DATA EXPLORATION AND PREPARATION

Introduction to Data Analytics (31250)

Assignment 2

Data Exploration and Preparation for insurance company dataset

Michelle Tanoto
13175144@student.uts.edu.au

Table of Contents

1A: Initial Data Exploration

The purpose of this section is to identify each attribute type in the dataset with the justification as well as identify the summarising properties and any interesting attributes.

Attributes

Quote_Id

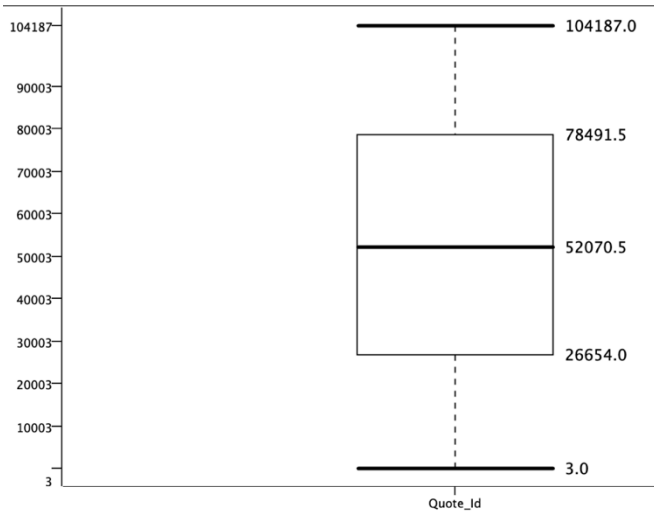


Figure 1 Box Plot of Quote_Id

Quote_Id is a nominal attribute type since the attribute values are distinct for each data points and based on the attribute name, it can be assumed it function as an identifier to distinguish each quote and it also does not present any meaningful order. Based on Figure 1, it can be seen that the Quote_Id ranges from 3 to 104187.

Quote_Date

Quote_Date is an interval attribute type as it possesses interval attribute type characteristics, namely the differences between the values are meaningful and there is no absolute zero point. For example, based on Figure 2, we can get the total amount of quotes during a specific period. The Quote_Date ranges from 2013-02-01 to 2018-05-18. In addition, the dates where it has the most quotes are 2013-08-05 and 2015-04-30 with 12 data points for each one of them.

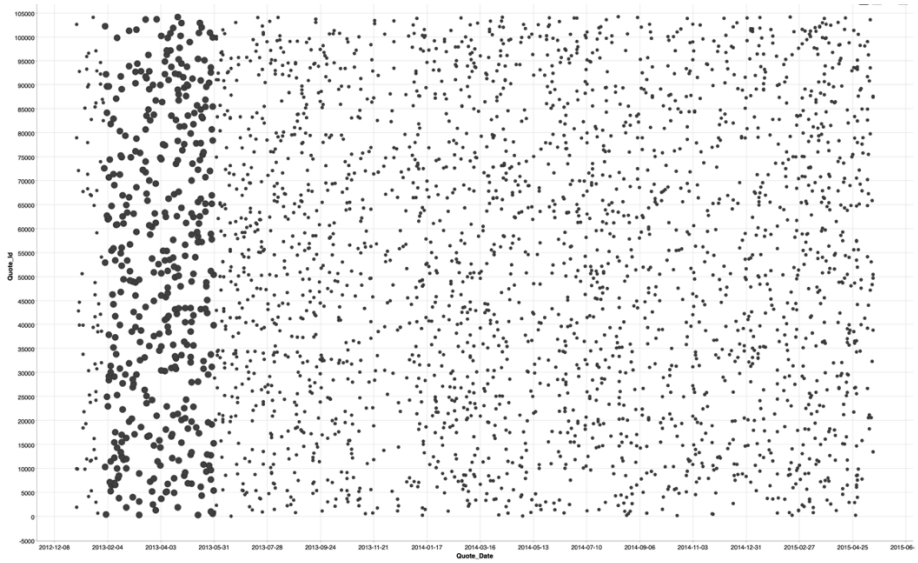


Figure 2 Scatter Plot of Quote_Date and Quote_Id

Quote_Flag

Quote_Flag is a nominal attribute type as the attribute values only act as labels and it is unordered. Based on Figure 3 it can be seen it only has two distinct values, such as '0' and '1' which means it is a binary attribute (attribute that only has 2 values/states). In summary, '0' is the most common value with 0.8037 proportion and a frequency of 2411. Table 1 shows the summary of the Quote_Flag frequency and distribution.

Quote_Flag	Proportion	Frequency
0	0.8037	2411
1	0.1963	589
Total	1	3000

Table 1 Quote_Flag frequency and distribution

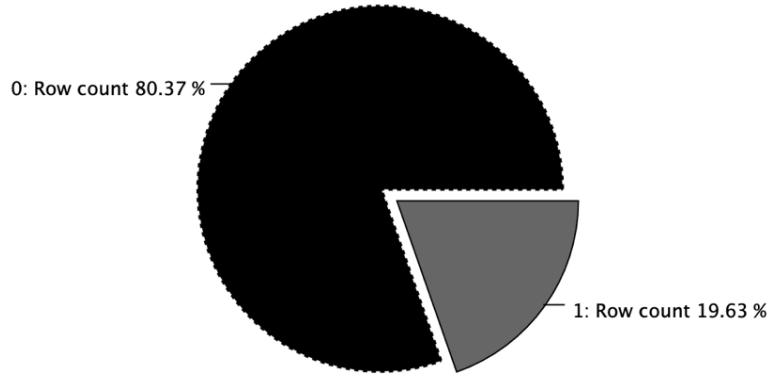


Figure 3 Pie Chart of Quote_Flag

Field_info1

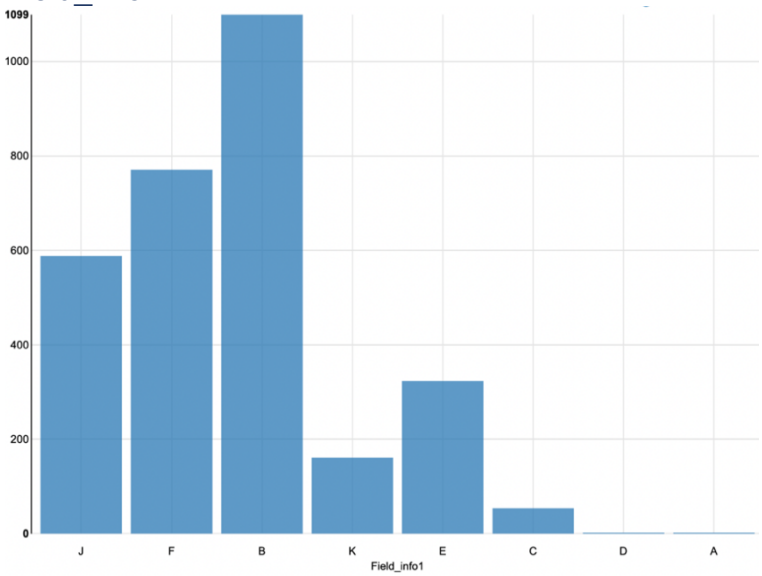


Figure 4 Bar Chart of Field_info1

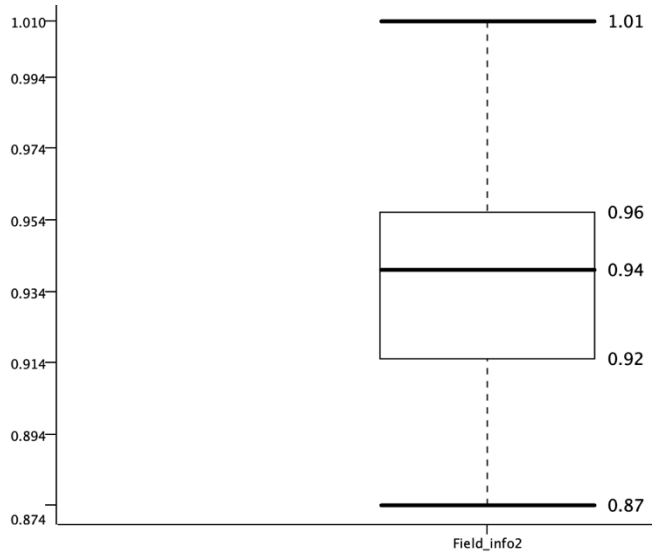
Field_info1 is a nominal attribute type as there is no ordering and it seems like it acts as labels. In Figure 4, It can be seen that the possible values of Field_info1 are A, B, C, D, E, F, J and K. Table 2 shows the summary of Field_info1 frequency and distribution. In summary, 'B' is the most common value with 0.3663 proportion and frequency of 1099.

Field_info1	Proportion	Frequency
A	0.0006	2
B	0.3663	1099
C	0.018	54
D	0.0006	2
E	0.1076	323

F	0.257	771
J	0.196	588
K	0.053	161
Total	1	3000

Table 2 Field_info1 frequency and distribution

Field_info2



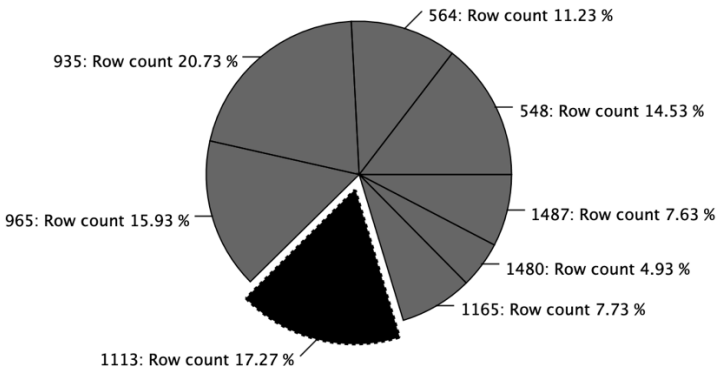
Field_info2 is a ratio attribute type as there is no negative value in the data points and it can be ordered. Figure 5 shows the box plot of Field_info2 with Table 3 showing the base statistics.

Figure 5 Box Plot of Field_info2

Statistic	Value
Range (Min – Max)	0.875 – 1.01
Mean	0.938
Q1	0.915
Median	0.94
Q3	0.957
Standard Deviation	0.036
Variance	0.001

Table 3 Field_info2 statistics

Field_info3



Field_info3 is nominal attribute type. Based on Figure 6, the value does not seem to be unique as the attribute values represent a specific value. Besides, the total occurrence of a specific value is too much to treat it as an arbitrary value. Therefore, the value behaves more as a category rather than numeric.

Figure 6 Pie Chart of Field_info3

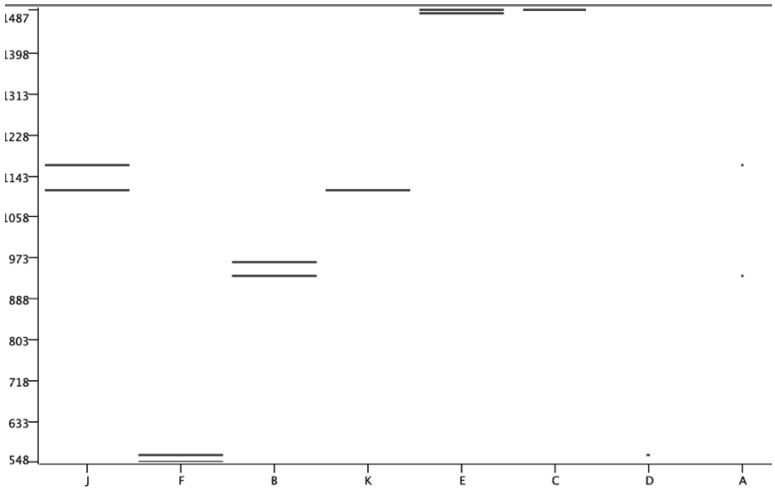


Figure 7 Scatter Plot of Field_info1 and Field_info3

Moreover, there seems to be a pattern found between Field_info1 and Field_info3 attribute. In Figure 7, it can be seen that each Field_info1 attribute value tend to reside near a specific Field_info3 attribute value although A has 2 values that are quite far which could be considered as outliers. Table 4 shows the range of each Field_info3 value.

Field_info3	Range
A	935 - 1165
B	935 – 965
C	1487
D	564
E	1480 - 1487
F	548 – 564
J	1113 – 1165
K	1113

Table 4 Field_info3 value range

Field_info4

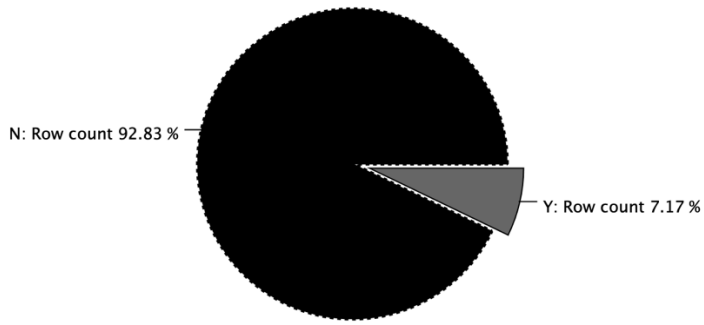


Figure 8 Pie Chart of Field_info4

Field_info4 is a nominal attribute type as it has no order and there is no distance between 'Y' and 'N'. Based on Figure 8, it can be seen that 'N' is the most common value with 0.928 proportion and 215 frequency. Table 5 shows the summary of Field_info4 frequency and distribution.

Field_info4	Proportion	Frequency
Y	0.072	2785
N	0.928	215
Total	1	3000

Table 5 Field_info4 frequency and distribution

Coverage_info1

Coverage_info1 is an interval attribute type as it seems to be ordered and the difference between the values are meaningful. Based on Figure 9, it can be observed

that the Coverage_info1 value are distributed fully in all [-1,25] range. Because it is fully distributed from -1 to 25, each point has equal distance from one another. Table 6 shows the summary statistics of the attribute.

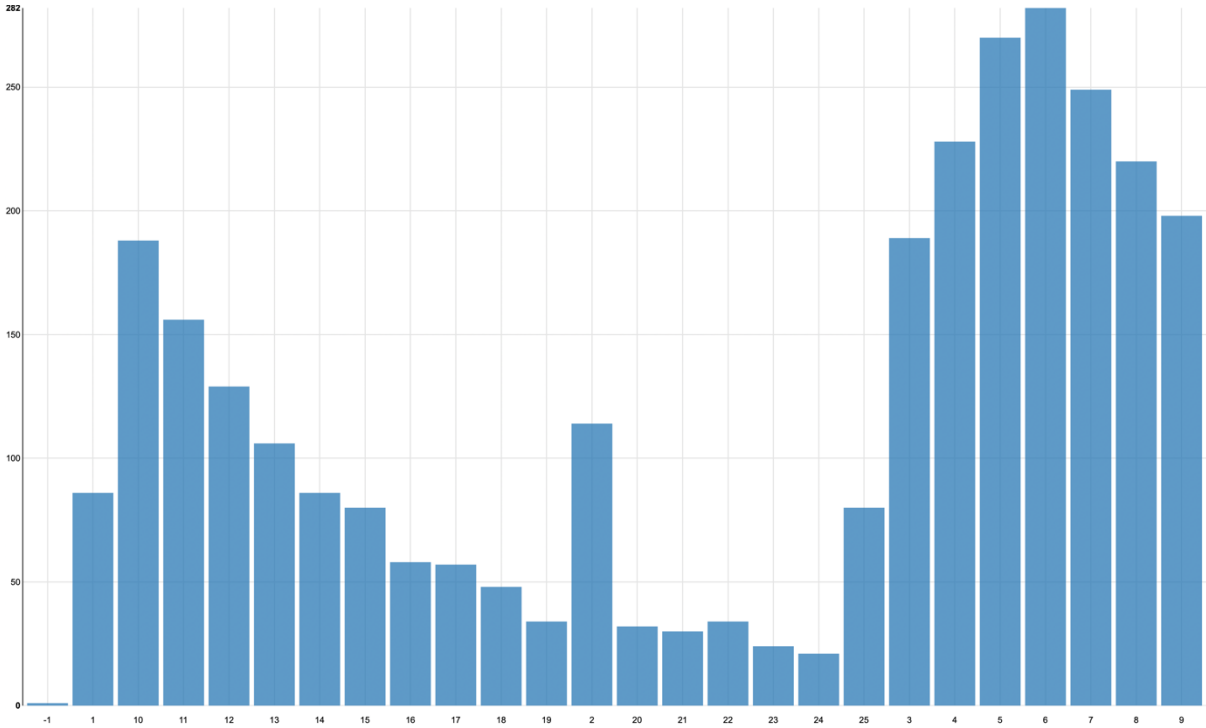


Figure 9 Bar Chart of Coverage_info1

Statistic	Value
Range (Min – Max)	(-1) – 25
Mean	9.175
Q1	5
Median	8
Q3	12
Standard Deviation	5.716
Variance	32.677

Table 6 Coverage_info1 statistics

Coverage_info2

Coverage_info2 is nominal attribute type. The occurrences of the same value are too much to assume it is just an arbitrary numeric data. Therefore, it can be assumed that it functions as a category. Based on Figure 10, it can be seen that it only has 4 distinct values, which consists of 1, 2, 22 and 25 with '22' as the most common value with 2451 frequency and 0.817 proportion. Table 7 shows the Coverage_info2 frequency and distribution.

Coverage_info2	Proportion	Frequency
1	0.0003	1
2	0.060	179
22	0.817	2451
25	0.123	369
Total	1	3000

Table 7 Coverage_info2 frequency and distribution

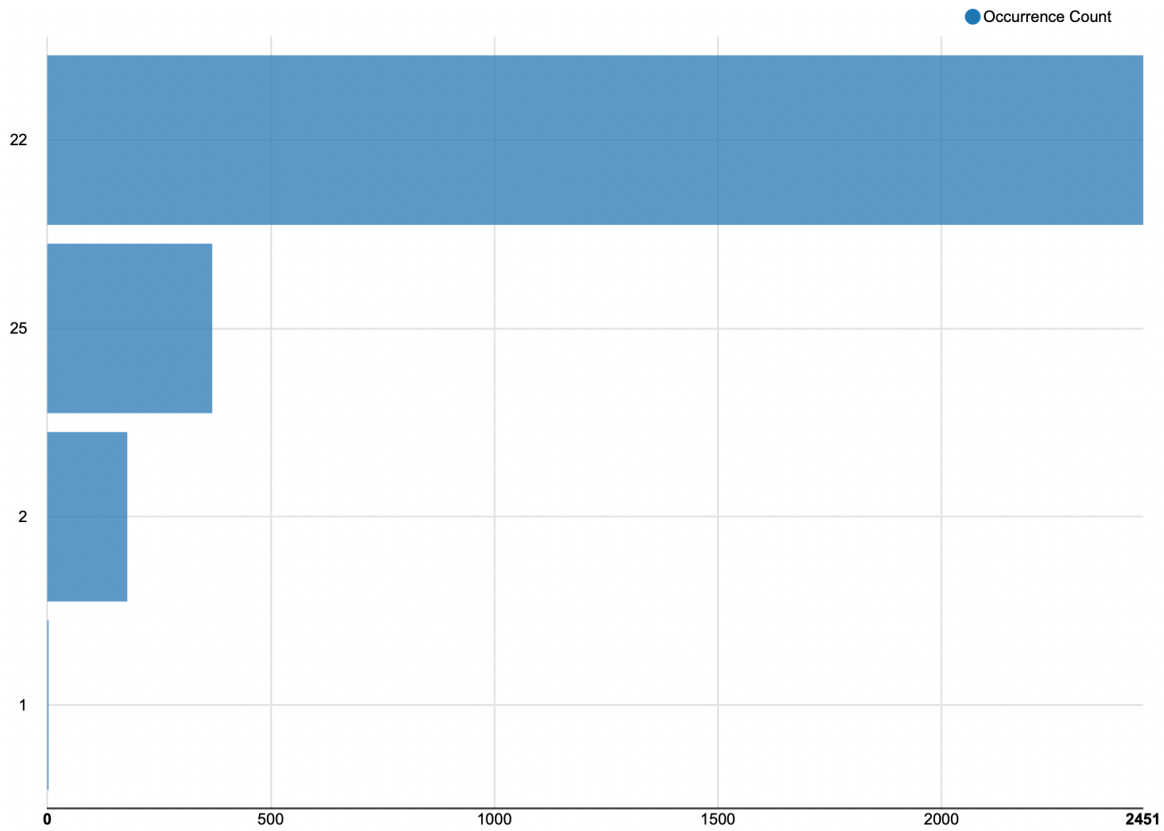


Figure 10 Bar Chart of Coverage_info2

Coverage_info3

Coverage_info3 is an ordinal attribute type. Based on Figure 11, the attribute values seem to be ordered from A – L although the magnitude between the values is not known. Table 8 shows the frequency and distribution of Coverage_info3.

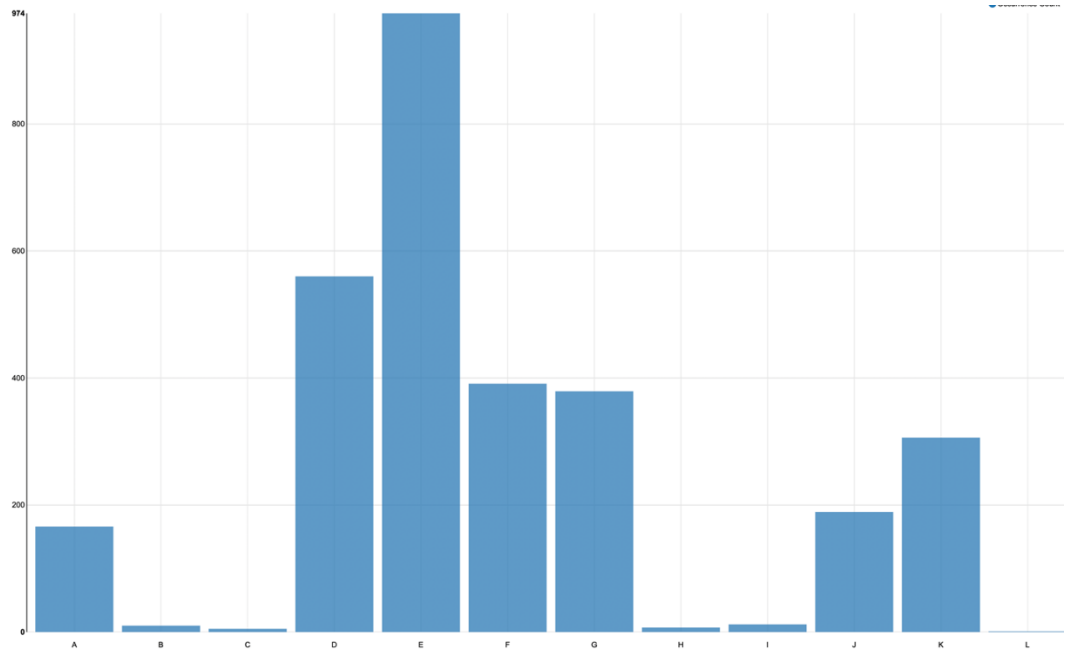


Figure 11 Bar Chart of Coverage_info3

Coverage_info3	Proportion	Frequency
A	0.055	166
B	0.003	10
C	0.001	5
D	0.187	560
E	0.325	974
F	0.130	391
G	0.126	379
H	0.002	7
I	0.004	12
J	0.063	189
K	0.102	306
L	0.0003	1
Total	1	3000

Table 8 Coverage_info3 frequency and distribution

Sales_info1

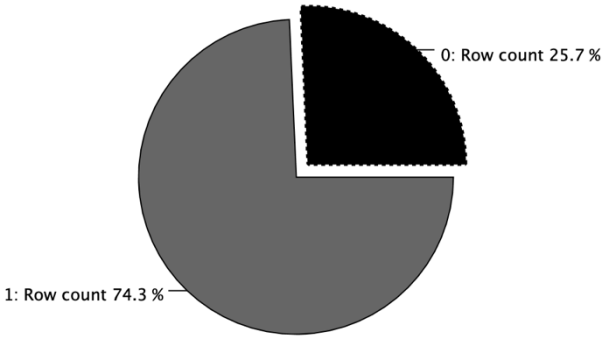


Figure 12 Pie Chart of Sales_info1

Sales_info1 is a nominal attribute type because there is no ordering and based on Figure 12 it can be seen it only has two distinct values, such as '0' and '1' which can be represented as categories. The attribute is highest on value '1' with 0.743 proportion and a frequency of 2229. Table 9 shows the summary of the Sales_info1 frequency and distribution.

Sales_info1	Proportion	Frequency
0	0.257	771
1	0.743	2229
Total	1	3000

Table 9 Sales_info1 frequency and distribution

Sales_info2

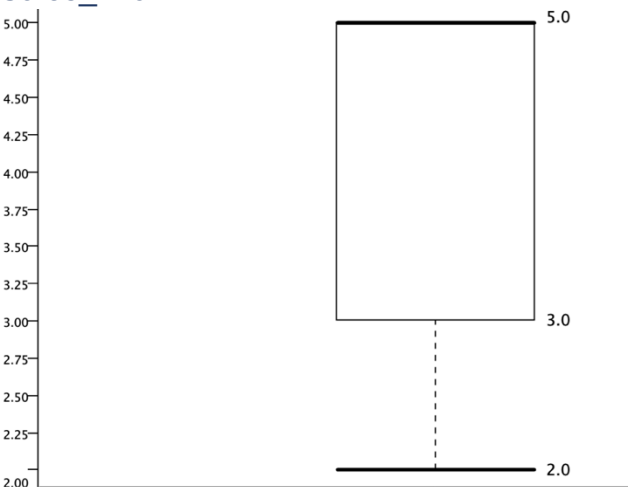
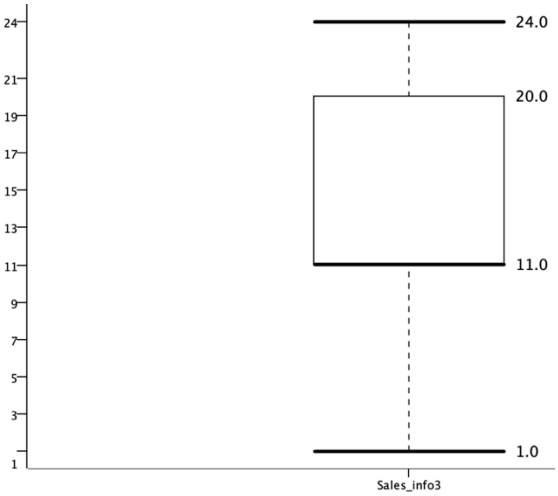


Figure 13 Box Plot of Sales_info2

Sales_info2 is a nominal attribute type because there are only four possible values which are 2,3,4 and 5. Therefore, it is reasonable to treat it as a category and there is no relationship or pattern found between the attribute and other attributes. Figure 13 shows the distribution of Sales_info2.

Sales_info3



Sales_info3 is an interval attribute type because it seems to be ordered from 1 – 24. Figure 14 shows the Sales_info3 attribute distribution and Table 10 shows the base statistics of Sales_info3 attribute.

Figure 14 Box Plot of Sales_info3

Statistic	Value
Range (Min – Max)	1 – 24
Mean	13.988
Q1	11
Median	11
Q3	20
Standard Deviation	6.31
Variance	39.817

Table 10 Sales_info3 statistics

Sales_info4

Sales_info4 is a nominal attribute as the values only act as labels and there is no ordering. Based on Figure 15, the range of values includes K, M, P, Q, R, T, V. In summary, the highest Sales_info4 attribute value is on K with 579 frequency and the lowest is M with 191 frequency as shown in Table 11.

Sales_info4	Proportion	Frequency
K	0.193	579
M	0.064	191
P	0.194	581
Q	0.150	449
R	0.008	239
T	0.173	519
V	0.147	442
Total	1	3000

Table 11 Sales_info4 frequency and distribution

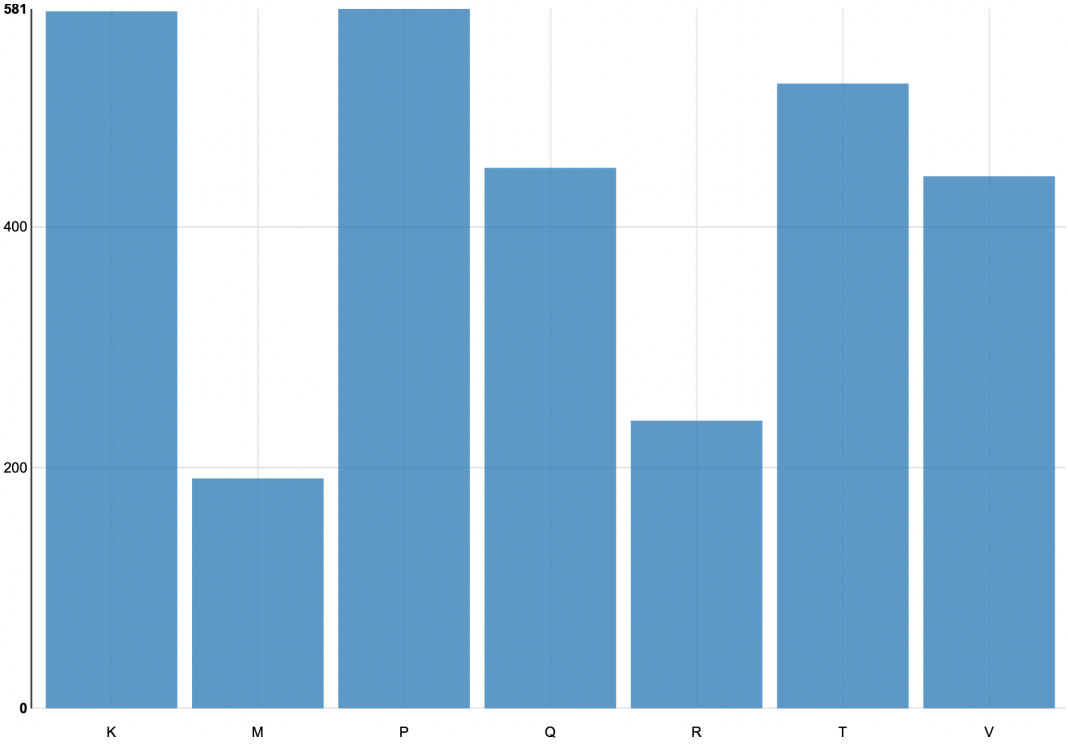


Figure 15 Bar Chart of Sales_info4

Sales_info5

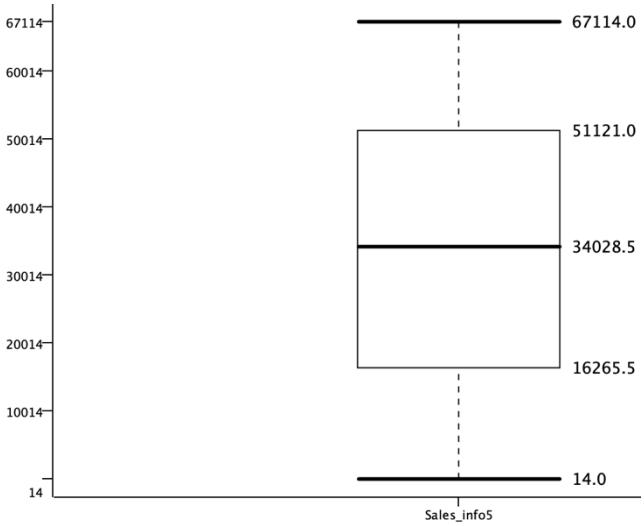


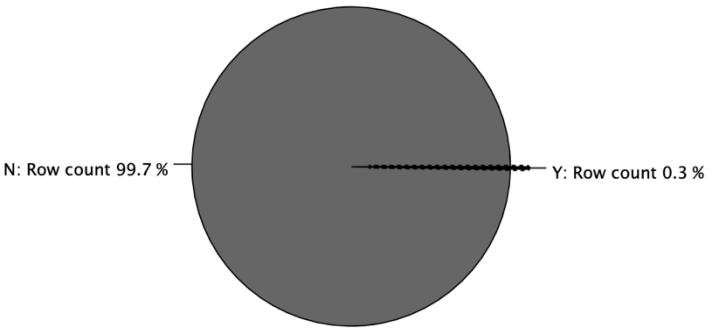
Figure 16 Box Plot of Sales_info5

Sales_info5 is an interval attribute type. There is no relationship found with other attributes as the values are quite different for each data points. Table 12 shows the base statistics of Sales_info5 and Figure 16 shows the attribute distribution.

Statistic	Value
Range (Min – Max)	14 – 67114
Mean	33841.4
Q1	16265.5
Median	34028.5
Q3	51121.0
Standard Deviation	19669.7
Variance	3.8689

Table 12 Sales_info5 statistics

Personal_info1



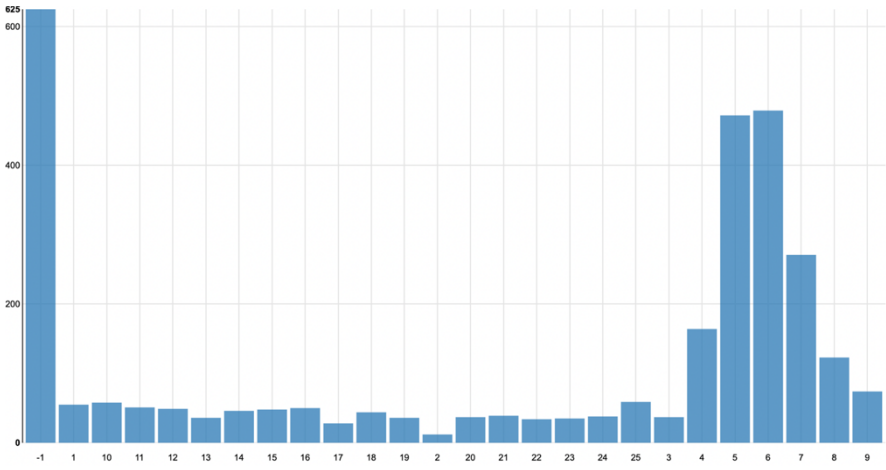
Personal_info1 is a nominal attribute type as it has no order and there is no distance between 'Y' and 'N'. Based on Figure 17, it can be seen that 'Y' is the most common value with 0.97 proportion and 2991 frequency. Table 13 shows the summary of Personal_info1 frequency and distribution.

Figure 17 Pie Chart of Personal_info1

Personal_info1	Proportion	Frequency
Y	0.97	2991
N	0.03	9
Total	1	3000

Table 13 Personal_info1 frequency and distribution

Personal_info2



Personal_info2 is an interval attribute type as it seems to be ordered and the distance between the values are meaningful. Based on Figure 18 it can be observed that the attribute values are distributed fully in all range [-1, 25].

Figure 18 Bar Chart of Personal_info2

Moreover, interestingly it has the same value range as Coverage_info1 attribute. Another interesting find is in Coverage_info1, '-1' value has the lowest frequency but in Personal_info2, '-1' has the most frequency out of all values. However, there seems to be no pattern found between these two attributes based on Figure 19.

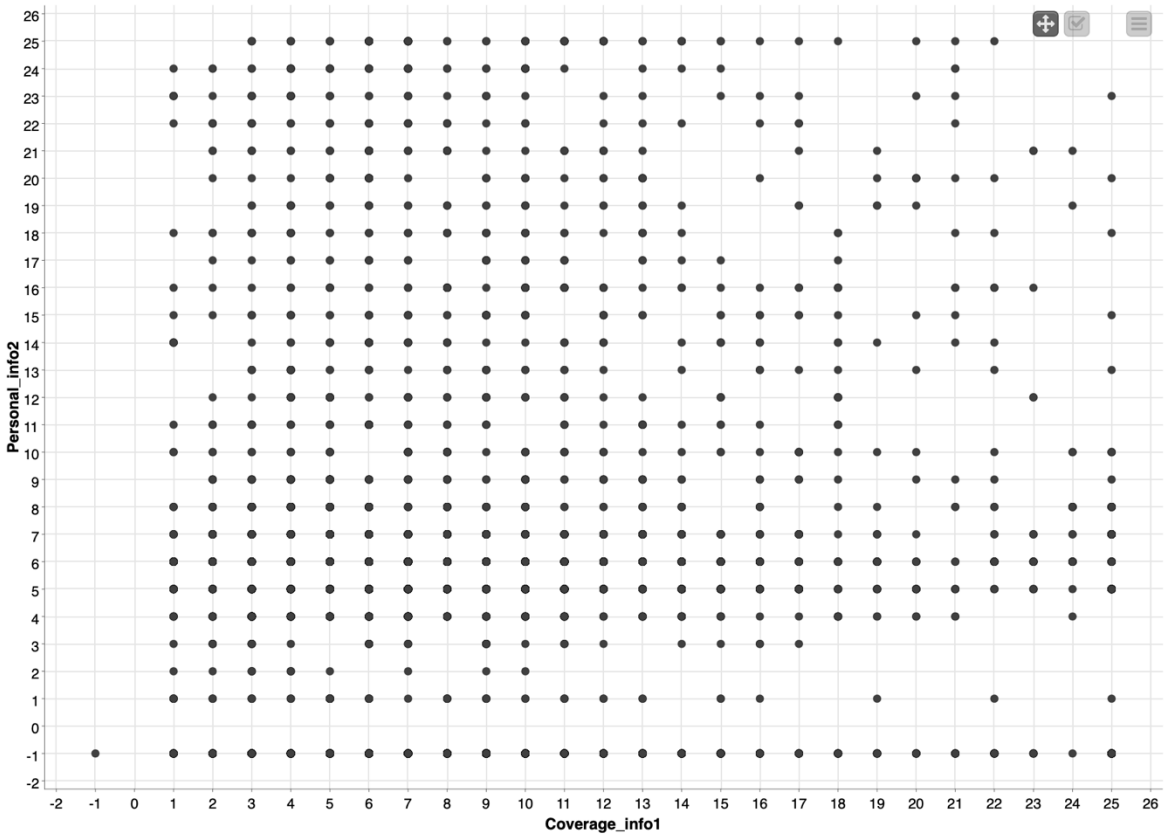


Figure 19 Scatter Plot of Coverage_info1 and Personal_info2

Personal_info3

Personal_info3 is a nominal attribute type since there is no ordering and it seems like it represents some kind of code which could be assumed used to categorise where each data points belongs to. Figure 20 shows the distribution of Personal_info3 attribute with 'ZA' having the most data points, counted 1427 and 'XK', 'YI' and 'ZU' having the lowest data points, counted 3.

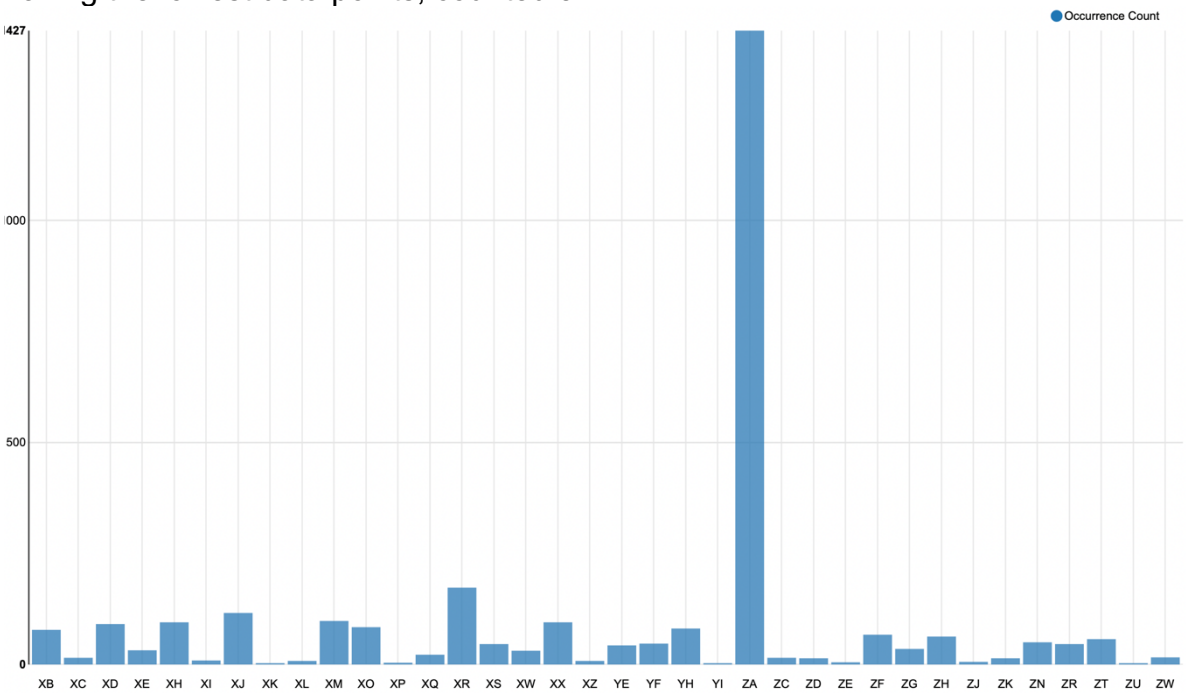


Figure 20 Bar Chart of Personal_info3

Personal_info4

Personal_info4 is a nominal attribute type. Based on Figure 21, It is not an interval attribute type as the distinct values are too less therefore it is more suitable to treat it as a categorical data. Table 14 shows the frequency and distribution of the attribute.

Personal_info4	Proportion	Frequency
0	0.9986	2996
1	0.0006	2
2	0.0003	1
6	0.0003	1
Total	1	3000

Table 14 Personal_info4 frequency and distribution

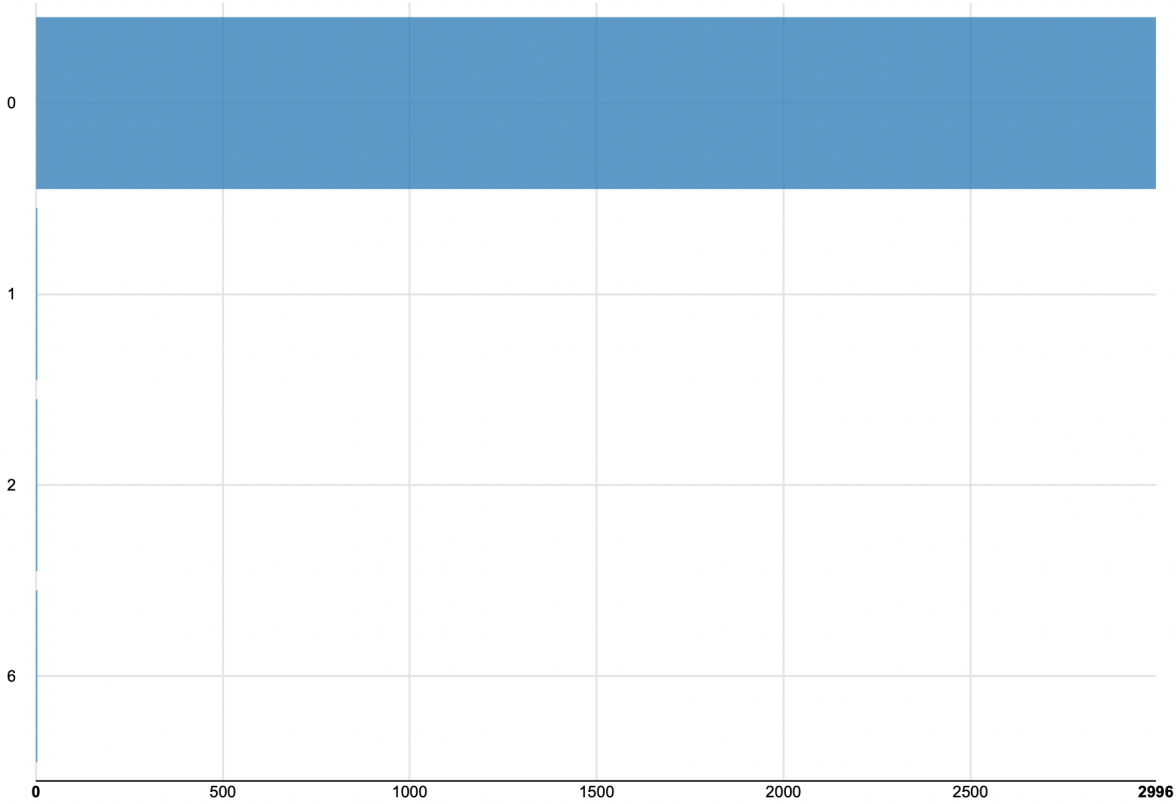
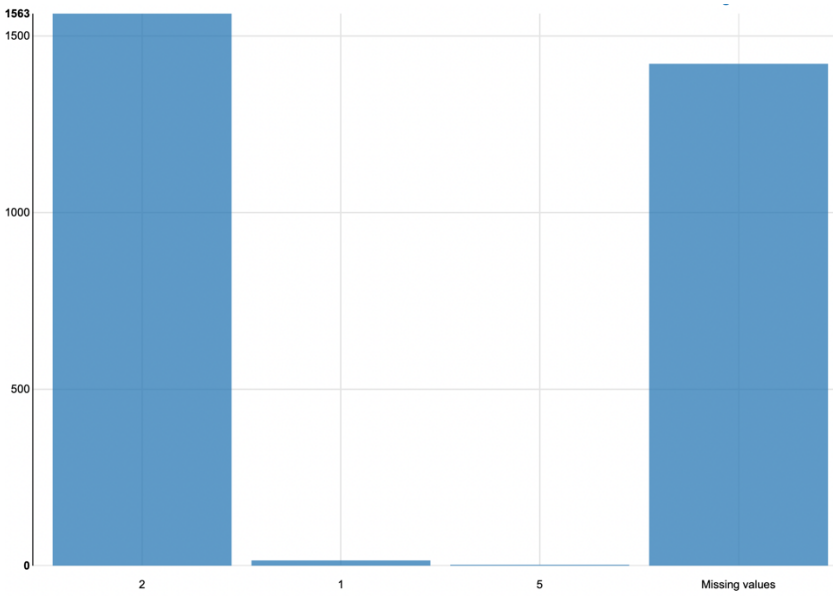


Figure 21 Bar Chart of Personal_info4

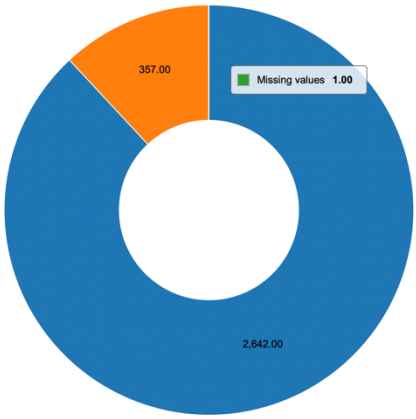
Personal_info5



Personal_info5 is assumed to be an nominal attribute type. Based on Figure 22, there are too many missing values to conclude what type of attribute it is.

Figure 22 Bar Chart of Personal_info5

Property_info1



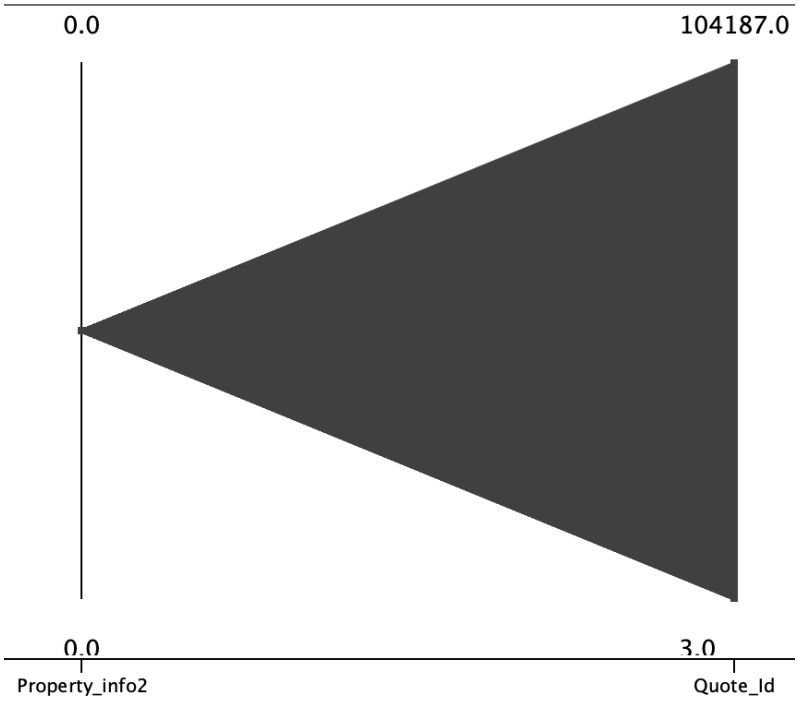
Property_info1 is a nominal attribute type as it has no order and there is no distance between 'Y' and 'N'. Based on Figure 23, it can be seen that 'N' attribute value has the most data points, although there is one missing value. Table 15 shows the summary of Property_info1 frequency and distribution.

Figure 23 Pie Chart of Property_info1

Property_info1	Proportion	Frequency
Y	0.88	2642
N	0.119	357
Missing Value	0.01	1
Total	1	3000

Table 15 Personal_info4 frequency and distribution

Property_info2



Property_info2 could be any attribute type as the data is too less to conclude. In this case, it is to be assumed as nominal attribute type. There is only one possible value in the attribute which makes it hard to determine the attribute type. Based on Figure 24, the possible value is only '0'.

Figure 24 Parallel Coordinates of Property_info2 and Quote_id

Property_info3

Property_info3 is a nominal attribute type since there is no ordering and it acts as labels. Figure 25 shows the distribution of Property_info3 attribute with O having the most data points, counted 875 and G has the lowest data points, counted 2.

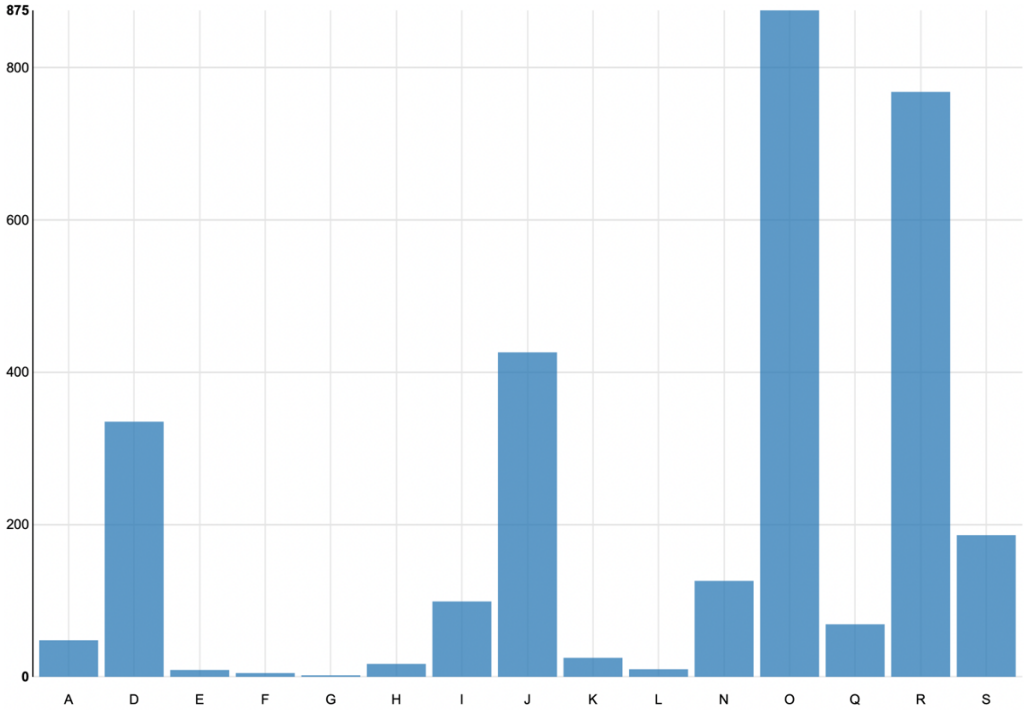


Figure 25 Bar Chart of Property_info3

Property_info4

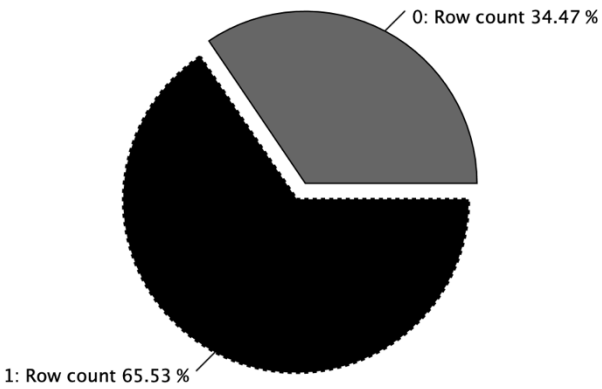


Figure 26 Pie Chart of Property_info4

Property_info4 is a nominal attribute because there is no ordering and based on Figure 26 it can be seen it only has two distinct values, such as '0' and '1' which can be represented as categories. The attribute is the highest on value '1' with 0.6553 proportion and a frequency of 1996. Table 1 shows the summary of the Property_info4 frequency and distribution.

Property_info4	Proportion	Frequency
0	0.3447	1034
1	0.6553	1966
Total	1	3000

Table 16 Property_info4 frequency and distribution

Property_info5

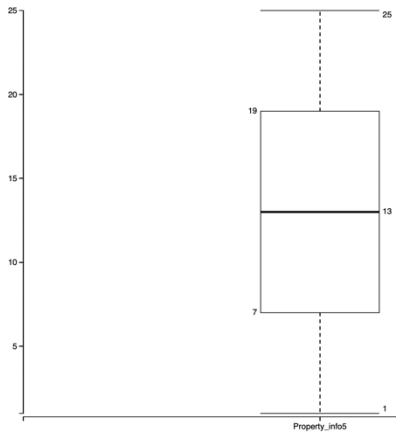


Figure 27 Box Plot of Field_info5

Property_info5 is an interval attribute type since it seems to be ordered with some records having higher Property_info5 value than others. In addition, all possible values in range [1,25] are present. Figure 27 shows the box plot of Property_info5 and Table 17 shows the summarising statistics of Property_info5.

Statistic	Value
Range (Min – Max)	1 – 25
Mean	13.0477
Q1	7
Median	13
Q3	19
Standard Deviation	7.2254
Variance	52.207

Table 17 Property_info5 statistics

Geographic_info1

Geographic_info1 is interval attribute type, based on Figure 28 it can be observed that the attribute values are distributed in range [1, 25]. Table 18 shows the statistics of Geographic_info1. Moreover, interestingly it has the same value sets as Property_info5. However, there seems to be no pattern found between these two attributes based on Figure 28.

Statistic	Value
Range (Min – Max)	1 – 25
Mean	7.398
Q1	2
Median	4
Q3	11
Standard Deviation	7.039
Variance	49.551

Table 18 Geographic_info1 statistics

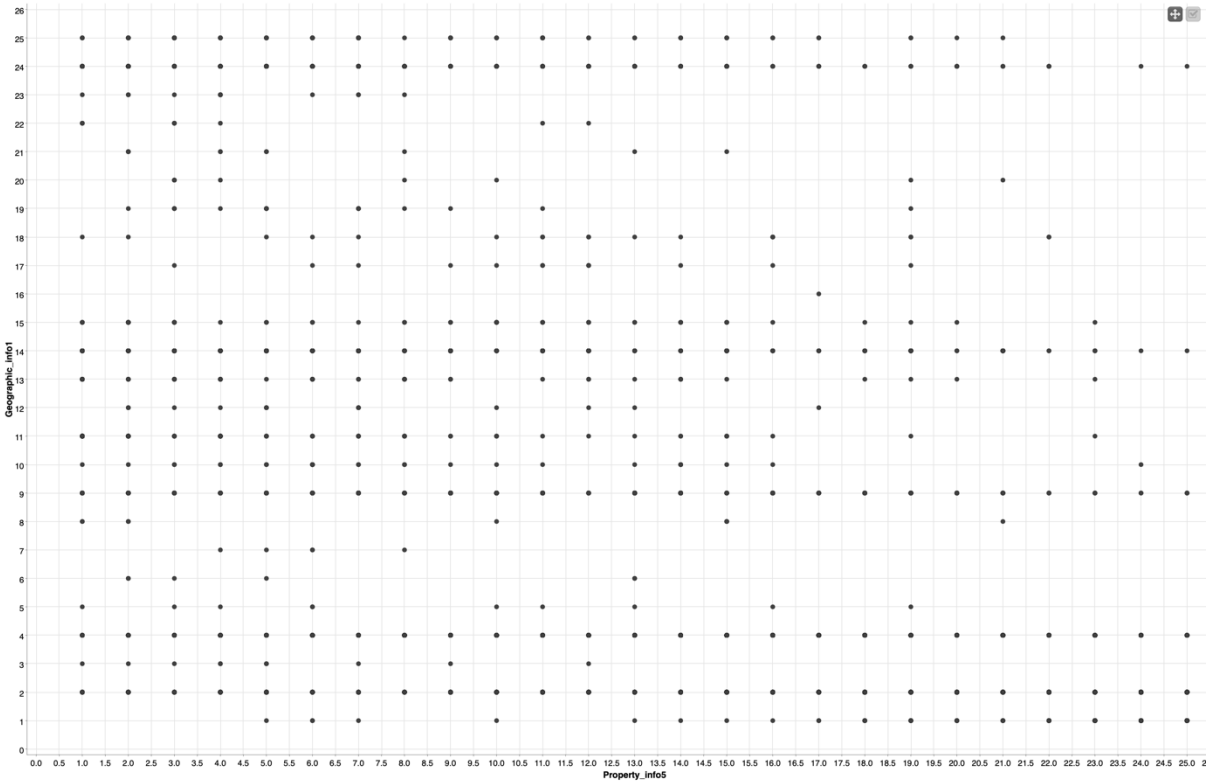


Figure 28 Scatter Plot of Geographic_info1 and Property_info5

Geographic_info2

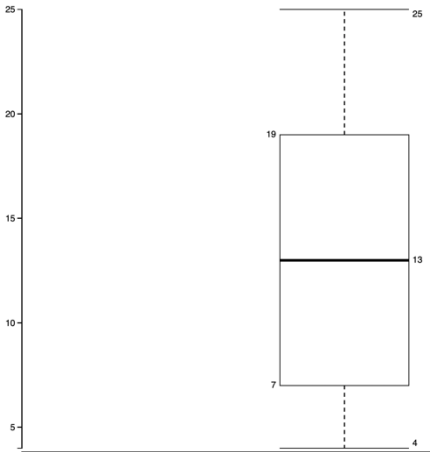


Figure 29 Box Plot of Geographic_info2

Geographic_info2 is an interval attribute type because it seems ordered and there are meaningful differences between two values. The interesting properties from this attribute is it has same median, Q1 and Q3 value with Property_info5. Figure 29 shows the distribution of the attribute. Moreover, Table 19 shows the statistics of Geographic_info2.

Statistic	Value
Range (Min – Max)	4 – 25
Mean	13.2977
Q1	7
Median	13
Q3	19
Standard Deviation	6.897
Variance	47.5649

Table 19 Geographic_info2 statistics

Geographic_info3

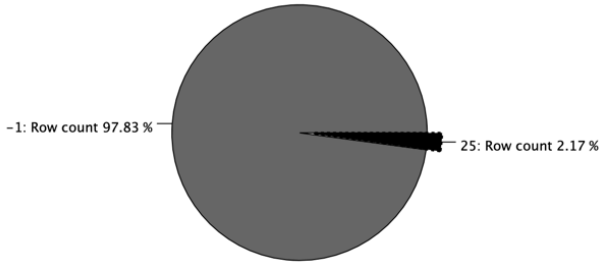


Figure 30 Pie Chart of Geographic_info3

Geographic_info3 is a nominal attribute type. Based on Figure 30, there is only two possible values which is '-1' and '25'. Therefore, it is reasonable to treat it as a category. In addition, an interesting property from this attribute is the two possible values are the min and max value of Coverage_info1 and Personal_info2 attribute.

Geographic_info4

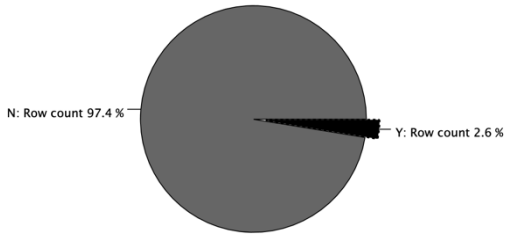


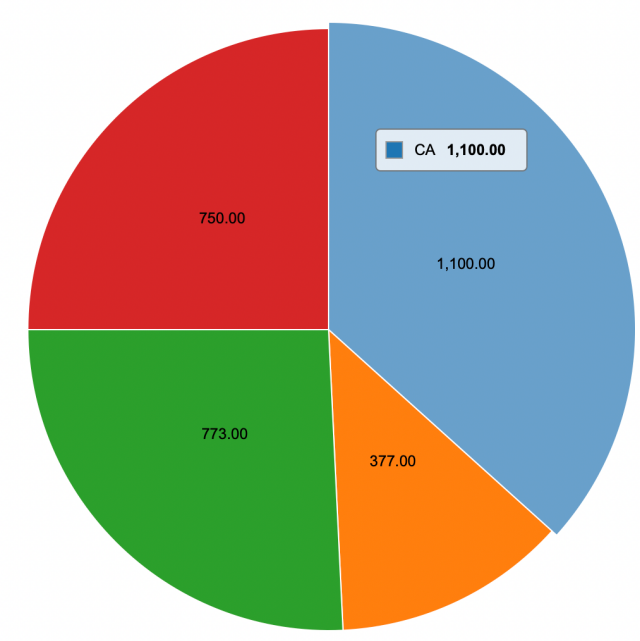
Figure 31 Pie Chart of Geographic_info4

Geographic_info4 is a nominal attribute type as it has no order and there is no distance between 'Y' and 'N'. Based on Figure 31, it can be seen that 'N' attribute value has the most data points with 2922 frequency and 0.974 distribution value. Table 20 shows the summary of Geographic_info4 frequency and distribution.

Property_info4	Proportion	Frequency
Y	0.026	78
N	0.974	2922
Total	1	3000

Table 20 Property_info4 frequency and distribution

Geographic_info5



Geographic_info5 is a nominal attribute type since there is no ordering and it seems like it represents state code in US which could be assumed used to categorise where each data points originates from. Figure 32 shows the distribution of Geographic_info5 attribute with 'CA' as the most common value with 1100 frequency and 0.3367 distribution value. Table 21 shows summary of Geographic_info5 frequency and distribution.

Figure 32 Pie Chart of Geographic_info5

Geographic_info5	Proportion	Frequency
CA	0.3667	1100
TX	0.25	750
NJ	0.2577	773
IL	0.1257	377
Total	1	3000

Table 21 Geographic_info5 frequency and distribution

1B: Data Pre-processing

The purpose of this section is to prepare the raw data to more understandable format. Multiple pre-processing techniques are used which includes binning, normalisation, discretisation and binarization.

Binning

This section applies binning techniques, namely equi-width binning and equi-depth binning to Property_info5 attribute.

Equi-width Binning

To perform equi-width binning of Property_info5 attribute, these are the number of steps to follow:

1. The data is sorted by Property_info5 attribute to get the minimum and maximum value of the attribute, the minimum is 1 and the maximum is 25.
2. Calculate range from minimum and maximum value which resulted in 24, using the formula:

$$Range(X) = Max(X) - Min(X)$$

The minimum and maximum value does not need to be adjusted (lowered or raised) as Property_info5 maximum value is same with Coverage_info1 and

Personal_info2 and the minimum value is also close, it is reasonable enough that these values does not need to be adjusted further.

3. The number of bins is 5, which is calculated using the formula:

$$Number\ of\ Bins = \frac{Range(X)}{bin\ width}$$

The reason that there are 5 bins with 5 values for each bin are because this is the ideal bin size as it is not too wide which could hide important details about distribution and also not too narrow which cause a lot of noise in the data. In addition, there is not much spread and variance produced by the histogram in KNIME as shown in Figure 33. Besides, all values from the range 1 – 25 are present, so it is safe to suggest there are 5 bins.

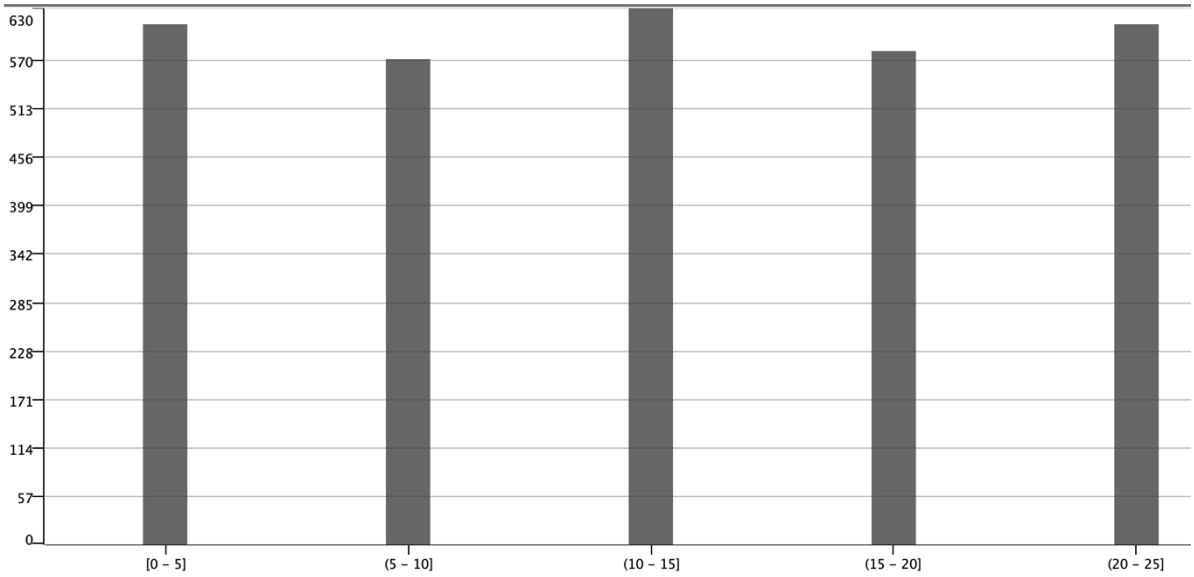


Figure 33 Histogram of Property_info5

4. The bin boundaries are calculated with KNIME Interactive Histogram node as seen in Table 22. As seen in the table, Bin 1 has 6 values rather than 5 values, this could be assumed as 5 values with [1,5] boundary as Property_info5 does not have any '0' value.

Equi-width	Bin boundary
Bin 1	[0,5]
Bin 2	(5,10]
Bin 3	(10,15]
Bin 4	(15,20]
Bin 5	(20,25]

Table 22 Equi-width of Property_info5

5. Apply a formula to the spreadsheet based on the bin boundaries found with KNIME using such chained IF statement formula:

```
=IF(AND(Y2>=0,Y2<=5),"Bin 1",IF(AND(Y2>5,Y2<=10),"Bin 2",IF(AND(Y2>10,Y2<=15),"Bin 3",IF(AND(Y2>15,Y2<=20),"Bin 4", IF(AND(Y2>20,Y2<=25),"Bin 5")))))
```

This formula essentially means that if the attribute value in the attribute sets is more than or equal to 0 and less than or equal to 5, it belongs to Bin 1 and so on.

6. Property_info5 values will be categorised on its bin according to the bin boundary as reflected in the excel spreadsheet.

Equi-depth Binning

To perform equi-depth binning of Property_info5 attribute, these are the number of steps to follow:

1. Calculate the number of records. Based on the excel files it is identified that it has 3000 records.
2. Based on the number of bins decided, the size of the bin is exactly 600. Although, equal frequency might not be possible due to repeated values.
3. Using KNIME, the values are assigned to the bins as shown in Table 23.

Equi-depth	Bin boundary
Bin 1	[0,5]
Bin 2	(5,10]
Bin 3	(10,15]
Bin 4	(15,20]
Bin 5	(20,25]

Table 23 Equi-depth of Property_info5

4. Apply a formula to the spreadsheet based on the bin boundaries found with KNIME Auto-Binner node using such chained IF statement formula:

   This formula essentially assigns each attribute value to the bin numbers based on the bin boundaries.

Normalisation

This section applies normalisation techniques, namely min – max normalisation and z-score normalisation to Sales_info5 attribute.

Min-max normalization

The Sales_info5 attribute range are 14 – 67114, the formula used to normalise the data with Min – max normalisation is:

$$X' = \frac{X - Min(X)}{Max(X) - Min(X)} \left(newMax(X) - newMin(X)\right) + newMin(X)$$

Where X is the original value, X' is the normalised value and $newMax(X)$ is 1 and $newMin(X)$ is 0. The formula is applied automatically with KNIME Normalizer (PPML) node and compared with the application to the spreadsheet. To get the same value with KNIME, apply the following formula to the spreadsheet. The result is then reflected in the excel spreadsheet.

```
=(O2-MIN($O$2:$O$3001))/(MAX($O$2:$O$3001)-MIN($O$2:$O$3001))
```

The formula above essentially means that get the whole Sales_info5 attribute values min and max. In this case, all the attribute values are represented as $O$2:$O$3001 and apply the min – max normalisation formula.

Z-score normalization

The formula used to normalise the data with Z-score normalisation is:

$$X' = \frac{X - Mean}{Standard\ Deviation}$$

Where X is the original value, and X' is the standard score which is used to show how far the mean of a data point is and measure how many standard deviations below or above the mean. The formula is applied automatically with KNIME Normaliser (PPML) node and compared with the application to the spreadsheet. To get the same value with KNIME, the mean and standard deviation of Sales_info5 attribute are calculated. Then, apply the following formula to get the Z-score normalization value. The result is then reflected in the excel spreadsheet.

```
=(O2-$AM$1)/$AM$2
```

The formula above essentially means that get the Mean from Sales_info5 attribute with the formula =AVERAGE(O2:O3001) and put it in $AM$1 column. Then, calculate the Standard Deviation with the formula =STDEV(O2:O3001) where O2:O3001 are the all attribute values of Sales_info5. Finally, apply the Z-score normalization formula.

Discretisation

The Coverage_Info1 attribute in the data set are discretised into the following categories, namely Basic, Low, Medium and High. The division of the attributes are done in KNIME with Histogram node as shown in Table 24.

Category	Range	Frequency
Basic	[-1 – 6]	1170
Low	(6 – 13]	1246
Medium	(13 – 20]	395
High	(20 – 27]	189

Table 24 Discretisation of Coverage_info1

To reflect the values in the spreadsheet, apply this excel formula based on the attribute range which is found with KNIME using such chained IF statement formula:

```
=IF(AND(H2>=-1,H2<=6),"Basic",IF(AND(H2>6,H2<=13),"Low",IF(AND(H2>13,H2<=20),"Medium",IF(AND(H2>20,H2<=27),"High"))))
```

This formula essentially means that if the attribute value in the attribute sets is more than or equal to -1 and less than or equal to 6, it belongs to “Basic” category and so on.

Binarisation

The Geographic_info5 attribute is binarized with four categories, CA, TX, NJ and IL. Each of the Geographic_info5 value have their own binarization column which is reflected in the excel spreadsheet. The formula is to apply IF function in excel for each category as shown below.

```
=IF(AD2="CA",1,0)
```

This formula means that if the attribute value is CA, return 1 if others, return 0. Then, this is applied for all categories that are present in Geographic_info5 attribute.

1C: Summary

In summary, the date range where it has most quotes are 05/08/2013 to 30/04/2015. In addition, it was found that Field_info1 distribution tend to reside near a specific Field_info3 attribute values. There are also some associations found that should be investigated more, namely the Personal_info2 attribute has same value range as Coverage_info1 attribute which is [-1, 25] but both attributes have one significant difference, Coverage_info1 attribute lowest frequency is at ‘-1’ attribute but Personal_info2 has ‘-1’ as the highest frequency in the dataset. However, there is no pattern found between these attributes. Then, Geographic_info1 and Property_info5 attribute also has same value range which is [1, 25] but there is also no pattern found between these two attributes. Moreover, it is suggested that 36.67% quotes originate from California (CA) based on Geographic_info5.

Currently, there is not much conclusion can be drawn due to the vagueness of the attribute names. It is highly recommended to re-evaluate the attributes provided in the dataset to provide better understanding and utilise the dataset fully.