Data analysis by using tools

Hojin Choi

I. Introduction

II. Initial data exploration

a. Identify attribute type
b. Summarize attribute

III. Data preprocessing

a. Basic statistics
b. Binning
c. Normalization
d. Discretise
e. Binarise

IV. Summary

I. Introduction

The reason for implementing this project is to analyze data using tools such as KNIME. I use R Studio to analyze given data. In order to analyze the given data, the following process is necessary - Data exploration, Data preprocessing, and Results. First step is data exploration. We need to identify attribute type – nominal, ordinal, interval, ratio and Identify the values of the summarizing properties for the attributes. After that, data processing is performed using KNIME etc. Second step is data preprocessing. It is a procedure that is carried out to obtain a sophisticated predictive analysis model as a necessary procedure before data analysis. The reason why data preprocessing is necessary is to examine whether the collected data has any missing parts, errors, or processing parts in data processing. After data preprocessing, we can get results and data analysis will be conducted.

Accordingly, I began to analyze the given data while complying with the above process. First step, identify attribute type.

II. Initial data exploration

a. Identify attribute type

Before analyzing data, we need to identify attribute type. This data has 39 attributes and 2003 instances. Each column can be set to the following type. Use R studio function “sum(is.na)” to know the number of NULLs and use “typeof()” function to look data types.

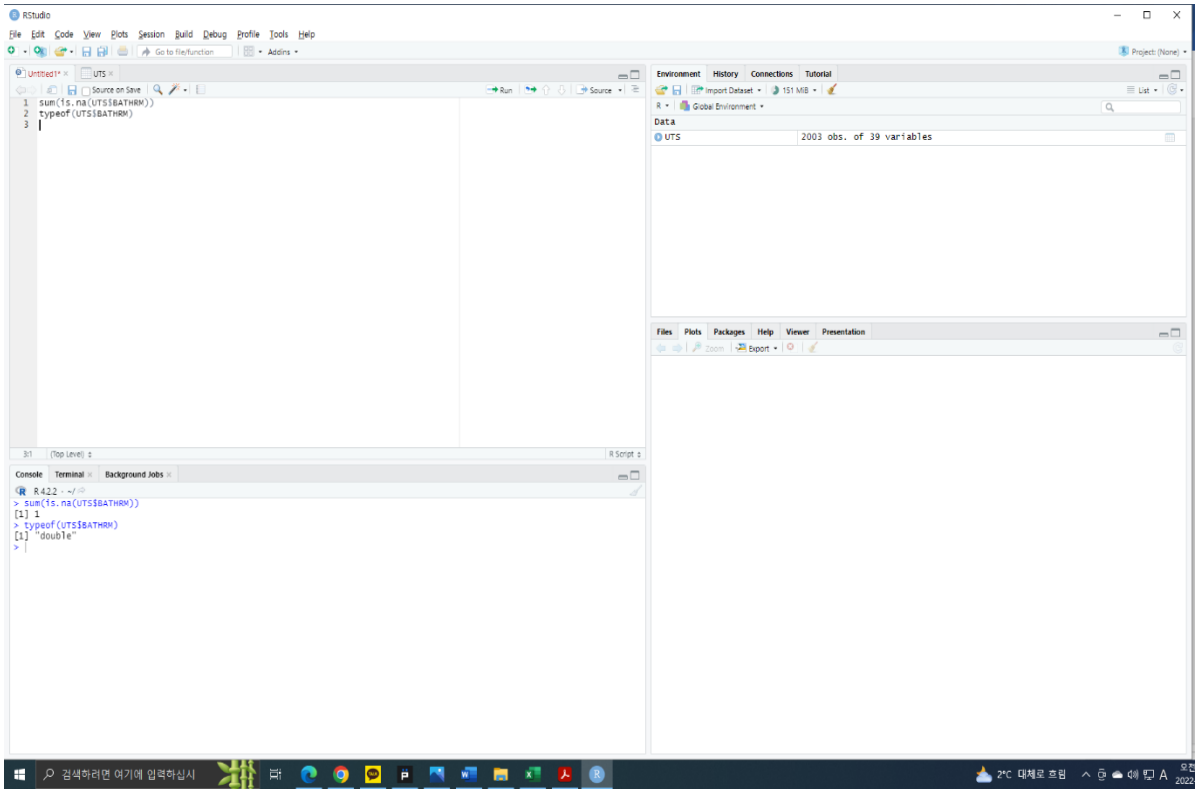


Figure 1 : R studio Function – sum(is.na) & typeof()

As a result of using the function, the following values were obtained.

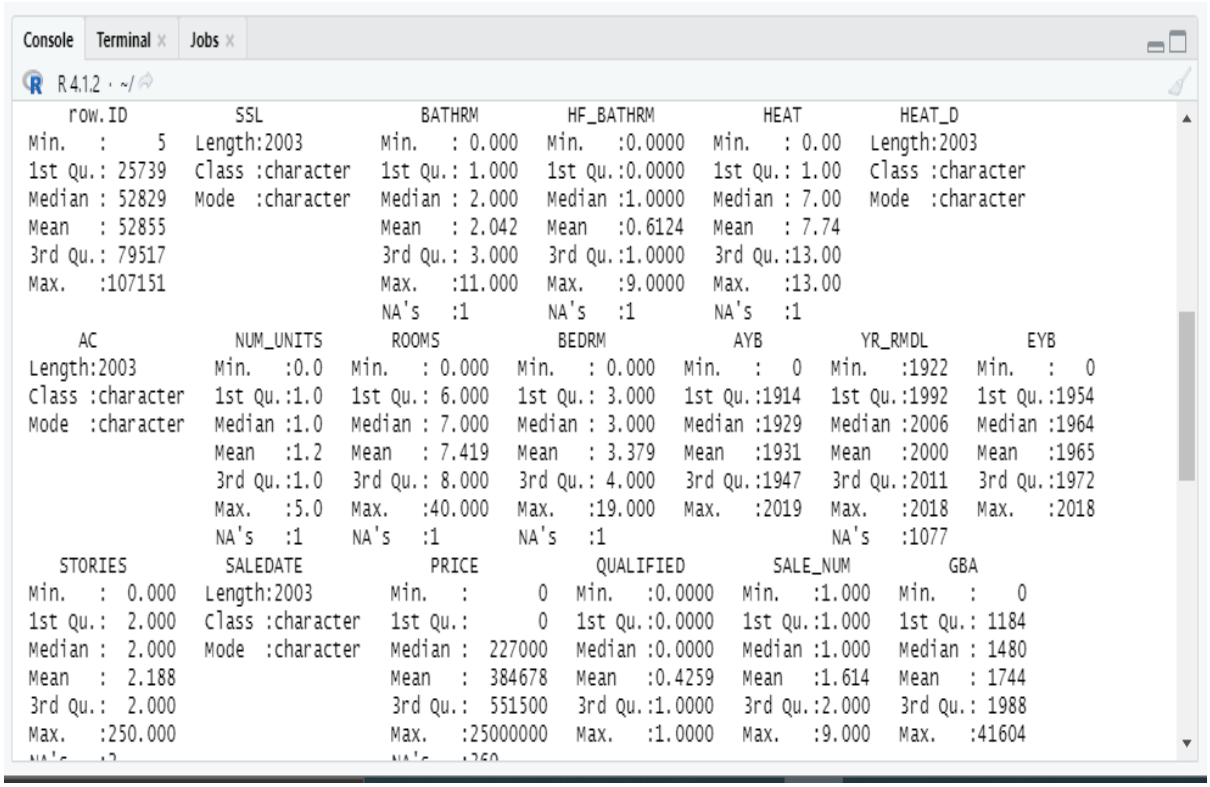
Column	NULL	Not NULL	Data Type
Row.ID	0	2003	Integer
SSL	0	2003	Character
BATHRM	1	2002	Double
HF_BATHRM	1	2002	Double
HEAT	1	2002	Double
HEAT_D	1	2002	Character
AC	0	2003	Double
NUM_UNITS	1	2002	Double
ROOMS	1	2002	Double
BEDRM	1	2002	Double
AYB	0	2003	Integer
YR_RDML	1077	926	Double
EYB	0	2003	Character
STORIES	2	2001	Double
SALEDATE	0	2003	Character
PRICE	369	1634	Double
QUALIFIED	0	2003	Integer
SALE_NUM	0	2003	Integer
GBA	0	2003	Integer
BLDG_NUM	0	2003	Integer
STYLE	1	2002	Double
STYLE_D	1	2002	Character
STRUCT	1	2002	Double
STRUCT_D	1	2002	Character
GRADE	1	2002	Double
GRADE_D	1	2002	Character
CNDTN	1	2002	Double
CNDTN_D	1	2002	Character
EXTWALL	1	2002	Double
EXTWALL_D	1	2002	Character
ROOF	1	2002	Double
ROOF_D	1	2002	Character
INTWALL	1	2002	float64
INTWALL_D	1	2002	object
KITCHENS	1	2002	float64
FIREPLACES	1	2002	float64
USECODE	0	2003	int64
LANDAREA	0	2003	int64
GIS_LAST_MOD_DTTM	0	2003	object

Table 1 : Dataset Information

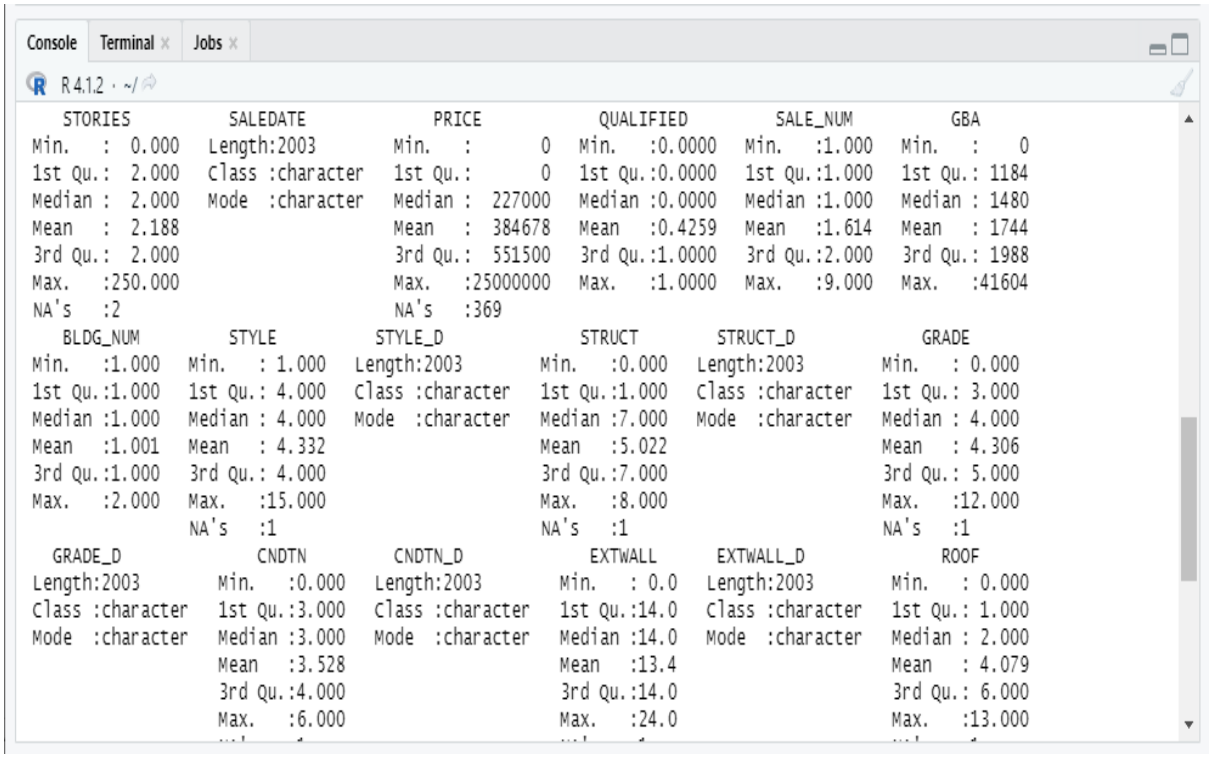
1. Two types of variables – numeric(integer, double), character
2. Some columns have “NULL” value → Missing value exists

b. Summarize attribute

We can easily obtain statistical information about the data by using the tool. Typical examples include mean, std, min, max, and quartiles. I found by using R studio and summarized them in a table. The results of data only include numeric data because we can't calculate character type.



```
Console  Terminal  Jobs
R 4.1.2 · ~/
     row.ID          SSL               BATHRM         HF_BATHRM          HEAT          HEAT_D
Min.   :     5   Length:2003        Min.   : 0.000   Min.   :0.0000   Min.   : 0.00   Length:2003
1st Qu.: 25739   Class :character   1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.: 1.00   Class :character
Median : 52829   Mode  :character   Median : 2.000   Median :1.0000   Median : 7.00   Mode  :character
Mean   : 52855                      Mean   : 2.042   Mean   :0.6124   Mean   : 7.74
3rd Qu.: 79517                      3rd Qu.: 3.000   3rd Qu.:1.0000   3rd Qu.:13.00
Max.   :107151                      Max.   :11.000   Max.   :9.0000   Max.   :13.00
                                    NA's   :1        NA's   :1        NA's   :1
      AC              NUM_UNITS         ROOMS            BEDRM            AYB          YR_RMDL          EYB
Length:2003        Min.   :0.0   Min.   : 0.000   Min.   : 0.000   Min.   :   0   Min.   :1922   Min.   :   0
Class :character   1st Qu.:1.0   1st Qu.: 6.000   1st Qu.: 3.000   1st Qu.:1914   1st Qu.:1992   1st Qu.:1954
Mode  :character   Median :1.0   Median : 7.000   Median : 3.000   Median :1929   Median :2006   Median :1964
                   Mean   :1.2   Mean   : 7.419   Mean   : 3.379   Mean   :1931   Mean   :2000   Mean   :1965
                   3rd Qu.:1.0   3rd Qu.: 8.000   3rd Qu.: 4.000   3rd Qu.:1947   3rd Qu.:2011   3rd Qu.:1972
                   Max.   :5.0   Max.   :40.000   Max.   :19.000   Max.   :2019   Max.   :2018   Max.   :2018
                   NA's   :1     NA's   :1        NA's   :1                       NA's   :1077
   STORIES         SALEDATE             PRICE          QUALIFIED         SALE_NUM          GBA
Min.   : 0.000   Length:2003        Min.   :       0   Min.   :0.0000   Min.   :1.000   Min.   :    0
1st Qu.: 2.000   Class :character   1st Qu.:       0   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 1184
Median : 2.000   Mode  :character   Median :  227000   Median :0.0000   Median :1.000   Median : 1480
Mean   : 2.188                      Mean   :  384678   Mean   :0.4259   Mean   :1.614   Mean   : 1744
3rd Qu.: 2.000                      3rd Qu.:  551500   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.: 1988
Max.   :250.000                     Max.   :25000000   Max.   :1.0000   Max.   :9.000   Max.   :41604
```



```
Console  Terminal  Jobs
R 4.1.2 · ~/
   STORIES         SALEDATE             PRICE          QUALIFIED         SALE_NUM          GBA
Min.   : 0.000   Length:2003        Min.   :       0   Min.   :0.0000   Min.   :1.000   Min.   :    0
1st Qu.: 2.000   Class :character   1st Qu.:       0   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 1184
Median : 2.000   Mode  :character   Median :  227000   Median :0.0000   Median :1.000   Median : 1480
Mean   : 2.188                      Mean   :  384678   Mean   :0.4259   Mean   :1.614   Mean   : 1744
3rd Qu.: 2.000                      3rd Qu.:  551500   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.: 1988
Max.   :250.000                     Max.   :25000000   Max.   :1.0000   Max.   :9.000   Max.   :41604
NA's   :2                           NA's   :369
   BLDG_NUM          STYLE          STYLE_D             STRUCT         STRUCT_D            GRADE
Min.   :1.000   Min.   : 1.000   Length:2003        Min.   :0.000   Length:2003        Min.   : 0.000
1st Qu.:1.000   1st Qu.: 4.000   Class :character   1st Qu.:1.000   Class :character   1st Qu.: 3.000
Median :1.000   Median : 4.000   Mode  :character   Median :7.000   Mode  :character   Median : 4.000
Mean   :1.001   Mean   : 4.332                      Mean   :5.022                      Mean   : 4.306
3rd Qu.:1.000   3rd Qu.: 4.000                      3rd Qu.:7.000                      3rd Qu.: 5.000
Max.   :2.000   Max.   :15.000                      Max.   :8.000                      Max.   :12.000
                NA's   :1                           NA's   :1                          NA's   :1
  GRADE_D             CNDTN         CNDTN_D            EXTWALL       EXTWALL_D            ROOF
Length:2003        Min.   :0.000   Length:2003        Min.   : 0.0   Length:2003        Min.   : 0.000
Class :character   1st Qu.:3.000   Class :character   1st Qu.:14.0   Class :character   1st Qu.: 1.000
Mode  :character   Median :3.000   Mode  :character   Median :14.0   Mode  :character   Median : 2.000
                   Mean   :3.528                      Mean   :13.4                      Mean   : 4.079
                   3rd Qu.:4.000                      3rd Qu.:14.0                      3rd Qu.: 6.000
                   Max.   :6.000                      Max.   :24.0                      Max.   :13.000
```

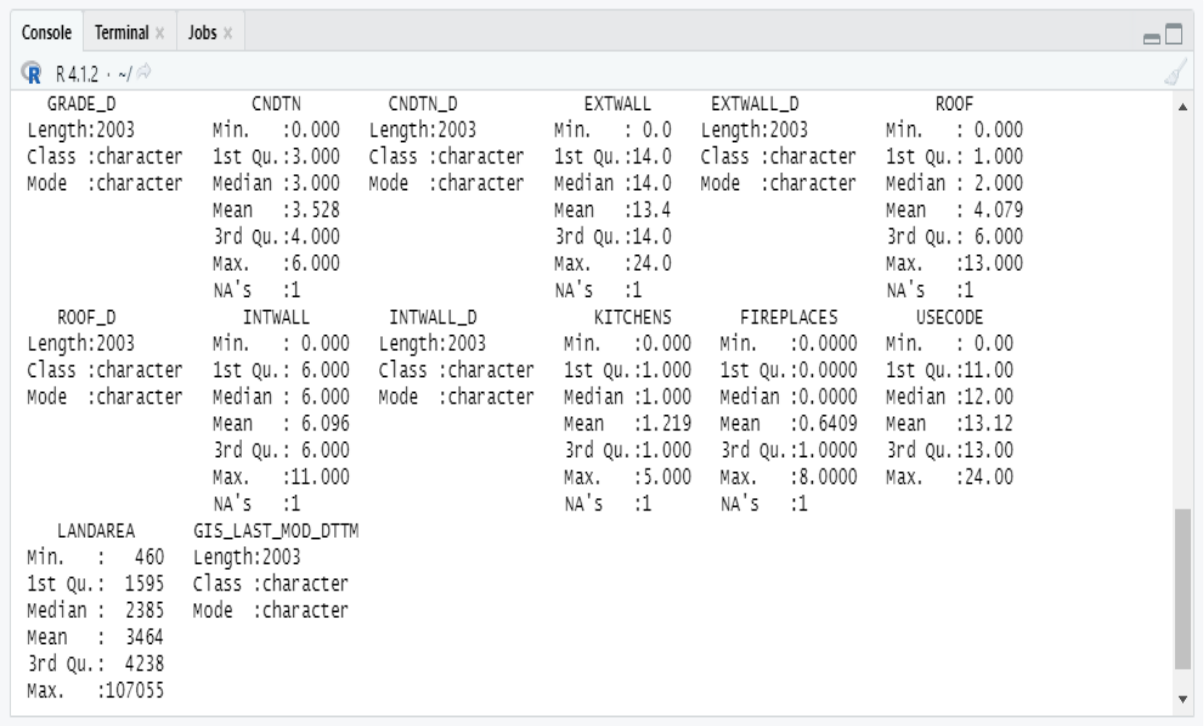


Figure 2 : R studio Function – summary()

The standard deviation was obtained as a function of sd(). Here, na.rm = true means that the missing value has been removed.

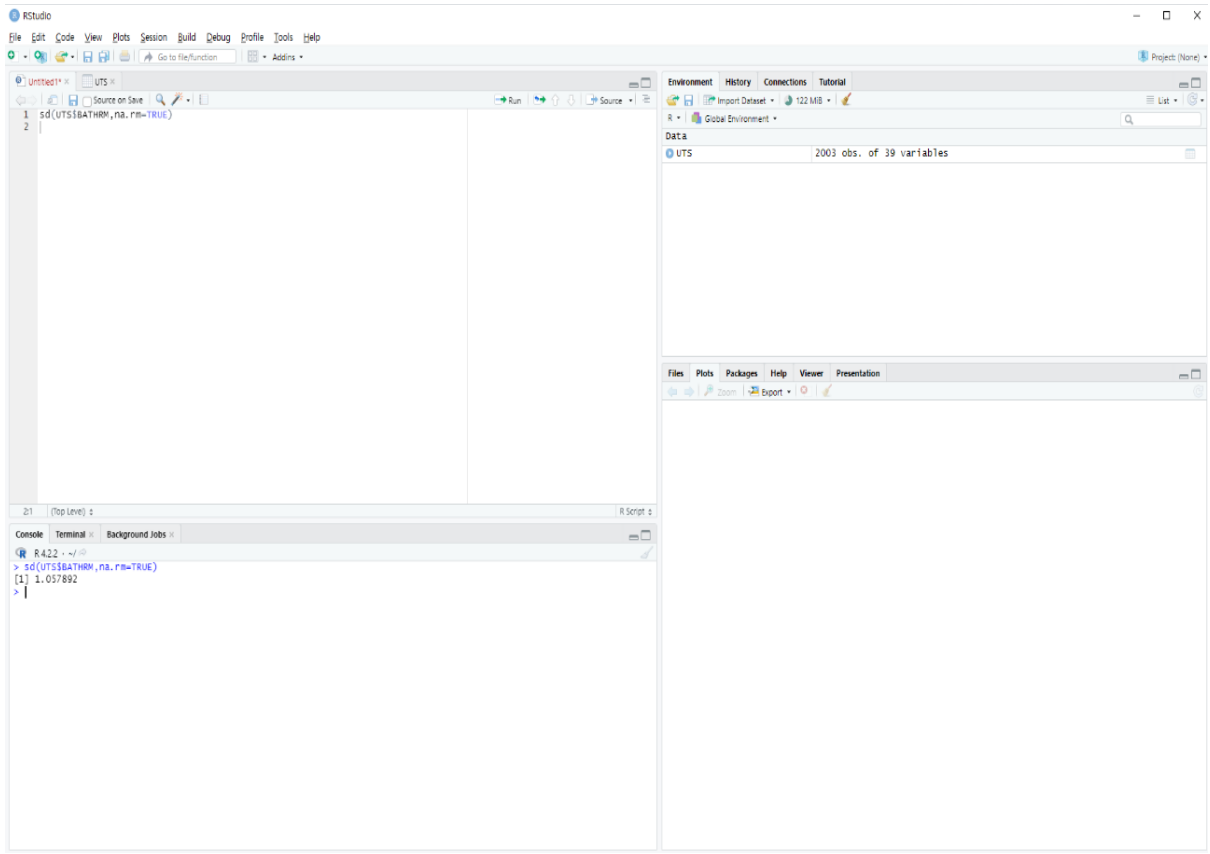


Figure 3 : R studio Function – sd()

As a result of using the function, the following values were obtained.

Column	mean	std	min	max	25%	50%	75%
row.ID	52855	31046.43	5	107151	25739	52829	79517
BATHRM	2.042	1.0579	0	11	1	2	3
HF_BATHRM	0.6124	0.6437	0	9	0	1	1
HEAT	7.74	4.9923	0	13	1	7	13
NUM_UNITS	1.2	0.6024	0	5	1	1	1
ROOMS	7.419	2.4973	0	40	6	7	8
BEDRM	3.379	1.1619	0	19	3	3	4
AYB	1931	80.2381	0	2019	1914	1929	1947
YR_RMDL	2000	15.9662	1922	2018	1992	2006	2011
EYB	1965	47.0105	0	2018	1954	1964	1972
STORIES	2.188	5.5583	0	250	2	2	2
PRICE	384678	781824.2	0	25000000	0	227000	551500
QUALIFIED	0.4259	0.4946	0	1	0	0	1
SALE_NUM	1.614	1.2565	1	9	1	1	2
GBA	1744	1280.754	0	41604	1184	1480	1988
BLDG_NUM	1.001	0.0316	1	2	1	1	1
STYLE	4.332	1.4899	1	15	4	4	4
STRUCT	5.022	2.8687	0	8	1	7	7
GRADE	4.306	1.3804	0	12	3	4	5
CNDTN	3.528	0.7302	0	6	3	3	4
EXTWALL	13.4	3.9957	0	24	14	14	14
ROOF	4.079	3.383	0	13	1	2	6
INTWALL	6.096	1.9021	0	11	6	6	6
KITCHENS	1.219	0.6133	0	5	1	1	1
FIREPLACES	0.6409	0.9263	0	8	0	0	1
USECODE	13.12	3.9712	0	24	11	12	13
LANDAREA	3464	4433.972	460	107055	1595	2385	4238

Table 2 : R studio summary results

Many items will be considered, but the most necessary information from the customer's point of view will be price information. Also, this assignment says “Use price attribute in data preprocessing.” This graph shows right tail test where price values exist near the left side. It means, zero or cheap prices are the mainstream and few are expensive.

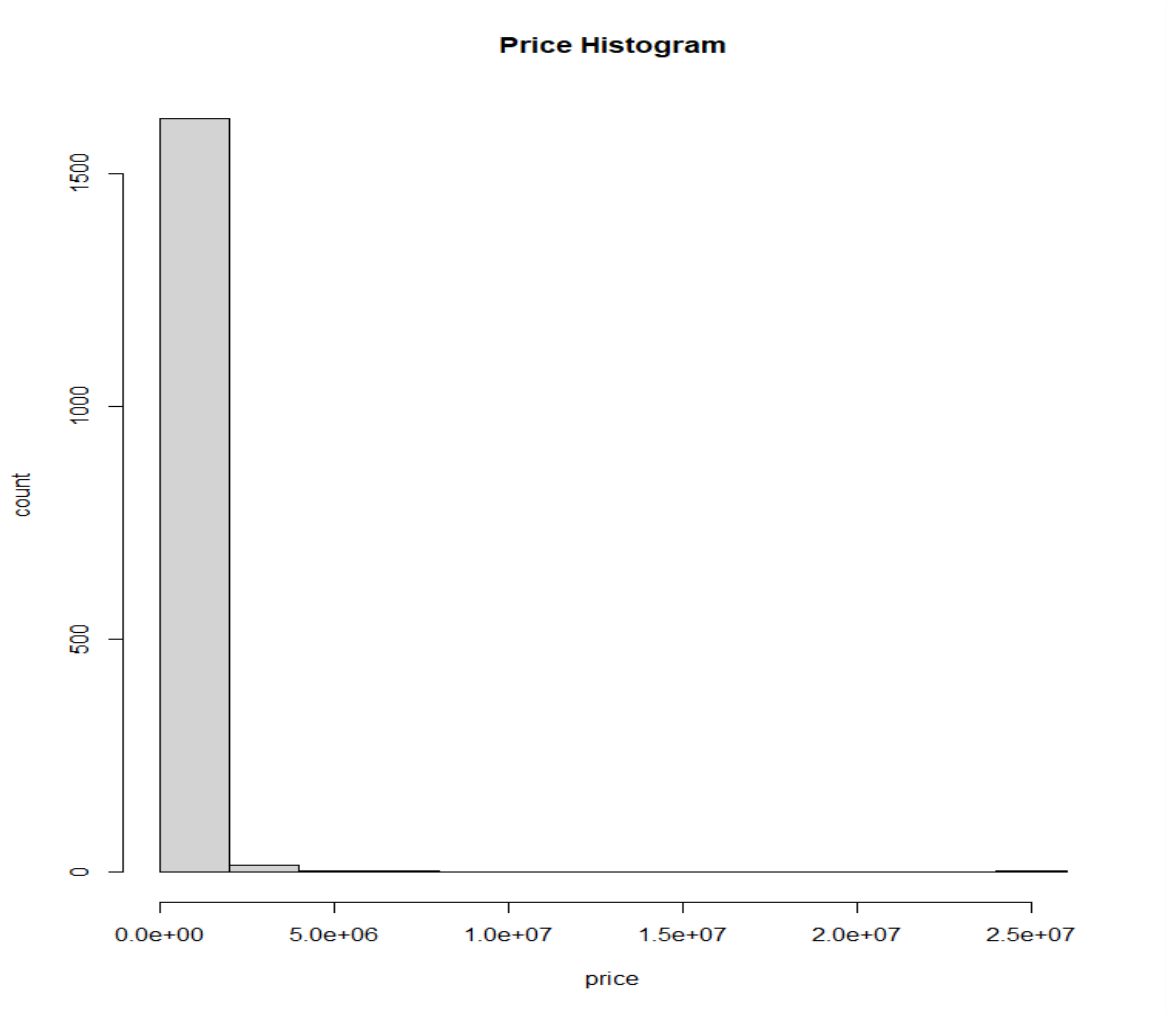


Figure 4 : R studio Histogram – hist()

III. Data Preprocessing

a. Basic statistics

The first thing to do is eliminate missing values. **U**se Excel and KNIME tools to accomplish data preprocessing. These tools are used because tools are convenient and easy to use.

1. Use Excel program

1) Delete "_D" attribute

"_D" means description of data. It is not necessary to analyze. What is needed immediately is numerical data analysis, and the explanation part can be viewed after the data analysis is finished.

2) Format "GIS_LAST_MOD_DTTM" : 0000-00-00

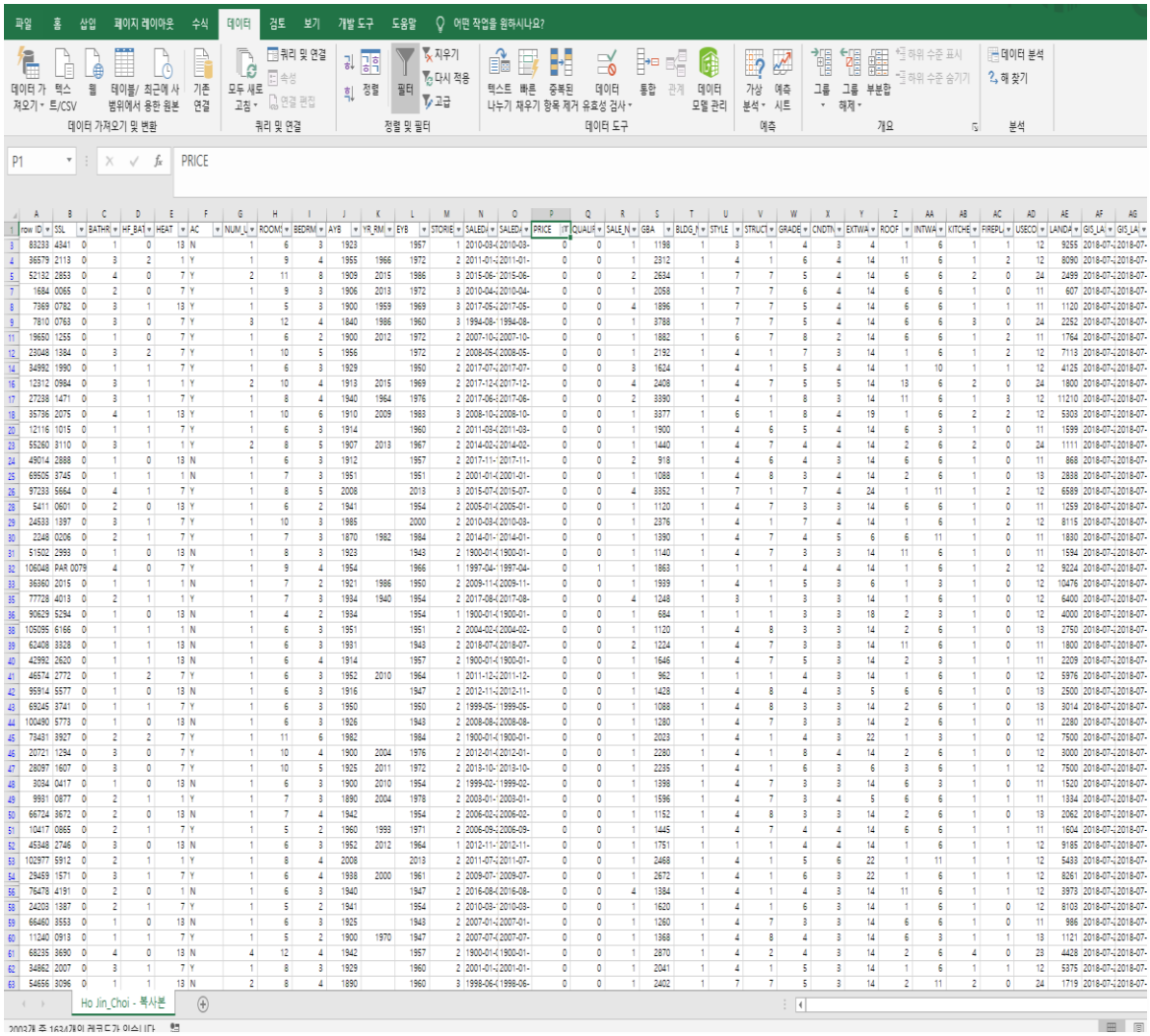
Use "left" function to perform date data. (remove time data : 00:00:00)

3) Convert instances to numeric

Use cell format(ctrl+1) and change into all numeric data.

4) Eliminate “PRICE” missing values

This project is being carried out to analyze the data of “PRICE” items. Removing missing values in the preprocessing process performed before data analysis is a must-go step. Use “filter” to remove blank data(NA) and sort by ascending order.



Sheet 1 : Data preprocessing results

2. Use KNIME tool

Excel tools are easy to use, but it’s not professional. There are limitations in analyzing complex data using Excel. Also, KNIME tools are more convenient then Excel. So, we use KNIME tools which enables professional and detailed data analysis, will be used.

1) Read csv file

Read the Excel file through CSV Reader.

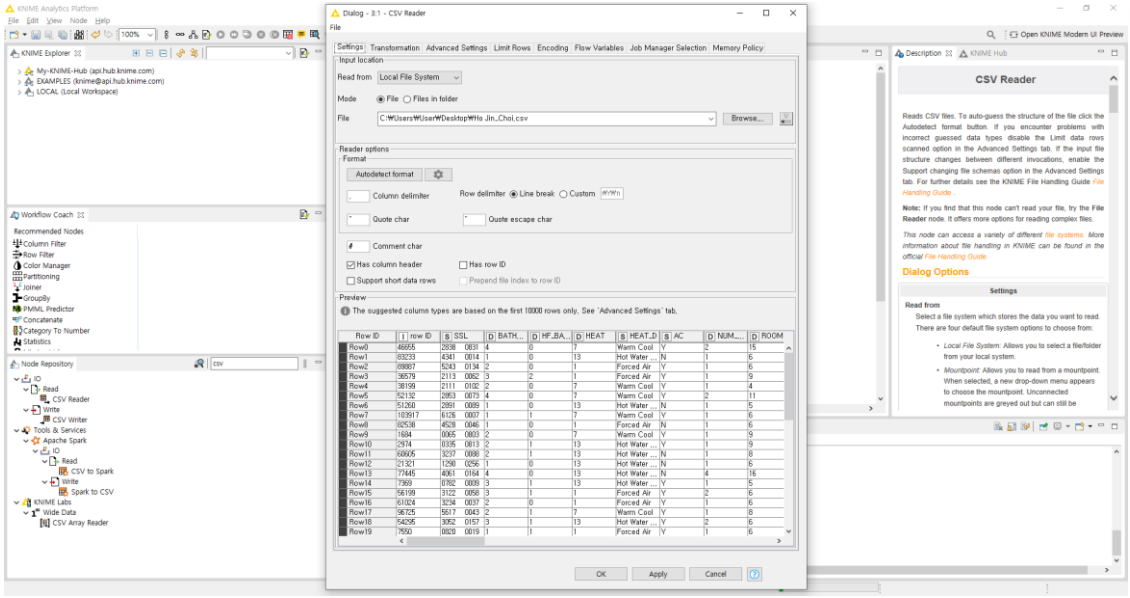
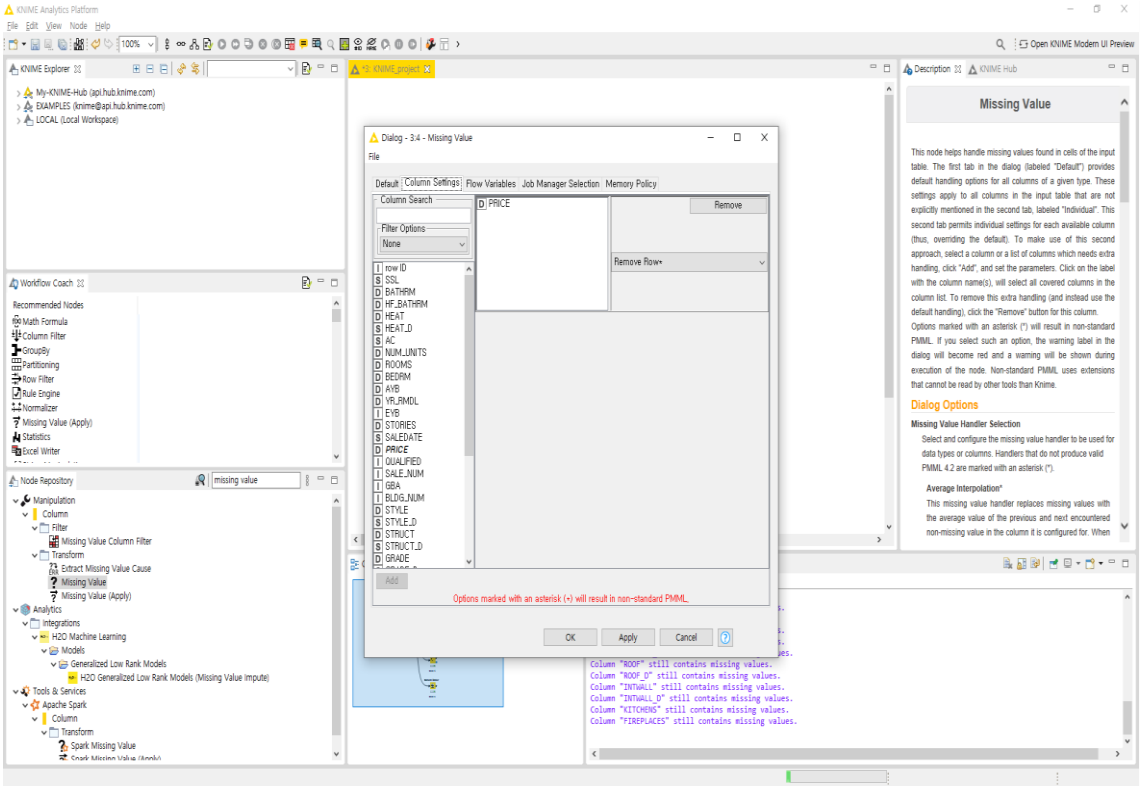


Figure 5 : KNIME – CSV Reader

2) Remove missing value

Perform the task using the “Missing value”. The row containing the missing value of the Price item was erased.



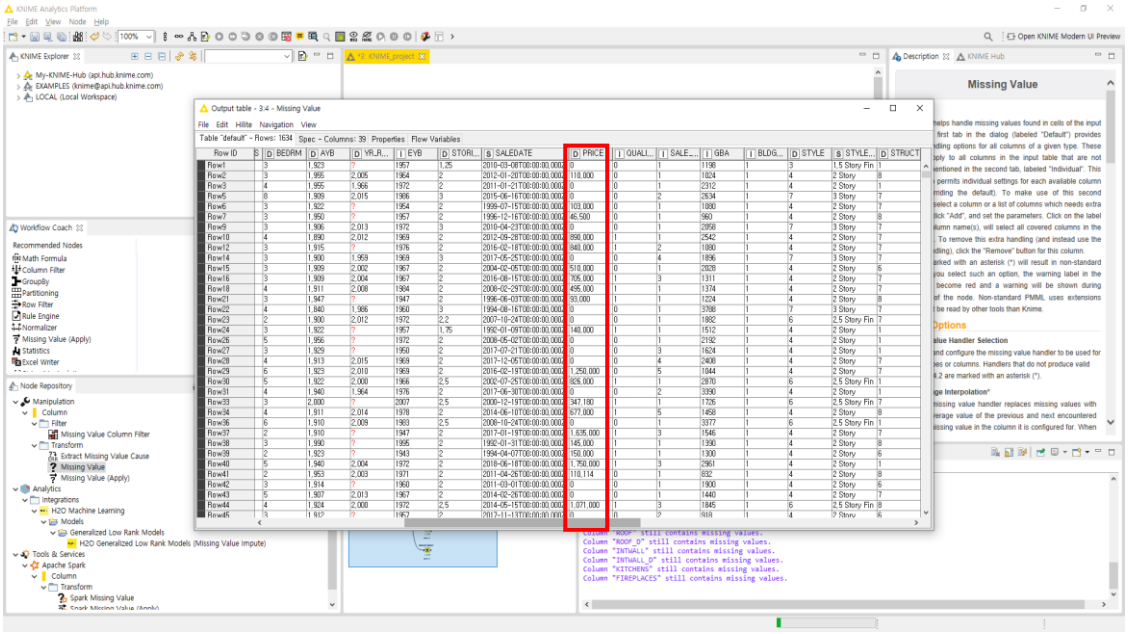


Figure 6 : KNIME – Missing value

3) Basic statistics

Perform the task using the “Statistics”. You can see several information by using KNIME.

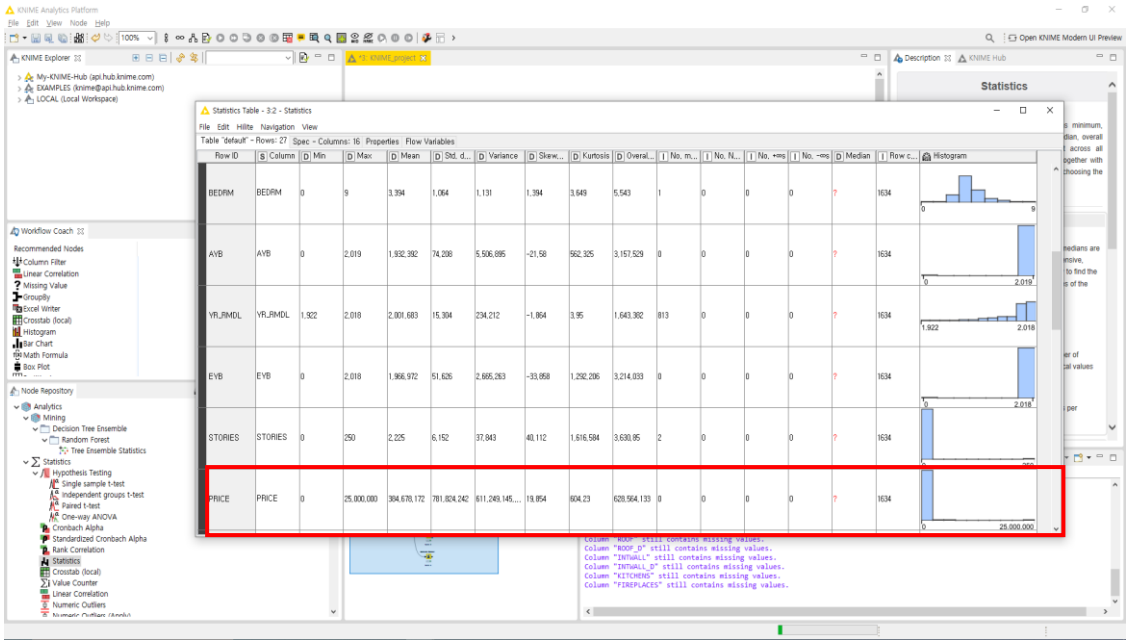


Figure 7 : KNIME – Statistics

As obtained in R studio, the mean, standard deviation, and quartile could be known, and the histogram could also be confirmed. In my opinion, KNIME tool is more convenient and useful than R studio. R studio needs more information to use function and harder to work. I was satisfied to be able to handle KNIME well while conducting this curriculum.

b. Binning

Binning means grouping a number of bin. The reason for using binning is to classify specific values by number or size. We use “auto-binner” to complete binning. It is a tool that can proceed with binning work very easily. If you specify the number of bins, KNIME tool automatically bins. There are two kinds of auto-binner – width, depth(frequency). I give number of bins = 5. The following results were obtained.

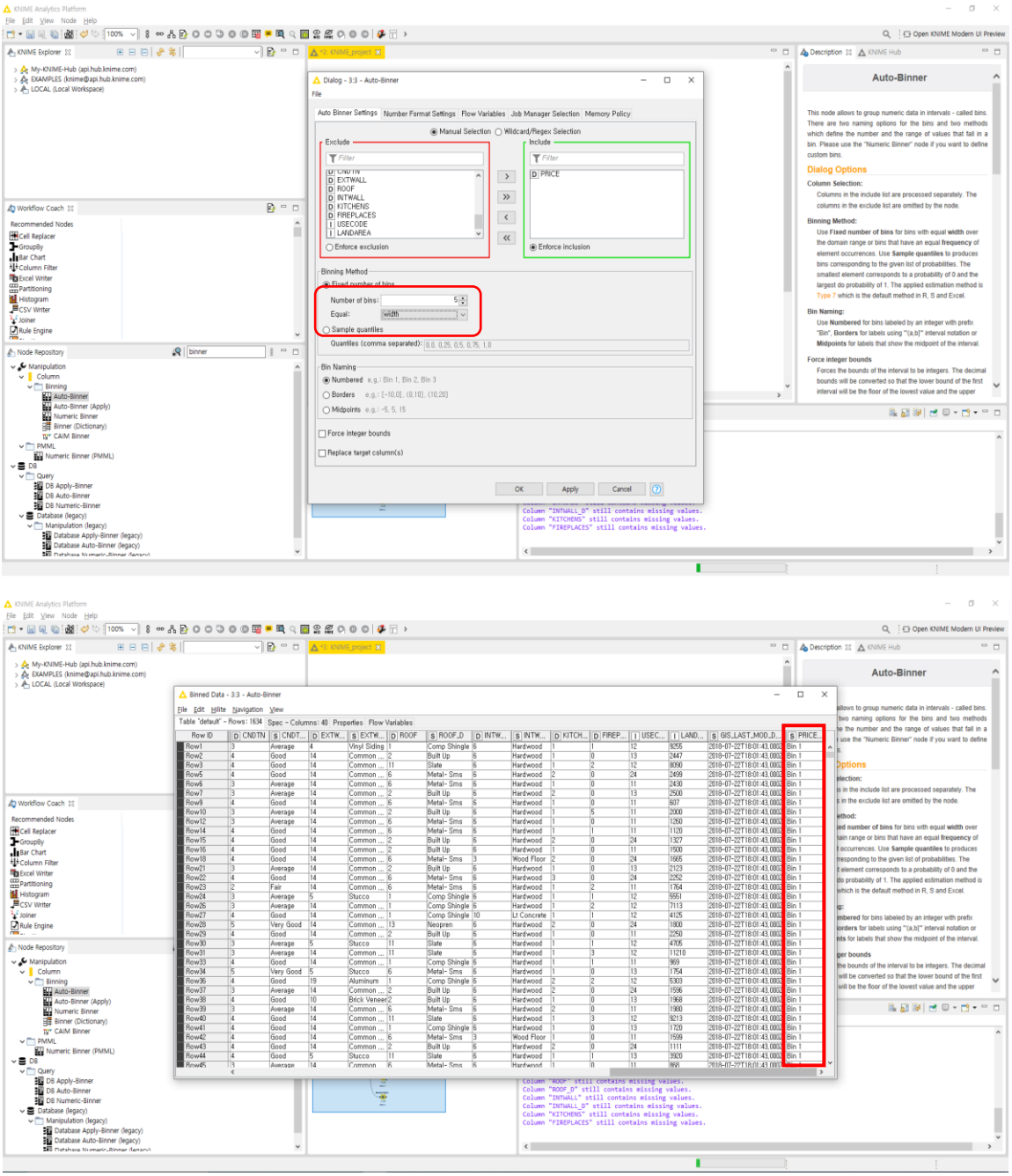


Figure 8 : KNIME – Auto-Binner(Width)

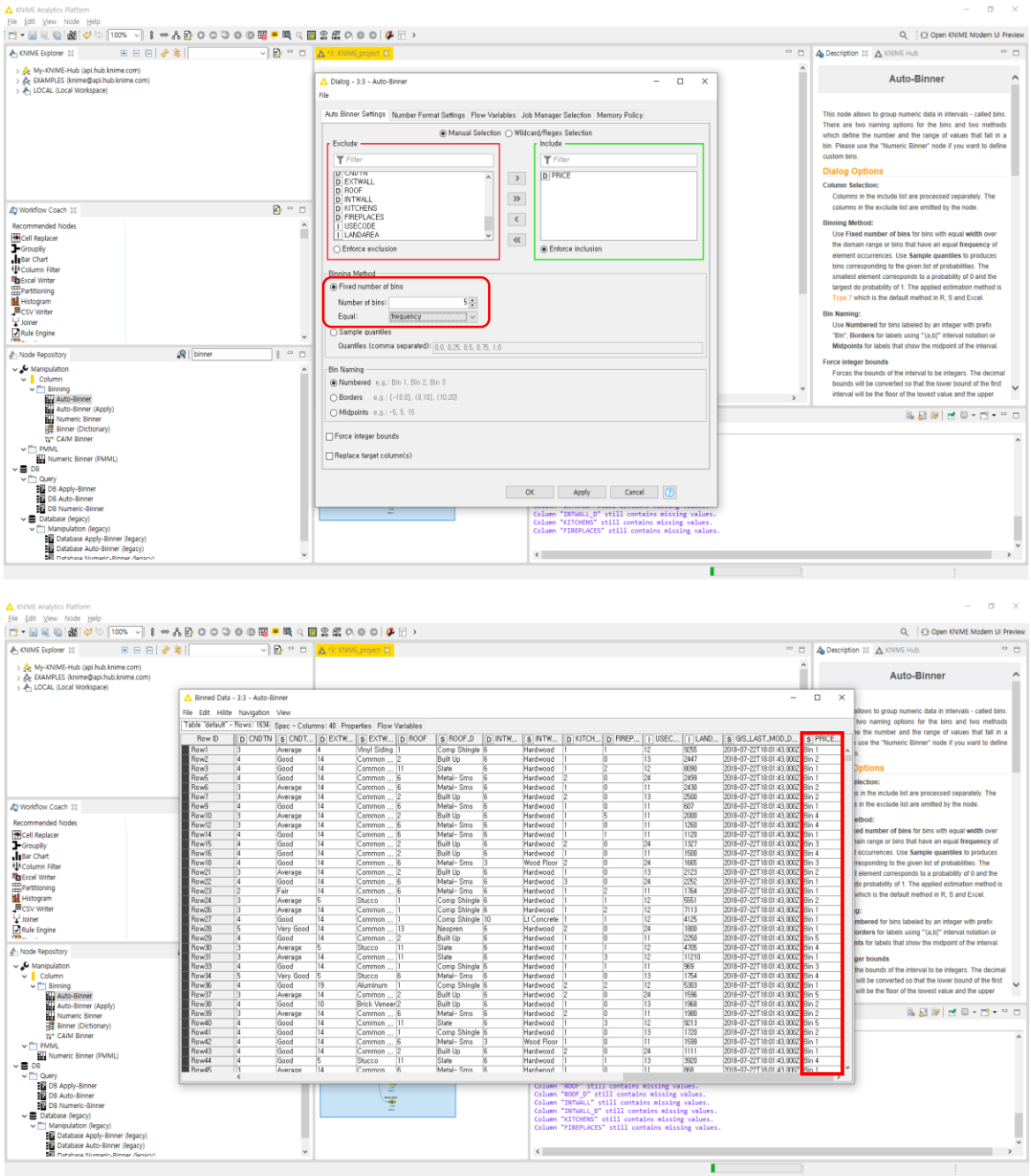


Figure 9 : KNIME – Auto-Binner(Frequency)

Since I give number of bins = 5, The results have Row1 to Row5. Row1 has the lowest number which means free or cheap and Row5 is the highest number which means so expensive.

c. Normalization

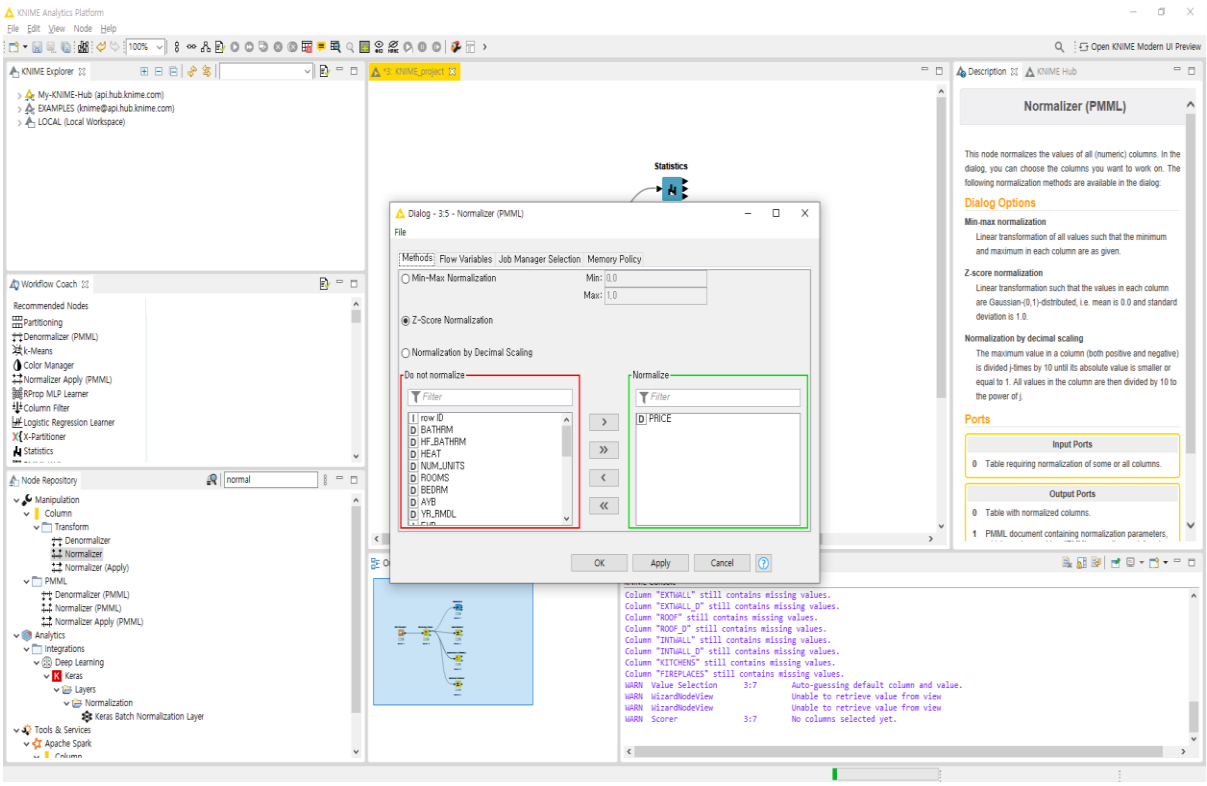
Normalization is the process of transforming data according to certain rules to make it easier to use. The reason for normalization is that a problem occurs when the scale of the feature of the data is severely different. There are two kinds of normalization – min-max normalization, z-score normalization.

1. min-max normalization

Minimum-maximum normalization is the most common method of normalizing data. For all features, convert each minimum value of 0, maximum value of 1, and other values between 0 and 1. Min-max normalization has advantage that all scales are same, but has the biggest disadvantage of not being able to handle outliers. So, I use z-score normalization to analyze given data.

2. z-score normalization

It means an operation to change the value of X to Z-score. The equation is as follows : (X - mean) / std. Unlike min-max normalization, it has the advantage of being able to solve outliers. If the value of feature matches the average, it will be normalized to 0, but if it is less than the average, it will be negative, and if it is greater than the average, it will be positive. The size of the negative and positive numbers calculated is determined by the std of the feature. So, if the standard deviation of the data is large (the value spreads widely), the normalized value approaches to zero.



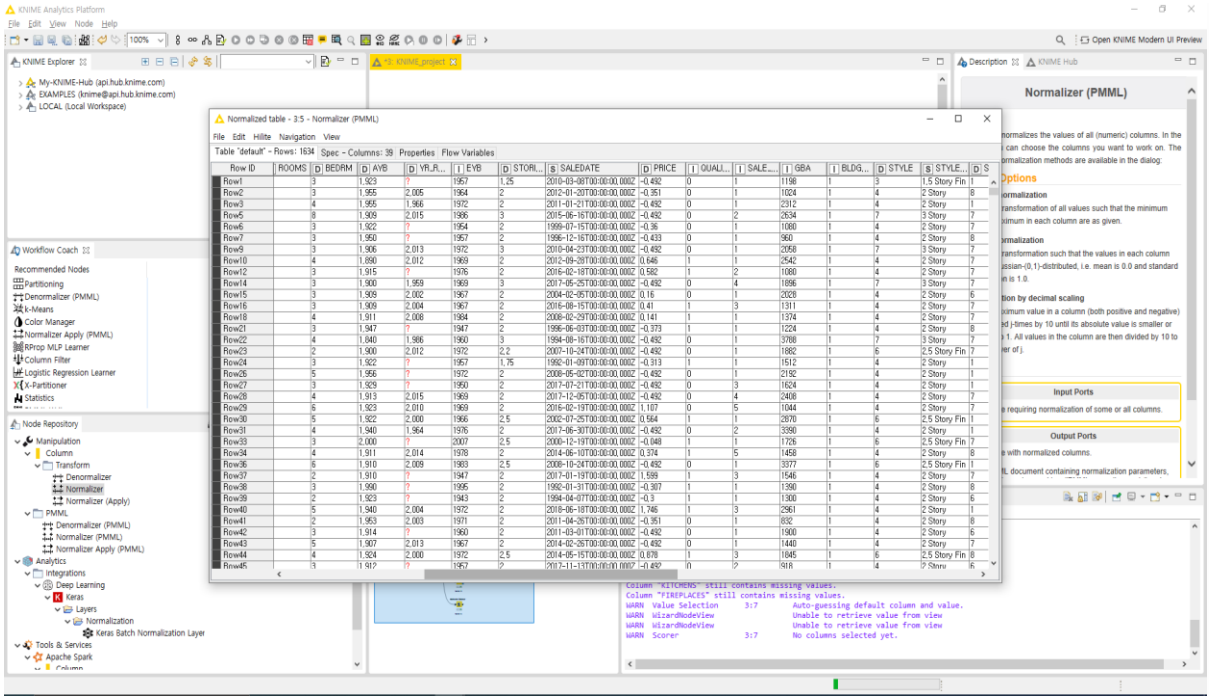
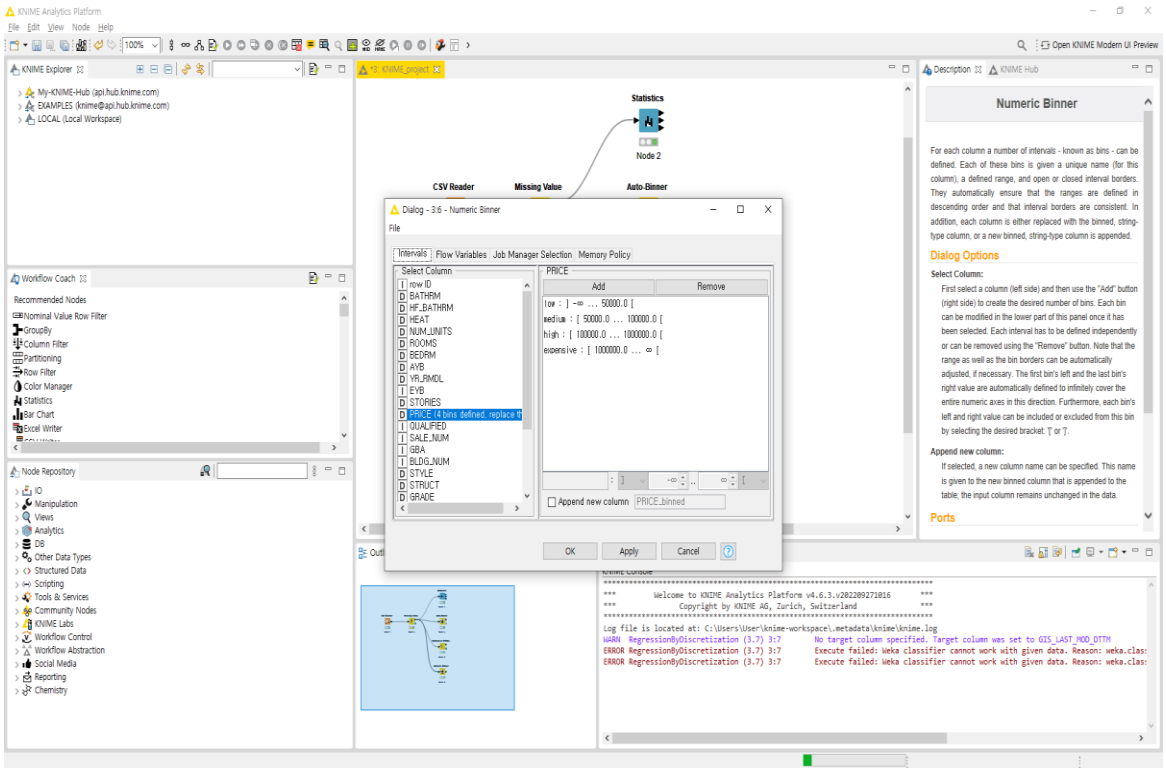


Figure 11 : KNIME – z-score normalization

d. Discretise

It is a system that performs binning by directly specifying a number interval. The number of bins should be described in detail in this system. Use “numeric-binning” to discretise following data. I designated the number range in four levels as follows – low, medium, high, and expensive.



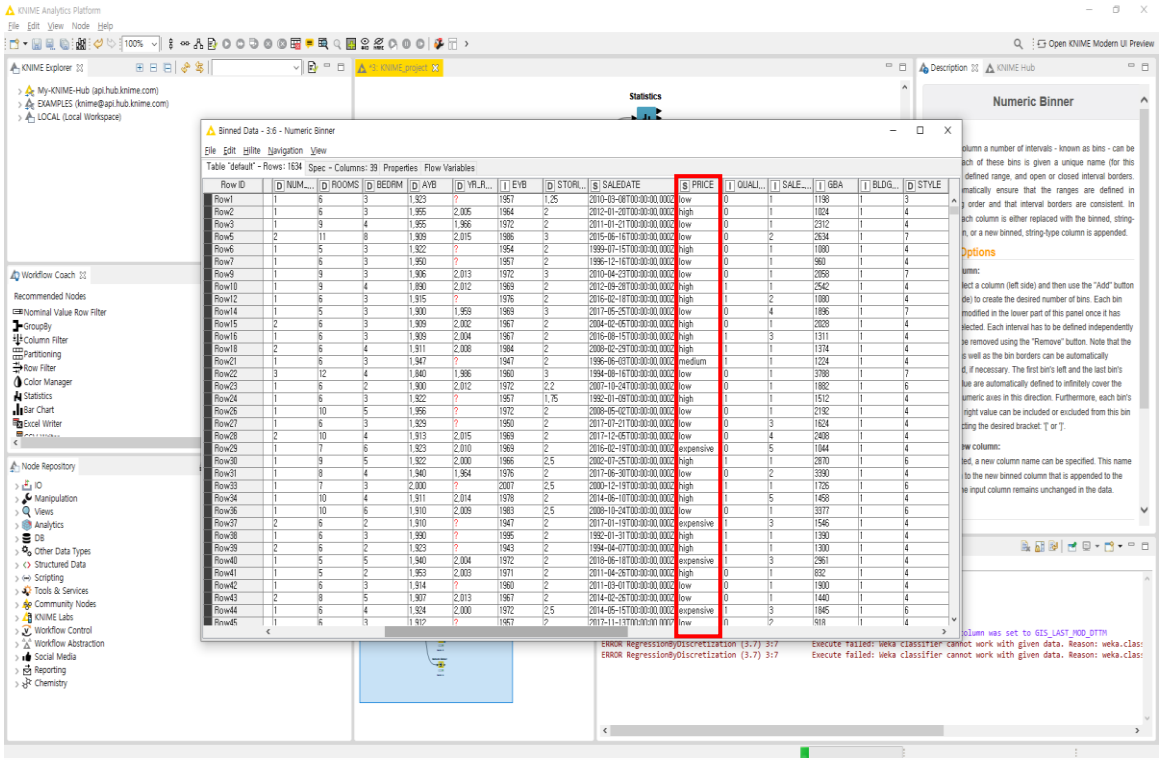
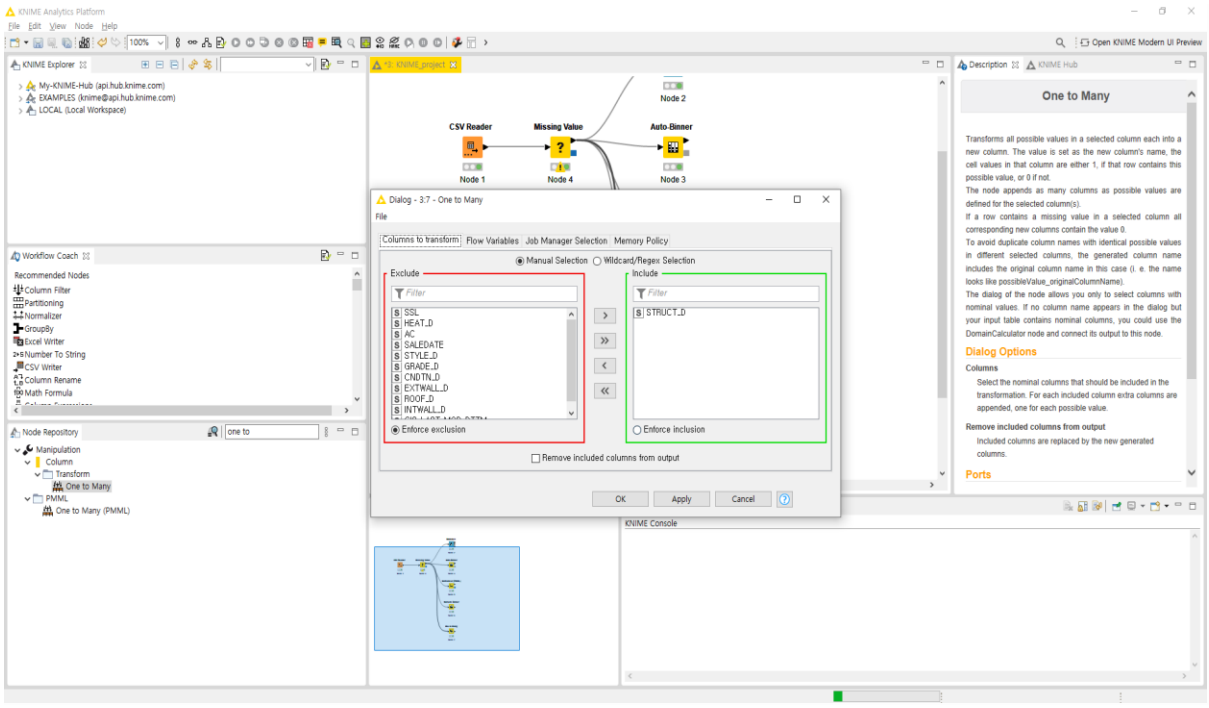


Figure 10 : KNIME – Numeric-Binner(Discretise)

e. Binarise

Since “STRUCT_D” is a character, not a numeric, it is difficult to use mathematically, and the effect of factors on the result value may be uneven. Therefore, we need to transform the character form into a numerical form. By binarise. Use “one to many” system to binarise “STRUCT_D.”



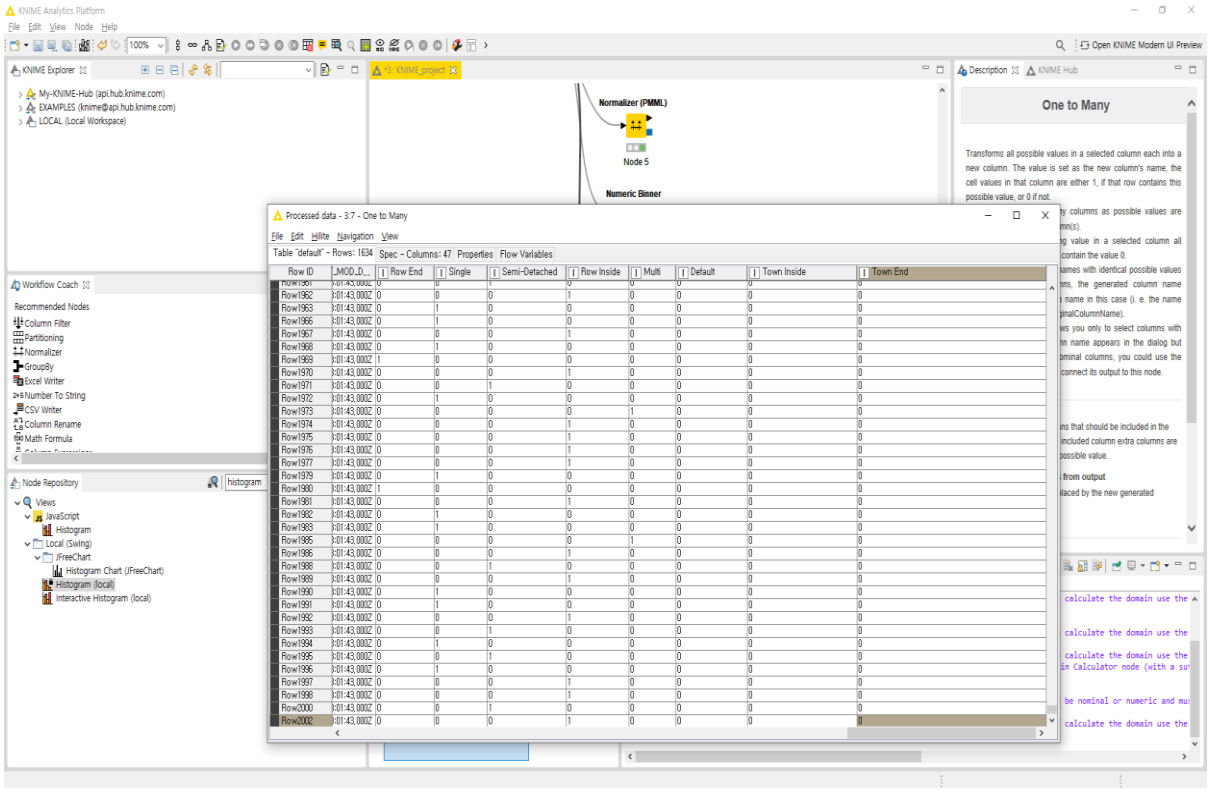
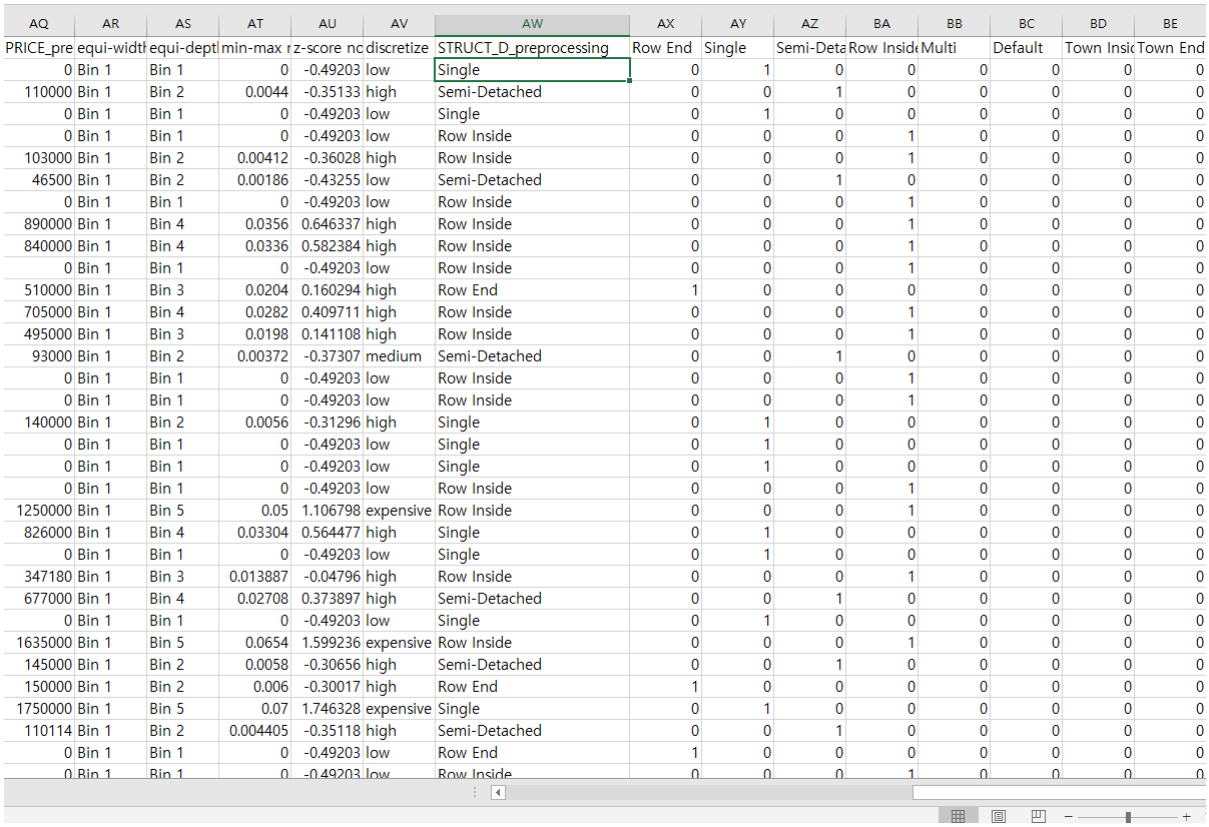


Figure 11 : KNIME – One to Many(Binarise)

Through binarise, we can check the data distribution of STRUCT_D.

IV. Conclusion

I put the previous values in the excel.



AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE
PRICE_pre	equi-width	equi-dept	min-max r	z-score nc	discretize	STRUCT_D_preprocessing	Row End	Single	Semi-Deta	Row Inside	Multi	Default	Town Insi	Town End
0	Bin 1	Bin 1	0	-0.49203	low	Single	0	1	0	0	0	0	0	0
110000	Bin 1	Bin 2	0.0044	-0.35133	high	Semi-Detached	0	0	1	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Single	0	1	0	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Row Inside	0	0	0	1	0	0	0	0
103000	Bin 1	Bin 2	0.00412	-0.36028	high	Row Inside	0	0	0	1	0	0	0	0
46500	Bin 1	Bin 2	0.00186	-0.43255	low	Semi-Detached	0	0	1	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Row Inside	0	0	0	1	0	0	0	0
890000	Bin 1	Bin 4	0.0356	0.646337	high	Row Inside	0	0	0	1	0	0	0	0
840000	Bin 1	Bin 4	0.0336	0.582384	high	Row Inside	0	0	0	1	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Row Inside	0	0	0	1	0	0	0	0
510000	Bin 1	Bin 3	0.0204	0.160294	high	Row End	1	0	0	0	0	0	0	0
705000	Bin 1	Bin 4	0.0282	0.409711	high	Row Inside	0	0	0	1	0	0	0	0
495000	Bin 1	Bin 3	0.0198	0.141108	high	Row Inside	0	0	0	1	0	0	0	0
93000	Bin 1	Bin 2	0.00372	-0.37307	medium	Semi-Detached	0	0	1	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Row Inside	0	0	0	1	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Row Inside	0	0	0	1	0	0	0	0
140000	Bin 1	Bin 2	0.0056	-0.31296	high	Single	0	1	0	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Single	0	1	0	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Single	0	1	0	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Row Inside	0	0	0	1	0	0	0	0
1250000	Bin 1	Bin 5	0.05	1.106798	expensive	Row Inside	0	0	0	1	0	0	0	0
826000	Bin 1	Bin 4	0.03304	0.564477	high	Single	0	1	0	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Single	0	1	0	0	0	0	0	0
347180	Bin 1	Bin 3	0.013887	-0.04796	high	Row Inside	0	0	0	1	0	0	0	0
677000	Bin 1	Bin 4	0.02708	0.373897	high	Semi-Detached	0	0	1	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Single	0	1	0	0	0	0	0	0
1635000	Bin 1	Bin 5	0.0654	1.599236	expensive	Row Inside	0	0	0	1	0	0	0	0
145000	Bin 1	Bin 2	0.0058	-0.30656	high	Semi-Detached	0	0	1	0	0	0	0	0
150000	Bin 1	Bin 2	0.006	-0.30017	high	Row End	1	0	0	0	0	0	0	0
1750000	Bin 1	Bin 5	0.07	1.746328	expensive	Single	0	1	0	0	0	0	0	0
110114	Bin 1	Bin 2	0.004405	-0.35118	high	Semi-Detached	0	0	1	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Row End	1	0	0	0	0	0	0	0
0	Bin 1	Bin 1	0	-0.49203	low	Row Inside	0	0	0	1	0	0	0	0

Each column means…

AQ : PRICE_preprocessing(eliminate “PRICE” missing value)

AR : equi-width(auto-binning)

AS : equi-depth(auto-binning)

AT : min-max normalization

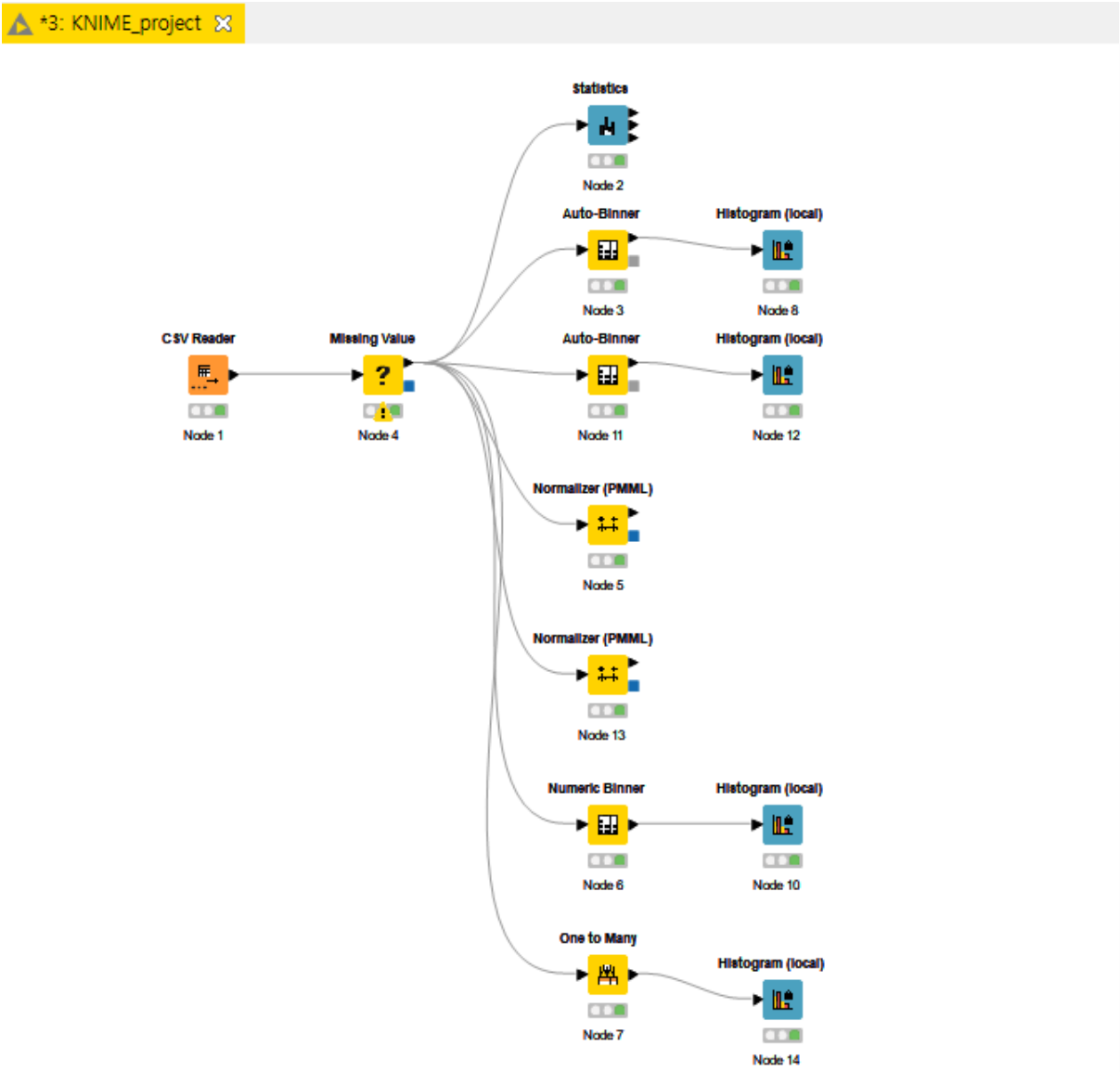
AU : z-score normalization

AV : discretise(numeric binning)

AW : STRUCT_D_preprocessing(eliminate “PRICE” missing value)

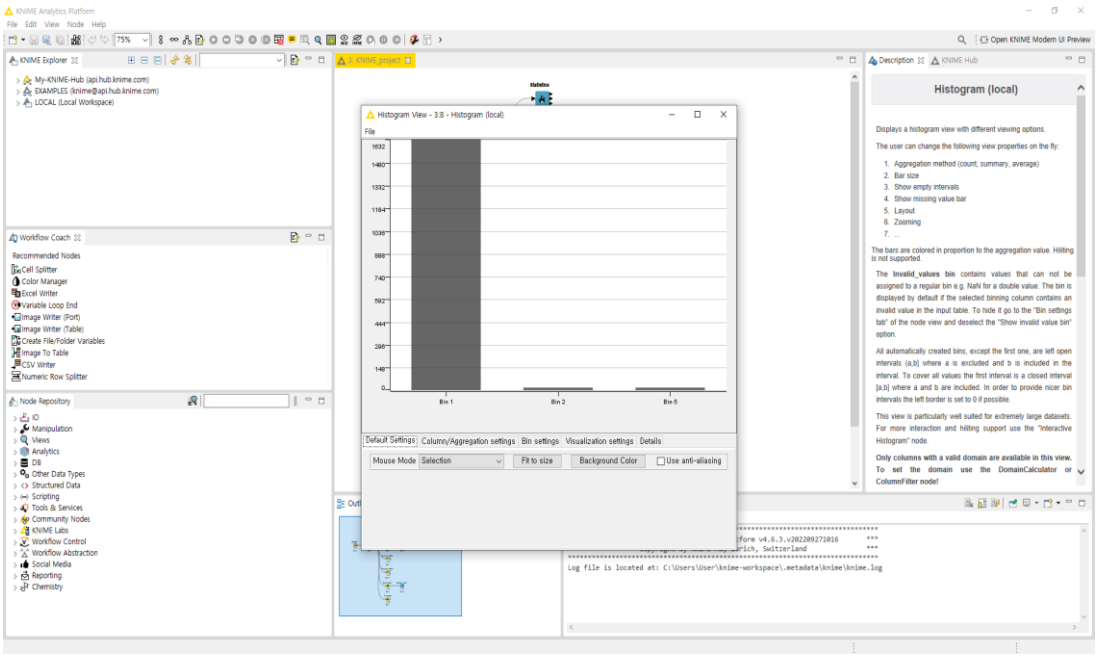
AX ~ BE : STRUCT_D binarise

And this is a KNIME workflow that was used for data analysis.



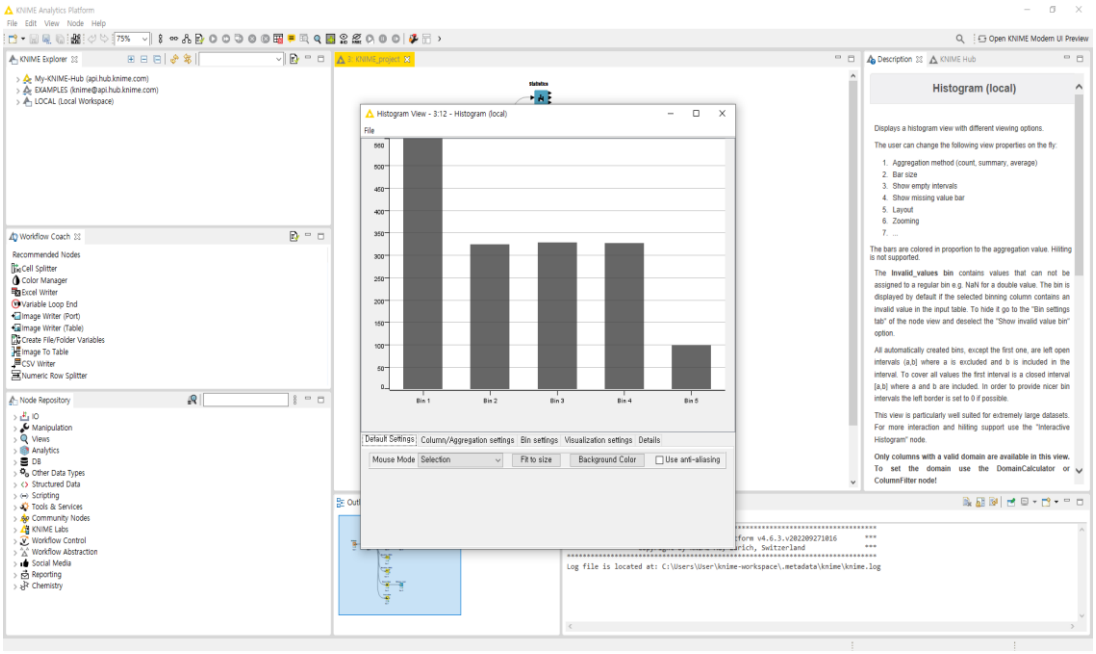
About binning : equi-width, equi-depth, discretise, and binarise, we can visualize the distribution of data, such as histogram. Histogram is a picture that plot arranged in columns (rectangle) with each section as the base and proportional to the frequency of appearance of the measured values in that section when the range of measurements is divided into several sections (classes). In KNIME, we can see histogram by using “Histogram” system.

1. Equi-width



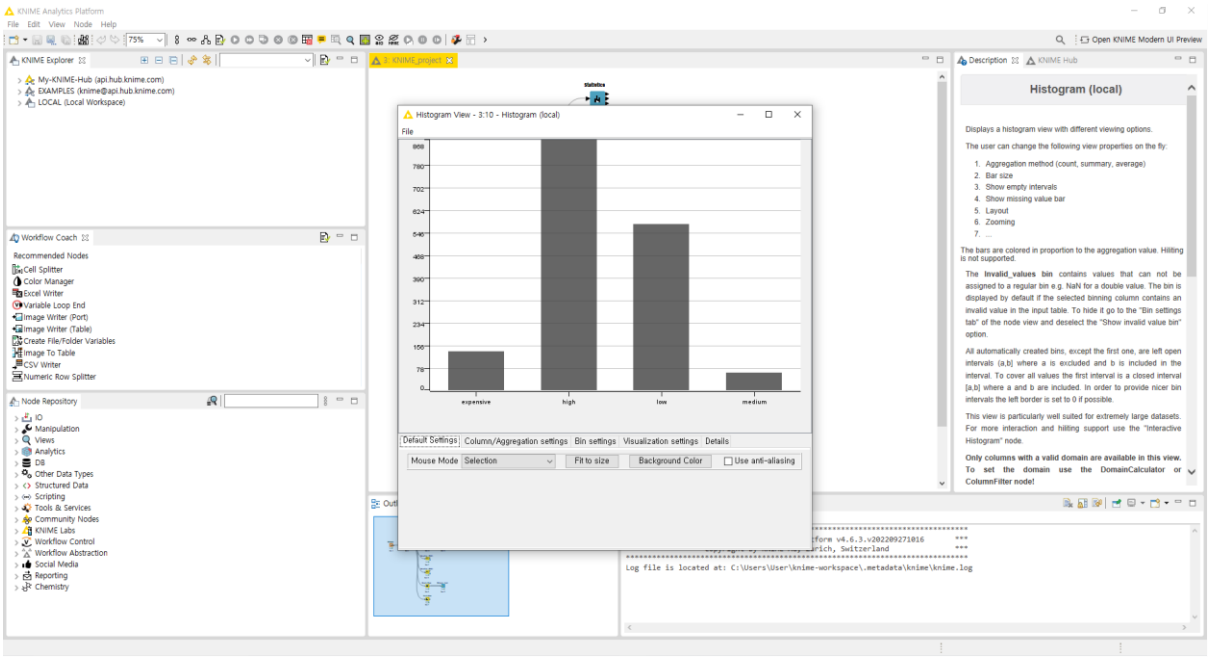
⇨ Binned by width - Bin1(1632), Bin2(1), Bin5(1)

2. Equi-depth



⇨ Binned by depth - Bin1(560), Bin2(323), Bin3(327), Bin4(326), Bin5(98)

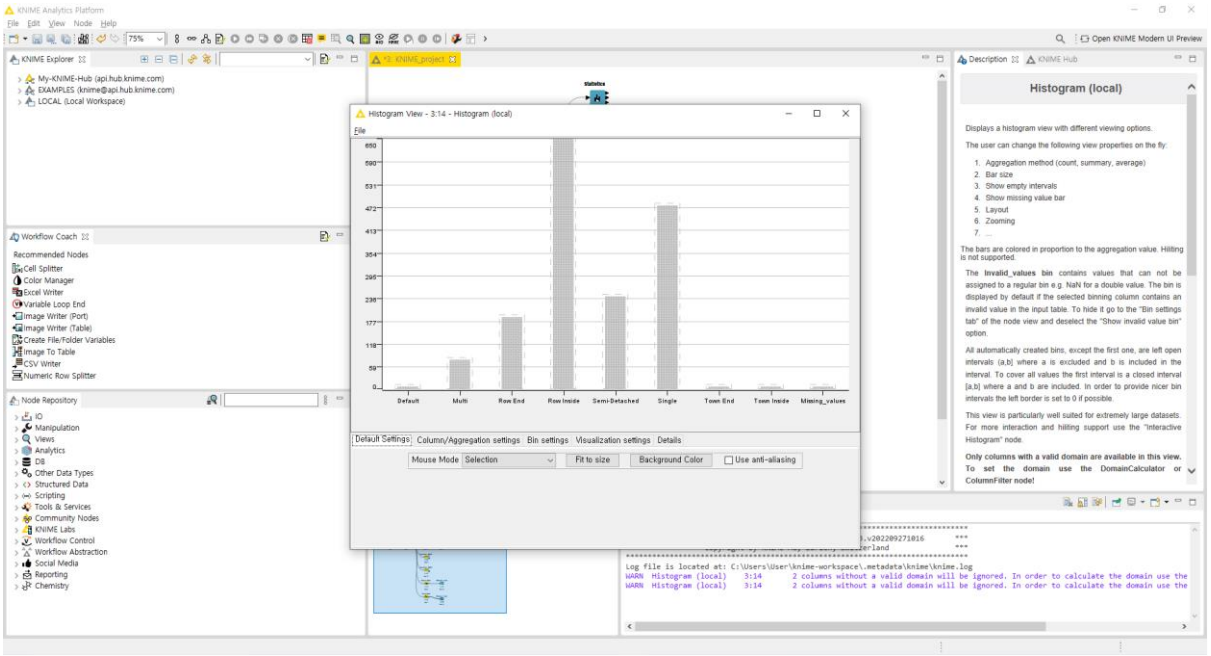
3. Discretise



The distribution of data according to the conditional expression can be checked.

⇨ Expensive(134), high(868), low(573), medium(59)

4. Binarise



Know the distribution of STRUCT_D

⇨ Row End(186), Single(476), Semi-Detached(239), Row Inside(650), Multi(76), Default(1), Town Inside(4), Town End(1)