Assignment 2 -31250 Introduction to Data Analytics

DATA EXPLORATION & PREPERATION

Piyush Vats

Student #13046899

Introduction

The aim of this report is to provide insight into provided data set regarding properties sold and relating attributes. A thorough analysis of each attribute will be performed, highlighting important statistical figures and interesting observations.

This report consists of three main parts; initial data exploration on an attribute level, pre-processing of the dataset to help further interpret the data and a summary where all findings and observations will be highlighted.

Initial Data Exploration

Each attribute is identified and explored via statistical analysis when possible and graphically represented using KNIME and its plethora of tools. Outliers are identified on an attribute basis with interesting observations noted. Observation regarding clusters will be noted at the end of this section.

Attributes

1. ID

Attribute Type:

Nominal. Consists of distinct numbers that have no relation to one another. Sole purpose is to uniquely identify the data.

Range:

3-107 – 107,012

2. SSL

Attribute Type:

Nominal.

The description provided for this attribute is insufficient. Through analysis, it was concluded that it cannot be ordered, thus value acts as just a label.

3. BATHRM

Attribute Type:

Interval.

The value of BATHRM signifies how many bathrooms said property consists of. As such, it can be used to order properties. Not only that, difference between values is meaningful.

Statistics:

Statistic	Value
Range	1 - 7

Median	2
Mean	2.022
Variance	1.136
Standard Deviation	1.066
1st Quartile	1
3rd Quartile	3

Distribution & Frequency:

As can be seen, all properties have at least 1 bathroom with 18 properties having 6 or more. The box plot shows that the interquartile range is between 1 and 3 and it can be concluded as a result that most properties have between 1 ad 3 bathrooms with 7 being an outlier.

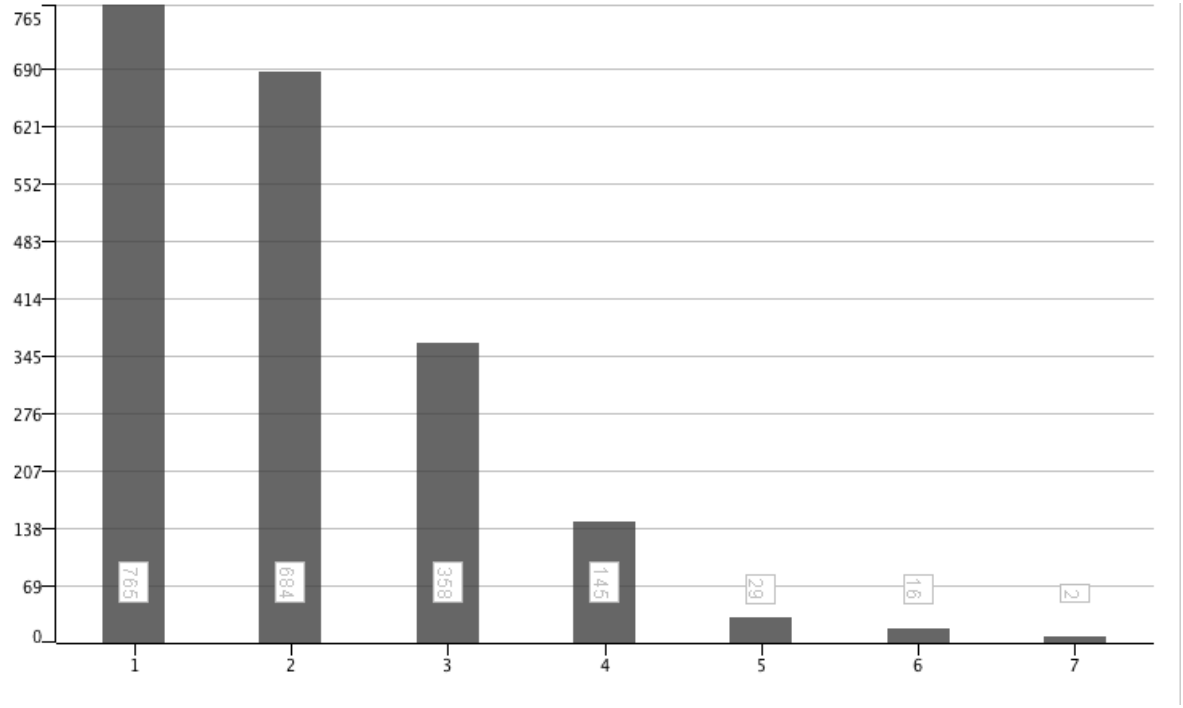

Figure 1 – Histogram distribution of No. of bathrooms

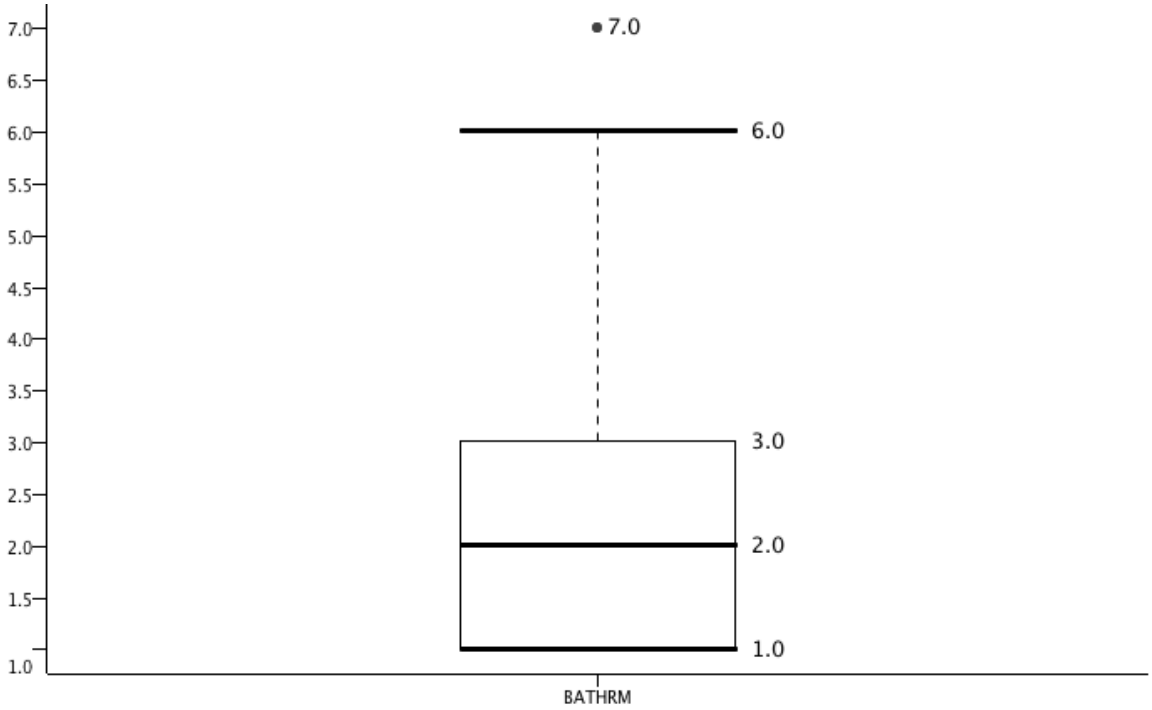

Figure 2 – Box Plot distribution of No. of bathrooms

4. HF BATHRM

Attribute Type:

Interval.

The value of HF BATHRM signifies how many half bathrooms said property consists of. As such, it can be used to order properties. Not only that, difference between values is meaningful.

Statistics:

Statistic	Value
Range	0 - 5
Median	1
Mean	.6118
Variance	0.38
Standard Deviation	.616
1st Quartile	0
3rd Quartile	1

Distribution & Frequency:

As can be seen, the most common no. of half bathrooms is actually 0, closely followed by 1. Interquartile range of 0 to 1 is evident on the box plot and it can be concluded as a result that data is heavily centred there and most properties have between 0 and 1 half bathrooms with 3 and 5 being outliers.

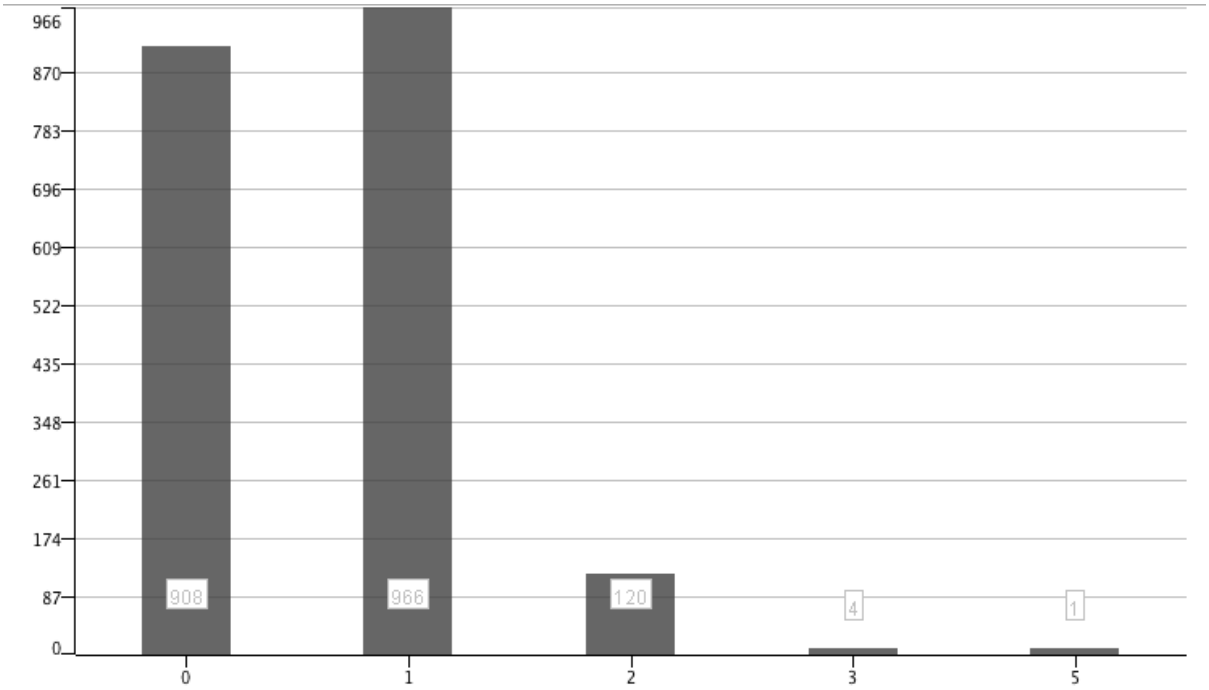

Figure 3 – Histogram distribution of No. of half bathrooms

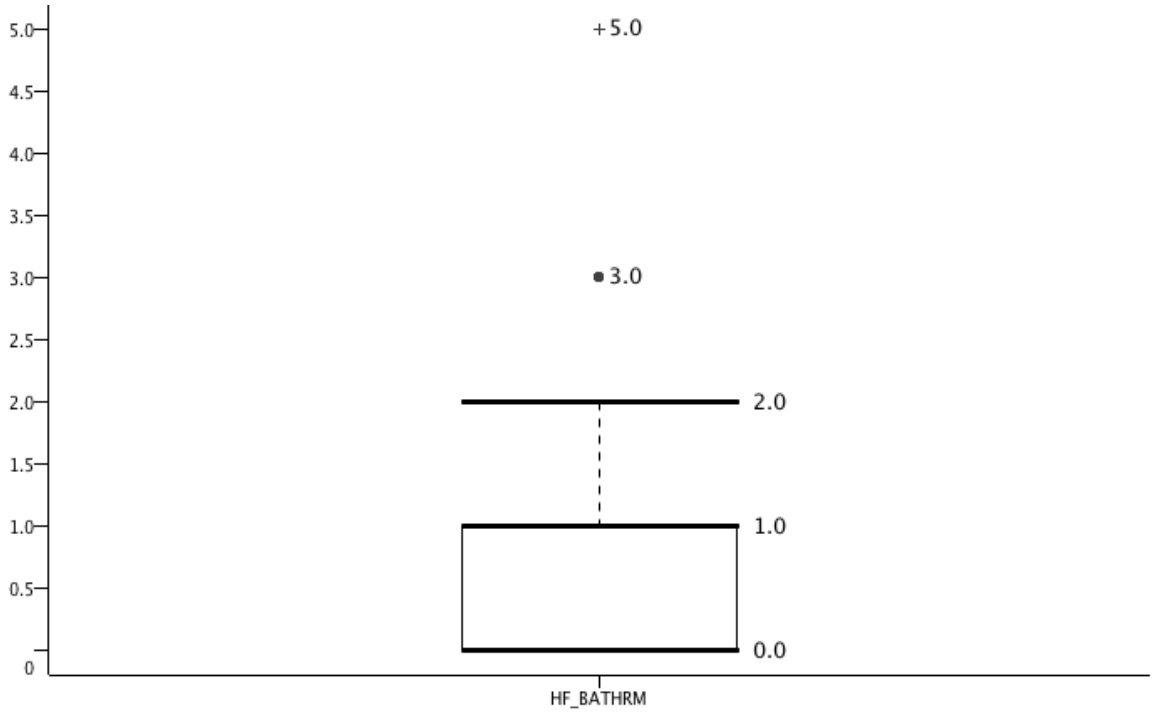

Figure 4 – Box Plot distribution of No. of half bathrooms

5. AC

Attribute Type:

Nominal.

The value of AC cannot be ordered. Although the attribute is represented by numerical values, they are mapped to non-quantitative Boolean, or more simply, yes (Y) or no (N).

Distribution & Frequency:

It can be very easily concluded from the below histogram and pie graph that more properties have Air conditioning (60.31%) than properties that do not (39.69%)

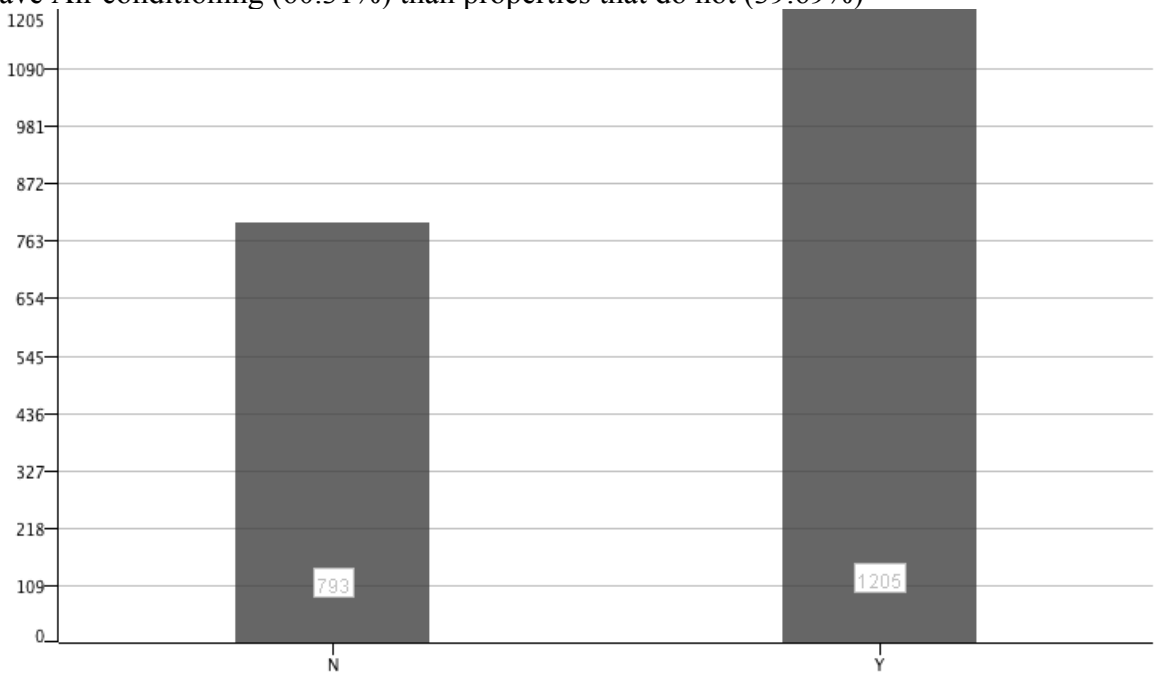

Figure 7 – Histogram distribution illustrating if property has AC or not.

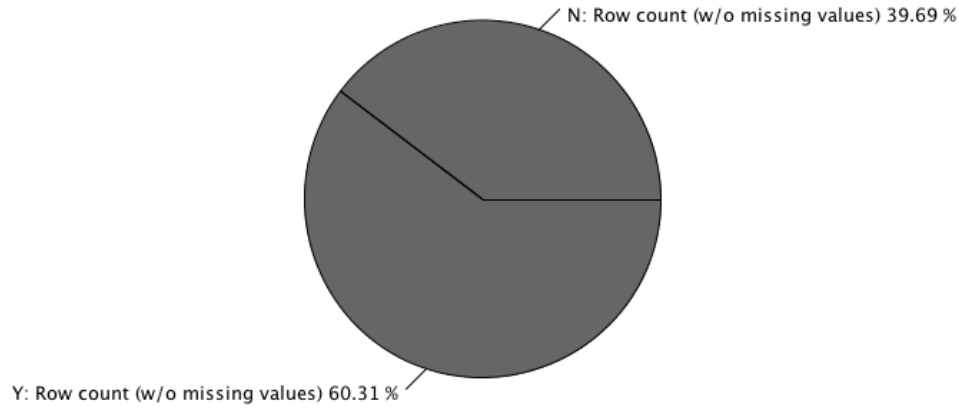

Figure 8 – Pie chart showing top three categories type of heating used by %

6. NUM_UNITS

Attribute Type:

Ratio.

The value of NUM_UNITS signifies how many units said property consists of. As such, it can be used to order properties. Further, exact difference between number of units is known. A property with 1 unit has less units than a property with 5 units with the difference being 4 units.

Statistics:

There was a bit of ambiguity with the data for this specific attribute. There were 6 properties with 0 units, which does not make logical sense. Every property must have at least 1 unit and as such, along with missing values, properties with NUM_UNITS = 0 were ignored from the stats and graphics presented below.

Statistic	Value
Range	1 - 4
Median	1
Mean	1.195
Variance	0.344
Standard Deviation	.586
1st Quartile	1
3rd Quartile	1

Distribution & Frequency:

As can be seen, the most common no. of half bathrooms is 1. Interquartile range of 1 to 1 is evident on the box plot and it can be concluded as a result that most properties have between 1 unit with 2, 3 and 4 being outliers.

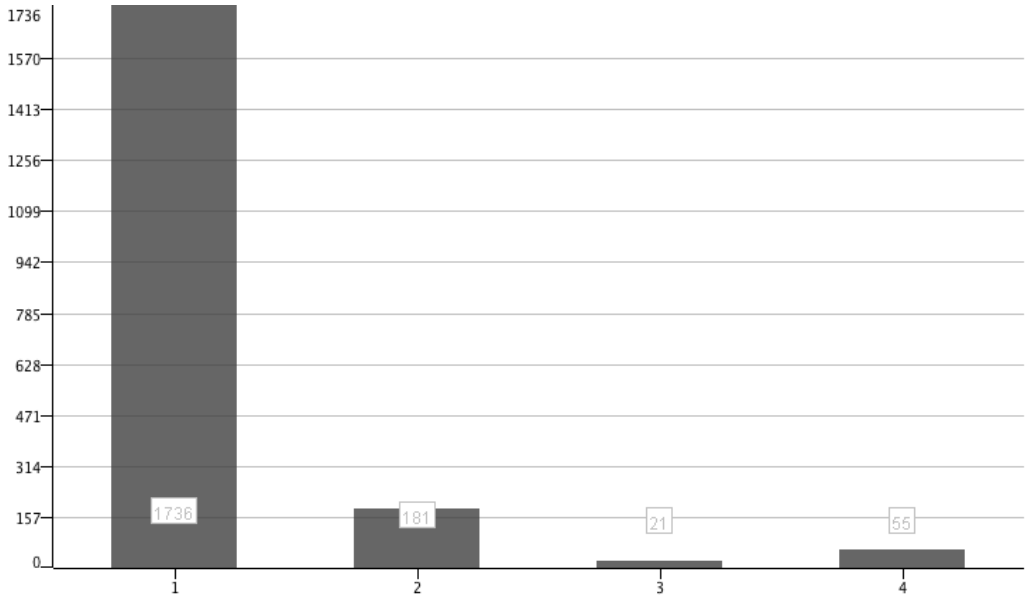

Figure 9 – Histogram distribution of number of units per property

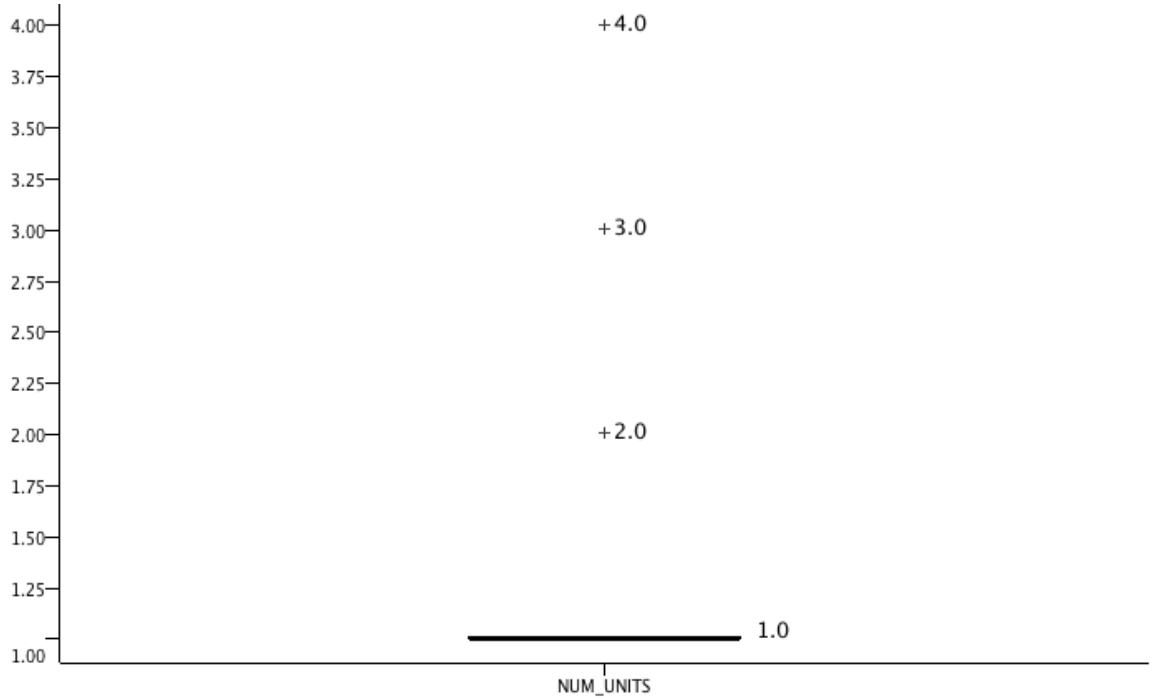

Figure 10 – Box plot of number of units per property

7. ROOMS

Attribute Type:

Ratio.

The value of ROOMS signifies how many rooms said property consists of. As such, it can be used to order properties. Further, exact difference between number of rooms is known. A property with 1 room has less rooms than a property with 5 rooms with the difference being 4 units.

Statistics:

Statistic	Value
Range	0 - 30
Median	7
Mean	7.353
Variance	5.486
Standard Deviation	2.342
1st Quartile	6
3rd Quartile	8

Distribution & Frequency:

As can be seen, the most common no. of rooms per property is 6 (637), followed by 7 (433) and 8 (276). Interquartile range of 6 to 8 observed in the box plot confirms this.

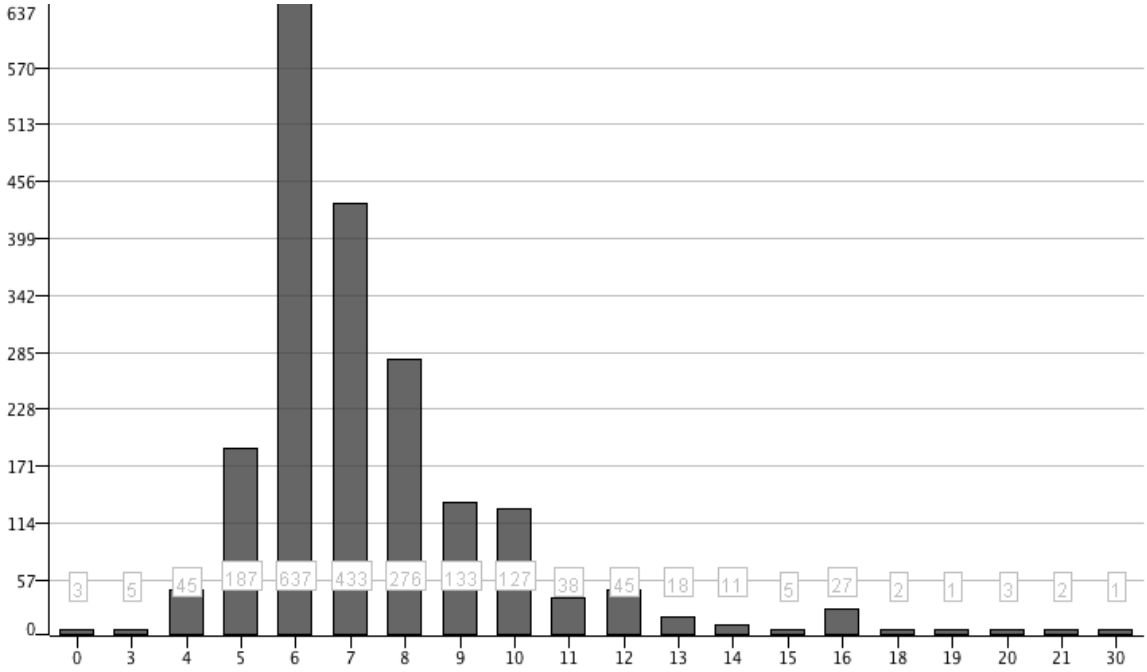

Figure 11 – Histogram distribution of number of rooms per property

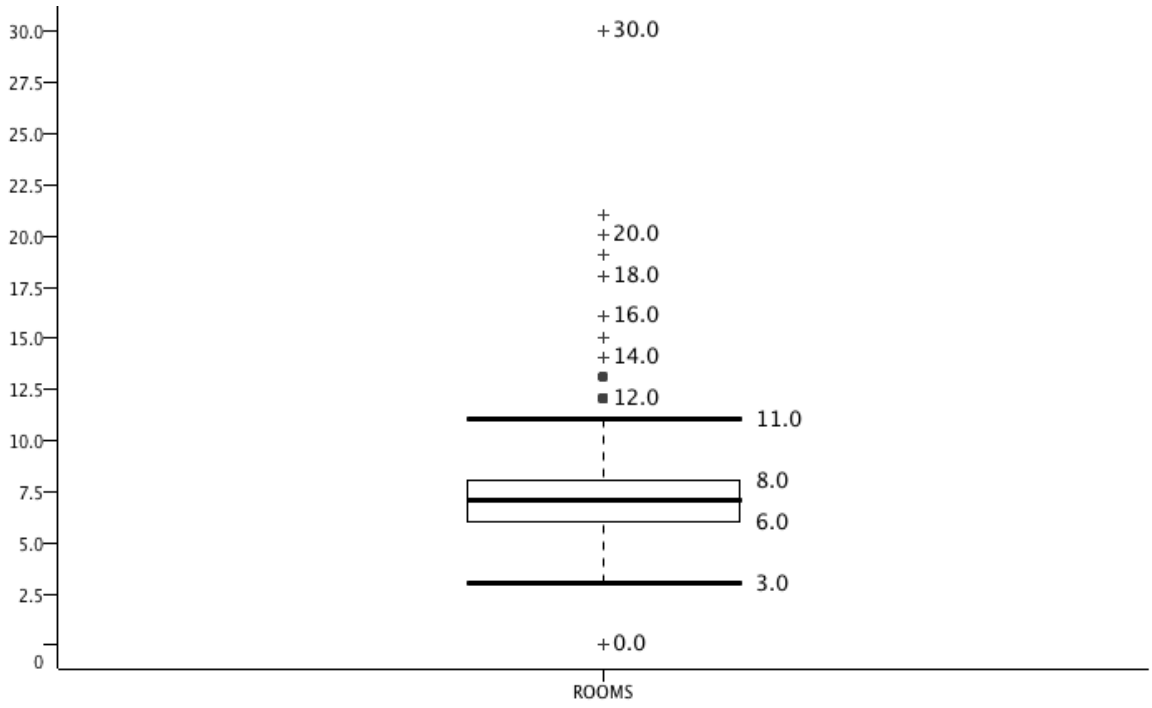

Figure 12 – Box plot of number of rooms per property

8. BEDRM

Attribute Type:

Ratio.

The value of BEDRM signifies how many bedrooms said property consists of. As such, it can be used to order properties. Further, exact difference between number of bedrooms is known.

A property with 1 room has less bedrooms than a property with 5 bedrooms with the difference being 4.

Statistics:

Statistic	Value
Range	0 – 12
Median	3
Mean	3.346
Variance	1.22
Standard Deviation	1.105
1st Quartile	3
3rd Quartile	4

Distribution & Frequency:

As can be seen, the no. of bedrooms per property is heavily centred between 3 (1001) and 4 (454) being the most popular. This is confirmed by the Interquartile range of 3 to 4 observed in the box

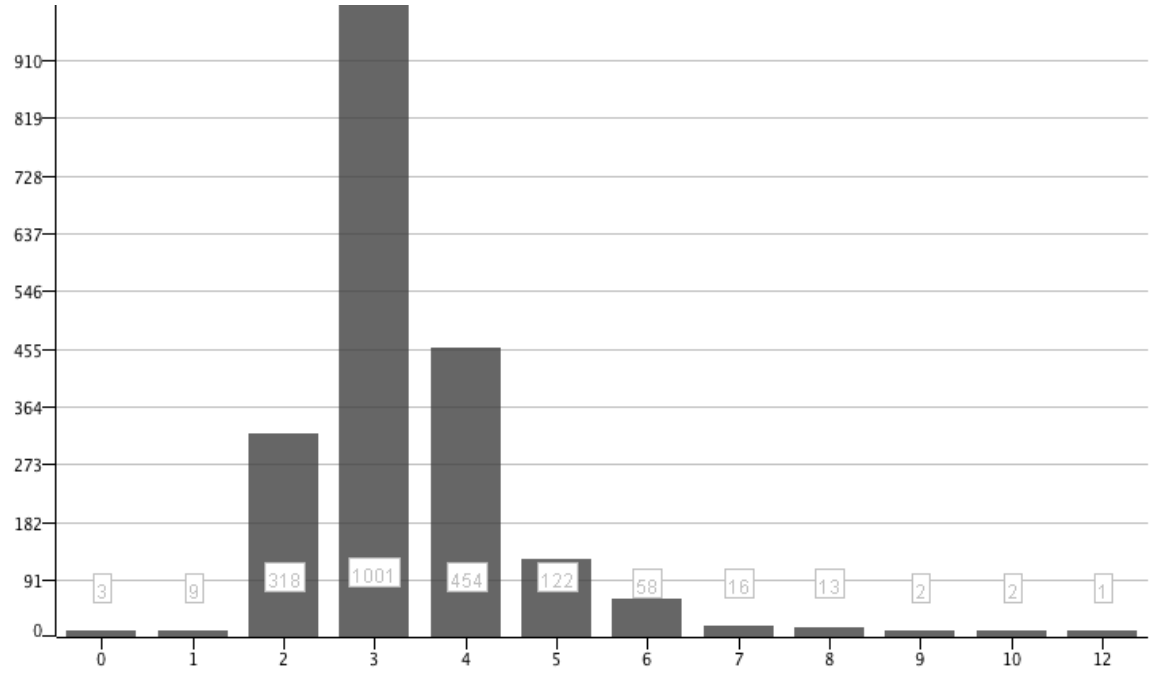

Figure 13 – Histogram distribution of number of bedrooms per property

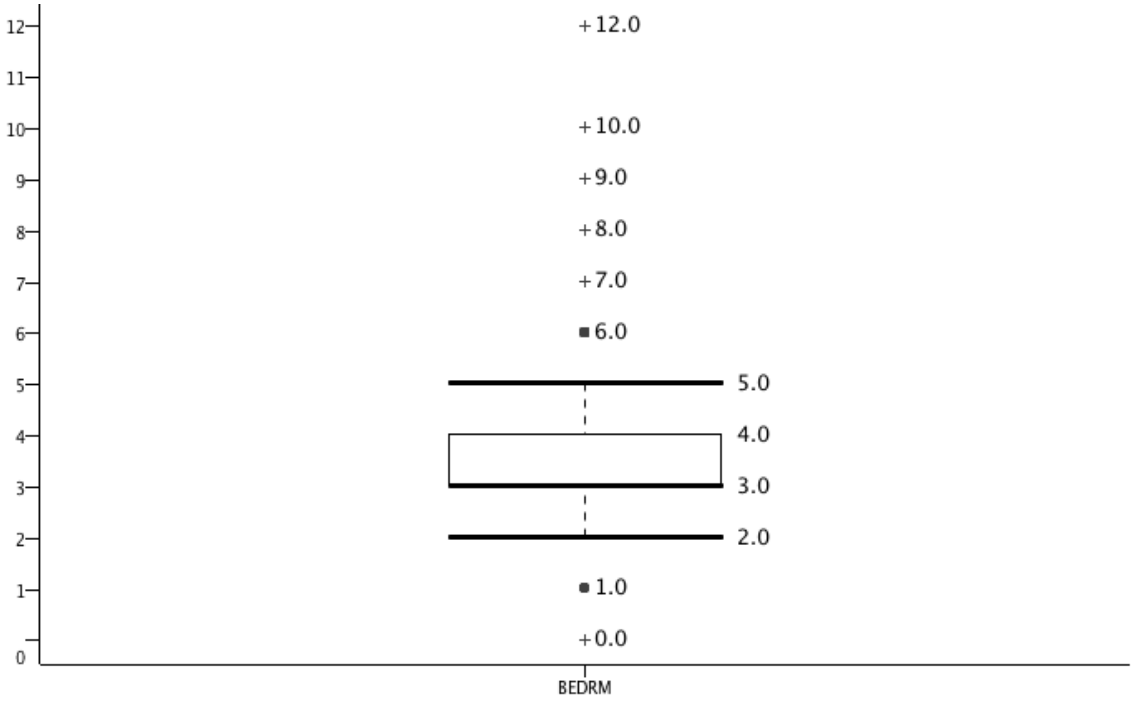

Figure 14 – Box plot of number of bedrooms per property

9. AYB

Attribute Type:

Interval.

The value of AYB signifies when the original building was constructed. As such, it can be used to order properties and the difference between construction years is valuable.

Statistics:

Columns with 0 value are assumed to be errors and excluded from the following statistical analysis.

Statistic	Value
Range	1800 – 2019
Median	1930
Mean	1934
Variance	247.573
Standard Deviation	15.734
1st Quartile	1915
3rd Quartile	1957

Distribution & Frequency:

As can be seen, there were two main peak construction periods, the first one around 1900 and the second around the 1925 mark, confirmed by both the pie chart and histogram. Peak construction period was between 1915 and 1957 (interquartile range) observed in the boxplot.

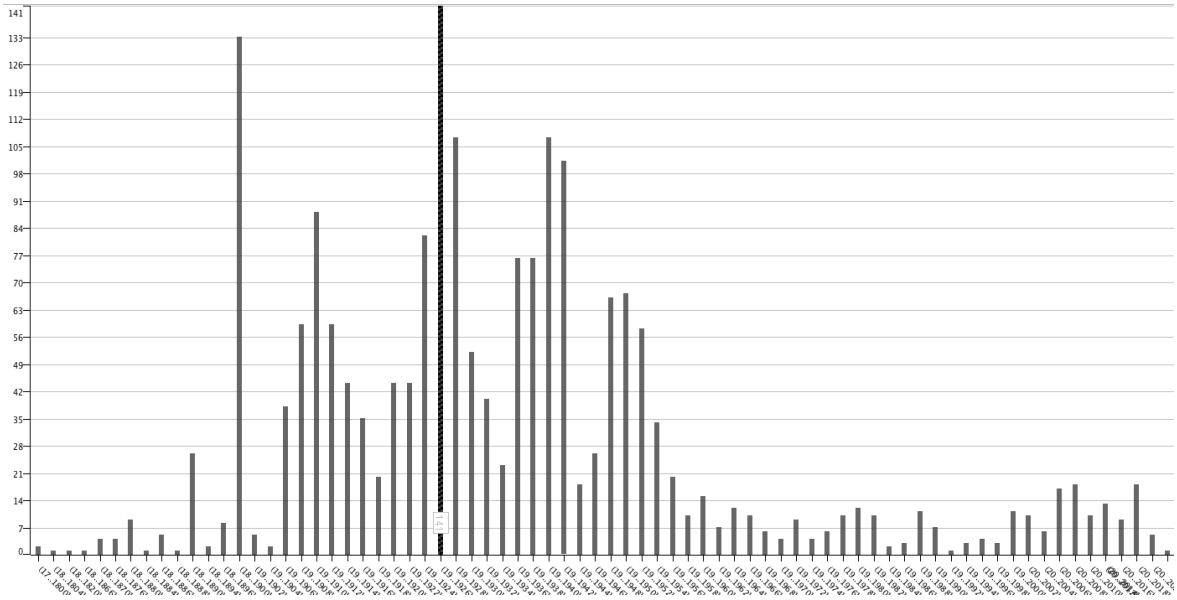

Figure 15 – Histogram distribution AYB

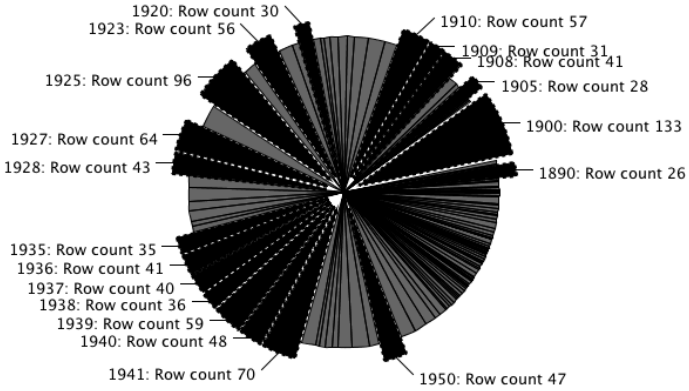

Figure 16 - Pie chart AYB

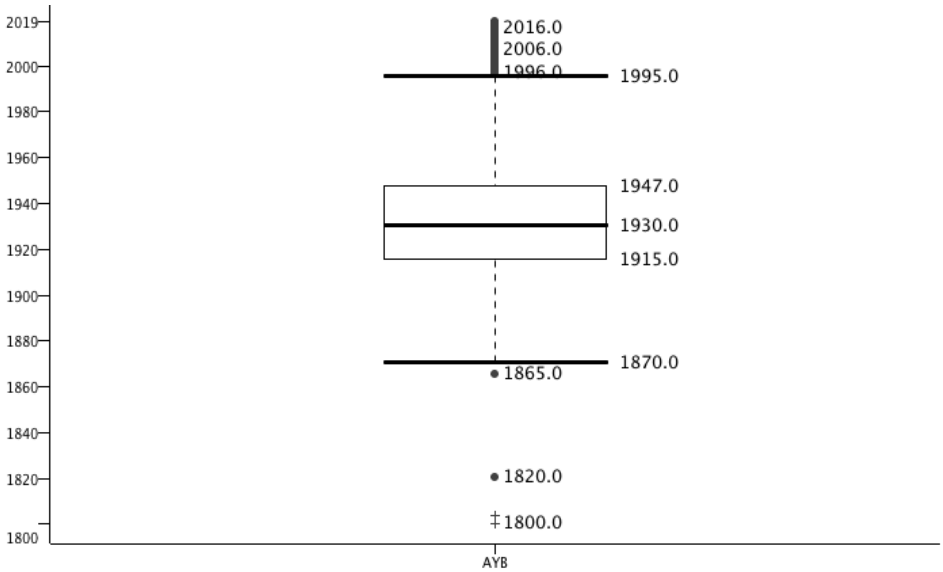

Figure 17 - Box Plot AYB

10. YR_RMDL

Attribute Type:

Interval.

The value of YR_RMDL signifies the last year internal remodelling was done on the original building. As such, it can be used to order properties in terms of when they were remodelled and the difference between remodelling years is a valuable insight.

Statistics:

Columns with 0 value are assumed to be errors and excluded from the following statistical analysis.

Statistic	Value
Range	1932 – 2018
Median	1999.9
Mean	1934
Variance	1.22
Standard Deviation	28.3
1st Quartile	1991
3rd Quartile	2011

Distribution & Frequency:

As can be seen in Figure 16, greater than 50% of the properties did not have data associated with the YR_RMDL. This could be interpreted in two ways. Firstly, missing value could mean that this data was not available for said property, and secondly, it could also be concluded that said property has not had any remodelling done since initial construction. For the purpose of this analysis, we will only analyse the statistics relevant to available data.

As can be seen from the interquartile range displayed on the box plot as well as the concentration of frequency on the histogram, most of the remodelling done on properties was between 1991 and 2011, peak being between 2010 and 2012 where 98 properties were remodelled. All remodelling done before 1961 are statistical outliers.

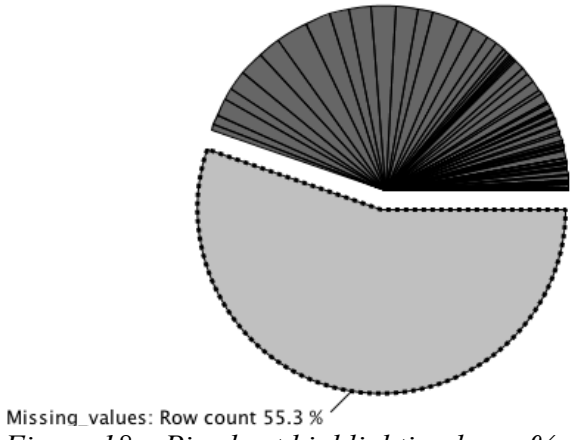

Figure 18 – Pie chart highlighting large % of missing values

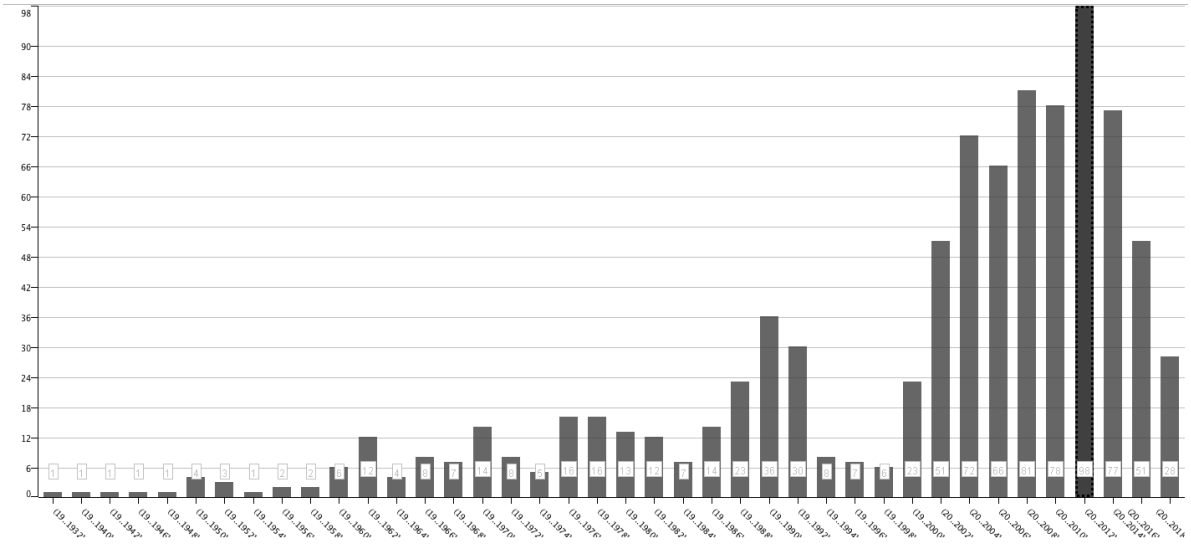

Figure 19 - Histogram distribution YR_RMDL

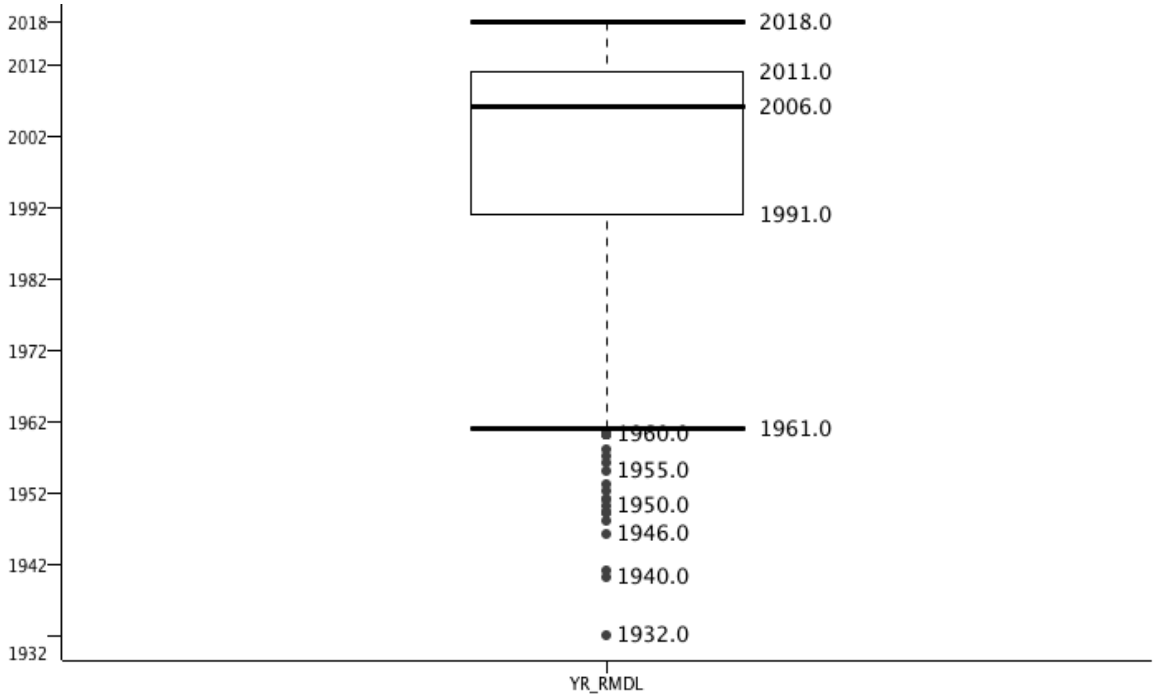

Figure 12 - Box Plot YR_RMDL

11. EYB

Attribute Type:

Interval.

The value of EYB signifies the last year any improvements were built into the original building. As such, it can be used to order properties and the difference between construction years is valuable.

Statistics:

Statistic	Value

Range	1943 – 2018
Median	1963
Mean	1965.982
Variance	268.861
Standard Deviation	16.397
1st Quartile	1955
3rd Quartile	1970

Distribution & Frequency:

It can be concluded by analysing the histogram that 1957 was the year with the highest number of properties (248) that had improvements built. It can also be noted that the majority of built improvements were made between 1943 and 1992 with the highest concentration being between 1955 and 1970 (Interquartile range).

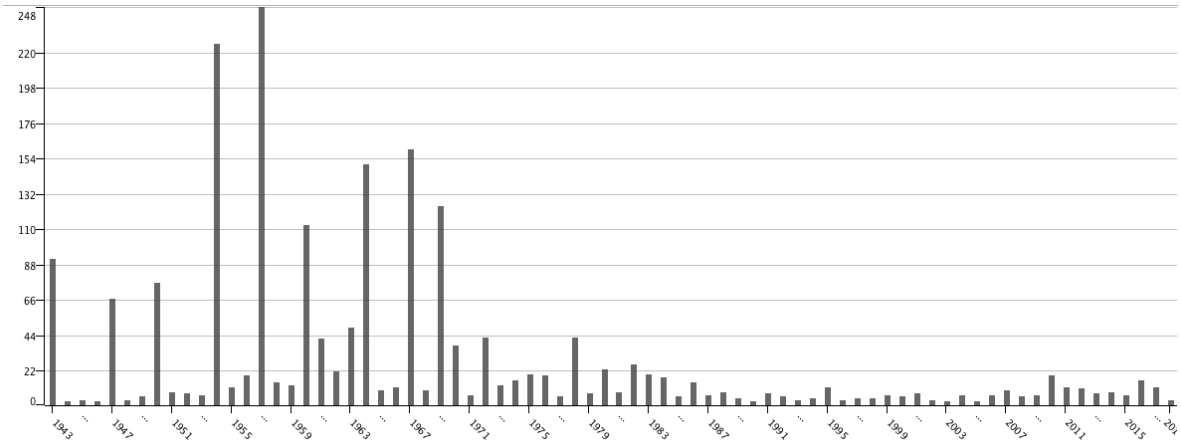

Figure 21 - Histogram distribution EYB

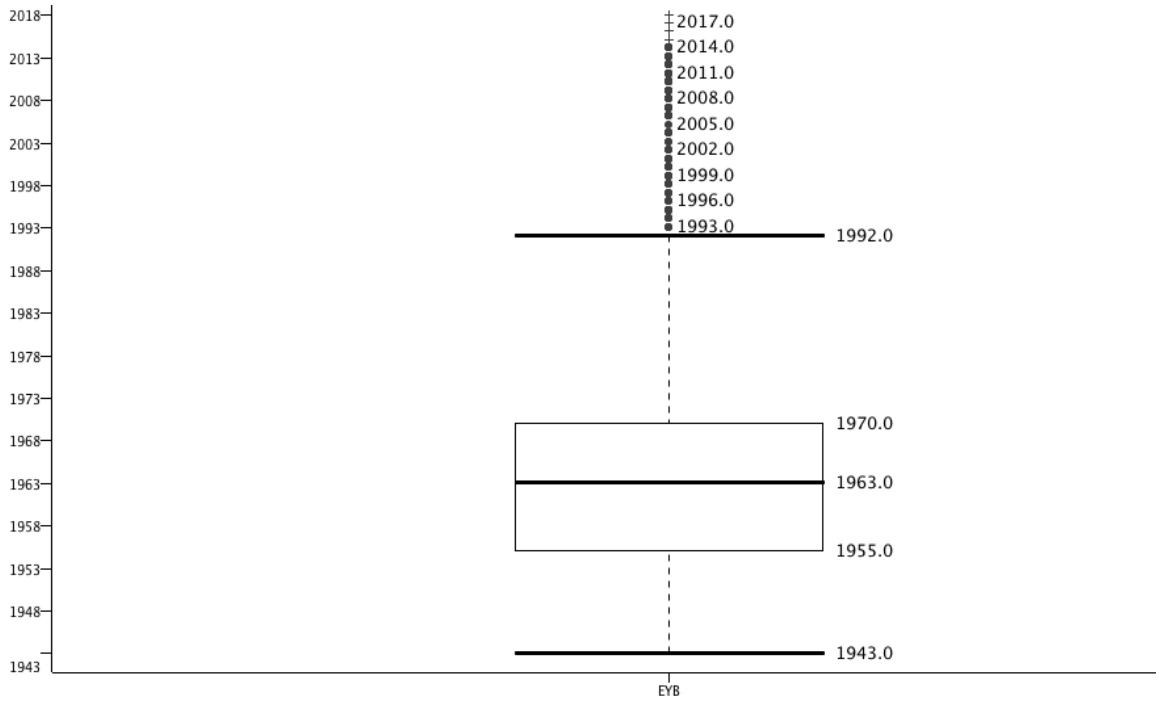

Figure 22 - Box Plot EYB

12. STORIES

Attribute Type:

Ratio.

The value of STORIES signifies the number of floors each property has. As such, it can be used to order properties and the difference as well as ratio between stories of two properties is valuable.

Statistics:

Statistic	Value
Range	1- 9
Median	2
Mean	2.071
Variance	0.184
Standard Deviation	0.428
1st Quartile	2
3rd Quartile	2

Distribution & Frequency:

It is unequivocally evident that the majority of properties have 2 stories. This is confirmed by examining the box plot and histogram figures below. To be exact, 1513 (75.69%) out of the 2000 properties have 2 floors, followed by 173(8.65%) properties that have 3 floors. There are obvious outliers with 1 property having 2.75, 3.75 and 2.75 floors respectively. Histogram figure is included below but not a great visual representation of the data. Pie chart and box plot are much more appropriate for this scenario.

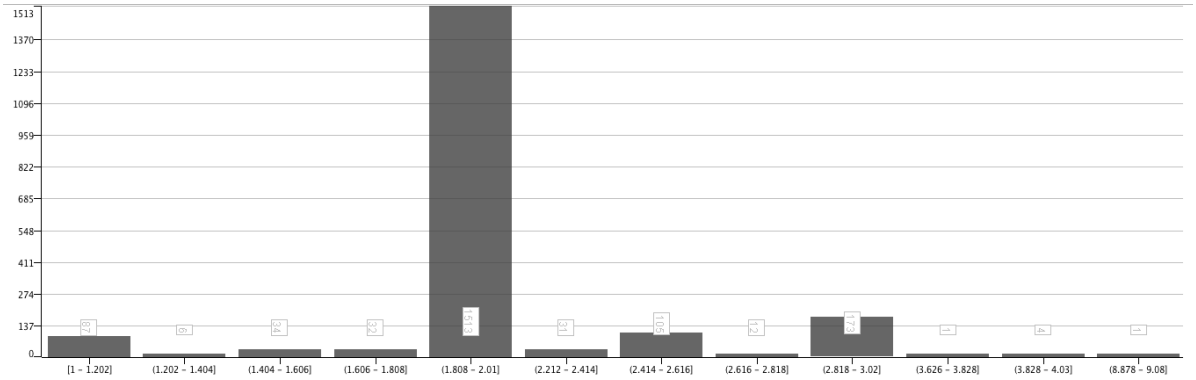

Figure 23 - Histogram distribution STORIES

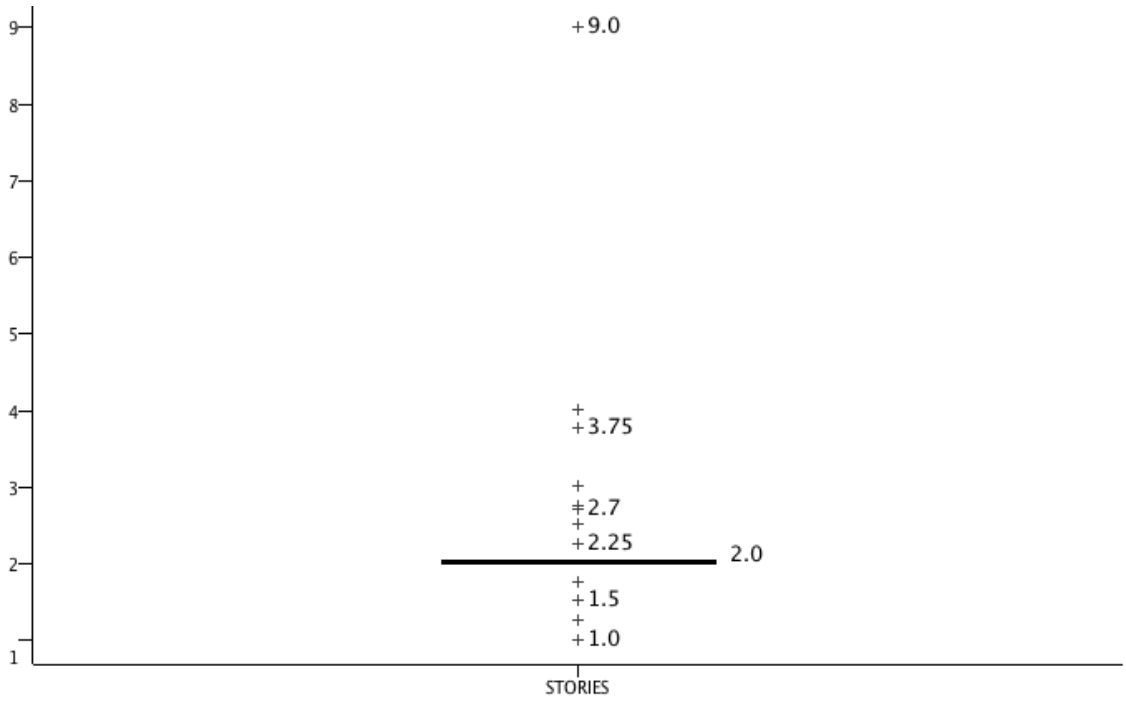

Figure 24 - Box Plot STORIES

Figure 25 - Pie Chart STORIES

13. SALEDATE

Attribute Type:

Interval.

The value of SALEDATE signifies the date a certain property was last sold. As such, it can be used to order properties by sale date and the difference in sale dates between properties is valuable.

Statistics:

To perform statistical analysis on sale dates, each date was broken down into its corresponding year and month. They were also grouped based on quarter to provide further insight.

On further analysis, it is evident that 426 properties have "SALEDATE" equal to 1900-01-01T00:00. It is highly unlikely that such a large percentage of properties were all sold at the same exact times which leads us to conclude that this was a placeholder date of sorts for properties' whose last sale date is not available. This does not mean that some of the mentioned properties were not sold on that date but there is no way to distinguish them. As such, these properties have been omitted from the following analysis.

BY YEAR:

Statistic	Value
Range	1965 - 2018
Median	2010
Mean	2008.685
Variance	50.731
Standard Deviation	7.123
1st Quartile	2004
3rd Quartile	2015

Distribution & Frequency:

It can be concluded by analysing the histogram that 2017 was the year with the highest number of properties (153) sold. It can also be noted that the majority of properties last sold were between 1990 and 2018 with the highest concentration being between 2004 and 2015 (Interquartile range). All properties sold before 1990 are outliers.

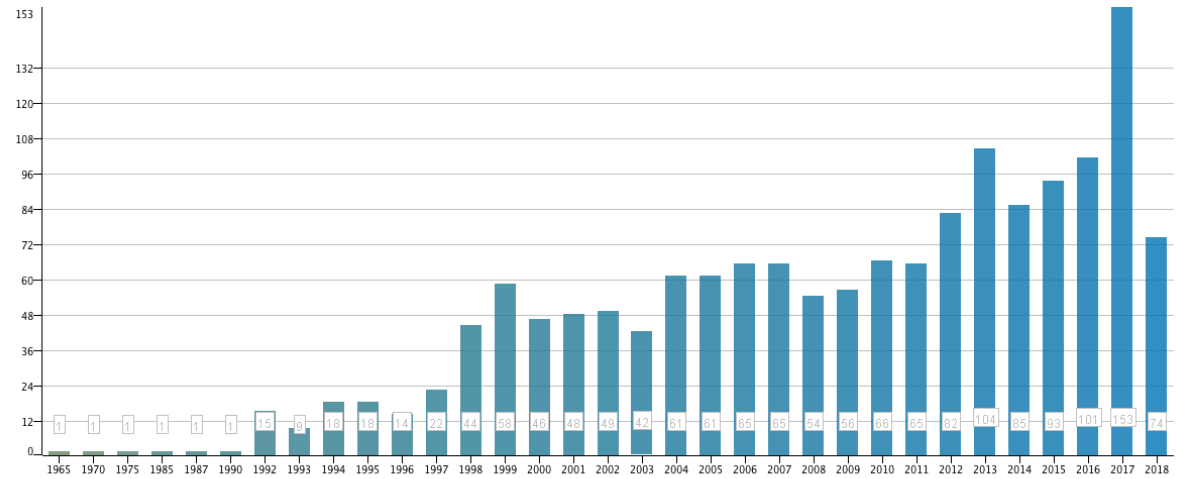

Figure 26 - Histogram distribution SALEDATE/YEAR

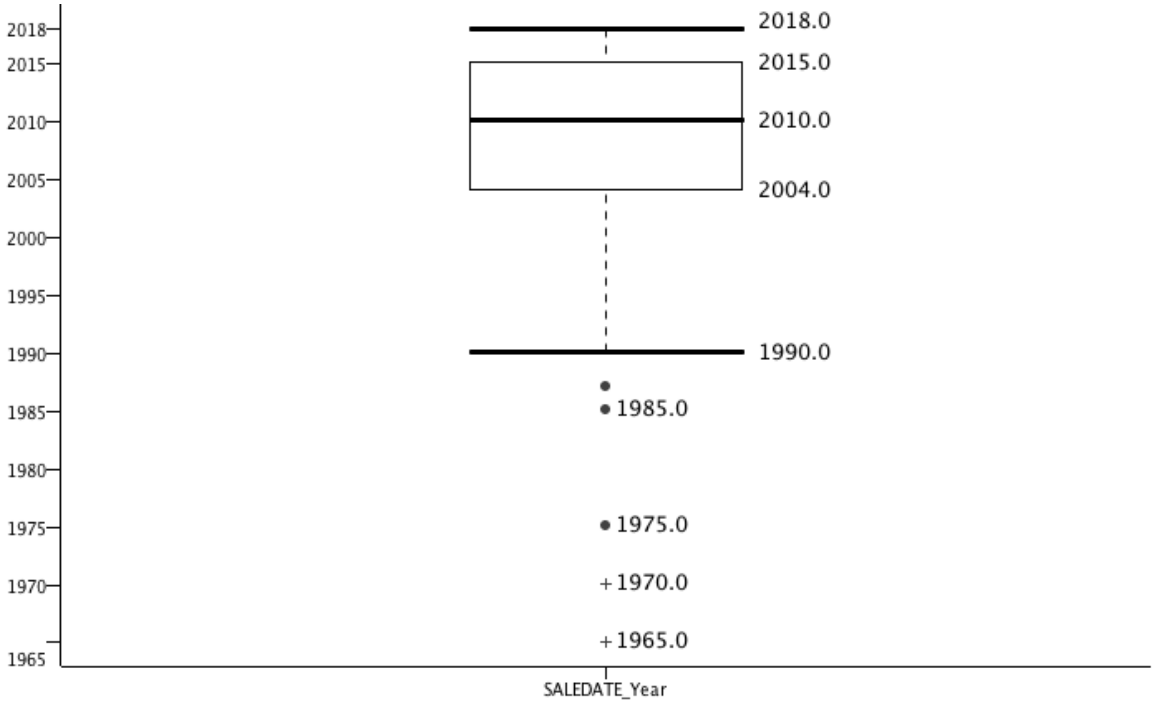

Figure 27 - Box Plot SALEDATE/YEAR

BY MONTH:

Statistic	Value
Range	1-12
Median	6
Mean	6.562
Variance	10.904
Standard Deviation	3.302
1st Quartile	4
3rd Quartile	9

Distribution & Frequency:

It can be concluded by analysing the histogram that June (Month #6) is the month with the highest number of properties (165) sold. Distribution is fairly even throughout the year with no statistical outliers.

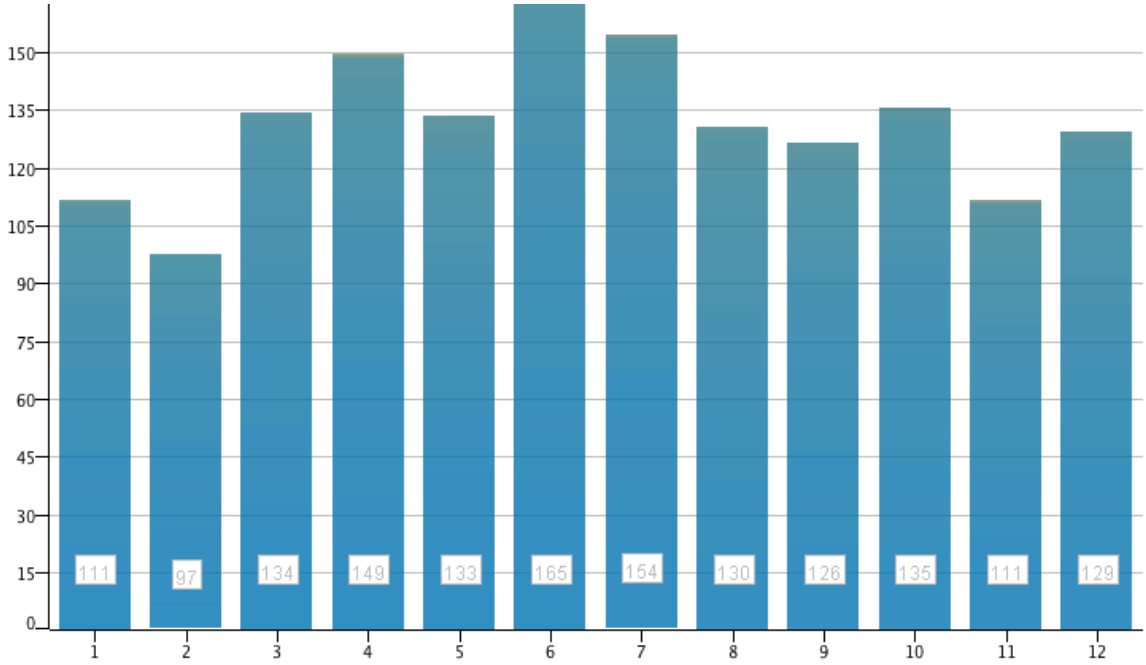

Figure 28 - Histogram distribution SALEDATE/MONTH

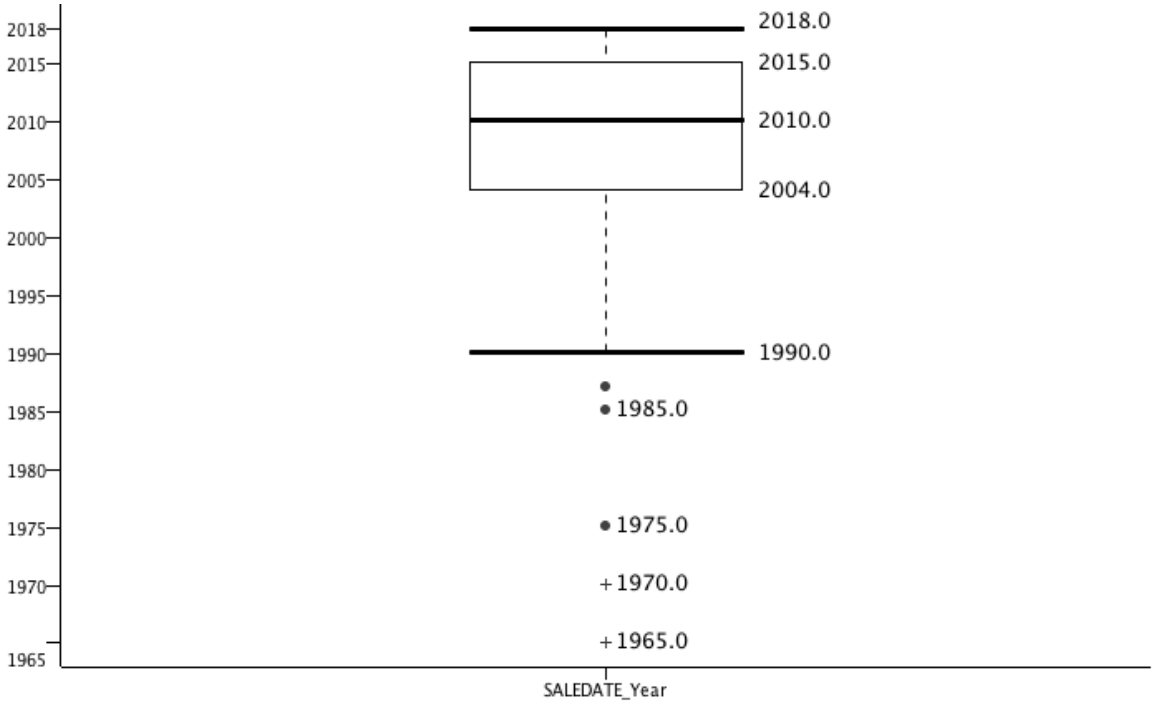

Figure 29 - Box Plot SALEDATE/MONTH

BY QUARTER:

Statistic	Value
Range	1- 4
Median	2
Mean	6.562
Variance	10.904
Standard Deviation	3.302

1st Quartile	2
3rd Quartile	3

Distribution & Frequency:

It can be concluded by analysing the histogram that the 2nd quarter is the month with the highest number of properties (447) sold. Distribution is fairly even throughout the year with no statistical outliers.

.

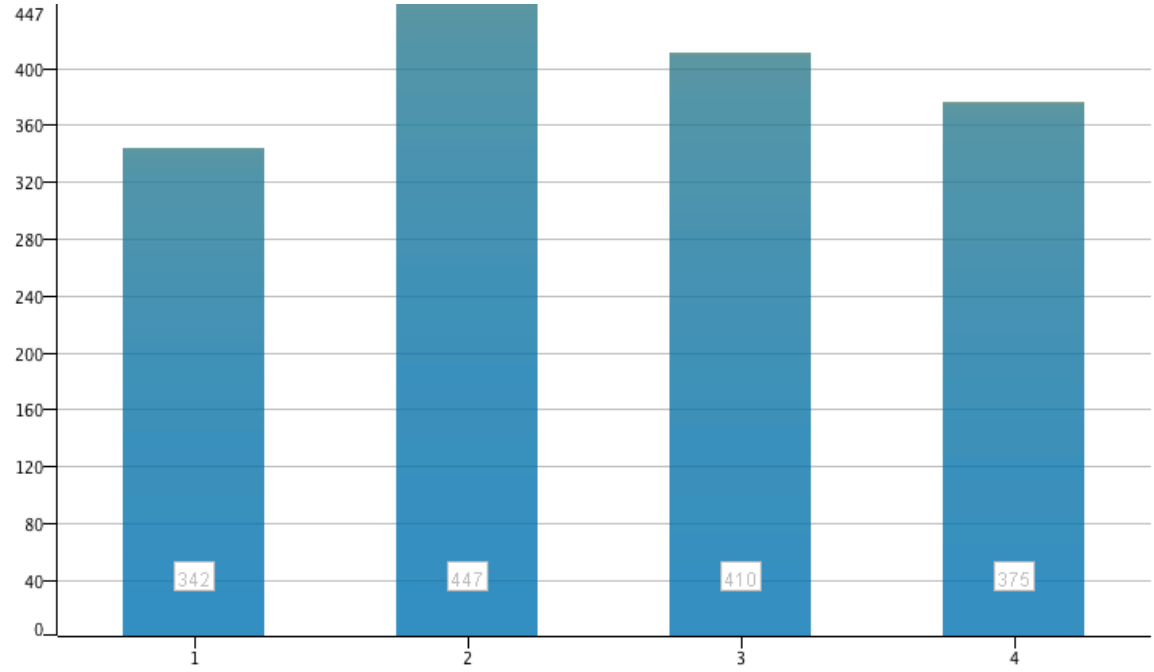

Figure 30 - Histogram distribution SALEDATE/QUARTER

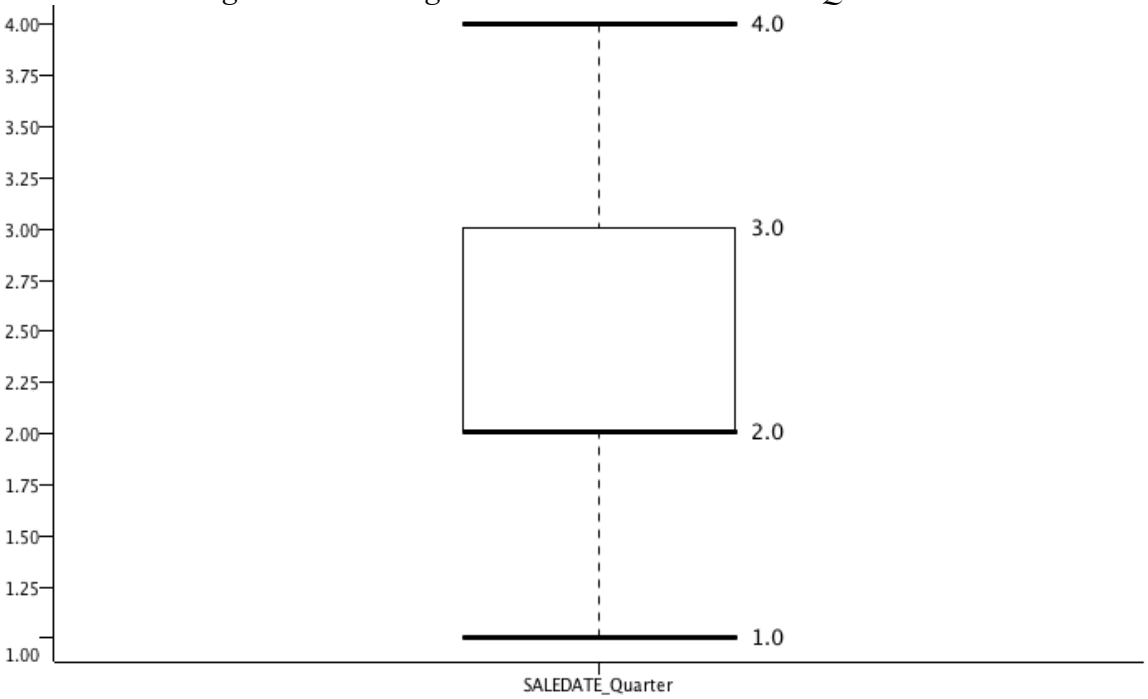

Figure 31 - Box Plot SALEDATE/QUARTER

14. PRICE

Attribute Type:

Ratio.

The value of PRICE signifies how much the property was sold for last in terms of dollar value (assumed). As such, it can be used to order properties. Further, exact difference between the price for two properties is known.

Statistics:

Properties with price = 0 but with SALEDATE information available have been assumed to be so because of this stat not available for several reasons. Keeping this assumption in mind, such properties have been excluded from the following analysis.

Statistic	Value
Range	9,125 – 5,500,000
Median	450,000
Mean	582,532.27
Variance	311,416,675,223.212
Standard Deviation	558,047.198
1st Quartile	241,750
3rd Quartile	746,250

Distribution & Frequency:

As can be seen from the boxplot, majority of the property prices are between 9,125 and 1,500,000 with the largest concentration being between 241,750 and 746,250. This is reflected in the histogram shown below. Property prices beyond are clearly outliers.

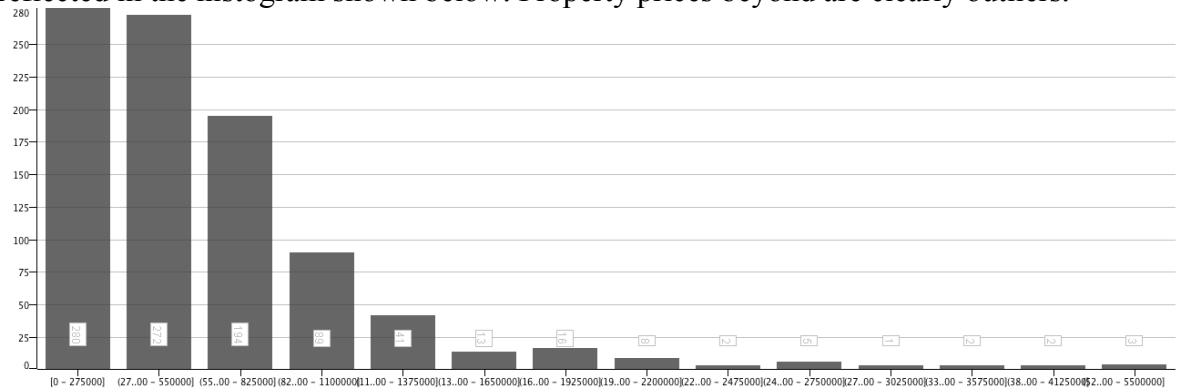

Figure 32 – Histogram distribution: Price when last sold per property

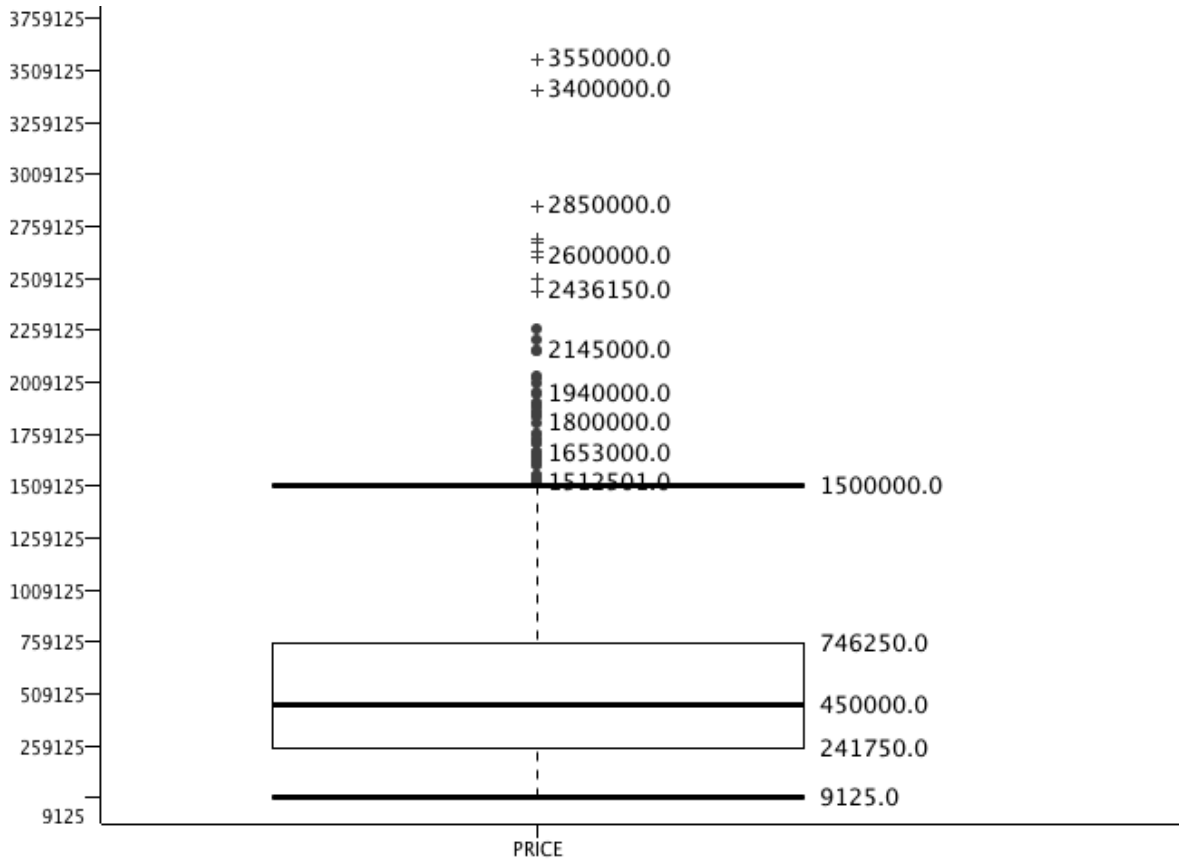

Figure 33 – Box plot: Price when last sold per property

15. HEAT

Attribute Type:

Nominal.

The value of Heat cannot be ordered. Although the attribute is represented by numerical values, they are mapped to non-quantitative equivalent values found in the HEAT_D column. Although one heating method could be classified as being more efficient than another, this information is not provided and as a conclusion, only acts as a label.

Mapping:

- No Data - 0
- Forced Air -1
- Air-Oil - 2
- Wall Furnace - 3
- Electric Rad - 4
- Elec Base Brd - 5
- Water Base Brd - 6
- Warm Cool - 7
- Ht Pump – 8
- Air Exchng - 10
- Gravity Furnac - 11
- Hot Water Rad- 13

Distribution & Frequency:

Mapping of non-quantitative attribute has already been provided in the HEAT_D column. It can be very easily concluded from the below histogram and pie graph that the most common heating solution used is “Hot Water Rad” (42.27%) and the least common being Forced Air, Elec Base Brd, Air Exchg & Gravity Furnac (0.05%).

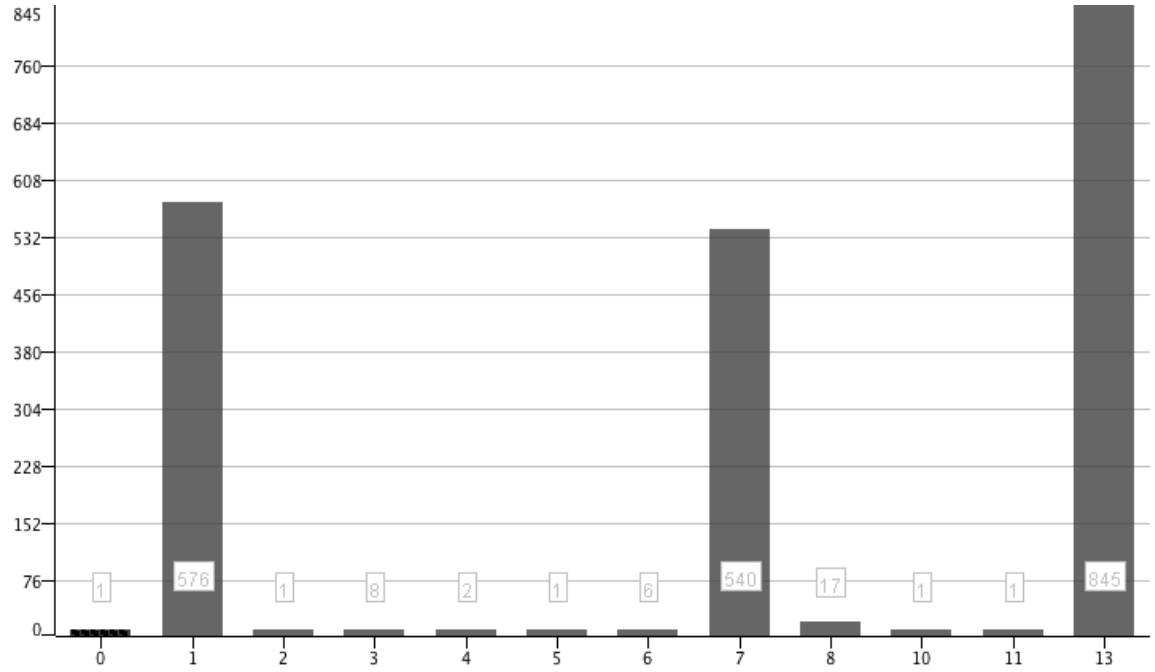

Figure 34 – Histogram distribution of type of heating used

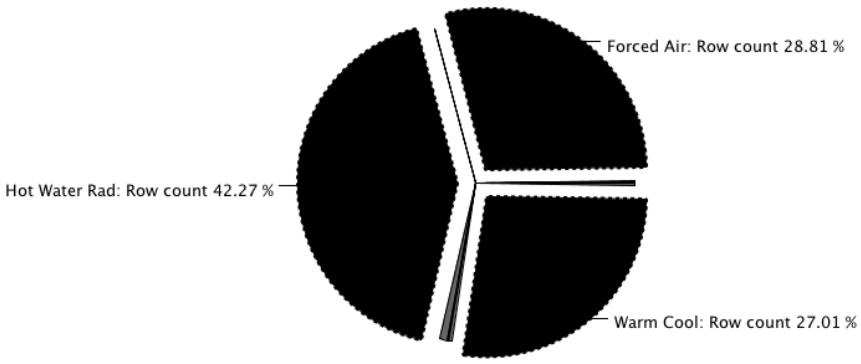

Figure 35 – Pie chart showing top three categories type of heating used by %

16. STYLE

Attribute Type:

Nominal.

The value of STYLE cannot be ordered. Although the attribute is represented by numerical values, they are mapped to non-quantitative equivalent values found in the STYLE_D column. Only acts as a label. Describes the style of said building

Distribution & Frequency:

Mapping of non-quantitative attribute has already been provided in the STYLE_D column. It can be very easily concluded from the below histogram and pie graph that the most common STYLE is "2 Story" with 1554 properties (77.77%) followed by "3 Story" with 174 properties (8.7%).

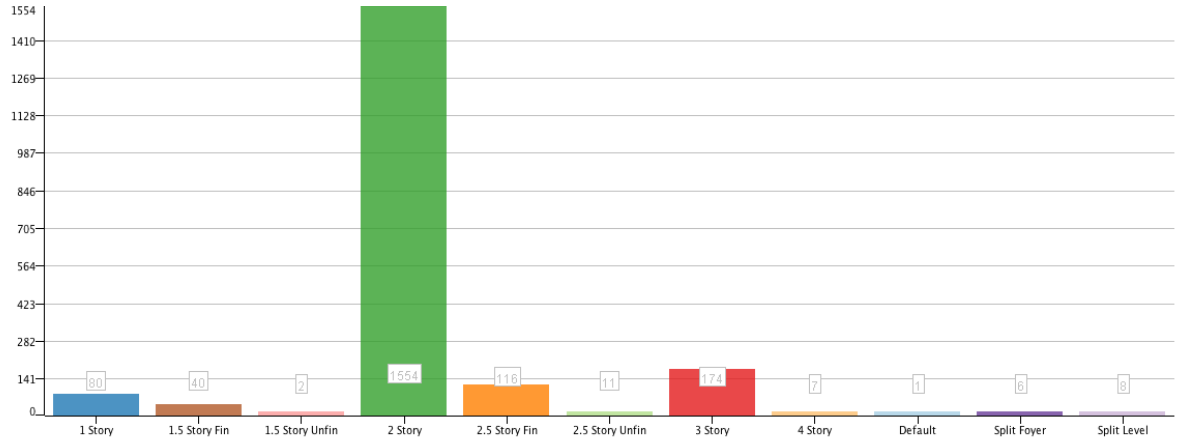

Figure 36 – Histogram distribution of STYLE categories

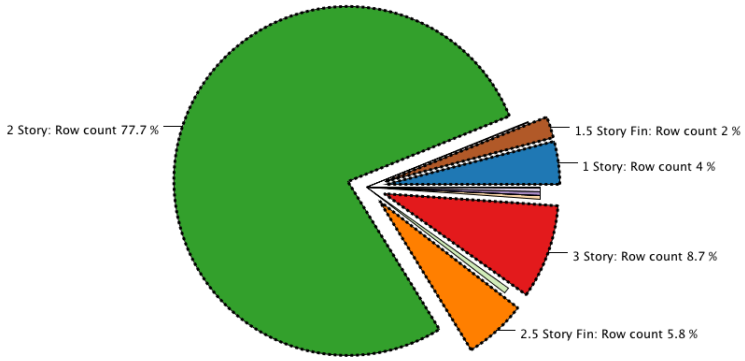

Figure 37 – Pie chart showing top STYLE categories by %

17. STRUCT

Attribute Type:

Nominal.

The value of STRUCT cannot be ordered. Although the attribute is represented by numerical values, they are mapped to non-quantitative equivalent values found in the STRUCT_D column. Only acts as a label. Describes the type of structure of said building.

Distribution & Frequency:

Mapping of non-quantitative attribute has already been provided in the STRUCT_D column. It can be very easily concluded from the below histogram and pie graph that the most common structure is "Row Inside" with 747 properties (37.37%) followed by "Single" with 598 properties (29.91%). The least common structure is "Town End" with only 1(0.05%) property with the structure.

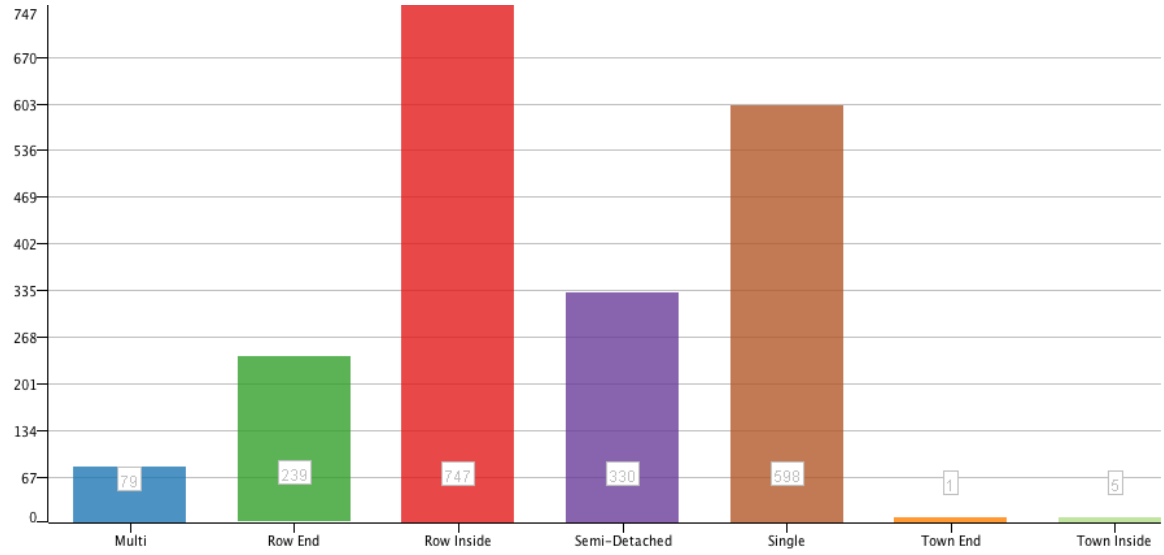

Figure 38 – Histogram distribution of STRUCT categories

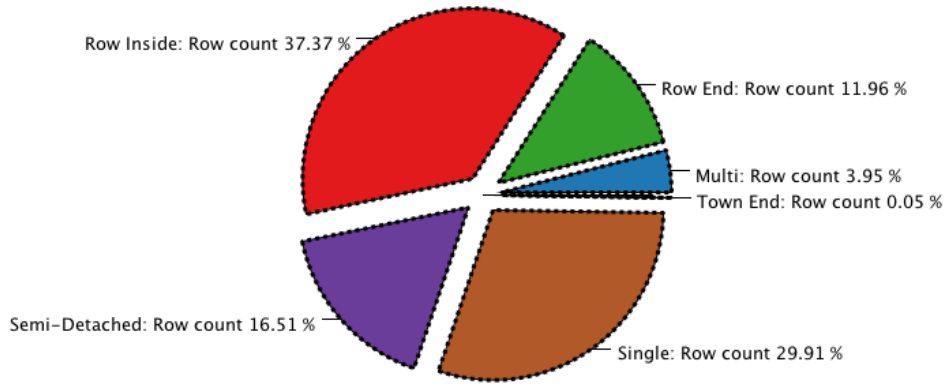

Figure 39 – Pie chart showing top STRUCT categories by %

18. GRADE

Attribute Type:

Nominal.

The value of GRADE cannot be ordered individually for properties. Although the attribute is represented by numerical values, they are mapped to non-quantitative equivalent values found in the GRADE_D column. Only acts as a label. Describes the quality of property.

Distribution & Frequency:

Mapping of non-quantitative attribute has already been provided in the GRADE_D column. It can be very easily concluded from the below histogram and pie graph that the most common GRADE is "Average" with 710 properties (35.52%) followed by "Good Quality" with 365 properties (18.26%). The least common structure is "Exceptional" with only 1(0.05%) property with the structure, being a clear outlier.

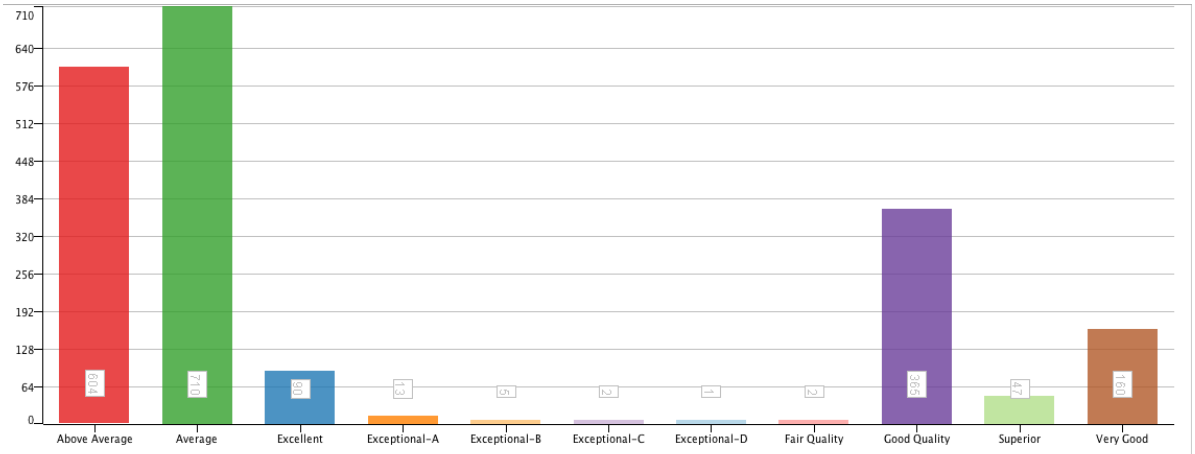

Figure 40 – Histogram distribution of GRADE categories

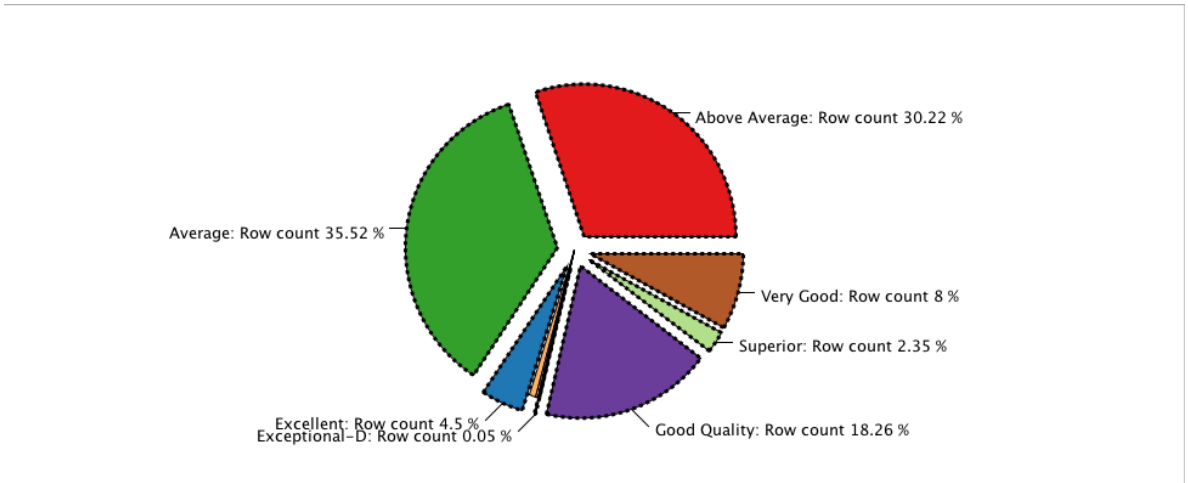

Figure 41 – Pie chart showing top GRADE categories by %

19. CNDTN

Attribute Type:

Nominal.

The value of CNDTN cannot be ordered individually for properties. Although the attribute is represented by numerical values, they are mapped to non-quantitative equivalent values found in the CNDTN_D column. Only acts as a label. Describes the condition of property.

Distribution & Frequency:

Mapping of non-quantitative attribute has already been provided in the CNDTN column. It can be very easily concluded from the below histogram and pie graph that the most common property condition is “Average” with 1162 properties (58.13%) followed by “Good” with 647 properties (32.37%). The least common structure is “Poor” with only 4 (0.2%) property with the condition, being a clear outlier.

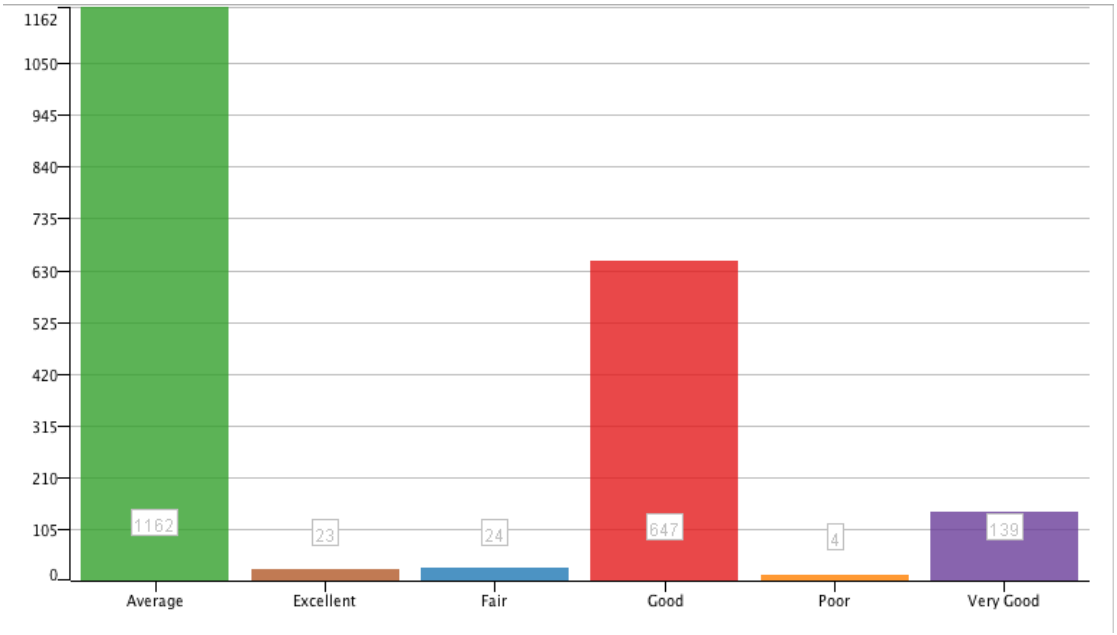

Figure 41 – Histogram distribution of CNDTN categories

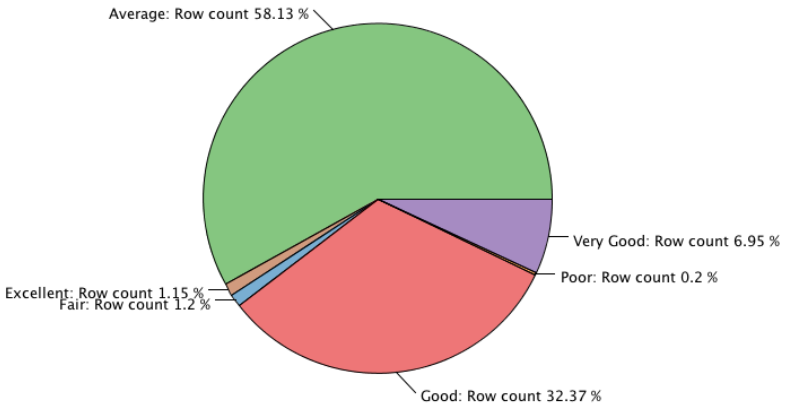

Figure 42 – Pie chart showing top CNDTN categories by %

20. EXTWALL

Attribute Type:

Nominal.

The value of EXTWALL cannot be ordered individually for properties. Although the attribute is represented by numerical values, they are mapped to non-quantitative equivalent values found in the EXTWALL column. Only acts as a label. Describes the type of material used in building the external wall of property.

Distribution & Frequency:

Mapping of non-quantitative attribute has already been provided in the EXTWALL column. It can be very easily concluded from the below histogram and pie graph that the most common material used for the external wall, by far, is "Common Brick" with 1500 properties (75.08%). The least common material used is "Stucco" and "Concrete" with only 1 (0.05%), being clear outliers.

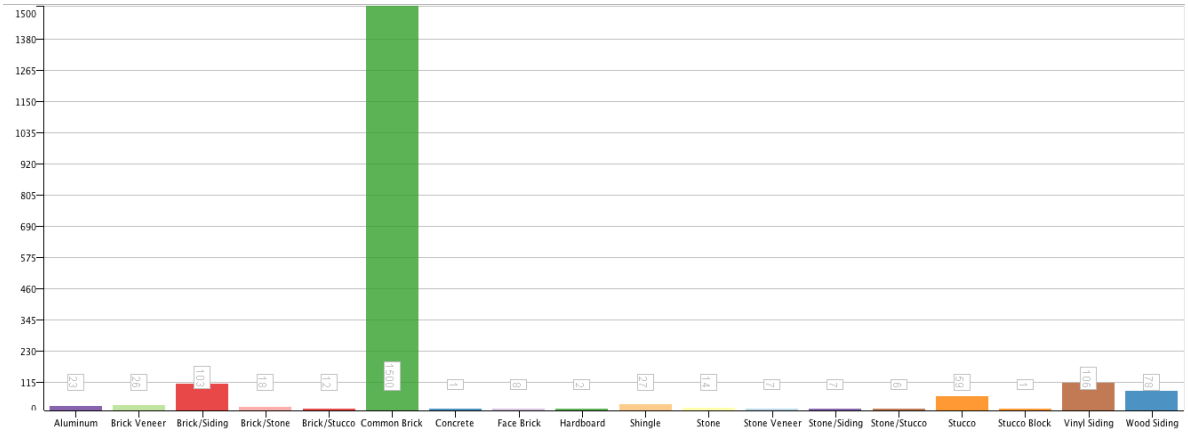

Figure 43 – Histogram distribution of EXTWALL *categories*

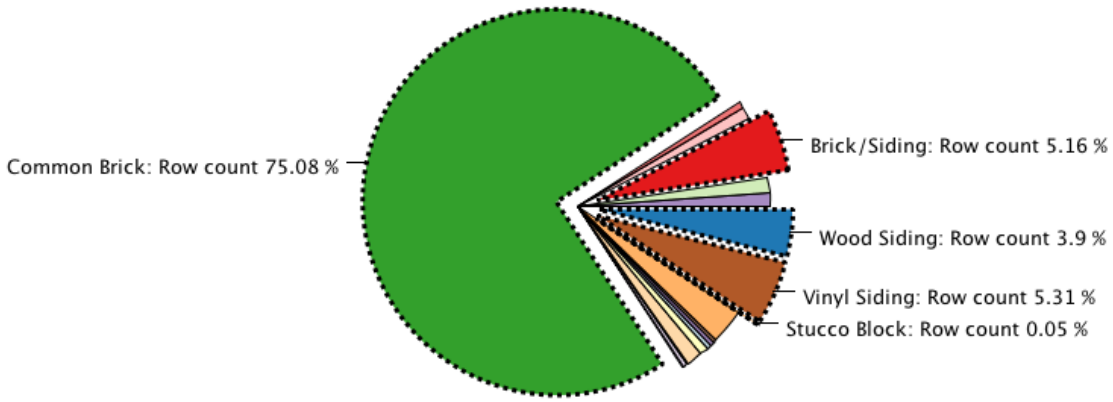

Figure 44 – Pie chart showing top EXTWALL *categories by %*

21. ROOF

Attribute Type:

Nominal.

The value of ROOF cannot be ordered individually for properties. Although the attribute is represented by numerical values, they are mapped to non-quantitative equivalent values found in the ROOF column. Only acts as a label. Describes the type of material used in building the roof of property.

Distribution & Frequency:

Mapping of non-quantitative attribute has already been provided in the ROOF column. It can be very easily concluded from the below histogram and pie graph that the most common materials used for the roof are "Built Up" (578), "Comp Shingle" (575) and "Metal- Sms" (560). Further breakdown is evident in histogram distribution. "Concrete Tile", "Concrete" and "Typical" are clear outliers with 1 property each using such roofing.

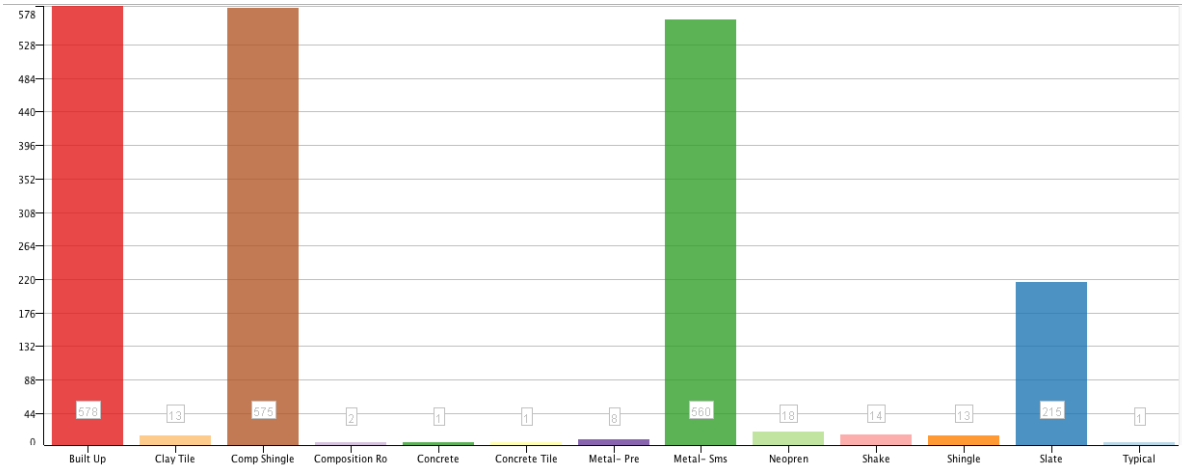

Figure 44 – Histogram distribution of ROOF *categories*

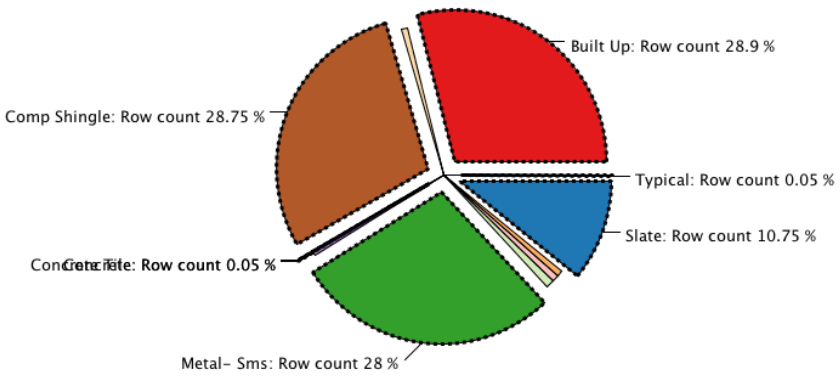

Figure 45 – Pie chart showing top ROOF *categories by %*

22. INTWALL

Attribute Type:

Nominal.

The value of INTWALL cannot be ordered individually for properties. Although the attribute is represented by numerical values, they are mapped to non-quantitative equivalent values found in the INTWALL column. Only acts as a label. Describes the type of material used in building the interior wall of property.

Distribution & Frequency:

Mapping of non-quantitative attribute has already been provided in the INTWALL column. It can be noted via below histogram and pie graph that the most common material used for the internal wall, by far, is "Hardwood" with 1559 properties (77.99%). The least common material used is "Vinyl Comp" and "Ceramic Tile" with only 1 (0.05%), being clear outliers.

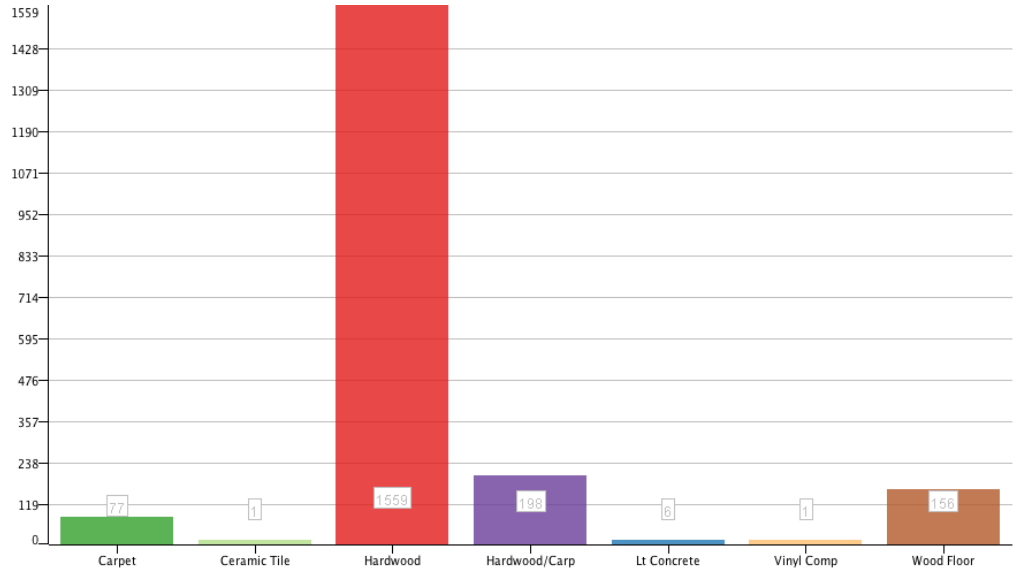

Figure 46 – Histogram distribution of INTWALL *categories*

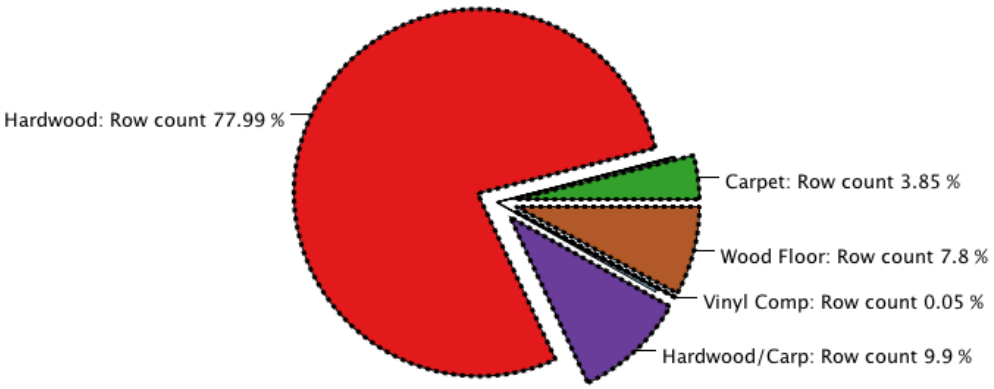

Figure 47 – Pie chart showing top INTWALL *categories by %*

23. KITCHENS

Attribute Type:

Ratio.

The value of KITCHENS signifies how many kitchens said property consists of. As such, it can be used to order properties. Further, exact difference between number of bedrooms is known. E.g. A property with 1 kitchen has less than a property with 5 with the difference being 4.

Statistics:

Statistic	Value
Range	0 – 6
Median	1
Mean	1.205
Variance	0.358
Standard Deviation	0.598
1st Quartile	1
3rd Quartile	1

Distribution & Frequency:

As can be seen, 1 is the most popular no. of kitchens per property with 1720 (86.4%) properties. There are clear outliers with 2 properties having 0 and 1 property having 6 kitchens respectively.

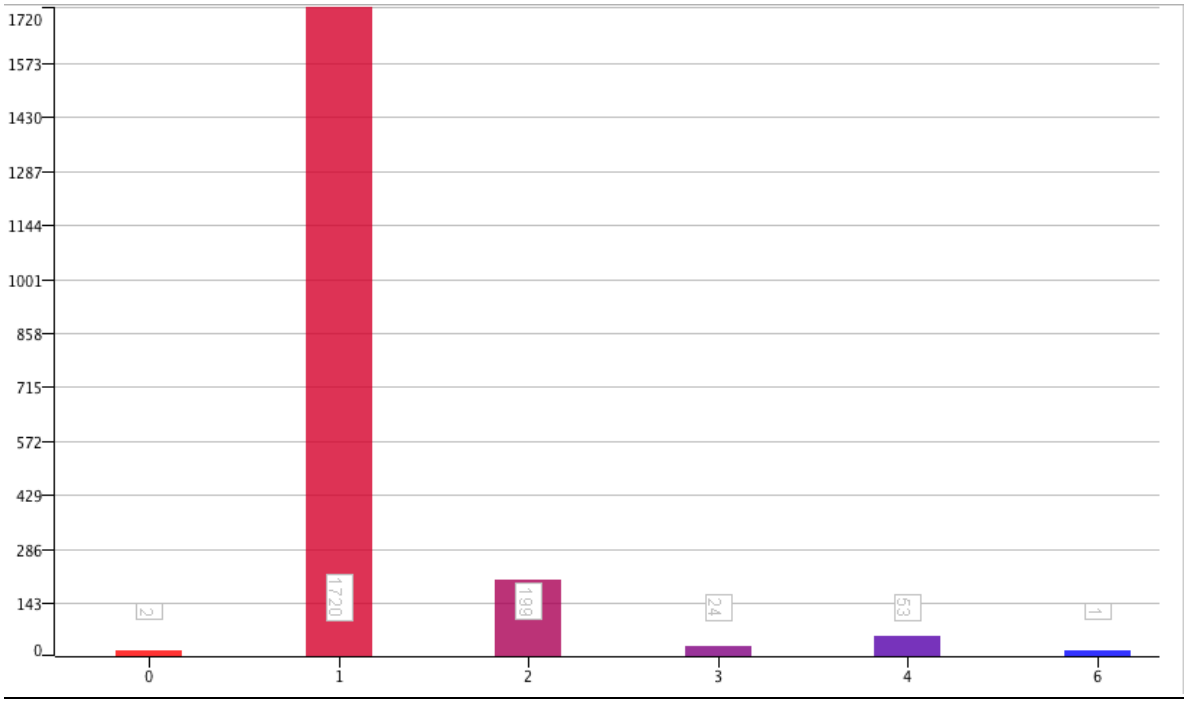

Figure 48 – Histogram distribution of number of kitchens per property

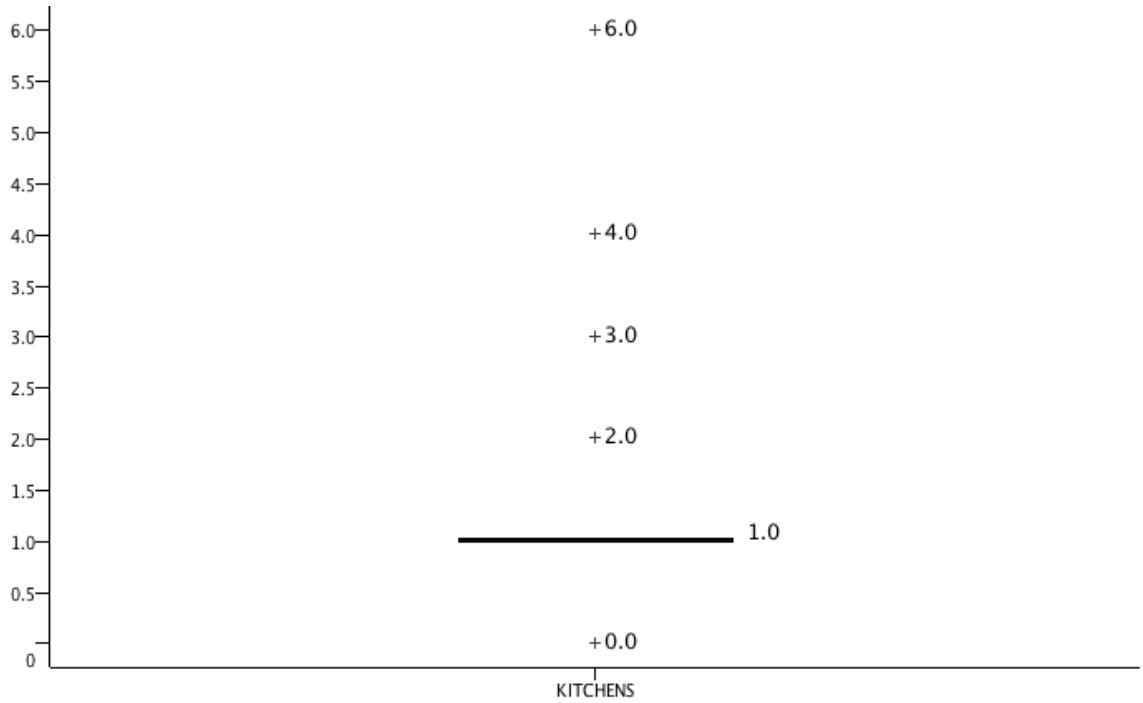

Figure 49 – Box plot of number of kitchens per property

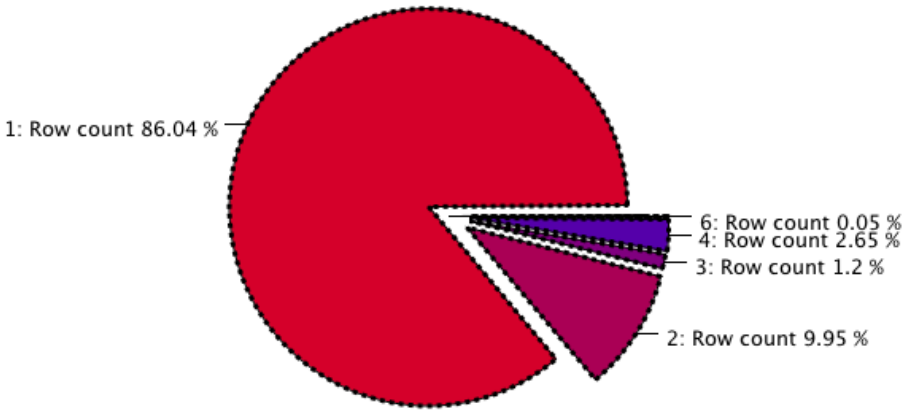

Figure 50 – Pie chart distribution of number of kitchens per property

24. QUALIFIED

Attribute Type:

Nominal.

The value of QUALIFIED cannot be ordered individually for properties. Only acts as a label. Enough information is not provided to conclude what this label signifies. Perhaps, it has to do with building regulations or certain standards.

Distribution & Frequency:

It can be noted via below histogram and pie graph that more properties are unqualified (1239) than properties that are qualified (761).

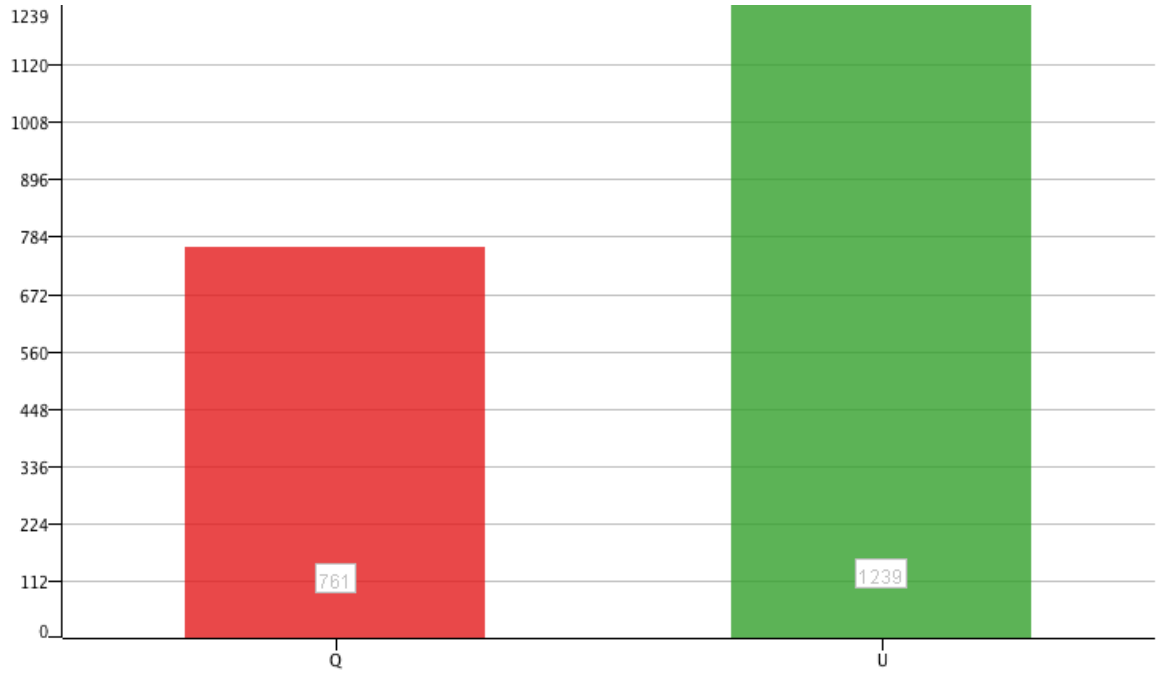

Figure 48 – Histogram distribution of QUALIFIED *categories*

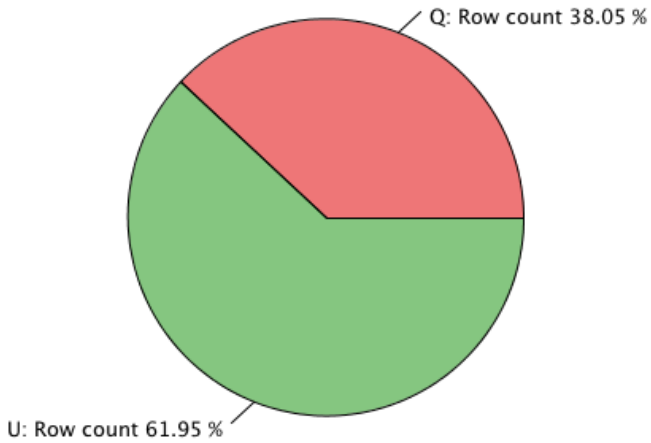

Figure 49 – Pie chart showing top QUALIFIED *categories by %*

25. SALE_NUM

Attribute Type:

Ratio.

The value of SALE_NUM signifies how many times the property has been sold in the past. As such, it can be used to order properties. Further, exact difference between number of times property is sold is known.

Statistics:

Statistic	Value
Range	1 - 10
Median	1
Mean	1.578
Variance	1.543
Standard Deviation	1.242
1st Quartile	1
3rd Quartile	1

Distribution & Frequency:

As can be seen, the most common no. of times a property has been sold is 1515 (75.75%). It is to be noted that 1 property has been sold 10 times and is a clear outlier. Further analysis to pin point reasons would be interesting.

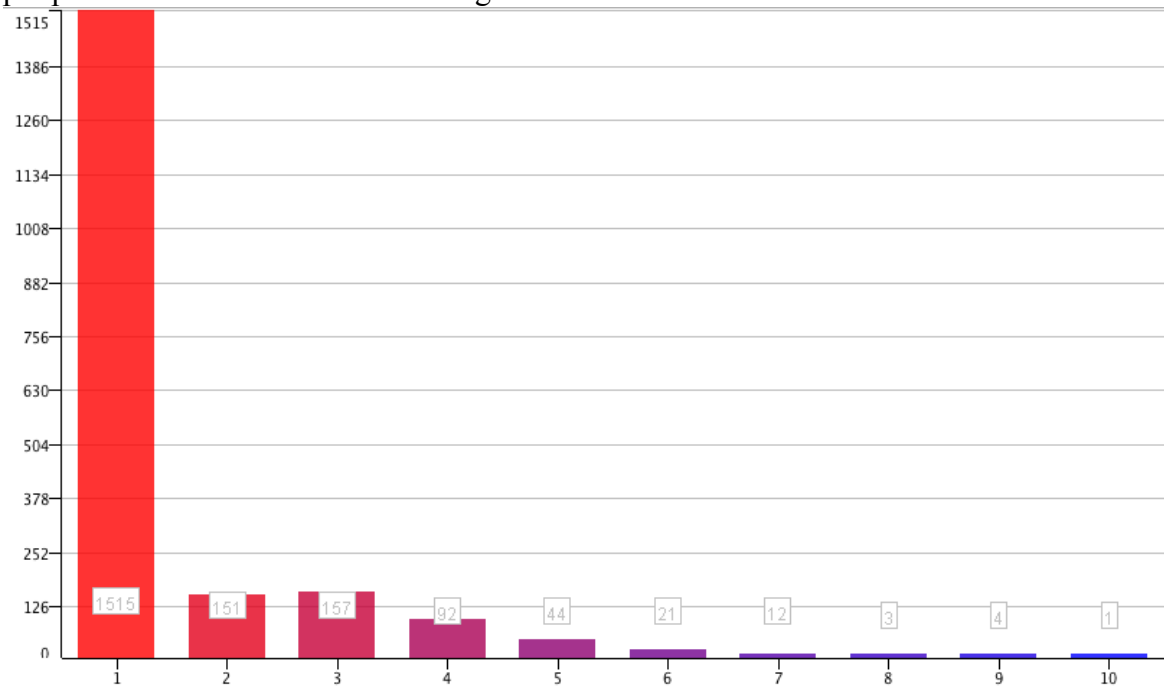

Figure 50 – Histogram distribution of number of times property has been sold

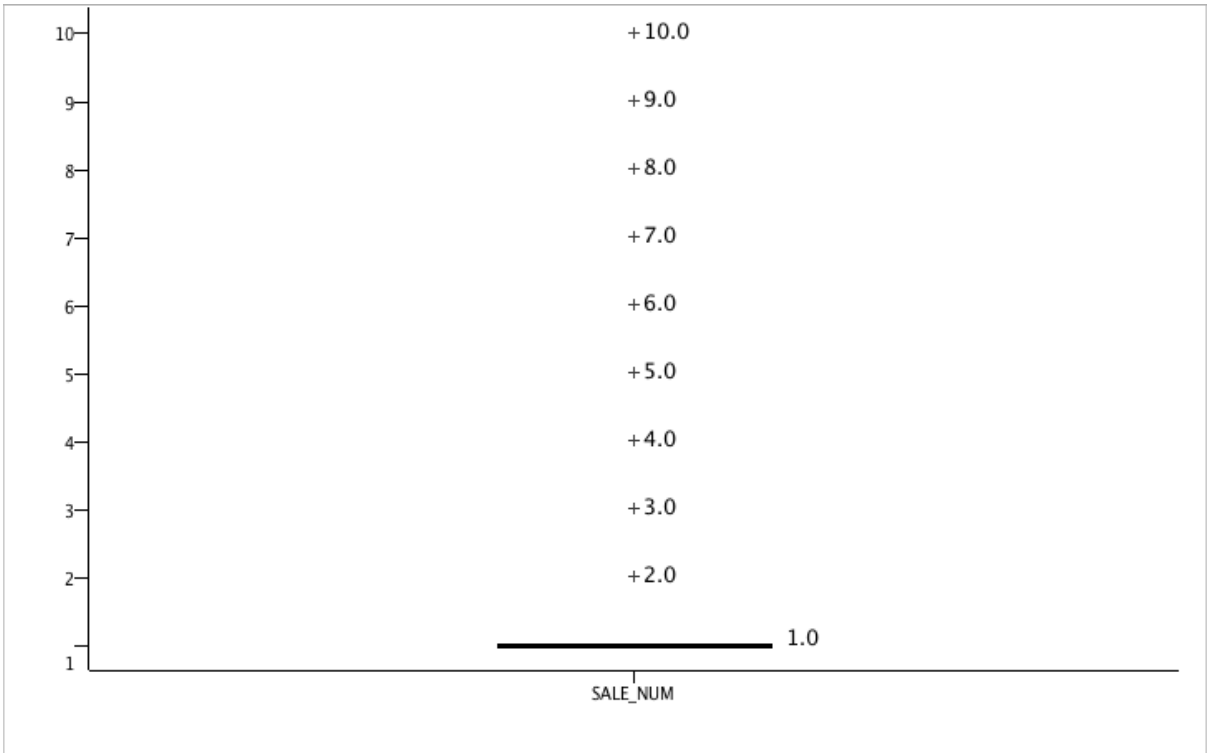

Figure 51 – Box plot of number times property has been sold

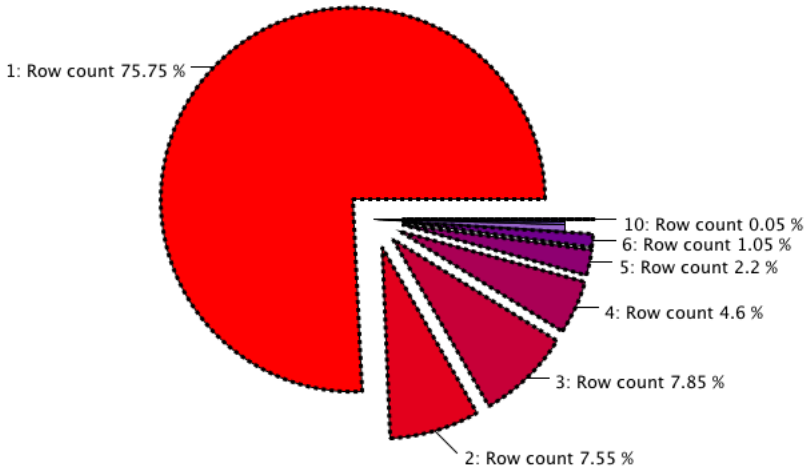

Figure 52 – Pie chart number times property has been sold

26. LANDAREA

Attribute Type:

Ratio.

The value of LANDAREA signifies the area of land the property is built on. As such, it can be used to order properties. Further, exact difference between land area is known and ratio of two land areas can easily be determined.

Statistics:

Statistic	Value
Range	520 – 43,809
Median	2,400
Mean	3,375.283
Variance	9,206,043.673
Standard Deviation	3,034.146
1st Quartile	1610
3rd Quartile	4120

Distribution & Frequency:

As can be seen from the box plot suggests the middle 50% of properties have land areas of 1610 – 4120 with the majority being between 520 -8100. Clear exceptions exist the largest land area being 43,809.

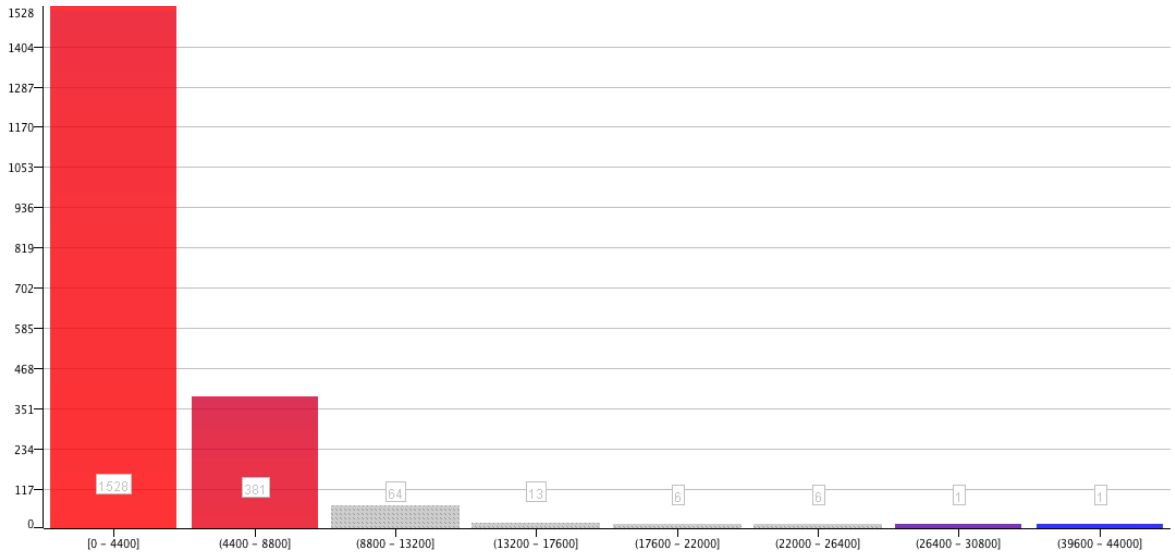

Figure 51 – Histogram distribution of property land area

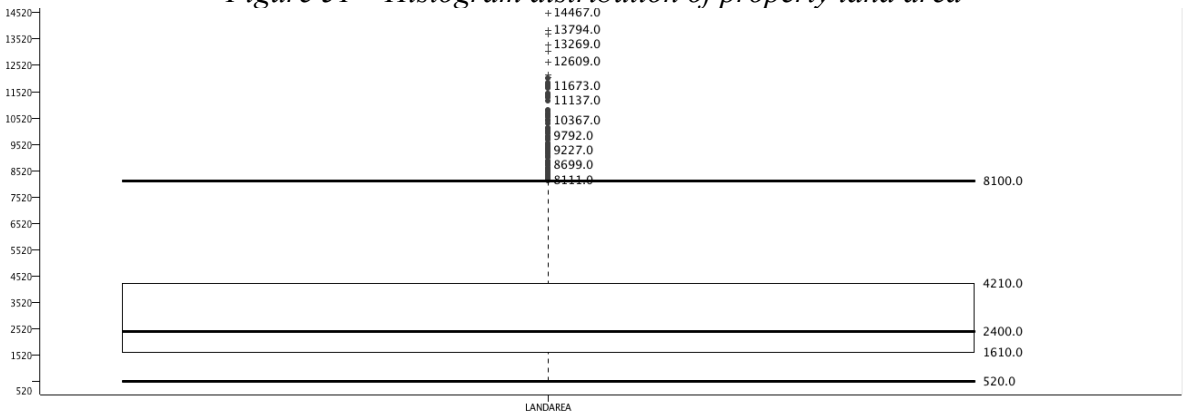

Figure 52 – Box plot of property land area

Clusters & Interesting Findings

Scatter plots were used to observer any clustering of data. Outliers are even more evident. Conclusions are detailed below of certain interesting scenarios.

1. Property Condition & Price

The hypothesis behind comparing condition and price of properties was the fact that the better the condition, the higher the price said property can be sold for. We need to keep in mind that other factors also play a part in deciding property price.

However, from figure 53, the clustering is mainly between the interquartile range of sold price highlighted in the section above. It is quite evident that for a property to demand an above average price, as demonstrated by the outliers, properties need to be in Good or better condition. Properties with Poor, Fair and Average conditions don't show even one instance of above average sold price.

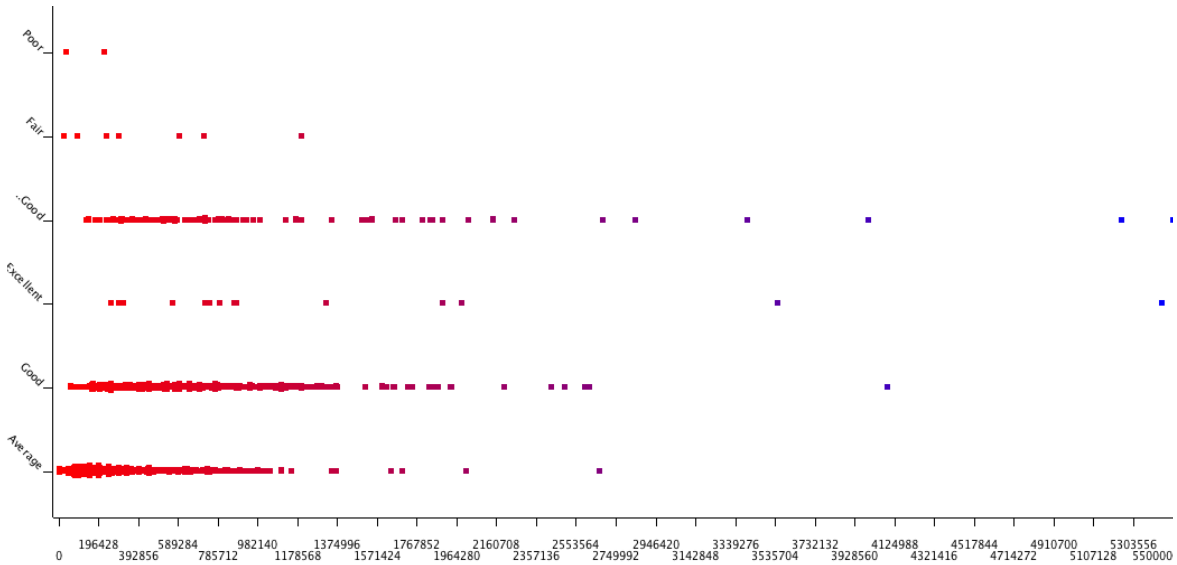

Figure 53 – Scatter Plot Price Vs Condition

2. Qualified & Price

It is to be noted that the majority of above average prices are for properties that are qualified (Blue). It is not certain what parameters need to be satisfied for a property to be qualified but form the scatter plot in Figure 54, it is evident that if price potential is to be maximised, property should be qualified.

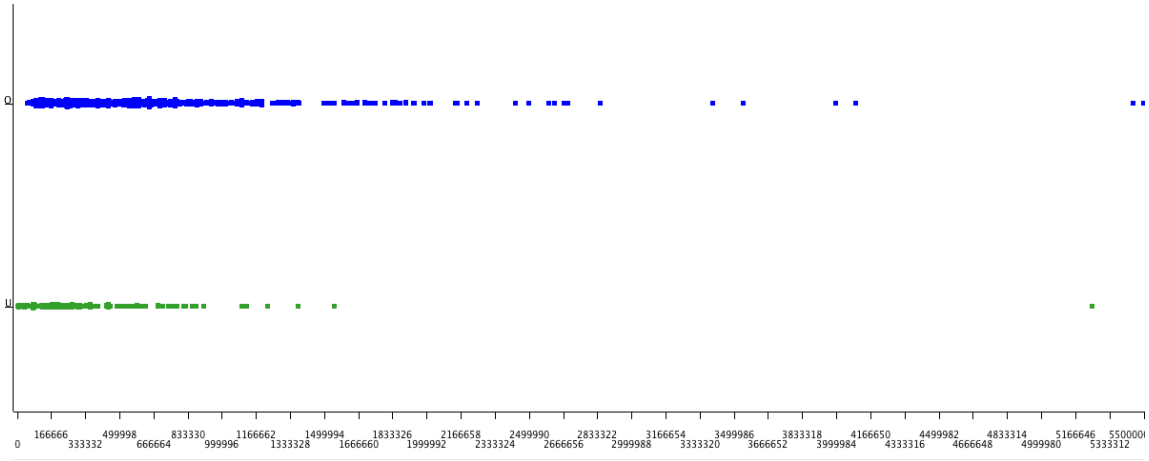

Figure 54 – Scatter Plot Price Vs Qualified

3. AC & Price

The big clusters as evident in the figure below are between average price range. Interestingly, properties seem to demand a higher price if they have air conditioning available as demonstrated by the outliers.

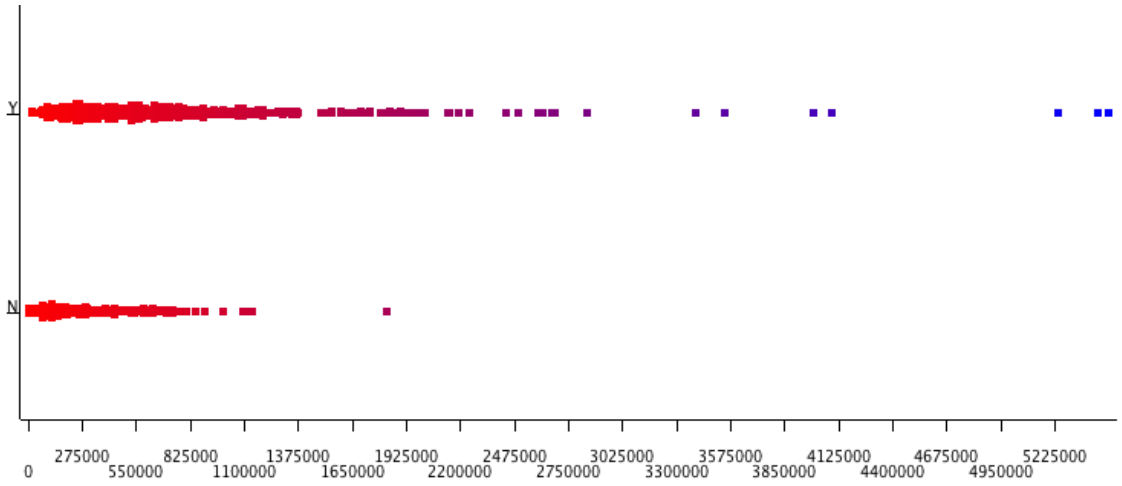

Figure 55 – Scatter Plot Price Vs Ac

4. Land Area & Price

The hypothesis behind analysing this pair was to try and prove the fact that larger land areas lead to a larger sold price. Main cluster can be noticed between the interquartile range for sold price. The hypothesis, however, was not confirmed after analysing the scatter plot in figure 56. L land areas, in instance, did garner a larger sold price, but so did properties with smaller areas. It can be concluded that sold price is not heavily dependent on land area and is most like influenced by more important factors such as location and condition.

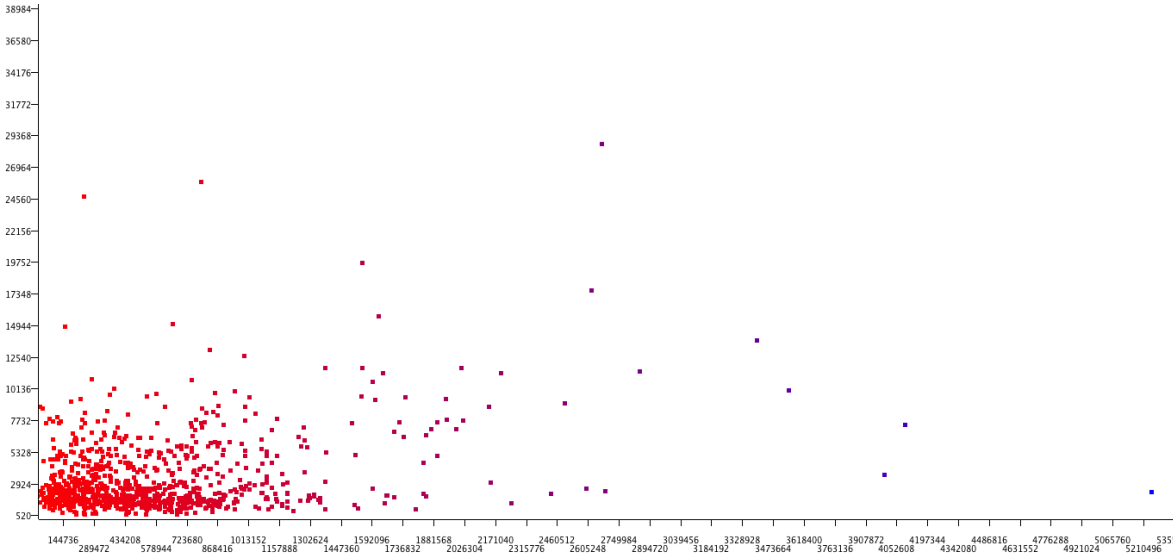

Figure 56 – Scatter Plot Price Vs Land Area

5. Remodelling & Price

As already concluded while analysing the year properties were remodelled, most of them went through this process between 1991 and 2011 as evident by the main cluster below. However, an interesting observation is the fact that most outliers in terms of sold price were remodelled between 2000-2018. Hence, it can be concluded that for a property to maximise sale price, it most likely needs to be remodelled.

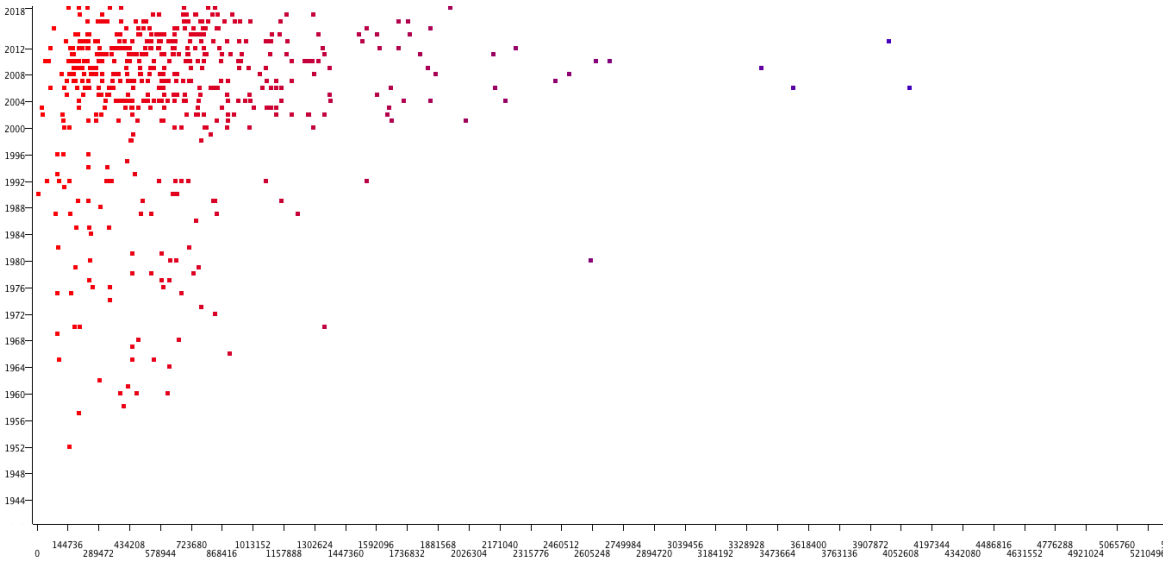

Figure 57 – Scatter Plot Price Vs Remodelling Data Pre-processing

All processed data is available in attached spreadsheet along with raw data. Each step has its own independent sheet with a similar label to the section names below.

Pre-processing

1. Equi – Width Binning

KNIME used to process raw data. Node setup is detailed in Figure 58 below.

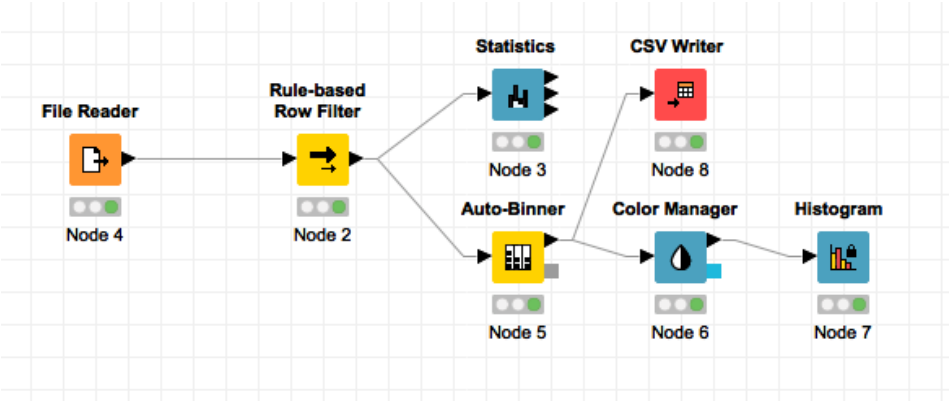

Figure 58 – Node setup for Binning

- Rows with 0 as price are omitted. These values are assumed to be default due to information not being available.
- Auto Binner node used. Price attribute selected for binning.
- Range of "Price" attribute is 9125 to 5,500,000. thus, Range 5,490,879. 5,490,879 rounded to 5,500,000 and divisible by 500,000. 5,500,000/500,000 = 11 and therefore number of bins chosen is 11 with width being 500,000.

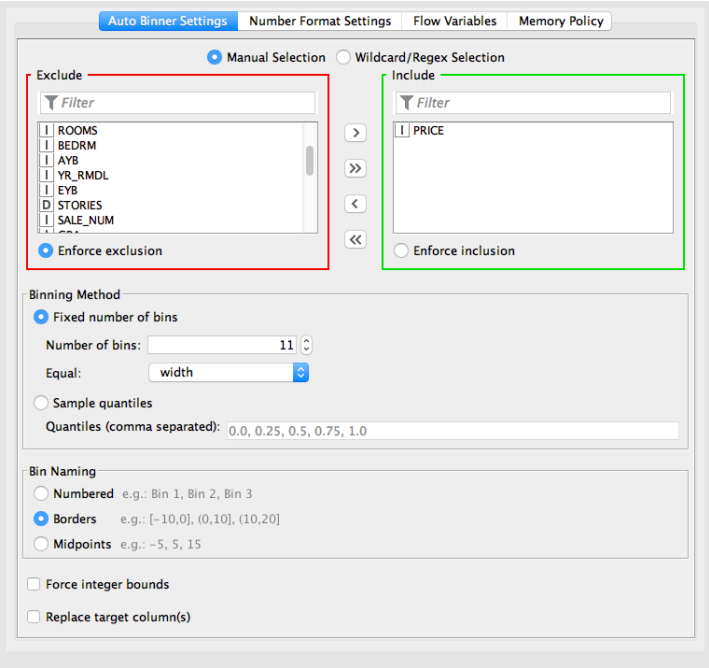

Figure 59 – Equi-width binning node setup

- Final step is to create a histogram distribution using the now binned data, shown below. As noted from Figure 60, most sold prices were placed in bin [0-500,00]. The least common bin is [4,000,000 – 4,500,000]

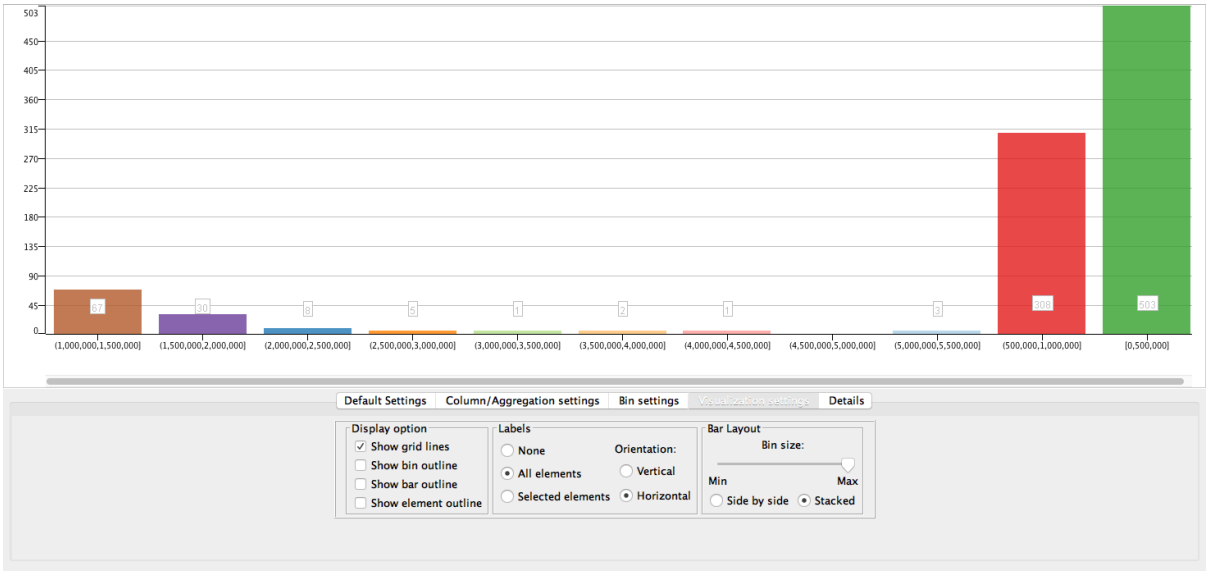

Figure 60 – Equi-width Price Histogram

2. Equi – Depth Binning

The node structure is exactly the same as the one used for Width binning with certain settings changed in the Binning node as highlighted below.

- The one setting changed was that of the Equal field, now being set to "frequency" also known as depth, followed by generation of histogram in Figure 61. As can be clearly seen in Figure 61, the distribution is a lot more even compared to equi-width binning.

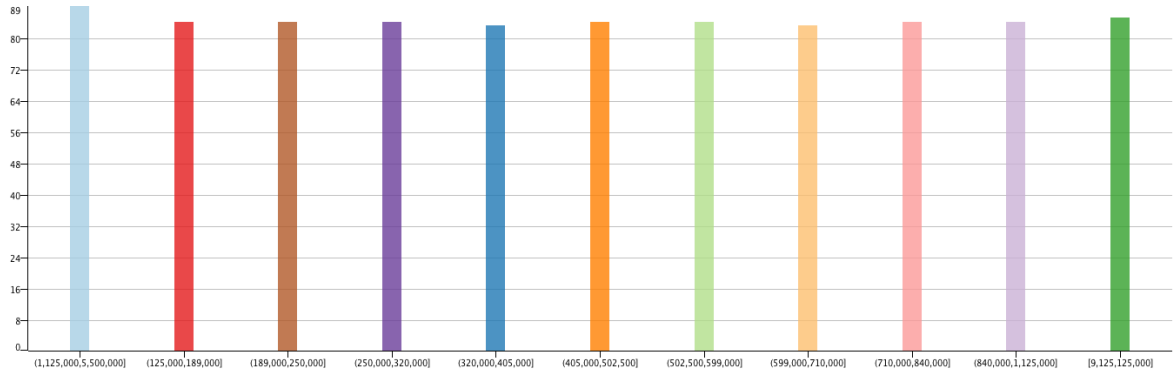

Figure 61 – Equi-depth Price Histogram

Normalise

KNIME used to normalise data using min-max & z-score normalisation. Processed data is included in attached excel workbook. Steps taken to achieve results outlined below.

Node setup highlighted below in Figure 61. Similar to binning, 0 price rows are omitted using the rule based row filter node. Generated results are stored in a csv file via the CSV writer node.

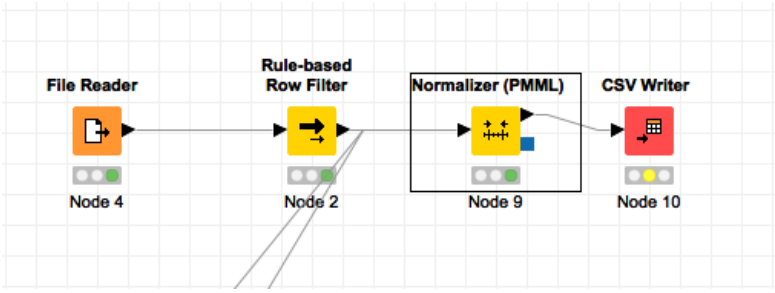

Figure 61 – Node setup for normalising price.

- Normaliser node is setup for min-max normalisation first and then to z-score normalisation, making sure to run the CSV writer once these settings are changed to save both results. Settings highlighted in Figure 62 and 63.

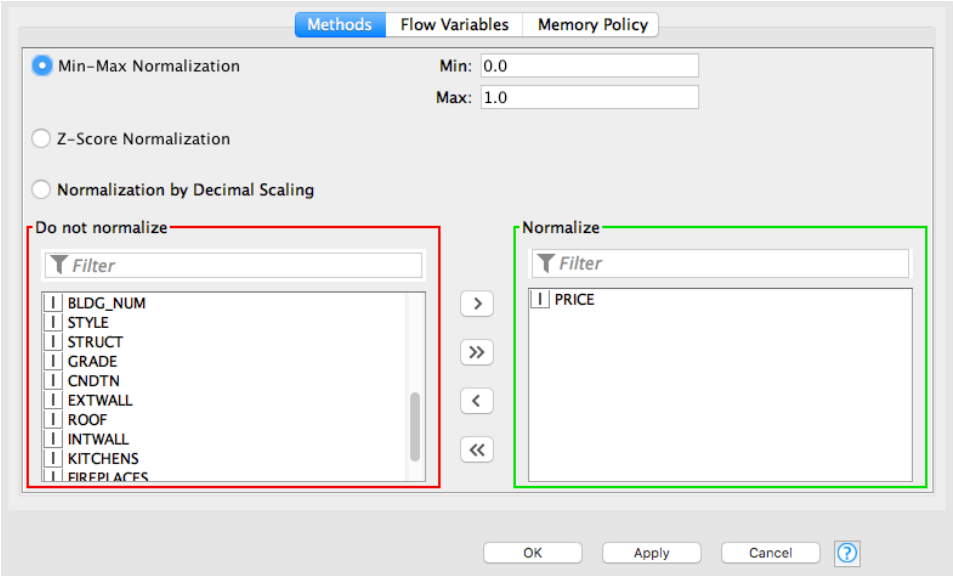

Figure 62 – Normaliser node setup for min-max

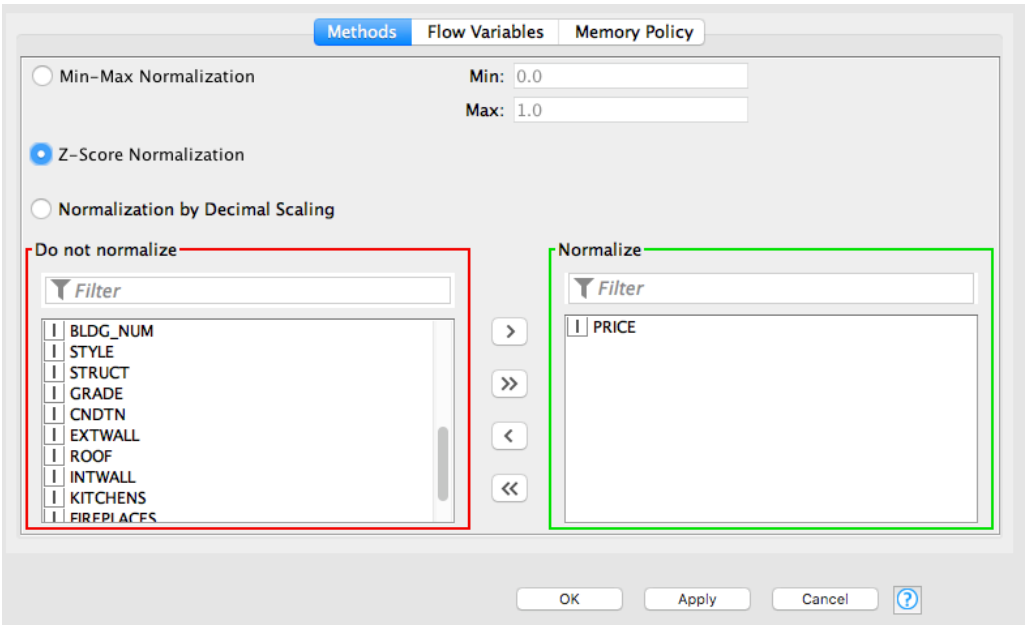

Figure 62 – Normaliser node setup for z-score

Discretise

Discretise is the process of categorising an attribute into predefined categories. For this instance, price has been categorised into the following: Low=0- 50k; Medium=51k-100k; High=101k-1000k; Expensive= 1001k+. Output data is attached to the excel workbook. Frequency of new categories provided below via a histogram plot in Figure 65.

Node setup using KNIME highlighted below in Figure 63.

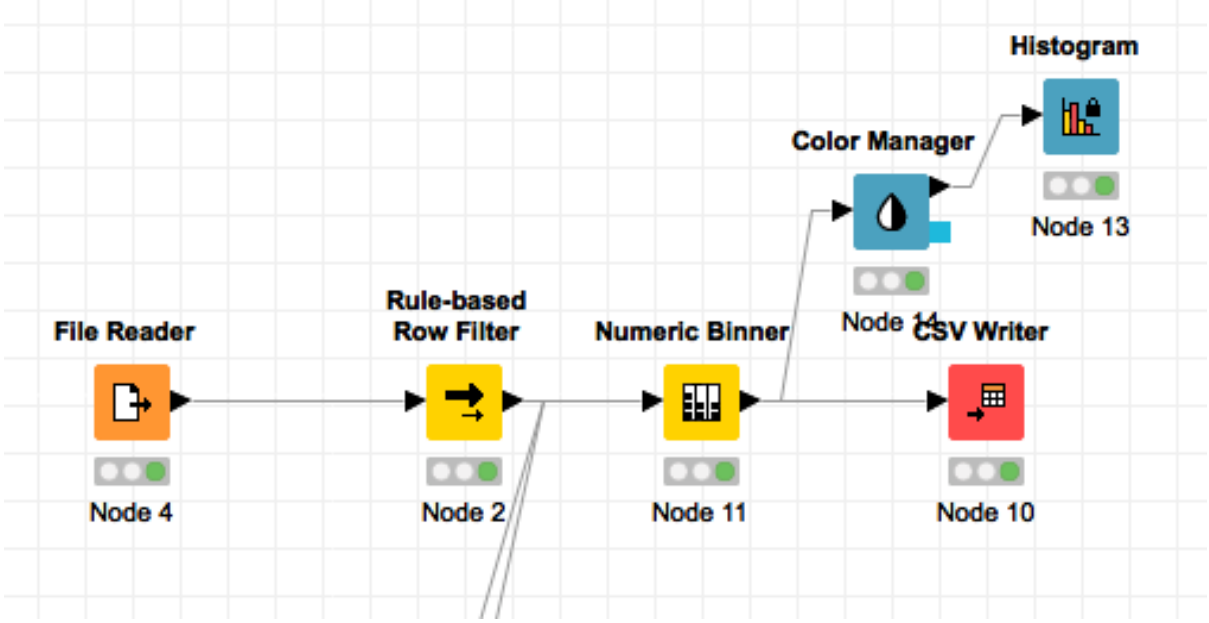

Figure 63 – Node setup for discretising price

- Rule based filter node used to filter out 0 price rows.
- Numeric binner node setup to bin price attribute according to specified categories. Settings highlighted in Figure 64

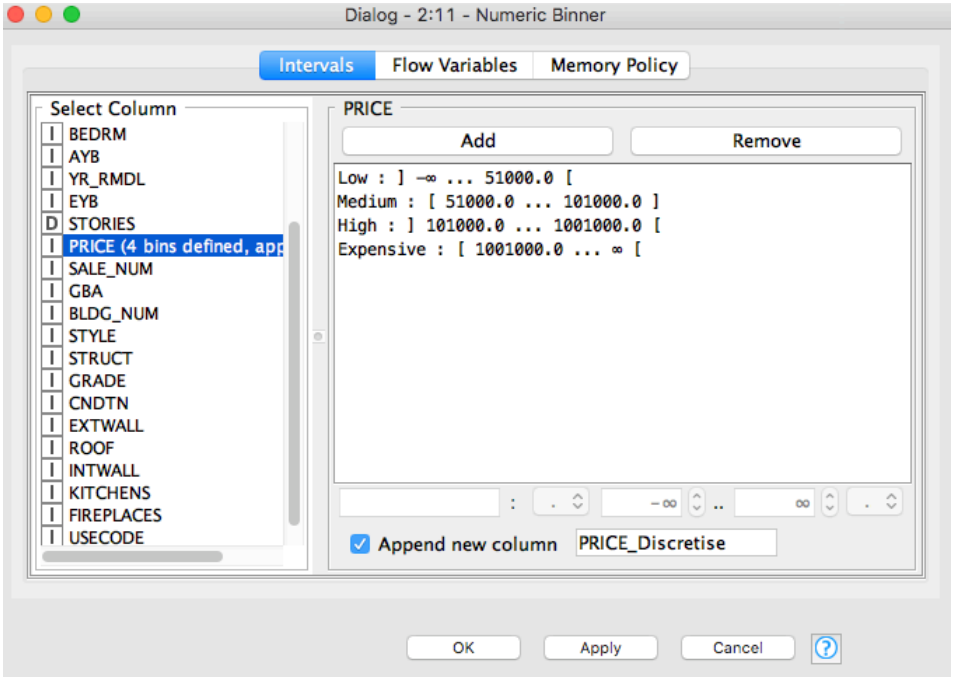

Figure 64 – Numeric node binner settings

- Resulting data saved using CSV write node.
-

Binarise

The premise behind binarising data is to have a 1(True) or 0 (False) result for a certain attribute. We are to binaries the STRUCT_D variable. This will be done using an excel workbook. Usually, strings such as the one provided for STRUCT_D need to be converted to a numerical format for the binarising process, but, this data is already provided to us in the STRUCT column. As such, we will use this column as the source.

- We need to obtain a binarisation column for each STRUCT_D category. Mapping and formulas used to do so are highlighted in the table below:

Assigned Value	STRUCT CATEGORY	Formula
1(AQ2)	Single	=IF(X11=$AQ$2,"1","0")
2(AQ3)	Multi	=IF(X11=$AQ$3,"1","0")
4(AQ4)	Town End	=IF(X11=$AQ$4,"1","0")
5(AQ5)	Town Inside	=IF(X11=$AQ$5,"1","0")
6(AQ6)	Row End	=IF(X11=$AQ$6,"1","0")
7(AQ7)	Row Inside	=IF(X11=$AQ$7,"1","0")
8(AQ8)	Semi Detached	=IF(X11=$AQ$8,"1","0")

Summary

A variety of observations were made about this particular dataset. They can be categorised into attribute values, clusters, distribution, outliers and topics of further interest.

Attributes, Distribution and Outliers

Detailed analysis of each attribute was done in the report previously. The most relevant attributes for this scenario seem to be ROOMS, PRICE, LANDAREA, CONDITION, YR_RMDL and BATHROOM.

ROOMS

The value of this attribute ranged from 0 to 30. The interquartile range of this attribute was heavily centred between 6-8 rooms with one property with 30 rooms being a clear outlier.

PRICE

This attribute seems to have a keen relationship with several other attributes and essential in further investigations in relation to the dataset. The price was heavily concentrated between 241,750 and 746,250 and ranging from 9,125 to 5,500,000.

LANDAREA

LANDAREA ranges from 520 – 43,809 with middle 50% of properties having land areas of 1610 – 4120 with the majority being between 520 -8100. Clear exceptions exist with the largest land area being 43,809.

CONDITION
Condition of properties was a nominal/categorical attribute. It was noted that the majority of properties were in Average condition (1162 or 58.13%) followed by "Good" with 647 properties (32.37%). The least common condition is "Poor" with only 4 (0.2%) property with the condition, being a clear outlier.

YR_RMDL
Most of the remodelling done on properties was between 1991 and 2011, peak period being between 2010 and 2012 where 98 properties were remodelled. All remodelling done before 1961 are statistical outliers.

BATHROOM
Range of bathrooms per property was 1-7. All properties had at least 1 bathroom with 18 properties having 6 or more. The box plot shows that the interquartile range is between 1 and 3 and it can be concluded as a result that most properties have between 1 and 3 bathrooms with 7 being an outlier.

Clusters & Outliers

Property Condition & Price

The clustering is mainly between the interquartile range of sold price highlighted in the section above and consists of a wide variety of property conditions. However, it is quite evident that for a property to demand an above average price, as demonstrated by the outliers, properties need to be in Good or better condition. Properties with Poor, Fair and Average conditions don't show even one instance of above average sold price.

Qualified & Price
It is to be noted from the cluster that the majority of above average prices are for properties that are qualified. It is not certain what parameters need to be satisfied for a property to be qualified but it is evident that if price potential is to be maximised, property should be qualified.

AC & Price
The big clusters as evident and are between average price range. Interestingly, properties seem to demand a higher price if they have air conditioning available as demonstrated by the outliers.

Remodelling & Price
It was evident while analysing when properties were remodelled that most of the, went through the process between 1991 and 2011 as evident by the main cluster. However, an interesting observation is the fact that most outliers in terms of sold price were remodelled between 2000-2018. Hence, it can be concluded that for a property to maximise sale price, it most likely needs to be remodelled.

Topics of Further Interest

Price & Land Area

The hypothesis behind analysing this pair was to try and prove the fact that larger land areas lead to a larger sold price. Main clusters can be noticed between the interquartile range for sold price. The hypothesis, however, was not confirmed after analysing the scatter plot in figure 56. Larger land areas, in instances, did garner a larger sold price, but so did properties with smaller areas. It can be concluded that sold price is not heavily dependent on land area and is most like influenced by more important factors such as location and condition.

It would be interesting to analyse and conclude the factors most likely to influence property sold price.

Conclusion

The topics of further interest should be analysed to come to a better understanding of factors influencing price of properties. This information can then be used to maximise asking price if and when a property is to be sold.