DATA EXPLORATION AND PREPARATION: ASSIGNMENT

UTS Data Analytics Foundations Micro-Credential

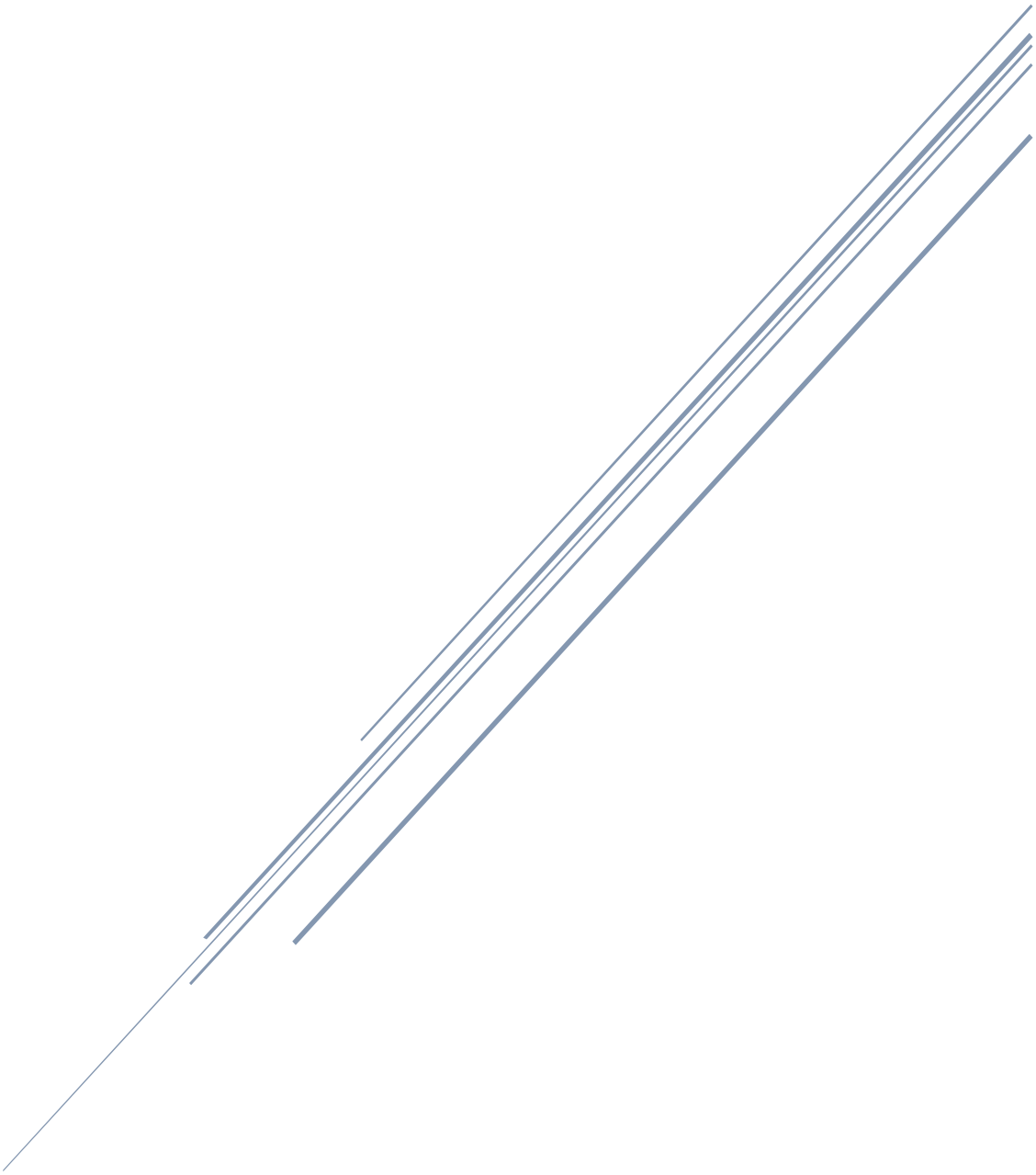

Table of Contents

Table of Figures

1. Introduction

Presented in this report is a detailed analysis of a property dataset with over 3000 unique observations on behalf of the Analytics Unit. A detailed explanation of the steps and methodology taken to provide these insights has also been provided in this report. Following the completion of this analysis, the findings, associations and data points of interest have been presented in the three major sections of this report.

Section 1A of this report examines the initial data exploration process including the identification of attribute types, a statistical evaluation of the attributes within the dataset and finally an exploration of the dataset itself uncovering any outliers and data clusters.

Similarly, section 1B of the report explains the data pre-processing stage of the analysis. This report will cover the importance of data cleansing and cover several data pre-processing techniques.

The final component of this report, Section 1C summarizes the findings of this data analysis. This section of the report will include the most important findings and present the key conclusions of the report in a succinct manner.

row ID	BATHRM	HF_BATHF	HEAT	HEAT_D	AC	NUM_UNI	ROOMS	BEDRM	AYB	YR_RMDL	EYB	STORIES	SALEDATE	PRICE	SALE_NUM	GBA
Row1	3	1	13	Hot Wate	Y	2	9	5	1910	2009	1984	3	2016-06-2	2100000	3	2522
Row63	3	1	13	Hot Wate	N	4	11	5	1948		1967	3	2013-04-2	0	1	2930
Row71	3	1	13	Hot Wate	Y	2	8	4	1910	2003	1972	3	2011-09-1	1680000	1	2556
Row93	3	1	7	Warm Coc	Y	2	11	6	1906	1973	1963	3	2011-01-1	0	1	2401
Row121	2	0	13	Hot Wate	N	2	8	3	1880	2002	1960	3	1900-01-01	T00:00:00.	1	2925
Row165	2	0	1	Forced Air	Y	2	6	3	1900	1996	1967	2	1998-09-0	0	1	792
Row187	1	1	7	Warm Coc	Y	1	6	2	1900	2001	1964	2	2006-08-1	575000	1	1200
Row202	2	0	5	Elec Base	N	2	6	2	1880		1957	2	2000-05-0	289000	1	696
Row243	3	1	13	Hot Wate	Y	2	7	4	1885	2009	1963	3	2016-05-2	0	5	3036
Row287	6	1	7	Warm Coc	Y	2	10	6	1900	2003	1972	3	2012-01-0	1231000	1	2838
Row332	3	1	7	Warm Coc	Y	1	7	3	1996		2004	3	2006-07-2	1210000	1	2408
Row346	1	0	13	Hot Wate	N	1	7	4	1885		1960	3	2007-07-3	0	1	2292
Row356	6	1	7	Warm Coc	Y	3	10	6	1900	2002	1984	3	2002-01-0	1125000	1	2319
Row384	1	0	13	Hot Wate	N	1	6	3	1907		1957	2	1996-04-3	232000	1	1354
Row395	4	0	13	Hot Wate	N	2	10	5	1905	1982	1972	2	1900-01-01	T00:00:00.	1	2440
Row400	3	0	7	Warm Coc	Y	3	10	6	1927	2008	1967	2	2010-05-1	1225000	1	1912
Row406	3	0	7	Warm Coc	Y	2	6	4	1895	1980	1972	3	2012-08-2	810000	1	1632
Row409	2	1	1	Forced Air	Y	2	7	4	1895	1988	1950	3	1900-01-01	T00:00:00.	1	1960
Row423	4	0	1	Forced Air	Y	4	13	6	1890	2004	1969	3	2008-12-2	0	1	1881
Row480	3	1	7	Warm Coc	Y	2	7	4	1900	2007	1972	3	2017-02-0	1505000	3	2924
Row514	4	1	13	Hot Wate	Y	2	12	5	1890	2005	1987	3	2018-03-2	0	5	2496
Row546	2	1	13	Hot Wate	Y	2	13	4	1890		1969	3	2018-01-1	1349750	2	2199
Row574	1	1	13	Hot Wate	N	1	6	3	1900	2012	1957	2	2000-01-2	349000	1	1324
Row668	4	0	7	Warm Coc	Y	4	16	8	1900		1957	2	2016-02-1	1120160	3	2928
Row682	1	0	13	Hot Wate	N	1	6	3	1900		1954	2	1900-01-0	0	1	1524
Row721	1	0	13	Hot Wate	N	1	6	2	1880	2011	1954	2	2002-10-1	0	1	888
Row738	3	1	1	Forced Air	Y	1	8	3	2001		2007	3	2013-09-0	879000	1	1832
Row768	2	0	13	Hot Wate	N	1	6	3	1910		1957	2	2016-06-2	675000	3	1350
Row773	1	0	13	Hot Wate	N	1	9	3	1900		1950	3	2017-07-2	0	4	3266
Row776	3	1	1	Forced Air	Y	2	8	4	1900	2010	1967	2	2009-12-0	0	1	1544
Row795	1	0	13	Hot Wate	Y	1	5	2	1900	2006	1964	2	1900-01-0	0	1	860
Row807	2	0	13	Hot Wate	N	1	6	2	1909		1954	2	2005-08-0	400000	1	1440
Row816	4	1	1	Forced Air	Y	1	9	5	1900	2006	1982	3	2007-12-2	1145000	1	3150
Row881	3	0	13	Hot Wate	Y	1	9	4	1885		1960	3	2001-10-1	250500	1	2685
Row895	2	0	13	Hot Wate	N	2	8	4	1909		1943	2	1900-01-01	T00:00:00.	1	1600

Figure 1 - The Dataset in CSV format before analysis

1.A Initial Data Exploration

1.A.I Dataset Attribute Type Identification and Justification

Table 1 below summarises the Attribute type selection (Nominal, Ordinal, Interval and Ratio) of the dataset and provides the justification for this selection.

Attribute	**Attribute Type**	**Justification**
row ID	Nominal	The row ID is a unique identifying attribute for each row of this dataset. Although numbers are provided, they are only labels with no further analysis able to be performed on this attribute.
BATHRM	Ratio	There is a clearly defined zero for this attribute - the absence of any bathrooms. The unit of measurement is the number of rooms and all mathematical operations can be performed.
HF_BATHRM	Ratio	There is a clearly defined zero for this attribute - the absence of any half bathrooms. The unit of measurement is the number of half-bathrooms and all mathematical operations can be performed.
HEAT	Nominal	This attribute only provides a simple descriptive code for heating. The values cannot be ordered and only limited statistical analysis can be performed.
HEAT_D	Nominal	This attribute only provides a simple descriptive label for heating. The values cannot be ordered and only limited statistical analysis can be performed.
AC	Nominal	This attribute provides only a simple ‘Y’ or ‘N’ description of the Air Conditioning status and doesn’t provide any comparative or ordinal information.
NUM_UNITS	Ratio	There is a clearly defined zero for this attribute - the absence of any units. All mathematical operations can be performed.
ROOMS	Ratio	There is a clearly defined zero for this attribute - the absence of any rooms. All mathematical operations can be performed.
BEDRM	Ratio	There is a clearly defined zero for this attribute - the absence of any bedrooms. All mathematical operations can be performed.
AYB	Interval	A calendar date and time of the last data modification, this ordered attribute allows the measurement of fixed and equal units. However, there is no defined origin of the scale.
YR_RMDL	Interval	A calendar date and time of the earliest time the main portion of the building being built, this ordered attribute allows the measurement of fixed and equal units. However, there is no defined origin of the scale.
EYB	Interval	A calendar date and time of the most recent modification, this ordered attribute allows the measurement of fixed and

		equal units. However, there is no defined origin of the scale.
STORIES	Ratio	There is a clearly defined zero for this attribute - the absence of any building stories. All mathematical operations can be performed.
SALEDATE	Interval	A calendar date and time of the sale date, this ordered attribute allows the measurement of fixed and equal units. However, there is no defined origin of the scale.
PRICE	Ratio	The price attribute has a clearly defined zero for the property and all mathematical operations can be performed.
SALE_NUM	Nominal	This attribute only provides a simple descriptive label for the sale number. The values cannot be ordered and only limited statistical analysis can be performed.
GBA	Ratio	There is a clearly defined zero for this attribute – no building area. All mathematical operations can be performed.
BLDG_NUM	Nominal	This attribute only provides a simple descriptive label for the building number. The values cannot be ordered and only limited statistical analysis can be performed.
STYLE	Nominal	This attribute only provides a simple descriptive label for the style code. The values cannot be ordered and only limited statistical analysis can be performed.
STYLE_D	Nominal	This attribute only provides a simple descriptive label for the building style. The values cannot be ordered and only limited statistical analysis can be performed.
STRUCT	Nominal	This attribute only provides a simple descriptive label for the structure code. The values cannot be ordered and only limited statistical analysis can be performed.
STRUCT_D	Nominal	This attribute only provides a simple descriptive label for the structure description. The values cannot be ordered and only limited statistical analysis can be performed.
GRADE	Interval	There is enough information in this attribute to compare and order the values, the values themselves are measured in fixed and equal units.
GRADE_D	Ordinal	Although a qualitative description, there is enough information in this attribute to compare and order the values.
CNDTN	Interval	There is enough information in this attribute to compare and order the values, the values themselves are measured in fixed and equal units.
CNDTN_D	Ordinal	Although a qualitative description, there is enough information in this attribute to compare and order the values.
EXTWALL	Nominal	This attribute only provides a simple descriptive label for the exterior wall code. The values cannot be ordered and only limited statistical analysis can be performed.

EXTWALL_D	Nominal	This attribute only provides a simple descriptive label for the exterior wall description. The values cannot be ordered and only limited statistical analysis can be performed.
ROOF	Nominal	This attribute only provides a simple descriptive label for the roof code. The values cannot be ordered and only limited statistical analysis can be performed.
ROOF_D	Nominal	This attribute only provides a simple descriptive label for the roof description. The values cannot be ordered and only limited statistical analysis can be performed.
INTWALL	Nominal	This attribute only provides a descriptive code for the interior wall. The values cannot be arranged in any specific order.
INTWALL_D	Nominal	This attribute only provides a descriptive label for the interior wall. The values cannot be arranged in any specific order.
KITCHENS	Ratio	There is a clearly defined zero for this attribute - the absence of any kitchens. All mathematical operations can be performed.
FIREPLACES	Ratio	There is a clearly defined zero for this attribute - the absence of any fireplaces. All mathematical operations can be performed.
USECODE	Nominal	The Usecode attribute only acts as a label, the values cannot be arranged in any specific order.
LANDAREA	Ratio	This attribute provides a measurement of the land area of the property. All mathematical operations can be performed and there is a clear zero point.
GIS_LAST_MOD_DTTM	Interval	A calendar date and time of the last data modification, this ordered attribute allows the measurement of fixed and equal units. However, there is no defined origin of the scale.
QUALIFIED	Nominal	This attribute provides only a binary ‘0’ or ‘1’ description of the ‘Qualified status’ and doesn’t provide any comparative or ordinal information.

Table 1 - Dataset Attribute Type

1.A.II Summarising properties of the Attributes

For each of the attributes in the dataset, statistics on the frequency, location and spread have been produced (where possible). The following section 1.A.III will provide commentary on any interesting datapoints found such as outliers, clusters and patterns.

BATHRM (Attribute: Ratio)

Min	0
Max	10
Range	10
Mean	2.014676
Median	2
Std Dev.	1.045727
Variance	1.093545
75th Percentile	3
25th Percentile	1
Missing Values	2

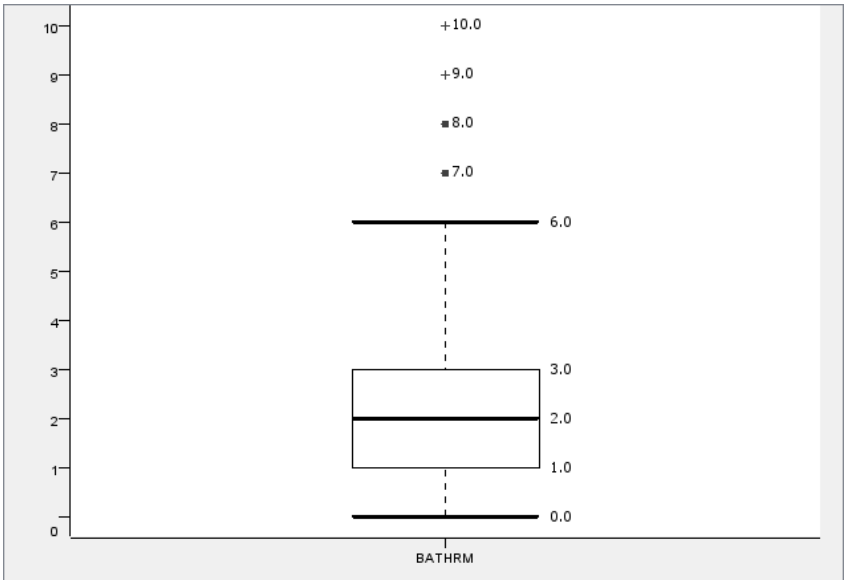

Figure 2 - Box Plot of BATHRM

Value	Frequency	Relative Frequency
0	3	0.10%
1	1121	37.39%
2	1045	34.86%
3	571	19.05%
4	211	7.04%
5	23	0.77%
6	18	0.60%
7	2	0.07%
8	2	0.07%
9	1	0.03%
10	1	0.03%

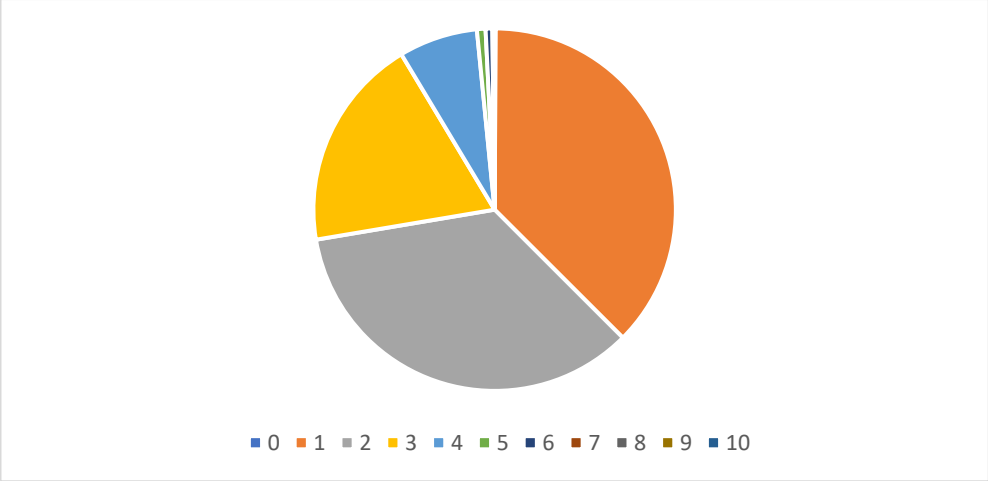

Figure 3 –Pie chart of BATHRM

HF_BATHRM (Attribute: Ratio)

Min	0
Max	4
Range	4
Mean	0.605404
Median	1
Std Dev.	0.615719
Variance	0.37911
75th Percentile	1
25th Percentile	0
Missing Values	2

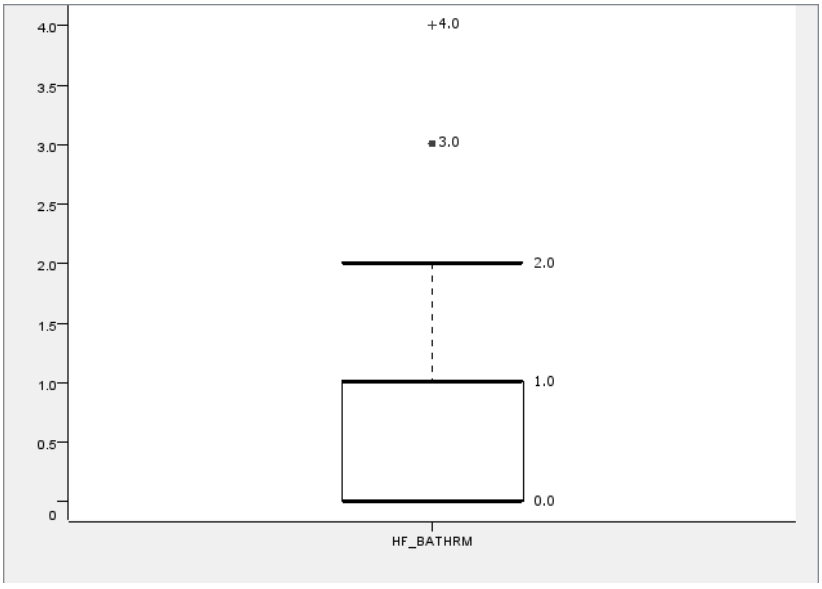

Figure 4 - Box Plot of HF_BATHRM

Value	Absolute Frequency	Relative Frequency
0	1381	46.06%
1	1430	47.70%
2	177	5.90%

3	9	0.30%
4	1	0.03%

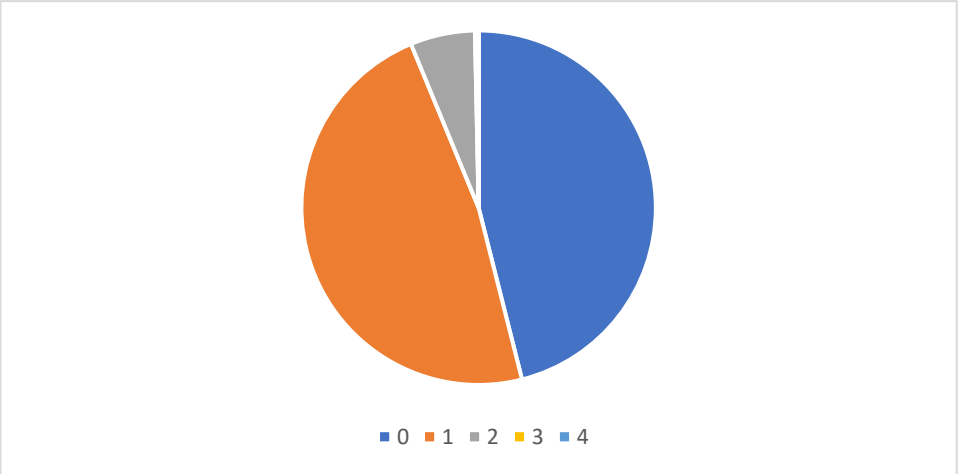

Figure 5 - Pie chart of HF_BATHRM

HEAT (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
0	4	0.13%
1	900	30.02%
2	2	0.07%
3	5	0.17%
4	1	0.03%
5	2	0.07%
6	4	0.13%
7	831	27.72%
8	23	0.77%
11	7	0.23%
13	1219	40.66%

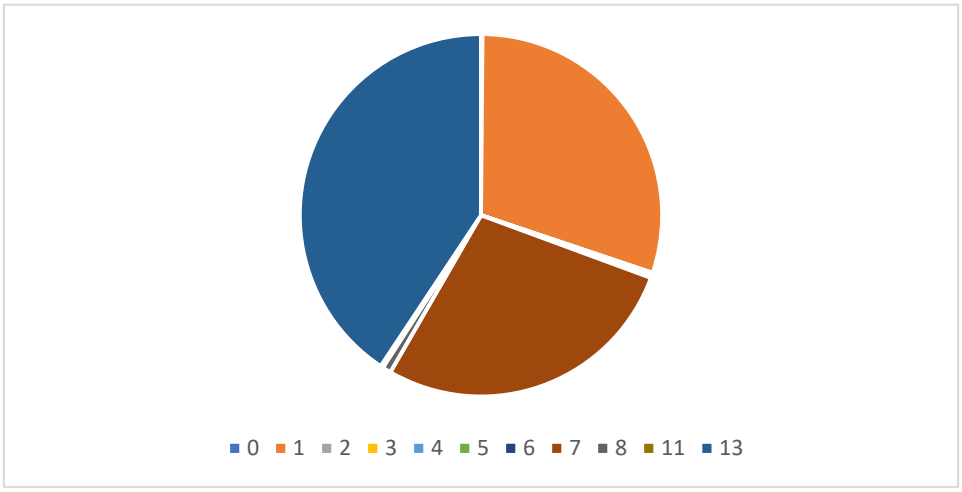

Figure 6 - Pie chart of HEAT

HEAT_D (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
Air-Oil	2	0.07%
Elec Base Brd	2	0.07%
Electric Rad	1	0.03%
Forced Air	900	30.02%
Gravity Furnac	7	0.23%
Hot Water Rad	1219	40.66%
Ht Pump	23	0.77%
No Data	4	0.13%
Wall Furnace	5	0.17%
Warm Cool	831	27.72%
Water Base Brd	4	0.13%

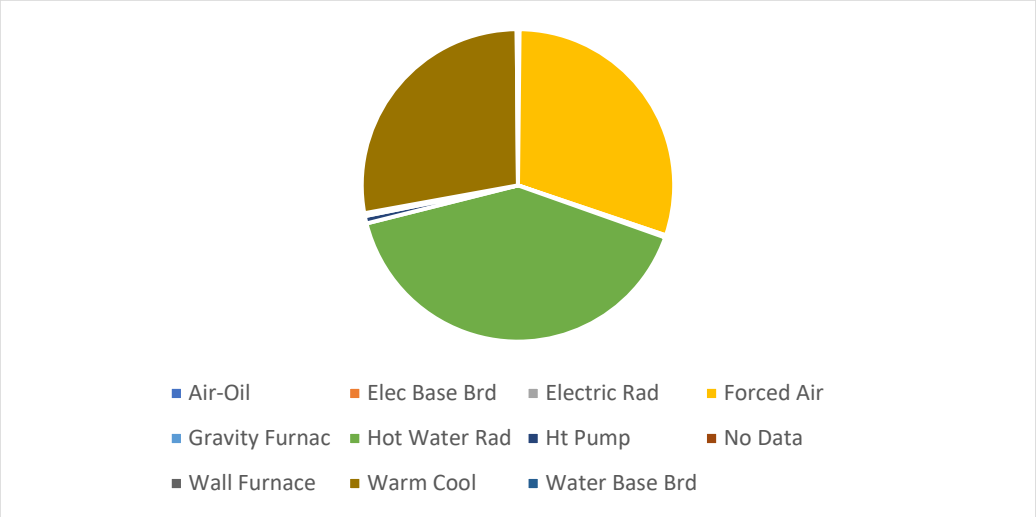

Figure 7 - Pie chart of HEAT_D

AC (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
0	1	0.03%
N	1078	35.96%
Y	1919	64.01%

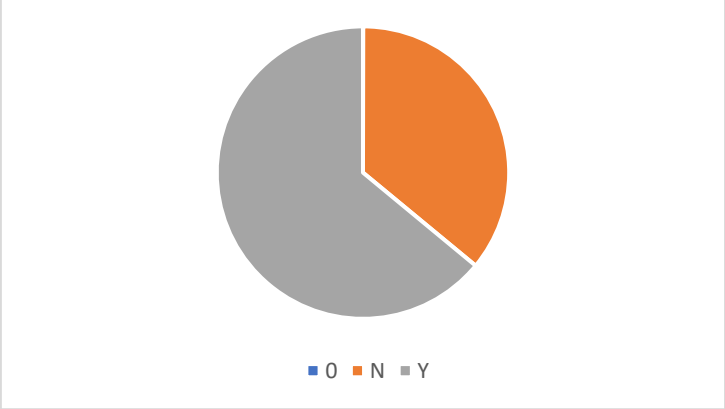

Figure 8 - Pie chart of AC

NUM_UNITS (Attribute: Ratio)

Min	0
Max	4
Range	4
Mean	1.211808
Median	1
Std Dev.	0.627078
Variance	0.393227
75th Percentile	2
25th Percentile	1
Missing Values	2

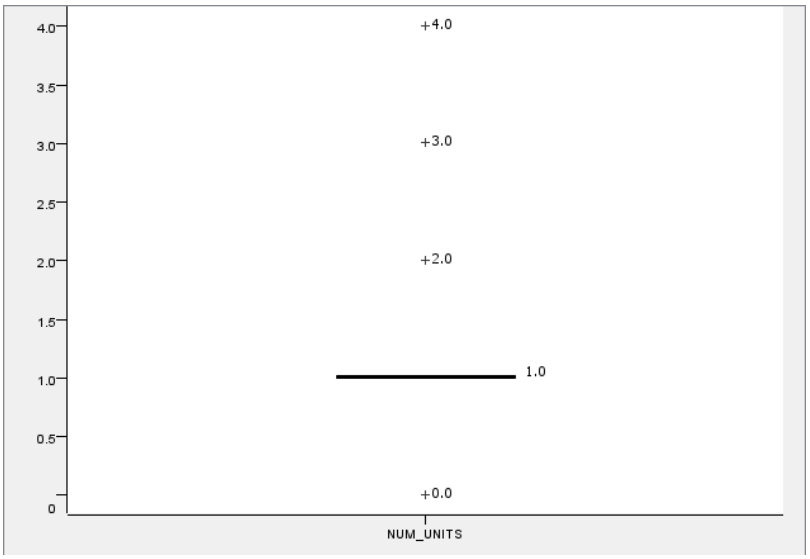

Figure 9 - Box Plot of NUM_UNITS

Value	Absolute Frequency	Relative Frequency
0	6	0.20%
1	2588	86.32%
2	263	8.77%
3	45	1.50%
4	96	3.20%

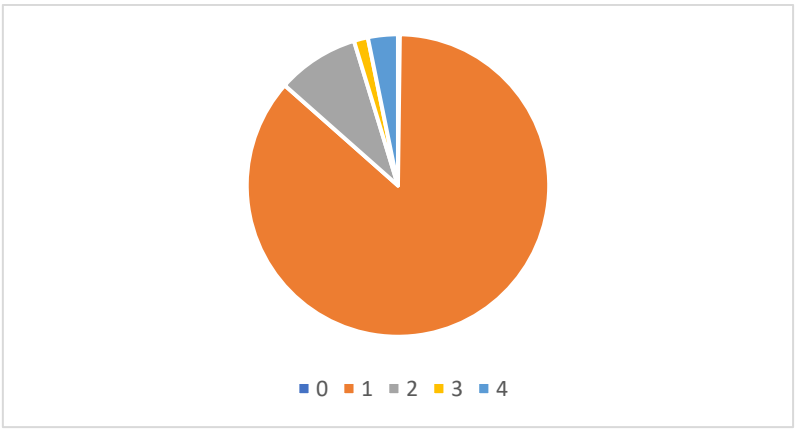

Figure 10 - Pie chart of NUM_UNITS

ROOMS (Attribute: Ratio)

Min	0
Max	23
Range	23
Mean	7.315877
Median	7
Std Dev.	2.309102
Variance	5.331953
75th Percentile	10
25th Percentile	5
Missing Values	2

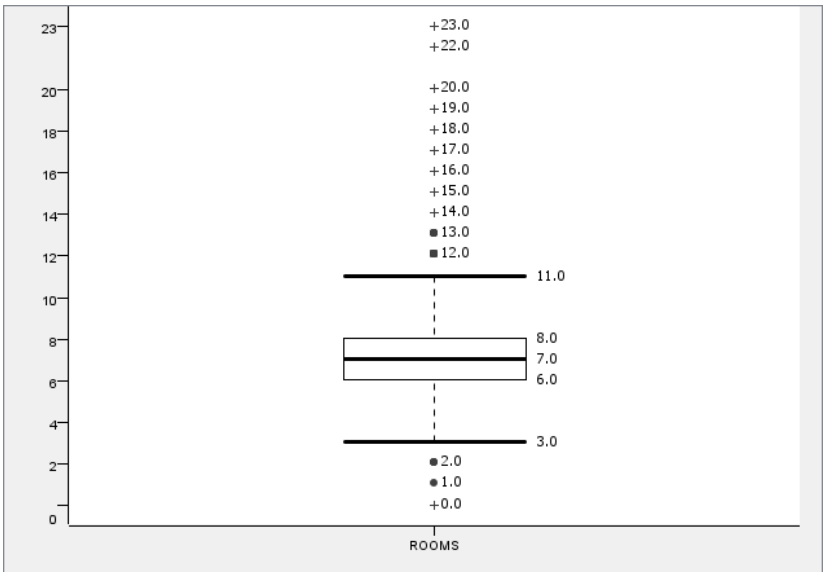

Figure 11- Box Plot of ROOMS

Value	Absolute Frequency	Relative Frequency
0	14	0.47%
1	1	0.03%
2	3	0.10%
3	8	0.27%
4	64	2.13%
5	259	8.64%
6	980	32.69%
7	636	21.21%
8	412	13.74%
9	209	6.97%
10	169	5.64%
11	69	2.30%
12	89	2.97%
13	17	0.57%
14	13	0.43%
15	4	0.13%
16	38	1.27%

17	2	0.07%
18	2	0.07%
19	1	0.03%
20	6	0.20%
22	1	0.03%
23	1	0.03%

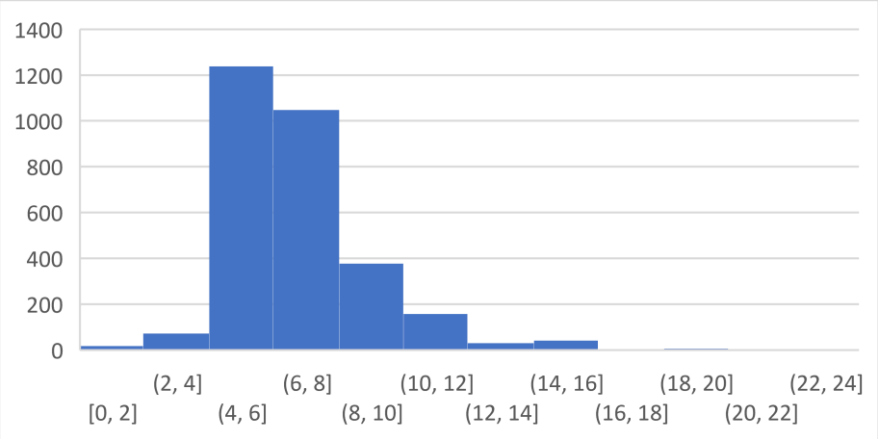

Figure 12- Histogram ROOMS

BEDRM (Attribute: Ratio)

Min	0
Max	11
Range	11
Mean	3.341561
Median	3
Std Dev.	1.089312
Variance	1.1866
90th Percentile	5
10th Percentile	2
Missing Values	2

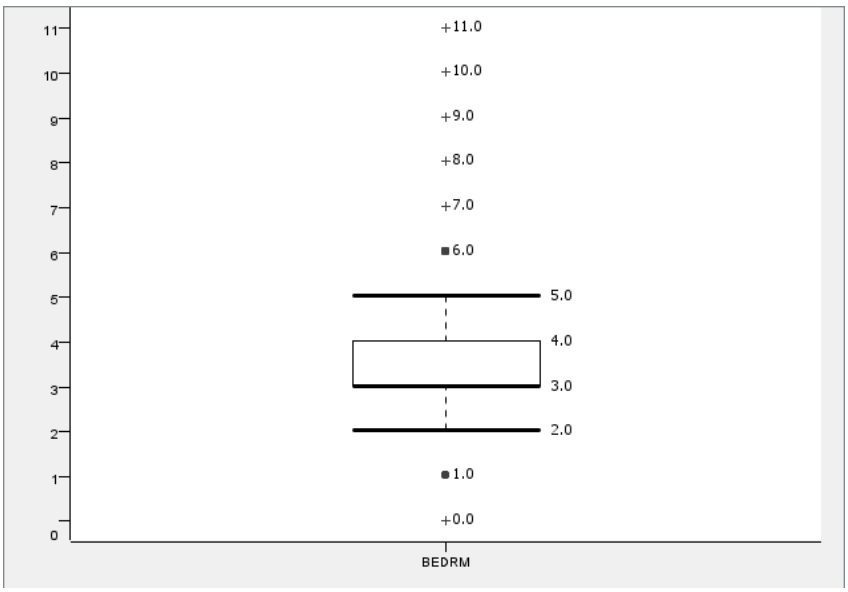

Figure 13 - Box Plot of BEDRM

Value	Absolute Frequency	Relative Frequency
0	12	0.40%
1	13	0.43%
2	428	14.28%
3	1563	52.13%
4	674	22.48%
5	173	5.77%
6	85	2.84%
7	23	0.77%
8	20	0.67%
9	3	0.10%
10	3	0.10%
11	1	0.03%

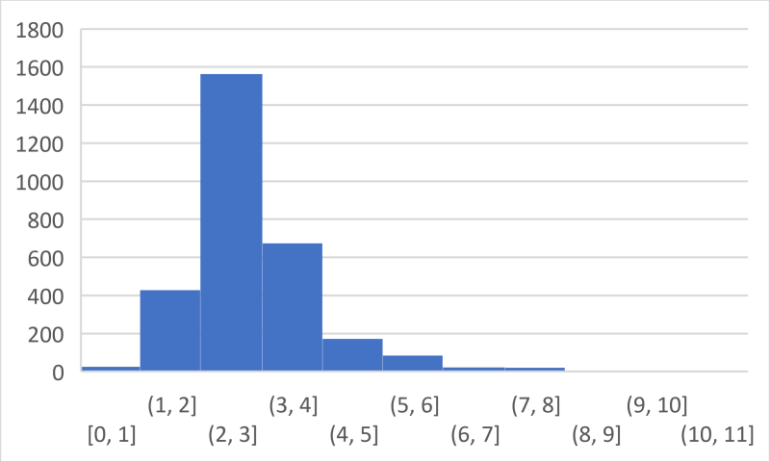

Figure 14 - Histogram of BEDRM

AYB (Attribute: Interval)

Min	0
Max	2019
Range	2019
Mean	1929.522667
Median	1929
Std Dev.	90.91969506
Variance	8266.39095
75th Percentile	1947
25th Percentile	1913
Missing Values	0

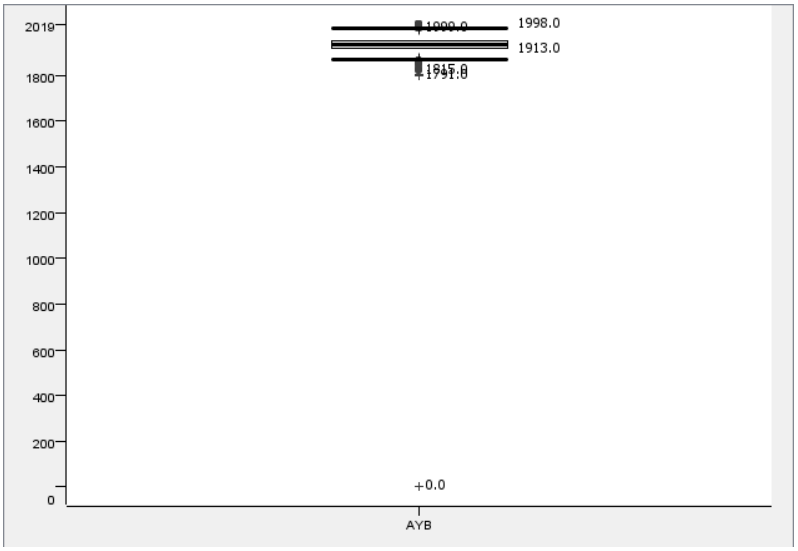

Figure 15 - Box plot of AYB

YR_RMDL (Attribute: Interval)

Min	1880
Max	2018
Range	138
Mean	2001.277
Median	2006
Std Dev.	14.62128
Variance	213.7818
75th Percentile	2011
25th Percentile	1995
Missing Values	1608

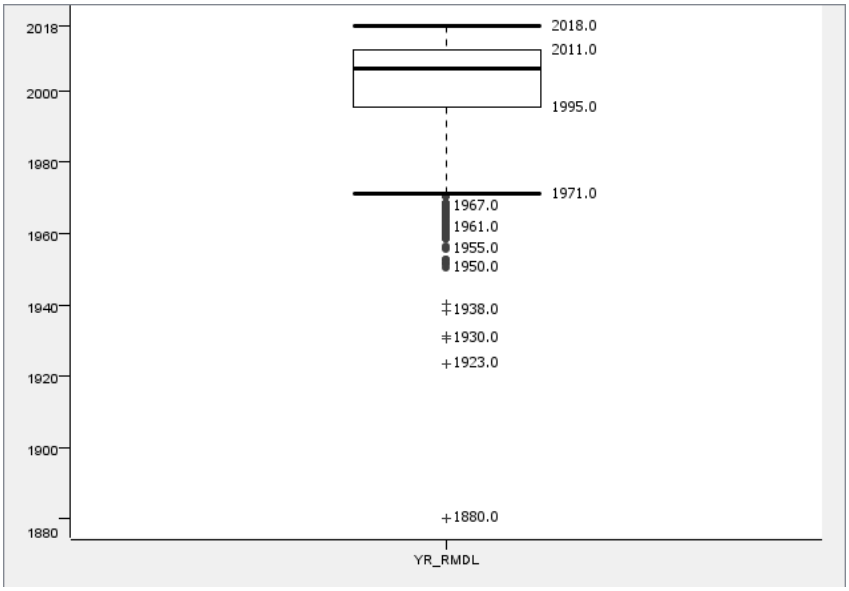

Figure 16 - Box plot of YR_RMDL

EYB (Attribute: Interval)

Min	0
Max	2018

Range	2018
Mean	1964.731
Median	1964
Std Dev.	53.20818555
Variance	2831.111009
75th Percentile	1970
25th Percentile	1957
Missing Values	0

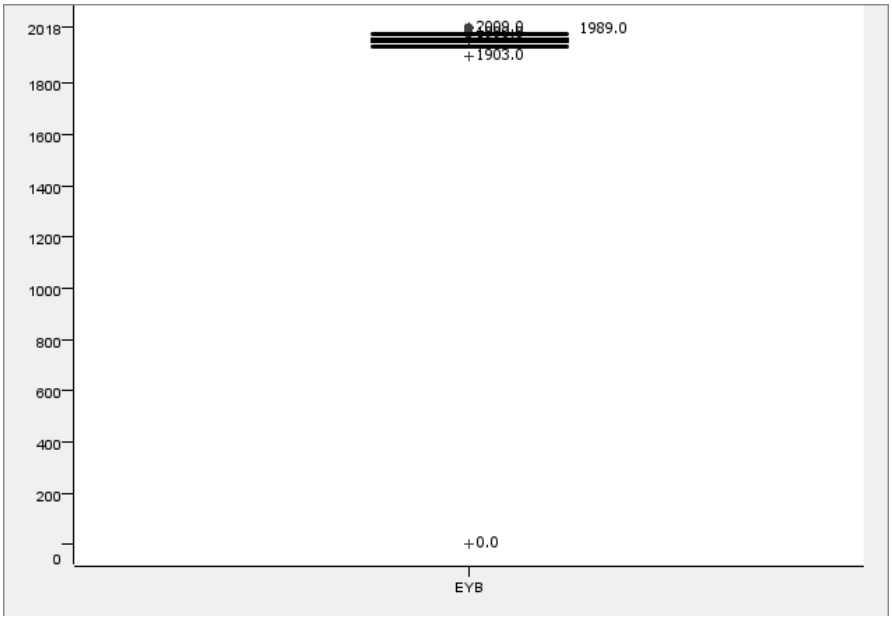

Figure 17 - Box plot of EYB

STORIES (Attribute: Ratio)

Min	0
Max	9
Range	9
Mean	2.075975976
Median	2
Std Dev.	0.468354458
Variance	0.219355898
75th Percentile	2
25th Percentile	2
Missing Values	3

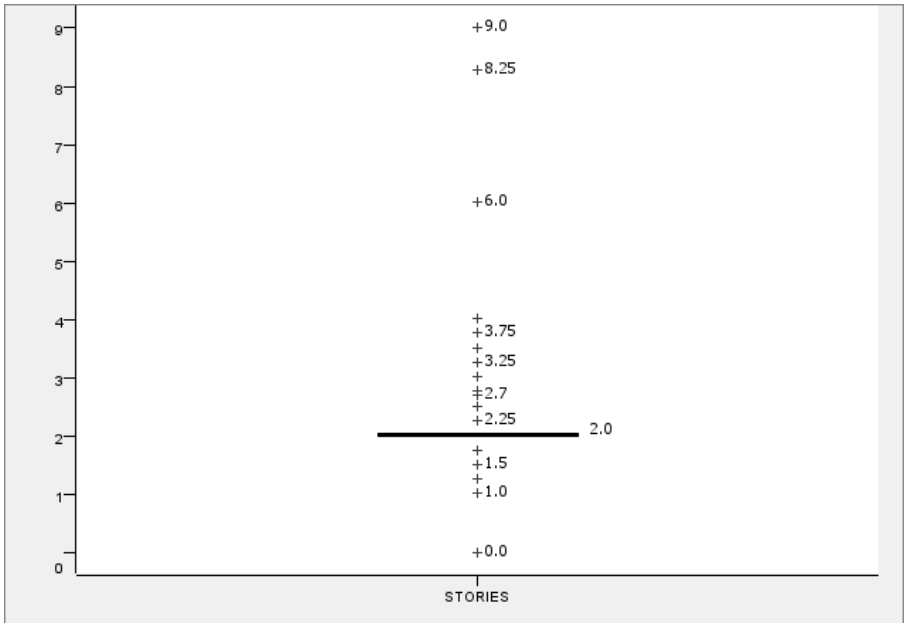

Figure 18 - Box Plot of STORIES

Value	Absolute Frequency	Relative Frequency
0	2	0.07%
1	128	4.27%
1.25	15	0.50%
1.5	66	2.20%
1.75	28	0.93%
2	2247	74.97%
2.25	65	2.17%
2.5	165	5.51%
2.7	1	0.03%
2.75	16	0.53%
3	241	8.04%
3.25	1	0.03%
3.5	5	0.17%
3.75	2	0.07%
4	11	0.37%
6	1	0.03%
8.25	1	0.03%
9	2	0.07%

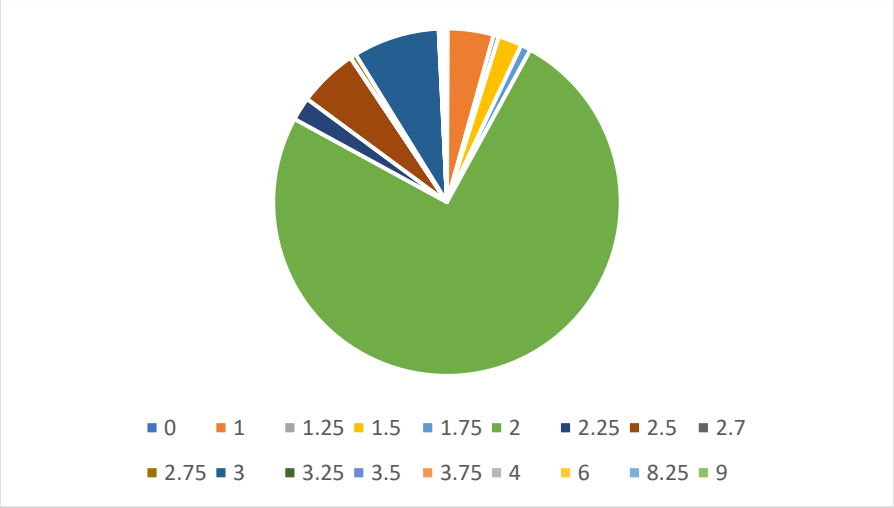

Figure 19 - Pie Chart STORIES

SALEDATE (Attribute: Interval)

Min	1/01/1900
Max	12/07/2018
Mean	30/05/1986
Median	4/02/2007
75th Percentile	25/06/2014
25th Percentile	21/05/1997
Missing Values	0

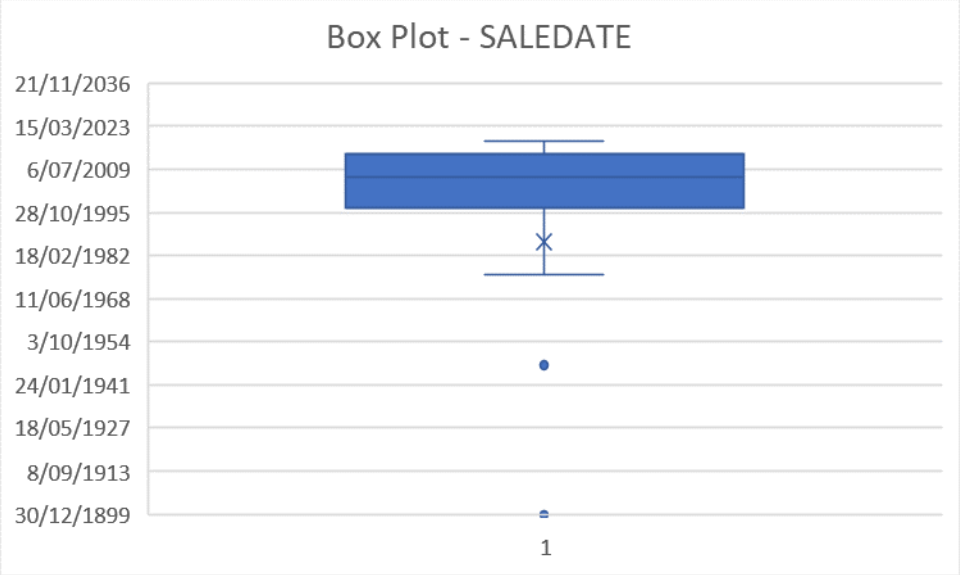

Figure 20 - Box Plot of SALEDATE

PRICE (Attribute: Ratio)

Min	0
Max	7000000
Range	7000000
Mean	371186.9549
Median	229260
Std Dev.	511695.4132

Variance	2.61832E+11
75th Percentile	564000
25th Percentile	0
Missing Values	519

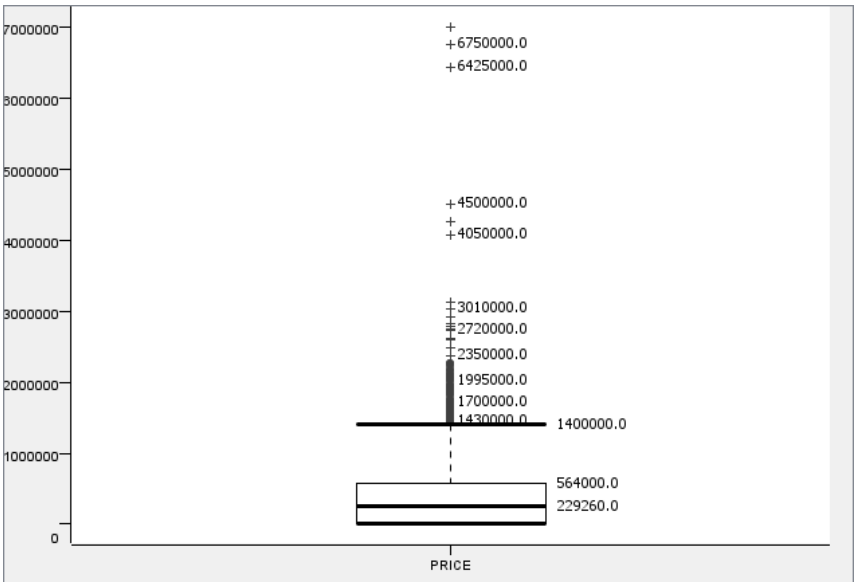

Figure 21 - Box Plot PRICE

SALE_NUM (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
1	2213	44.68%
2	203	8.20%
3	258	15.63%
4	180	14.54%
5	84	8.48%
6	39	4.72%
7	14	1.98%
8	4	0.65%
9	3	0.55%
14	1	0.28%
15	1	0.30%

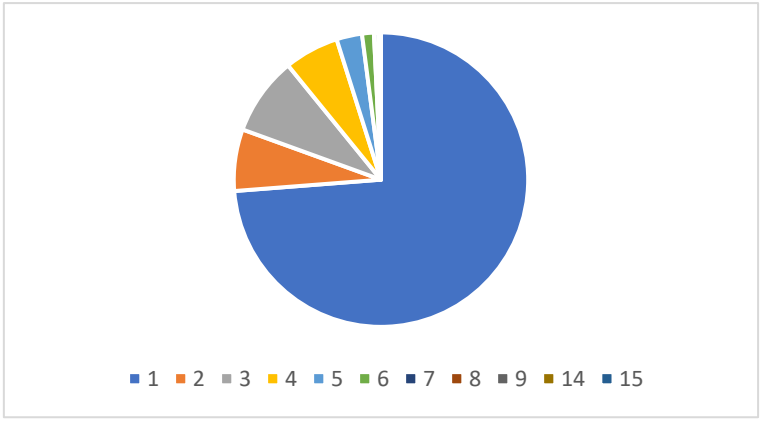

Figure 22- Pie Chart SALE_NUM

GBA (Attribute: Ratio)

Min	0
Max	10661
Range	10661
Mean	1696.739333
Median	1478
Std Dev.	833.9294631
Variance	695438.3495
75th Percentile	1968
25th Percentile	1188
Missing Values	0

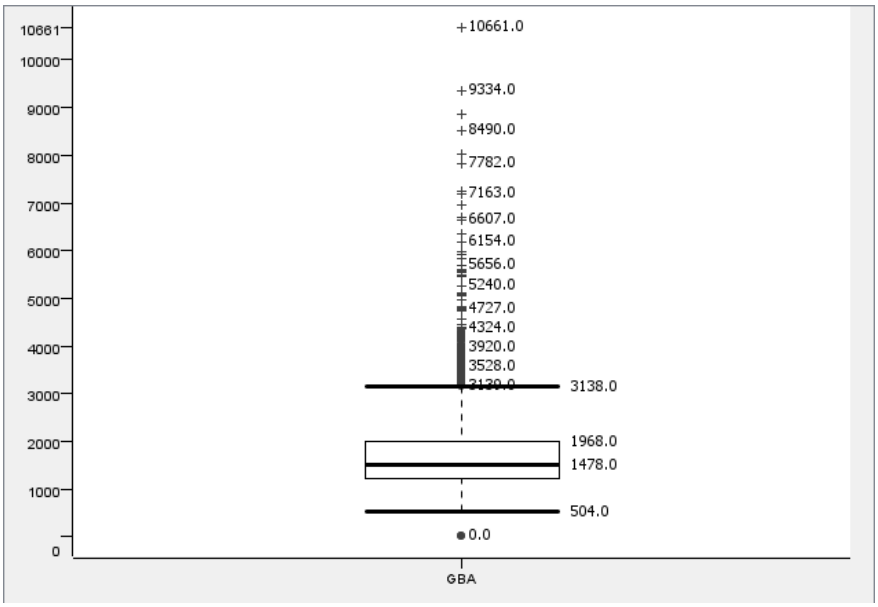

Figure 23 - Box Plot GBA

BLDG_NUM (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
1	2997	99.90%
2	3	0.10%

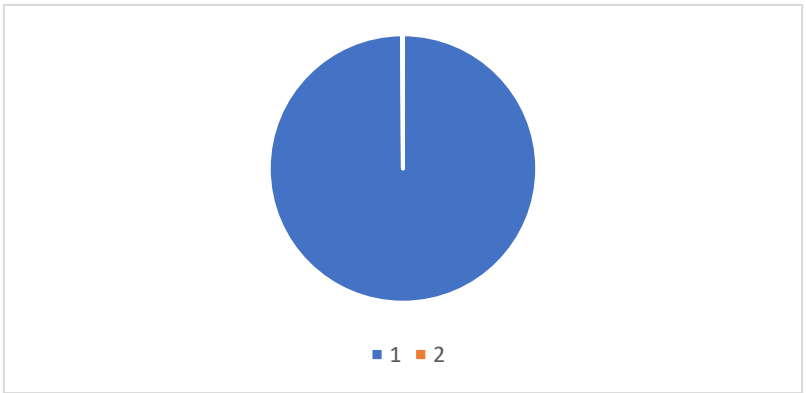

Figure 24 - Pie Chart BLDG_NUM

STYLE (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
0	2	0.07%
1	119	3.97%
2	1	0.03%
3	81	2.70%
4	2288	76.32%
5	24	0.80%
6	193	6.44%
7	252	8.41%
9	7	0.23%
10	14	0.47%
13	1	0.03%
14	7	0.23%
15	9	0.30%

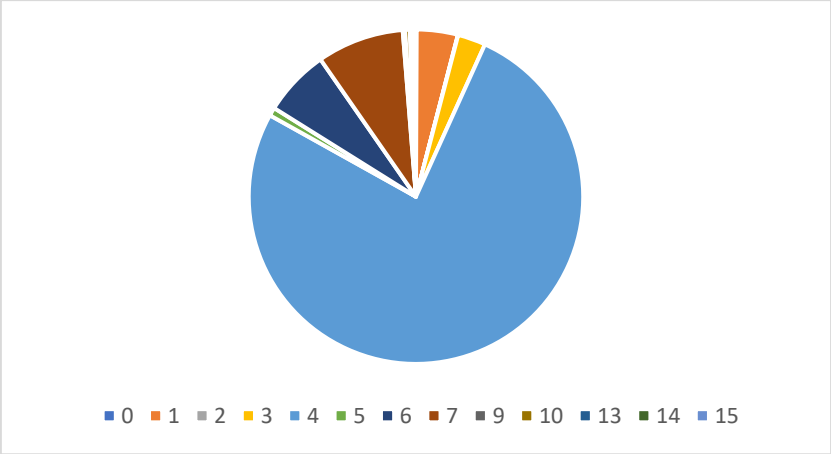

Figure 25 - Pie Chart STYLE

STYLE_D (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
1 Story	119	3.97%
1.5 Story Fin	81	2.70%
1.5 Story Unfin	1	0.03%
2 Story	2288	76.32%
2.5 Story Fin	193	6.44%
2.5 Story Unfin	24	0.80%
3 Story	252	8.41%
3.5 Story Fin	7	0.23%
4 Story	14	0.47%
Bi-Level	1	0.03%
Default	2	0.07%
Split Foyer	9	0.30%
Split Level	7	0.23%

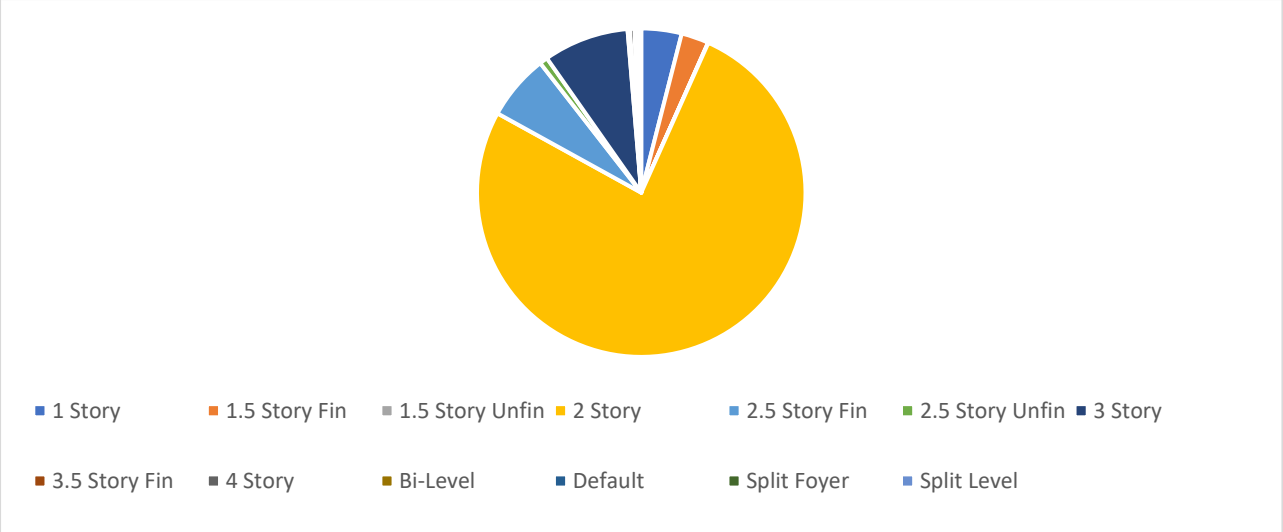

Figure 26 - Pie Chart STYLE_D

STRUCT (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
1	882	29.42%
2	137	4.57%
4	2	0.07%
5	6	0.20%
6	329	10.97%
7	1154	38.49%
8	488	16.28%

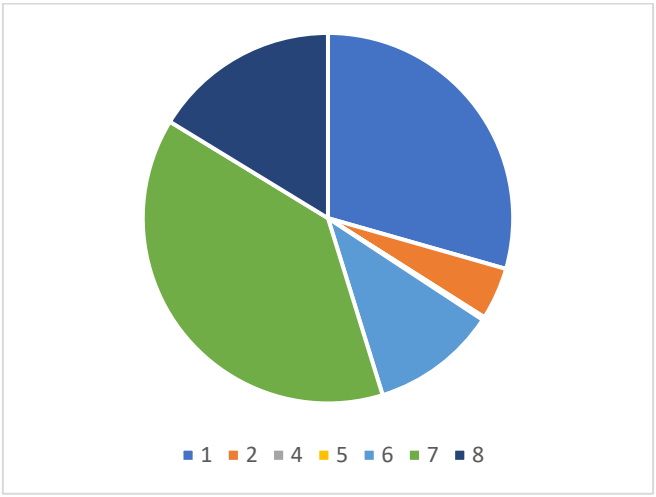

Figure 27 - Pie Chart STRUCT

STRUCT_D (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
Multi	137	4.57%
Row End	329	10.97%
Row Inside	1154	38.49%
Semi-Detached	488	16.28%
Single	882	29.42%

Town End	2	0.07%
Town Inside	6	0.20%

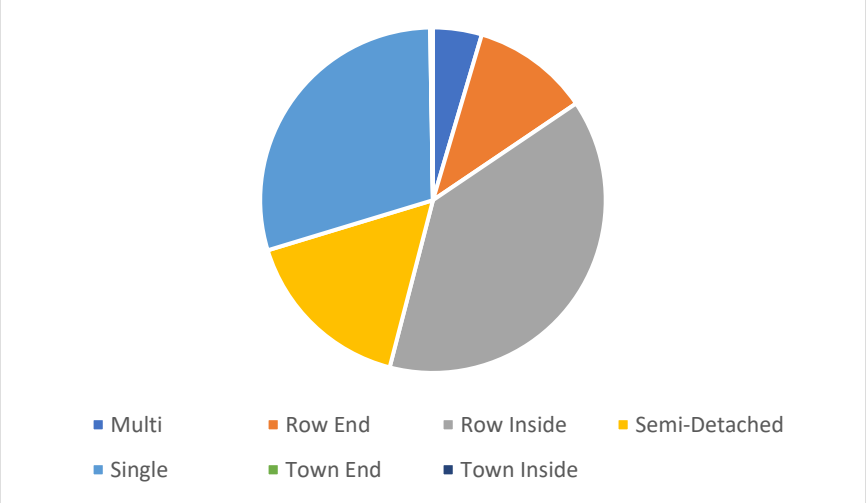

Figure 28 - Pie Chart STRUCT_D

GRADE (Attribute: Interval)

Min	2
Max	12
Range	10
Mean	4.253836
Median	4
Std Dev.	1.354663
Variance	1.835112
75th Percentile	5
25th Percentile	3
Missing Values	2

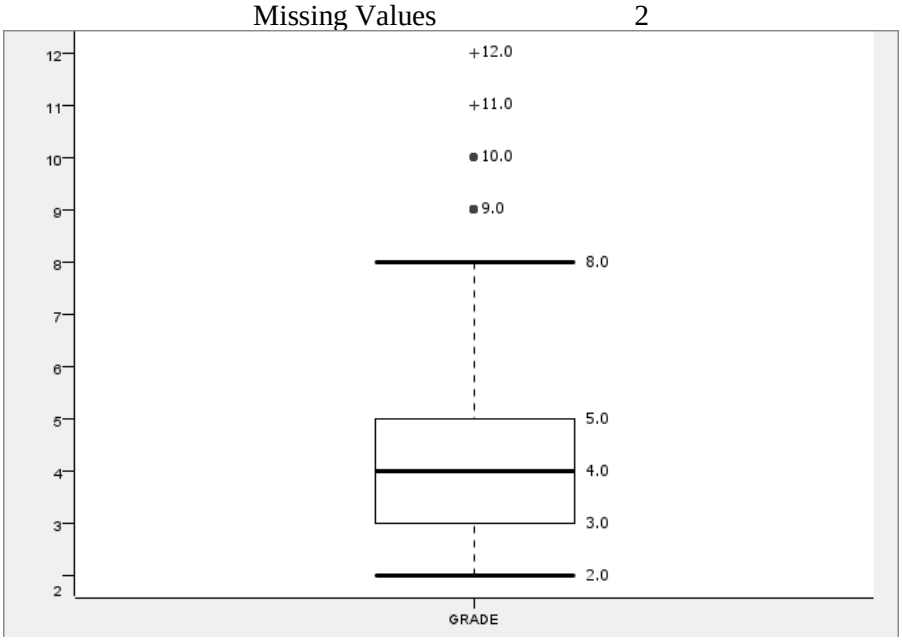

Figure 29 - Box Plot GRADE

Value	Absolute Frequency	Relative Frequency
2	4	0.13%

3	1079	35.99%
4	871	29.05%
5	587	19.58%
6	263	8.77%
7	87	2.90%
8	78	2.60%
9	19	0.63%
10	4	0.13%
11	5	0.17%
12	1	0.03%

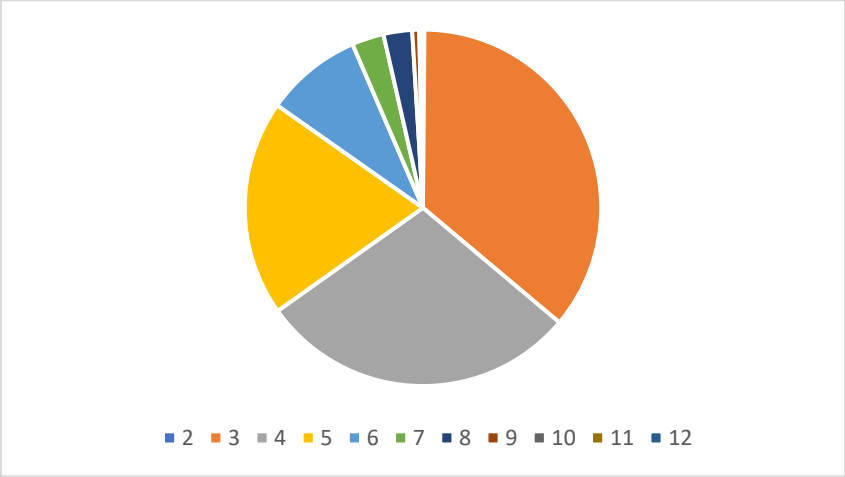

Figure 30 - Pie Chart GRADE

GRADE_D (Attribute: Ordinal)

Value	Absolute Frequency	Relative Frequency
Above Average	871	29.05%
Average	1079	35.99%
Excellent	87	2.90%
Exceptional-A	19	0.63%
Exceptional-B	4	0.13%
Exceptional-C	5	0.17%
Exceptional-D	1	0.03%
Fair Quality	4	0.13%
Good Quality	587	19.58%
Superior	78	2.60%
Very Good	263	8.77%

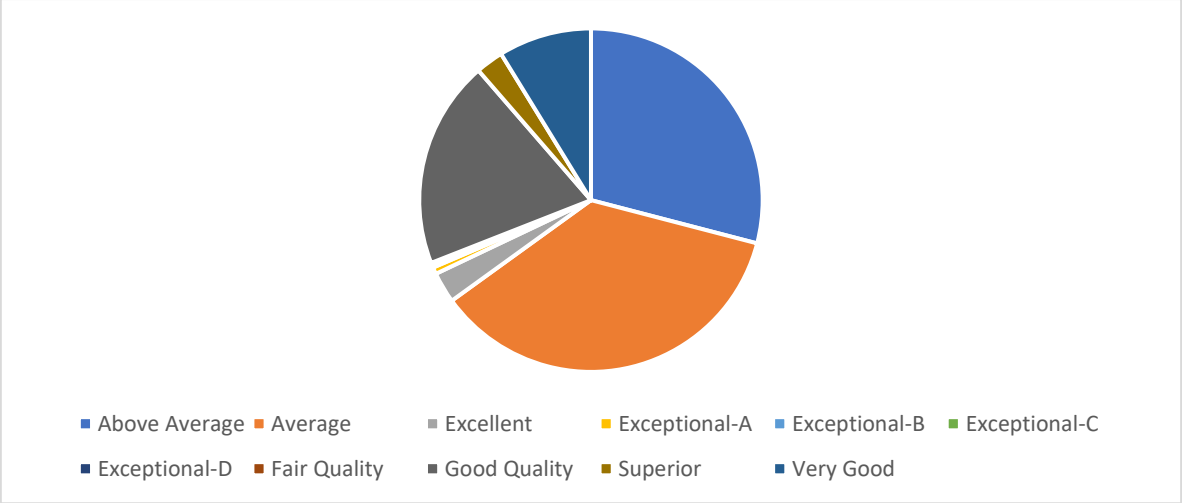

Figure 31 - Pie Chart GRADE_D

CNDTN (Attribute: Interval)

Min	1
Max	6
Range	5
Mean	3.524349566
Median	3
Std Dev.	0.702187459
Variance	0.493067228
75th Percentile	4
25th Percentile	3
Missing Values	2

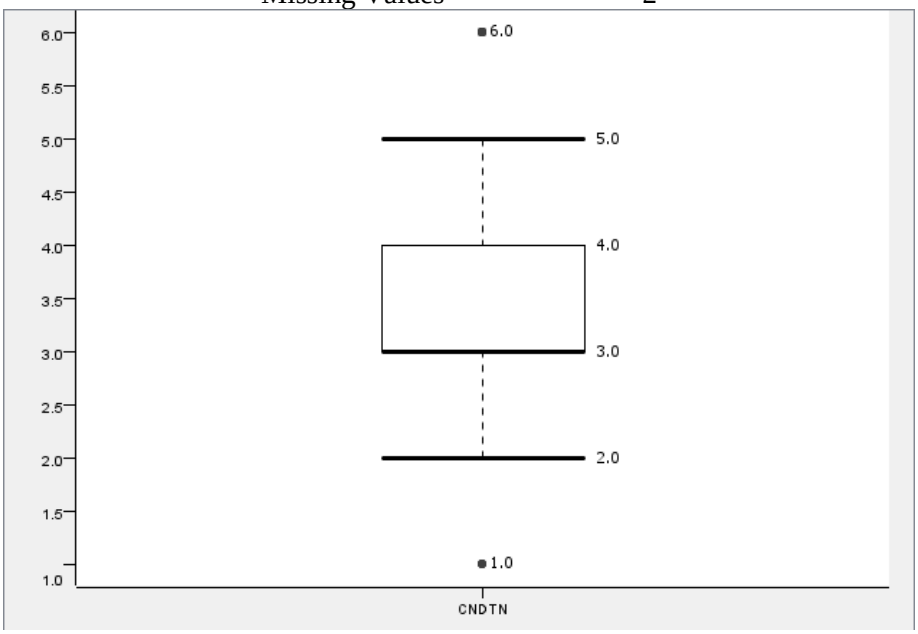

Figure 32 - Box Plot CNDTN

Value	Absolute Frequency	Relative Frequency
1	4	0.13%
2	35	1.17%

3	1630	54.37%
4	1075	35.86%
5	222	7.40%
6	32	1.07%

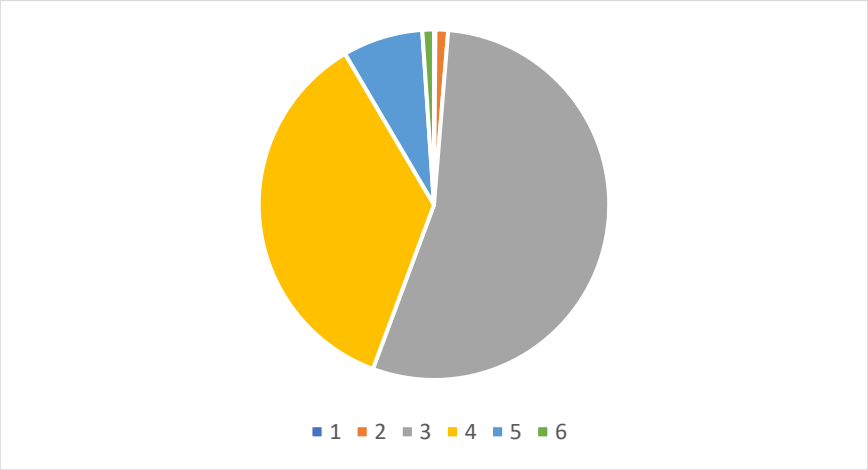

Figure 33 - Pie Chart CNDTN

CDTN_D (Attribute: Ordinal)

Value	Absolute Frequency	Relative Frequency
Average	1630	54.37%
Excellent	32	1.07%
Fair	35	1.17%
Good	1075	35.86%
Poor	4	0.13%
Very Good	222	7.40%

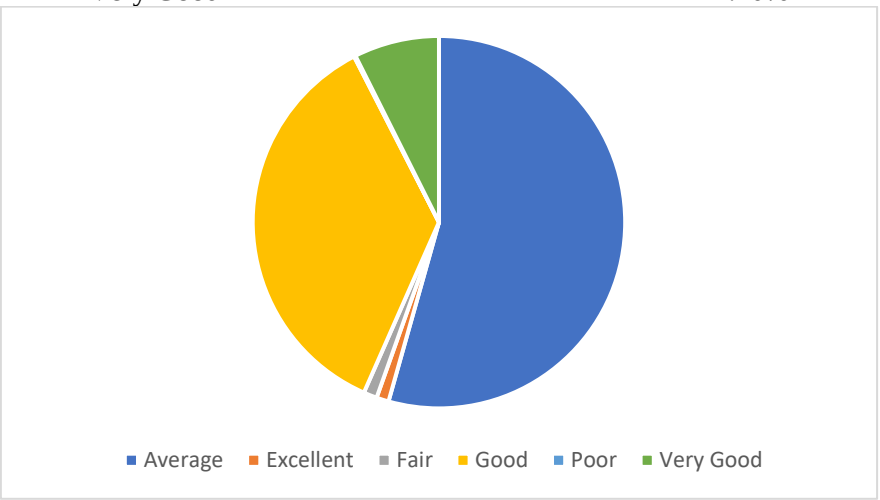

Figure 34 - Pie Chart CNDTN_D

EXTWALL (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
0	1	0.03%
2	1	0.03%
3	4	0.13%
4	147	4.90%

5	88	2.94%
6	142	4.74%
7	28	0.93%
10	32	1.07%
11	5	0.17%
12	1	0.03%
13	3	0.10%
14	2280	76.05%
15	12	0.40%
17	20	0.67%
18	6	0.20%
19	27	0.90%
20	13	0.43%
21	18	0.60%
22	152	5.07%
23	8	0.27%
24	10	0.33%

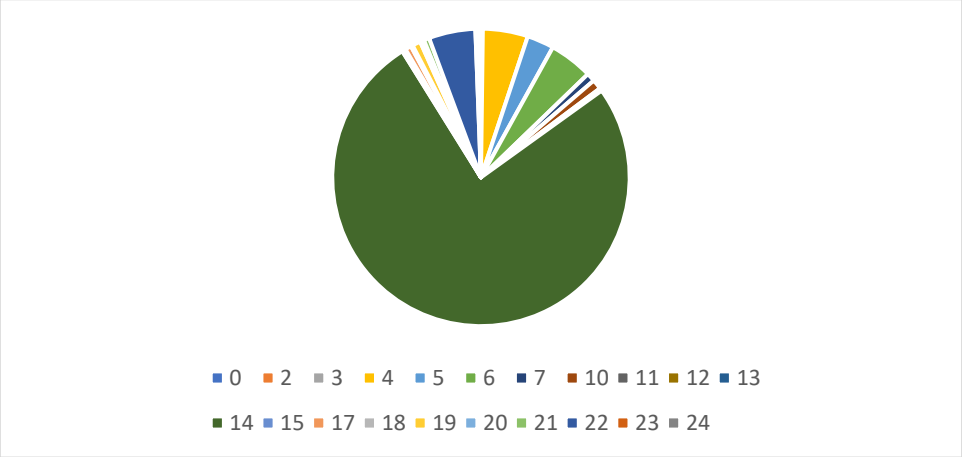

Figure 35 - Pie Chart EXTWALL

EXTWALL_D (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
Aluminum	27	0.90%
Brick Veneer	32	1.07%
Brick/Siding	152	5.07%
Brick/Stone	13	0.43%
Brick/Stucco	18	0.60%
Common Brick	2280	76.05%
Concrete	6	0.20%
Concrete Block	1	0.03%
Default	1	0.03%
Face Brick	12	0.40%
Hardboard	1	0.03%
Metal Siding	4	0.13%
Shingle	28	0.93%

Stone	20	0.67%
Stone Veneer	5	0.17%
Stone/Siding	10	0.33%
Stone/Stucco	8	0.27%
Stucco	88	2.94%
Stucco Block	3	0.10%
Vinyl Siding	147	4.90%
Wood Siding	142	4.74%

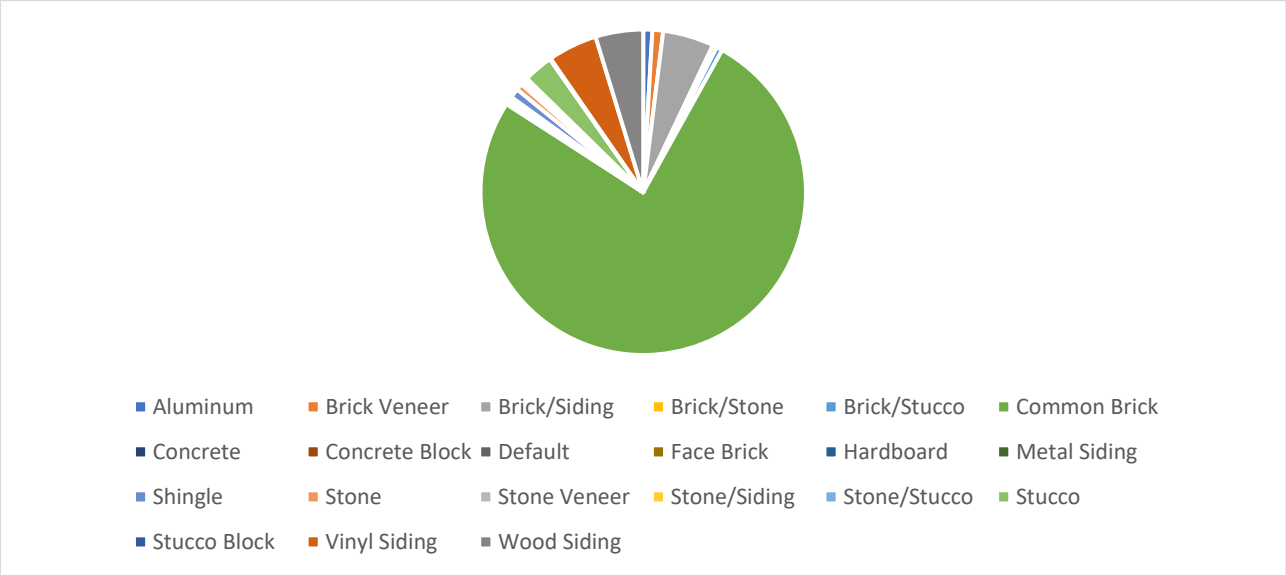

Figure 36 - Pie Chart EXTWALL_D

ROOF (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
0	6	0.20%
1	851	28.39%
2	906	30.22%
3	11	0.37%
4	34	1.13%
5	6	0.20%
6	807	26.92%
7	1	0.03%
8	3	0.10%
10	22	0.73%
11	309	10.31%
13	42	1.40%

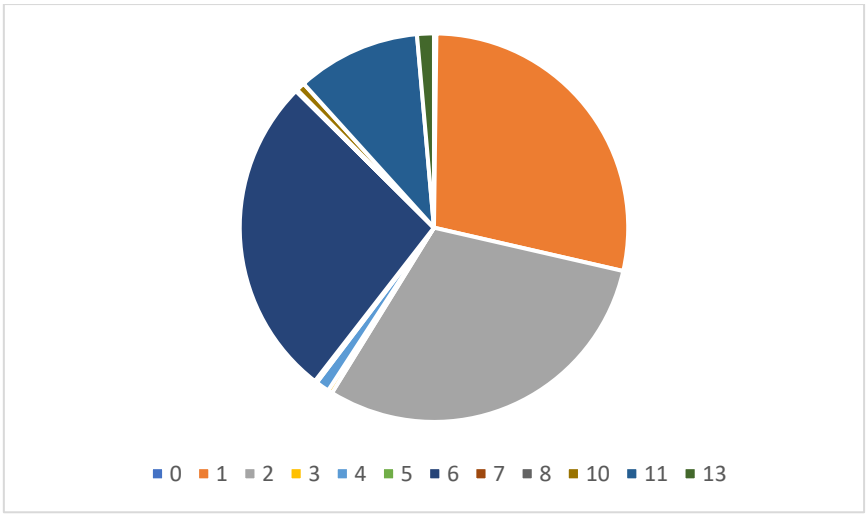

Figure 37 - Pie Chart ROOF

ROOF_D (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
Built Up	906	30.22%
Clay Tile	22	0.73%
Comp Shingle	851	28.39%
Composition Ro	3	0.10%
Metal- Cpr	1	0.03%
Metal- Pre	6	0.20%
Metal- Sms	807	26.92%
Neopren	42	1.40%
Shake	34	1.13%
Shingle	11	0.37%
Slate	309	10.31%
Typical	6	0.20%

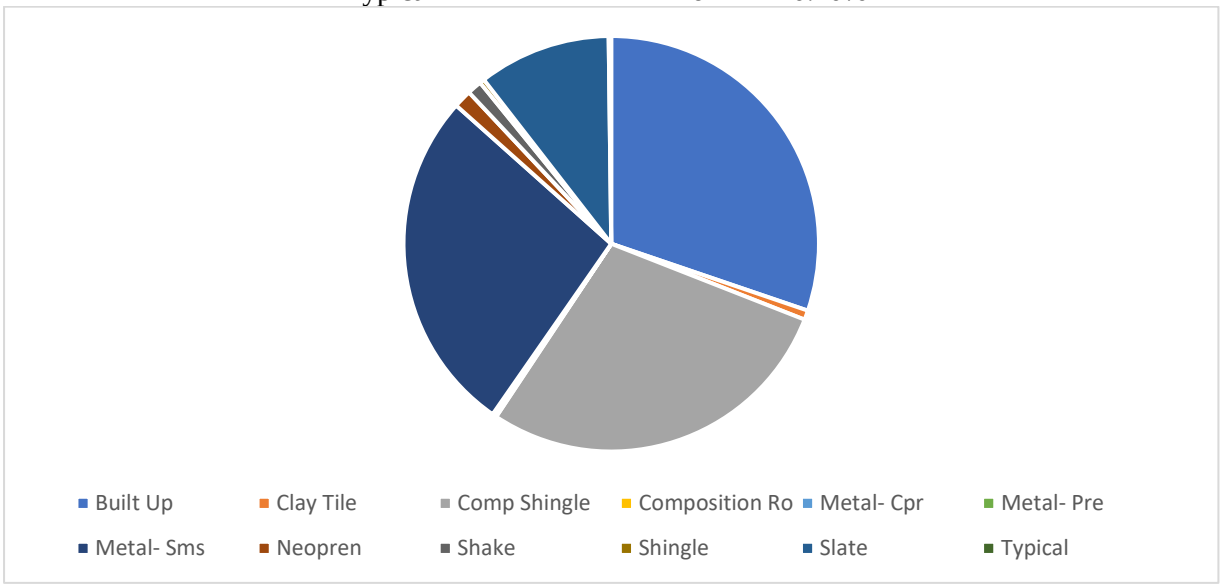

Figure 38 - Pie Chart ROOF_D

INTWALL (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
0	2	0.07%
2	104	3.47%
3	263	8.77%
4	1	0.03%
6	2316	77.25%
10	4	0.13%
11	308	10.27%

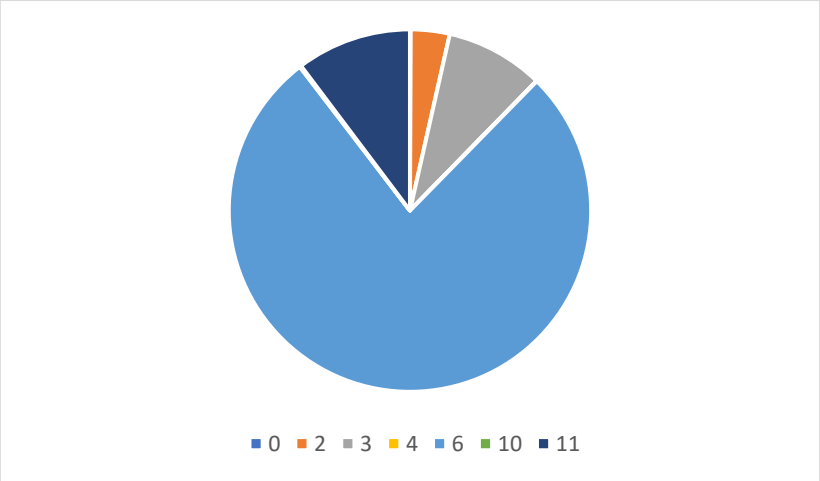

Figure 39 - Pie Chart INTWALL

INTWALL_D (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
Carpet	104	3.47%
Ceramic Tile	1	0.03%
Default	2	0.07%
Hardwood	2316	77.25%
Hardwood/Carp	308	10.27%
Lt Concrete	4	0.13%
Wood Floor	263	8.77%

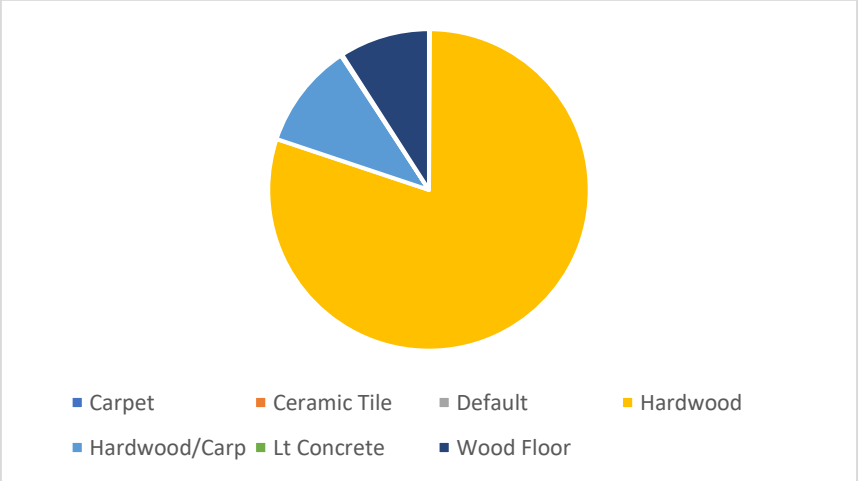

Figure 40 - Pie Chart INTWALL_D

KITCHENS (Attribute: Ratio)

Min	0
Max	44
Range	44
Mean	1.241494
Median	1
Std Dev.	1.007009
Variance	1.014067
75th Percentile	1
25th Percentile	1
Missing Values	2

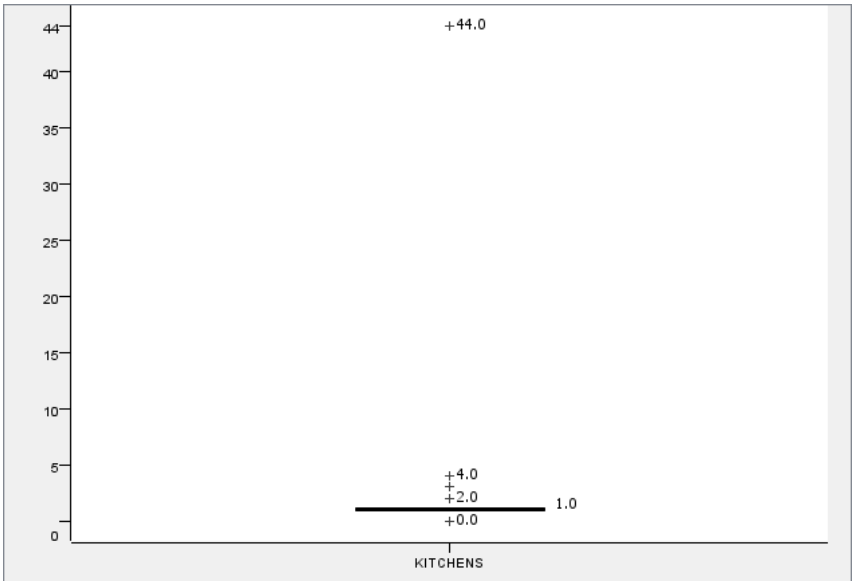

Figure 41 - Box Plot KITCHENS

Value	Absolute Frequency	Relative Frequency
0	6	0.20%
1	2543	84.82%
2	306	10.21%
3	45	1.50%
4	97	3.24%
44	1	0.03%

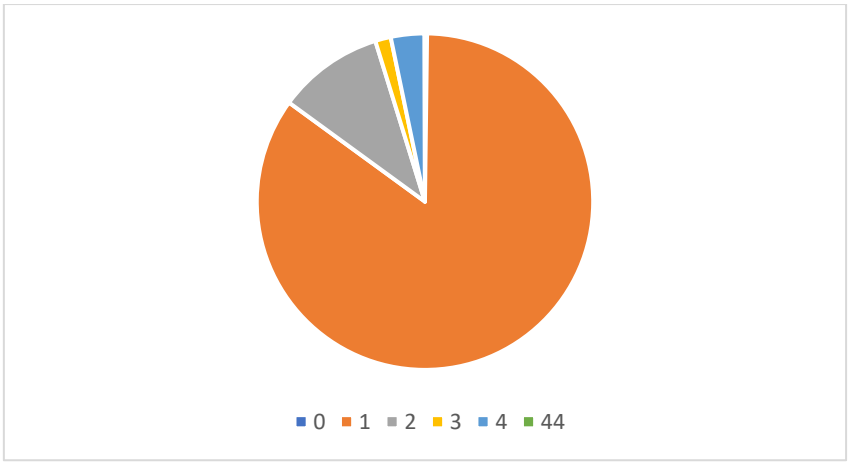

Figure 42 - Pie Chart KITCHENS

FIREPLACES (Attribute: Ratio)

Min	0
Max	6
Range	6
Mean	0.596398
Median	0
Std Dev.	0.831006
Variance	0.690571
75th Percentile	1
25th Percentile	0
Missing Values	2

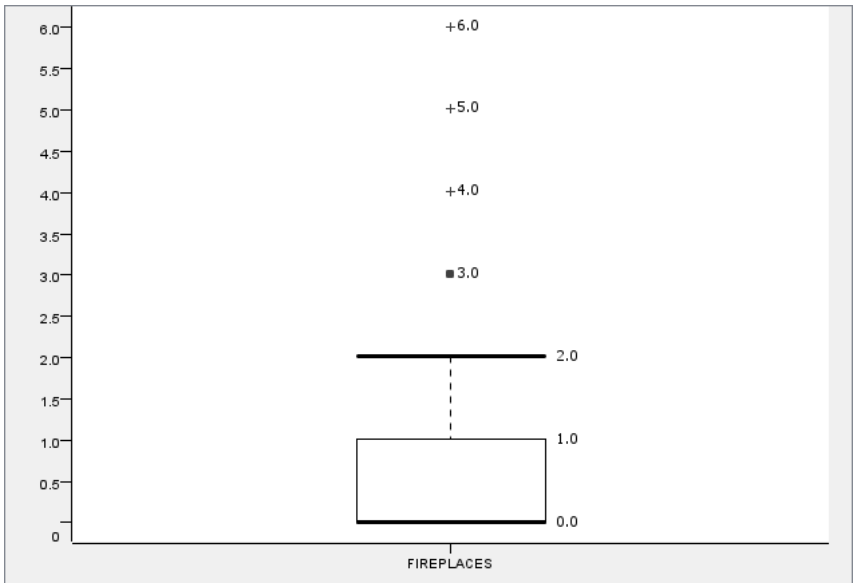

Figure 43 - Box Plot FIREPLACES

Value	Absolute Frequency	Relative Frequency
0	1699	56.67%
1	943	31.45%
2	264	8.81%

3	60	2.00%
4	25	0.83%
5	5	0.17%
6	2	0.07%

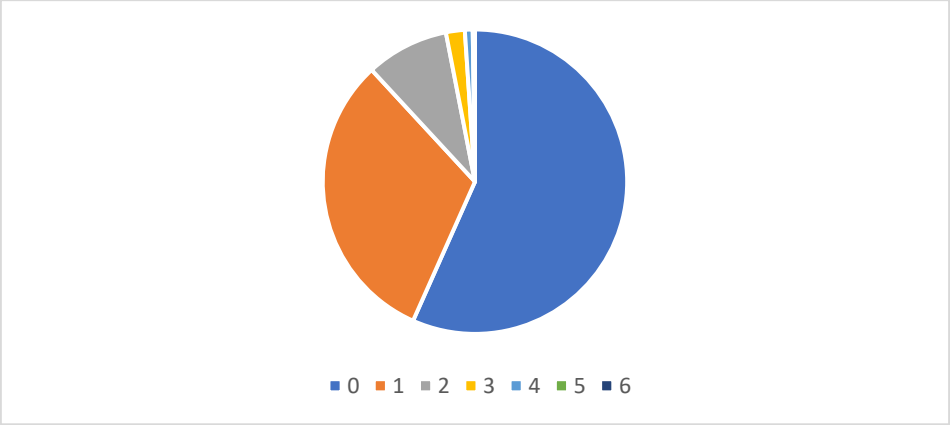

Figure 44 - Pie Chart FIREPLACES

USECODE (Attribute: Nominal)

Value	Absolute Frequency	Relative Frequency
0	2	0.07%
11	1277	42.57%
12	867	28.90%
13	484	16.13%
19	1	0.03%
23	128	4.27%
24	241	8.03%

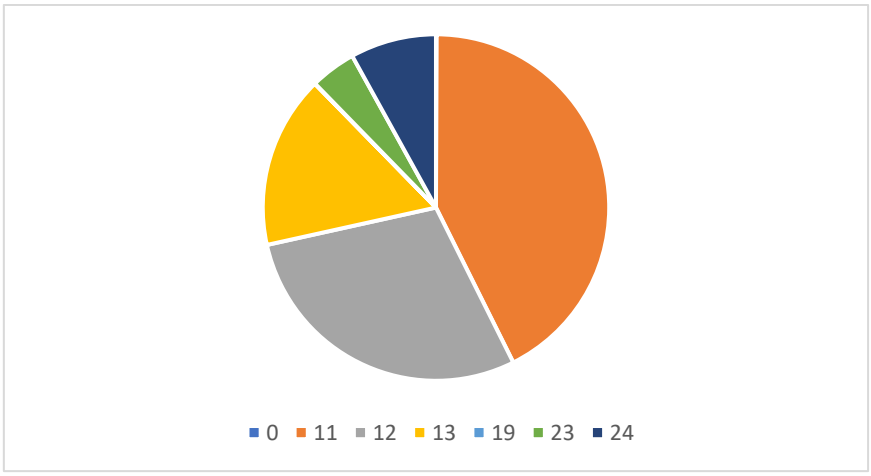

Figure 45 - Pie Chart USECODE

LANDAREA (Attribute: Ratio)

Min	52
Max	198634
Range	198582
Mean	3331.893667
Median	2335

Std Dev.	4531.531913
Variance	20534781.48
75th Percentile	4000
25th Percentile	1600
Missing Values	0

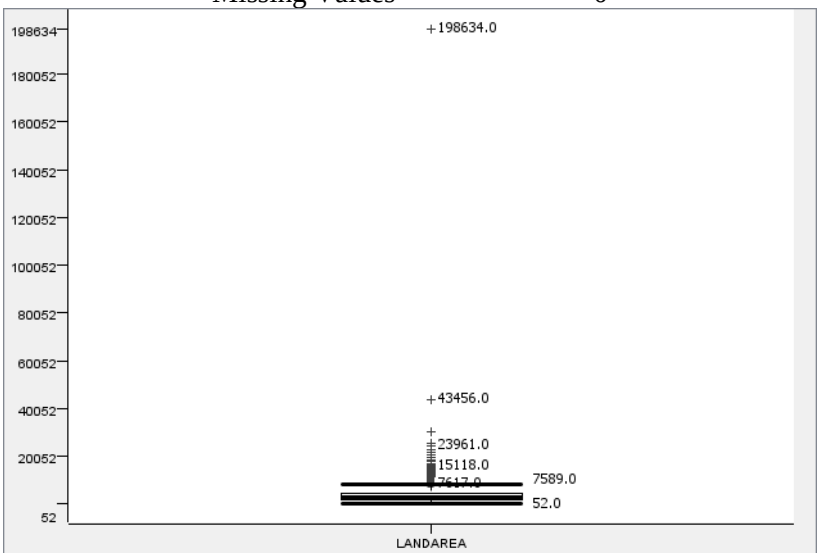

Figure 46 - Box Plot LANDAREA

GIS_LAST_MOD_DTTM (Attribute: Interval)

Value	Absolute Frequency	Relative Frequency
2018-07-22T18:01:43.000Z	3000	100.00%

QUALIFIED

Value	Absolute Frequency	Relative Frequency
0	1700	56.67%
1	1300	43.33%

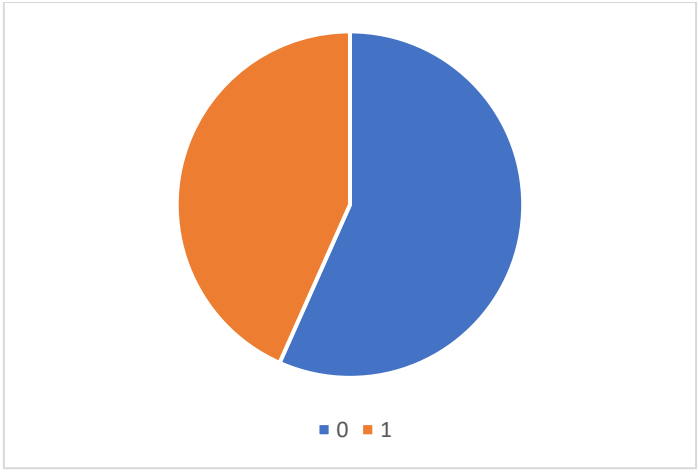

Figure 47 - Pie Chart QUALIFIED

1.A.III Dataset Exploration

Following the statistical and frequency analysis presented in section 1.A.II the following points of interest and outliers were found:

Attribute	General Commentary
BATHRM	Potential outliers for the BATHRM attribute include the 0, 7, 8, 9 and 10 values. Most of the values in this attribute appear to be clustered around the median with a small standard deviation.
HF_BATHRM	Properties with 3 or 4 half bathrooms appear to be outliers in this dataset.
HEAT/HEAT_D	As expected, these two attributes are strongly correlated. Most properties appear to have either Forced Air, Hot Water Rad or Warm Cool Heating (~98%).
AC	There appears to be one incorrectly entered value in the dataset for this attribute with value '0'.
NUM_UNITS	This attribute is normally distributed around 1 unit.
ROOMS	This attribute is normally distributed around 7 rooms, there are some potential outliers above 14 rooms.
BEDRM	This attribute is normally distributed around 3 bedrooms, there are some potential outliers above 7 bedrooms.
AYB	There is one possible error value (0) in this attribute.
YR_RMDL	The handful of values prior to 1938 appear to be potential outliers
EYB	There is one possible error value (0) in this attribute.
STORIES	Most instances in the dataset appear to have 2 Stories.
SALEDATE	There is a large gap between the median and the mean. There are a few outliers prior to 1969
PRICE	This attribute has quite a large range and a large number of 0 values. There are 6 major outliers in this dataset.
GBA	This attribute is normally distributed around the median with a very long tail of outlier values.
BLDG_NUM	This attribute is almost entirely uniform apart from three values of '2'.
STYLE/ STYLE_D	Most instances in the dataset appear to have 2 Stories.
GRADE/ GRADE_D	Both GRADE Attributes are normally distributed with values centred around a '3' or 'Average' grade.

CNDTN/ CNDTN_D	Both CNDTN Attributes are normally distributed with values centred around a '3' or 'Average' grade.
EXTWALL/ EXTWALL_D	Common Brick or a combination of brick with another material was found in the overwhelming majority of instances in the dataset.
INTWALL/ INTWALL_D	Hardwood or a combination of Hardwood with Carp are the most common material found in the properties in the dataset (~87.5%).
KITCHENS	Most instances in the dataset appear to have less than two kitchens. There is one entry with a value of '44' which could be an error.
FIREPLACES	Most instances in the dataset appear to have less than two fireplaces.
LANDAREA	The values are generally clustered around the median with a handful of outliers. The '198634' value appears to be an incorrectly entered value or a significant outlier as it is several orders of magnitude larger than the next highest value
GIS_LAST_ MOD_DTTM	This is the only attribute in the dataset that is completely uniform with all instances having the same value of '2018-07-22T18:01:43.000Z'.

Linear Correlation

The Linear Correlation node in Knime has been used to identify attributes within the dataset that have strong correlations. It can be observed in the figure below that there are some interesting correlations (positive and negative) between the following attribute pairs: GBA/ROOMS (0.6696), ROOMS/BEDRM (0.7003), USECODE/NUM_UNITS (0.8186), STORIES/STYLE (0.6182), CNDTH/EYB (0.6055), KITCHENS/NUM_UNITS (0.6445), LANDAREA/STRUCT (-0.404).

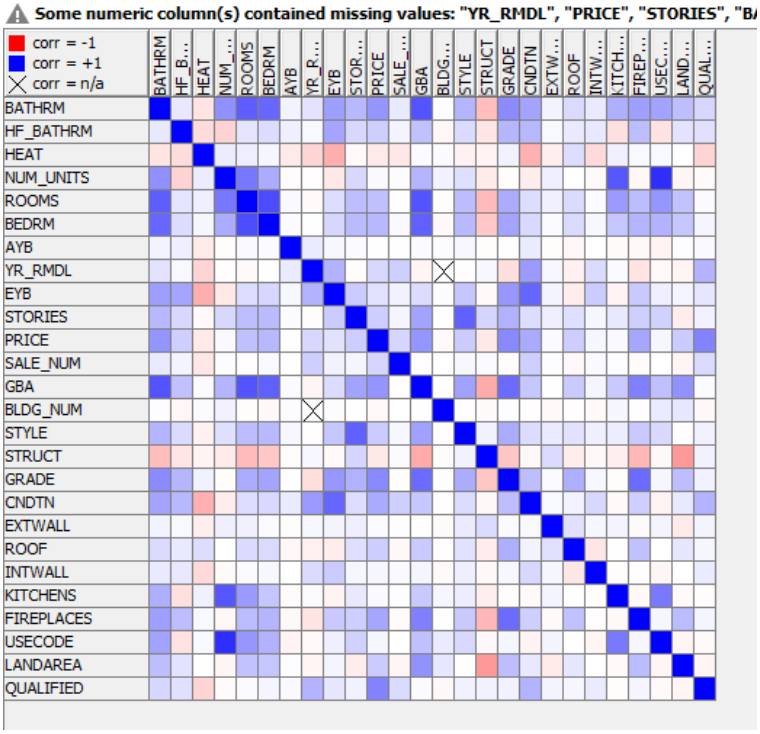

Figure 48 - Linear Correlation of Dataset Attributes

Setup of Clustering Analysis

Following the linear correlation analysis, the k-Means Clustering Node and Scatter Plot tool are used to visually highlight potential clusters within the dataset. The k-Means method with k=3 was selected over the Hierarchical approach due to its time and computational efficiency.

To accommodate the k-Means method in Knime the missing values in the dataset are modified using the 'Missing-Value' tool. Missing string values were replaced with the most frequent value and missing number values were replaced with the mean.

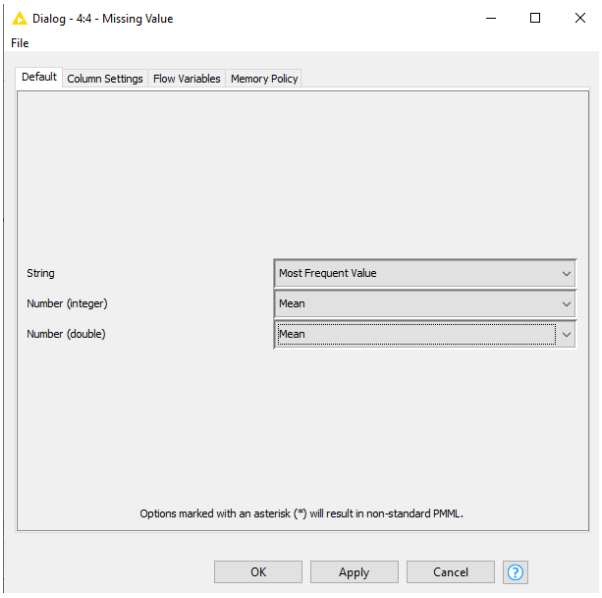

Figure 49 - Manipulating missing values in Knime

Comparison of Price to Condition and Rooms

Price was found to be correlated to the condition and grade attributes. From the cluster analysis we find three distinct clusters: 'Average', 'Very Good' and the much wider 'Exceptional'.

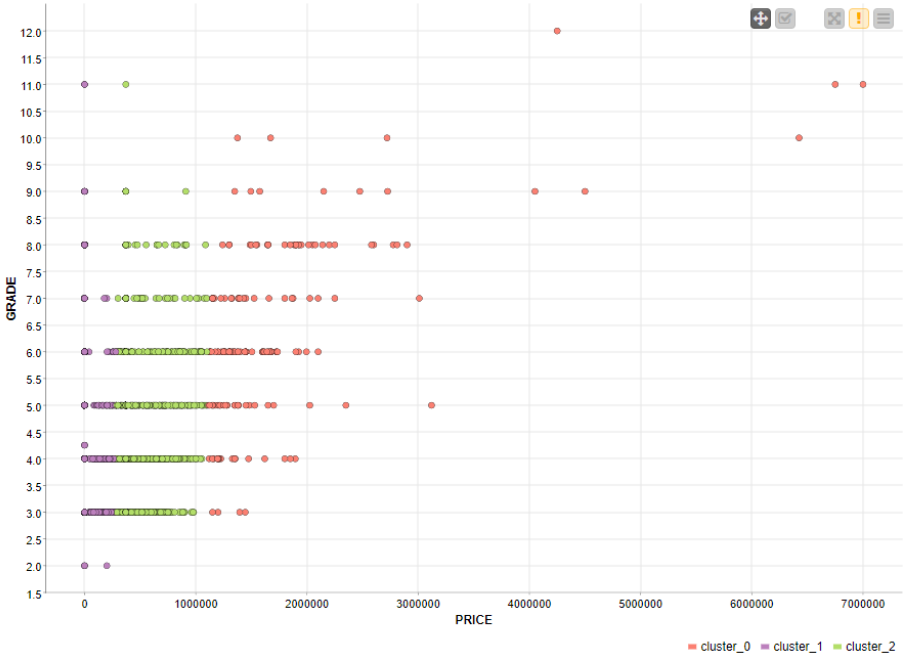

Figure 50 - Clustering Grade and Price

A similar pattern emerges when Rooms and Price are put through the k-means clustering analysis. Typically, as the price of the property increases so too does the number of rooms.

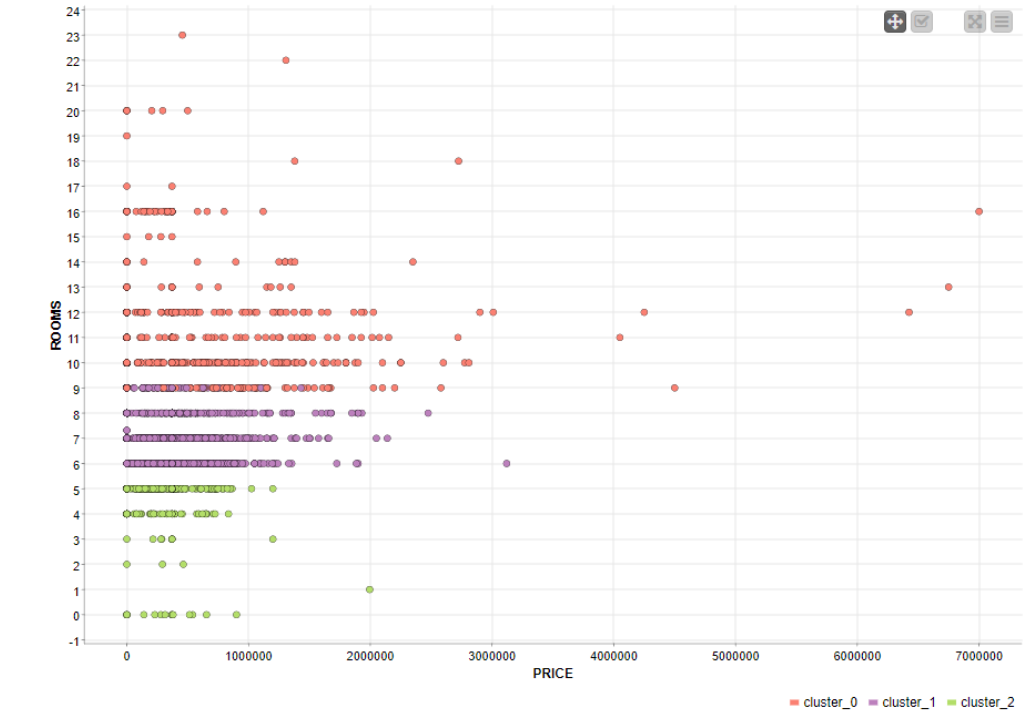

Figure 51 - Clustering Rooms and Price

Comparison of Land Area to Building Area

Building area tends to be much smaller than the total land area, indicating that properties with front/back yards were prevalent in the dataset. Properties with larger building footprints tended to take a larger proportion of the total land area.

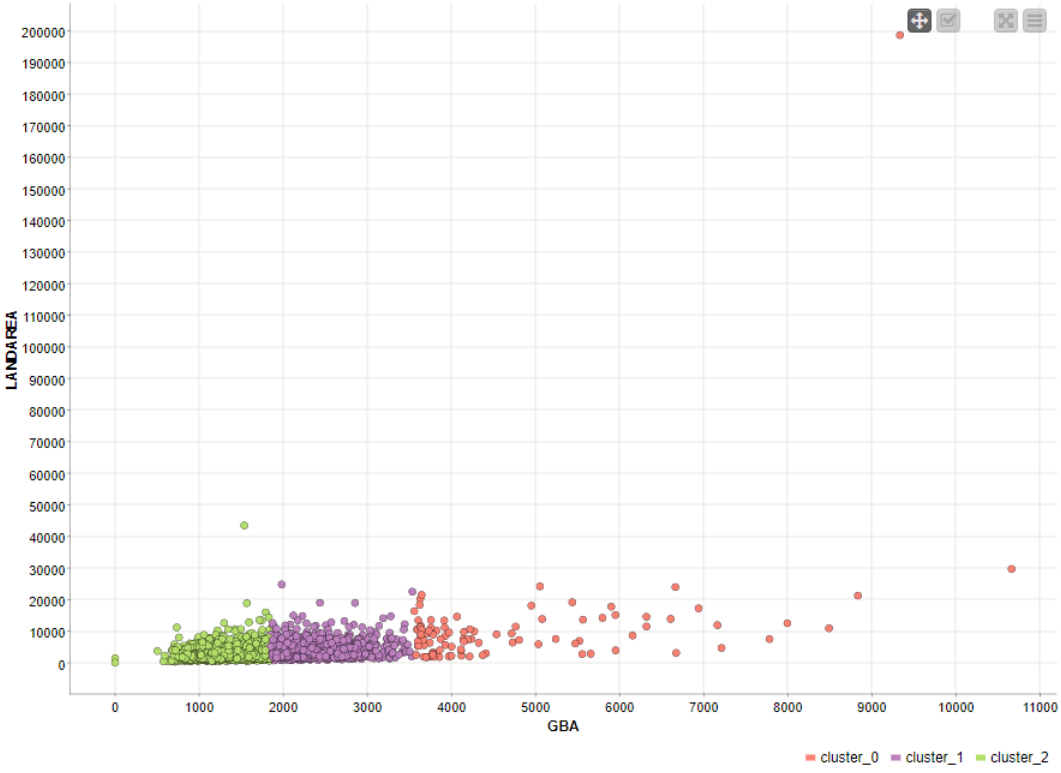

Figure 52 - Clustering Land Area to Building Area

Comparison of Price to Building Area

Prices appear to be clustered around the averages, the price and area are somewhat correlated with some discrepancies in the upper range outliers.

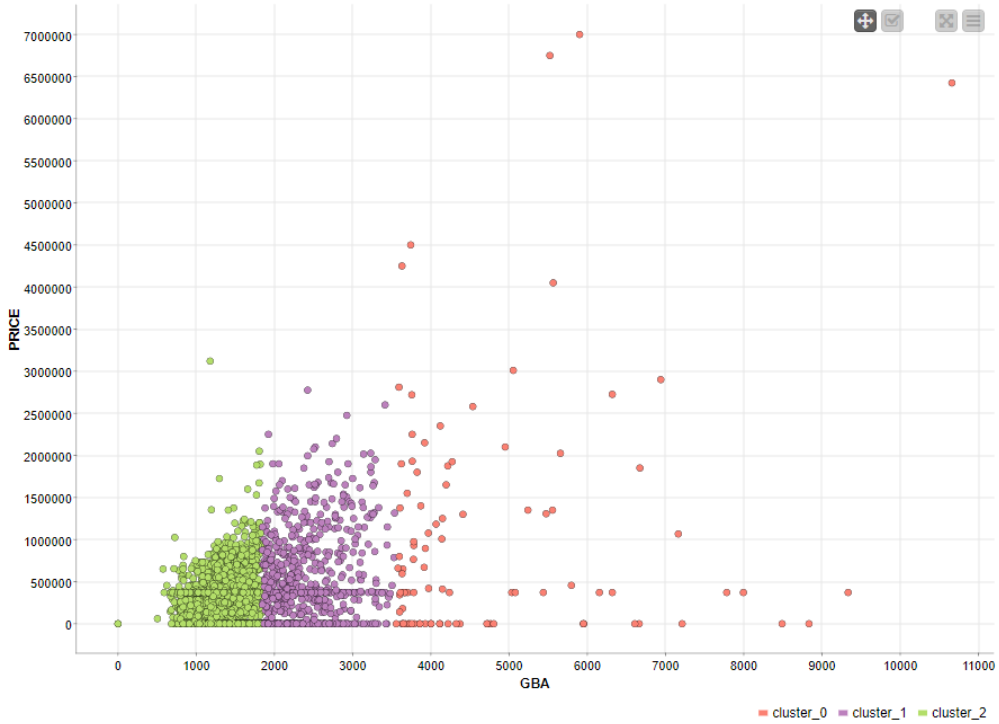

Figure 53 - Clustering GBA and Price

Clustering of Heat Description and AC

Another interesting observation was the comparison of Heating against Air conditioning. Properties with 'Hot Water Rad' tended not to have any Air conditioning. In contrast properties with 'Warm Cool' or 'Forced Air' tended to have AC units.

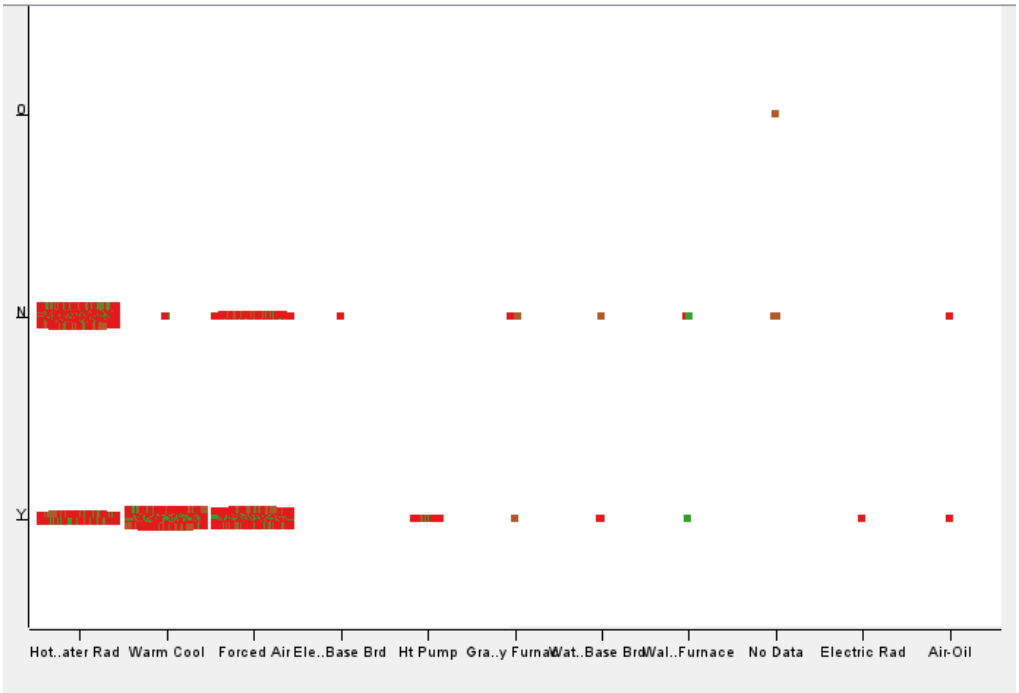

Figure 54 - Clustering AC and Heat_D

1.B Data Pre-processing

Insightful data analysis relies on data that has been properly prepared and is free of errors, this section of the report explains the use of several data pre-processing methods and the benefits of utilising these methods.

Data mining methods and machine learning techniques are out of scope for this report, however should further analysis take place it is necessary to understand the importance of this data pre-processing step.

1.B.I Utilising Binning techniques on the Price Attribute

Histograms are often used to provide an efficient graphical representation of the distribution of numerical data by sorting and categorising values into a selection of appropriately sized "bins".

A number of different methods to select the appropriate number of bins for the Price attribute were considered, these methods included: the Square-Root choice, Sturges formula and the Freedman-Diaconis rule

Ultimately, the Freedman-Diaconis rule was used due to the large number of outliers in the dataset and non-normal distribution.

The number of bins (k) and the bin width (h) are given by the formulas:

$$k = \frac{\max x - \min x}{h} \qquad h = 2\frac{IQR}{n^{\frac{1}{3}}}$$

Using the values calculated in section 1.A.II and excluding the missing values:

$$h = 2\frac{564000}{2481^{\frac{1}{3}}} = 83323.37 \qquad k = \frac{7000000-0}{83323.37} = \sim 84$$

Using the Auto-Binner and histogram nodes in Knime:

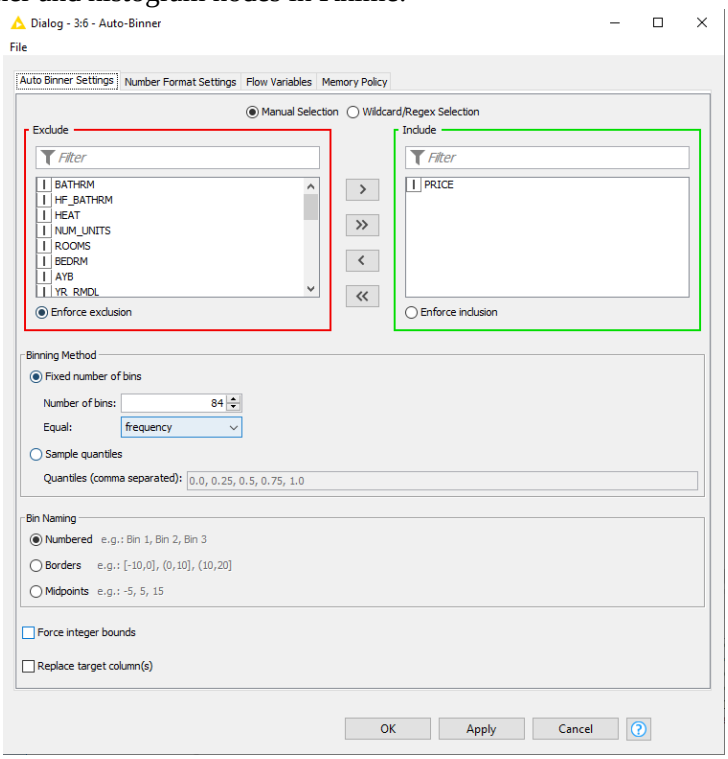

Figure 55 - Equi-Depth Bins

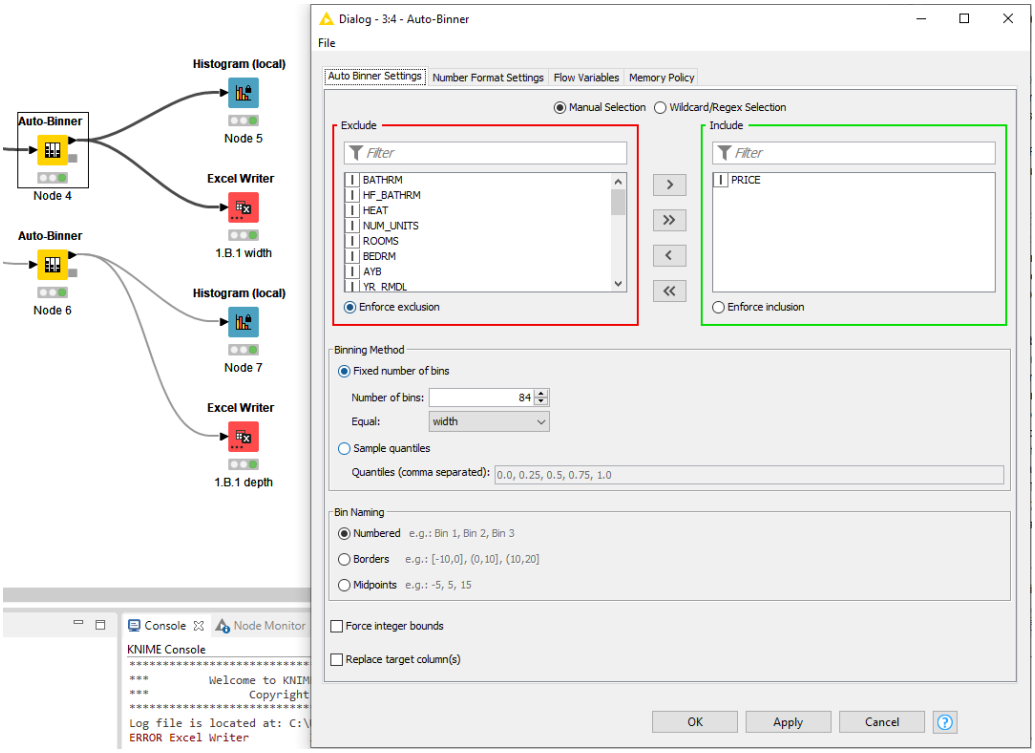

Figure 56 – Equi-width Bins

1.B.II Normalising the Price Attribute

In this report the min-max normalisation and z-score normalisation techniques have been used to standardise the values of the Price attribute. The data normalisation process attempts to provide equal weight to each value in a sample, effectively reducing the effect of units of measurement.

The min-max normalisation method performs a linear transformation over the source data, scaling the values to a range between 0.0 and 1.0. For this range, the formula may be simply express as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Similarly, the z-score normalisation method normalises values using the mean (μ) and standard deviation (σ). The z-score normalisation formula is given as:

$$x' = \frac{x - \mu}{\sigma}$$

In Knime the Normalizer node has been used twice (once with each technique) to produced normalised results.

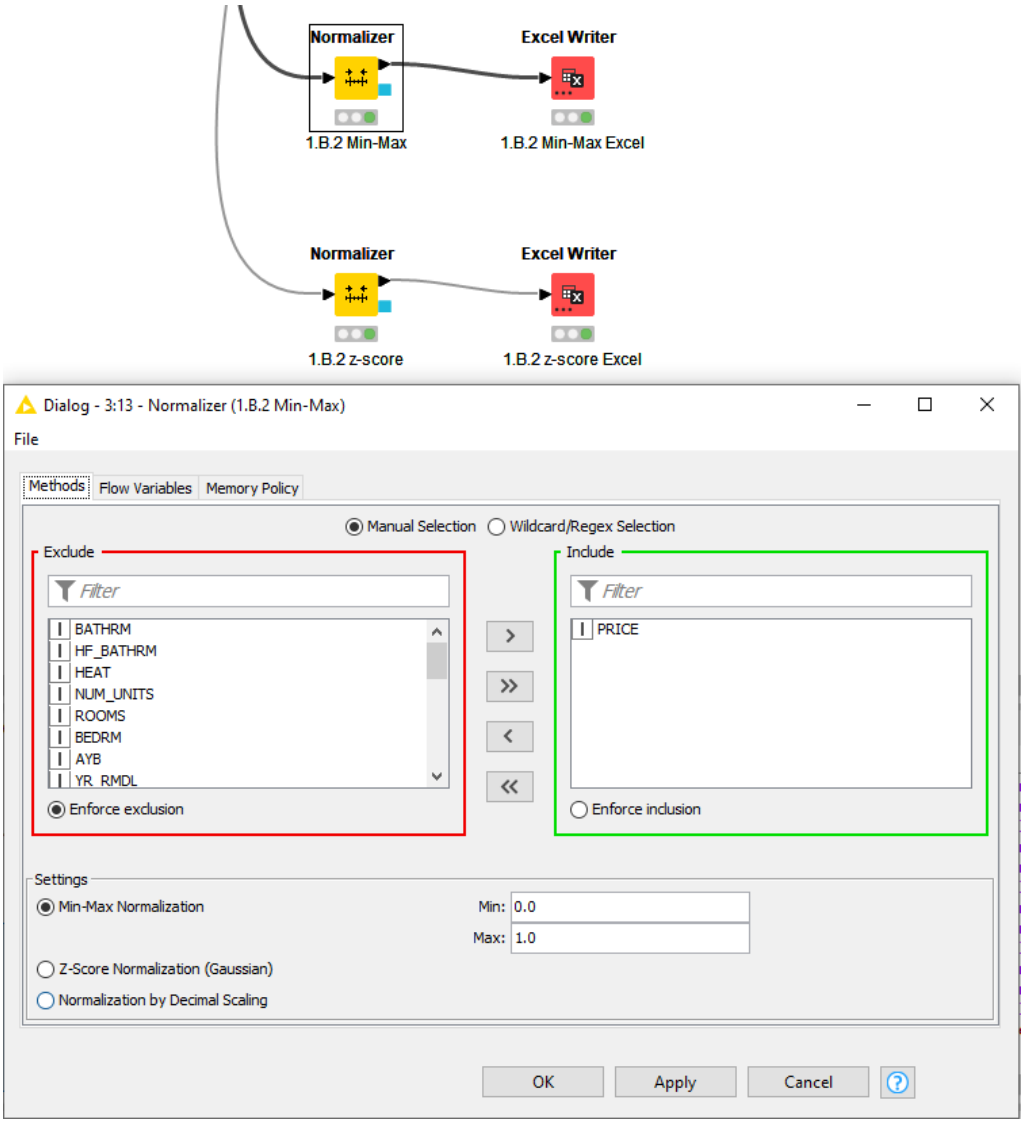

Figure 57 - Normalisation of Price in Knime

1.B.III Discretisation of the Price Attribute

Discretisation refers to the process of transforming a continuous attribute into a categorical or discrete attribute. In this analysis the Price attribute of each row will be discretised into four categories to provide a relative measure of the expense of the property.

As specified in the report brief, the following categories will be used to categorise the Price attribute: Low, Medium, High and Expensive (e.g.: Low=0-50k; Medium=50k-100k; High=100k-1000k; Expensive= 1000k+).

The results of this discretisation process are listed in the table below, highlighting the frequency of each bin:

Bin	Frequency
Low	871
Medium	64
High	1346
Expensive	200

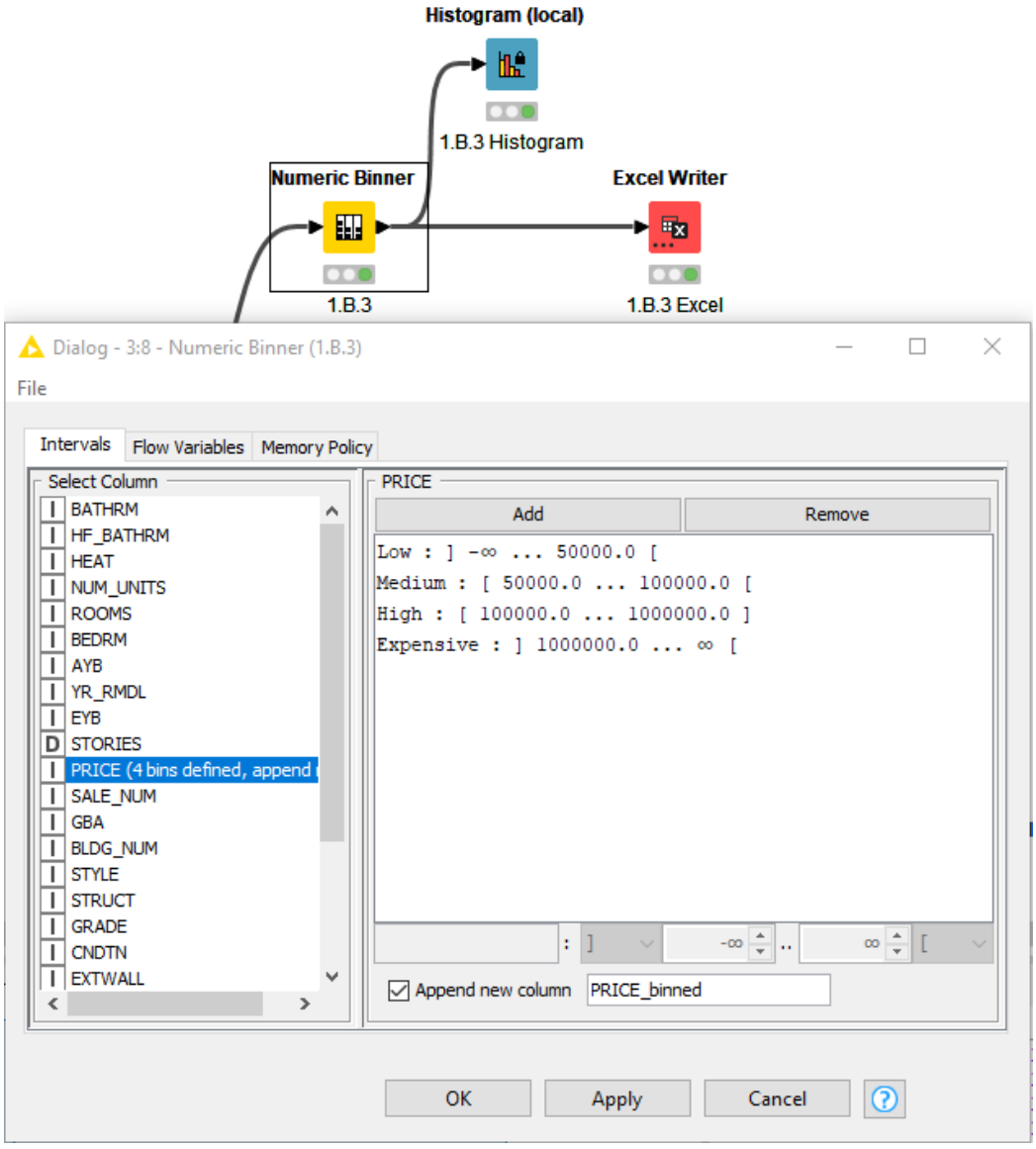

Figure 58 - Discretisation in Knime

1.B.IV Binarisation of the Struct_D Attribute

Binarisation is the process of transforming a continuous attribute into on or more binary vectors. This process is often used as a pre-processing step for further analysis via data mining algorithms or machine learning methods.

In this report the STRUCT_D attribute has been binarised into a further set of seven variables. In Knime the "One to Many" node has been used for the Binarisation process as shown in the figure below. The results have been then exported to Excel for further analysis.

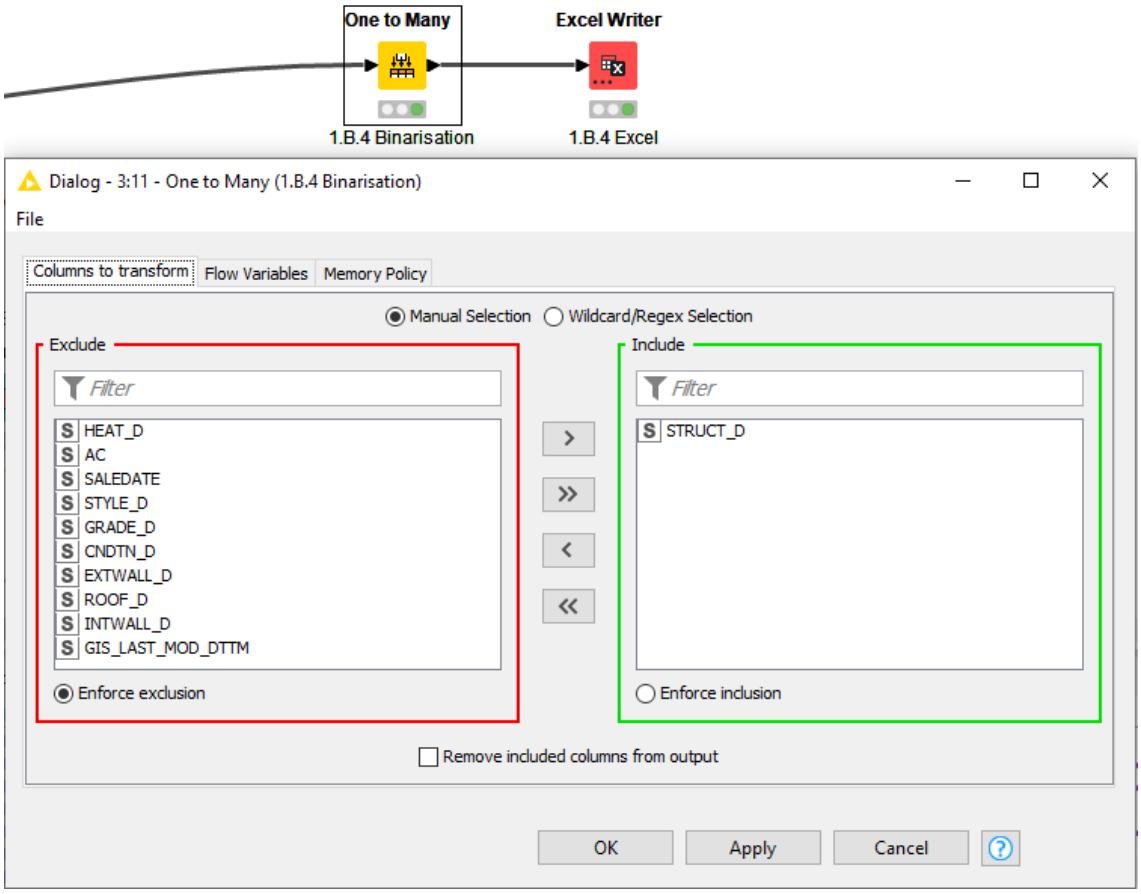

Figure 59 - Binarisation of the STRUCT_D attribute

1.C Summary

This report has provided an in-depth explanation of the steps and methodology required to conduct the initial data exploration and data pre-processing steps of the analytics process. In putting together this report a number of interesting findings were noted.

Unsurprisingly there was a strong linear correlation between the gross building area and the number of rooms, kitchens, bedrooms, fireplaces. In essence larger properties had more features. In the same vein, larger properties tended to have higher condition and grade ratings as well as fetch higher prices.

From the statistical analysis it was interesting to find that the overwhelming majority of properties in the dataset were 2 stories, with a brick or brick combination construction. In terms of layouts most properties tended to have a single kitchen, one or two bathrooms and on average three bedrooms. The most expensive property in the dataset which was a major outlier had 16 rooms, 7 bedrooms and predictably graded very highly.

It is important to note that in this analysis there were some deficiencies in the quality of the dataset particularly with the Price and Year Remodelled attributes. Over half of the YR_RMDL values were missing and nearly 20% of the price values were blank. Curiously nearly a third of the values in the Price attribute were also found to be zero possibly as a result of issues in the data collection process.

In the second major section of this report, data pre-processing steps were taken to clean up the PRICE and STRUCT_D attributes. The discretisation of the property prices revealed that most of the property prices were considered ‘high’.

Should further analysis in the form of data mining or machine learning take place the dataset has already been cleaned up in the data pre-processing step and the data exploration exercise has revealed some interesting inter-attribute relationships.