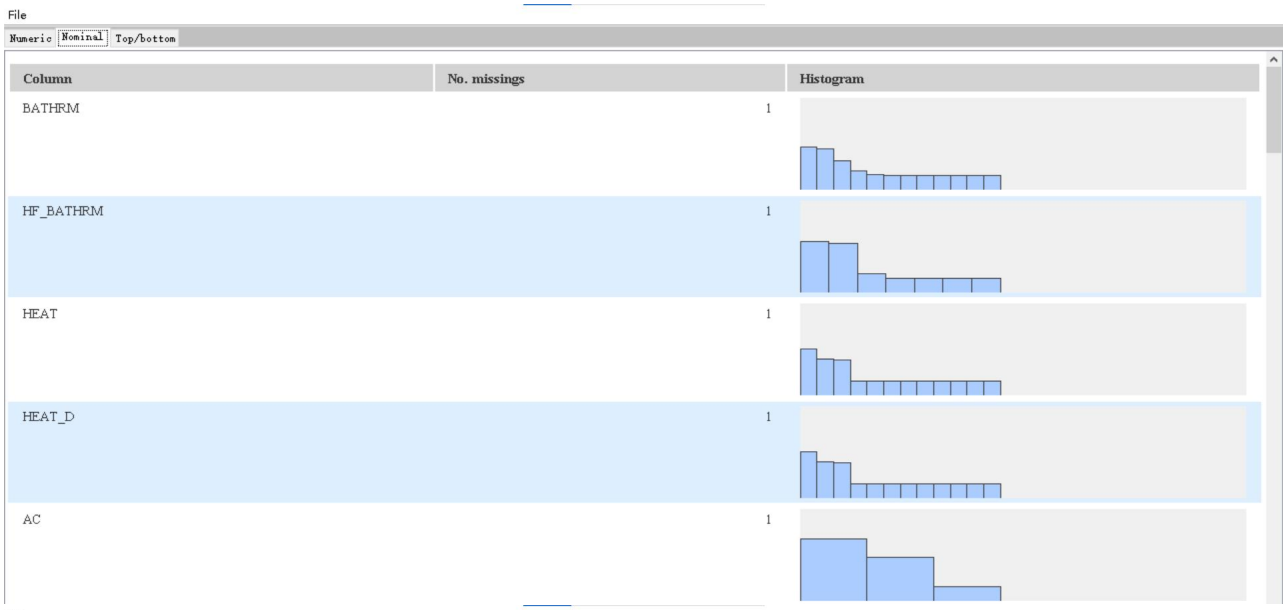


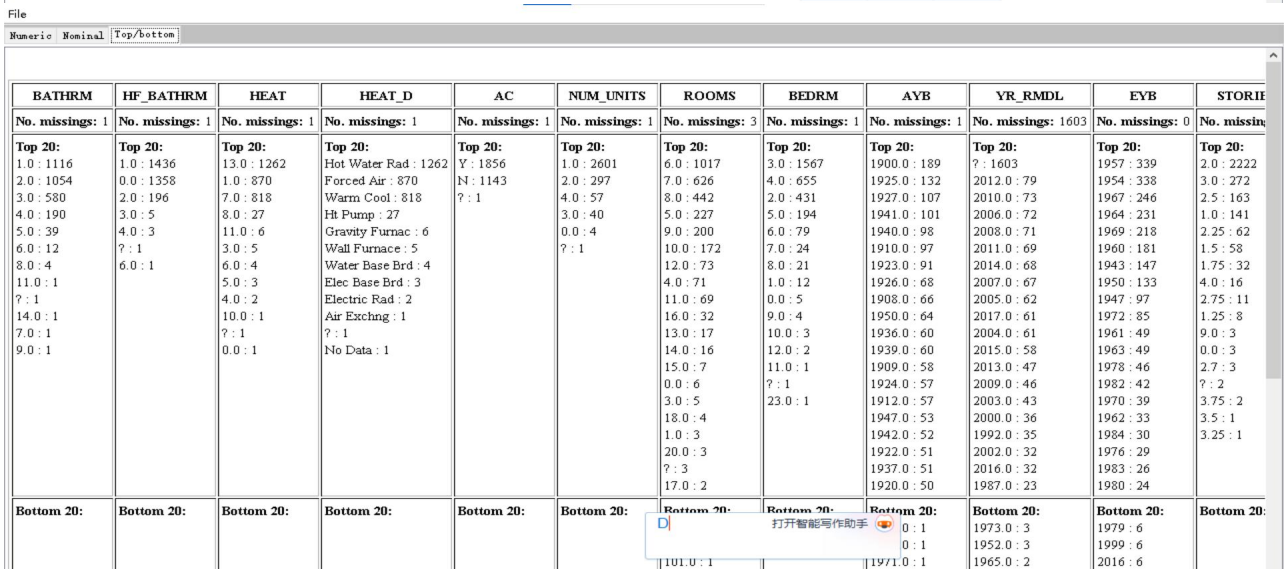


Result:



Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
BATHRM	1	2.022	?	14	1.0713	1.7286	8.6461	1	0	0	
HF_BATHRM	0.0	0.6205	?	6	0.6311	0.8063	1.9049	1	0	0	
HEAT	0.0	7.7879	?	13	5.0078	-0.252	-1.5223	1	0	0	
NUM_UNITS	0.0	1.1814	?	4	0.5403	3.5073	13.3821	1	0	0	
ROOMS	0.0	7.3557	?	101	2.8613	12.9658	388.8575	3	0	0	





BATHRM	HF_BATHRM	HEAT	HEAT_D	AC	NUM_UNITS	ROOMS	BEDRM	AYB	YR_RMDL	EYB	STORIE
No. missings: 1	No. missings: 1	No. missings: 1	No. missings: 1	No. missings: 1	No. missings: 1	No. missings: 3	No. missings: 1	No. missings: 1	No. missings: 1603	No. missings: 0	No. missin
Top 20: 1.0 : 1116 2.0 : 1054 3.0 : 580 4.0 : 190 5.0 : 39 6.0 : 12 8.0 : 4 11.0 : 1 ? : 1 14.0 : 1 7.0 : 1 9.0 : 1	Top 20: 1.0 : 1436 0.0 : 1358 2.0 : 196 3.0 : 5 4.0 : 3 ? : 1 6.0 : 1	Top 20: 13.0 : 1262 1.0 : 870 7.0 : 818 8.0 : 27 11.0 : 6 3.0 : 5 6.0 : 4 5.0 : 3 4.0 : 2 10.0 : 1 ? : 1 0.0 : 1	Top 20: Hot Water Rad : 1262 Forced Air : 870 Warm Cool : 818 Ht Pump : 27 Gravity Furnac : 6 Wall Furnace : 5 Water Base Brd : 4 Elec Base Brd : 3 Electric Rad : 2 Air Exchng : 1 ? : 1 No Data : 1	Top 20: Y : 1856 N : 1143 ? : 1	Top 20: 1.0 : 2601 2.0 : 297 4.0 : 57 3.0 : 40 0.0 : 4 ? : 1	Top 20: 6.0 : 1017 7.0 : 626 8.0 : 442 5.0 : 227 9.0 : 200 10.0 : 172 12.0 : 73 4.0 : 71 11.0 : 69 16.0 : 32 13.0 : 17 14.0 : 16 15.0 : 7 0.0 : 6 3.0 : 5 18.0 : 4 1.0 : 3 20.0 : 3 ? : 3 17.0 : 2	Top 20: 3.0 : 1567 4.0 : 655 2.0 : 431 5.0 : 194 6.0 : 79 7.0 : 24 8.0 : 21 1.0 : 12 0.0 : 5 9.0 : 4 10.0 : 3 12.0 : 2 11.0 : 1 ? : 1 23.0 : 1	Top 20: 1900.0 : 189 1925.0 : 132 1927.0 : 107 1941.0 : 101 1940.0 : 98 1910.0 : 97 1923.0 : 91 1926.0 : 68 1908.0 : 66 1950.0 : 64 1936.0 : 60 1939.0 : 60 1909.0 : 58 1924.0 : 57 1912.0 : 57 1947.0 : 53 1942.0 : 52 1922.0 : 51 1937.0 : 51 1920.0 : 50	Top 20: ? : 1603 2012.0 : 79 2010.0 : 73 2006.0 : 72 2008.0 : 71 2011.0 : 69 2014.0 : 68 2007.0 : 67 2005.0 : 62 2017.0 : 61 2004.0 : 61 2015.0 : 58 2013.0 : 47 2009.0 : 46 2003.0 : 43 2000.0 : 36 1992.0 : 35 2002.0 : 32 2016.0 : 32 1987.0 : 23	Top 20: 1957 : 339 1954 : 338 1967 : 246 1964 : 231 1969 : 218 1960 : 181 1943 : 147 1950 : 133 1947 : 97 1972 : 85 1961 : 49 1963 : 49 1978 : 46 1982 : 42 1970 : 39 1962 : 33 1984 : 30 1976 : 29 1983 : 26 1980 : 24	Top 20: 2.0 : 2222 3.0 : 272 2.5 : 163 1.0 : 141 2.25 : 62 1.5 : 58 1.75 : 32 4.0 : 16 2.75 : 11 1.25 : 8 9.0 : 3 0.0 : 3 2.7 : 3 ? : 2 3.75 : 2 3.5 : 1 3.25 : 1
Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20:	Bottom 20: 0 : 1 0 : 1 1971.0 : 1	Bottom 20: 1973.0 : 3 1952.0 : 3 1965.0 : 2	Bottom 20: 1979 : 6 1999 : 6 2016 : 6	Bottom 20:

2.2 Utilizing “GroupBy” to get “median” value

2.2.1 Judgment attributes

The reason for selecting the property:

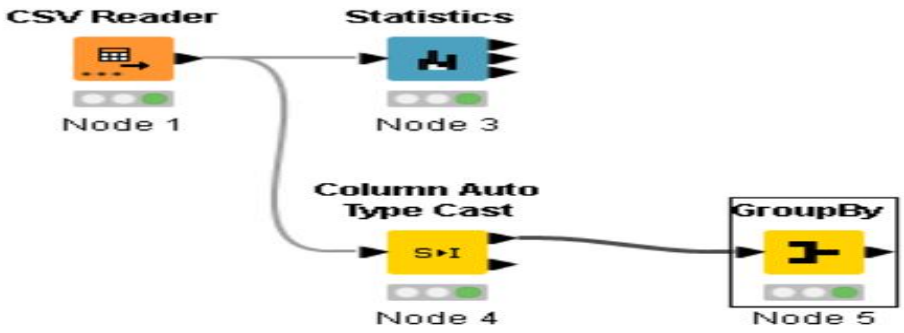
1. The selling price of the property: Properties with a high **selling price** may be more popular and valuable than properties with a lower selling price.
For example: PRICE

2. Attributes of house facilities: A house with facilities such as **air conditioning, heating, fireplaces and kitchen** will be more popular than a house without it, and the more **rooms and bedrooms and bathrooms** will be more popular than less. A house with **a larger land area** will be more valuable than a house with a smaller land area. The better the room **condition and grade**, the higher the likelihood of popularity.
For example: BATHRM, HF_BATHRM, HEAT, NUM_UNITS, ROOMS, BEDRM, GBA, STYLE, GRADE, CNDTN, Kitchens, Fireplaces, LANDAREA, QUALIFIED

3. Date: The **later the house** is built, the more likely it is to be popular. The **later the renovation** is built, the more popular it is likely to be.

For example: AYB, YR_RMDL, EYB

2.2.2 Operate



File

Settings | Description | Flow Variables | Job Manager Selection | Memory Policy

Groups | Manual Aggregation | Pattern Based Aggregation | Type Based Aggregation

Aggregation settings

Available columns

D	BATHRM
D	HF_BATHRM
D	HEAT
S	HEAT_D
S	AC
D	NUM_UNITS
D	ROOMS
D	BEDRM
D	AYB
D	YR_RMDL
I	EYB
D	STORIES
S	SALEDATE
D	PRICE
I	SALE_NUM
I	GBA
I	BLDG_NUM
D	STYLE
S	STYLE_D

Select

add >>

add all >>

<< remove

<< remove all

To change multiple columns use right mouse click for context menu.

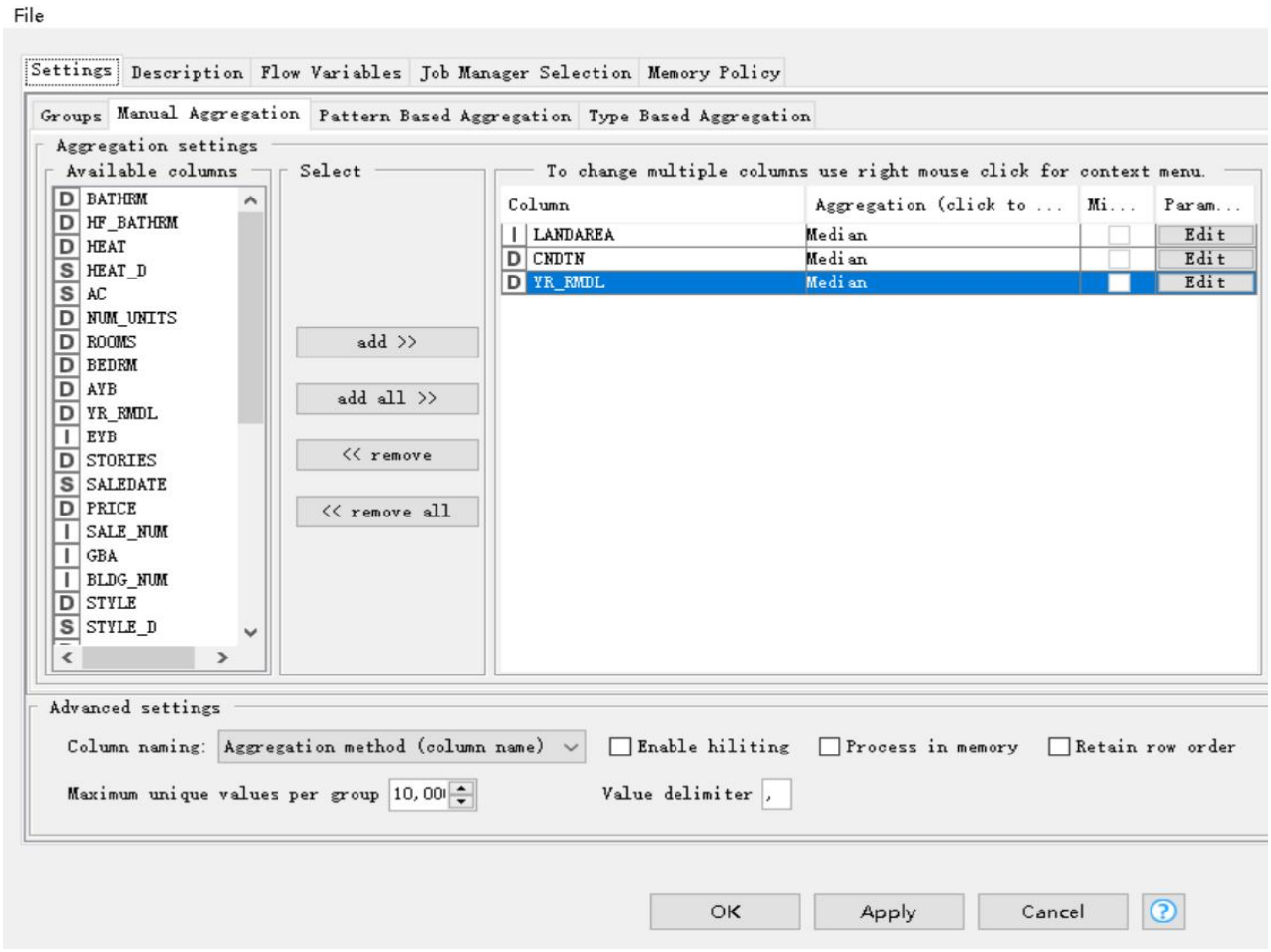
	Column	Aggregation (click to ...	Mi...	Param...
D	BATHRM	Median		Edit
D	HF_BATHRM	Median		Edit
D	ROOMS	Median		Edit
D	NUM_UNITS	Median		Edit
D	HEAT	Median		Edit
D	BEDRM	Median		Edit
D	AYB	Median		Edit
S	AC	First	☑	
I	EYB	Median		Edit
S	SALEDATE	First	☑	
D	PRICE	Median		Edit
I	GBA	Median		Edit
D	STYLE	Median		Edit
D	GRADE	Median		Edit
D	KITCHENS	Median		Edit
D	FIREPLACES	Median		Edit
I	QUALIFIED	Median		Edit

Advanced settings

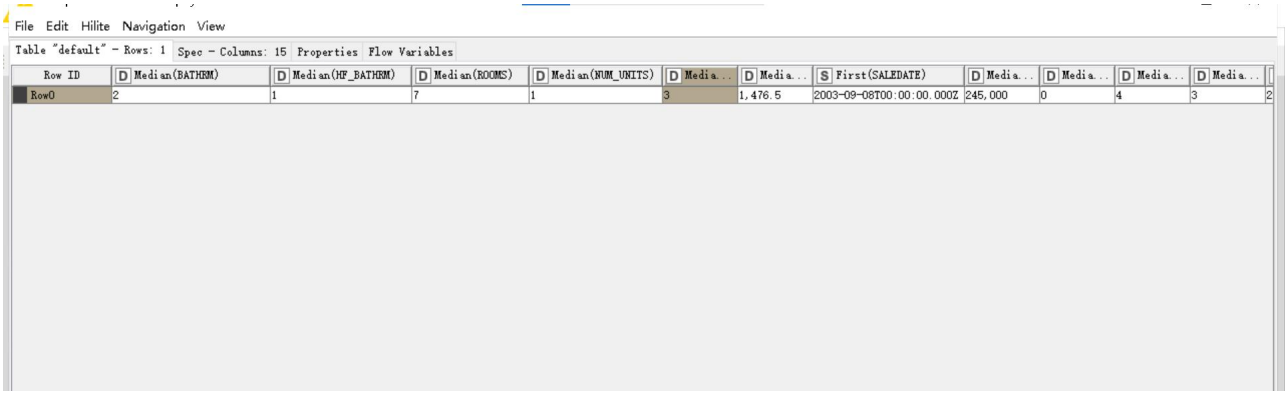
Column naming: Aggregation method (column name) | Enable hiliting | Process in memory | Retain row order

Maximum unique values per group 10,000 | Value delimiter ,

OK | Apply | Cancel



2.2.3 Result



3. Recorded data

1. BATHRM

Statistics	Value
Mean	2.022
Median	2
Minimum Value	1
Maximum Value	14
Standard deviation	1.071
Variance	1.148
Skewness	1.729
Kurtosis	8.646
Overall sum	6,064

No. missing	1
Row count	3000

2. HF_BATHRM

Statistics	Value
Mean	0.621
Median	1
Minimum Value	0
Maximum Value	6
Standard deviation	0.631
Variance	0.398
Skewness	0.806
Kurtosis	1.905
Overall sum	1,861
No. missing	1
Row count	3000

3. HEAT

Statistics	Value
Mean	7.788
Median	7
Minimum Value	0
Maximum Value	13
Standard deviation	5.008
Variance	25.078
Skewness	-0.252
Kurtosis	-1.522
Overall sum	23,356
No. missing	1
Row count	3000

4. NUM_UNITS

Statistics	Value
Mean	1.181
Median	1
Minimum Value	0
Maximum Value	4
Standard deviation	0.54
Variance	0.292
Skewness	3.507
Kurtosis	13.382
Overall sum	3,543
No. missing	1
Row count	3000

5. ROOMS

Statistics	Value
Mean	7.356
Median	7
Minimum Value	0.0

Maximum Value	101
Standard deviation	2.861
Variance	8.187
Skewness	12.966
Kurtosis	388.858
Overall sum	22,045
No. missing	3
Row count	3000

6. BEDRM

Statistics	Value
Mean	3.367
Median	3
Minimum Value	0.0
Maximum Value	23
Standard deviation	1.168
Variance	1.363
Skewness	3.156
Kurtosis	31.989
Overall sum	10,099
No. missing	1
Row count	3000

7. AYB

Statistics	Value
Mean	1928.647
Median	1928
Minimum Value	0
Maximum Value	2019
Standard deviation	102.613
Variance	10,529.406
Skewness	-17.213
Kurtosis	318.898
Overall sum	5,784,013
No. missing	1
Row count	3000

8. YR_RMDL

Statistics	Value
Mean	2,001.188
Median	2006
Minimum Value	1,923
Maximum Value	2,018
Standard deviation	14.982
Variance	224.452
Skewness	-1.504
Kurtosis	2.169
Overall sum	2,795,659
No. missing	1603
Row count	3000

9. EYB

Statistics	Value
Mean	1,965.426
Median	1964
Minimum Value	0
Maximum Value	2,018
Standard deviation	39.439
Variance	1,555.471
Skewness	-41.197
Kurtosis	2,057.741
Overall sum	5,896,277
No. missing	0
Row count	3000

10. PRICE

Statistics	Value
Mean	391,045.743
Median	245,000
Minimum Value	0
Maximum Value	8,500,000
Standard deviation	562,599.823
Variance	316,518,560,278.7
Skewness	4.967
Kurtosis	47.743
Overall sum	957,279,978
No. missing	552
Row count	3000

11. GBA

Statistics	Value
Mean	1,712.341
Median	1476.5
Minimum Value	0
Maximum Value	16,658
Standard deviation	894.985
Variance	800,997.899
Skewness	4.038
Kurtosis	37.345
Overall sum	5,137,024
No. missing	0
Row count	3000

12. STYLE

Statistics	Value
Mean	4.376
Median	4
Minimum Value	0
Maximum Value	99

Standard deviation	2.281
Variance	5.203
Skewness	24.369
Kurtosis	989.513
Overall sum	13,123
No. missing	1
Row count	3000

13. GRADE

Statistics	Value
Mean	4.27
Median	4
Minimum Value	2
Maximum Value	12
Standard deviation	1.396
Variance	1.948
Skewness	1.503
Kurtosis	3.043
Overall sum	12,806
No. missing	1
Row count	3000

14. CNDTN

Statistics	Value
Mean	3.526
Median	3
Minimum Value	1
Maximum Value	6
Standard deviation	0.699
Variance	0.488
Skewness	0.9
Kurtosis	0.873
Overall sum	10,574
No. missing	1
Row count	3000

15. Kitchens

Statistics	Value
Mean	1.201
Median	1
Minimum Value	0
Maximum Value	4
Standard deviation	0.55
Variance	0.303
Skewness	3.312
Kurtosis	11.999
Overall sum	3,601
No. missing	1
Row count	3000

16. Fireplaces

Statistics	Value
Mean	0.627
Median	0
Minimum Value	0
Maximum Value	9
Standard deviation	0.895
Variance	0.8
Skewness	2.101
Kurtosis	7.5
Overall sum	1,879
No. missing	1
Row count	3000

17. LANDAREA

Statistics	Value
Mean	3,586.204
Median	2,314.5
Minimum Value	266
Maximum Value	691,817
Standard deviation	13,076.041
Variance	170,982,843.379
Skewness	48.815
Kurtosis	2,561.739
Overall sum	10,758,612
No. missing	0
Row count	3000

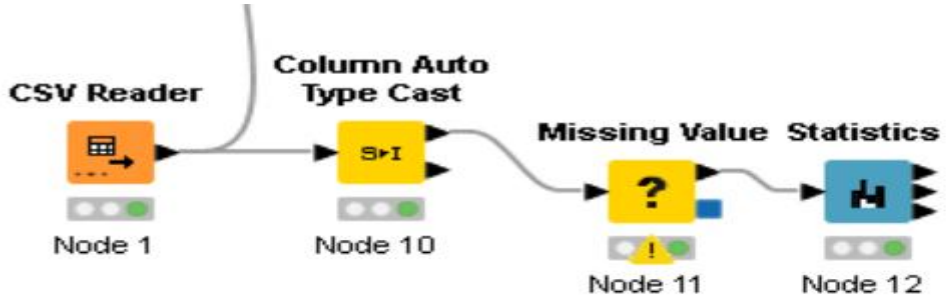
18. QUALIFIED

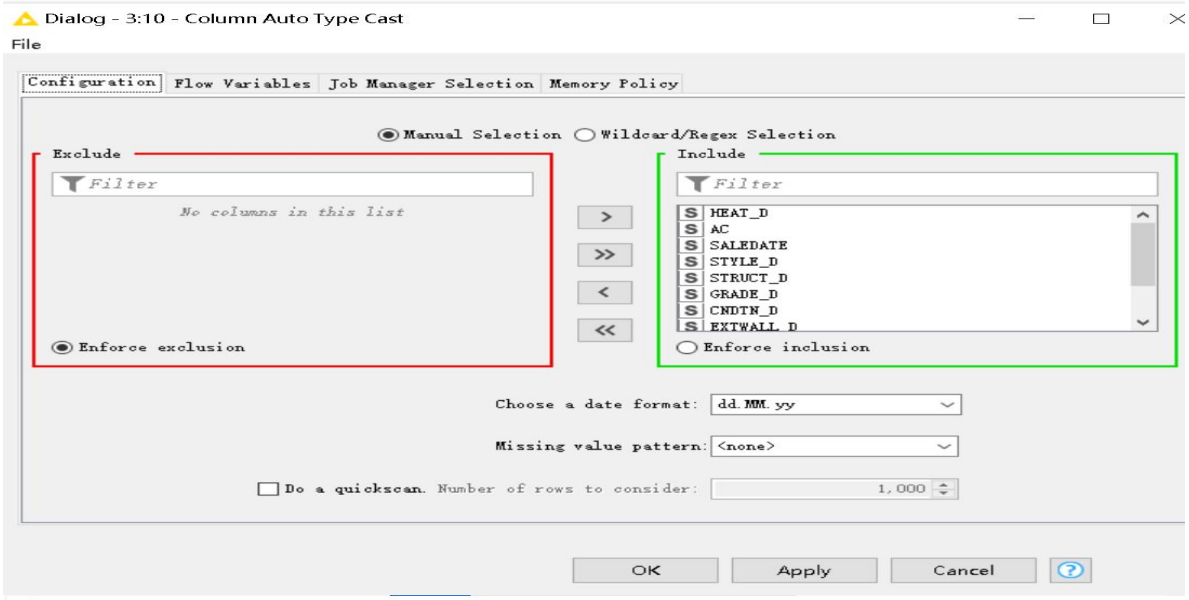
Statistics	Value
Mean	0436
Median	0
Minimum Value	0
Maximum Value	1
Standard deviation	0.496
Variance	0.246
Skewness	0.26
Kurtosis	-1.934
Overall sum	1,307
No. missing	0
Row count	3000

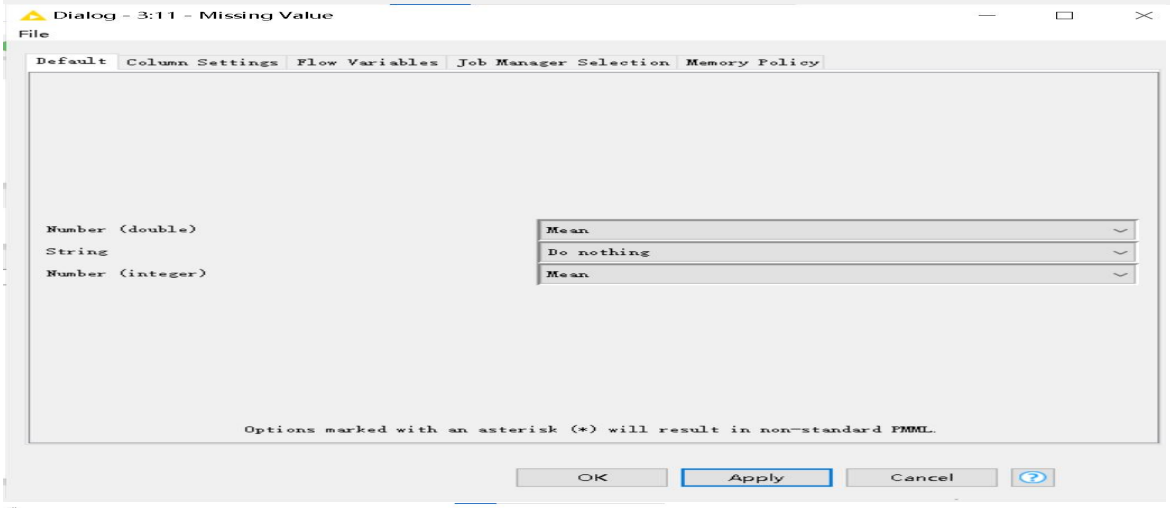
4. Dealing with the missing values

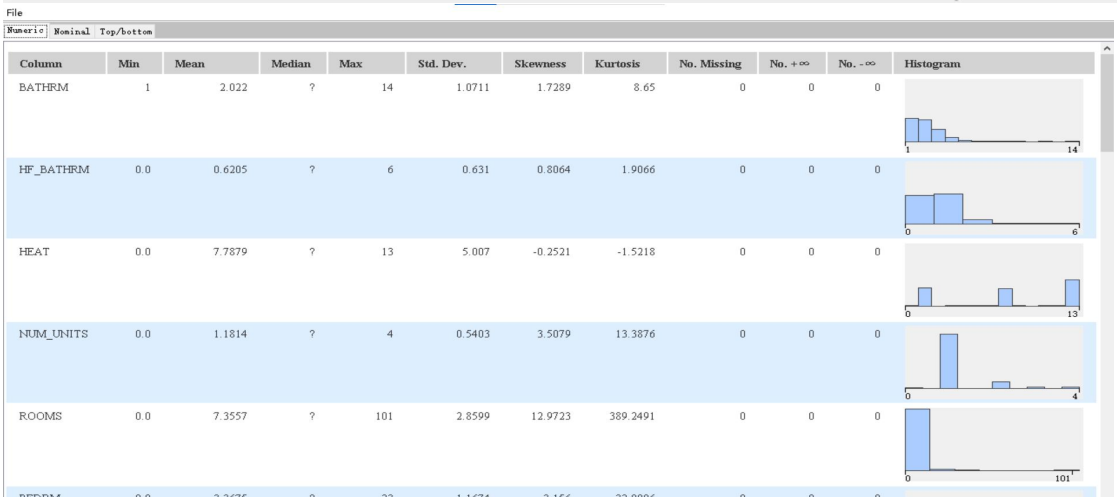
To solve the missing numbers in the data table using the mean or median instead of missing values, the tool used here is 'missing value'.

4.1 Use "Mean" to replace missing data









Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
BATHRM	1	2.022	?	14	1.0711	1.7289	8.65	0	0	0	
HF_BATHRM	0.0	0.6205	?	6	0.631	0.8064	1.9066	0	0	0	
HEAT	0.0	7.7879	?	13	5.007	-0.2521	-1.5218	0	0	0	
NUM_UNITS	0.0	1.1814	?	4	0.5403	3.5079	13.3876	0	0	0	
ROOMS	0.0	7.3557	?	101	2.8599	12.9723	389.2491	0	0	0	

File

Numeric | Nominal | Top/bottom

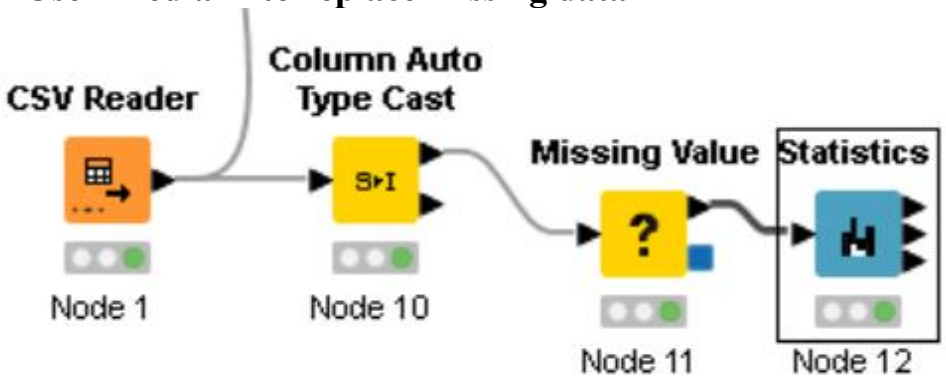
GBA	0.0	1,712.3413	?	16,658	894.9849	4.038	37.3445	0	0	0
BLDG_NUM	1	1.0007	?	2	0.0258	38.7105	1,497.4973	0	0	0
STYLE	0.0	4.3758	?	99	2.2805	24.3734	989.843	0	0	0
STRUCT	1	5.076	?	8	2.8483	-0.6208	-1.46	0	0	0
GRADE	2	4.2701	?	12	1.3954	1.5037	3.0445	0	0	0

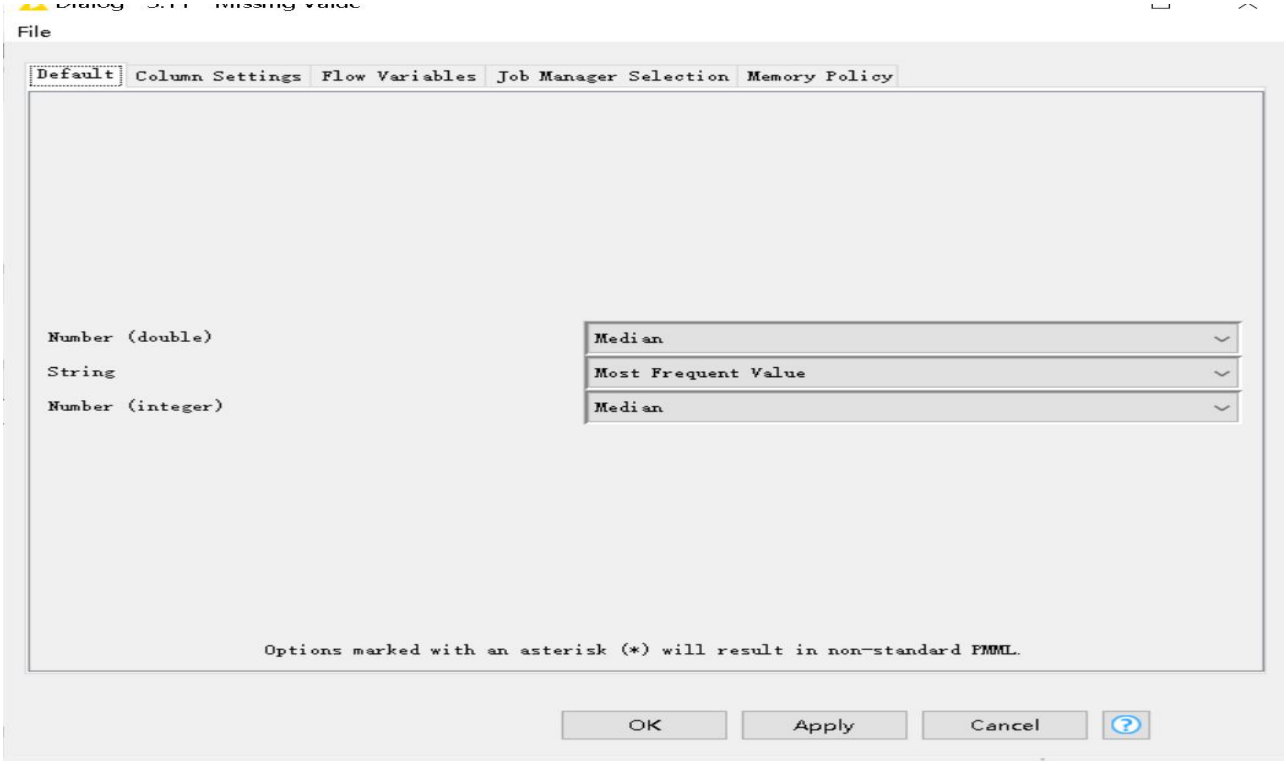
File

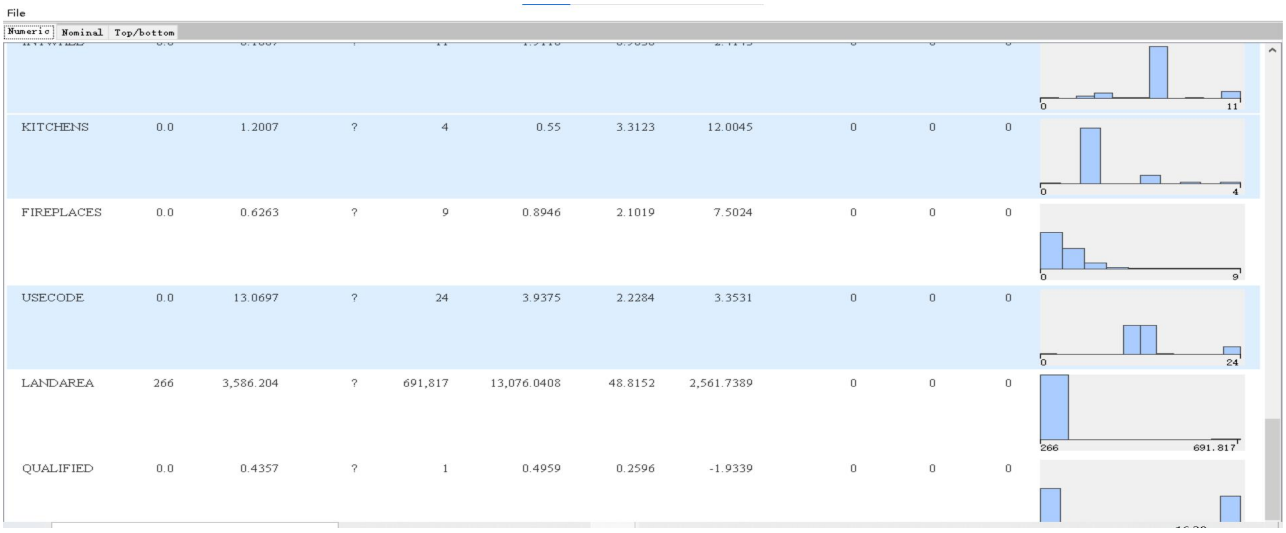
Numeric | Nominal | Top/bottom

Column	No. missings	Histogram
BATHRM	0	
HF_BATHRM	0	
HEAT	0	
HEAT_D	1	
AC	1	

4.2 Use “Median” to replace missing data





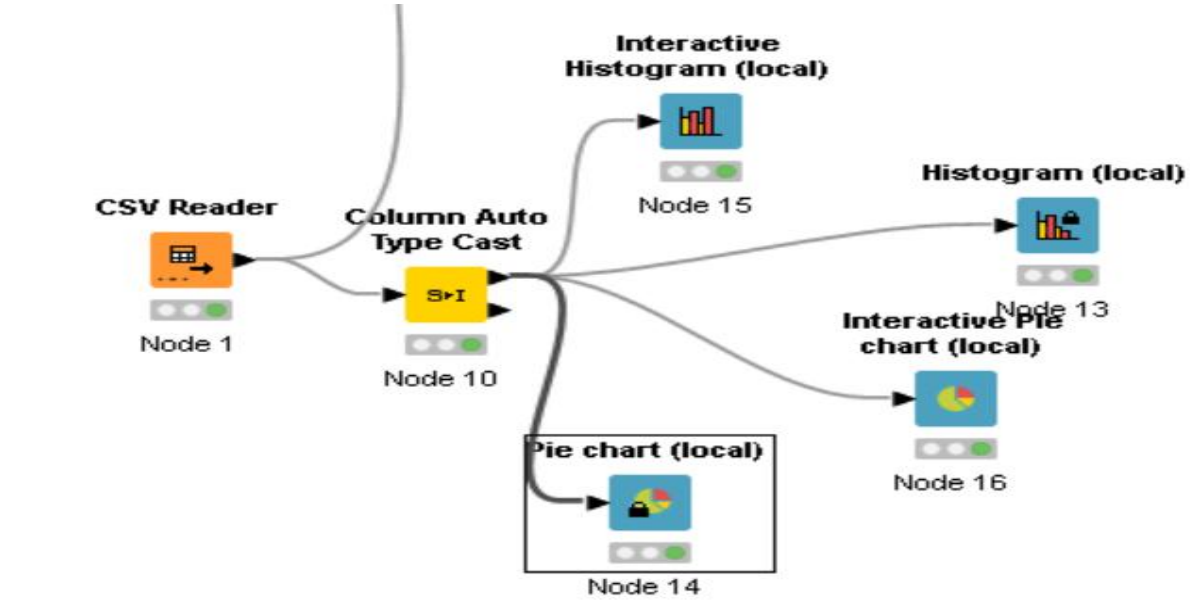


5. Visualized by frequency graph and pie charts

The methods used here contain: “Histogram” and “Pie chart”.

5.1 BATHRM

The pie chart (figure 2) and Frequency Graph(figure 1) below show the number of bedrooms, from 1 to 14. Only one bathroom was the largest proportion, with 37.2% of a total of 1,116 houses. The second largest was 2 bathrooms, with only 1,054 new homes accounting for 35.13 percent.Houses with three and four bedrooms accounted for 19.33% and 6.33%, respectively. Fewer houses own other numbers of bathrooms. Using the pie chart (figure 2) can clearly see the relatively large type of bathroom. But when the number of bathroom is greater than 5, the proportion of bathroom is less, there will be multiple data accumulation together is not easy to view.



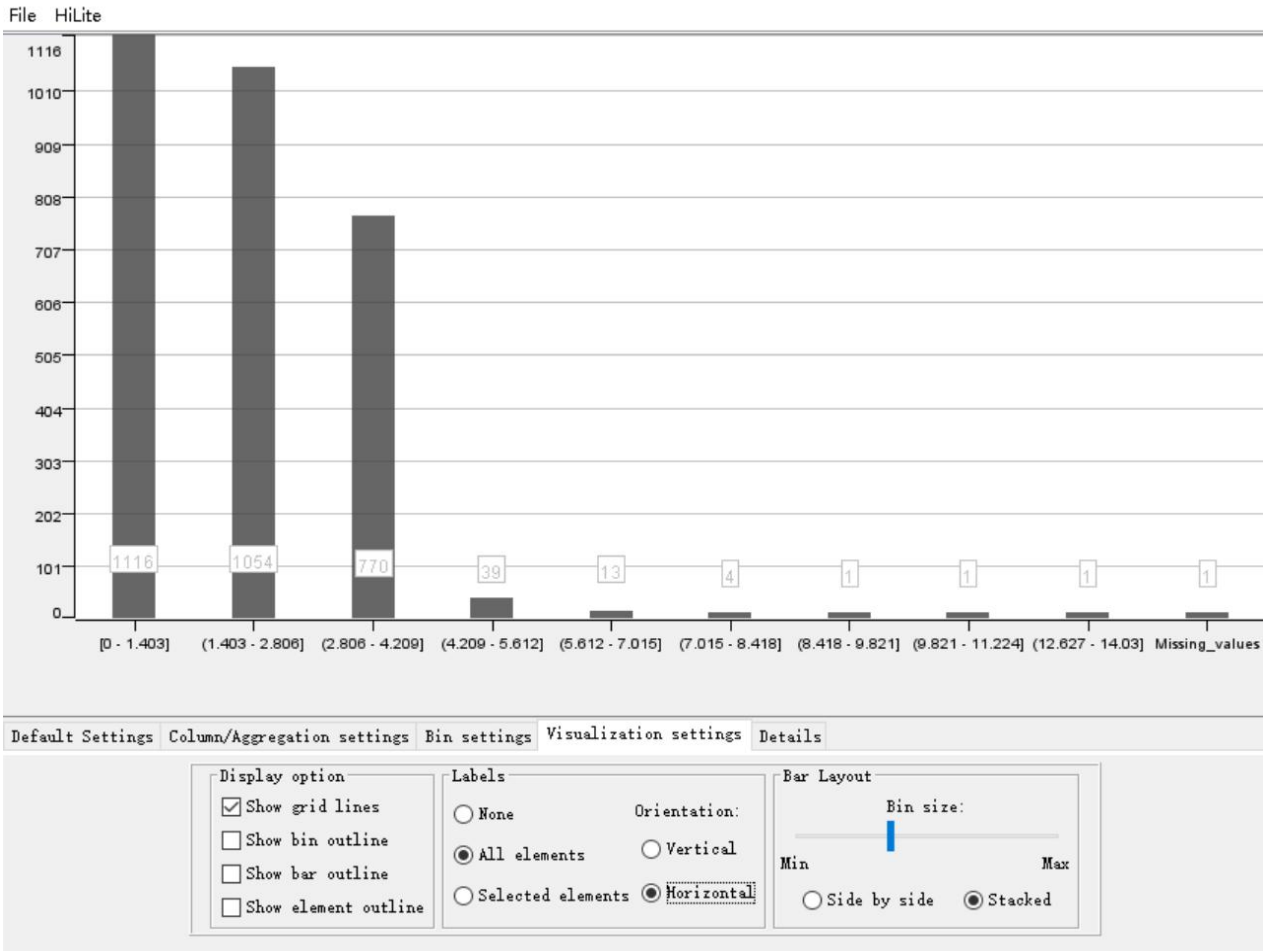


Figure 1 *Frequency Graph for BATHRM*

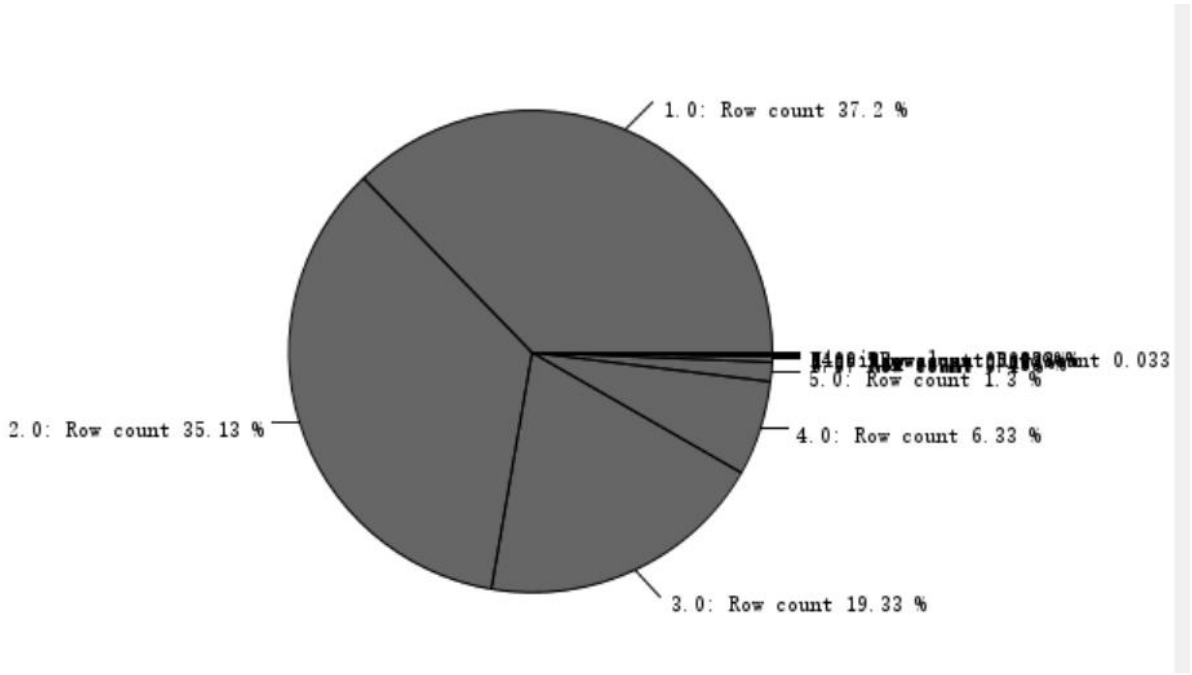


Figure 2 Pie Chart for BATHRM

5.2 NUM_UNITS

The Frequency Graph(figure 3) and pie chart (figure 4) below show the number of units , from 0 to 4. The largest number was only one unit, with 86.7% of 2,605 houses. Conversely, the smallest number of houses with 3 units that can be clearly seen from the pie chart was 1.33%. The proportion of houses with two and four units was 9.9% and 1.9%, and the number was 297 and 57, respectively. In this data, the data in the pie chart will be clearer than that in the frequency graph.It can be seen from here that the vast majority of houses have a unit.

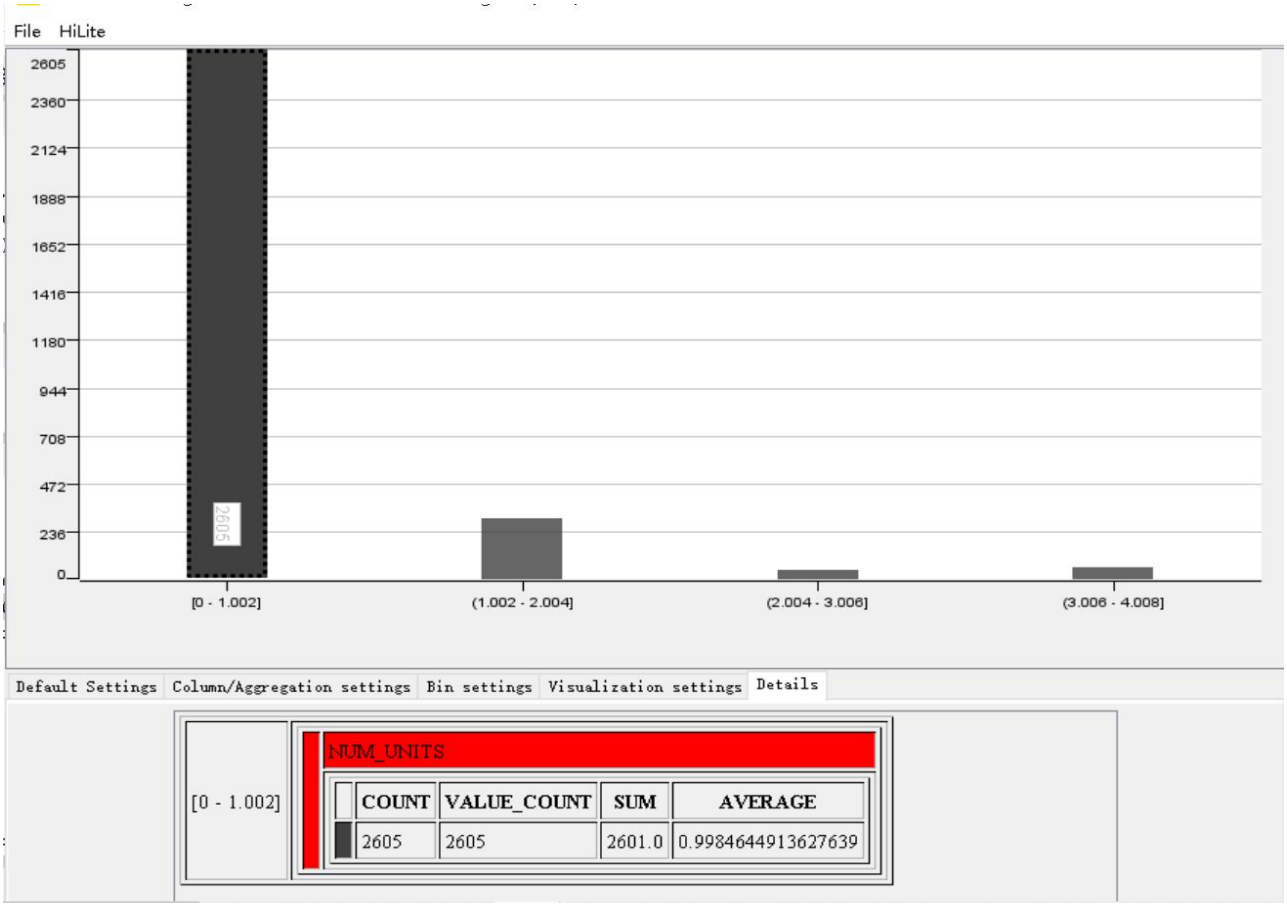


Figure 3 Frequency Graph for NUM_UNITS

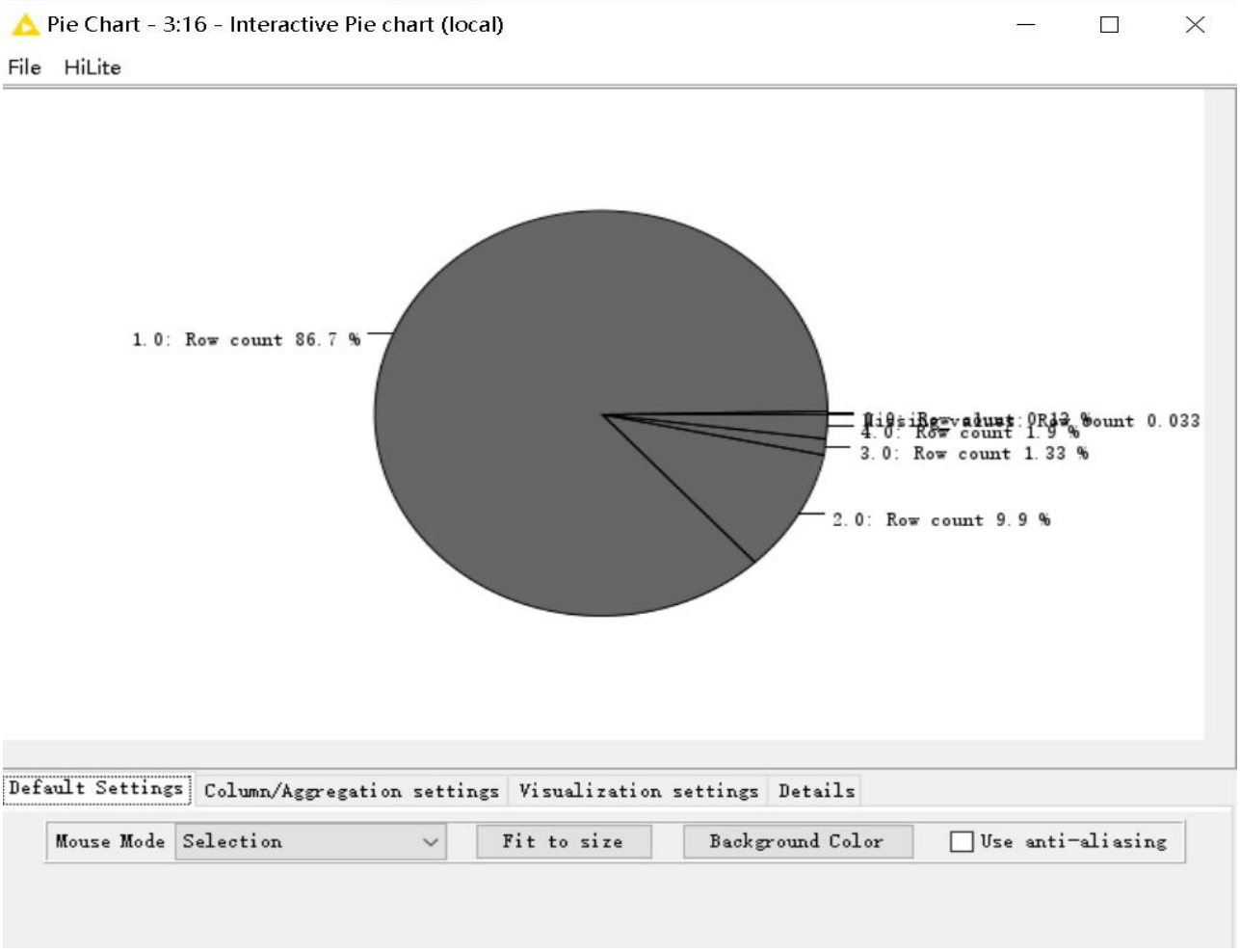


Figure 4 Pie Chart for NUM_UNITS

5.3 ROOMS

The Frequency Graph(figure 5) and pie chart (figure 6) below show the number of rooms , from 0 to 101. From the pie chart, we can clearly see that the most abundant type of house uses houses with six rooms, with 33.9% owning 1,017 houses. Obviously, the second largest proportion is 626 houses with 7 rooms in 20.87%. In this data, because the data is scattered, the frequency chart can clearly see the scattered general area, and the pie chart can clearly see the proportion of rooms.From the trends in the chart, it can be predicted that the number of rooms that the house is most likely to have is 3 to 11.

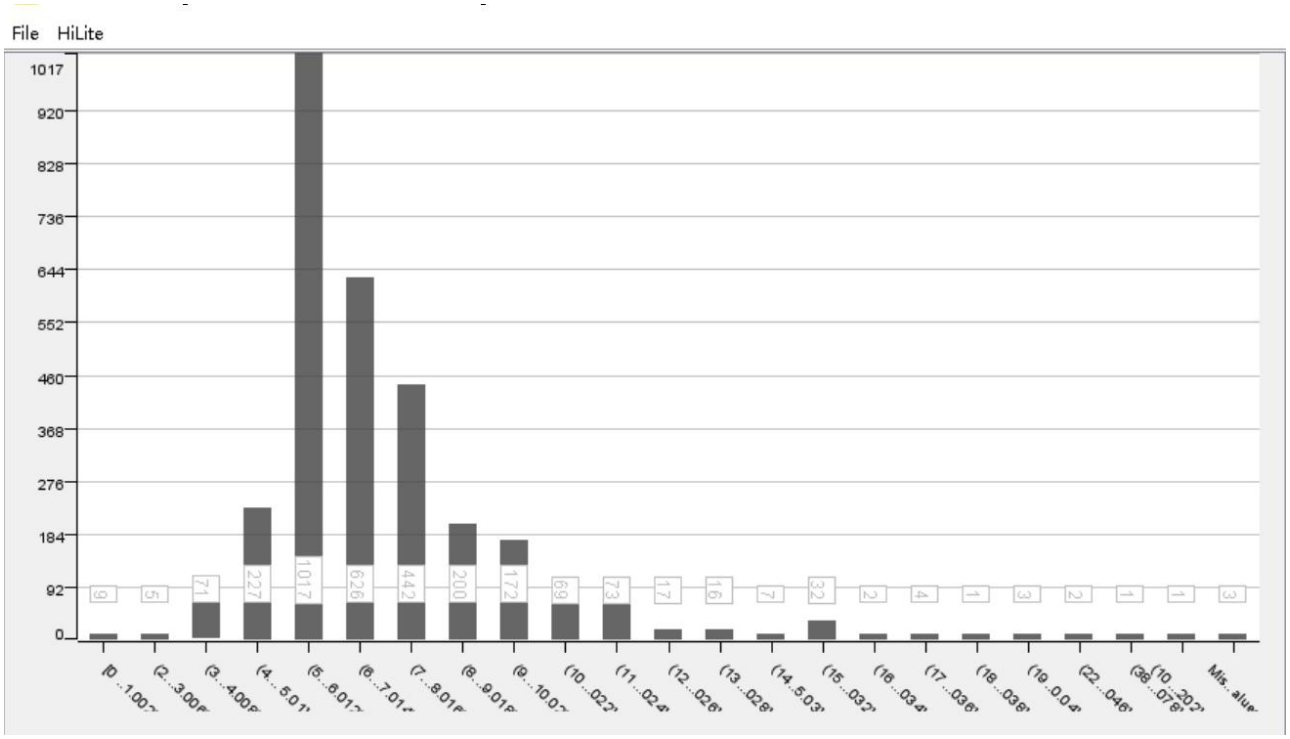


Figure 5 Frequency Graph for ROOMS

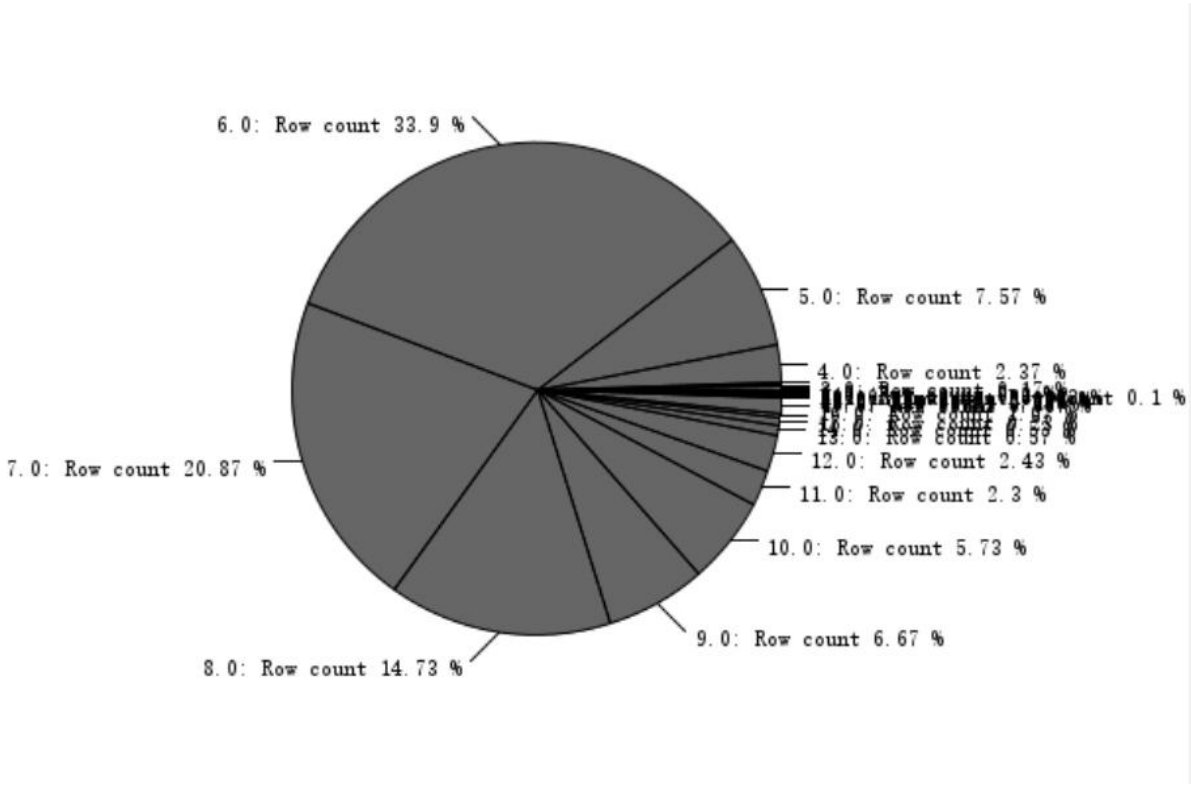


Figure 6 Pie Chart for ROOMS

5.4 BEDRM

The Frequency Graph(figure 7) and pie chart (figure 8) below show the number of bedroom , from 0 to 24. We can clearly see through the pie chart (figure 8) that the highest proportion of houses with three bedrooms is 52.23% and 1,567 houses. The proportion of houses with four bedrooms is 21.83%, and the proportion of houses with two bedrooms is 14.37%, and their number is 655 and 431, respectively. 6.47% of houses with five bedrooms and 2.63% with six bedrooms. Therefore, the vast majority of rooms are usually concentrated between two and six.

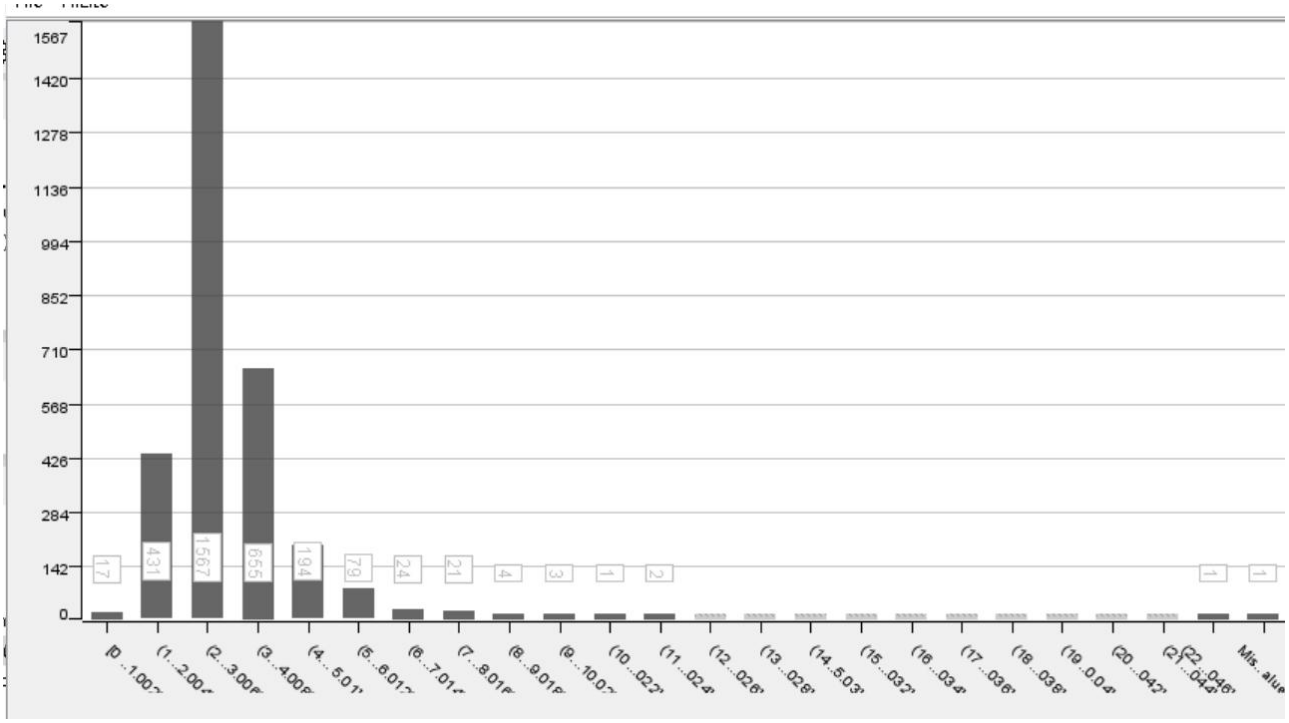


Figure 7 Frequency Graph for BEDRM

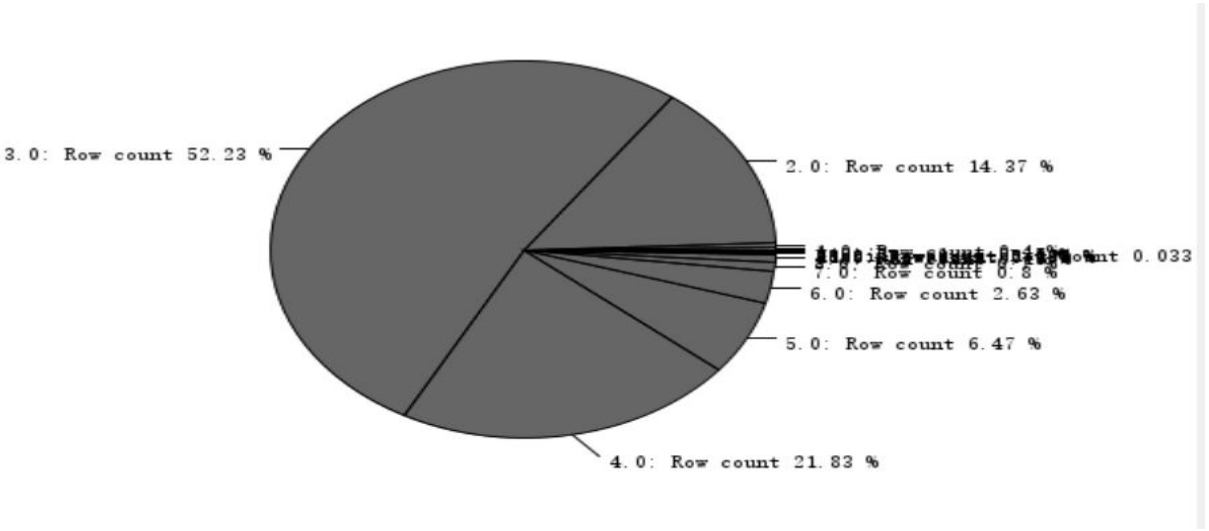


Figure 8 Pie Chart for BEDRM

5.5 GRADE_D

The Frequency Graph(figure 9) and pie chart (figure 10) below show Grade description. Through the frequency graph, it can be clearly seen that the average proportion is the largest, accounting for 35.7% of all room types, but the proportion of rooms below the average is very high, accounting for 28.97%, and the third proportion is good quality, accounting for 20.11%. In this data, a pie chart will more directly represent the area where each data is distributed.

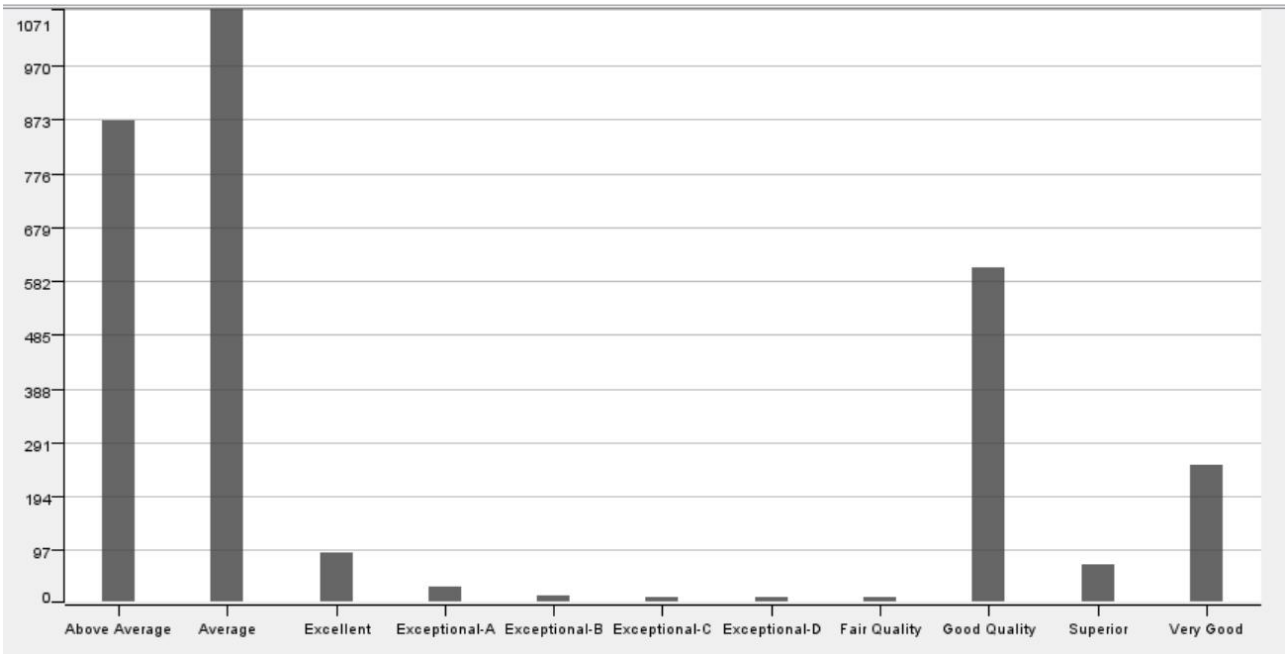


Figure 9 Frequency Graph for GRADE_D

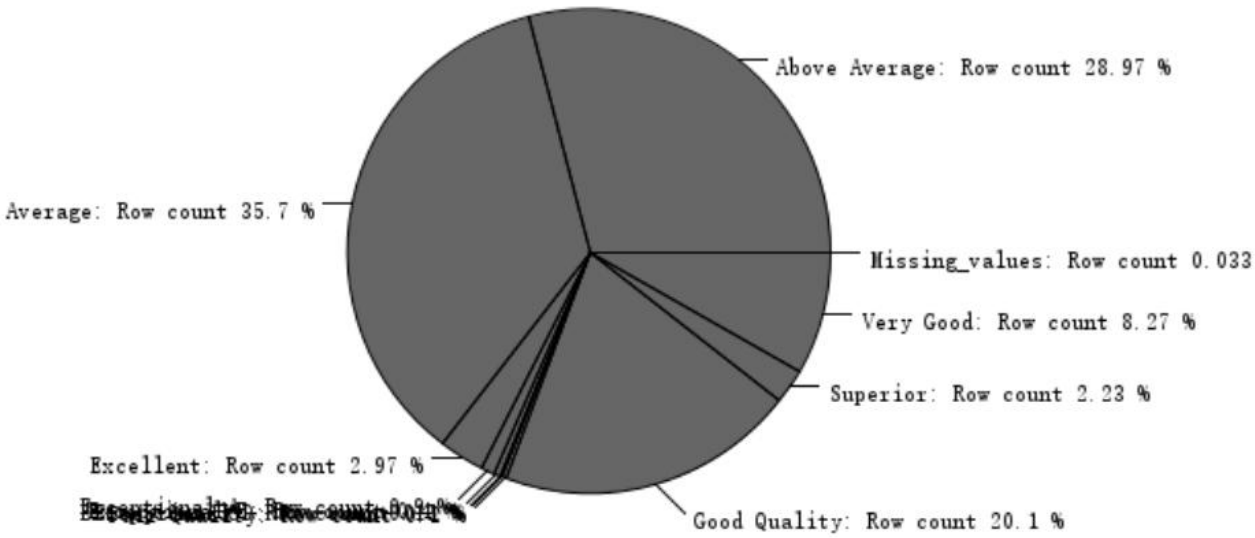


Figure 10 Pie Chart for GRADE_D

5.6 CNDTN_D

The Frequency Graph (figure 11) and pie chart (figure 12) below show Condition description. From the frequency distribution chart and pie chart, it can be clearly seen that the average proportion reached 54.73% at most, the number was 1642 houses, and the second place was a good house with 35.77%, with 1073 houses. In third place are very good rooms with 7.3%, the other rooms are excellent and fail as well as the frequency of poor to account for very little. As can be inferred from the chart, most of its house condition are concentrated in average and good.

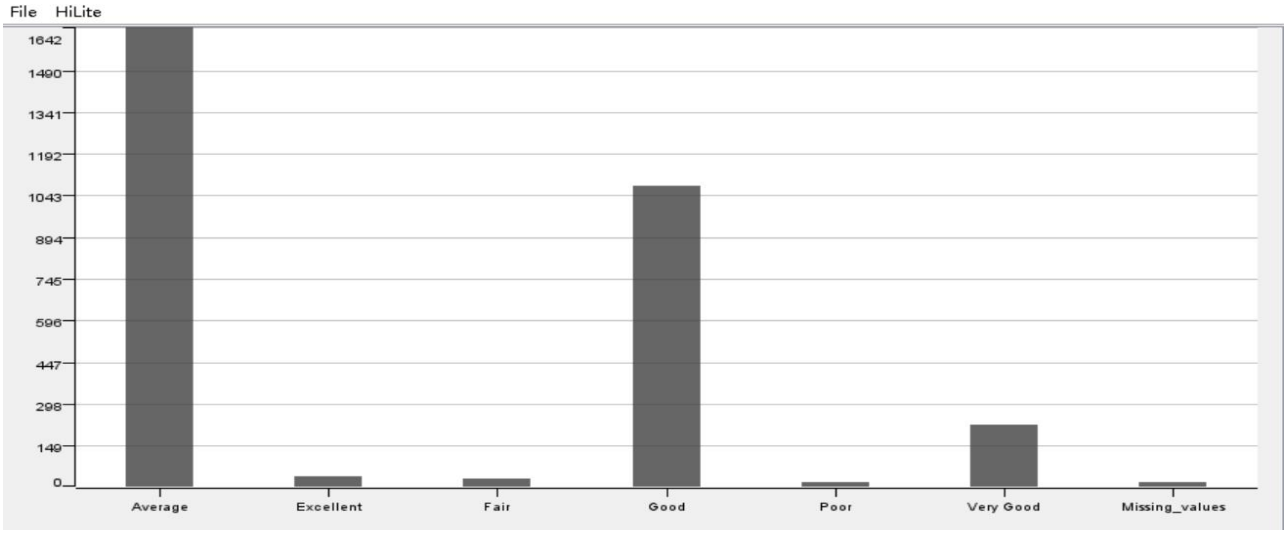


Figure 11 Frequency Graph for CNDTN_D

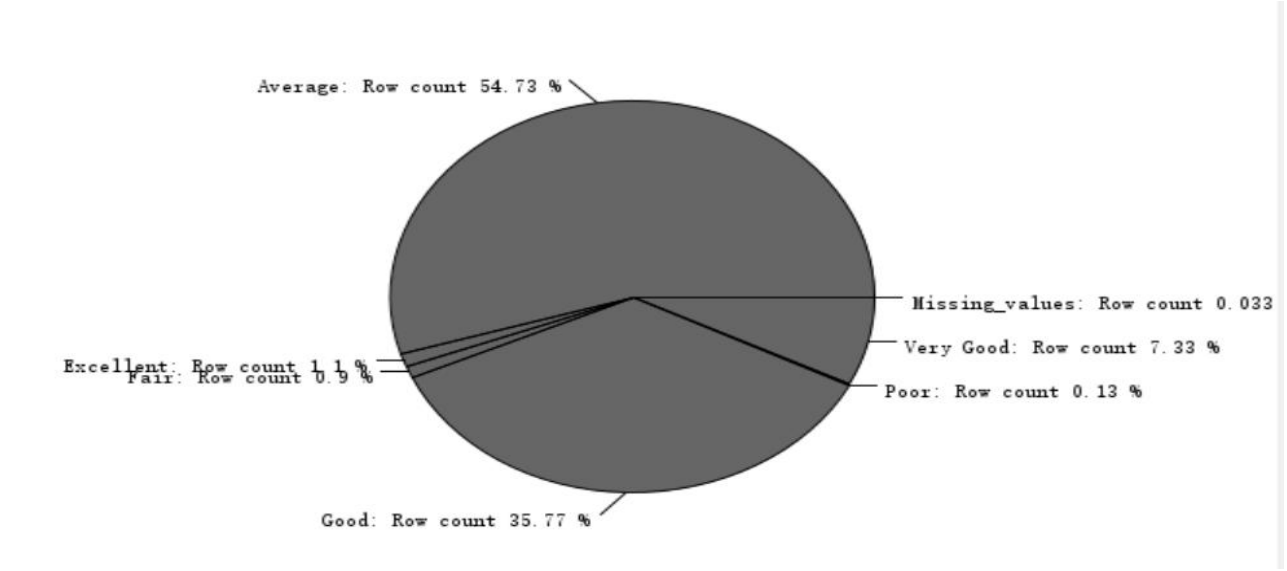


Figure 12 Pie Chart for CNDTN_D

5.7 KITCHENS

The Frequency Graph (figure 13) and pie chart (figure 14) below show kitchens. From the pie chart, it can be clearly seen that the proportion of houses with only one kitchen reaches 84.97%, and the proportion of houses with two kitchens is as high as 11.77%. In this data, the pie chart will be clearly clear and intuitive over the frequency chart. And we can roughly guess that the vast majority of houses only have one kitchen.

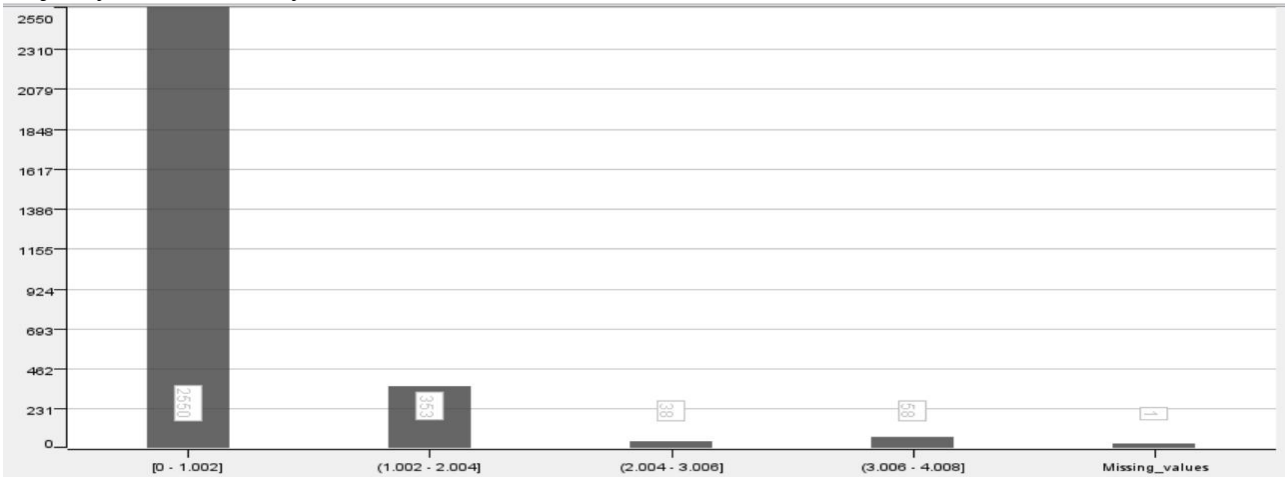


Figure 13 Frequency Graph for KITCHENS

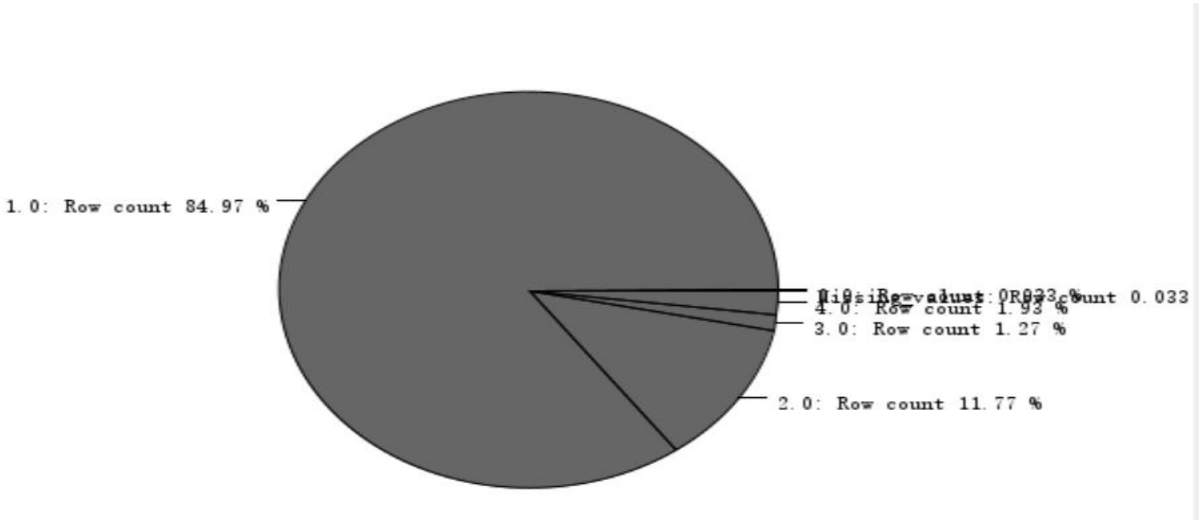


Figure 14 Pie Chart for KITCHENS

5.8 FIREPLACES

The Frequency Graph (figure 15) and pie chart (figure 16) below show FIREPLACES number. From chart 15 and 16, we can clearly see the rooms without fireplaces, accounting for the highest proportion of 56.17%, the number of which is 1685. The number of houses with a fireplace ranked second with 30.8%, with 924. In this data, a pie chart is more visually intuitive than a frequency chart.

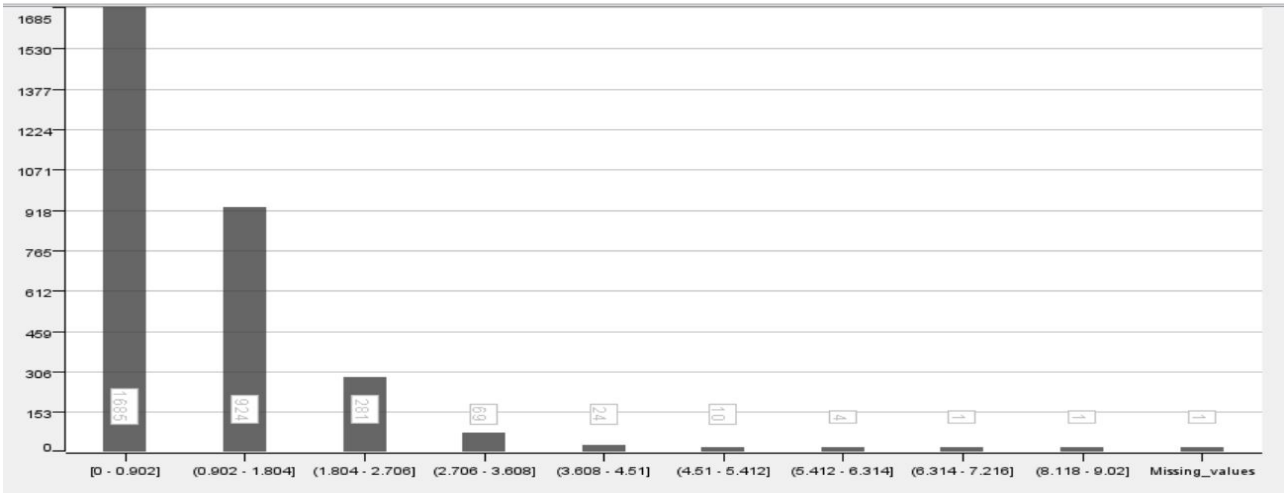


Figure 15 Frequency Graph for FIREPLACES

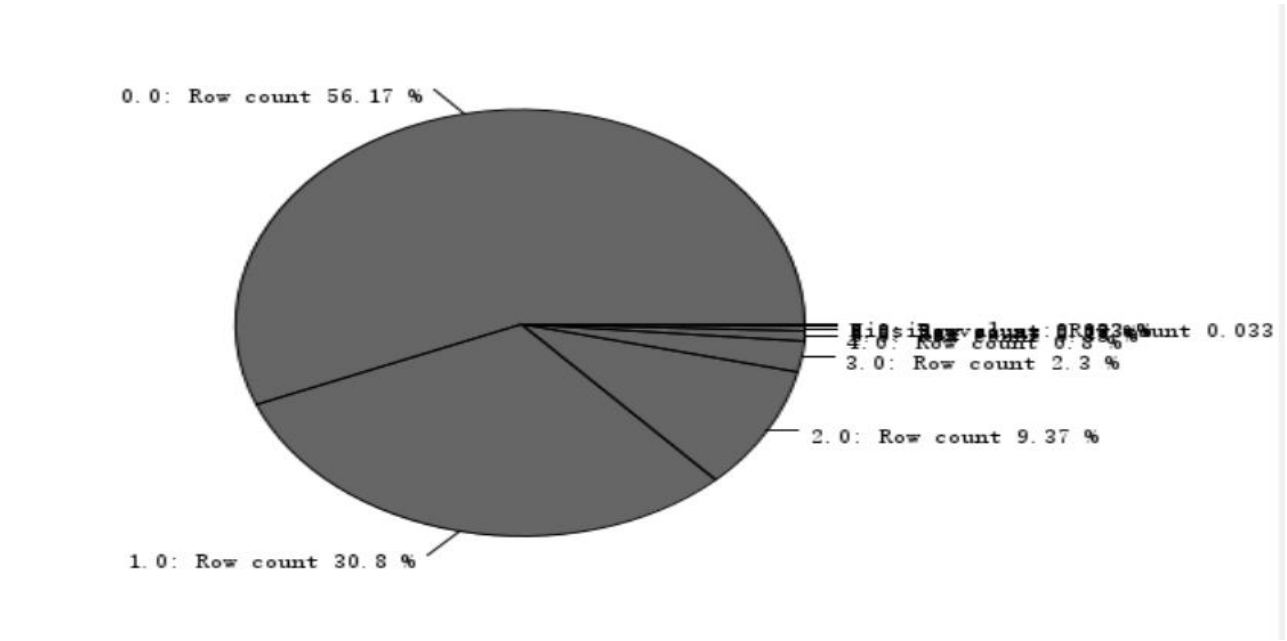


Figure 16 Pie Chart for FIREPLACES

5.9 QUALIFIED

The Frequency Graph (figure 17) and pie chart (figure 18) below show Qualified (Q) and unqualified (U). Utilising that logic in this data, '0' refers to Qualified (Q) and '1' refers to the unqualified (U). From the chart, it can be clearly seen that the proportion of qualified is up to 56.43%, the number is 1693, the number of unqualified accounts for 43.57%, and the number is 1307. This data pie chart will show the proportion more clearly and visually than the frequency chart.

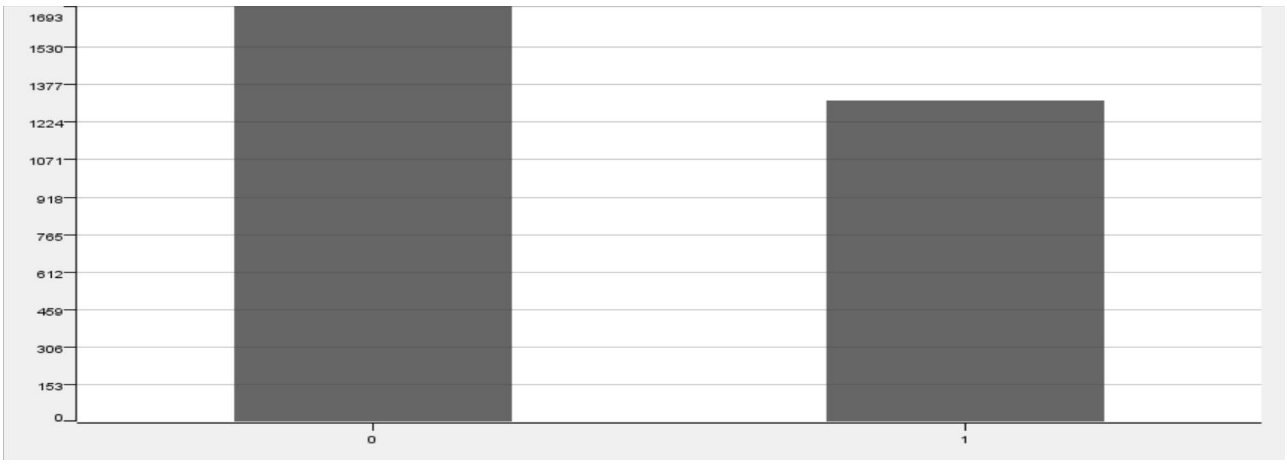


Figure 17 Frequency Graph for Qualified

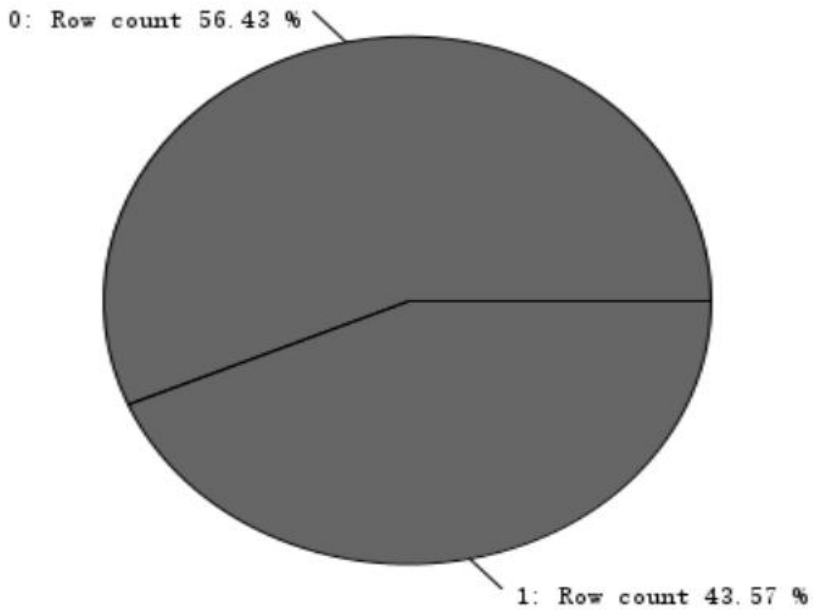
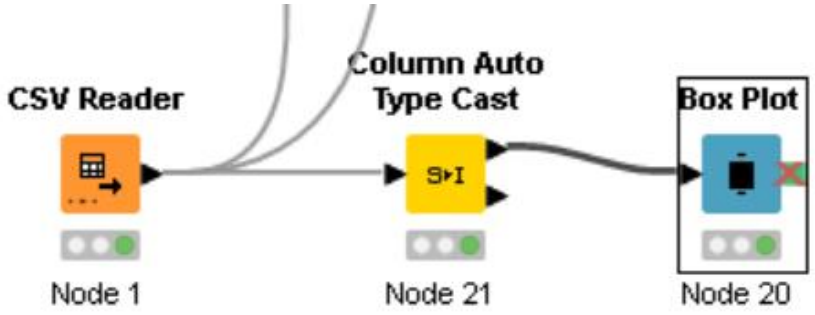
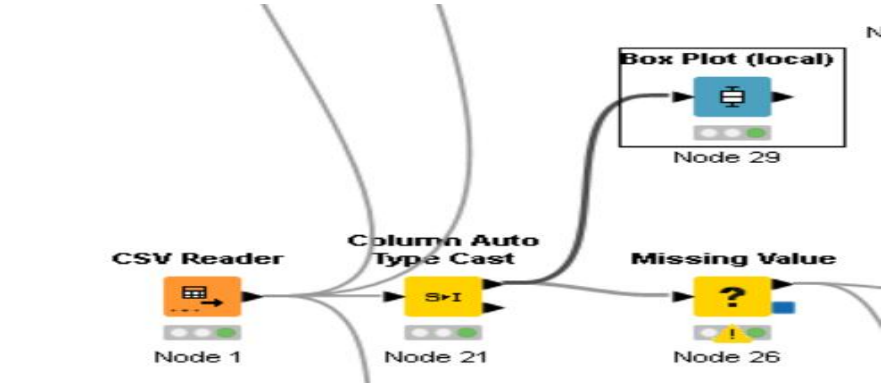


Figure 18 Pie Chart for Qualified

6. Exploration

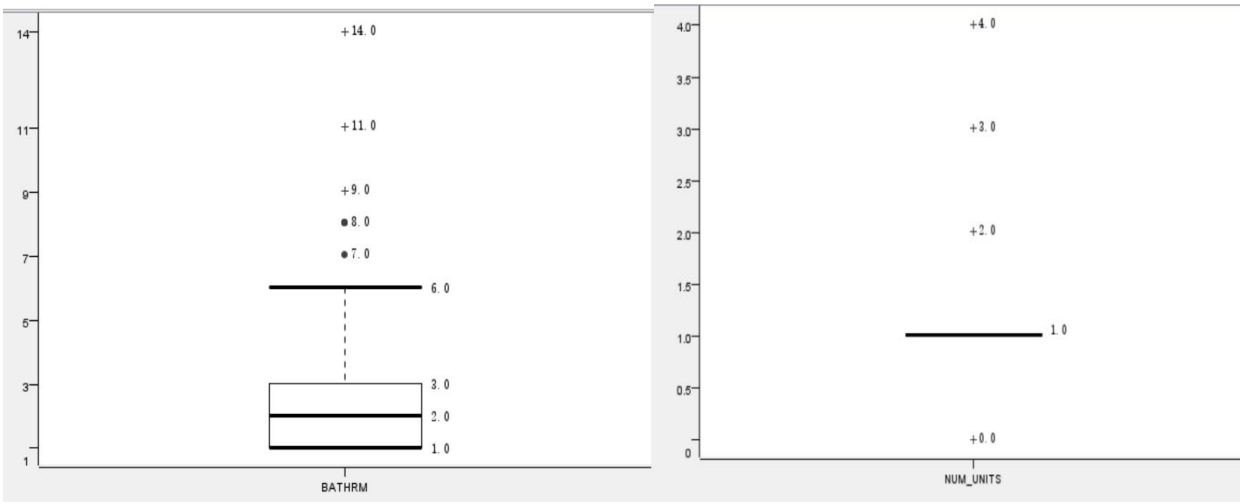
Using the tool "Box Plot" to view the correlation of different attributes. Boxplots are a commonly used data visualization method commonly used to view the distribution and outliers of data. In the chart below, we can clearly see the distribution of the data, outliers, median quartiles, and intergroup comparisons. X is thought to have multiple outliers, o is thought to have a single outlier in the data, both of which need to be observed by researchers to see if they are potential anomalies. We can see the most common types of houses and the distribution of quantities of different properties through the following charts.





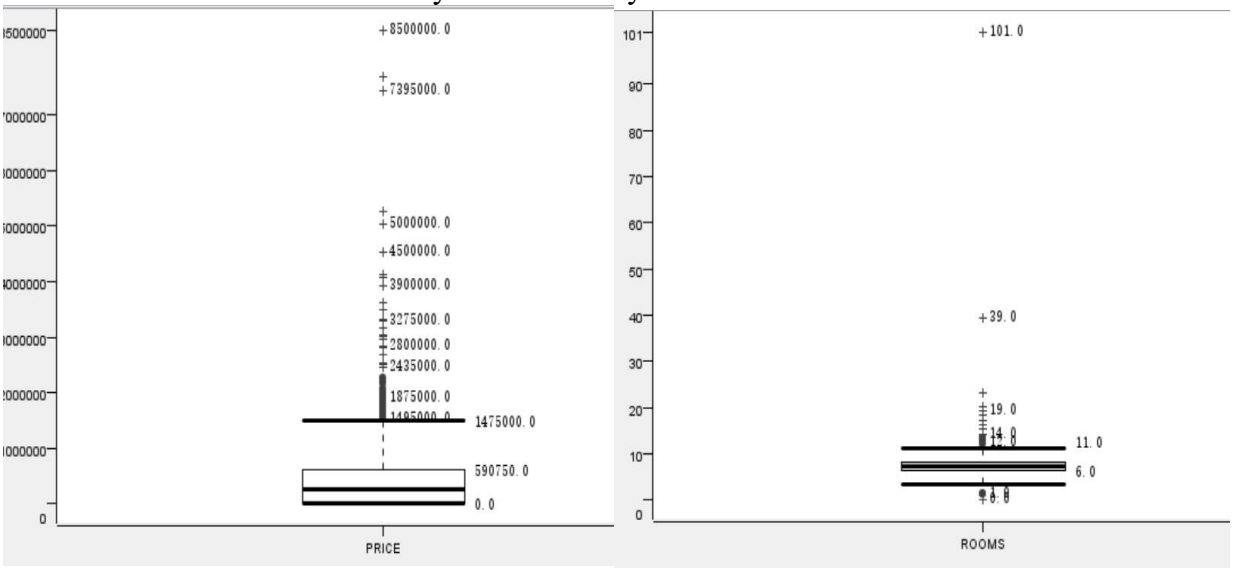
BATHRM and NUM_UNITS

From the chart below, we can clearly see that the normal is to have one to three bathrooms is the norm, and the number of bathrooms with 10 to 14 is a minority. The most commonly built house is to own a unit.



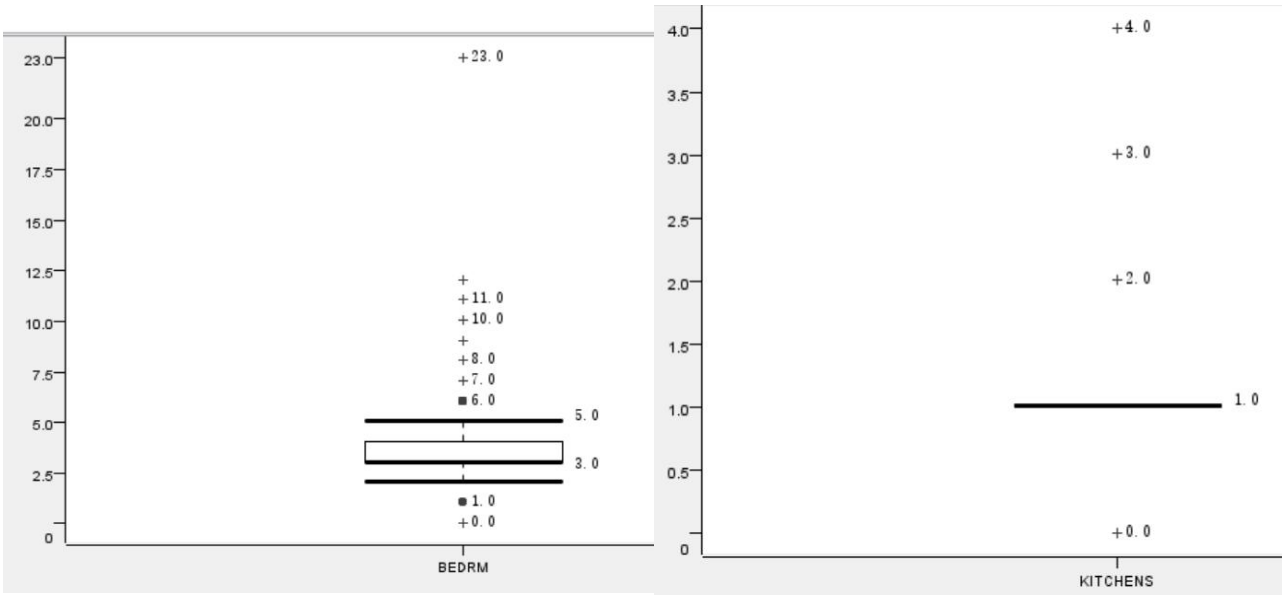
PRICE and ROOMS

The most commonly built house is to own a unit. The most commonly built house is to own 6 or 11 rooms. One of the houses that may be an anomaly has 101 rooms.



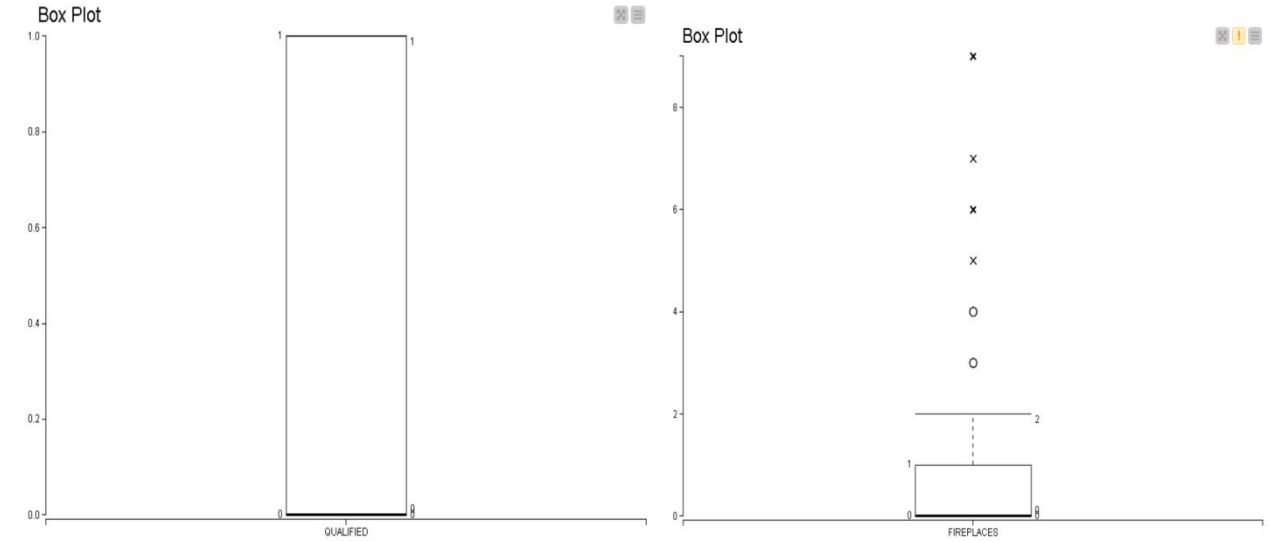
BEDRM and Kitchens

The amount of room bedrooms is usually distributed between three and five, and the number of kitchens is usually one.



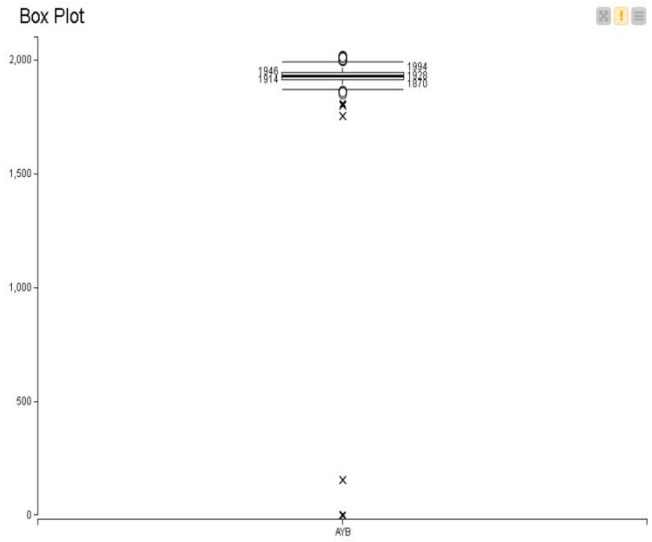
Qualified and FIREPLACES

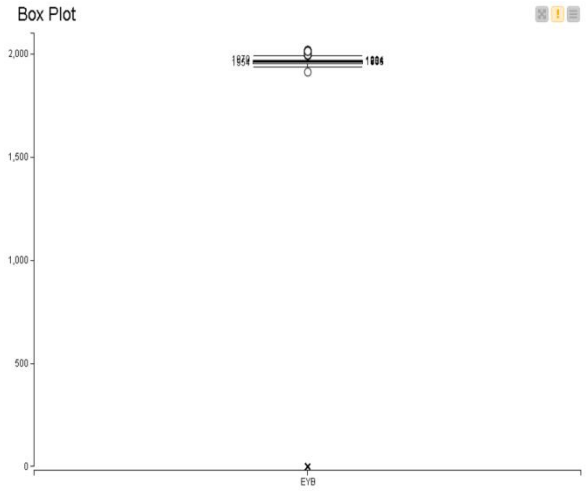
Houses usually have one fireplace and a small number have two.



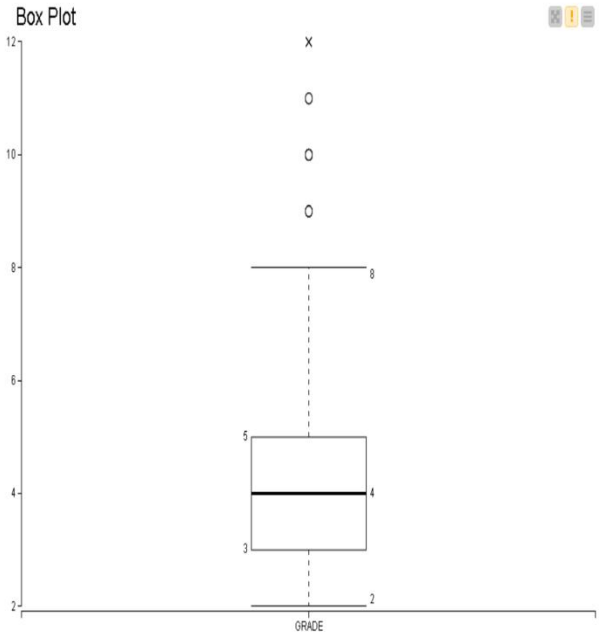
AYB and EYB

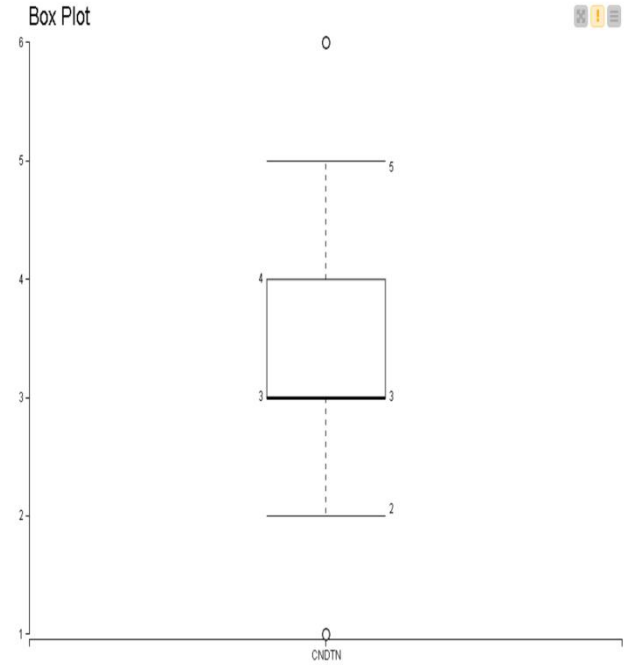
AYB and EYB are usually concentrated around 1870-2000 years.



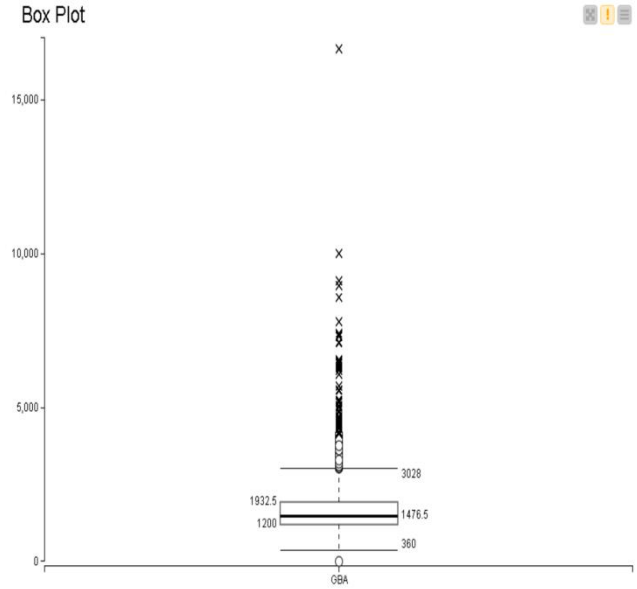


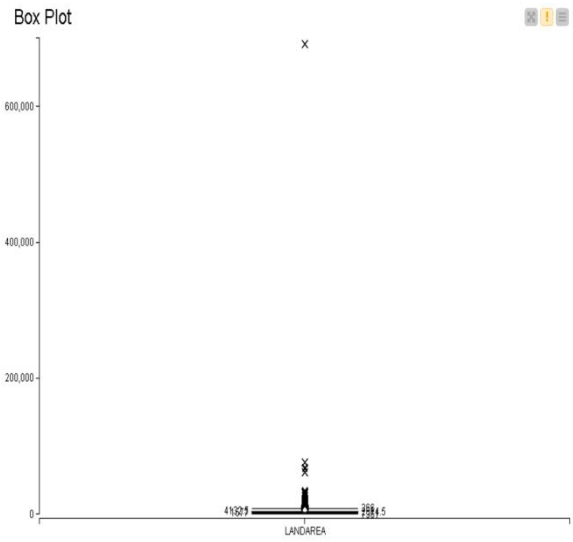
GRADE and CNDTN





GBA and LANDAREA

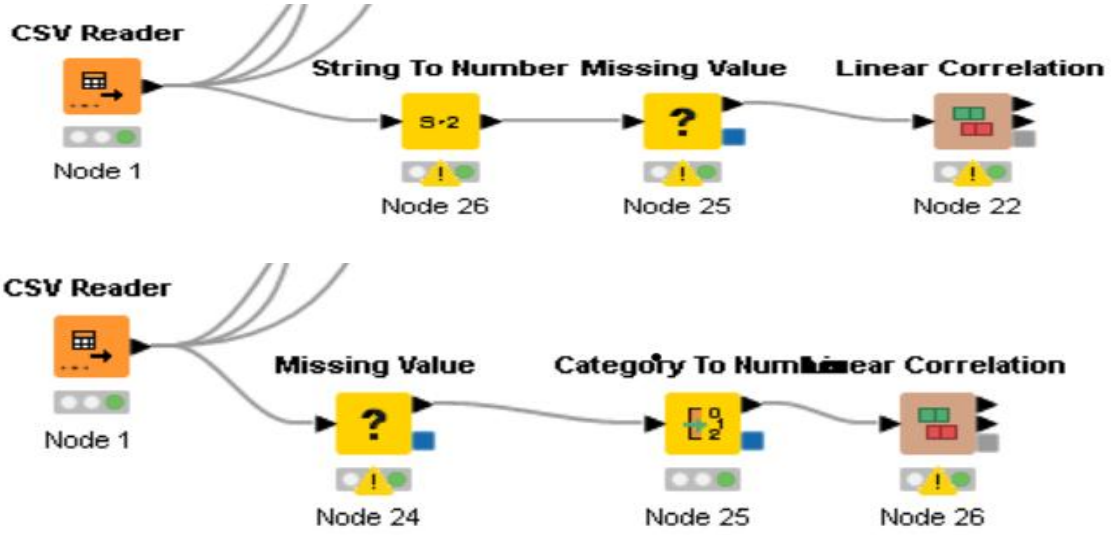




7. Identify the correlations

7.1 Using "Linear Correlation" and “Rank Correlation”

Using the tool "Linear Correlation" and “Rank Correlation” to view the correlation of different attributes. As shown in the figure below, blue represents high relevance, and red represents low relevance. The high linear relationship between NUM_UNITS and Kitchen can be clearly seen from Figures 19. The high linear relationship between KITCHENS, USERCODE and NUM_UNIT can be clearly seen from Figures 19. KITCHENS and FIRPLACE have a high correlation in figure 19.



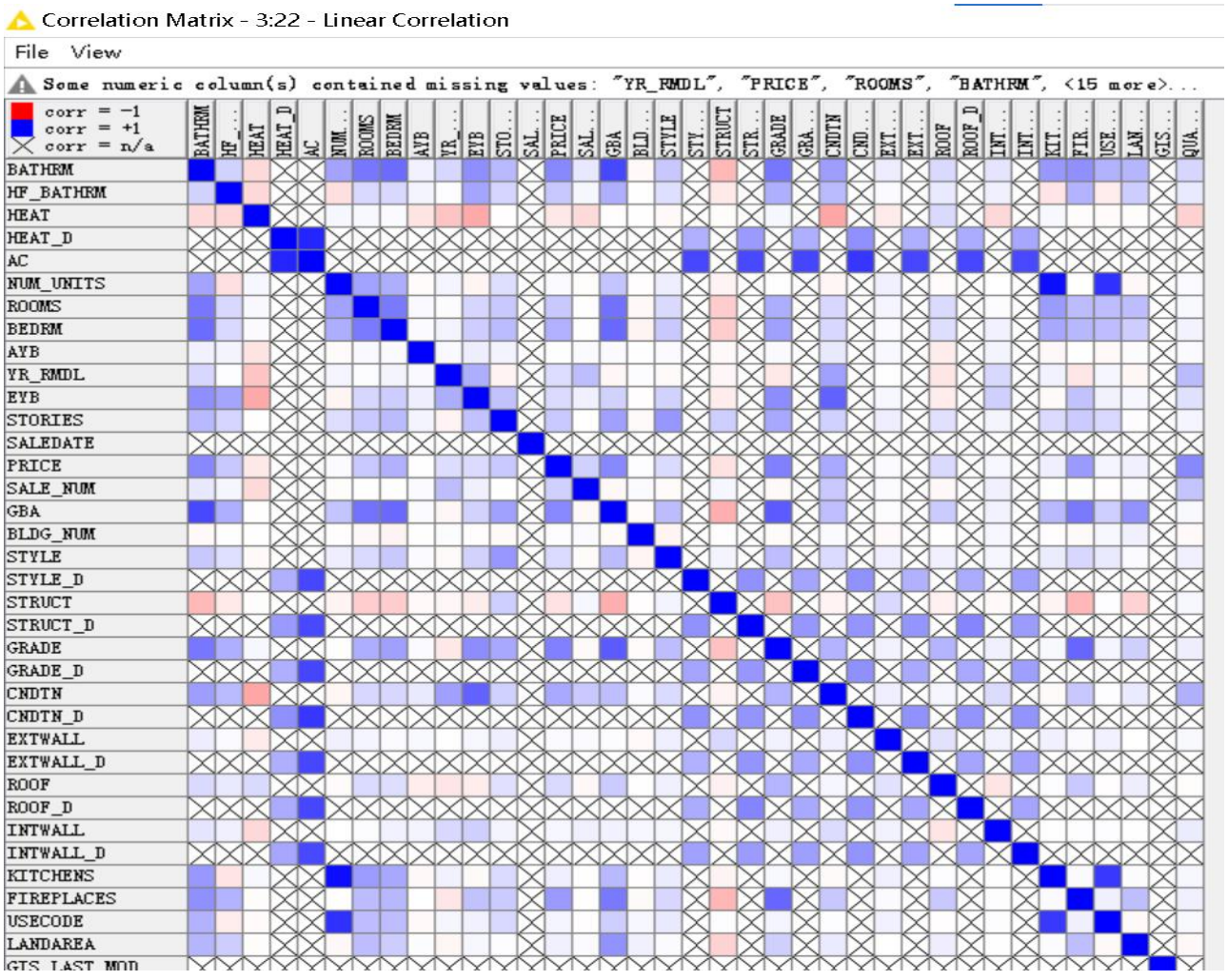


Figure 19 Linear Correlation graph

In Figure 20, we clearly can see STORIES and STYPLE have a strong correlation. NUM_UNITS and Kitchen have a high relationship. BAthrm and ROOMS Have a high relationship.

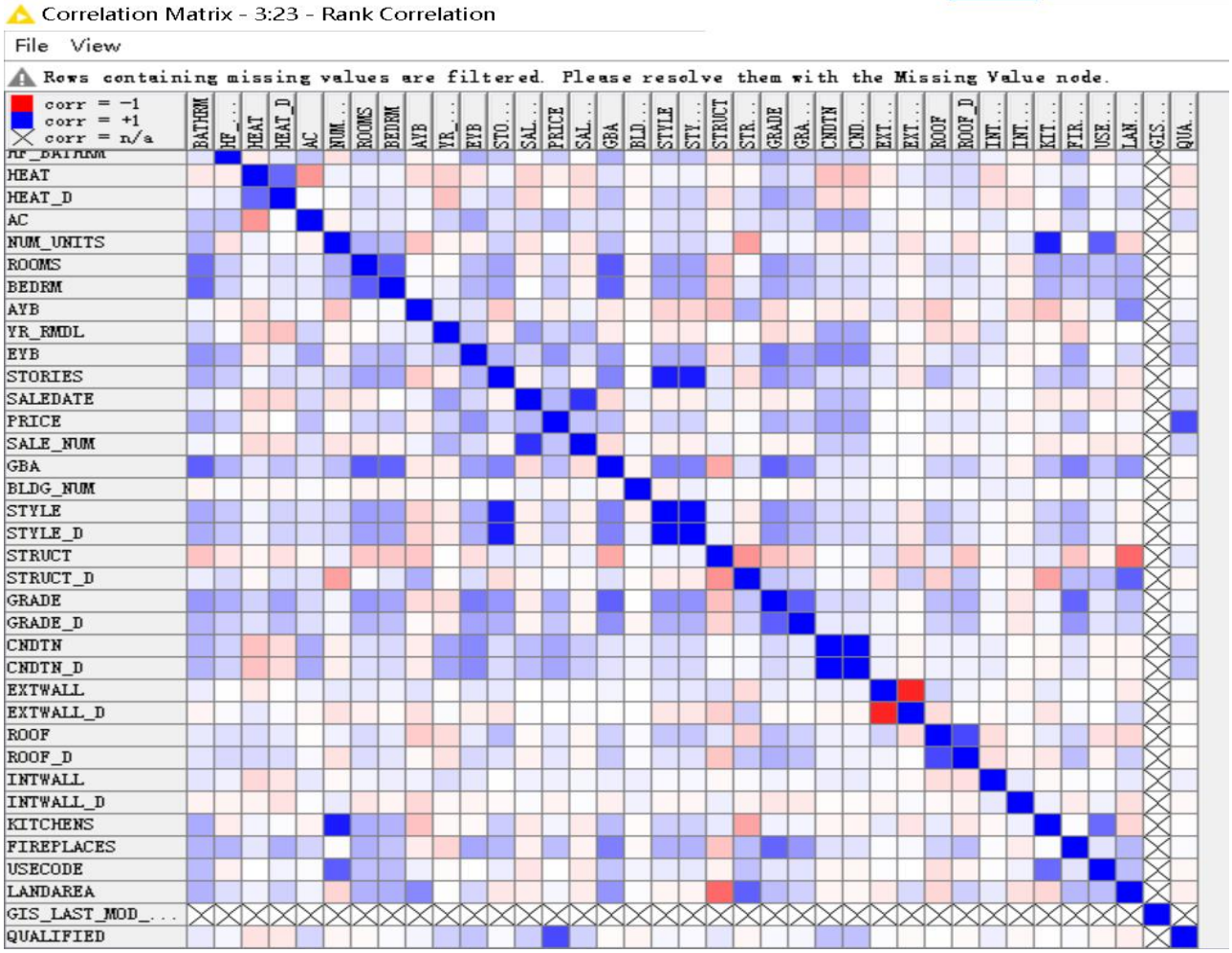
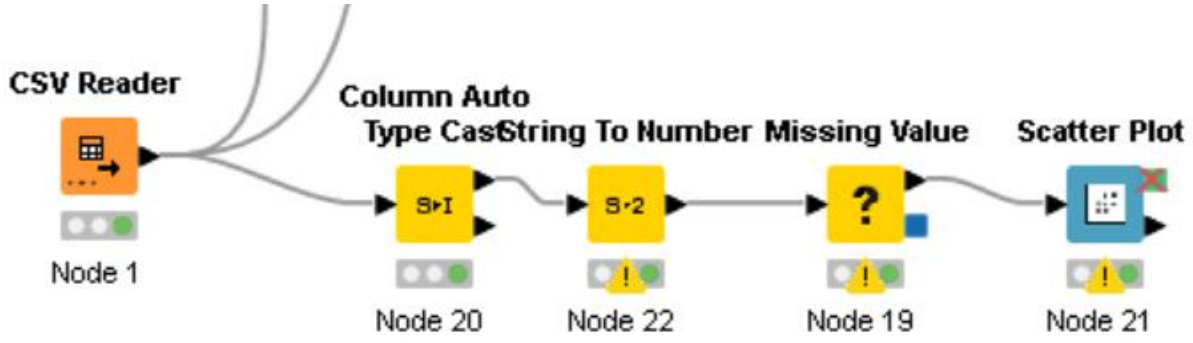


Figure 20 Rank Correlation graph

7.2 Using"Scatter Plot"

Using the tool "Scatter Plot" and showing the relationship between two quantitative variables. We can determine whether there is a positive correlation, a negative correlation, or no relationship by observing scatter plot.



We can see from the figure (figure 21) below that the interesting phenomenon is that as the value of the **style** increases, the value of the **story** increases.

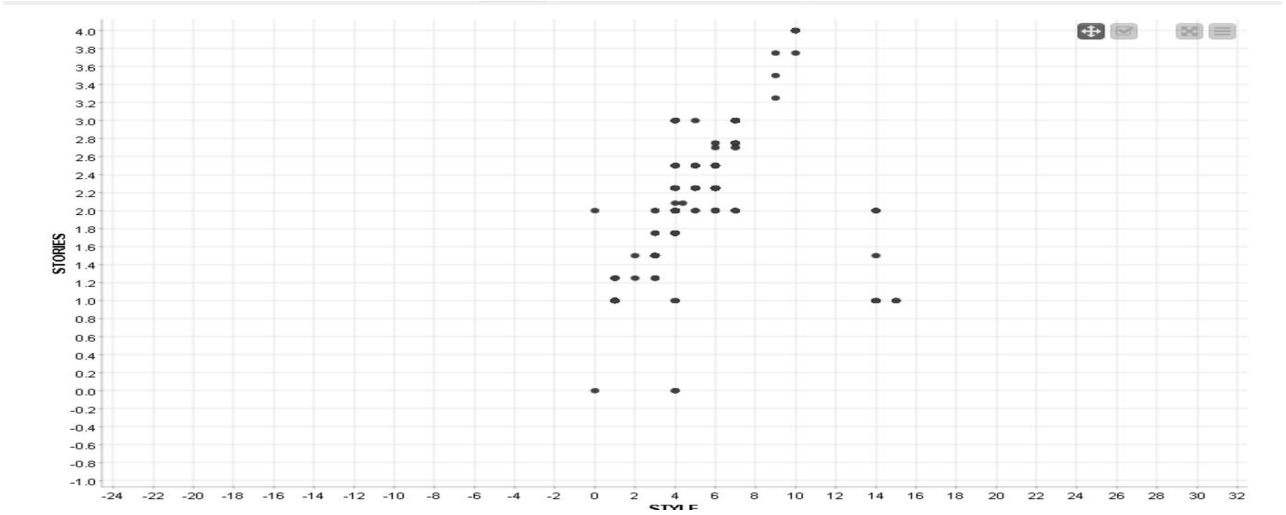
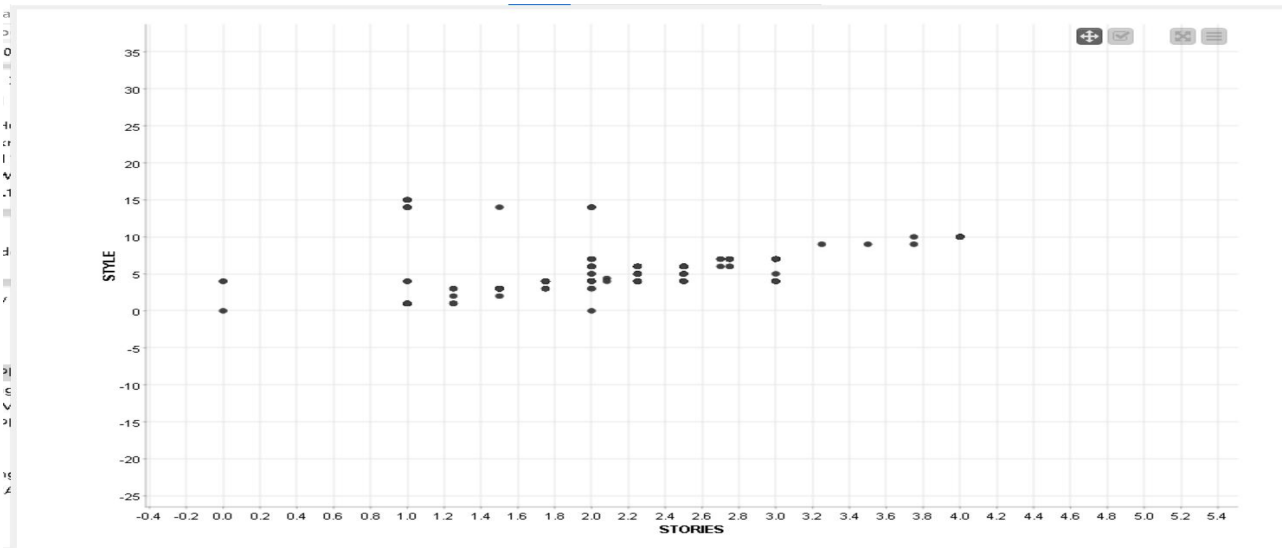


Figure 21 Scatter Plot-- Style VS Stories



This diagram(figure 21) shows that the more **bathroom**s a house has, the more **rooms** it is. This is in line with the normal situation in our lives, since more rooms mean more people who may live together, and more bathrooms are needed to provide a better and more comfortable environment.

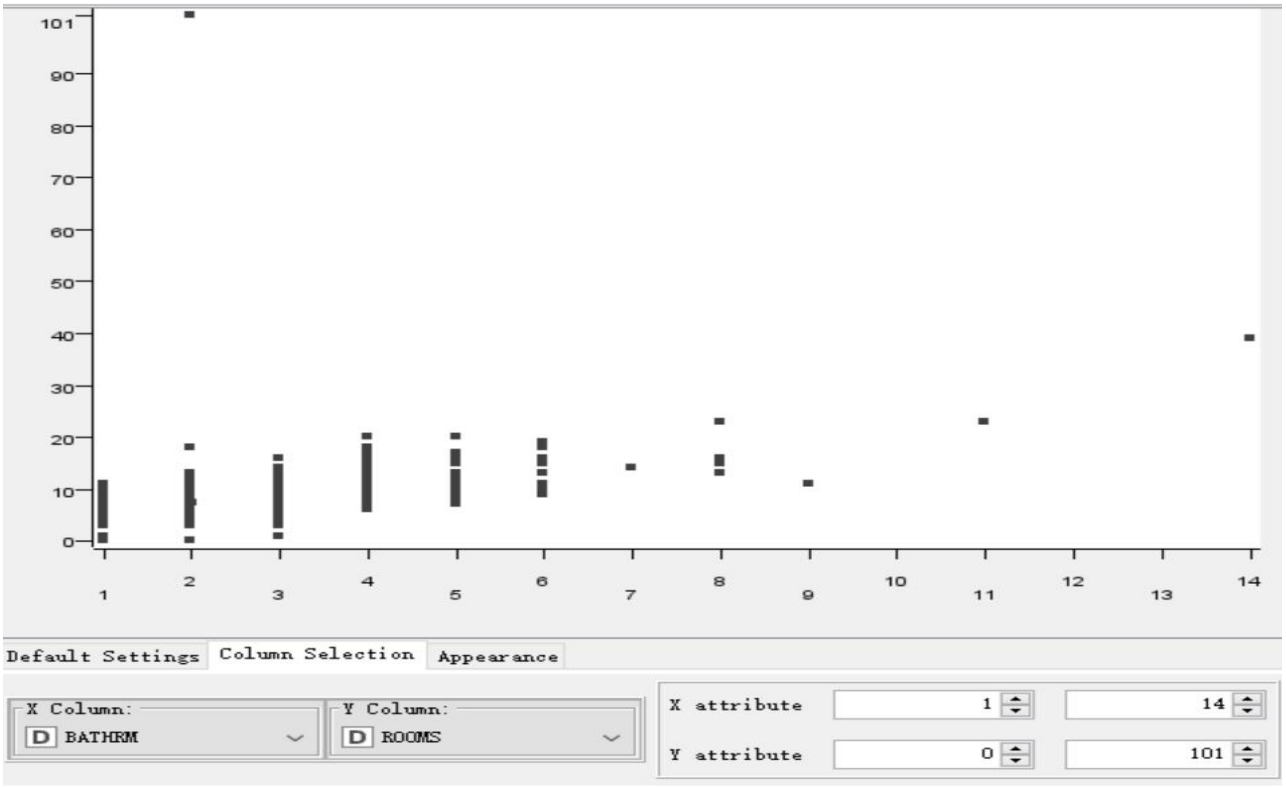


Figure 22 Scatter Plot-- Bathrm VS ROOMs

From the following figure (figure 23), we can see that the interesting phenomenon is that if you get a higher **grade**, the higher the chance of getting a high **price**.

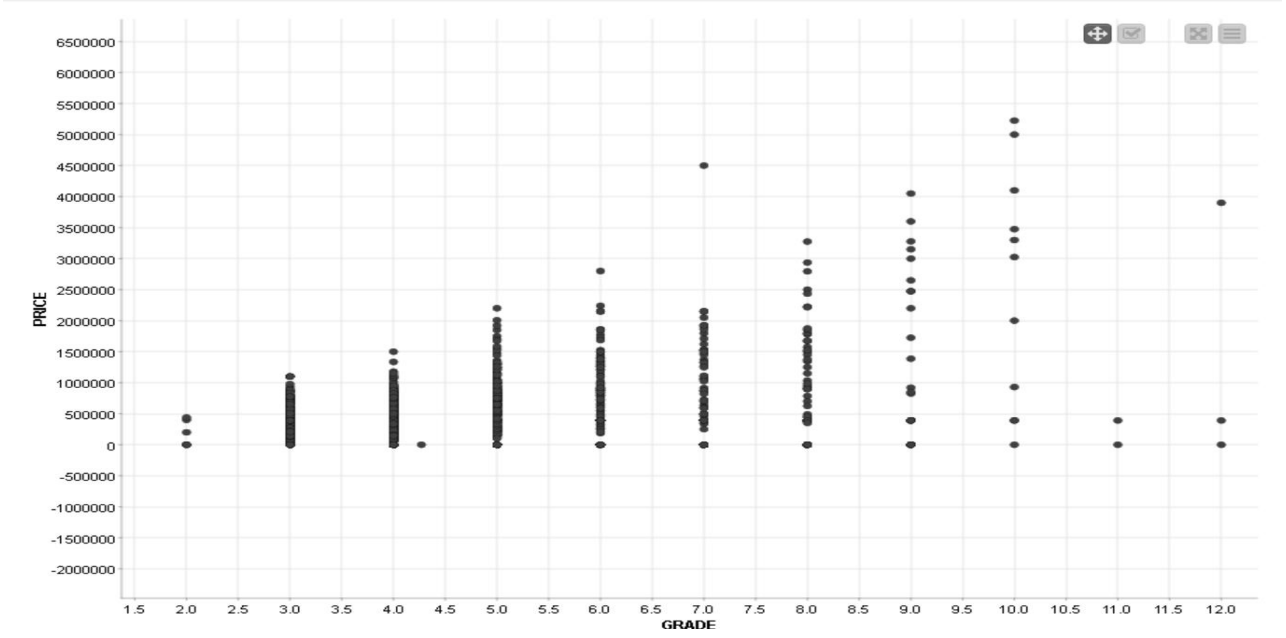


Figure 23 Scatter Plot-- GRADE VS Price

Through the following figure (figure 24), we can find that the number of more **units** will be higher, but the higher the **price** may be when the unit is one. The illustrative unit is not the reason why people decide how high or low the price is.

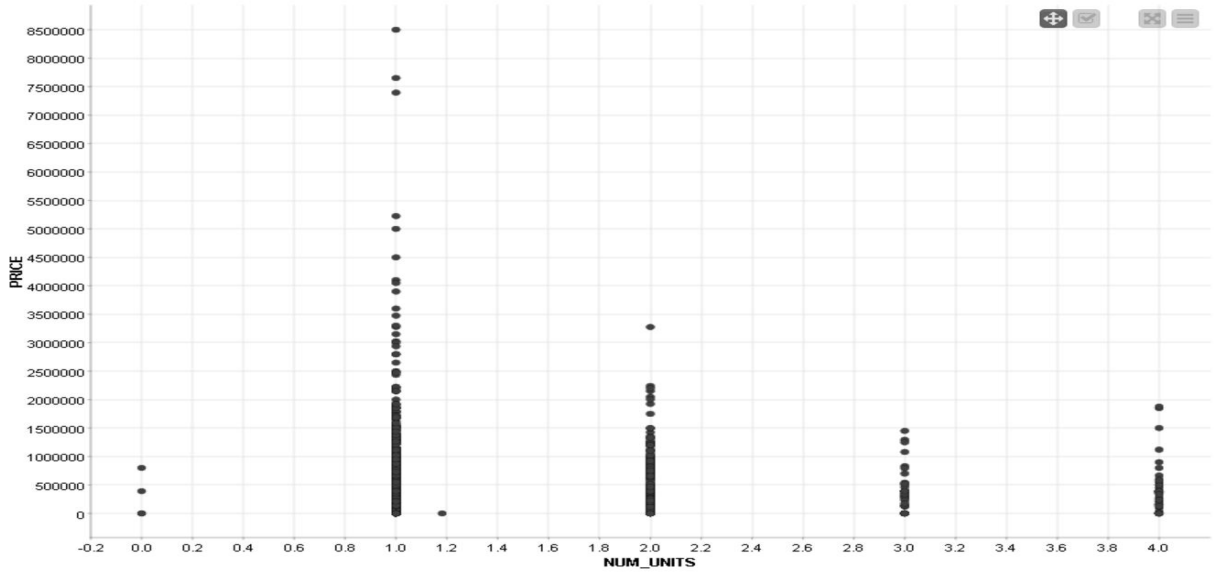


Figure 24 Scatter Plot-- NUM_UNITS VS Price

Through the following figure (figure 25), the EYB's time is mainly concentrated after 1940, which means that the vast majority of buildings were not remodeled until after 1940. After 1940, part of the building **EYB** and **A Y B** are of the same year, which means that the part of the earliest construction of the building is the same as the remodeled part. After 1950, the later the AYB year, the higher the probability that the EYB year is later.

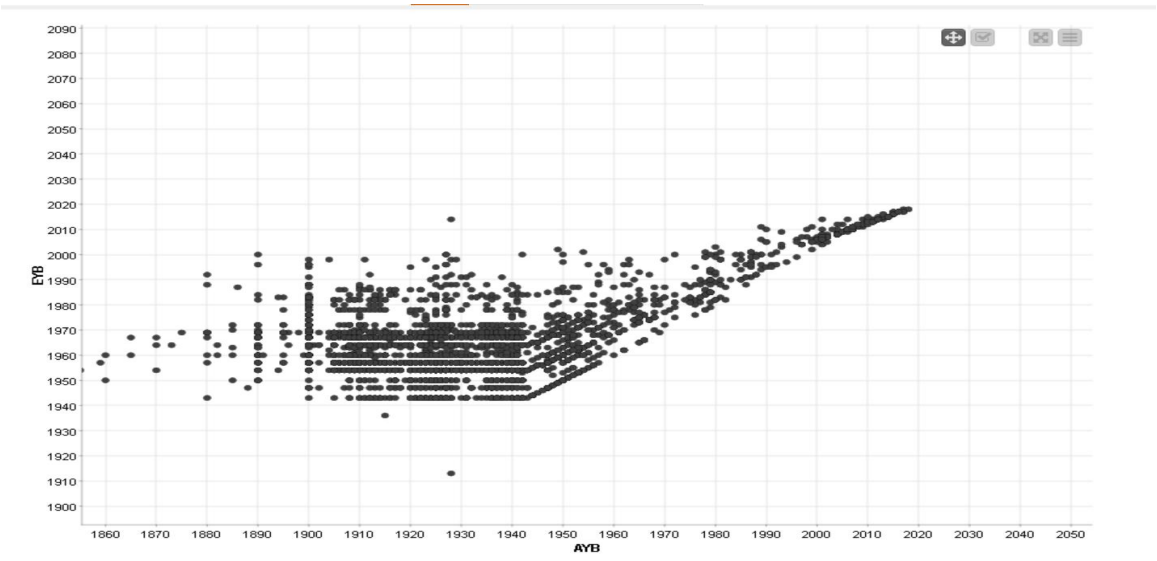


Figure 25 Scatter Plot-- AYB VS EYB

Through the following figure (figure 26), there is a linear relationship between **rooms** and GBA, and the **GBA** area will become larger as the number of rooms increases. This means that the greater the number of rooms, the larger the total floor area.

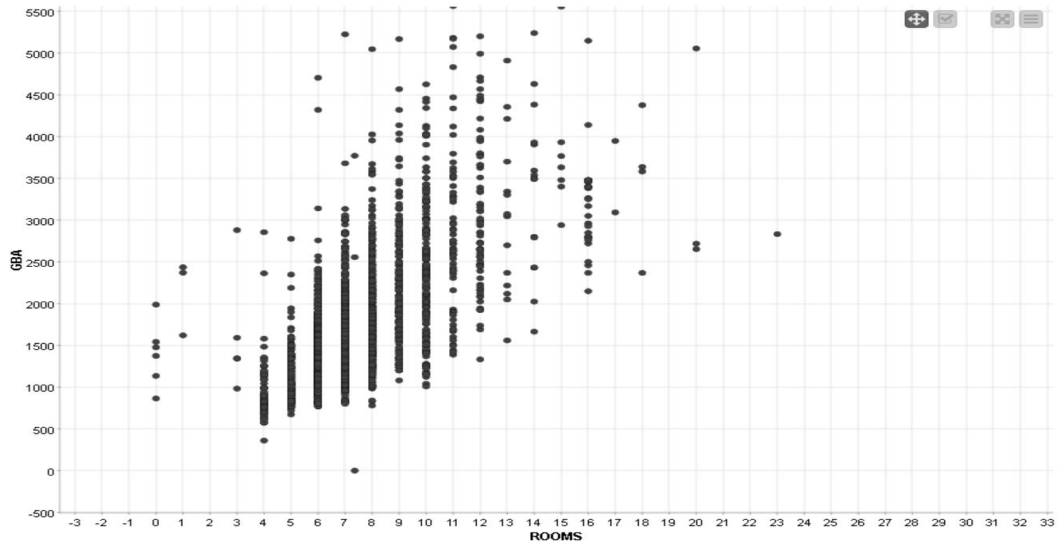
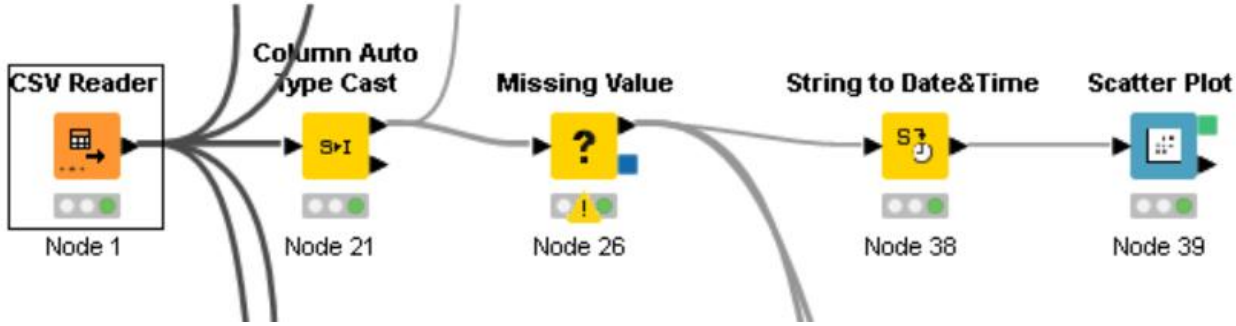


Figure 26 Scatter Plot-- GBA VS Rooms



As you can see from the figure 27 below, the later the **sale date**, the higher the **price**, which is in line with the concept that the newer the house is more valuable in real life

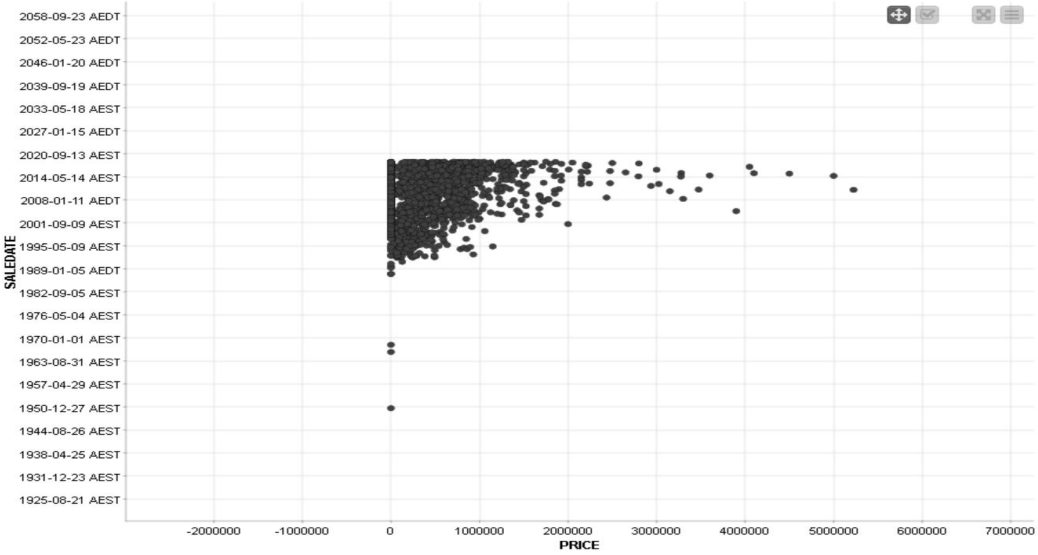


Figure 27 Scatter Plot-- SALEDATE VS Price

From Figure 28, it can be seen that the **price** has a certain relationship with **GBA**, and when **GBA** is larger, the higher the probability of price.

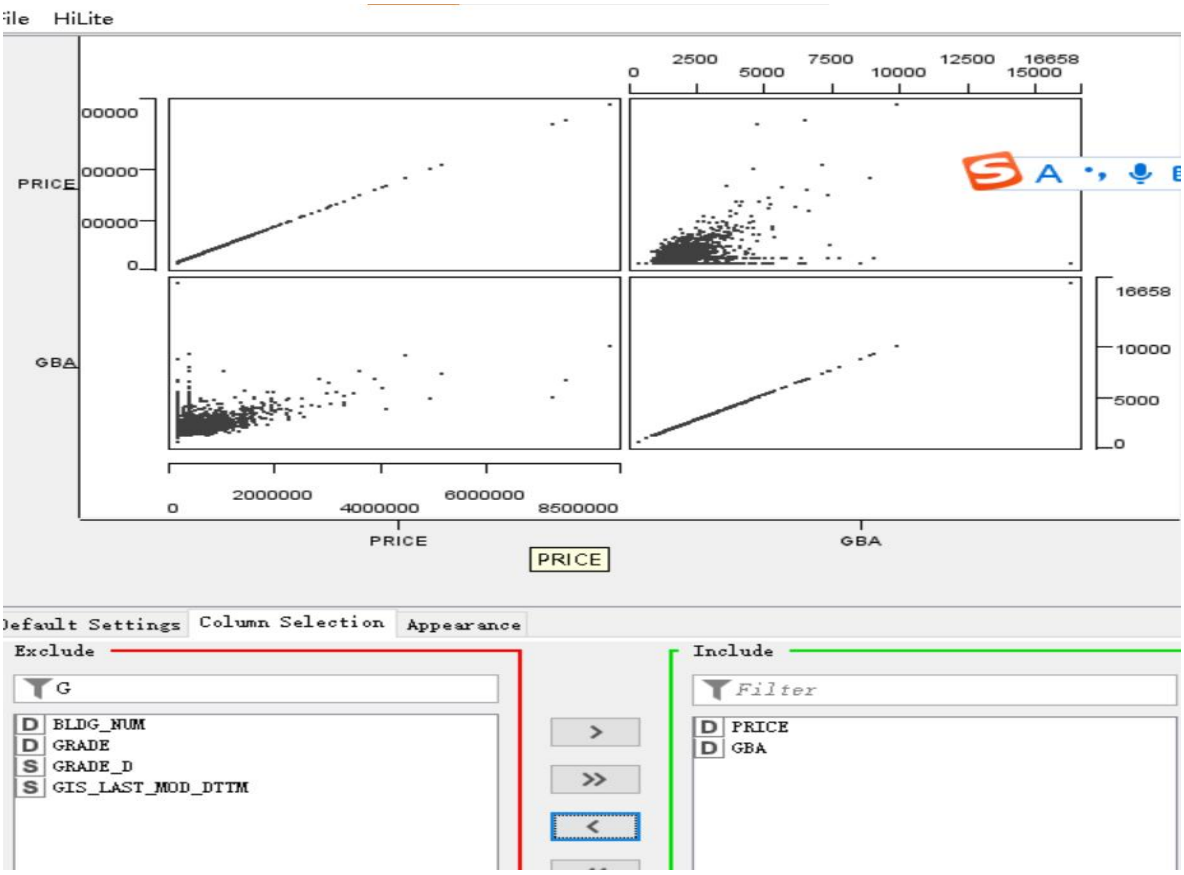
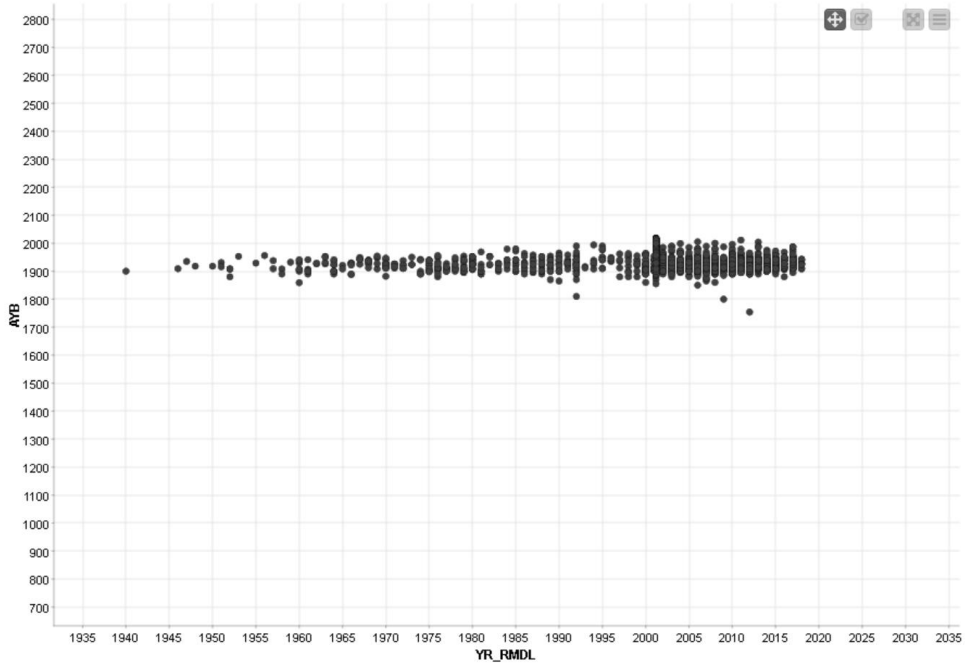
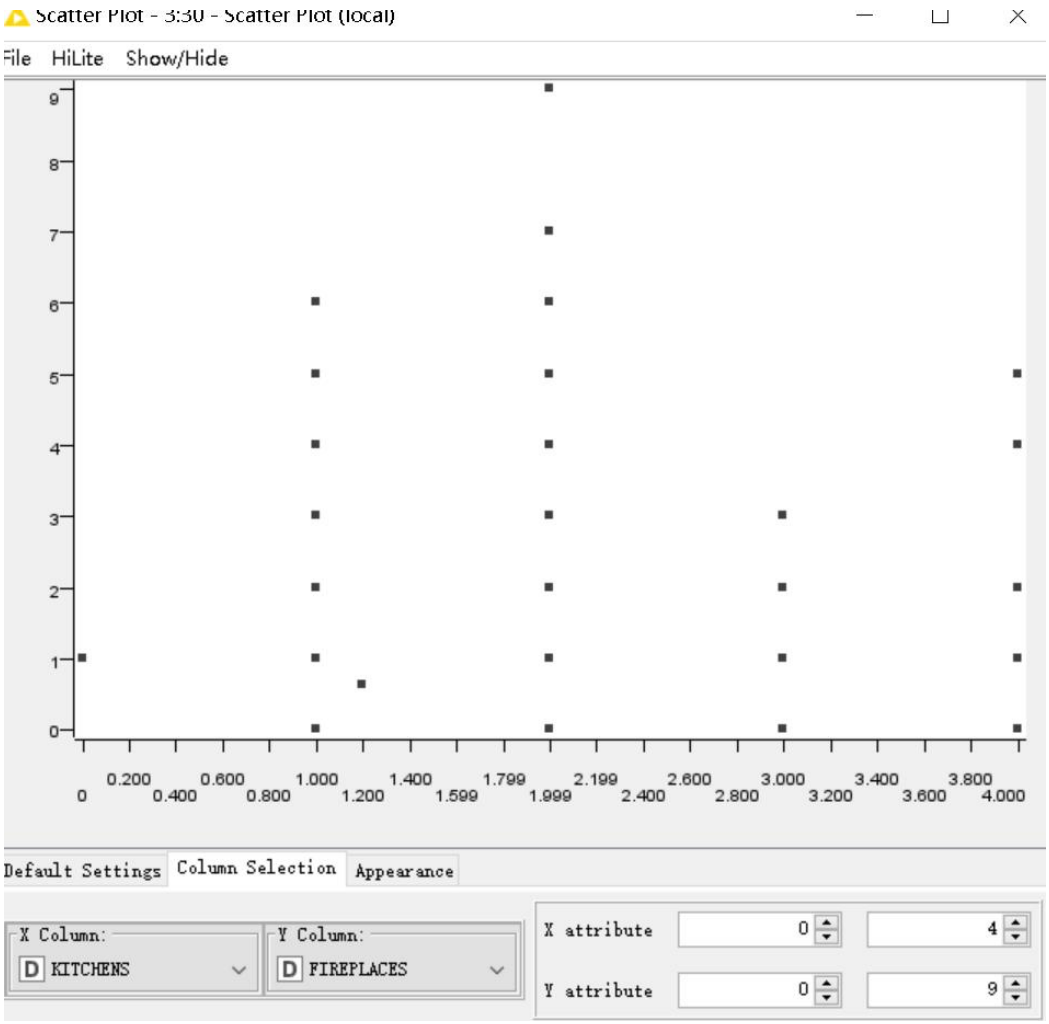


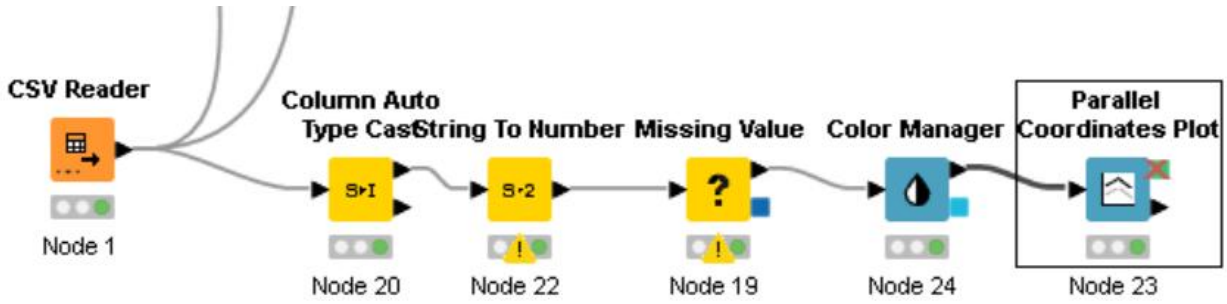
Figure 28 Scatter Plot-- GBA VS Price

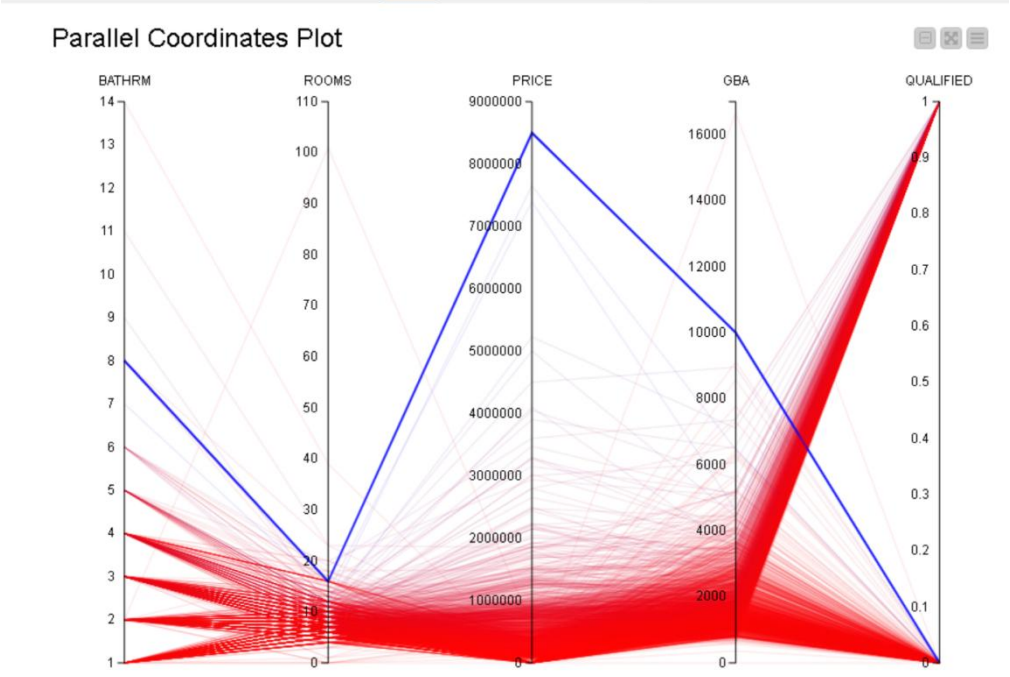




7.3 Using “Parallel coordinate”

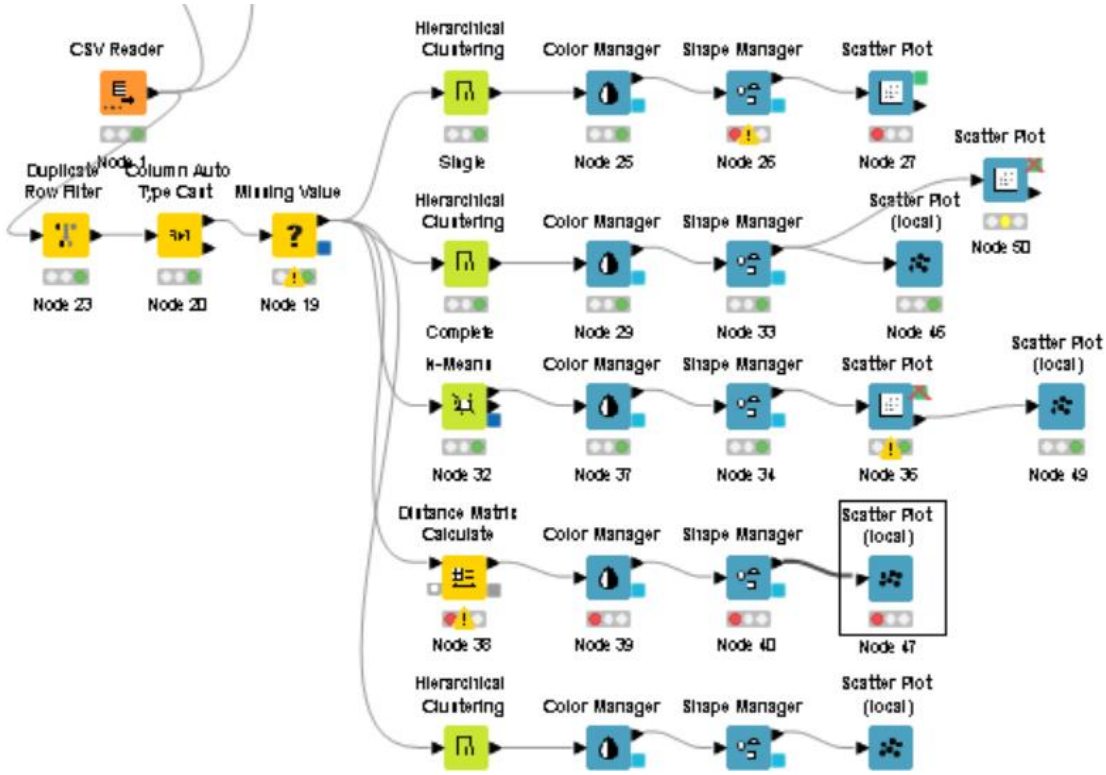
Parallel coordinate graphs can help us visually observe the relationship and trend between different numerical variables. “Color manage” is used to color the maximum and minimum values blue.





8. Explore dataset and identify any outliers, clusters of similar instances

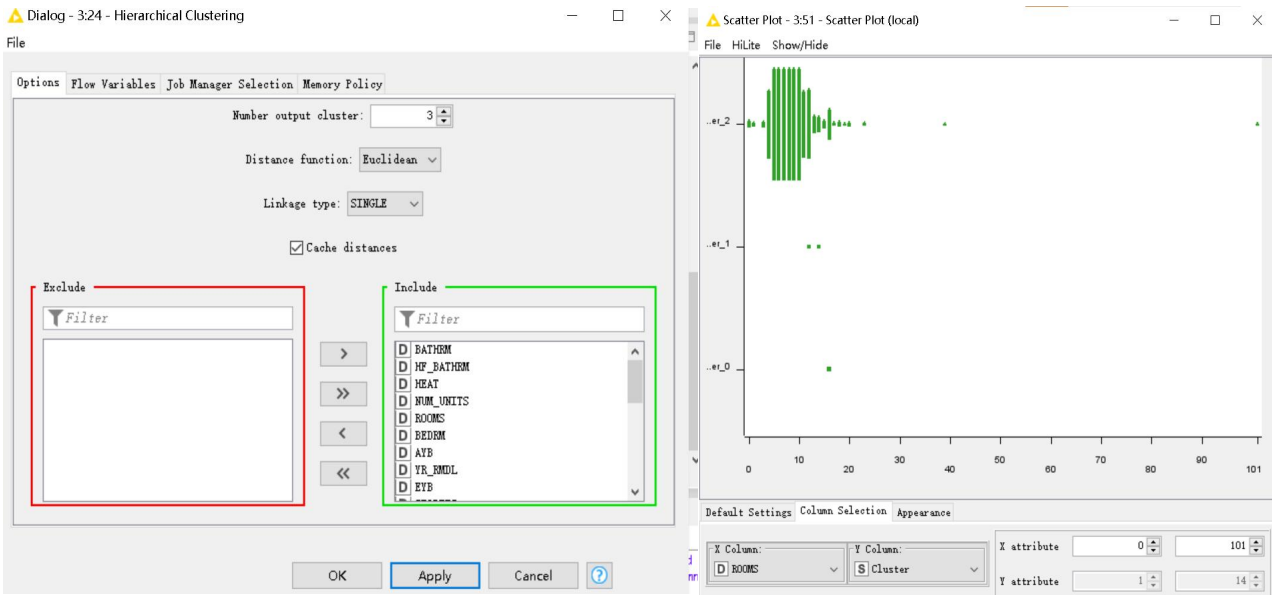
It is a clustering algorithm used to group data and objects, and different connection methods can be selected by adjusting the linkage type to calculate the distance between clusters.

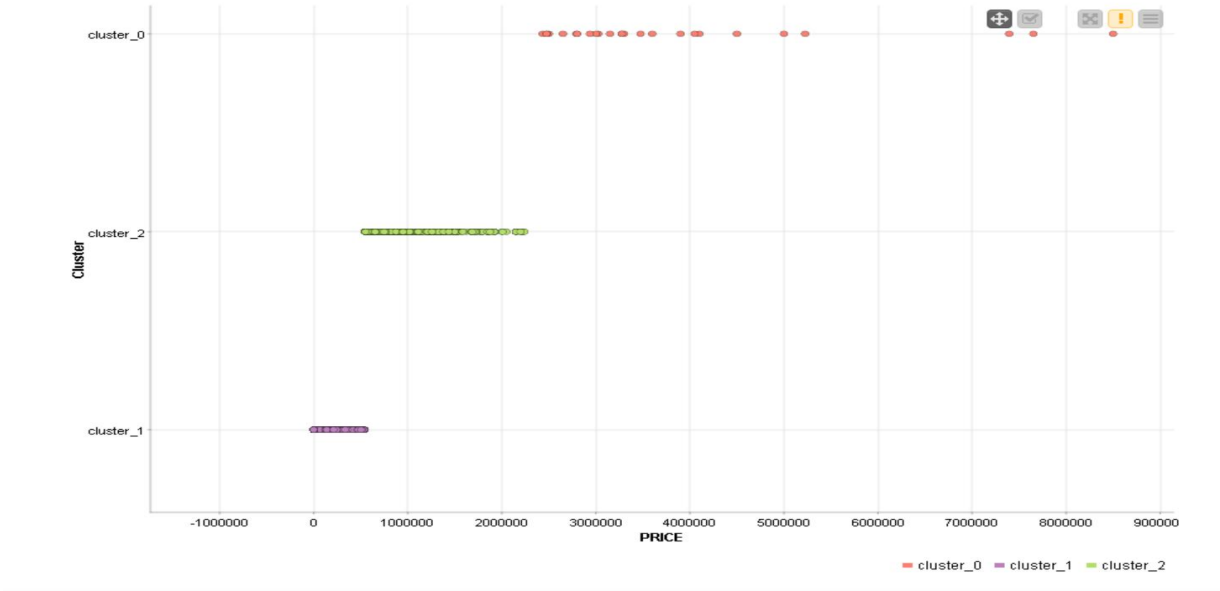


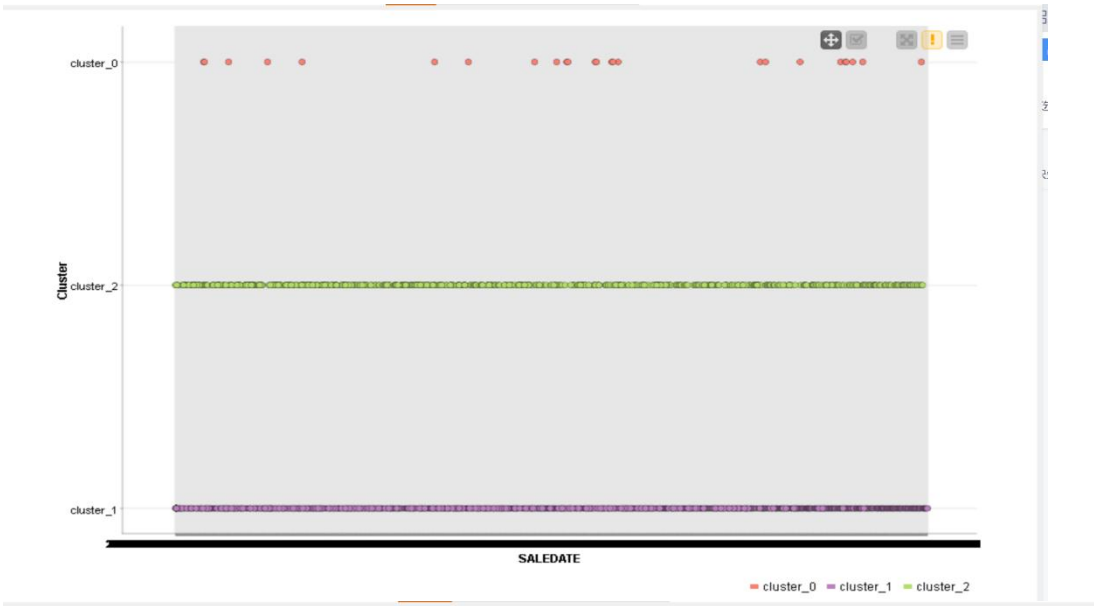
8.1 Hierarchical Clustering (Single)

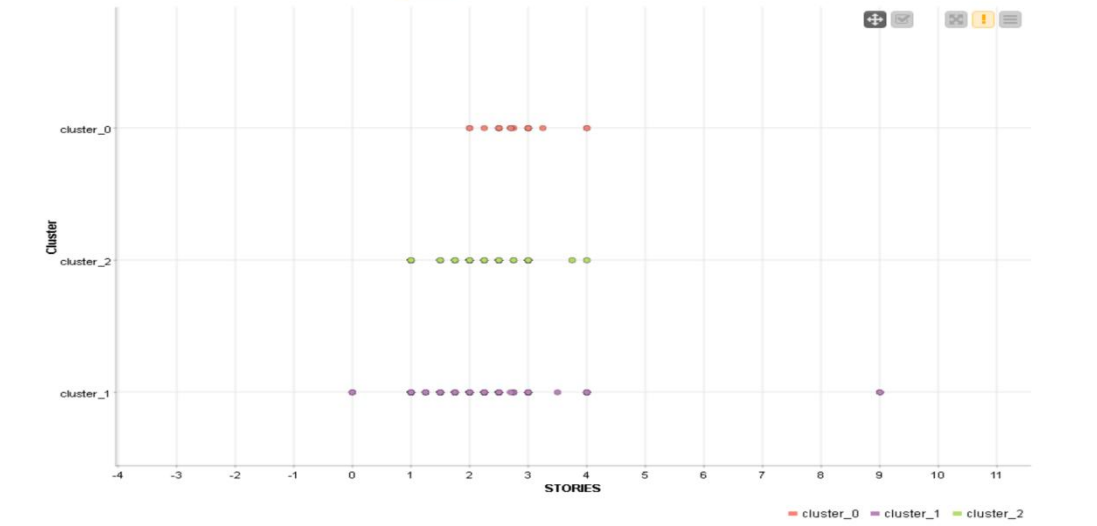
Setting up 3 output clusters, with the distance function Euclidean and Linkage type “STNPLE”. It is the distance between the two closest data points in two combined data as the distance between the two combined data points, which is susceptible to extreme values.

Assignment 2 Yongyan Liu 14214338



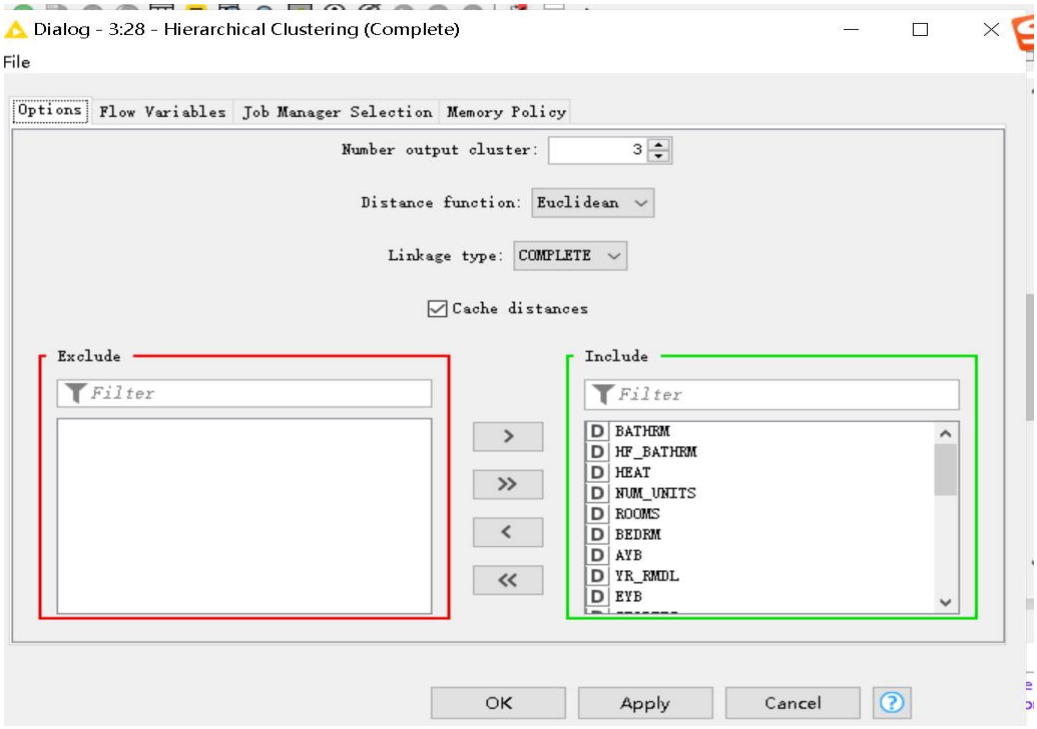


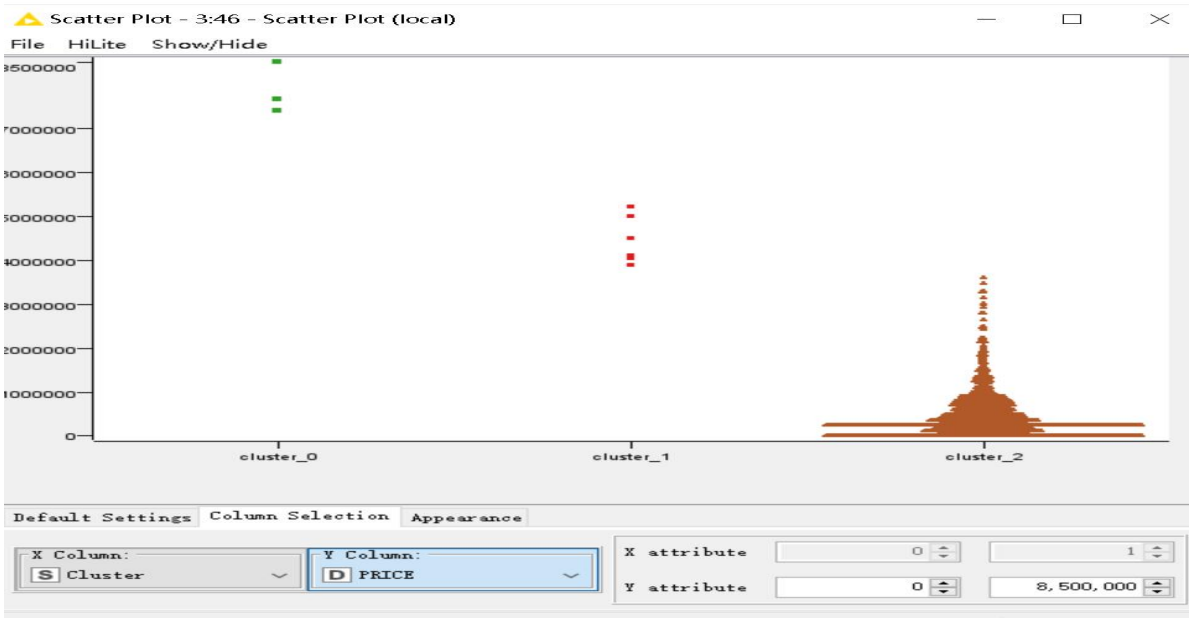




8.2 Hierarchical Clustering (COMPLETE)

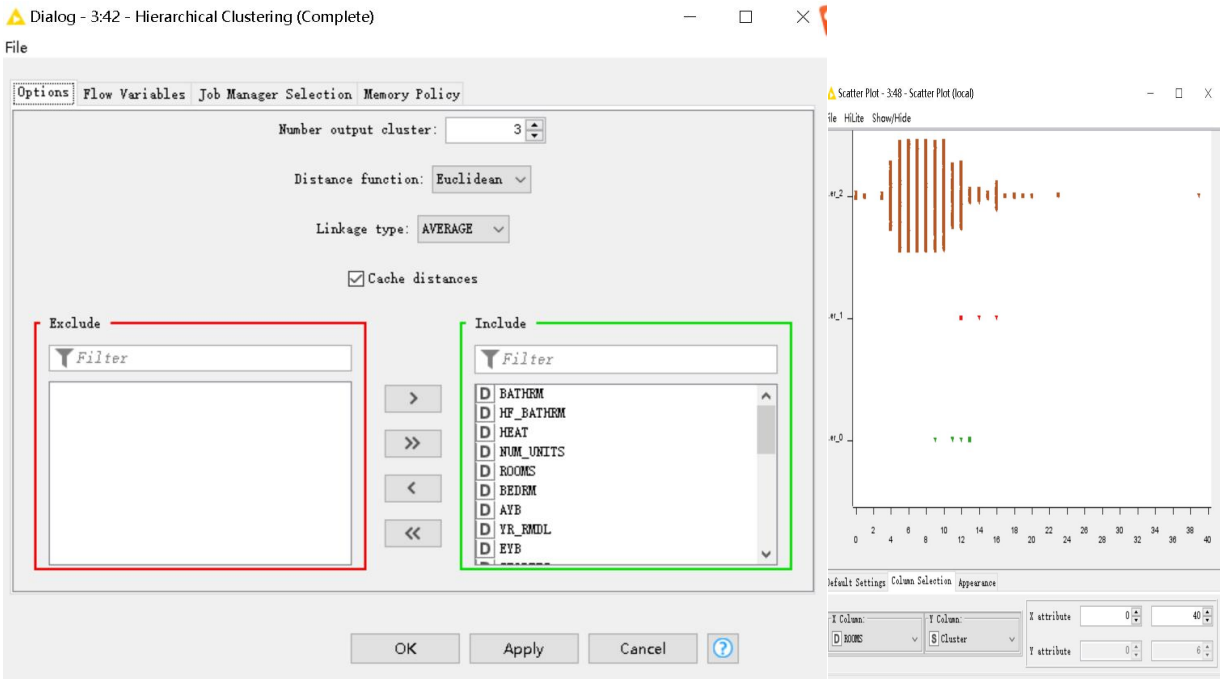
Setting up 3 output clusters, the distance function “Euclidean” and Linkage type “COMPLETE”. “Complete” means that the distance of its masses is determined by the two farthest points in a group, and the advantage of this algorithm is that it can identify more scattered groups, but it is susceptible to noise and outlier images. It shows that there is only few data point in both cluster_0 and cluster_1. Therefore, this data is not suitable for completed structured clustering.

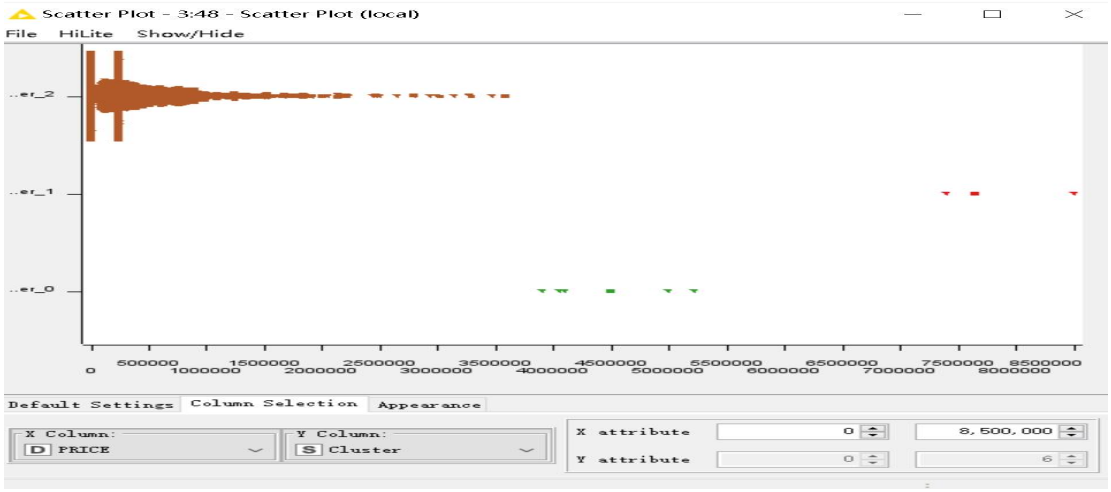

Dialog - 3:28 - Hierarchical Clustering (Complete)
File
Options
Flow Variables
Job Manager Selection
Memory Policy
Number output cluster:
3
Distance function:
Euclidean
Linkage type:
COMPLETE
Cache distances
Exclude
Filter
Include
Filter
BATHRM
HF_BATHRM
HEAT
NUM_UNITS
ROOMS
BEDRM
AYB
YR_RMDL
EYB
OK
Apply
Cancel


Scatter Plot - 3:46 - Scatter Plot (local)
File
HiLite
Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings
Column Selection
Appearance
X Column:
Cluster
Y Column:
PRICE
X attribute
Y attribute
8,500,000

8. 3 Hierarchical Clustering (Average)

Setting up 3 output clusters, the distance function “Euclidean” and Linkage type “Average”. Average is the distance in the cluster is determined by the average of the distances between points in the cluster. When two clusters merge, the distance between them is averaged by the distance between all points, and this algorithm is good for eliminating anomalies, but may ignore more scattered clusters.

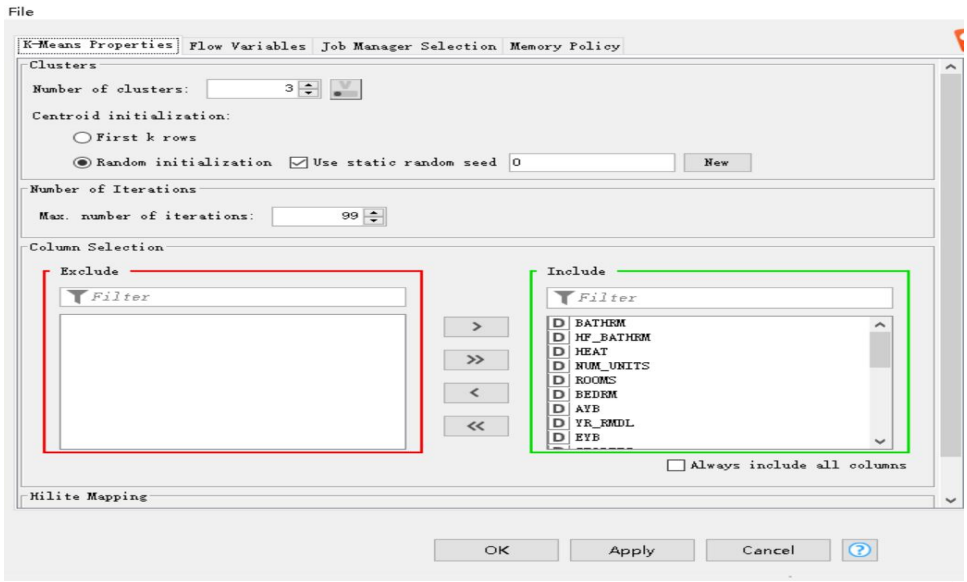


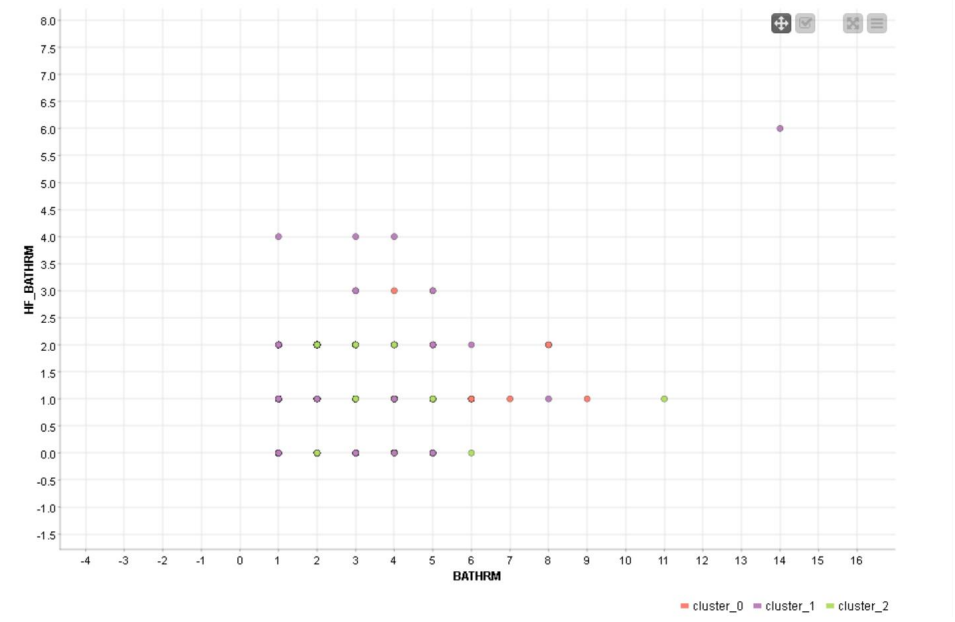


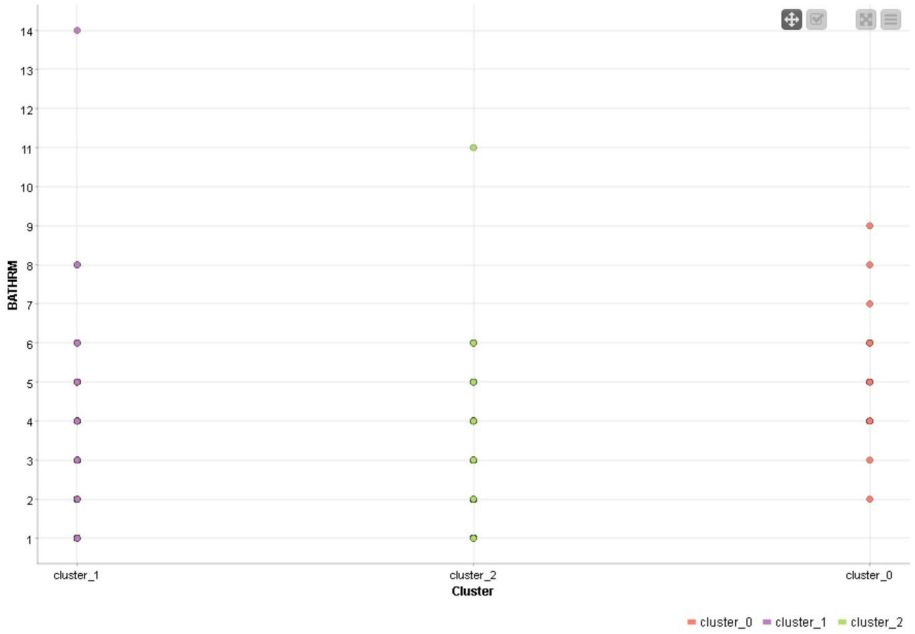
8. 4 K-Means

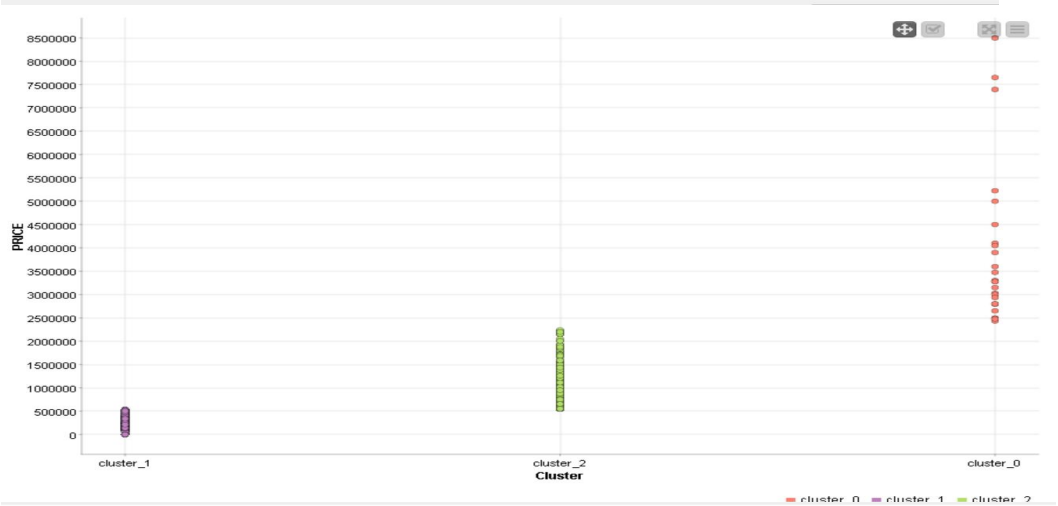
The goal of the K-mean algorithm is to divide a set of data points into K cluster classes, so that the sum of squared errors of cluster classes is minimized. This method requires random selection of K points as initial particles, so it is easy to fall into local optimization.

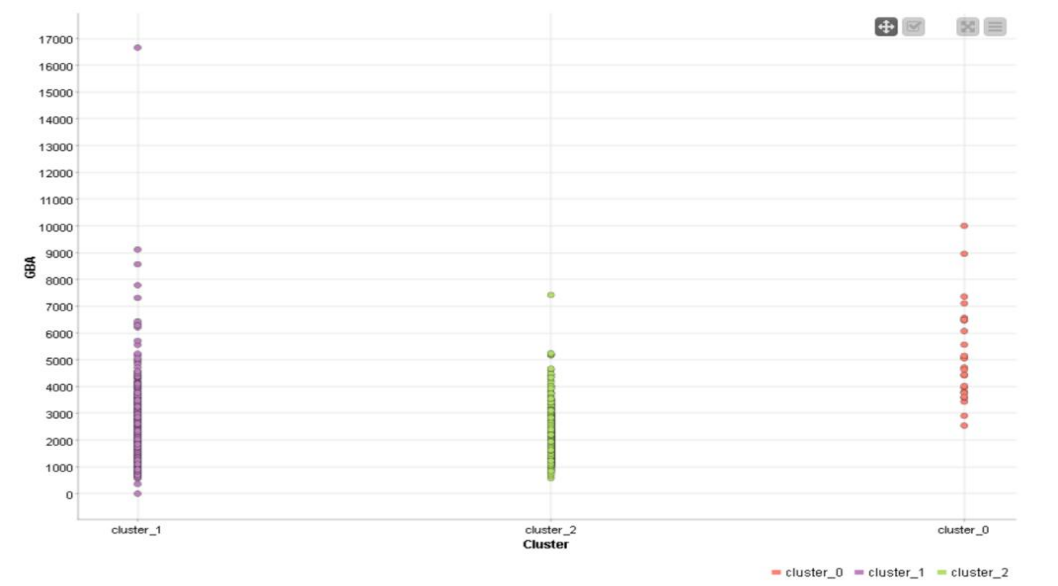
Setting up 3 output clusters.

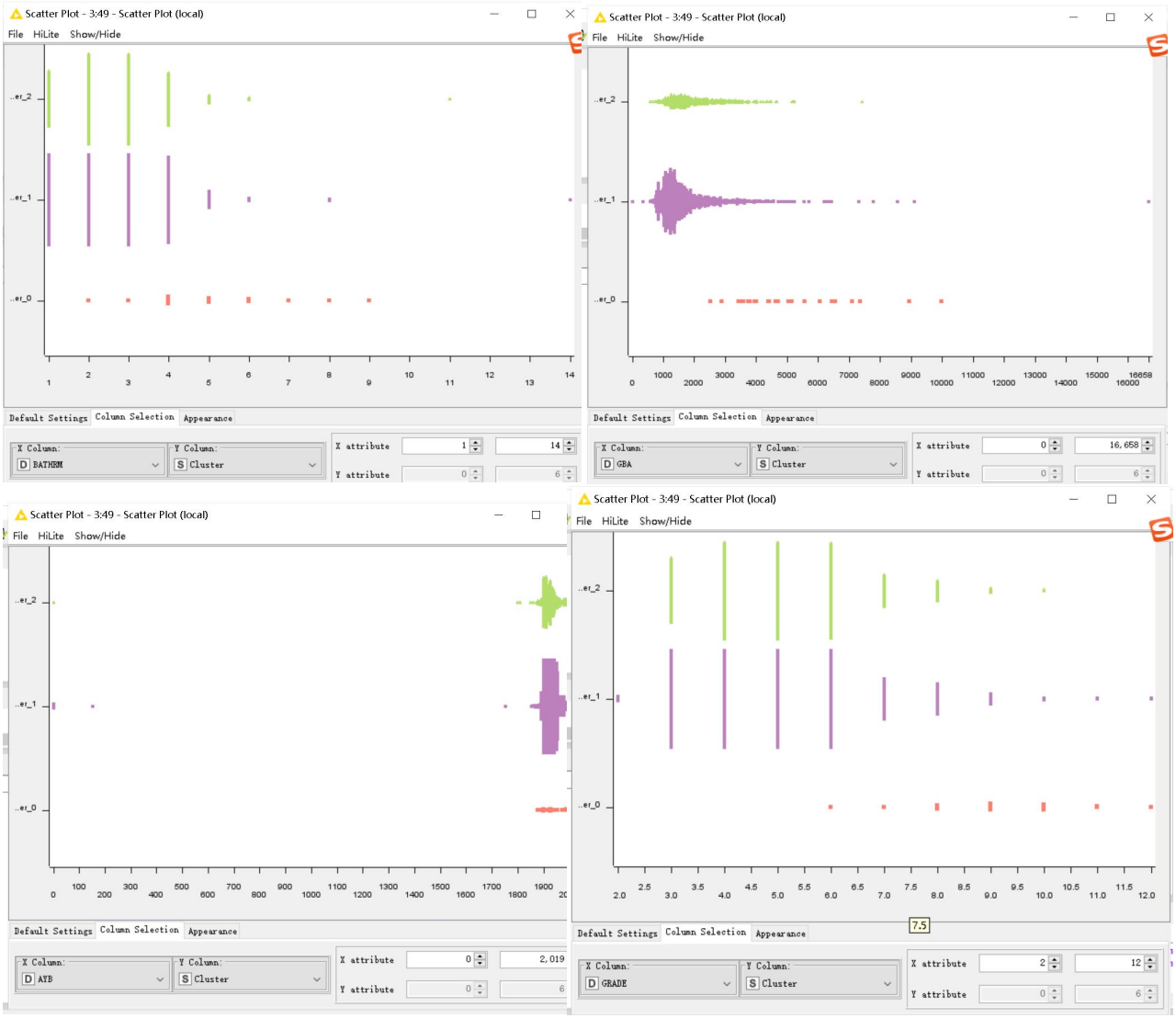





BATHRM
14
13
12
11
10
9
8
7
6
5
4
3
2
1
cluster_1
cluster_2
cluster_0
Cluster
cluster_0
cluster_1
cluster_2


PRICE
8500000
8000000
7500000
7000000
6500000
6000000
5500000
5000000
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
cluster_1
cluster_2
cluster_0
Cluster


GBA
17000
16000
15000
14000
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
cluster_1
cluster_2
cluster_0
Cluster
cluster_0
cluster_1
cluster_2



1B Data Pre-processing

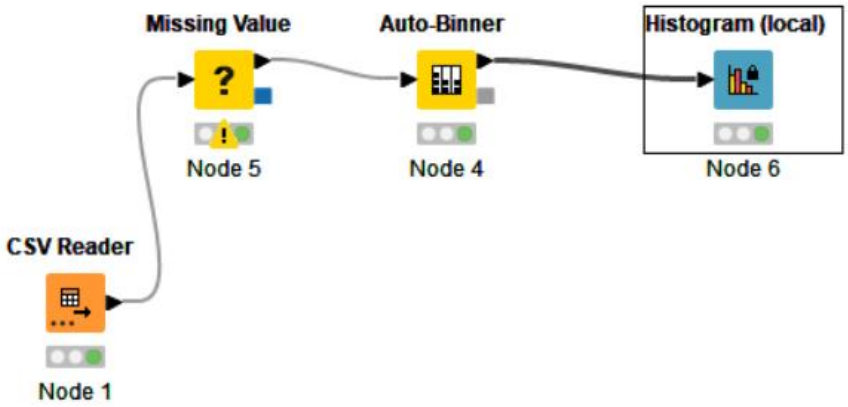
1. Binning techniques

Binning technique is a data preprocessing technique that divides continuous data into a limited number of discrete data or intervals. The following will use the same depth and same width technique to smooth the Value property value.

1.1 Equi-width binning

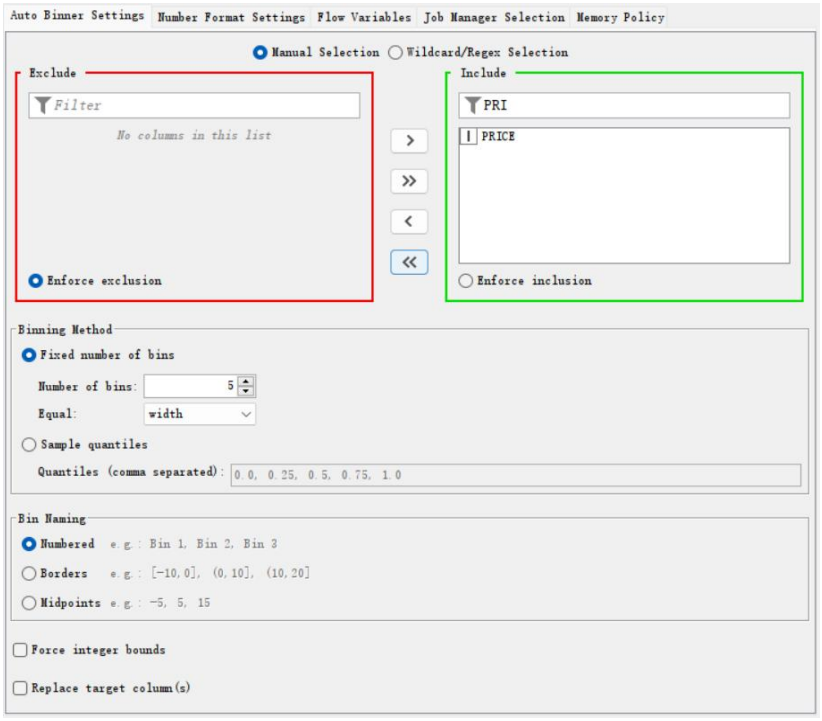
Using “Auto-Binner” and “Histogram(local)”. Auto-Binner is used to Equi-width binning and Histogram is used to view the distribution of binning results.

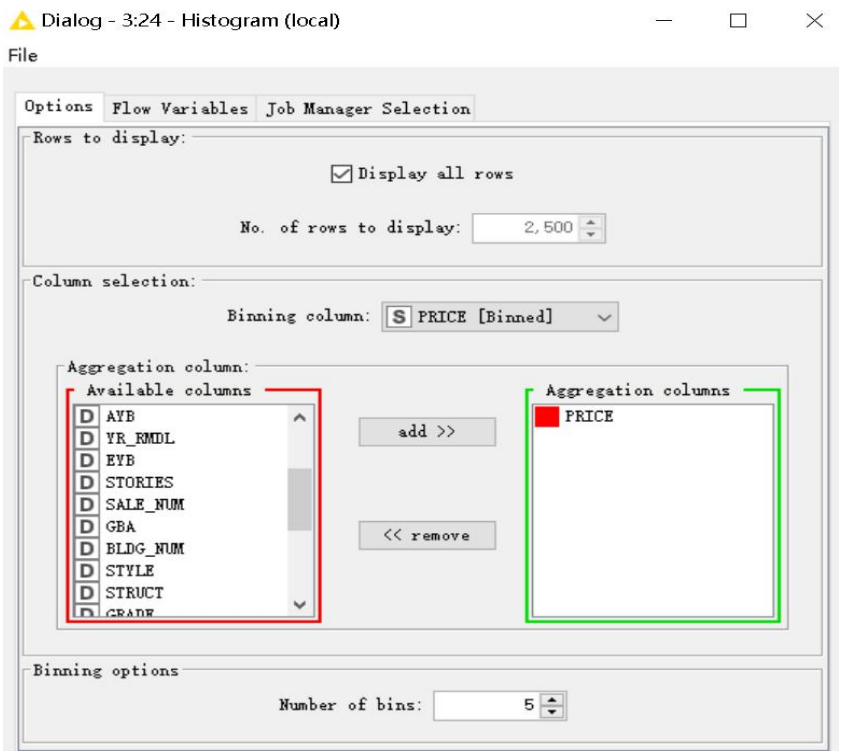
1.



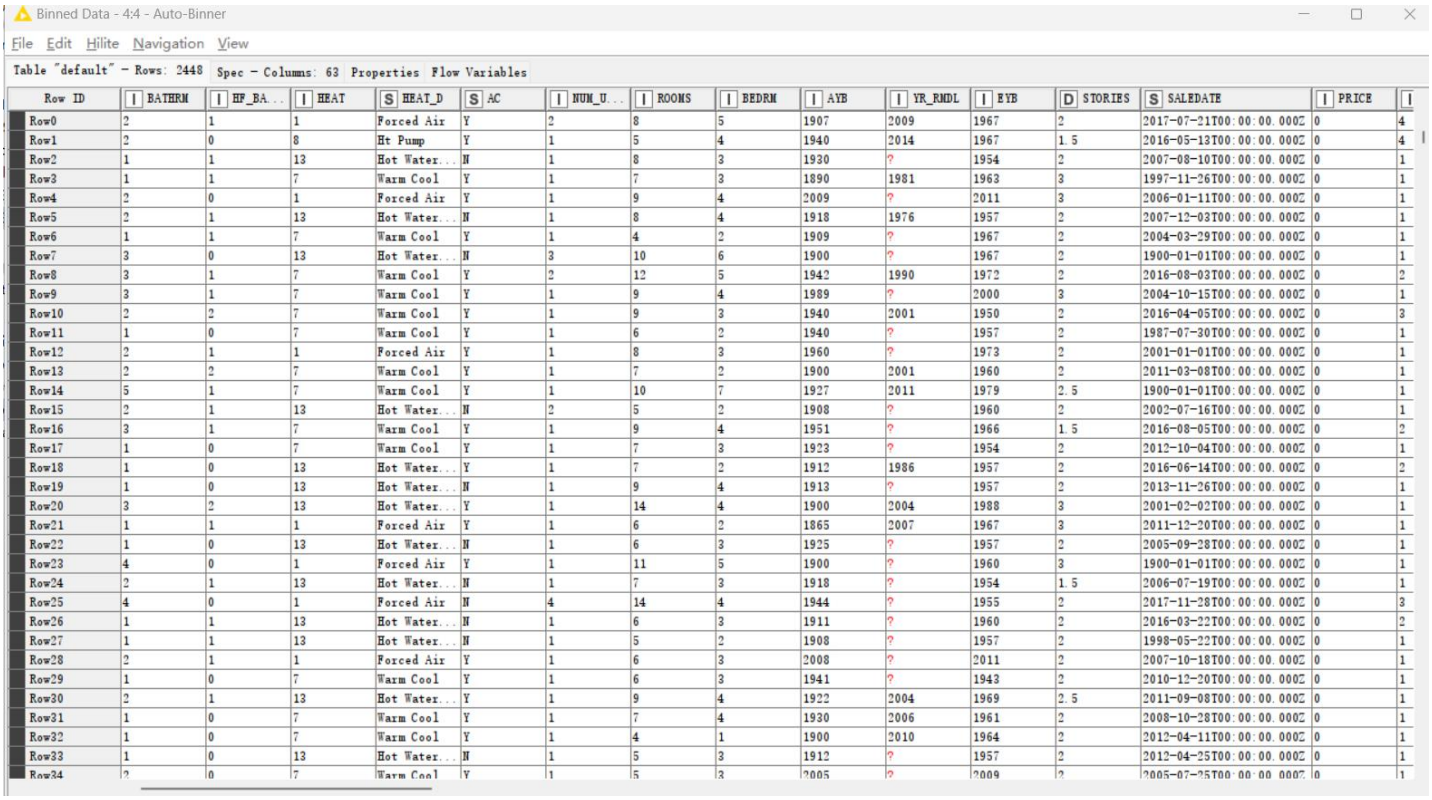
2. Parameter setting

The reason why the number of bins is 5 is that when the number of bins is greater than 5, empty boxes will appear.



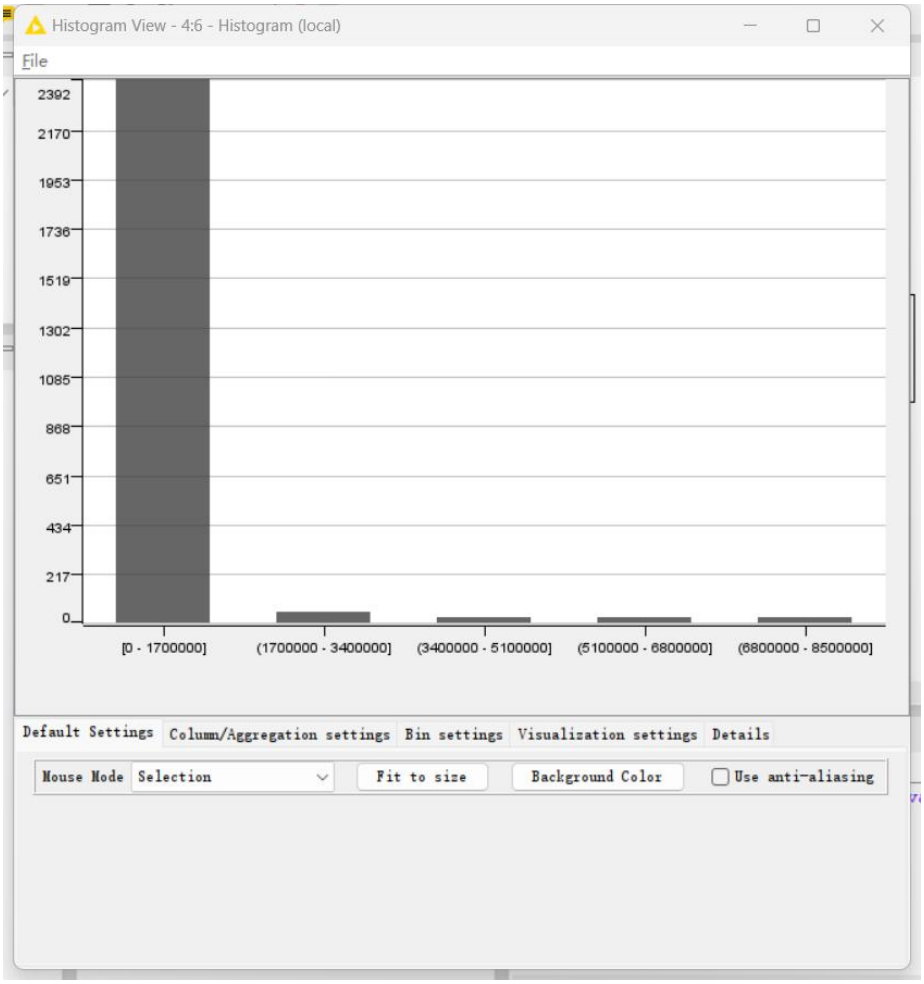


3. Result of auto-binner：



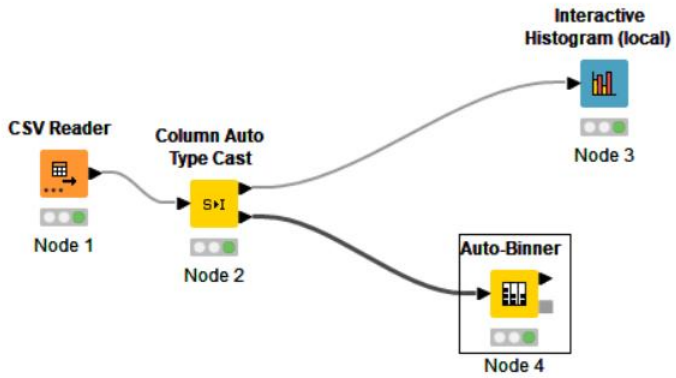
Row ID	BATHRM	HF_BA...	HEAT	HEAT_D	AC	NUM_U...	ROOMS	BEDRM	AYB	YR_RMDL	EYB	STORIES	SALEDATE	PRICE	
Row0	2	1	1	Forced Air	Y	2	8	5	1907	2009	1967	2	2017-07-21T00:00:00.000Z	0	4
Row1	2	0	8	Ht Pump	Y	1	5	4	1940	2014	1967	1.5	2016-05-13T00:00:00.000Z	0	4
Row2	1	1	13	Hot Water...	N	1	8	3	1930	?	1954	2	2007-08-10T00:00:00.000Z	0	1
Row3	1	1	7	Warm Cool	Y	1	7	3	1890	1981	1963	3	1997-11-26T00:00:00.000Z	0	1
Row4	2	0	1	Forced Air	Y	1	9	4	2009	?	2011	3	2006-01-11T00:00:00.000Z	0	1
Row5	2	1	13	Hot Water...	N	1	8	4	1918	1976	1957	2	2007-12-03T00:00:00.000Z	0	1
Row6	1	1	7	Warm Cool	Y	1	4	2	1909	?	1967	2	2004-03-29T00:00:00.000Z	0	1
Row7	3	0	13	Hot Water...	N	3	10	6	1900	?	1967	2	1900-01-01T00:00:00.000Z	0	1
Row8	3	1	7	Warm Cool	Y	2	12	5	1942	1990	1972	2	2016-08-03T00:00:00.000Z	0	2
Row9	3	1	7	Warm Cool	Y	1	9	4	1989	?	2000	3	2004-10-15T00:00:00.000Z	0	1
Row10	2	2	7	Warm Cool	Y	1	9	3	1940	2001	1950	2	2016-04-05T00:00:00.000Z	0	3
Row11	1	0	7	Warm Cool	Y	1	6	2	1940	?	1957	2	1987-07-30T00:00:00.000Z	0	1
Row12	2	1	1	Forced Air	Y	1	8	3	1960	?	1973	2	2001-01-01T00:00:00.000Z	0	1
Row13	2	2	7	Warm Cool	Y	1	7	2	1900	2001	1960	2	2011-03-08T00:00:00.000Z	0	1
Row14	5	1	7	Warm Cool	Y	1	10	7	1927	2011	1979	2.5	1900-01-01T00:00:00.000Z	0	1
Row15	2	1	13	Hot Water...	N	2	5	2	1908	?	1960	2	2002-07-16T00:00:00.000Z	0	1
Row16	3	1	7	Warm Cool	Y	1	9	4	1951	?	1966	1.5	2016-08-05T00:00:00.000Z	0	2
Row17	1	0	7	Warm Cool	Y	1	7	3	1923	?	1954	2	2012-10-04T00:00:00.000Z	0	1
Row18	1	0	13	Hot Water...	Y	1	7	2	1912	1986	1957	2	2016-06-14T00:00:00.000Z	0	2
Row19	1	0	13	Hot Water...	N	1	9	4	1913	?	1957	2	2013-11-26T00:00:00.000Z	0	1
Row20	3	2	13	Hot Water...	Y	1	14	4	1900	2004	1988	3	2001-02-02T00:00:00.000Z	0	1
Row21	1	1	1	Forced Air	Y	1	6	2	1865	2007	1967	3	2011-12-20T00:00:00.000Z	0	1
Row22	1	0	13	Hot Water...	N	1	6	3	1925	?	1957	2	2005-09-28T00:00:00.000Z	0	1
Row23	4	0	1	Forced Air	Y	1	11	5	1900	?	1960	3	1900-01-01T00:00:00.000Z	0	1
Row24	2	1	13	Hot Water...	N	1	7	3	1918	?	1954	1.5	2006-07-19T00:00:00.000Z	0	1
Row25	4	0	1	Forced Air	N	4	14	4	1944	?	1955	2	2017-11-28T00:00:00.000Z	0	3
Row26	1	1	13	Hot Water...	N	1	6	3	1911	?	1960	2	2016-03-22T00:00:00.000Z	0	2
Row27	1	1	13	Hot Water...	N	1	5	2	1908	?	1957	2	1998-05-22T00:00:00.000Z	0	1
Row28	2	1	1	Forced Air	Y	1	6	3	2008	?	2011	2	2007-10-18T00:00:00.000Z	0	1
Row29	1	0	7	Warm Cool	Y	1	6	3	1941	?	1943	2	2010-12-20T00:00:00.000Z	0	1
Row30	2	1	13	Hot Water...	Y	1	9	4	1922	2004	1969	2.5	2011-09-08T00:00:00.000Z	0	1
Row31	1	0	7	Warm Cool	Y	1	7	4	1930	2006	1961	2	2008-10-28T00:00:00.000Z	0	1
Row32	1	0	7	Warm Cool	Y	1	4	1	1900	2010	1964	2	2012-04-11T00:00:00.000Z	0	1
Row33	1	0	13	Hot Water...	N	1	5	3	1912	?	1957	2	2012-04-25T00:00:00.000Z	0	1
Row34	2	0	7	Warm Cool	Y	1	5	3	2005	?	2009	2	2005-07-25T00:00:00.000Z	0	1

4. Result of histogram



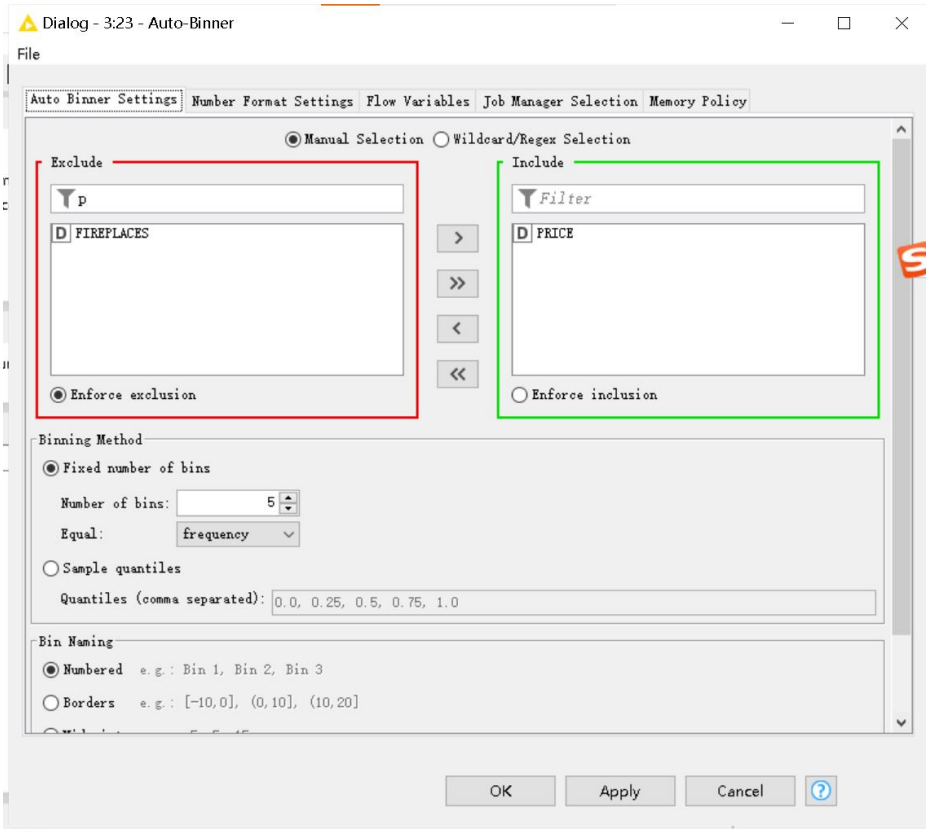
1. 2 Equi-depth binning

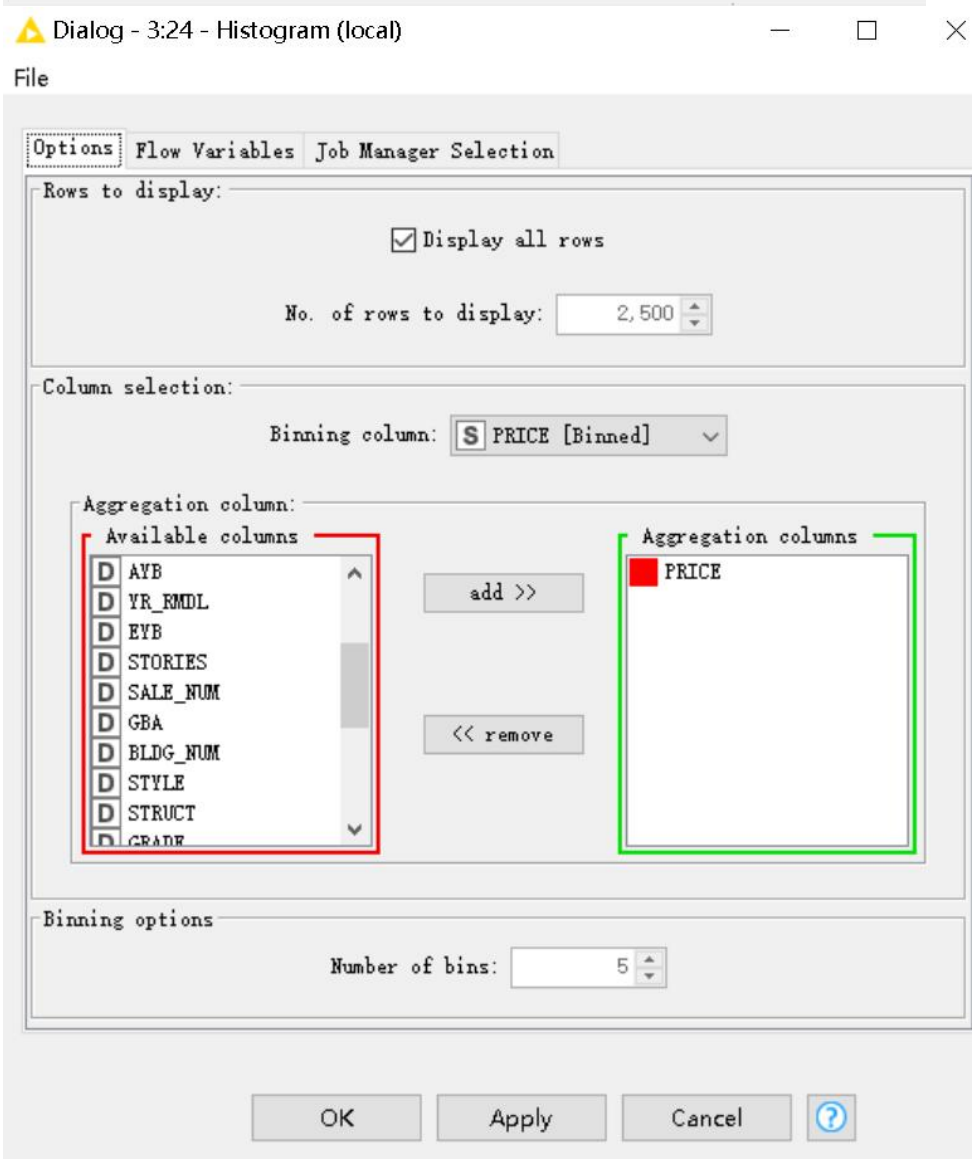
Using “Auto-Binner” and “Interactive Histogram”. Auto-Binner is used to Equi-depth binning and Histogram is used to view the distribution of binning results.



2. Parameter setting

The reason why the number of bins is 5 is that when the number of bins is greater than seven, empty boxes will appear. The number of five is chosen because the data will be evenly distributed.


Dialog - 3:23 - Auto-Binner
File
Auto Binner Settings
Number Format Settings
Flow Variables
Job Manager Selection
Memory Policy
Manual Selection
Wildcard/Regex Selection
Exclude
p
FIREPLACES
Enforce exclusion
Include
Filter
PRICE
Enforce inclusion
Binning Method
Fixed number of bins
Number of bins:
5
Equal:
frequency
Sample quantiles
Quantiles (comma separated):
0.0, 0.25, 0.5, 0.75, 1.0
Bin Naming
Numbered
e.g.: Bin 1, Bin 2, Bin 3
Borders
e.g.: [-10,0], (0,10], (10,20]
OK
Apply
Cancel


Dialog - 3:24 - Histogram (local)
File
Options
Flow Variables
Job Manager Selection
Rows to display:
Display all rows
No. of rows to display:
2,500
Column selection:
Binning column:
PRICE [Binned]
Aggregation column:
Available columns
AYB
YR_RMDL
EYB
STORIES
SALE_NUM
GBA
BLDG_NUM
STYLE
STRUCT
add >>
<< remove
Aggregation columns
PRICE
Binning options
Number of bins:
5
OK
Apply
Cancel

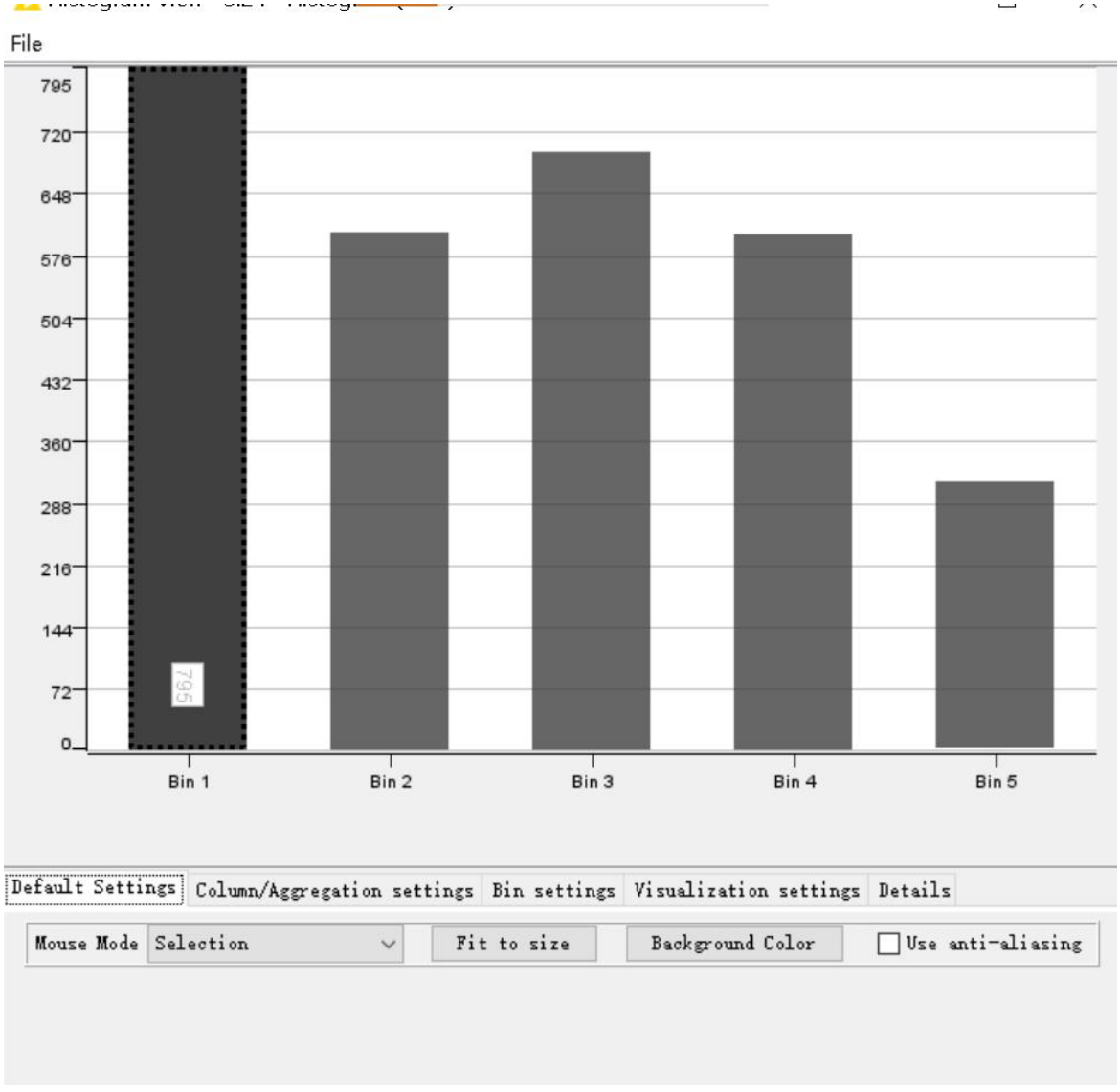
2. Result of auto-binner:

File Edit Hilite Navigation View

Table "default" – Rows: 3000 Spec – Columns: 38 Properties Flow Variables

Row ID	S CNDTN_D	D EXTWALL	S EXTWA...	D ROOF	S ROOF_D	D INTWALL	S INTWA...	D KITCHENS	D FIREP...	D USECODE	D LANDAREA	S GIS_LAST_MOD_DTTM	D QUALI...	S PRICE...
Row0	Average	14	Common Brick	6	Metal- Sms	6	Hardwood	1	0	11	1,545	2018-07-22T18:01:43.000Z	1	Bin 2
Row1	Average	14	Common Brick	6	Metal- Sms	6	Hardwood	1	0	13	2,102	2018-07-22T18:01:43.000Z	0	Bin 2
Row2	Average	6	Wood Siding	1	Comp Shingle	6	Hardwood	1	0	12	3,860	2018-07-22T18:01:43.000Z	0	Bin 3
Row3	Average	14	Common Brick	2	Built Up	11	Hardwood/...	2	0	11	1,500	2018-07-22T18:01:43.000Z	0	Bin 1
Row4	Average	4	Vinyl Siding	4	Shake	3	Wood Floor	1	0	11	870	2018-07-22T18:01:43.000Z	0	Bin 3
Row5	Good	14	Common Brick	6	Metal- Sms	3	Wood Floor	1	0	11	1,172	2018-07-22T18:01:43.000Z	1	Bin 5
Row6	Good	14	Common Brick	2	Built Up	3	Wood Floor	2	1	24	870	2018-07-22T18:01:43.000Z	0	Bin 2
Row7	Good	14	Common Brick	6	Metal- Sms	6	Hardwood	1	1	11	1,634	2018-07-22T18:01:43.000Z	1	Bin 4
Row8	Good	14	Common Brick	2	Built Up	11	Hardwood/...	1	0	13	2,005	2018-07-22T18:01:43.000Z	1	Bin 2
Row9	Average	14	Common Brick	5	Metal- Pre	0	Default	1	0	11	1,150	2018-07-22T18:01:43.000Z	1	Bin 3
Row10	Average	14	Common Brick	1	Comp Shingle	6	Hardwood	1	1	12	5,000	2018-07-22T18:01:43.000Z	0	Bin 1
Row11	Average	6	Wood Siding	6	Metal- Sms	6	Hardwood	1	0	11	1,600	2018-07-22T18:01:43.000Z	0	Bin 3
Row12	Average	22	Brick/Siding	1	Comp Shingle	3	Wood Floor	1	0	11	1,900	2018-07-22T18:01:43.000Z	0	Bin 2
Row13	Good	4	Vinyl Siding	1	Comp Shingle	6	Hardwood	1	0	12	2,520	2018-07-22T18:01:43.000Z	1	Bin 3
Row14	Good	5	Stucco	1	Comp Shingle	6	Hardwood	1	0	12	2,822	2018-07-22T18:01:43.000Z	1	Bin 2
Row15	Good	22	Brick/Siding	1	Comp Shingle	11	Hardwood/...	1	0	12	10,266	2018-07-22T18:01:43.000Z	0	Bin 3
Row16	Average	14	Common Brick	6	Metal- Sms	11	Hardwood/...	1	0	11	2,224	2018-07-22T18:01:43.000Z	1	Bin 2
Row17	Average	14	Common Brick	2	Built Up	6	Hardwood	1	1	13	2,691	2018-07-22T18:01:43.000Z	0	Bin 1
Row18	Average	14	Common Brick	6	Metal- Sms	6	Hardwood	4	0	23	3,808	2018-07-22T18:01:43.000Z	1	Bin 4
Row19	Good	14	Common Brick	6	Metal- Sms	6	Hardwood	2	1	11	1,935	2018-07-22T18:01:43.000Z	1	Bin 4
Row20	Fair	14	Common Brick	6	Metal- Sms	6	Hardwood	1	0	13	2,838	2018-07-22T18:01:43.000Z	0	Bin 3
Row21	Average	14	Common Brick	6	Metal- Sms	6	Hardwood	1	1	11	930	2018-07-22T18:01:43.000Z	0	Bin 1
Row22	Very Good	14	Common Brick	13	Neopren	11	Hardwood/...	1	0	11	978	2018-07-22T18:01:43.000Z	0	Bin 1
Row23	Average	5	Stucco	11	Slate	6	Hardwood	1	1	12	4,250	2018-07-22T18:01:43.000Z	0	Bin 1
Row24	Good	14	Common Brick	1	Comp Shingle	6	Hardwood	1	1	11	1,180	2018-07-22T18:01:43.000Z	1	Bin 4
Row25	Average	14	Common Brick	2	Built Up	6	Hardwood	1	0	13	2,811	2018-07-22T18:01:43.000Z	0	Bin 3
Row26	Average	14	Common Brick	6	Metal- Sms	6	Hardwood	1	0	11	1,209	2018-07-22T18:01:43.000Z	1	Bin 2
Row27	Average	14	Common Brick	2	Built Up	6	Hardwood	1	0	11	1,500	2018-07-22T18:01:43.000Z	0	Bin 3
Row28	Good	14	Common Brick	2	Built Up	6	Hardwood	2	0	24	1,575	2018-07-22T18:01:43.000Z	0	Bin 2
Row29	Good	14	Common Brick	13	Neopren	6	Hardwood	1	0	11	875	2018-07-22T18:01:43.000Z	1	Bin 4
Row30	Good	14	Common Brick	1	Comp Shingle	11	Hardwood/...	2	0	24	1,462	2018-07-22T18:01:43.000Z	0	Bin 4
Row31	Average	14	Common Brick	6	Metal- Sms	6	Hardwood	1	0	11	1,399	2018-07-22T18:01:43.000Z	1	Bin 4
Row32	Good	14	Common Brick	10	Clay Tile	6	Hardwood	1	1	12	6,397	2018-07-22T18:01:43.000Z	0	Bin 3
Row33	Very Good	14	Common Brick	6	Metal- Sms	6	Hardwood	1	0	11	2,305	2018-07-22T18:01:43.000Z	1	Bin 4

3. Result of Histogram:

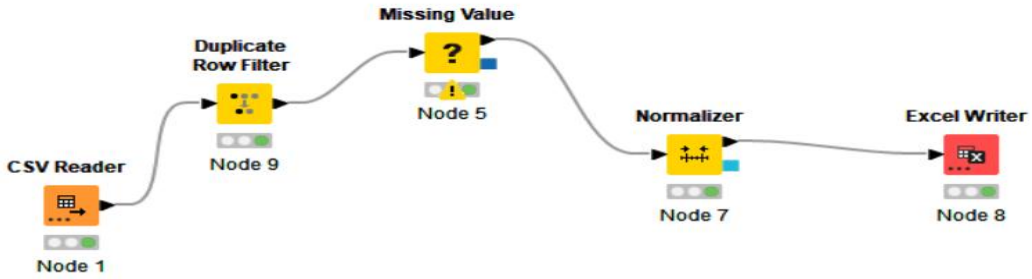


In contrast, using equi-depth bins is more advantageous because the data is more evenly distributed and easier for analysis. Equi-depth binning divides the quantity into the same number of groups for binning, which is conducive to the situation that the data distribution is concentrated or there are obvious outliers. Equal-width bins are divided into the same width, which is conducive to a more uniform distribution of data, but the numerical distribution is uneven in the price attribute.

2. Normalize：

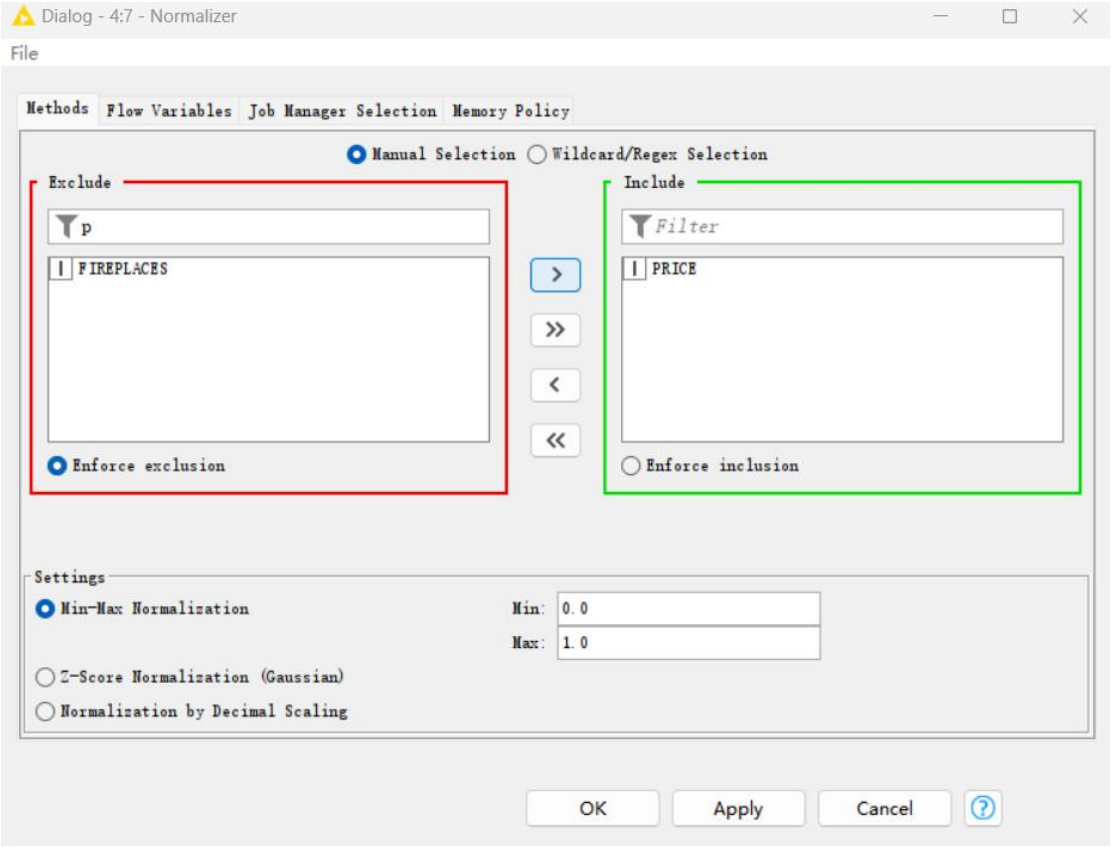
2.1 Min-Max Normalization

The purpose of standardization is to convert values into uniform units of data, reduce dimensions and reduce bias and confirmation weights, so that the data are comparable and standardized. The method is to map the maximum value in the original data to one, the minimum value to zero, and the others only map to a number from zero to one and distribute the data between zero and one. Using “Normailzer” with the Min-Max Normalization setting and “Excel Writer” to output result.

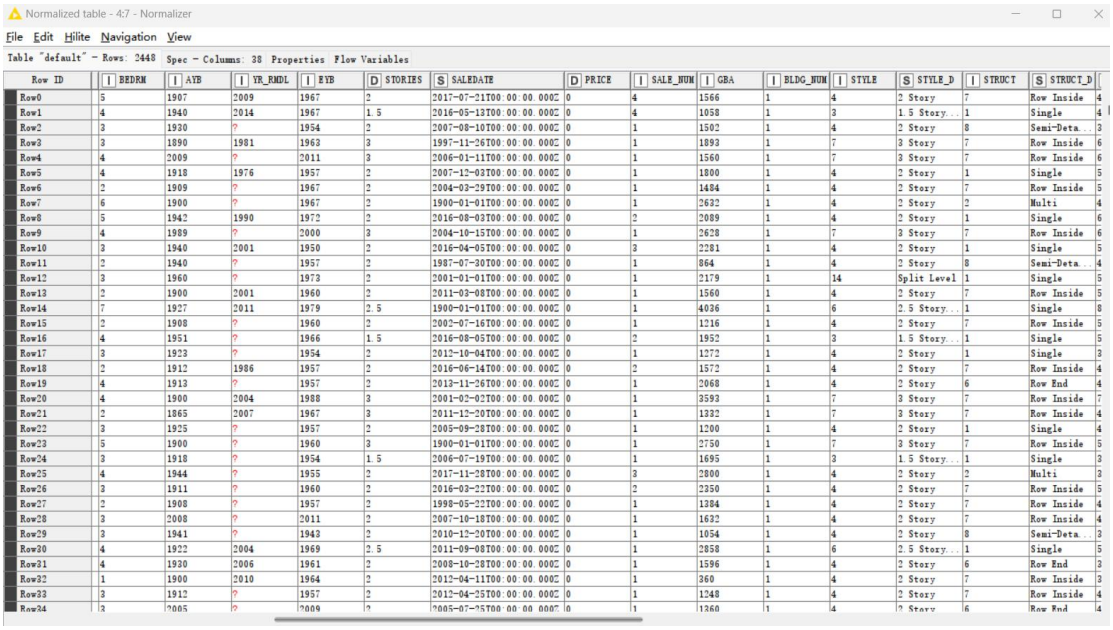


2. Parameter setting

Set min is 0 and max is 1. The significance of scaling relative to maximum minimum is that it is easier to train the model and keep none of the data negative.



3.Result of Min-Max Normalization：


Normalized table - 4:7 - Normalizer

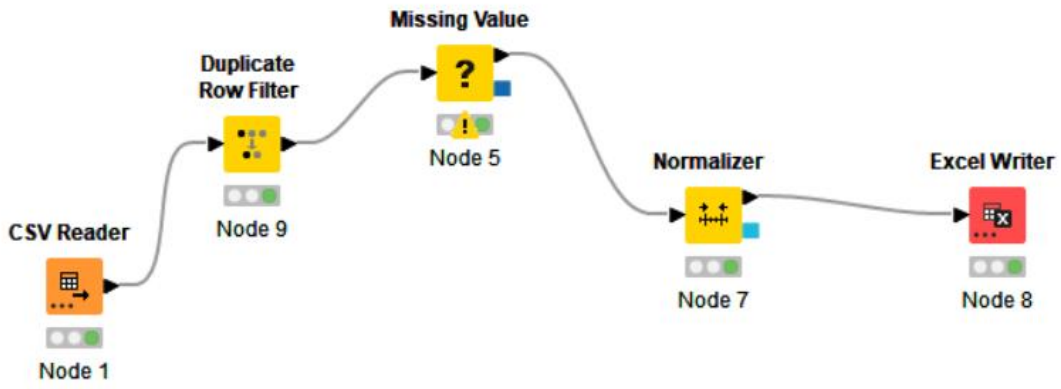
File Edit Hilite Navigation View

Table "default" – Rows: 2448 Spec – Columns: 38 Properties Flow Variables

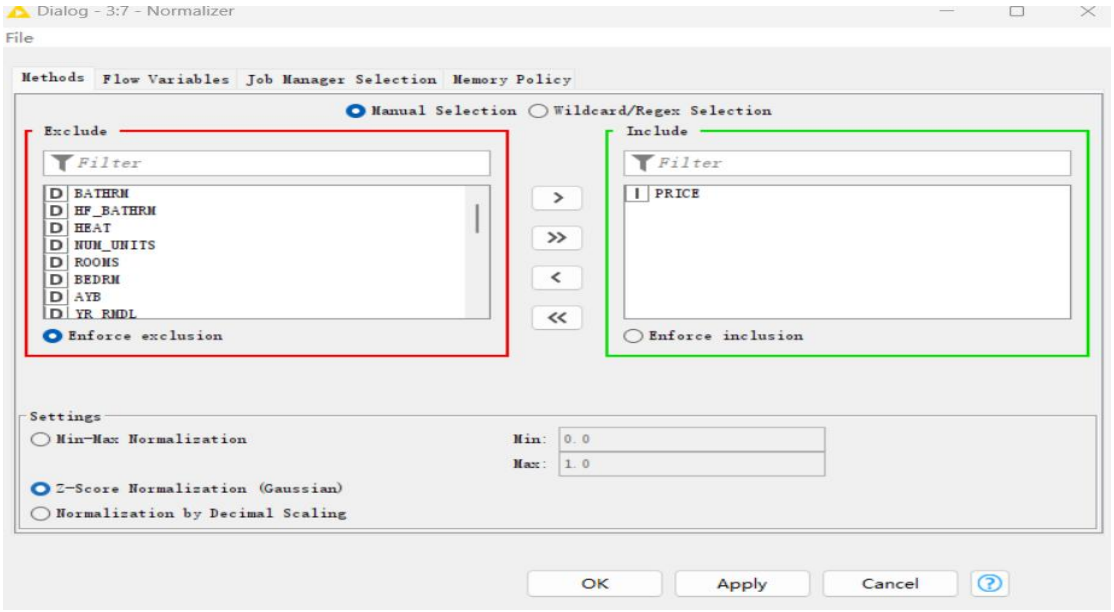
Row ID	BEDRM	AYB	YR_RMDL	EYB	STORIES	SALEDATE	PRICE	SALE_NUM	GBA	BLDG_NUM	STYLE	STYLE_D	STRUCT	STRUCT_D
Row0	5	1907	2009	1967	2	2017-07-21T00:00:00.000Z	0	4	1566	1	4	2 Story	7	Row Inside
Row1	4	1940	2014	1967	1.5	2016-05-13T00:00:00.000Z	0	4	1058	1	3	1.5 Story...	1	Single
Row2	3	1930	?	1954	2	2007-08-10T00:00:00.000Z	0	1	1502	1	4	2 Story	8	Semi-Deta...
Row3	3	1890	1981	1963	3	1997-11-26T00:00:00.000Z	0	1	1893	1	7	3 Story	7	Row Inside
Row4	4	2009	?	2011	3	2006-01-11T00:00:00.000Z	0	1	1560	1	7	3 Story	7	Row Inside
Row5	4	1918	1976	1957	2	2007-12-03T00:00:00.000Z	0	1	1800	1	4	2 Story	1	Single
Row6	2	1909	?	1967	2	2004-03-29T00:00:00.000Z	0	1	1484	1	4	2 Story	7	Row Inside
Row7	6	1900	?	1967	2	1900-01-01T00:00:00.000Z	0	1	2632	1	4	2 Story	2	Multi
Row8	5	1942	1990	1972	2	2016-08-03T00:00:00.000Z	0	2	2089	1	4	2 Story	1	Single
Row9	4	1989	?	2000	3	2004-10-15T00:00:00.000Z	0	1	2628	1	7	3 Story	7	Row Inside
Row10	3	1940	2001	1950	2	2016-04-05T00:00:00.000Z	0	3	2281	1	4	2 Story	1	Single
Row11	2	1940	?	1957	2	1987-07-30T00:00:00.000Z	0	1	864	1	4	2 Story	8	Semi-Deta...
Row12	3	1960	?	1973	2	2001-01-01T00:00:00.000Z	0	1	2179	1	14	Split Level	1	Single
Row13	2	1900	2001	1960	2	2011-03-08T00:00:00.000Z	0	1	1560	1	4	2 Story	7	Row Inside
Row14	7	1927	2011	1979	2.5	1900-01-01T00:00:00.000Z	0	1	4036	1	6	2.5 Story...	1	Single
Row15	2	1908	?	1960	2	2002-07-16T00:00:00.000Z	0	1	1216	1	4	2 Story	7	Row Inside
Row16	4	1951	?	1966	1.5	2016-08-05T00:00:00.000Z	0	2	1952	1	3	1.5 Story...	1	Single
Row17	3	1923	?	1954	2	2012-10-04T00:00:00.000Z	0	1	1272	1	4	2 Story	1	Single
Row18	2	1912	1986	1957	2	2016-06-14T00:00:00.000Z	0	2	1572	1	4	2 Story	7	Row Inside
Row19	4	1913	?	1957	2	2013-11-26T00:00:00.000Z	0	1	2068	1	4	2 Story	6	Row End
Row20	4	1900	2004	1988	3	2001-02-02T00:00:00.000Z	0	1	3593	1	7	3 Story	7	Row Inside
Row21	2	1865	2007	1967	3	2011-12-20T00:00:00.000Z	0	1	1332	1	7	3 Story	7	Row Inside
Row22	3	1925	?	1957	2	2005-09-28T00:00:00.000Z	0	1	1200	1	4	2 Story	1	Single
Row23	5	1900	?	1960	3	1900-01-01T00:00:00.000Z	0	1	2750	1	7	3 Story	7	Row Inside
Row24	3	1918	?	1954	1.5	2006-07-19T00:00:00.000Z	0	1	1695	1	3	1.5 Story...	1	Single
Row25	4	1944	?	1955	2	2017-11-28T00:00:00.000Z	0	3	2800	1	4	2 Story	2	Multi
Row26	3	1911	?	1960	2	2016-03-22T00:00:00.000Z	0	2	2350	1	4	2 Story	7	Row Inside
Row27	2	1908	?	1957	2	1998-05-22T00:00:00.000Z	0	1	1384	1	4	2 Story	7	Row Inside
Row28	3	2008	?	2011	2	2007-10-18T00:00:00.000Z	0	1	1632	1	4	2 Story	7	Row Inside
Row29	3	1941	?	1943	2	2010-12-20T00:00:00.000Z	0	1	1054	1	4	2 Story	8	Semi-Deta...
Row30	4	1922	2004	1969	2.5	2011-09-08T00:00:00.000Z	0	1	2858	1	6	2.5 Story...	1	Single
Row31	4	1930	2006	1961	2	2008-10-28T00:00:00.000Z	0	1	1596	1	4	2 Story	6	Row End
Row32	1	1900	2010	1964	2	2012-04-11T00:00:00.000Z	0	1	360	1	4	2 Story	7	Row Inside
Row33	3	1912	?	1957	2	2012-04-25T00:00:00.000Z	0	1	1248	1	4	2 Story	7	Row Inside
Row34	3	2005	?	2009	2	2005-07-25T00:00:00.000Z	0	1	1360	1	4	2 Story	6	Row End

2.2 Z-Score Normalization

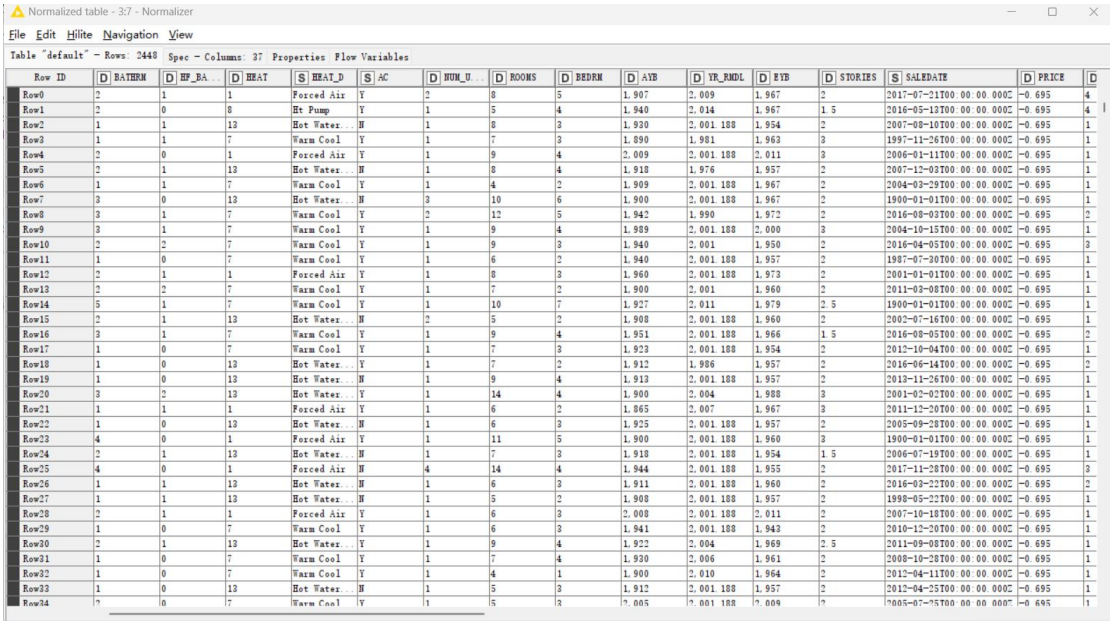
Z-Score Normalization is normalized to a distribution with a mean of 0 and a standard deviation of one by subtracting the mean from the original data and dividing by the standard deviation. This model can reduce the impact of noise in the data on the model and improve the stability and accuracy of the model, but this method may be affected only by extremes, resulting in poor standardization.



2. Parameter setting



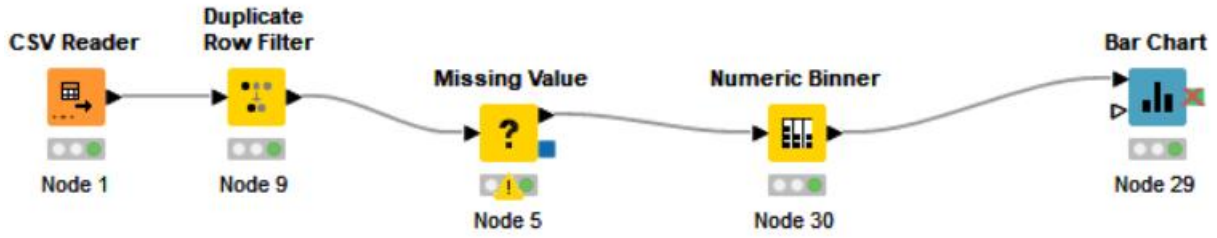
3.Result of Z-Score Normalization：

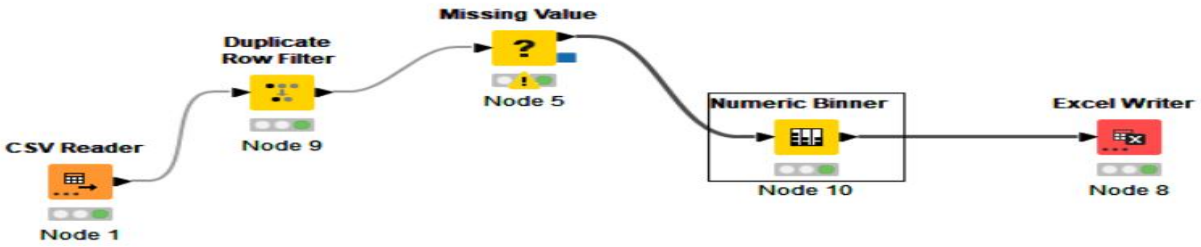


Row ID	BATHRM	HF_BA...	HEAT	HEAT_D	AC	NUM_U...	ROOMS	BEDRM	AYB	YR_RNDL	EYB	STORIES	SALEDATE	PRICE	D
Row0	2	1	1	Forced Air	Y	2	8	5	1.907	2.009	1.967	2	2017-07-21T00:00:00.000Z	-0.695	4
Row1	2	0	8	Ht Pump	Y	1	5	4	1.940	2.014	1.967	1.5	2016-05-13T00:00:00.000Z	-0.695	4
Row2	1	1	13	Hot Water...	N	1	8	3	1.930	2.001.188	1.954	2	2007-08-10T00:00:00.000Z	-0.695	1
Row3	1	1	7	Warm Cool	Y	1	7	3	1.890	1.981	1.963	3	1997-11-26T00:00:00.000Z	-0.695	1
Row4	2	0	1	Forced Air	Y	1	9	4	2.009	2.001.188	2.011	3	2006-01-11T00:00:00.000Z	-0.695	1
Row5	2	1	13	Hot Water...	N	1	8	4	1.918	1.976	1.957	2	2007-12-03T00:00:00.000Z	-0.695	1
Row6	1	1	7	Warm Cool	Y	1	4	2	1.909	2.001.188	1.967	2	2004-03-29T00:00:00.000Z	-0.695	1
Row7	3	0	13	Hot Water...	N	3	10	6	1.900	2.001.188	1.967	2	1900-01-01T00:00:00.000Z	-0.695	1
Row8	3	1	7	Warm Cool	Y	2	12	5	1.942	1.990	1.972	2	2016-08-03T00:00:00.000Z	-0.695	2
Row9	3	1	7	Warm Cool	Y	1	9	4	1.989	2.001.188	2.000	3	2004-10-15T00:00:00.000Z	-0.695	1
Row10	2	2	7	Warm Cool	Y	1	9	3	1.940	2.001	1.950	2	2016-04-05T00:00:00.000Z	-0.695	3
Row11	1	0	7	Warm Cool	Y	1	6	2	1.940	2.001.188	1.957	2	1987-07-30T00:00:00.000Z	-0.695	1
Row12	2	1	1	Forced Air	Y	1	8	3	1.960	2.001.188	1.973	2	2001-01-01T00:00:00.000Z	-0.695	1
Row13	2	2	7	Warm Cool	Y	1	7	2	1.900	2.001	1.960	2	2011-03-08T00:00:00.000Z	-0.695	1
Row14	5	1	7	Warm Cool	Y	1	10	7	1.927	2.011	1.979	2.5	1900-01-01T00:00:00.000Z	-0.695	1
Row15	2	1	13	Hot Water...	N	2	5	2	1.908	2.001.188	1.960	2	2002-07-16T00:00:00.000Z	-0.695	1
Row16	3	1	7	Warm Cool	Y	1	9	4	1.951	2.001.188	1.966	1.5	2016-08-05T00:00:00.000Z	-0.695	2
Row17	1	0	7	Warm Cool	Y	1	7	3	1.923	2.001.188	1.954	2	2012-10-04T00:00:00.000Z	-0.695	1
Row18	1	0	13	Hot Water...	Y	1	7	2	1.912	1.986	1.957	2	2016-06-14T00:00:00.000Z	-0.695	2
Row19	1	0	13	Hot Water...	N	1	9	4	1.913	2.001.188	1.957	2	2013-11-26T00:00:00.000Z	-0.695	1
Row20	3	2	13	Hot Water...	Y	1	14	4	1.900	2.004	1.988	3	2001-02-02T00:00:00.000Z	-0.695	1
Row21	1	1	1	Forced Air	Y	1	6	2	1.865	2.007	1.967	3	2011-12-20T00:00:00.000Z	-0.695	1
Row22	1	0	13	Hot Water...	N	1	6	3	1.925	2.001.188	1.957	2	2005-09-28T00:00:00.000Z	-0.695	1
Row23	4	0	1	Forced Air	Y	1	11	5	1.900	2.001.188	1.960	3	1900-01-01T00:00:00.000Z	-0.695	1
Row24	2	1	13	Hot Water...	N	1	7	3	1.918	2.001.188	1.954	1.5	2006-07-19T00:00:00.000Z	-0.695	1
Row25	4	0	1	Forced Air	N	4	14	4	1.944	2.001.188	1.955	2	2017-11-28T00:00:00.000Z	-0.695	3
Row26	1	1	13	Hot Water...	N	1	6	3	1.911	2.001.188	1.960	2	2016-03-22T00:00:00.000Z	-0.695	2
Row27	1	1	13	Hot Water...	N	1	5	2	1.908	2.001.188	1.957	2	1998-05-22T00:00:00.000Z	-0.695	1
Row28	2	1	1	Forced Air	Y	1	6	3	2.008	2.001.188	2.011	2	2007-10-18T00:00:00.000Z	-0.695	1
Row29	1	0	7	Warm Cool	Y	1	6	3	1.941	2.001.188	1.943	2	2010-12-20T00:00:00.000Z	-0.695	1
Row30	2	1	13	Hot Water...	Y	1	9	4	1.922	2.004	1.969	2.5	2011-09-08T00:00:00.000Z	-0.695	1
Row31	1	0	7	Warm Cool	Y	1	7	4	1.930	2.006	1.961	2	2008-10-28T00:00:00.000Z	-0.695	1
Row32	1	0	7	Warm Cool	Y	1	4	1	1.900	2.010	1.964	2	2012-04-11T00:00:00.000Z	-0.695	1
Row33	1	0	13	Hot Water...	N	1	5	3	1.912	2.001.188	1.957	2	2012-04-25T00:00:00.000Z	-0.695	1
Row34	?	0	7	Warm Cool	Y	1	5	3	2.005	2.001.188	2.009	2	2005-07-25T00:00:00.000Z	-0.695	1

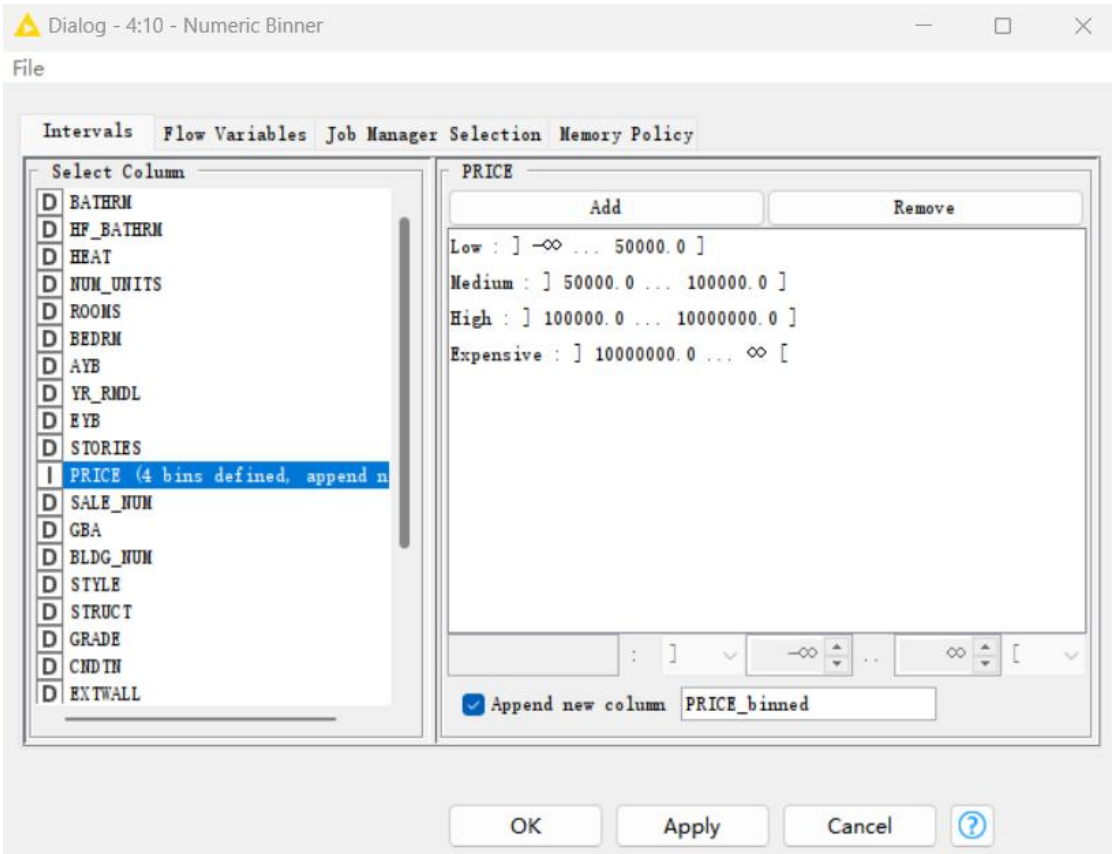
3. Discretise the "PRICE"

In the process of converting discretized continuous data into discrete data, it can be used for feature engineering or processing the input data of some computer learning algorithms. In this data, we split the price into high, medium, low and expensive (e.g.: Low=0-50k; Medium=50k-100k; High=100k-1000k; Expensive= 1000k+). Using “Numeric Binner” to discretise and “Excel Writer” to output result.





2. Parameter setting



3. Result of Discretise：

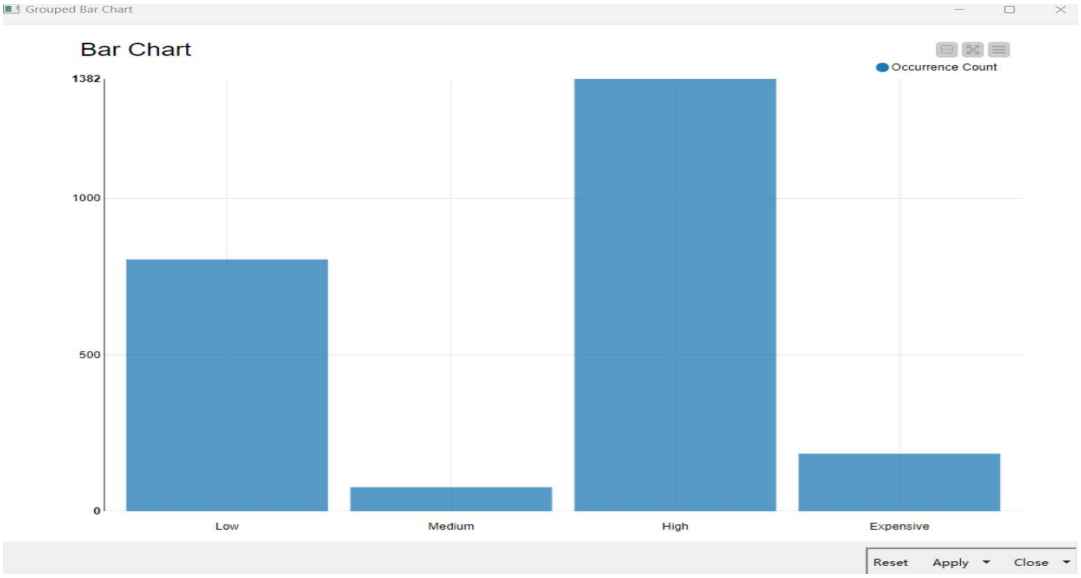
Binned Data - 3:10 - Numeric Binner

File Edit Hilite Navigation View

Table "default" – Rows: 2448 Spec – Columns: 38 Properties Flow Variables

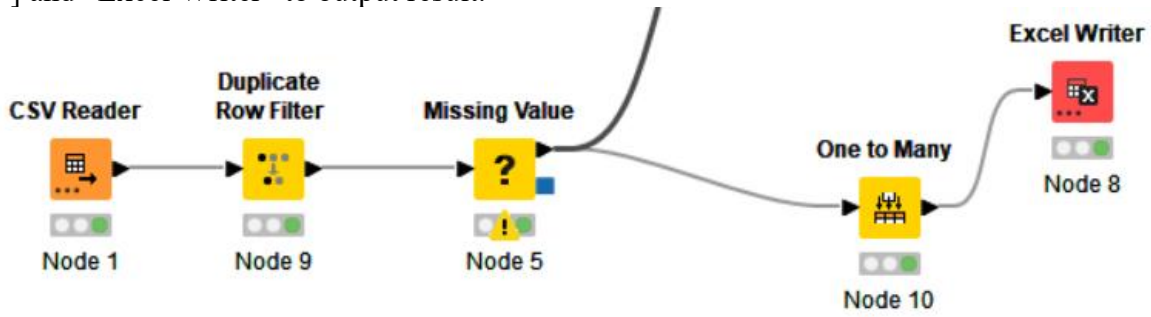
Row ID	HEAT_D	AC	NUM_U...	ROOMS	BEDRM	AYB	YR_RNDL	EYB	STORIES	SALEDATE	PRICE	PRICE_binned	
Row0	Forced Air	Y	2	8	5	1.907	2.009	1.967	2	2017-07-21T00:00:00.000Z	0	Low	4
Row1	Ht Pump	Y	1	5	4	1.940	2.014	1.967	1.5	2016-05-13T00:00:00.000Z	0	Low	4
Row2	Hot Water...	N	1	8	3	1.930	2.001 188	1.954	2	2007-08-10T00:00:00.000Z	0	Low	1
Row3	Warm Cool	Y	1	7	3	1.890	1.981	1.963	3	1997-11-26T00:00:00.000Z	0	Low	1
Row4	Forced Air	Y	1	9	4	2.009	2.001 188	2.011	3	2006-01-11T00:00:00.000Z	0	Low	1
Row5	Hot Water...	N	1	8	4	1.918	1.976	1.957	2	2007-12-03T00:00:00.000Z	0	Low	1
Row6	Warm Cool	Y	1	4	2	1.909	2.001 188	1.967	2	2004-03-29T00:00:00.000Z	0	Low	1
Row7	Hot Water...	N	3	10	6	1.900	2.001 188	1.967	2	1900-01-01T00:00:00.000Z	0	Low	1
Row8	Warm Cool	Y	2	12	5	1.942	1.990	1.972	2	2016-08-03T00:00:00.000Z	0	Low	2
Row9	Warm Cool	Y	1	9	4	1.989	2.001 188	2.000	3	2004-10-15T00:00:00.000Z	0	Low	1
Row10	Warm Cool	Y	1	9	3	1.940	2.001	1.950	2	2016-04-05T00:00:00.000Z	0	Low	3
Row11	Warm Cool	Y	1	6	2	1.940	2.001 188	1.957	2	1987-07-20T00:00:00.000Z	0	Low	1
Row12	Forced Air	Y	1	8	3	1.960	2.001 188	1.973	2	2001-01-01T00:00:00.000Z	0	Low	1
Row13	Warm Cool	Y	1	7	2	1.900	2.001	1.960	2	2011-03-08T00:00:00.000Z	0	Low	1
Row14	Warm Cool	Y	1	10	7	1.927	2.011	1.979	2.5	1900-01-01T00:00:00.000Z	0	Low	1
Row15	Hot Water...	N	2	5	2	1.908	2.001 188	1.960	2	2002-07-16T00:00:00.000Z	0	Low	1
Row16	Warm Cool	Y	1	9	4	1.951	2.001 188	1.966	1.5	2016-08-05T00:00:00.000Z	0	Low	2
Row17	Warm Cool	Y	1	7	3	1.923	2.001 188	1.954	2	2012-10-04T00:00:00.000Z	0	Low	1
Row18	Hot Water...	Y	1	7	2	1.912	1.986	1.957	2	2016-06-14T00:00:00.000Z	0	Low	2
Row19	Hot Water...	N	1	9	4	1.913	2.001 188	1.957	2	2013-11-26T00:00:00.000Z	0	Low	1
Row20	Hot Water...	Y	1	14	4	1.900	2.004	1.988	3	2001-02-02T00:00:00.000Z	0	Low	1
Row21	Forced Air	Y	1	6	2	1.865	2.007	1.967	3	2011-12-20T00:00:00.000Z	0	Low	1
Row22	Hot Water...	N	1	6	3	1.925	2.001 188	1.957	2	2005-09-28T00:00:00.000Z	0	Low	1
Row23	Forced Air	Y	1	11	5	1.900	2.001 188	1.960	3	1900-01-01T00:00:00.000Z	0	Low	1
Row24	Hot Water...	N	1	7	3	1.918	2.001 188	1.954	1.5	2006-07-19T00:00:00.000Z	0	Low	1
Row25	Forced Air	N	4	14	4	1.944	2.001 188	1.955	2	2017-11-28T00:00:00.000Z	0	Low	3
Row26	Hot Water...	N	1	6	3	1.911	2.001 188	1.960	2	2016-03-22T00:00:00.000Z	0	Low	2
Row27	Hot Water...	N	1	5	2	1.908	2.001 188	1.957	2	1998-05-22T00:00:00.000Z	0	Low	1
Row28	Forced Air	Y	1	6	3	2.008	2.001 188	2.011	2	2007-10-18T00:00:00.000Z	0	Low	1
Row29	Warm Cool	Y	1	6	3	1.941	2.001 188	1.943	2	2010-12-20T00:00:00.000Z	0	Low	1
Row30	Hot Water...	Y	1	9	4	1.922	2.004	1.969	2.5	2011-09-08T00:00:00.000Z	0	Low	1
Row31	Warm Cool	Y	1	7	4	1.930	2.006	1.961	2	2008-10-28T00:00:00.000Z	0	Low	1
Row32	Warm Cool	Y	1	4	1	1.900	2.010	1.964	2	2012-04-11T00:00:00.000Z	0	Low	1
Row33	Hot Water...	N	1	5	3	1.912	2.001 188	1.957	2	2012-04-25T00:00:00.000Z	0	Low	1
Row34	Warm Cool	Y	1	5	3	2.005	2.001 188	2.009	2	2005-07-25T00:00:00.000Z	0	Low	1

4. Bar chart

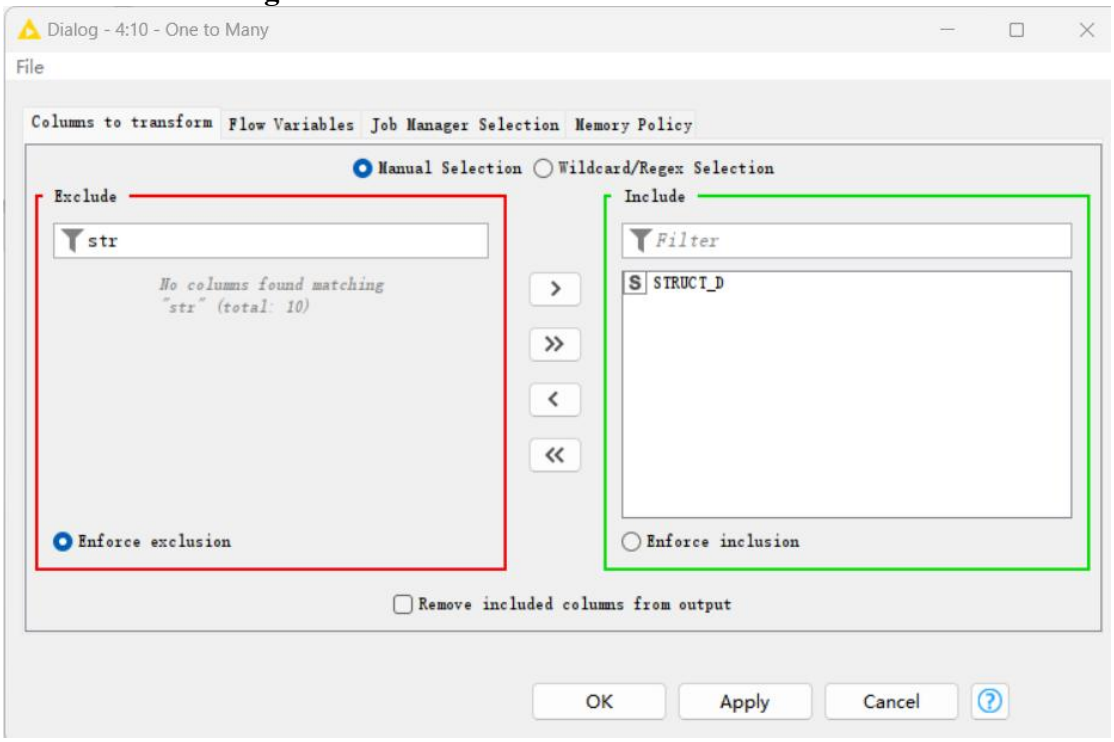


5. Binarisation

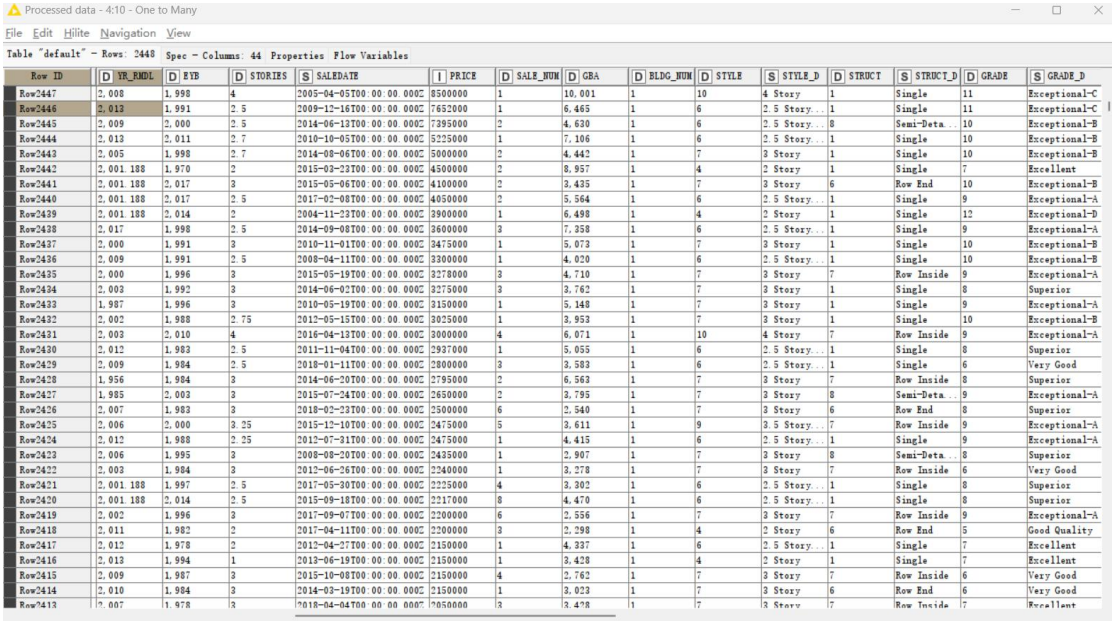
Binarisation is the process of converting continuous values into binary data. This method is based on a threshold and sets all values less than or equal to that threshold to 0 and all values greater than that threshold to 1. Using “One to many” to binarise the "STRUCT_D" variable [with values "0" or "1"] and “Excel Writer” to output result.



2. Parameter setting



3. Result of Binarisation：


Processed data - 4:10 - One to Many

File Edit Hilite Navigation View

Table "default" – Rows: 2448 Spec – Columns: 44 Properties Flow Variables

Row ID	YR_RMDL	EYB	STORIES	SALEDATE	PRICE	SALE_NUM	GBA	BLDG_NUM	STYLE	STYLE_D	STRUCT	STRUCT_D	GRADE	GRADE_D
Row2447	2,008	1,998	4	2005-04-05T00:00:00.000Z	8500000	1	10,001	1	10	4 Story	1	Single	11	Exceptional-C
Row2446	2,013	1,991	2.5	2009-12-16T00:00:00.000Z	7652000	1	6,465	1	6	2.5 Story...	1	Single	11	Exceptional-C
Row2445	2,009	2,000	2.5	2014-06-13T00:00:00.000Z	7395000	2	4,630	1	6	2.5 Story...	8	Semi-Deta...	10	Exceptional-B
Row2444	2,013	2,011	2.7	2010-10-05T00:00:00.000Z	5225000	1	7,106	1	6	2.5 Story...	1	Single	10	Exceptional-B
Row2443	2,005	1,998	2.7	2014-08-06T00:00:00.000Z	5000000	2	4,442	1	7	3 Story	1	Single	10	Exceptional-B
Row2442	2,001.188	1,970	2	2015-03-23T00:00:00.000Z	4500000	2	8,957	1	4	2 Story	1	Single	7	Excellent
Row2441	2,001.188	2,017	3	2015-05-06T00:00:00.000Z	4100000	2	3,435	1	7	3 Story	6	Row End	10	Exceptional-B
Row2440	2,001.188	2,017	2.5	2017-02-08T00:00:00.000Z	4050000	2	5,564	1	6	2.5 Story...	1	Single	9	Exceptional-A
Row2439	2,001.188	2,014	2	2004-11-23T00:00:00.000Z	3900000	1	6,498	1	4	2 Story	1	Single	12	Exceptional-D
Row2438	2,017	1,998	2.5	2014-09-08T00:00:00.000Z	3600000	3	7,358	1	6	2.5 Story...	1	Single	9	Exceptional-A
Row2437	2,000	1,991	3	2010-11-01T00:00:00.000Z	3475000	1	5,073	1	7	3 Story	1	Single	10	Exceptional-B
Row2436	2,009	1,991	2.5	2008-04-11T00:00:00.000Z	3300000	1	4,020	1	6	2.5 Story...	1	Single	10	Exceptional-B
Row2435	2,000	1,996	3	2015-05-19T00:00:00.000Z	3278000	3	4,710	1	7	3 Story	7	Row Inside	9	Exceptional-A
Row2434	2,003	1,992	3	2014-06-02T00:00:00.000Z	3275000	3	3,762	1	7	3 Story	1	Single	8	Superior
Row2433	1,987	1,996	3	2010-05-19T00:00:00.000Z	3150000	1	5,148	1	7	3 Story	1	Single	9	Exceptional-A
Row2432	2,002	1,988	2.75	2012-05-15T00:00:00.000Z	3025000	1	3,953	1	7	3 Story	1	Single	10	Exceptional-B
Row2431	2,003	2,010	4	2016-04-13T00:00:00.000Z	3000000	4	6,071	1	10	4 Story	7	Row Inside	9	Exceptional-A
Row2430	2,012	1,983	2.5	2011-11-04T00:00:00.000Z	2937000	1	5,055	1	6	2.5 Story...	1	Single	8	Superior
Row2429	2,009	1,984	2.5	2018-01-11T00:00:00.000Z	2800000	3	3,583	1	6	2.5 Story...	1	Single	6	Very Good
Row2428	1,956	1,984	3	2014-06-20T00:00:00.000Z	2795000	2	6,563	1	7	3 Story	7	Row Inside	8	Superior
Row2427	1,985	2,003	3	2015-07-24T00:00:00.000Z	2650000	2	3,795	1	7	3 Story	8	Semi-Deta...	9	Exceptional-A
Row2426	2,007	1,983	3	2018-02-23T00:00:00.000Z	2500000	6	2,540	1	7	3 Story	6	Row End	8	Superior
Row2425	2,006	2,000	3.25	2015-12-10T00:00:00.000Z	2475000	5	3,611	1	9	3.5 Story...	7	Row Inside	9	Exceptional-A
Row2424	2,012	1,988	2.25	2012-07-31T00:00:00.000Z	2475000	1	4,415	1	6	2.5 Story...	1	Single	9	Exceptional-A
Row2423	2,006	1,995	3	2008-08-20T00:00:00.000Z	2435000	1	2,907	1	7	3 Story	8	Semi-Deta...	8	Superior
Row2422	2,003	1,984	3	2012-06-26T00:00:00.000Z	2240000	1	3,278	1	7	3 Story	7	Row Inside	6	Very Good
Row2421	2,001.188	1,997	2.5	2017-05-30T00:00:00.000Z	2225000	4	3,302	1	6	2.5 Story...	1	Single	8	Superior
Row2420	2,001.188	2,014	2.5	2015-09-18T00:00:00.000Z	2217000	8	4,470	1	6	2.5 Story...	1	Single	8	Superior
Row2419	2,002	1,996	3	2017-09-07T00:00:00.000Z	2200000	6	2,556	1	7	3 Story	7	Row Inside	9	Exceptional-A
Row2418	2,011	1,982	2	2017-04-11T00:00:00.000Z	2200000	3	2,298	1	4	2 Story	6	Row End	5	Good Quality
Row2417	2,012	1,978	2	2012-04-27T00:00:00.000Z	2150000	1	4,337	1	6	2.5 Story...	1	Single	7	Excellent
Row2416	2,013	1,994	1	2013-06-19T00:00:00.000Z	2150000	1	3,428	1	4	2 Story	1	Single	7	Excellent
Row2415	2,009	1,987	3	2015-10-08T00:00:00.000Z	2150000	4	2,762	1	7	3 Story	7	Row Inside	6	Very Good
Row2414	2,010	1,984	3	2014-03-19T00:00:00.000Z	2150000	1	3,023	1	7	3 Story	6	Row End	6	Very Good
Row2413	2,007	1,978	3	2018-04-04T00:00:00.000Z	2050000	3	3,478	1	7	3 Story	7	Row Inside	7	Excellent

1C Summary:

The most important findings of this report include the following:

1. The same data that appears in the data may be duplicated. There are 551 empty data and 795 zero data in the price attribute, resulting in the average median, maximum and minimum values being greatly affected by extreme values, which may be incomplete or wrong in collecting data. All of the data collected were sold on January 1, 1900, and the selling price was either empty or zero, possibly anomalous.

Using the mean or median method in the statistics instead of vacancies can only lead to untruthful and misguided data models. These two properties should be subjected to more rigorous exploration and visual inspection in the future.

2. In the process of identifying correlations, it was found that Style and stories have a strong linear relationship, bathroom and Room have a certain relationship, and grade and price also have a certain association, and the correlation of these attributes should be more studied and more rigorous visual inspection in the future.

3. In the settings, there is a row A Y B is 2019, but EY B years old data is zero, but E YB time will not be earlier than A Y B and most of the other data in this message is blank, so we need to determine whether this row of data is caused by human factors caused by erroneous data.

4. By utilising hierarchical clustering , we found that these data are optimally classified by K-means. In the future, more in-depth research can be carried out by adjusting the model parameters.

5. During the visualization process, it was found that the price data was not very strong correlated with other data, and could be affected by extreme values and error values. In the future, more careful exploration and rigorous collection and investigation of such data will be required.

6. Through the visual chart, it can be inferred that the price is related to the selling time, and the later the selling time, the higher the probability of obtaining a high price. And price has a certain relationship with GBA, and when GBA is larger, the higher the probability of price. In the future, more careful exploration and rigorous collection and investigation of such data will be required.