31250 Introduction to Data Analytics

Assignment #2

Data Exploration and Preparation

Florian Lubitz
FEIT
13688799
florian.lubitz@student.uts.edu.au

Contents

1 Initial data exploration

1.1 Attribute features

1.1.1 Quote_Id

Attribute type: nominal The "Quote_Id" is a customer number. It is neither orderable nor is the number a quantitative value. As the id is a unique key, there is no reason to perform any statistic analysis on this attribute.

1.1.2 Quote_Date

Attribute type: interval "Quote_Date" seems to be a date. These have a fixed sequence and plus and minus operations can be performed with them.

Statistic	Value
Missing Values	None
Minimum	02.01.2013
Maximum	18.05.2015
Mean	23.03.2014
Std. Dev.	245 d

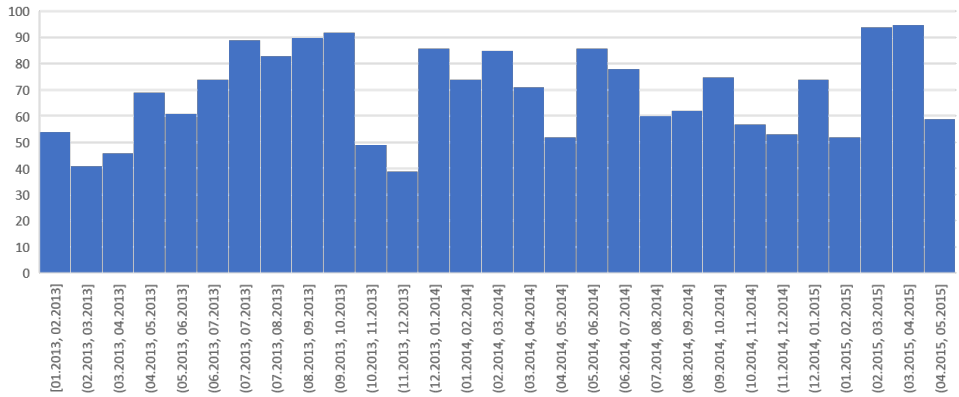

Figure 1: Histogramm Quote_Date

1.1.3 Quote_Flag

Attribute type: nominal (dichotomous) The description of the attributes declares "Quote_Flag" as information about whether an insurance was purchased. This can be only answered with yes or no.

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
0	1605	0.8025
1	395	0.1975

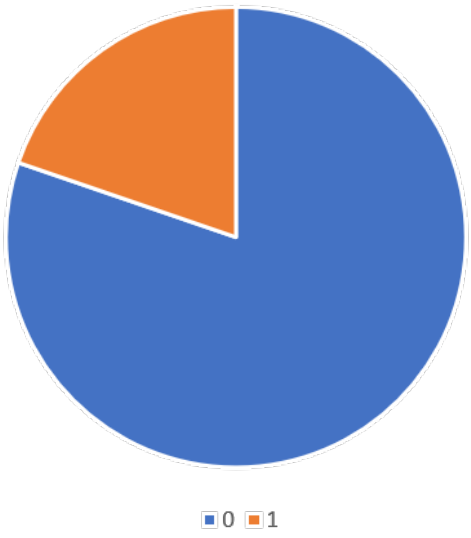

Figure 2: Pie Graph Quote_Flag

1.2 Field_Info1

Attribute type: nominal These characters are not close enough together to asssume, they belong to a ordererd attribute.

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
J	390	0.365
F	540	0.27
B	730	0.195
E	200	0.1
C	39	0.0505
K	101	0.0195

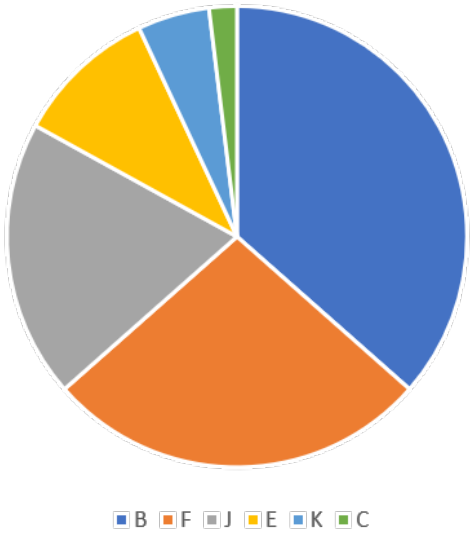

Figure 3: Pie Graph Field_Info1

1.3 Field_Info2

Attribute type: ratio All values of "Field_Info2" are numbers between 0 and 2 and have a precision of 4 decimals. This suggests that these values are on a zero-based scale.

Statistic	Value
Missing Values	None
Minimum	0.8746
Maximum	1.0101
Mean	0.9391
Std. Dev.	0.0373
90th Perc.	1.00051
10th Perc.	0.8922

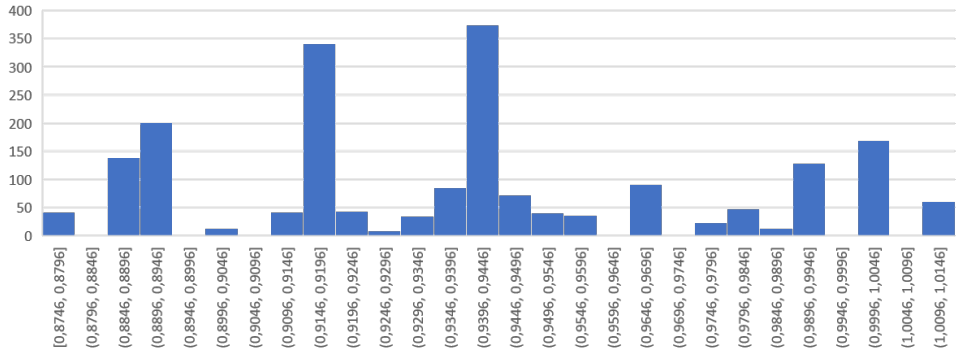

Figure 4: Histogramm Field_Info2

1.4 Field_Info3

Attribute type: ratio The values in this attribute are all numeric range in a range of around 1000. Given the size of the numbers this could be a attribute containing monetary values.

Statistic	Value
Missing Values	None
Minimum	548
Maximum	1487
Mean	950.5195
Std. Dev.	290.62
90th Perc.	901480
10th Perc.	548

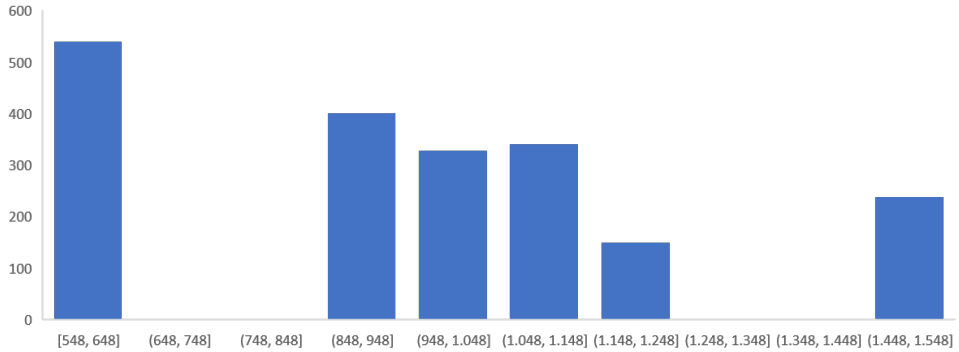

Figure 5: Histogramm Field_Info3

1.5 Field_Info4

Attribute type: nominal (dichotomous) In this Attribute only the values "Y" and "N" appear. Those values are often used as a short version on "Yes" and "No". This field therefore seems to be a yes-no answer.

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
N	1860	0.93
Y	140	0.07

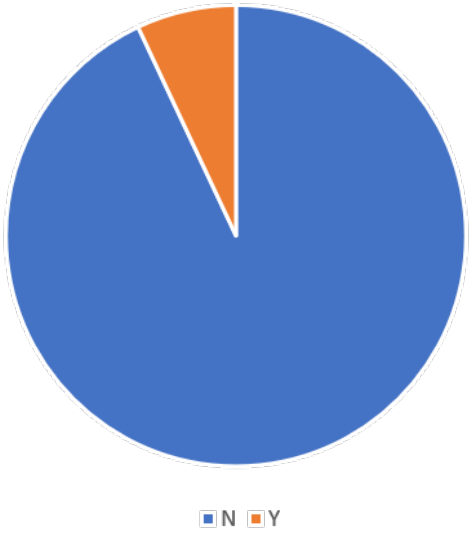

Figure 6: Pie Graph Field_Info4

1.6 Coverage_Info1

Attribute type: ordinal This attribute contains values from 1 to 25. These could be different values in a specified order but not a numeric value.

Statistic	Value
Missing Values	None
Numeric Outliers	74

Value	Absolute Frequency	Relative Frequency
-1	2	0.001
1	42	0.021
2	81	0.0405
3	118	0.059
4	149	0.0745
5	189	0.0945
6	182	0.091
7	185	0.0925
8	148	0.074
9	151	0.0755
10	113	0.0565
11	106	0.053
12	83	0.0415
13	67	0.0335
14	62	0.031
15	53	0.0265
16	49	0.0245
17	35	0.0175
18	37	0.0185
19	19	0.0095
20	18	0.009
21	19	0.0095
22	18	0.009
23	12	0.006
24	11	0.0055
25	51	0.0255

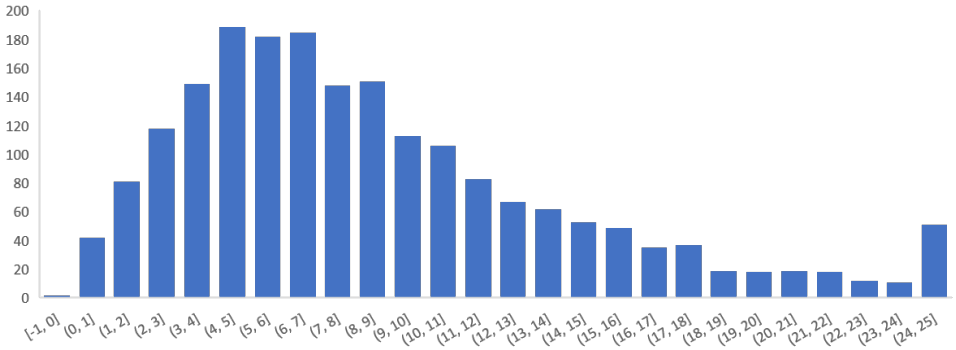

Figure 7: Histogramm Coverage_Info1

1.7 Coverage_Info2

Attribute type: ordinal As before, this attribute contains values from 1 to 25. These could be different values in a specified order but not a numeric value but we only got datapoints with values at 1,2, 22 and 25, which is quite interesting as only 3 customers selected value 1

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
1	3	0.0015
2	125	0.0625
22	1620	0.81
25	252	0.126

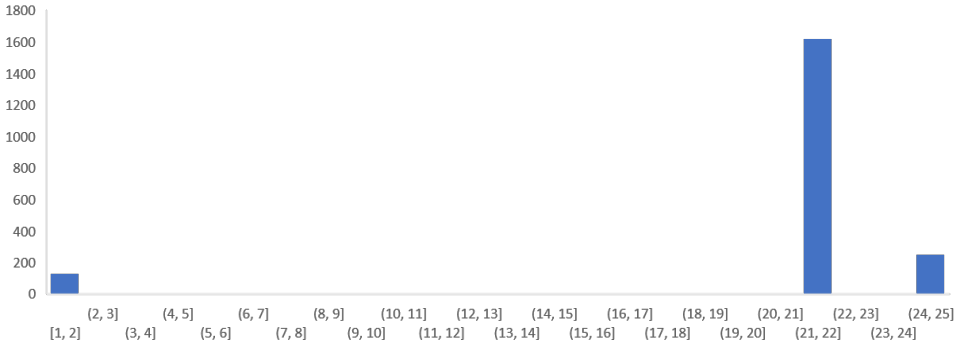

Figure 8: Histogramm Coverage_Info2

1.8 Coverage_Info3

Attribute type: ordinal This field contains letters from the beginning of the alphabet. This indicates for these values to be ordered in that order.

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
A	116	0.058
B	4	0.002
C	3	0.0015
D	388	0.194
E	673	0.3365
F	221	0.1105
G	245	0.1225
H	2	0.001
I	7	0.0035
J	110	0.055
K	229	0.1145
L	2	0.001

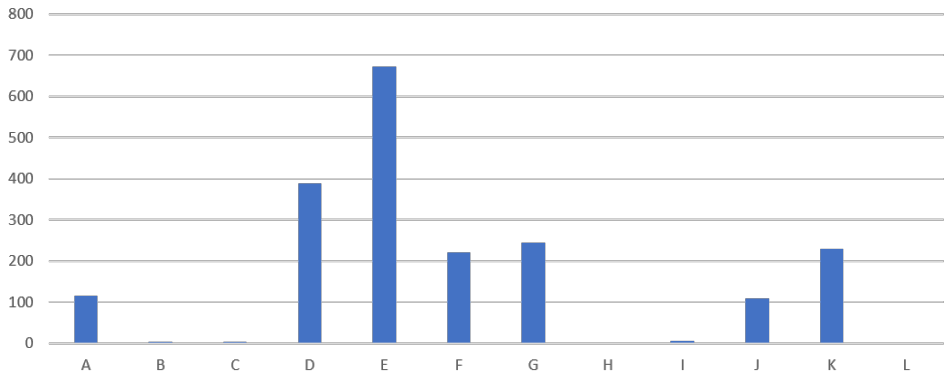

Figure 9: Histogramm Coverage_Info3

1.9 Sales_Info1

Attribute type: nominal (dichotomous)

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
1	1490	0.745
0	510	0.255

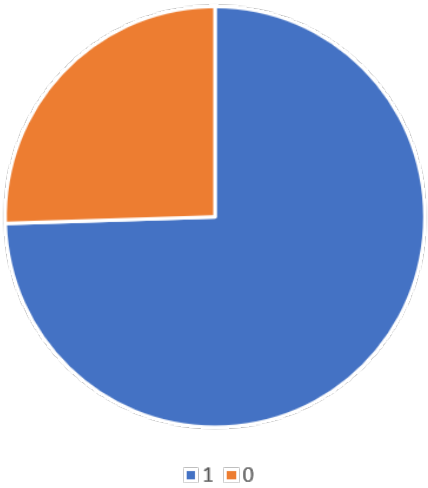

Figure 10: Pie Graph Sales_Info1

1.10 Sales_Info2

Attribute type: ordinal This Attributes contains numbers from 2 to 5. Probably, the provided data is missing the value 1. With that value these could be values ordered from 1 to 5.

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
2	93	0.0465
3	519	0.2595
4	295	0.1475
5	1093	0.5465

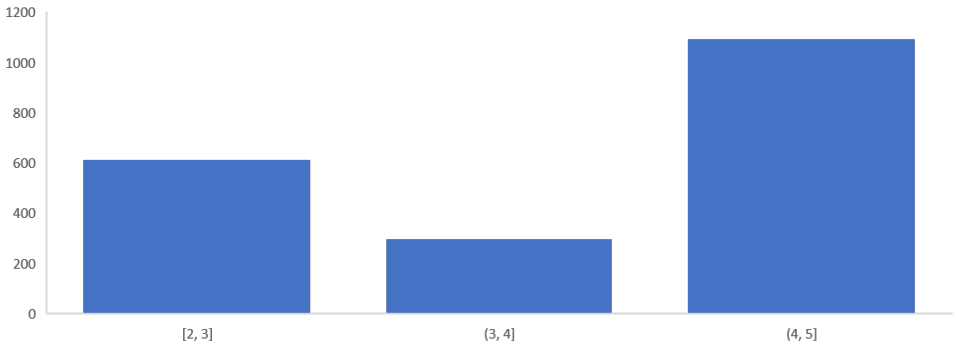

Figure 11: Histogramm Sales_Info2

1.11 Sales_Info3

Attribute type: ordinal Like Coverage_Info1 this is a ordinal attribute, but we are missing some values inside the provided range.

Statistic	Value
Missing Values	None

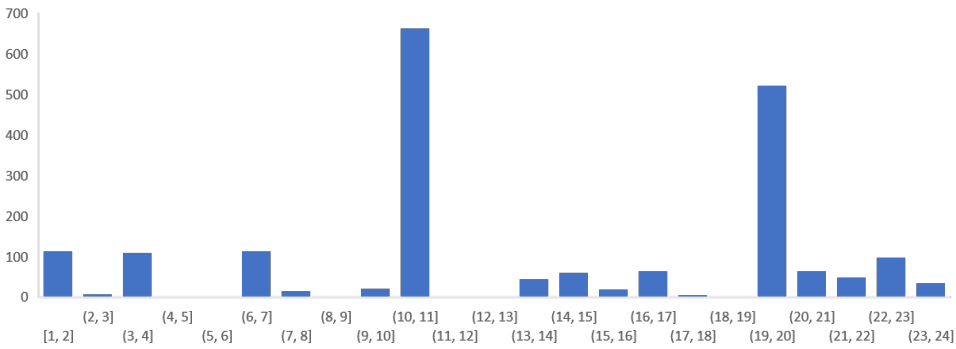

Figure 12: Histogramm Sales_Info3

1.12 Sales_Info4

Attribute type: nominal These letters seem to have no recognisable order.

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
K	410	0.205
P	376	0.188
T	332	0.166
Q	312	0.156
V	298	0.149
R	156	0.078
M	116	0.058

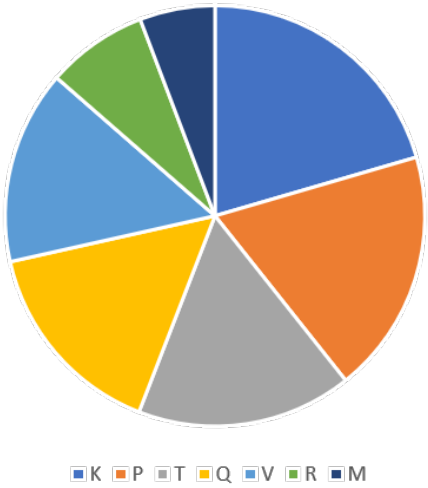

Figure 13: Pie Graph Sales_Info4

1.13 Sales_Info5

Attribute type: ratio The values in this attribute are distributed very evenly and range in a big range. Given the size of the numbers this could be a attribute containing monetary values.

Statistic	Value
Missing Values	None
Minimum	1
Maximum	67162
Mean	33785.54
Std. Dev.	18863.43
90th Perc.	59741.7
10th Perc.	7828.9

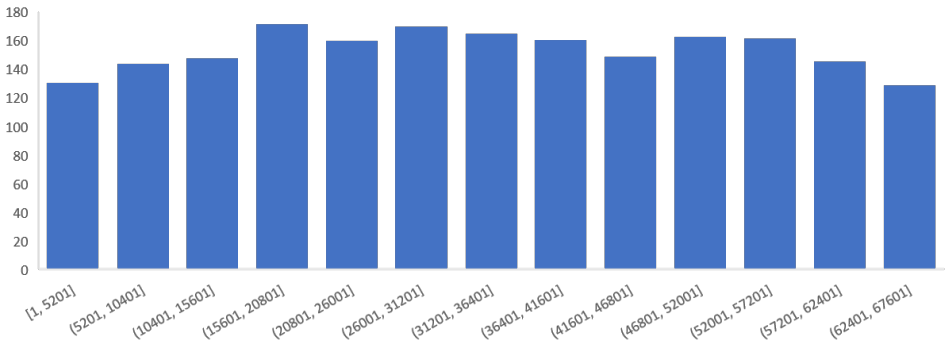

Figure 14: Histogramm Sales_Info5

1.14 Personal_Info1

Attribute type: nominal (dichotomous)

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
N	1993	0.9965
Y	7	0.0035

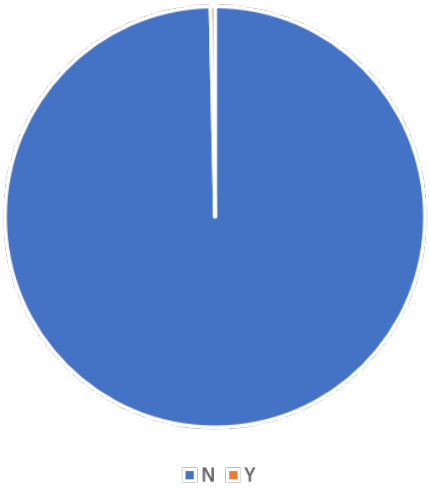

Figure 15: Pie Graph Personal_Info1

1.15 Personal_Info2

Attribute type: ordinal Like Coverage_Info1, this attribute probably also contains ordered values, but in the range between 1 and 25. And the value -1 means "no answer".

Statistic	Value
Missing Values	None

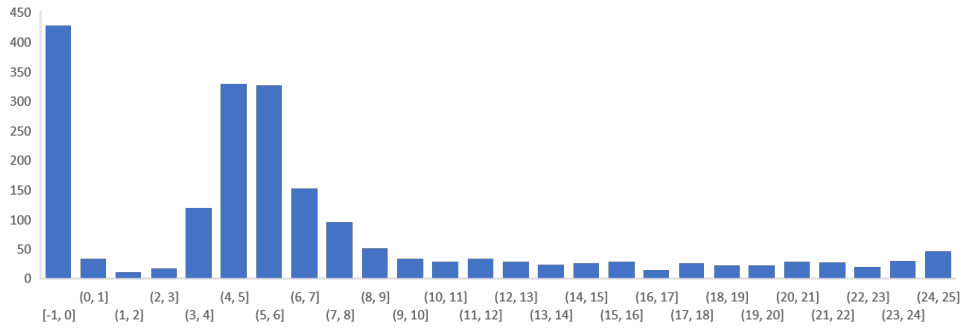

Figure 16: Histogramm Personal_Info2

1.16 Personal_Info3

Attribute type: nominal These two letter combinations could be order in a alphabetic order, but they seem to start at a random place inside the alphabet. Therefore this attribute will be treated as nominal.

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
ZA	937	0.4685
XR	104	0.052
XM	81	0.0405
XJ	73	0.0365
XD	63	0.0315
XX	62	0.031
XB	59	0.0295
YH	58	0.029
XH	54	0.027
ZT	51	0.0255
XO	50	0.025
ZF	46	0.023
ZR	40	0.02
ZN	39	0.0195
ZH	37	0.0185
XS	35	0.0175
YF	29	0.0145
XW	23	0.0115
ZG	20	0.01
XE	20	0.01
ZW	18	0.009
YE	17	0.0085
ZC	16	0.008
XC	14	0.007
XQ	14	0.007
XL	7	0.0035
ZE	7	0.0035
ZJ	6	0.003
XI	5	0.0025
ZD	5	0.0025
ZK	4	0.002
XZ	4	0.002
ZU	2	0.001

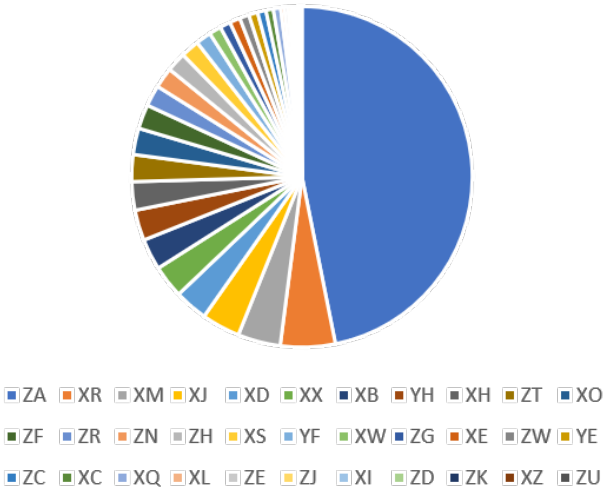

Figure 17: Pie Graph Personal_Info3

1.17 Personal_Info4

Attribute type: nominal (dichotomous) All those values are 0 except for one, which is 1.

Statistic	Value
Missing Values	None
Outliers	1

Value	Absolute Frequency	Relative Frequency
0	1999	0.9995
1	1	0.001

1.18 Personal_Info5

Attribute type: ordinal This Attribute has no value in a lot of Cases. This Attribute could have up to 5 different, ordered values (1 - 5). This Attribute will probably not get looked at during prediction analysis.

Statistic	Value
Missing Values	966

1.19 Property_Info1

Statistic	Value
Missing Values	1

Value	Absolute Frequency	Relative Frequency
N	1738	0.869
Y	261	0.1305

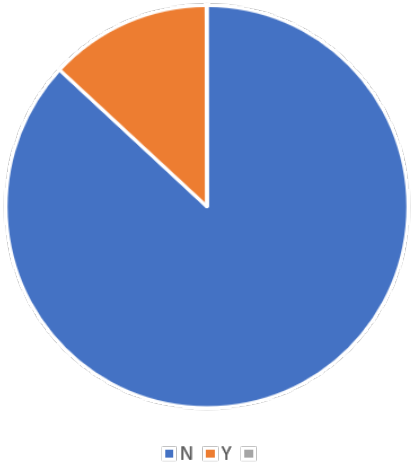

Figure 18: Pie Graph Property_Info1

Attribute type: nominal (dichotomous)

1.20 Property_Info2

Attribute type: nominal All those values are 0

Statistic	Value
Missing Values	None

1.21 Property_Info3

Attribute type: nominal random letters

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
O	573	0.2865
R	533	0.2665
J	285	0.1425
D	198	0.099
S	146	0.073
N	92	0.046
I	77	0.0385
A	31	0.0155
Q	30	0.015
E	10	0.005
H	9	0.0045
K	6	0.003
F	5	0.0025
L	3	0.0015
G	2	0.001

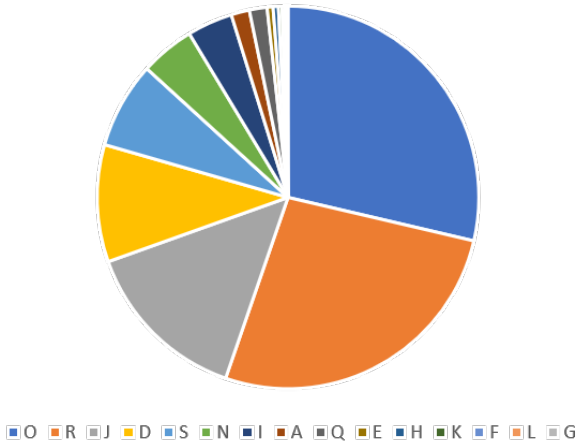

Figure 19: Pie Graph Property_Info3

1.22 Property_Info4

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
1	1364	0.682
0	636	0.318

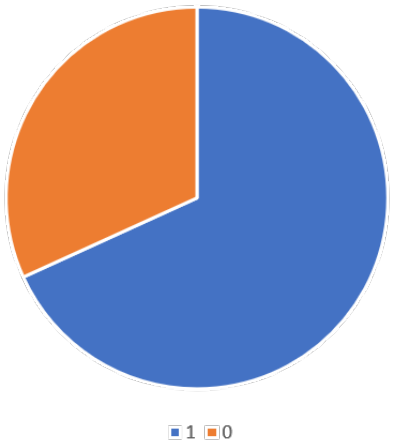

Figure 20: Pie Graph Property_Info4

Attribute type: nominal (dichotomous)

1.23 Property_Info5

Attribute type: ordinal This Attribute too contains values from 1 to 25. I defined those as ordinal before.

Statistic	Value
Missing Values	None

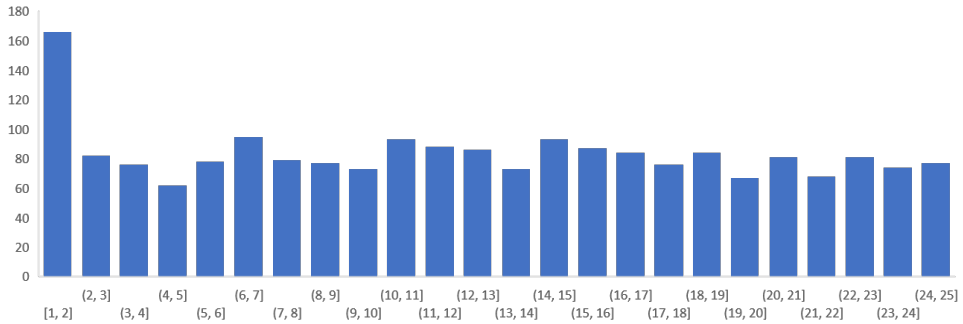

Figure 21: Histogramm Property_Info5

1.24 Geographic_Info1

Attribute type: ordinal This Attribute too contains values from 1 to 25. I defined those as ordinal before.

Statistic	Value
Missing Values	None
Ouliers	58

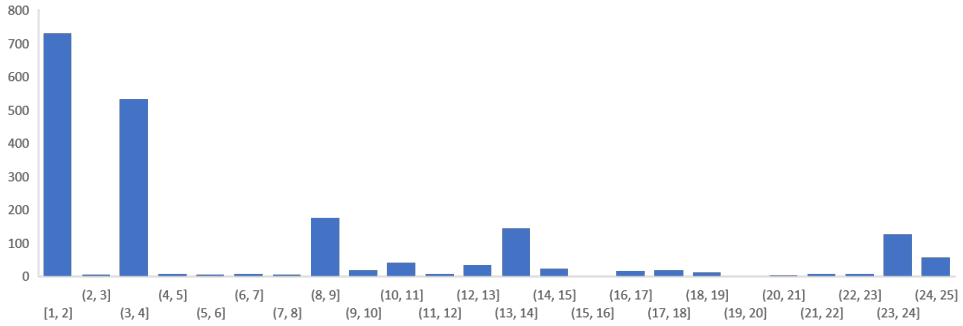

Figure 22: Histogramm Geographic_Info1

1.25 Geographic_Info2

Attribute type: ordinal This Attribute too contains values from 1 to 25. I defined those as ordinal before.

Statistic	Value
Missing Values	None

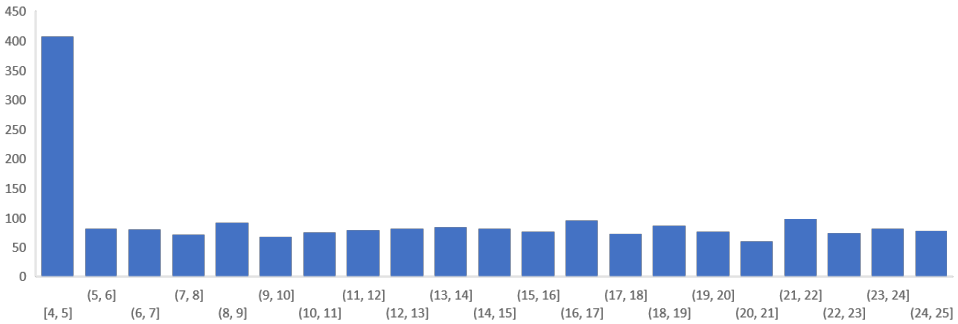

Figure 23: Histogramm Geographic_Info2

1.26 Geographic_Info3

Attribute type: nominal These values are very interessting. They could belong to a ordered attribute like "Personal_Info2". Nevertheless, it is defined as nominal because too many values are missing.

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
-1	1942	0.971
25	58	0.029

1.27 Geographic_Info4

Attribute type: nominal (dichotomous)

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
N	1961	0.9805
Y	39	0.0195

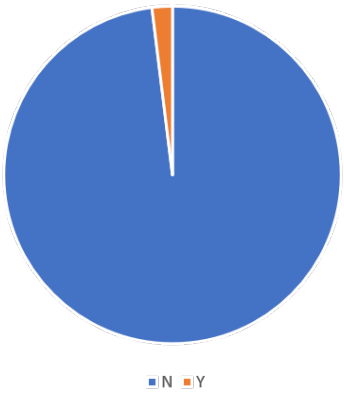

Figure 24: Pie Graph Geographic_Info4

1.28 Geographic_Info5

Attribute type: nominal These values could be the states California, New Jersey, Texas and Illinois. They also could represent any other value and don't have any specific order.

Statistic	Value
Missing Values	None

Value	Absolute Frequency	Relative Frequency
CA	730	0.365
NJ	540	0.27
TX	491	0.2455
IL	239	0.1195

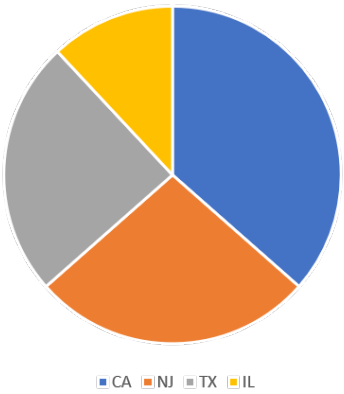

Figure 25: Pie Graph Geographic_Info5

1.29 Interesting Attributes

Field_Info3 All values in this attribute are either on the side or in the middle of the range.

Field_Info4 Normal distribution, but there is a peak at the end.

Coverage_Info2 On this attribute we only have datapoints with four different values allthough the range of those values suggets 25 options have been available.

Geographic_Info2 & Property_Info 5 Both of these values show a even distribution with a spike at the begin of their ranges.

Personal_Info3 Almost 50% of the values are "ZA".

Property_Info2 All values are 0

Personal_Info4 All values are 0 except for one 1

1.29.1 Linear Correlations

There are some interessting correlations in this dataset. Exspecially Geographic_Info5 seems to have some strong correllations with Field_Info2 and Coverage_Info3. Field_Info1 seems to be correlatted to Field_Info4 and Field_Info2 has as negative Correlation with Field_Info3.

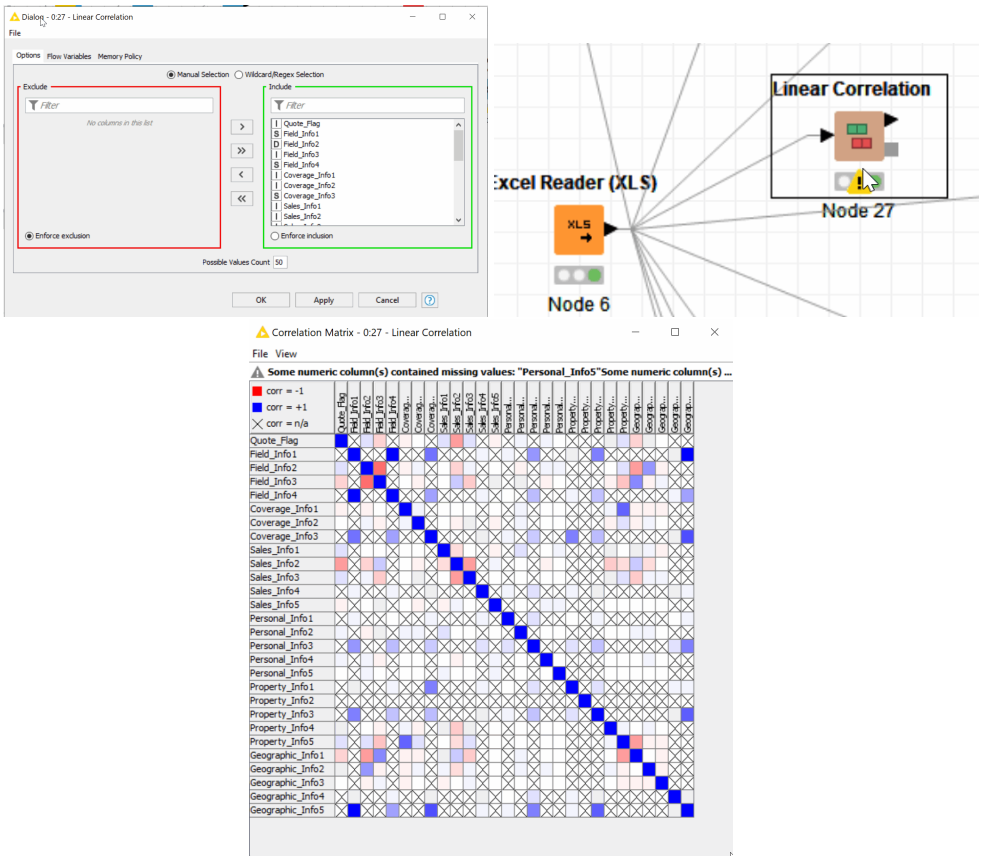

Figure 26: Screenshots Linear Correlation

2 Data Preprocessing

2.1 Binning of Property_Info5

To bin the different values of Property_Info5 we have to decide how many bins we want to create. To do so we can use the following formula.

$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil$$

Where h is the Freedman–Diaconis rule:

$$h = 2\frac{\text{IQR}}{n^{1/3}}$$

Using this we can calculate the number of bins k to be

$$\begin{aligned}
\text{IQR} &= 19 - 7 = 12 \\
n &= 2000 \\
h &= 2\frac{\text{IQR}}{n^{1/3}} = 2\frac{12}{2000^{1/3}} \approx 1.9 \\
k &= \left\lceil \frac{\max x - \min x}{h} \right\rceil = \left\lceil \frac{25 - 1}{1.9} \right\rceil = 13
\end{aligned}$$

With this number of bins, we can use Knime to bin our datapoints. The resulting binned data can be found inside of `data.xlsx`.

Figure 27: Screenshots Binning

2.2 Normalization of Sales_Info5

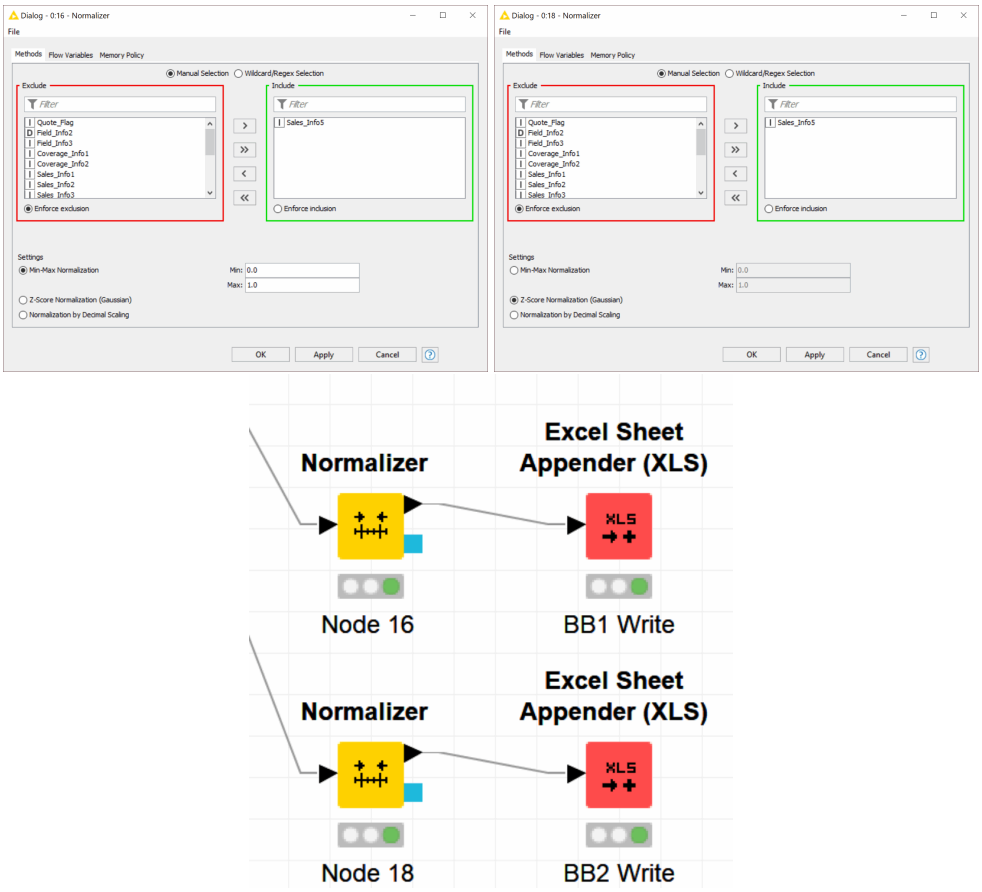

Figure 28: Screenshots Normalization

2.3 Discretization of Coverage_Info1

Since the values in Coverage_Info1 range from 1 to 25, it makes sense to set the range of each category to 6 and the thresholds for the four categories as follows:

$$\begin{aligned} \text{Basic} &=]-\infty, 7[\\ \text{Low} &=]7, 13[\\ \text{Medium} &=]13, 19[\\ \text{High} &=]19, \infty[\end{aligned}$$

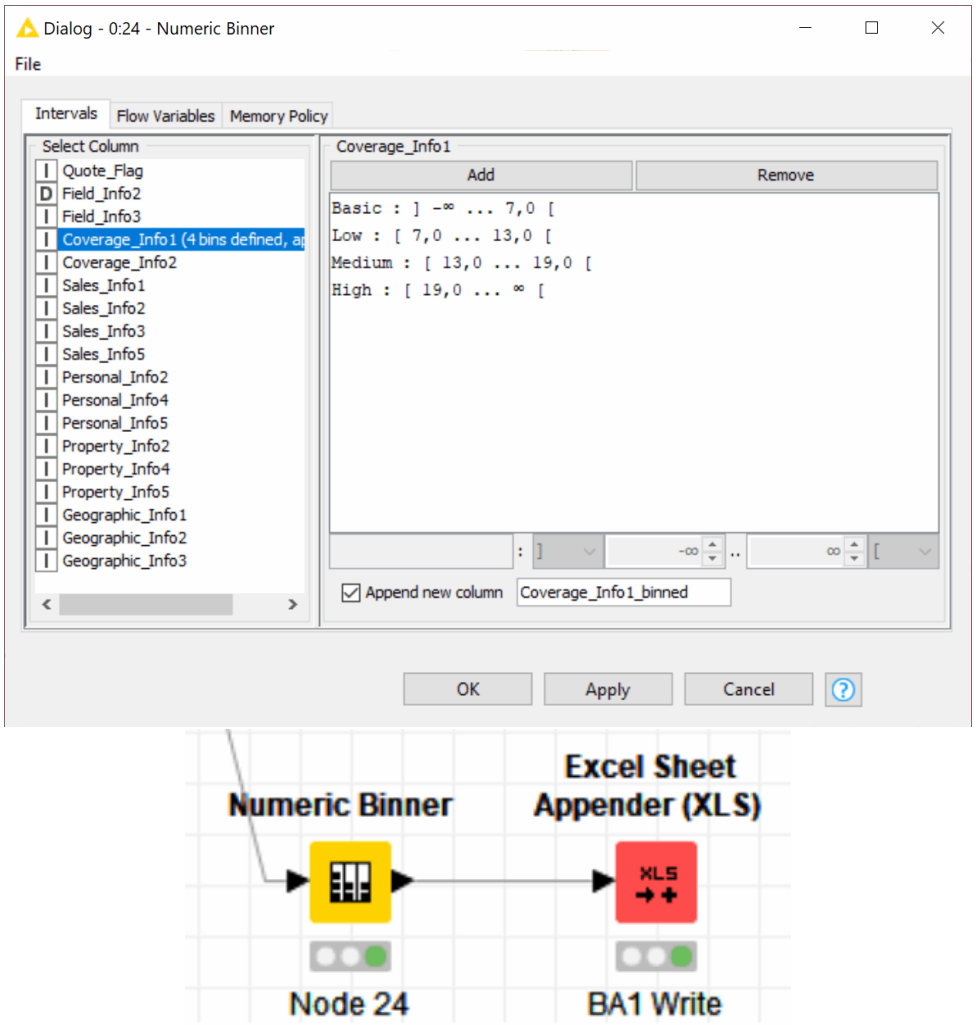

Figure 29: Screenshots Discretization

2.4 Binarization of Geographic_Info5

To binarise the Attribute I used the "One to Many" Node in Knime. The result of the binarization can be found inside the attached excel-workbook.

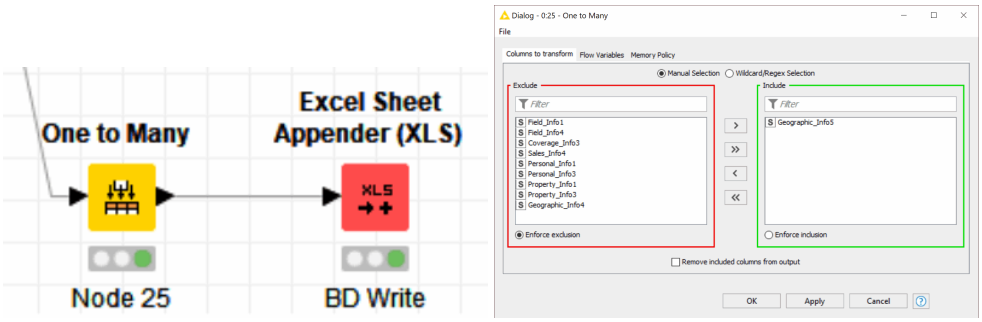

Figure 30: Screenshots Binarization

3 Summary

Field_Info3 All values in this attribute are either on the side or in the middle of the range.

Field_Info4 Normal distribution, but there is a peak at the end.

Coverage_Info2 On this attribute we only have datapoints with four different values allthough the range of those values suggets 25 options have been available.

Geographic_Info2 & Property_Info 5 Both of these values show a even distribution with a spike at the begin of their ranges.

Personal_Info3 Almost 50% of the values are "ZA".

Property_Info2 All values are 0

Personal_Info4 All values are 0 except for one 1

3.0.1 Linear Correlations

There are some interessting correlations in this dataset. Exspecially Geographic_Info5 seems to have some strong correllations with Field_Info2 and Coverage_Info3. Field_Info1 seems to be correlatted to Field_Info4 and Field_Info2 has as negative Correlation with Field_Info3.

List of Figures