Data Exploration and Preparation

Assignment Task 2 – Introduction to Data Analytics

Ahmed Khursheed - 13477878

Table of Contents:

1. Identifying Attribute type

In the given data set, there are 30 attributes of different types. To work with the data set, we must try to first understand what each attribute represents, its values and how it may be of use as we proceed towards gaining insight on the data.

Name	Type	Reason
Quote_ID	Nominal	The values in this column are customer quote ID numbers which can not be ordered in any way; therefore, it is a nominal attribute.
Quote_Date	Interval	The values are dates which can be measured from one to the other, therefore it is an interval.
Quote_Flag	Nominal	The values for this attribute are 1 and 0; a binary type of attribute amongst which we can not measure or order, thus it is a nominal.
Field_info1	Ordinal	F, J, B, K, E, and C are the occurrences of this
Field_info2	Ratio	All the values seem to be measured from a fixed 0 point and multiplication and division can be performed on the values as well.
Field_info3	Nominal/Ordinal	The values consist of positive integers which seem to be repeated quite often, most likely denoting some category or a code to a type of information which can be grouped. If the numbers have some ranking system, we can classify it as an ordinal attribute. Therefore, it is a nominal/ordinal attribute.
Field_info4	Nominal	The values of Y and N are most likely labels, which is a nominal attribute. It likely represents Yes and No.
Coverage_info1	Interval	The values are numeric but since there is a -1 present, it cannot be a ratio since division will not be possible.
Coverage_info2	Interval	The values are numeric, although in this data set, this attribute does not contain a negative value, the ranges for coverage_info1 and coverage_info2 are similar and I think they represent the same sort of data. Therefore, I suggest that this is an interval too.
Coverage_info3	Ordinal	The values consist of alphabetical symbols which might be labels and could mean something. Therefore, we could order them, and it is ordinal.
Sales_info1	Nominal	The values seem to be of binary type i.e., 1 and 0. It is only providing us information to differ one from the other.

Sales_info2	Ratio	The values here are real numbers and range from 1 to 5. We can see a lot of repetition in the values and therefore it seems like a range from 1-5 scale. It is of ratio type since we can divide and multiply and there are no negative values.
Sales_info3	Ratio	There are no negative values within the dataset, and range is from 1-24. So, we can say that these values are measures from a 0 point. Therefore, this is a ratio.
Sales_info4	Ordinal	The values consist of alphabetical symbols which might be labels and could mean something. Therefore, we could order them, and it is ordinal.
Sales_info5	Ratio	The values provide a wide range of real numbers, most likely linked to purchases made by consumers. With no negative values, the data attribute is Ratio.
Personal_info1	Nominal	The values provided mostly likely represents "No" and “Yes” type of a flag, however we see that N is the value for all the customers.
Personal_info2	Interval	The value is provided within a small data pool, with both negative and positive values; the starting point is not zero. Thus, the attribute is Interval.
Personal_info3	Ordinal	The values here provide a few different alphabetical entries as symbols. The values could be grouped and ordered by their corresponding alphabet sequence.
Personal_info4	Nominal	The only value here is 0, and it could be something that all the customers have in common.
Personal_info5	Not identifiable	The only value here we see is 2 and a lot of customers do not have this entered, i.e., the values are missing and there is not enough data to decide what this is.
Property_info1	Nominal	The values of Y and N are most likely labels, which is a nominal attribute. It likely represents Yes and No.
Property_info2	Nominal	The only value here is 0, and it could be something that all the properties have in common.
Property_info3	Ordinal	The value provided showcases alphabetic as symbols. The values of property_info3 are classified, so its values are not known. Instead, the sequence of alphabets demonstrates the importance of each data ranked, possibly in alphabetical order. It’s of ordinal type.
Property_info4	Nominal	The values seem to be of binary type i.e., 1 and 0. It is only providing us information to differ one from the other. It is of nominal type.

Property_info5	Interval	The values provide a small range of real numbers, most likely linked to property values or ownership. With -1 being a negative value present, the data attribute is interval.
Geographic_info1	Ratio	The values provide a small range of real numbers. With no negative values and possessing the ability to be multiplied, divided, added, and subtracted, it is a ratio.
Geographic_info2	Ratio	The values provide a small range of real numbers. With no negative values and possessing the ability to be multiplied, divided, added, and subtracted, it is a ratio.
Geographic_info3	Nominal	The values provide only 2 numbers, most likely linked to geographical info. With -1 being a negative value present and 25 the other value, the data attribute is nominal.
Geographic_info4	Nominal	The values of Y and N are most likely labels, which is a nominal attribute. It likely represents Yes and No.
Geographic_info5	Nominal	The values given here are alphabetic symbols which I think might be for different states in USA or any other geographic relevant geographical information according to which the customers can be grouped.

2. Data Summary

Here is a basic summary of the statistics of the attributes in the dataset.

Attribute	Minimum	Maximum	Median	Mean	Standard Deviation	Variance (Spread between data)
Quote_Id	28	104211	52531.5	52658.75	29838.54	890338268.71
Quote_Date	01/01/13	12/05/15	-	-	-	-
Quote_Flag	0	1	0	0.192	0.394	0.155
Field_info1	-	-	-	-	-	-
Field_info2	0.875	1.01	0.94	0.938	0.038	0.001
Field_info3	548	1487	965	956.286	288.377	83161.505
Field_info4	-	-	-	-	-	-
Coverage_info1	-1	25	8	8.899	5.516	30.429
Coverage_info2	1	25	22	21.129	5.002	25.02
Coverage_info3	-	-	-	-	-	-
Sales_info1	0	1	1	0.736	0.441	0.195
Sales_info2	2	5	5	4.222	0.98	0.961
Sales_info3	1	24	11	14.092	6.276	39.384
Sales_info4	-	-	-	-	-	-
Sales_info5	24	67155	33031.5	33303.473	19488.751	379811430.313
Personal_info1	-	-	-	-	-	-
Personal_info2	-1	25	6	6.895	6.647	44.185
Personal_info3	-	-	-	-	-	-
Personal_info4	0	1	0	0	0.018	0
Personal_info5	1	2	2	1.995	0.071	0.005
Property_info1	-	-	-	-	-	-
Property_info2	0	0	0	0	0	0
Property_info3	-	-	-	-	-	-
Property_info4	0	1	1	0.67	0.47	0.221
Property_info5	-1	25	13	12.828	7.299	53.277
Geographic_info1	1	25	4	7.415	7.062	49.871
Geographic_info2	4	25	13	13.092	6.884	47.391
Geographic_info3	-1	25	-1	-0.402	3.898	15.195
Geographic_info4	-	-	-	-	-	-
Geographic_info5	-	-	-	-	-	-

3. Exploring Attributes

The descriptive statistics of the attributes are given in section 2. In this section, the emphasis will on visualization of attributes to gain some meaningful information from the raw data.

3.1. Quote_Id

Since quote id is a nominal attribute which refers to the customer, they are all unique values, no such computation and analytics can be performed on it which might yield useful insight into the data.

3.2. Quote_Date

Since the values are of date type, we cannot visualize the data but some basic information like the range of date could be found which is 01/01/2013 – 12/05/2015. It is an interval type attribute since we can calculate the differences between dates. However, here is something below that could be done with the given data.

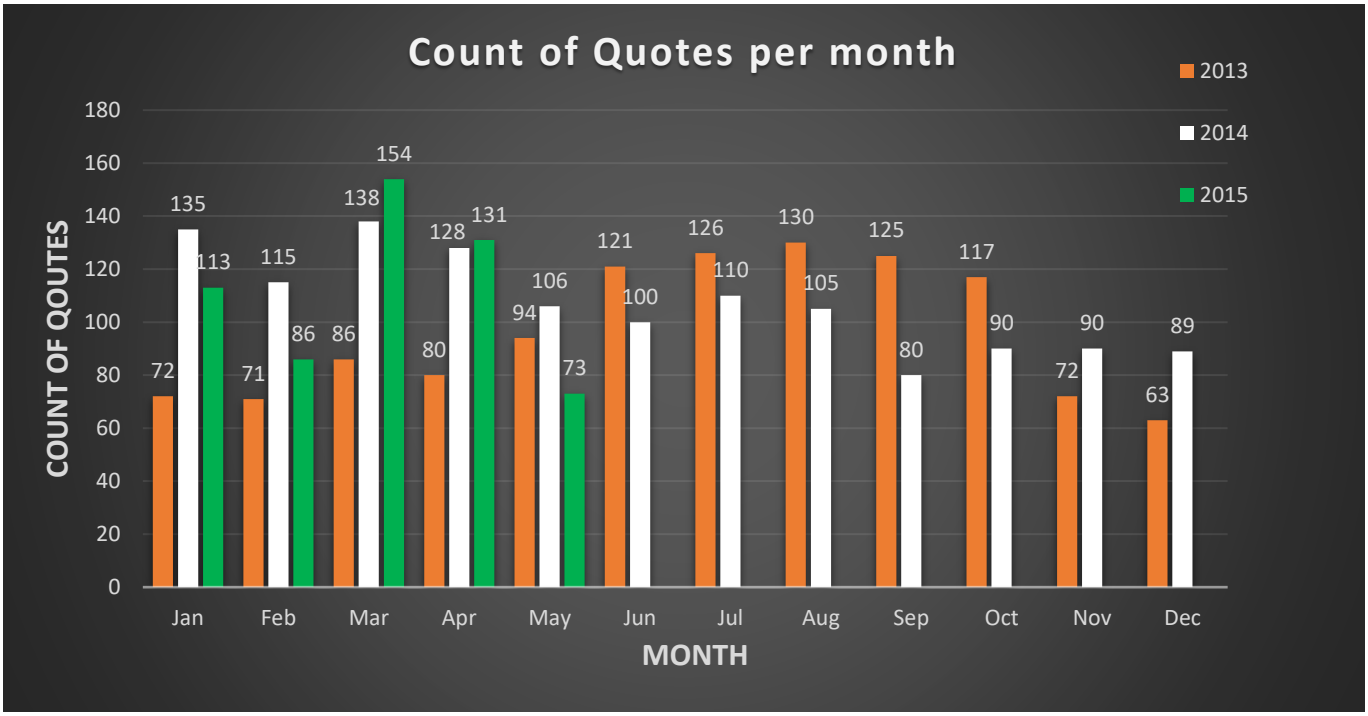


FIGURE 1. NUMBER OF QUOTES GIVEN EVERY MONTH

3.3. Quote_Flag

Quote Flag is an ordinal data type, we can only differ one from the other and upon exploring the data we find that only 19.23% of people bought the insurance policy while 80.77% people only took the quote and did not purchase the policy as seen in the figure below.

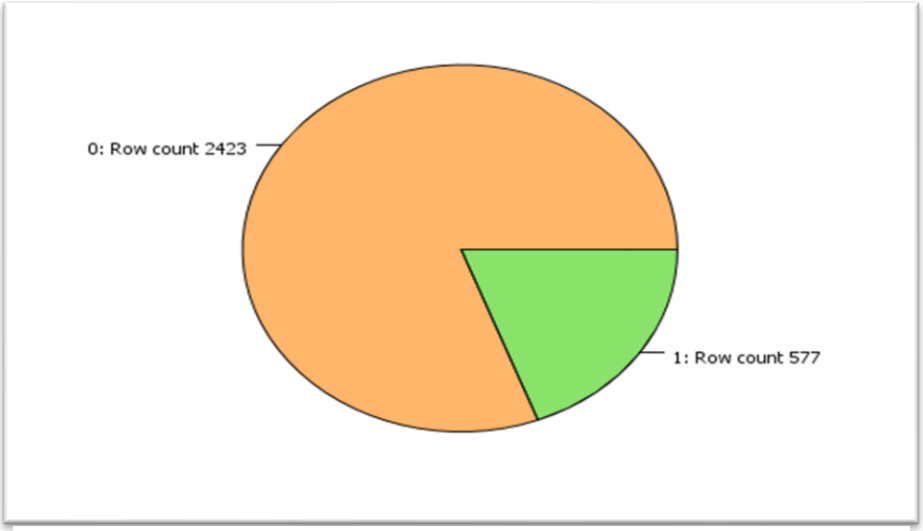


FIGURE 2. COUNT OF POLICIES BOUGHT(QUOTE_FLAG)

3.4. Field_info1

This attribute as we can see below can be represented in this form, if the letters denote some form of ranking system, we can see that it is an ordinal attribute, if the letters are random denoting just a category but are not comparable, it is a nominal attribute. We can see that categories: A, C, D, E and K were the least chosen by people.

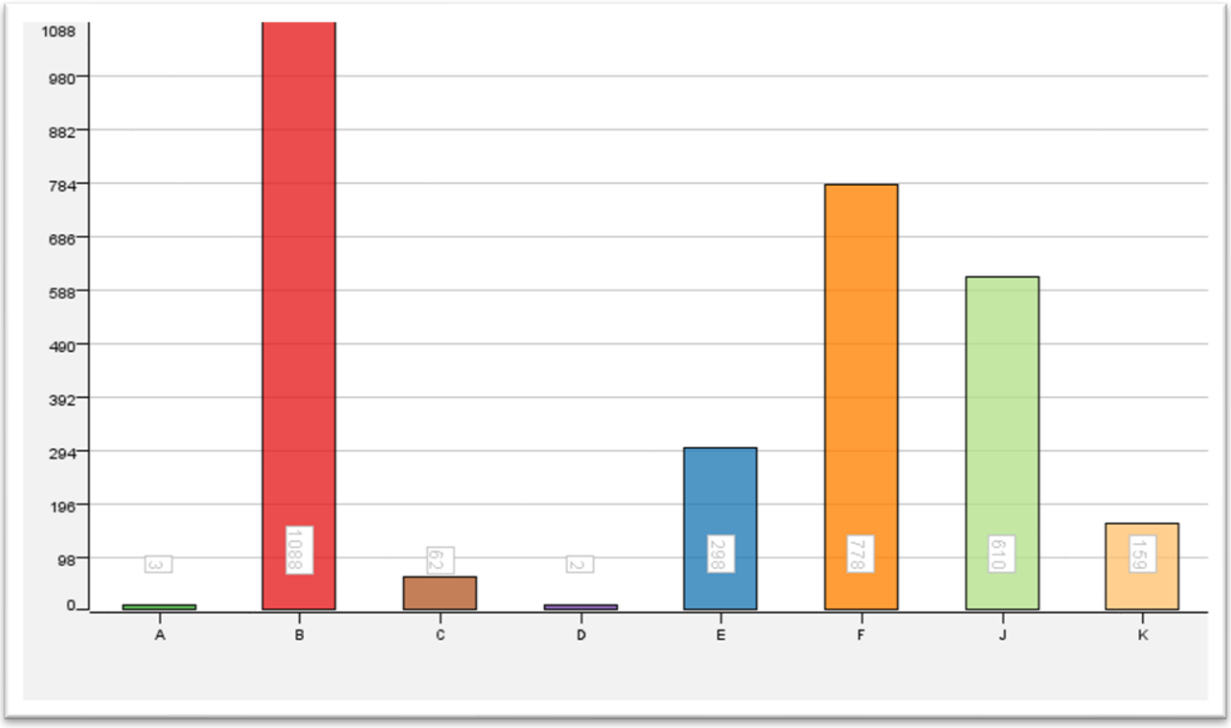


FIGURE 3. HISTOGRAM FOR ATTRIBUTE FIELD_INFO1

3.5. Field_info2

Attribute field_info2 looks like a normalized column around ±1. As we can see below in figure 4 the minimum value is 0.87 and maximum is 1.01. The median is 0.94 and the lower/25th percentile is 0.92 and upper/75th percentile is 0.97.

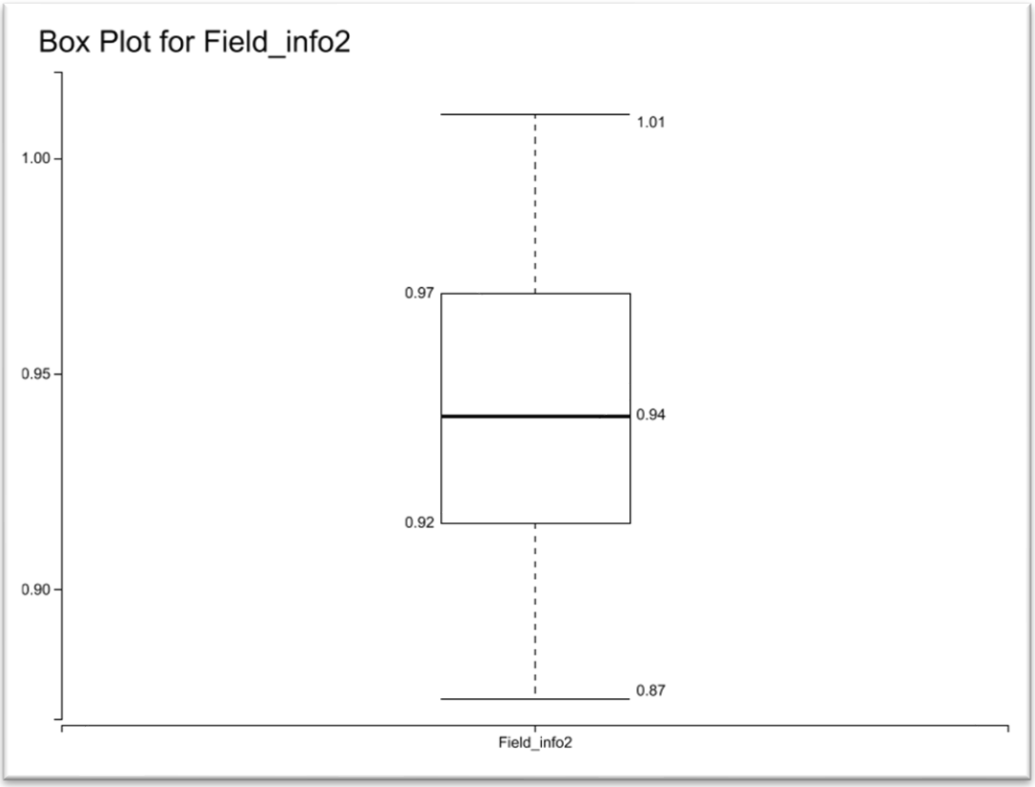


FIGURE 4. BOX PLOT FOR FIELD_INFO2

3.6. Field_info3

For this attribute we see that there are 8 values which are repeated quite often, most likely values that might denote a code for something related to field_info3. We see that 935 is the most common category followed by 1113 and 965. 1480 is the least common category.

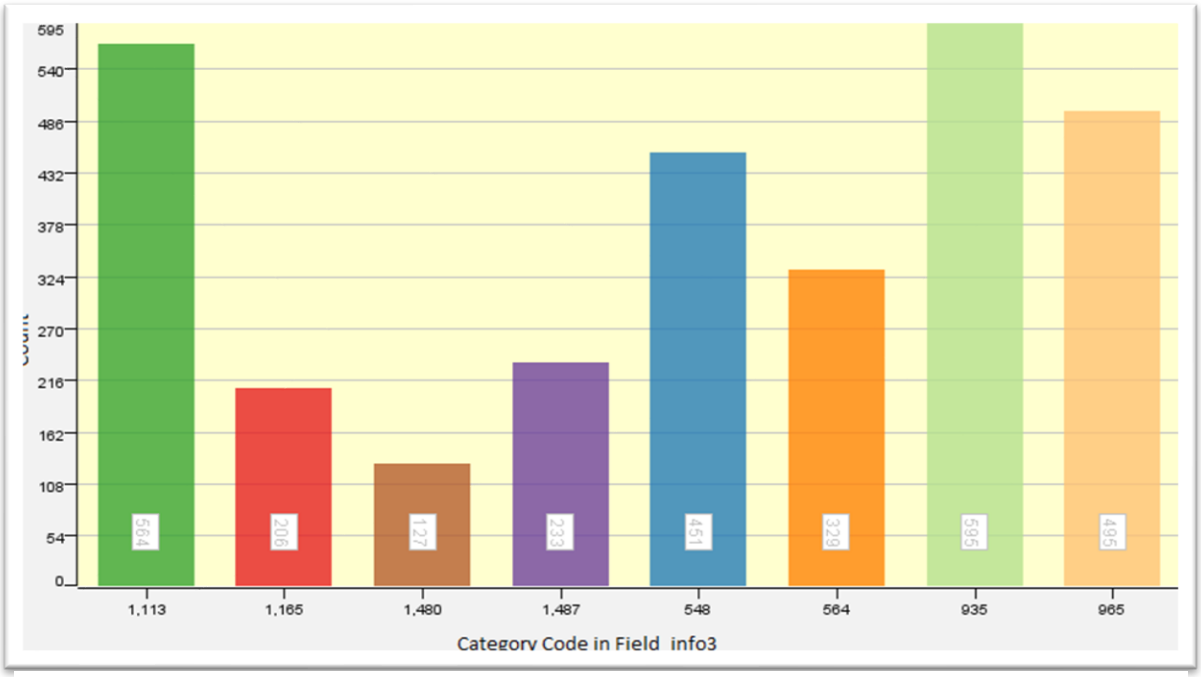
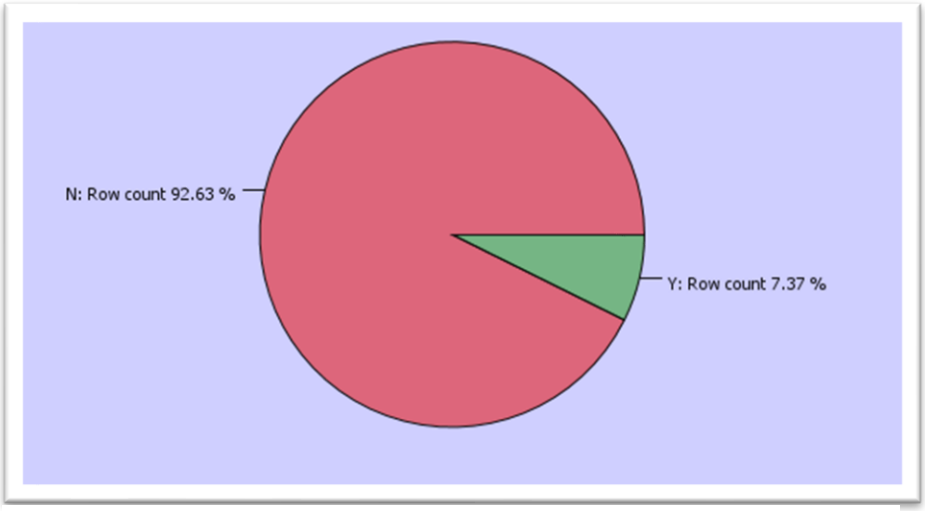


FIGURE 5. HISTOGRAM FOR FIELD_INFO3

3.7. Field_info4



Field_info4 consists of a Yes/No flag with the count of 2779 for No which is 92.63% and 221 for Yes which is 7.37%.

FIGURE 6. PIE CHART FOR FIELD_INFO4

3.8. Coverage_info1

Here the attribute coverage_info1 has been visualized using a box plot. As we can see below in figure 7 the minimum value is -1 and maximum is 25. The median is 8 and the lower/25th percentile is 5 and upper/75th percentile is 11. We can also see that there are 5 outliers: 21, 22, 23, 24 and 25.

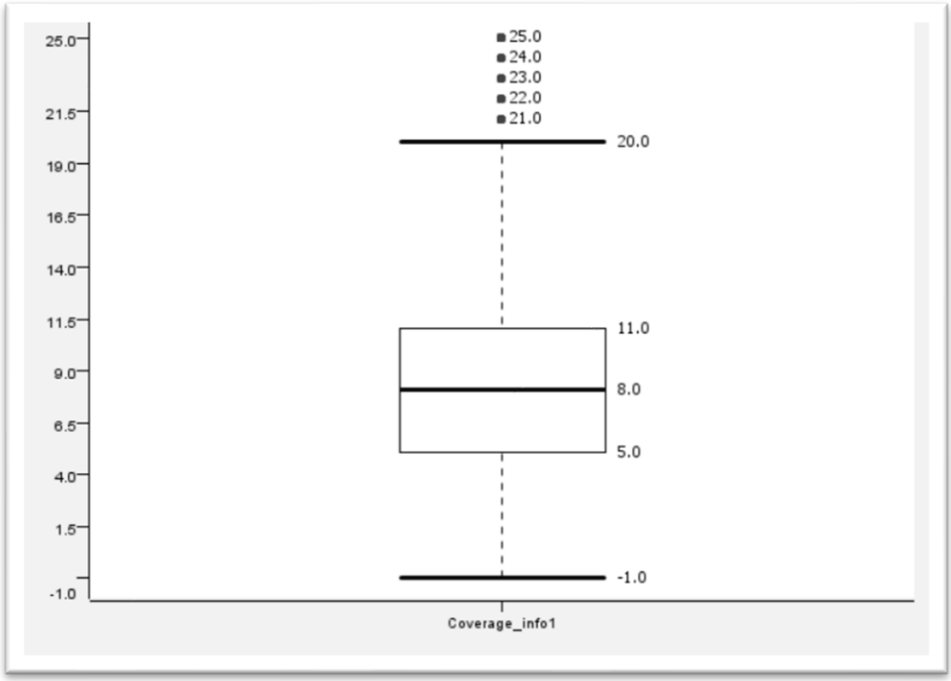


FIGURE 7. BOX PLOT FOR COVERAGE_INFO1

3.9. Coverage_info2

Coverage_info2 has been visualized below in figure 8 using a pie chart since the values were too far and distinct despite the range being from 1~25, the distribution was not even. Therefore, it could not be visualized any other way. We see that 22 is the prominent value with the most occurrences followed by 25 and 2. However, 1 only has 1 occurrence.

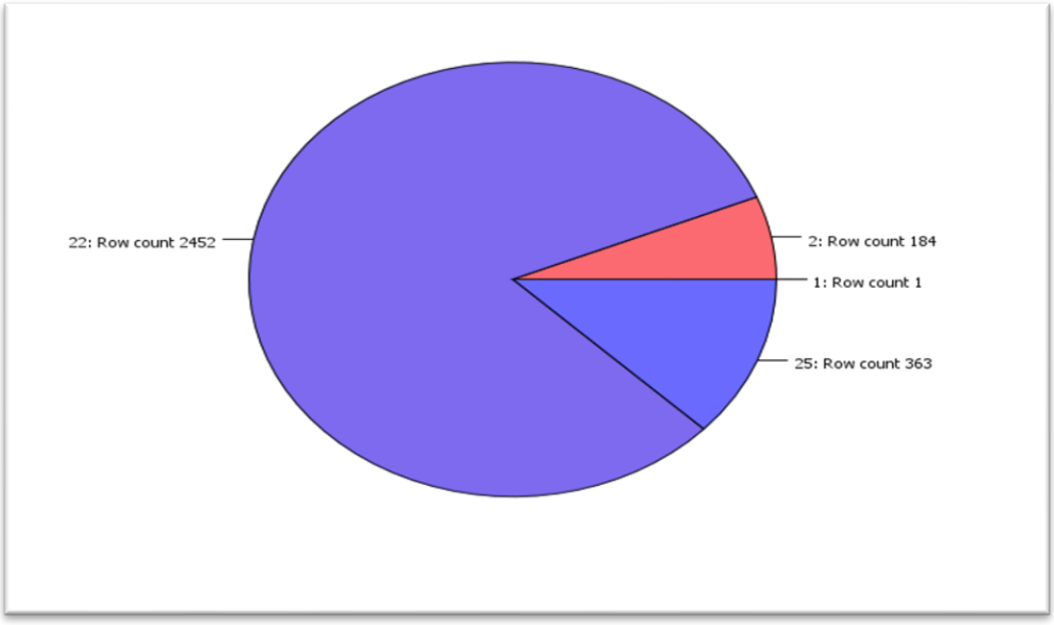


FIGURE 8. PIE CHART FOR COVERAGE_INFO2

3.10. Coverage_info3

Coverage_info3 has a categorical spread with alphabets denoting some order. It has been visualized below in figure 9 using a histogram.

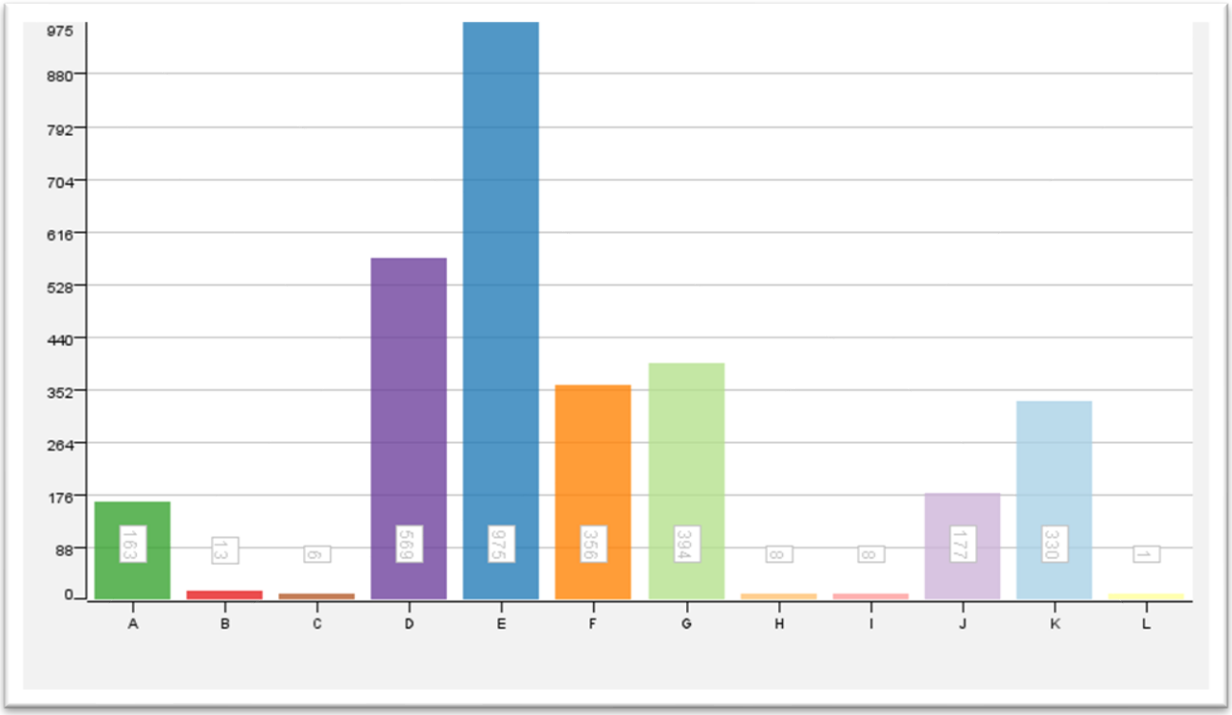


FIGURE 9. HISTOGRAM FOR COVERAGE_INFO3

3.11. Sales_info1

The attribute sales_info1 is a binary attribute with 793 (26.43%) and 2207 (73.57%) occurrences of 0 and 1, respectively. It has been visualised below in figure 10.

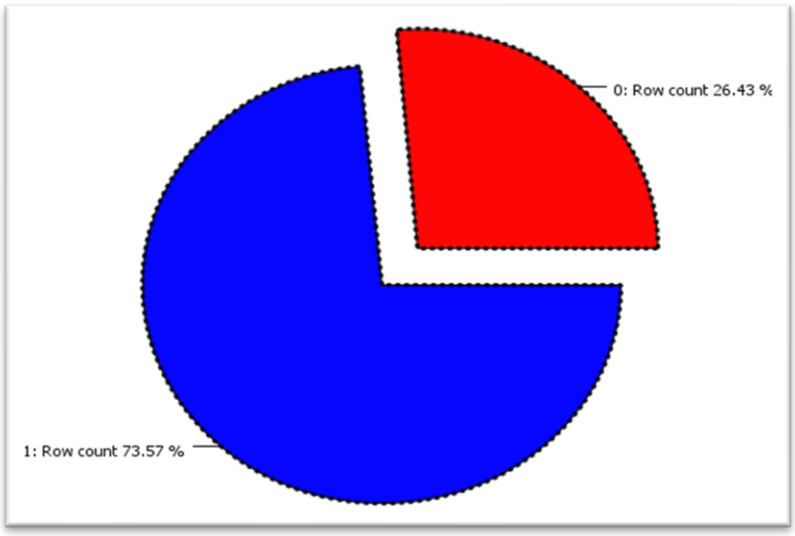


FIGURE 10. PIE CHART FOR SALES_INFO1

3.12. Sales_info2

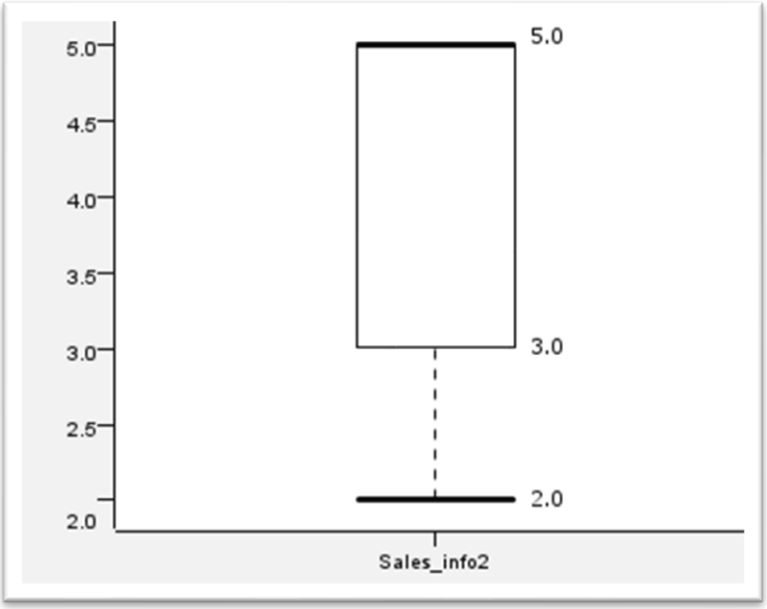


FIGURE 11. BOX PLOT FOR SALES_INFO2

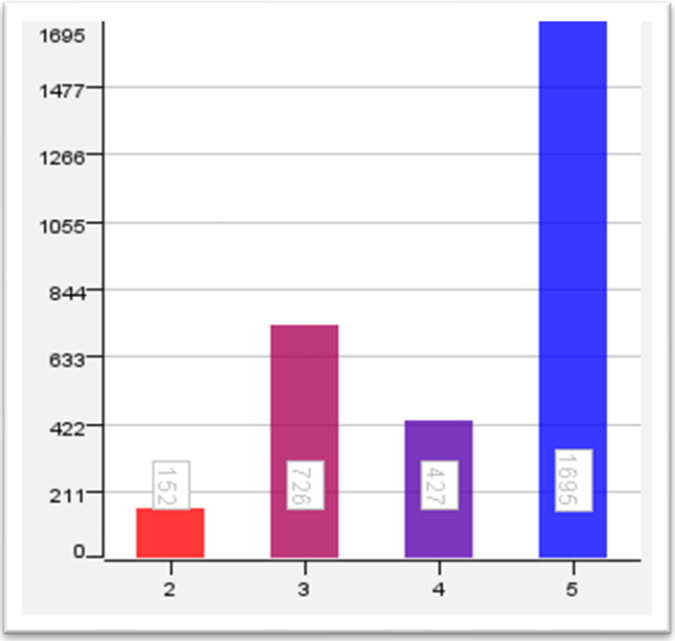


FIGURE 12. HISTOGRAM FOR SALES_INFO2

Sales_info2 has been visualised using a box plot and histogram (figure 11 & 12 respectively). The minimum value is 2 and maximum is 5. The median is 5 and the lower/25th percentile is 3 and upper/75th percentile is 5.

3.13. Sales_info3

For the attribute sales_info3 in figures 13 and 14 below we see that 11 has 1038 occurrences i.e., almost 1/3rd of the dataset. As we can see below in figure 13 and 14 the minimum value is 1 and maximum is 24. The median is 11 and the lower/25th percentile is 11 and upper/75th percentile is 20.

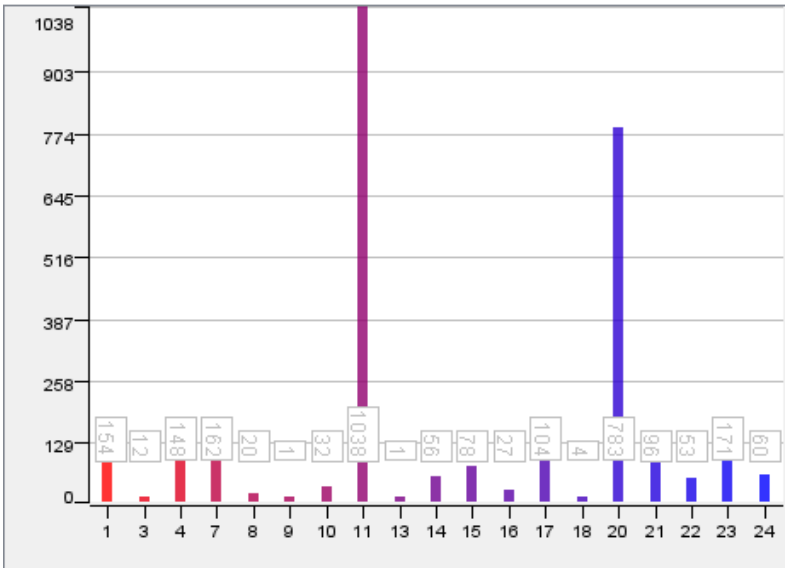


FIGURE 13. HISTOGRAM FOR SALES_INFO3

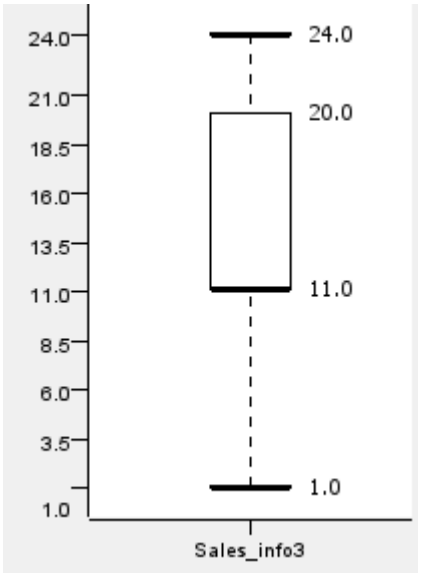


FIGURE 14. BOX PLOT FOR SALES_INFO3

3.14. Sales_info4

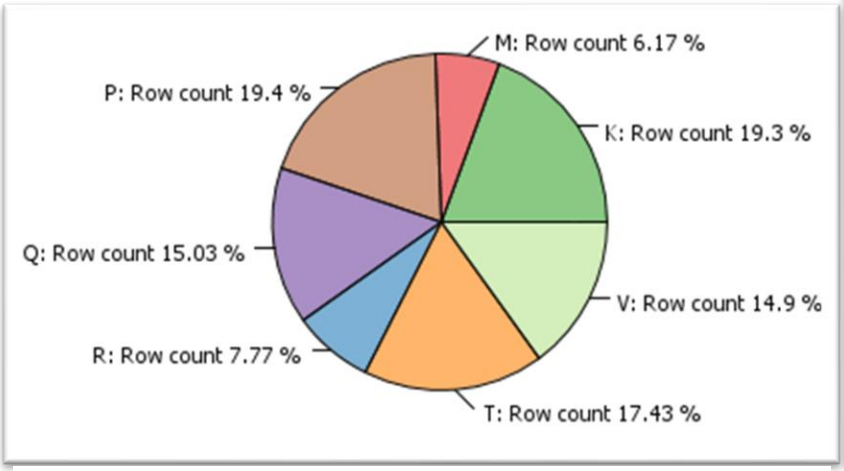


FIGURE 15. PIE CHART FOR SALES_INFO4

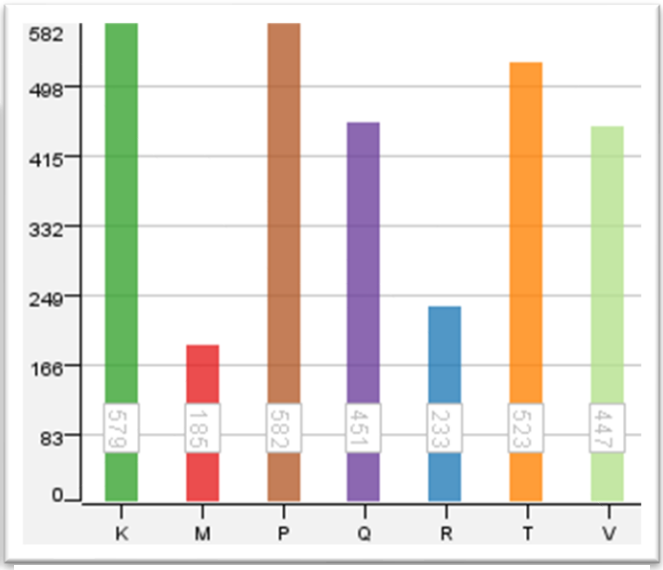


FIGURE 16. HISTOGRAM FOR SALES_INFO4

As we can see in the figures above, sales_info4 denotes some alphabetical category which can be ordered alphabetically. The distribution is almost even amongst all categories except for M and R.

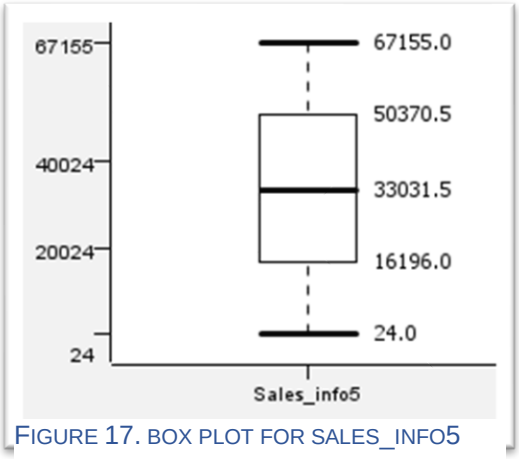


FIGURE 17. BOX PLOT FOR SALES_INFO5

3.15. Sales_info5

This attribute consists of positive integers. As we can see in figure 17 the minimum value is 24 and maximum is 67155. The median is 33031.5 and the lower/25th percentile is 16196 and upper/75th percentile is 50370.5. We can also see that there are no outliers and the distribution is almost even.

3.16. Personal_info1

Personal_info1 is a binary nominal type of attribute in the form of Yes/No. There are only 14 occurrences of Yes, meanwhile 2986 entries of No.

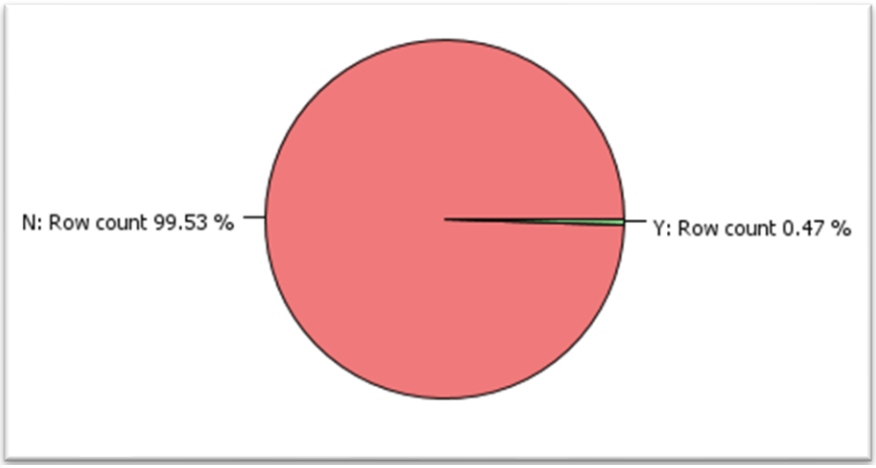


FIGURE 18. PIE CHART FOR PERSONAL_INFO1

3.17. Personal_info2

The data for the attribute personal_info2 has a lot of variance. As we can see on the right in figure 19 the minimum value is -1 and maximum is 25. The median is 6 and the lower/25th percentile is 4 and upper/75th percentile is 8.5. We can also see that there are some outliers: 16, 17, 18, 19, 20, 21, 22, 23, 24 and 25.

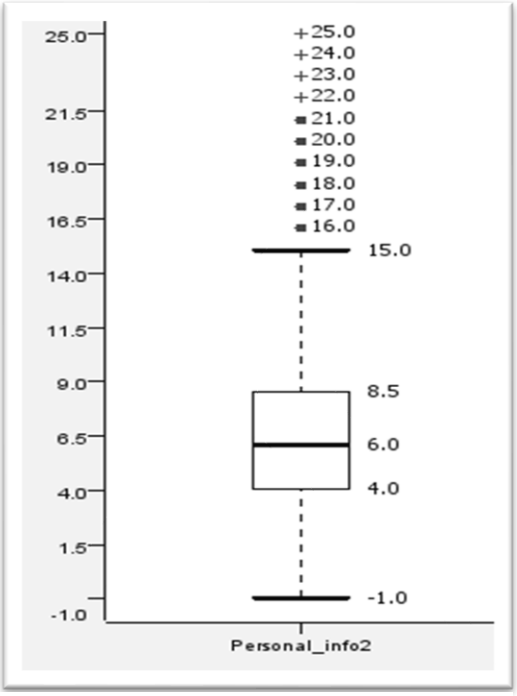


FIGURE 19. BOX PLOT FOR PERSONAL_INFO2

3.18. Personal_info3

As we can see in figure 20 below, there are many alphabetical categories for personal_info3. The values seem to follow an alphabetical pattern with the first letter being X, Y, Z and a second letter following it in alphabetical order. Category ZA has 1416 occurrences which is 47.2% of the dataset.

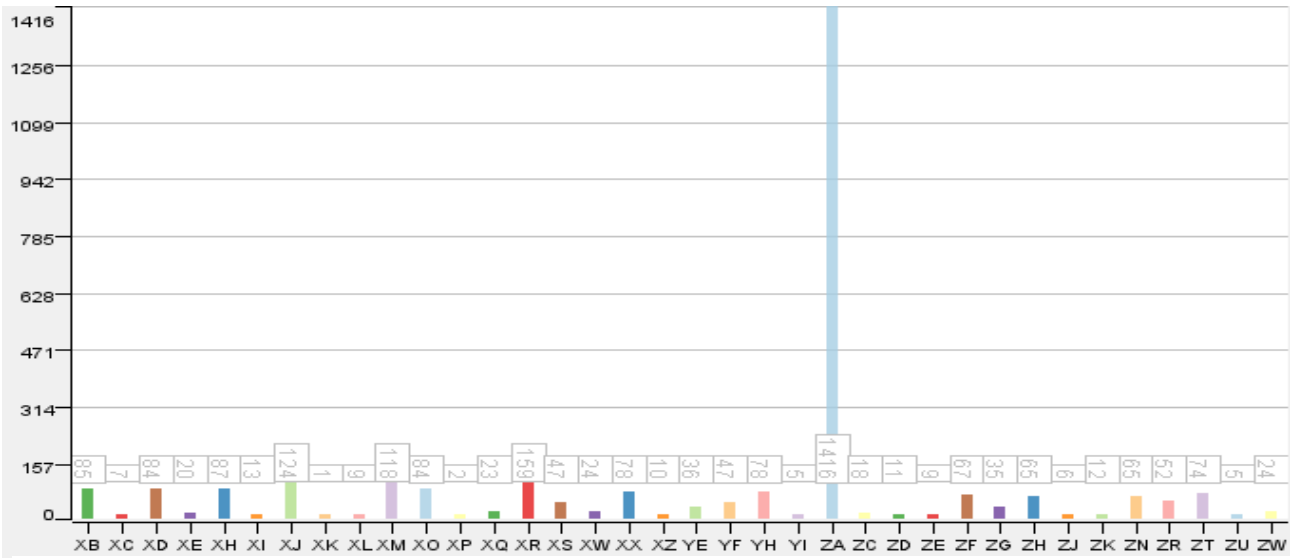


FIGURE 20. HISTOGRAM FOR PERSONAL_INFO3

3.19. Personal_info4

The attribute personal_info4 is a binary nominal attribute with 2999 (99.97%) and 1 (0.033%) occurrence(s) of 0 and 1, respectively. It has been visualised below in figure 21.

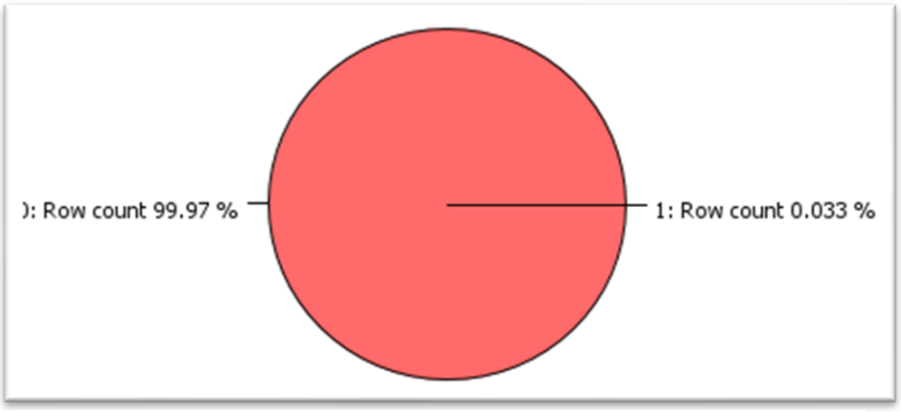


FIGURE 21. PIE CHART FOR PERSONAL_INFO4

3.20. Personal_info5

No analysis and useful information can be obtained from personal_info5. It only has 3 distinct values. As we can see in figure 22 and 23, since it has 1437 missing values, 8 occurrences of 1 and 1555 occurrences of 2. Since it has a lot of missing values, we cannot gain any useful insight.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Personal_info5	1	1.9949	?	2	0.0714	-13.8834	190.9944	1,437	0	0	

FIGURE 22. DESCRIPTIVE STATISTICS FOR PERSONAL_INFO5

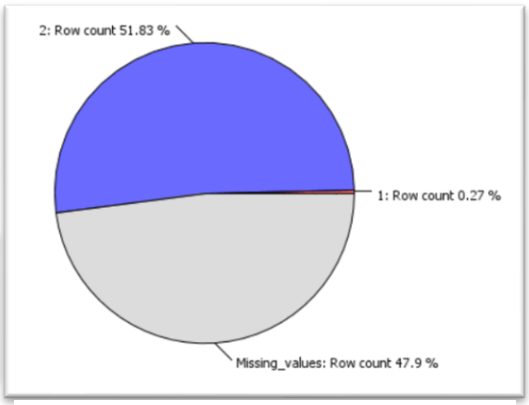


FIGURE 23. PIE CHART FOR PERSONAL_INFO5

3.21. Property_info1

Property_info1 is a binary nominal attribute and consists of a Yes/No flag with the count of 2627 for No which is 87.57% and 373 for Yes which is 12.43%.

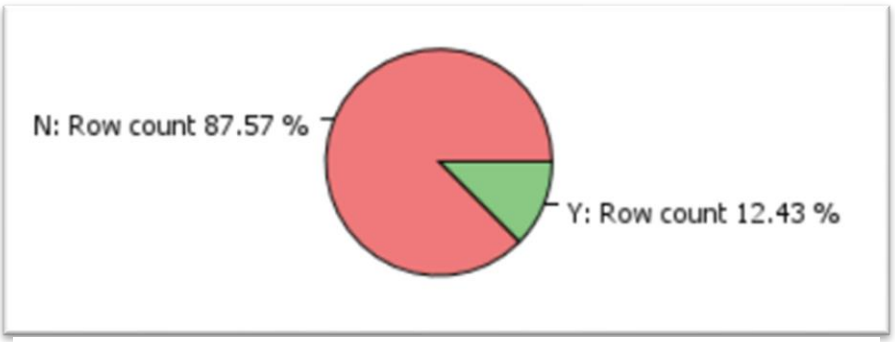


FIGURE 24. PIE CHART FOR PROPERTY_INFO1

3.22. Property_info2

There is no data available for property_info2 as seen in the statistics below, all the entries for this attribute are 0.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Property_info2	0.0	0.0	?	0.0	0.0	0.0	0.0	0	0	0	

FIGURE 25. DESCRIPTIVE STATISTICS FOR PROPERTY_INFO2

3.23. Property_info3

Property_info3 consists of alphabetical categories of data which can be ordered alphabetically as seen in figure 26. The values are unevenly distributed with the top 3 categories O, R, and J making up almost 68% of the total of 15 categories.

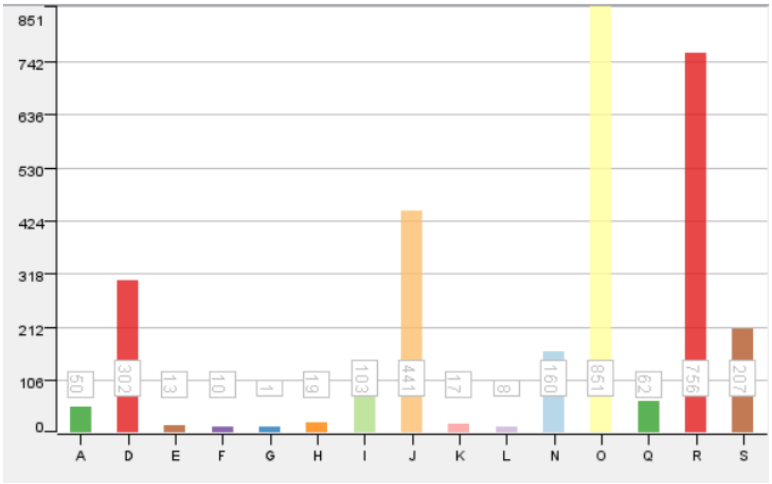


FIGURE 26. HISTOGRAM FOR PROPERTY_INFO3

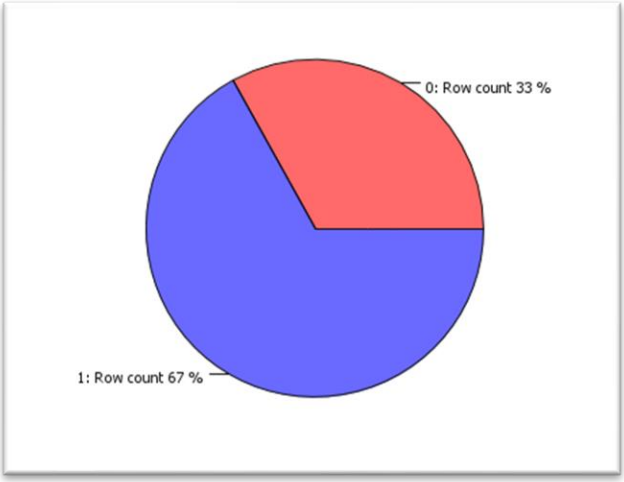


FIGURE 27. PIE CHART FOR PROPERTY_INFO4

3.24. Property_info4

The attribute personal_info4 is a binary attribute with 990 (33%) and 2010 (67%) occurrence(s) of 0 and 1, respectively. It has been visualised below in figure 27.

3.25. Property_info5

The attribute has an even distribution. As we can see on the right in figure 28 the minimum value is -1 and maximum is 25. The median is 13 and the lower/25th percentile is 6 and upper/75th percentile is 19.

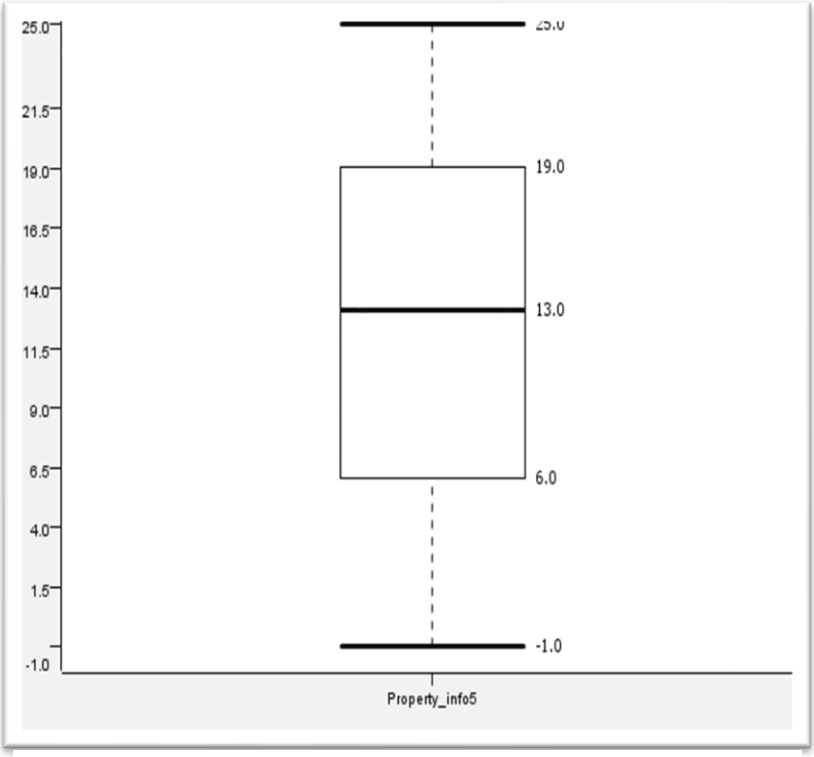


FIGURE 28. BOX PLOT FOR PROPERTY_INFO4

3.26. Geographic_info1

Geographic_info1 has positive integer values, therefore it is a ratio type attribute. As we can see on the right in figure 29 the minimum value is 2 and maximum is 25. The median is 4 and the lower/25th percentile is 2 and upper/75th percentile is 11. We can also see an outlier present; 25.

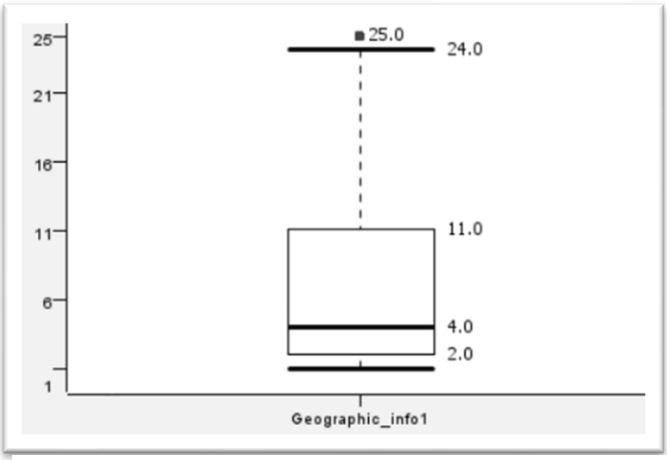


FIGURE 29. BOX PLOT FOR GEOGRAPHIC_INFO1

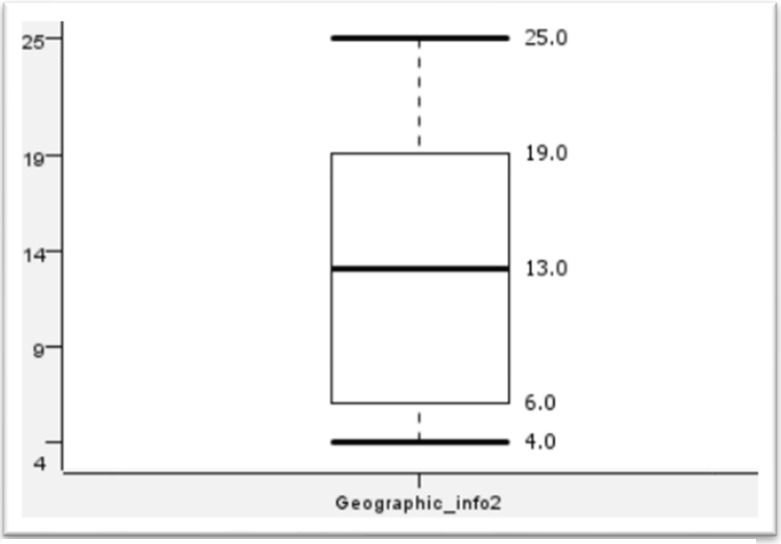


FIGURE 30. BOX PLOT FOR GEOGRAPHIC_INFO2

3.27. Geographic_info2

Geographic_info2 has positive integer values, much like geographic_info1 therefore it is a ratio type attribute. This attribute is much evenly distributed as compared to geographic_info1. As we can see on the left in figure 30 the minimum value is 4 and maximum is 25. The median is 13 and the lower/25th percentile is 6 and upper/75th percentile is 11.

3.28. Geographic_info3

For geographic_info3 we cannot say much as there are only 2 distinct values (-1 & 25) and there is a significant gap between them. For now, we can only say that it is a nominal type of attribute. The count is 2931 for -1 which makes up 97.7% and 69 for 25 which makes up 2.3% as seen in figure 31.

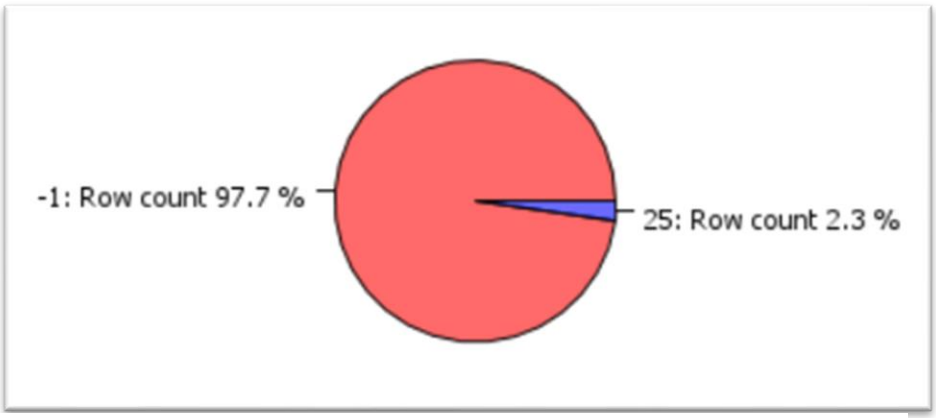


FIGURE 31. PIE CHART FOR GEOGRAPHIC_INFO3

3.29. Geographic_info4

Geographic_info4 is a nominal attribute and consists of a Yes/No flag with the count of 2945 for No which is 98.17% and 55 for Yes which is 1.83% as seen in figure 32.

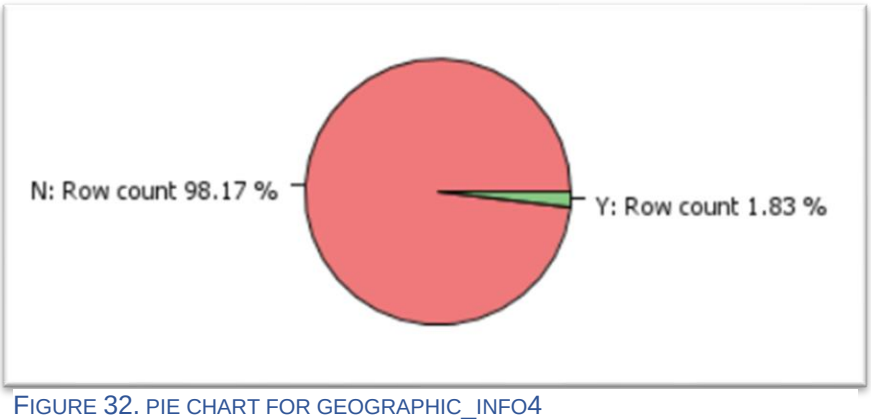


FIGURE 32. PIE CHART FOR GEOGRAPHIC_INFO4

3.30. Geographic_info5

Geographic_info5 consists of alphabetical categories of data which can be categorised by their code as seen in figure 33. This attribute seems to tell the state of residence possibly maybe(?) of the person purchasing the insurance in US. The codes might stand for California, Illinois, New Jersey, and Texas.

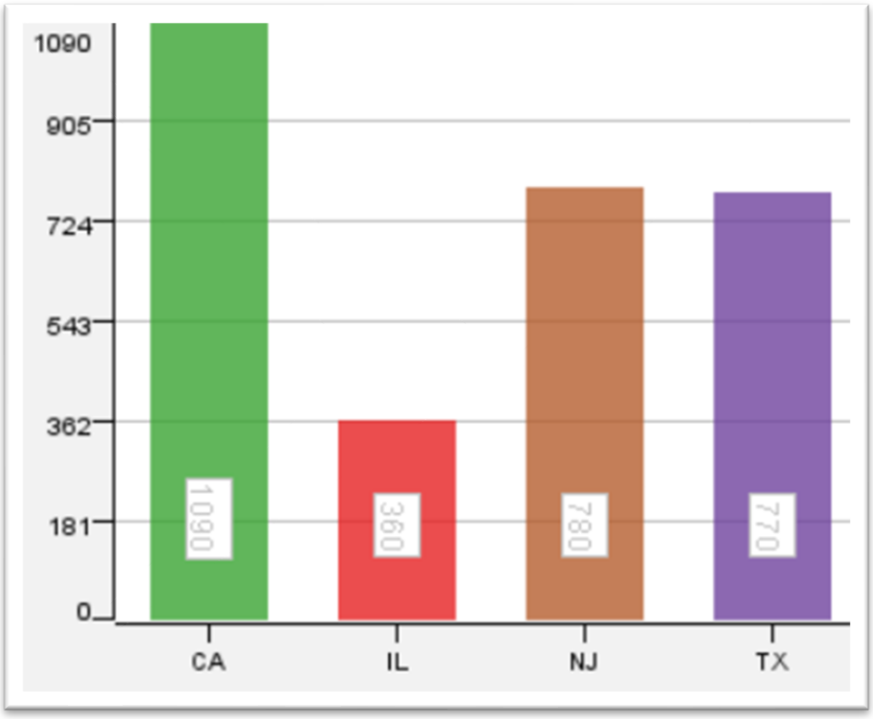


FIGURE 33. HISTOGRAM FOR GEOGRAPHIC_INFO5

4. Interacting with the dataset using KNIME

To understand more about a dataset and what it might come in use for, the relationship between attributes must be investigated. For the following purpose, one of the ways is that we have the nodes Rank Correlation and Matrix Correlation which have algorithms which can indicate relationships amongst attributes. To read this map, the darker blue the color on the grid is, the more 2 attributes are related.

4.1. Rank Correlation Matrix

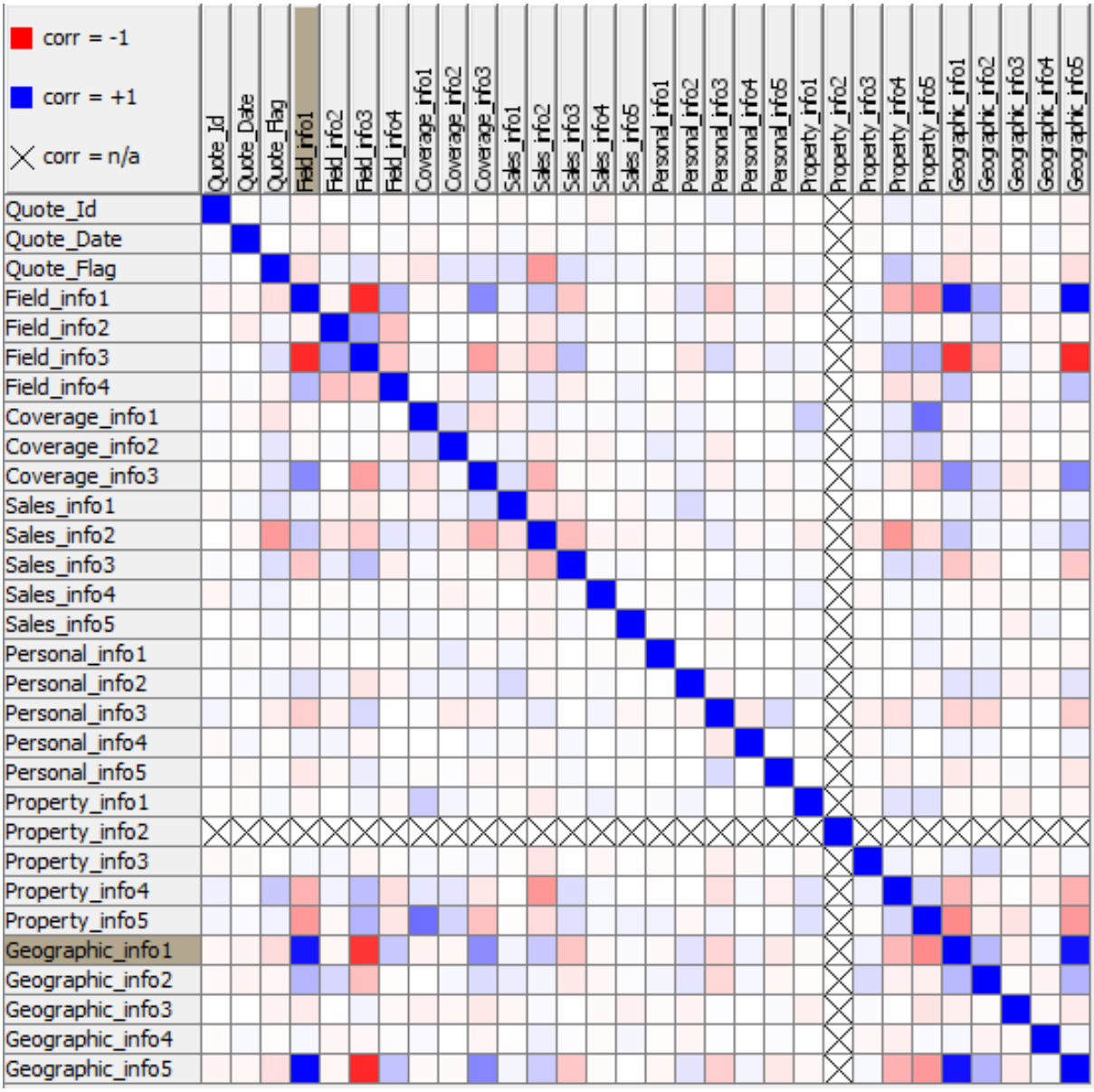


FIGURE 34. RANK CORRELATION MATRIX FOR THE DATASET

4.2. Linear Correlation Matrix

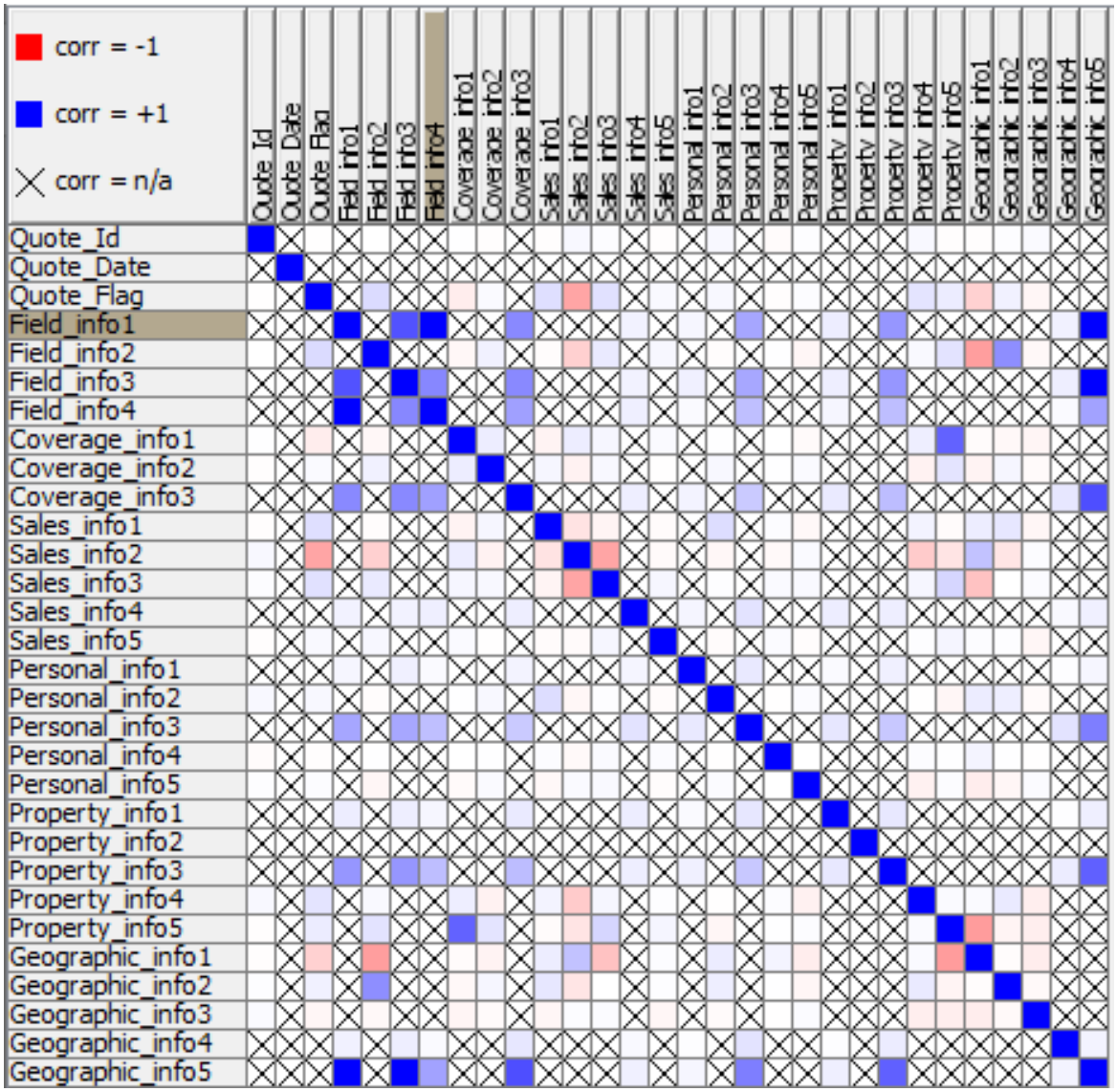


FIGURE 35. LINEAR CORRELATION MATRIX FOR THE DATASET

On reading these 2 matrices, we find that there are relationships between the following attributes:

- Field_info1 & Field_info4 1 (Field_info1 - Field_info4)
- Field_info1 & Geographic_info1 0.9253 (Field_info1 - Geographic_info1)
- Field_info1 and Geographic_info5 0.9998 (Field_info1 - Geographic_info5)
- Geographic_info1 and Geographic_info5 0.9331 (Geographic_info1 - Geographic_info5)
- Field_info3 and Geographic_info5 1 (Field_info3 - Geographic_info5)

To understand the relationship between the attributes, we will make a scatter plot for each of them using the scatter plot node on KNIME.

4.1.1. Field_info1 and Field_info4

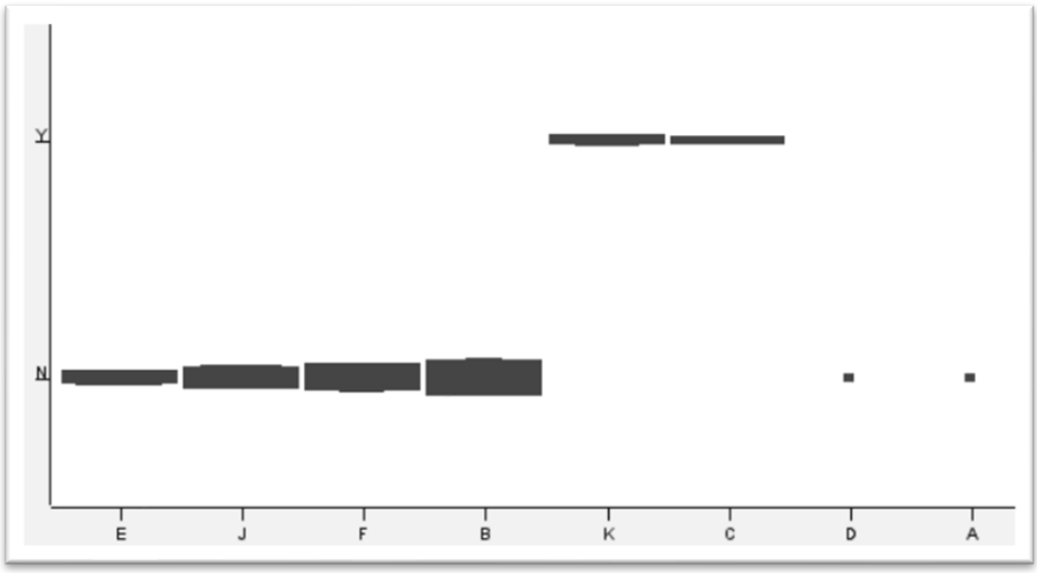


FIGURE 36. SCATTER PLOT FOR FIELD_INFO1 & FIELD_INFO4

In figure 36, we can see the relationship between Field_info1 and field_info4 quite evidently that only the K and C categories from field_info1 were the ones to choose yes in field_info4 and the rest chose no.

4.1.2. Field_info1 and Geographic_info1

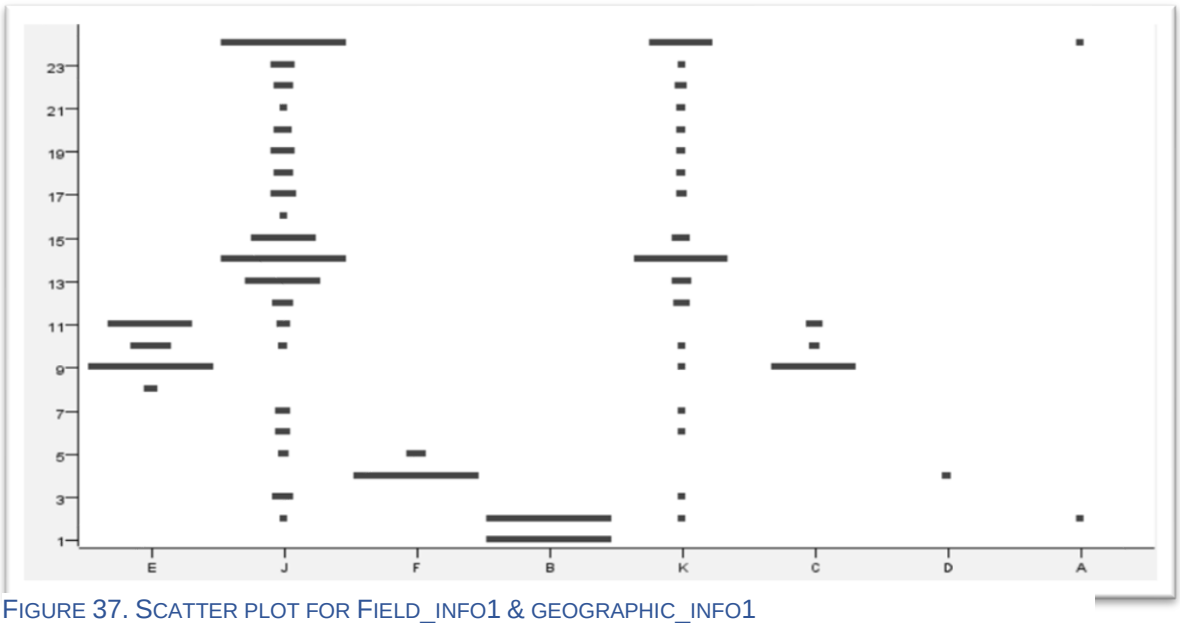


FIGURE 37. SCATTER PLOT FOR FIELD_INFO1 & GEOGRAPHIC_INFO1

In figure 37, we can see the relationship between Field_info1 and Geographic_info1. We see that only the categories F and K have values more than that of 12 and rest all the categories have values less than 12 other than one outlier in A.

4.1.3. Field_info1 and Geographic_info5

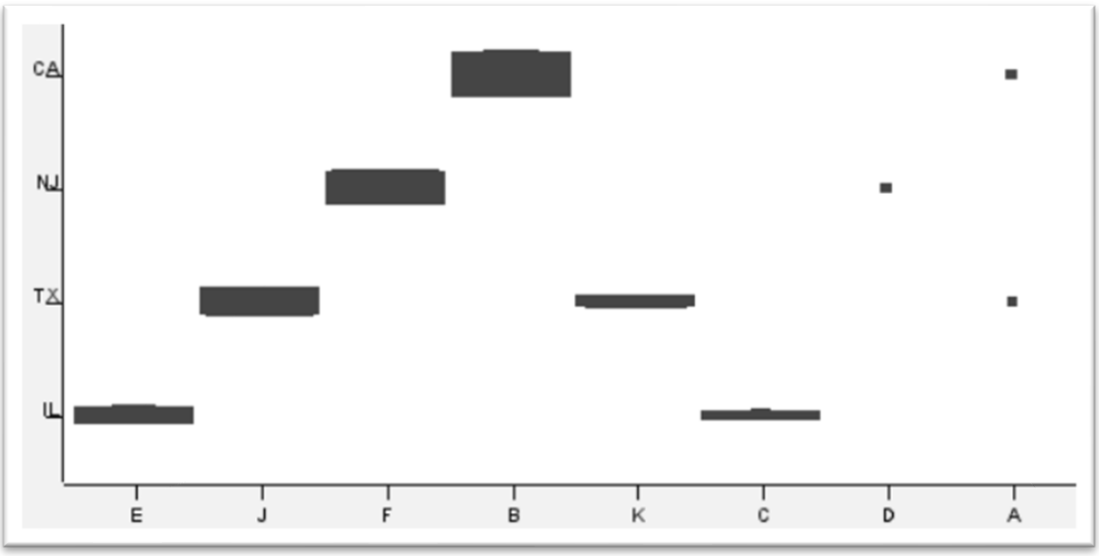


FIGURE 38. SCATTER PLOT FOR FIELD_INFO1 & GEOGRAPHIC_INFO5

In figure 38, we can see the relationship between Field_info1 and Geographic_info5 and that there is no overlap in these groups. We see that:

- the categories E and C from field_info1 belong to IL geographic_info5.
- the categories J, K and A from field_info1 belong to TX geographic_info5.
- the categories F and D from field_info1 belong to NJ geographic_info5.
- the categories B and A from field_info1 belong to CA geographic_info5.

4.1.4. Geographic_info1 and Geographic_info5

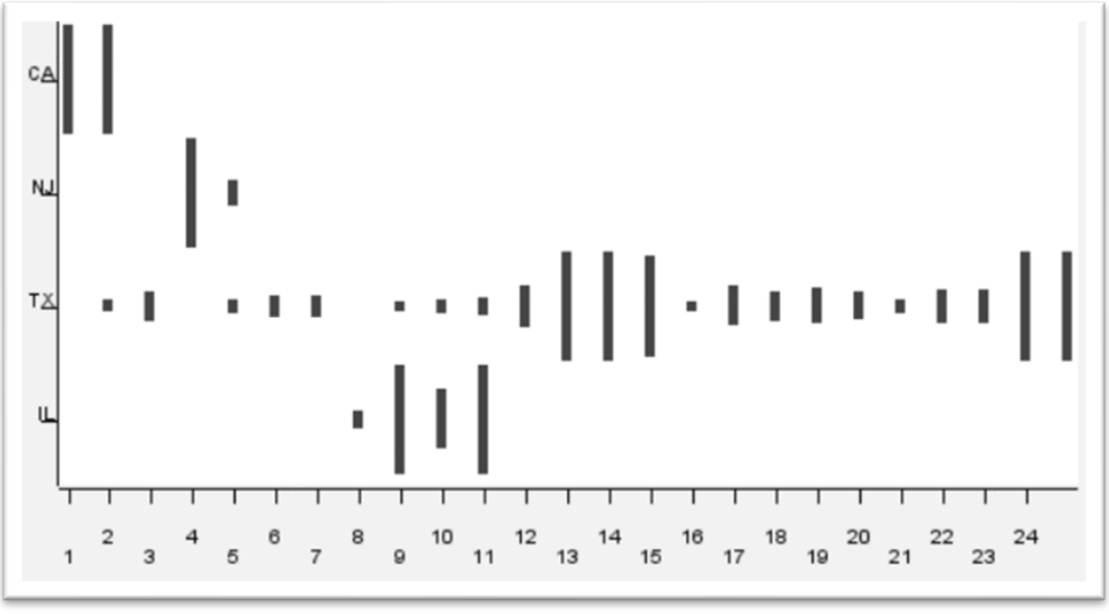


FIGURE 39. SCATTER PLOT FOR GEOGRAPHIC_INFO1 AND GEOGRAPHIC_INFO5

In figure 39, we can see the relationship between Geographic_info1 and Geographic_info5 and that there is no overlap in these groups. We see that:

- the values of 1 and 2 from geographic_info1 belong to CA geographic_info5.
- the NJ category in geographic_info5 only consists of the value 4 and 5.

- the IL category only consists of values 8, 9, 10, and 11 among which 9 and 11 occur the most.
- the rest of the values from our 1-25 range are from the TX category from geographic_info5.

4.1.5. Field_info3 and Geographic_info5

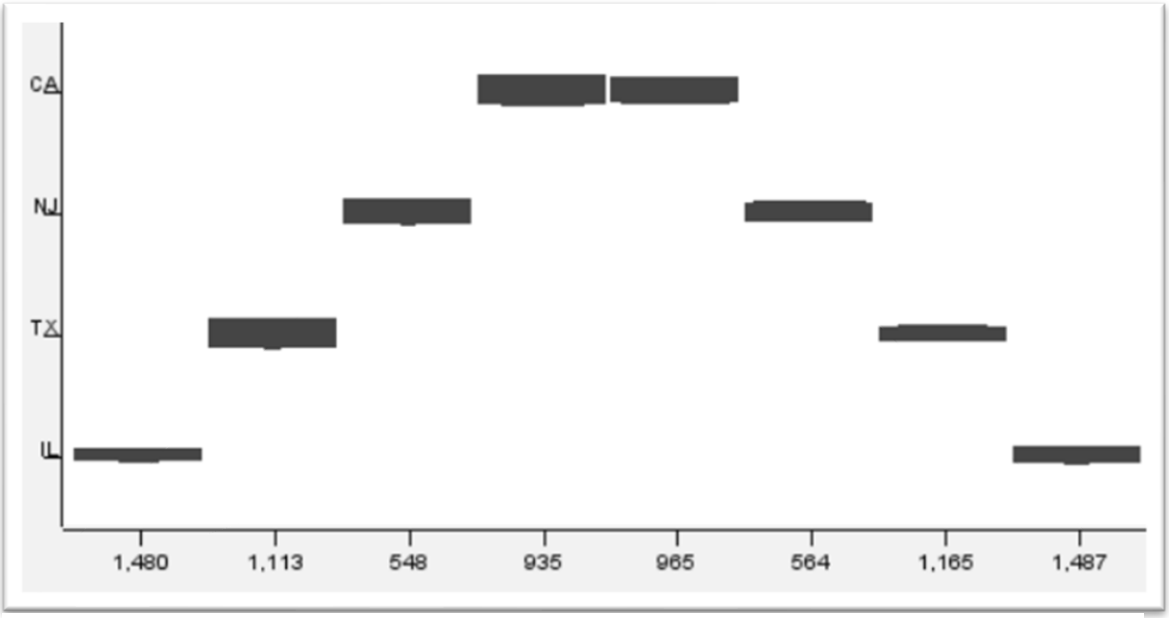


FIGURE 40. SCATTER PLOT FOR FIELD_INFO3 AND GEOGRAPHIC_INFO5

In figure 40, we can see the relationship between Field_info3 and Geographic_info5 and we can clearly see that from the scatter plot as well. We can conclude that:

- the values/category of 1487 and 1480 from field_info3 belong to IL geographic_info5.
- the values/category of 1113 and 1165 from field_info3 belong to TX geographic_info5.
- the values/category of 548 and 564 from field_info3 belong to NJ geographic_info5.
- the values/category of 935 and 965 from field_info3 belong to CA geographic_info5.

5. Data Preprocessing (1B)

Certain Data Preprocessing methods were also used on the dataset as follows.

5.1 Binning Attribute Property_info5

The number of bins I have decided to go with is 5. Since the values range from -1 to 25, 5 bins seemed like the right choice with a range of 5 in each one.

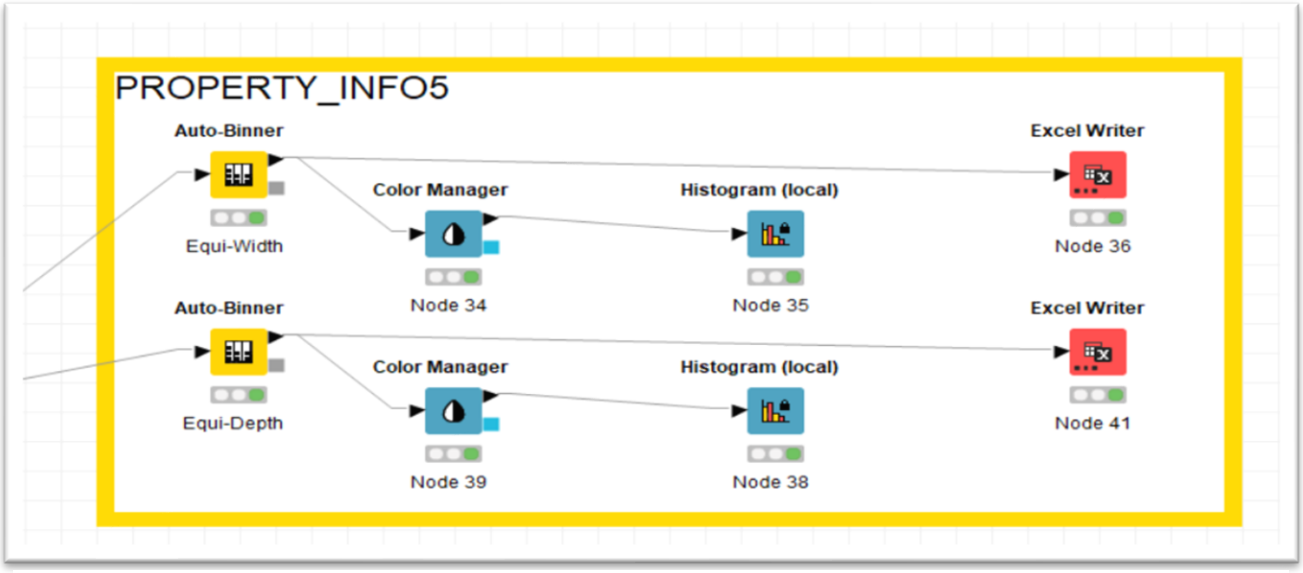


FIGURE 41. KNIME WORKFLOW FOR BINNING

5.1.1 Equi-Width Binning

As shown in fig 41, firstly the File Reader node was used to read the csv file (not in the image) and then Auto-Binner node was used with the width setting set to 5 to bin the attribute Property_info5. Then using the color manager node to set a color for the histogram. And finally, histogram in figure 42 was formed after executing. Further, an excel writer node was used to extract the results.

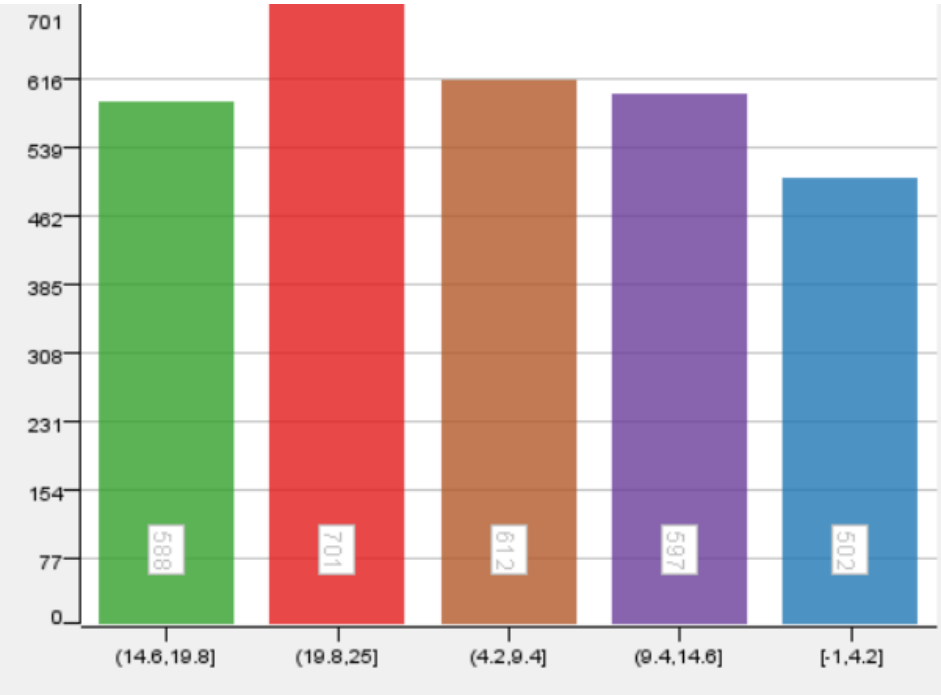


FIGURE 42. HISTOGRAM FOR EQUI-WIDTH BINNING OF PROPERTY_INFO5

5.1.2 Equi-Depth Binning

As shown in fig 41, firstly the File Reader node was used to read the csv file (not in the image) and then Auto-Binner node was used with the frequency setting set to 5 to bin the attribute Property_info5. Then using the color manager node to set a color for the histogram. And finally, histogram in figure 43 was formed after executing. Further, an excel writer node was used to extract the results.

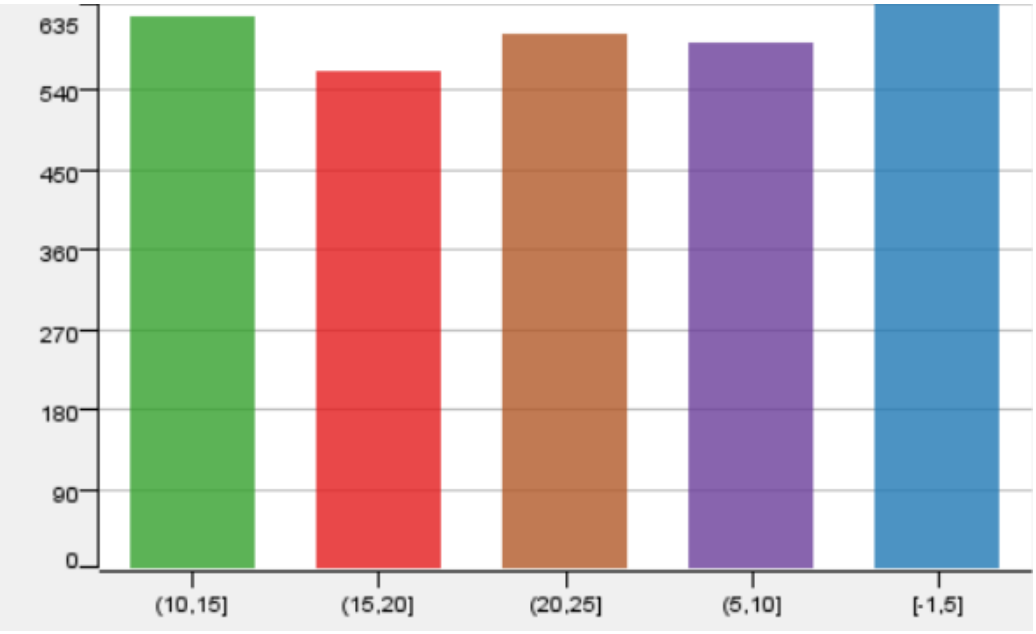


FIGURE 43. HISTOGRAM FOR EQUI-DEPTH BINNING OF PROPERTY_INFO5

5.2 Normalizing Attribute Sales_info5

When dealing with large numbers, it can get messy sometimes and normalization is a technique used to avoid that and to work with the numbers in an easy way and reduce them in a relative manner.

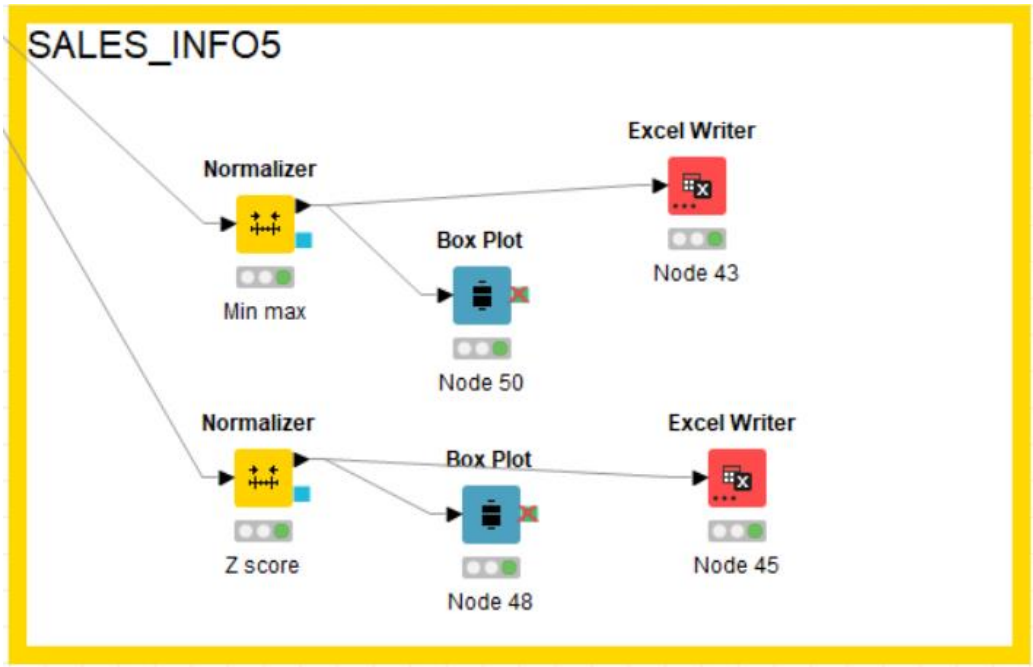


FIGURE 44. WORKFLOW FOR NORMALIZING IN KNIME

5.2.1 Min-Max Normalization

Sales_info5
0.059
0.805
0.546
0.135
0.387
0.444
0.697
0.402
0.339
0.546
0.231
0.607
0.045
0.294
0.349
0.791
0.043

FIGURE 45. MIN-MAX NORMALIZATION RESULT

As shown in the figure 44, the File Reader node is used to read the csv file (not in the image) and then the normalizer node Is used to normalize the sales_info5 attribute, Selecting the Min-Max Normalization in settings. And finally, the Excel writer node to export the normalized data into a excel file (results attached in the excel file). A sample can be seen in figure 45.

5.2.2 Z-Score Normalization

Z-Score normalization is the normalization based on the mean and standard deviation of the attribute.

This normalization was done following the same a method as min-max normalization except changing the settings to Z-score (results attached in the excel file). A sample can be seen in figure 46.

Sales_info5
-1.506
1.064
0.172
-1.242
-0.373
-0.18
0.695
-0.321
-0.54
0.172
-0.914
0.383

FIGURE 46. Z-SCORE NORMALIZATION RESULT

5.3 Discretizing Attribute Coverage_info1

The attribute has been categorised into 4 different categories:

1. Basic: (<5)
2. Low: [5-10)
3. Medium: [10-15)
4. High: (15<)

As shown in the fig 47, Using File Reader node the csv file is read (not in the image) then the Numeric Binner node is used to categorise into the 4 categories where we enter the values ourselves. And finally, Excel Writer node is used to export into a excel file. A histogram of the result can be seen in figure 48.

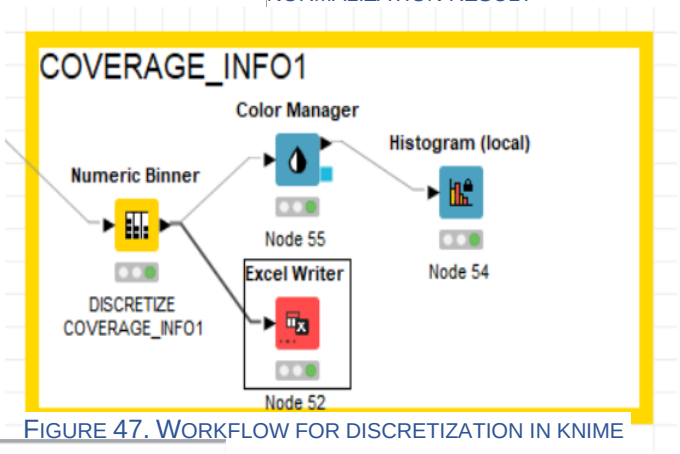


FIGURE 47. WORKFLOW FOR DISCRETIZATION IN KNIME

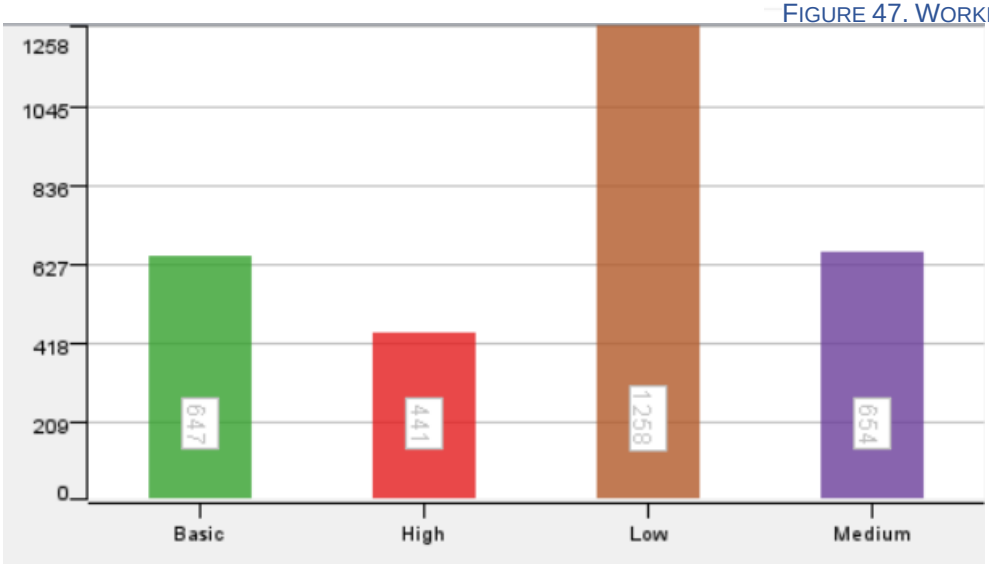


FIGURE 48. HISTOGRAM FOR POST DISCRETISED COVERAGE_INFO1

5.4 Binarizing Attribute Geographic_info5

Binarization is the method of transforming data features of any entity into vectors of binary numbers so that it is easier to read the data and identify existence of the data.

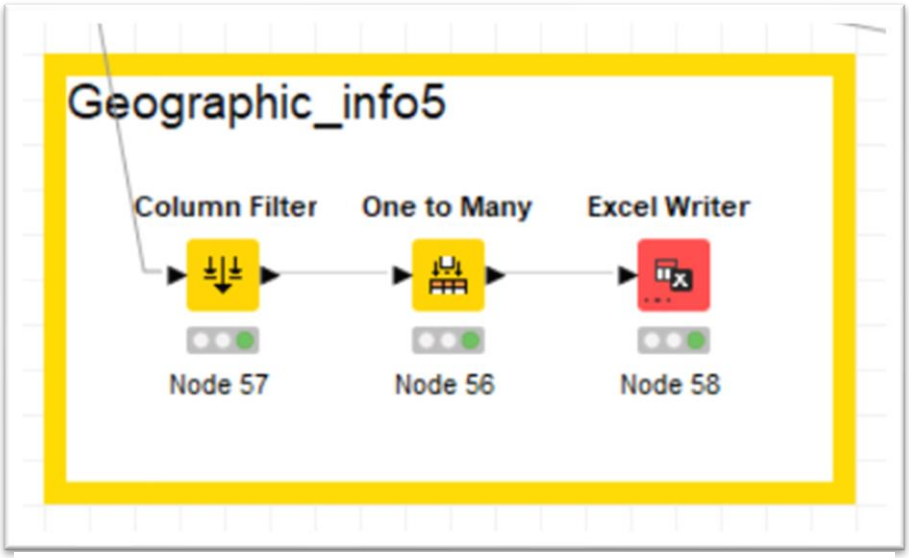


FIGURE 49. BINARIZING WORKFLOW IN KNIME

As shown in the figure 49, File reader node is used to read the csv file (not shown in the image) then the Column Filter node is used to select the Geographic_info5 column. And then using the One to Many node to binaries the data. Finally, the Excel Writer node is used to export the data to a excel file. A sample can be seen in figure 50.

S Geogra...	I IL	I TX	I NJ	I CA
IL	1	0	0	0
TX	0	1	0	0
NJ	0	0	1	0
CA	0	0	0	1
CA	0	0	0	1
CA	0	0	0	1

FIGURE 50. SAMPLE OF BINARIZED GEOGRAPHIC_INFO5

6. Summary

After studying and working with the dataset, there needs to be more information gathered about what the attributes represent to study them more and establish relationships amongst them since at the moment they seem highly unrelated. However, with the present data that we have some relationships between two attributes could be established and observed:

- Field_info1 & Field_info4
- Field_info1 & Geographic_info1
- Field_info1 and Geographic_info5
- Geographic_info1 and Geographic_info5
- Field_info3 and Geographic_info5