Homework 2: Data Pre-processing with Python

Sonam Dawani | 002512473
MS in DS and Analytics - BDML
Georgia State University
Atlanta, Georgia
sdawani1@student.gsu.edu

Task1 – How many attributes do you have in each of the saved files?

Ans – Number of attributes in Quantitative: 9

Number of attributes in Others: 5

Task2 – Create Summary Table in Data Quality Report for Continuous Features. Are the heat maps of the covariance and correlation tables any different? Should they be? Can you tell me about any observations I made about quantitative attributes so far? Which of the generated results helped you made those observations?

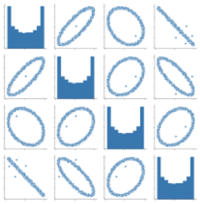
Ans – *Table 1 Data Quality Report for Continuous (at last)*

For histogram number of bins taken is 20. This is because this value is not too small for the given data hence do not result very low resolution. Also the value is not too large which could cause empty bins. So by taking 20 bins we are able see the required observations and pattern.

There is difference in covariance and correlation heat map. Covariance and correlation both are metrics to check the relationship between attributes. If there is absolutely no relation between attributes both sets to value 0. Going forward, both increases with positive relation between attributes and decreases from 0 if there is negative relationship. Still the heat maps for covariance and correlation are different and that's because their magnitude differs. Covariance can take values from -infinity to +infinity, whereas correlation is kind of normalized covariance, hence its value lies in specific range [-1,1] and depicts the extent of relation.

<u>Observation on quantitative attributes</u>:

Attr 8,9,10,11 have almost the same data, but is shuffled in a particular manner. This is evident by few facts. First, the scatter plot have symmetry graphs. Second, they have same range, quartile and standard deviation

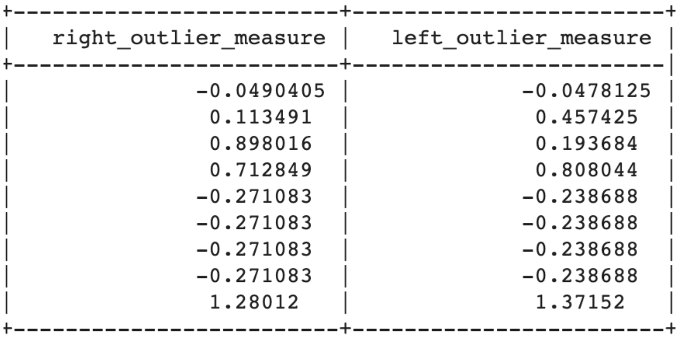


Que 3 – Find outliers (state clearly what method you picked, and provide some short rationale for your choice if possible), and implement clamp transformation on them. Then normalize the data (state clearly what method you picked, and provide your rationale if possible). Now, generate box plots and SPLOMs again and compare with the related results from Task 2. Discuss what you observed, and try to provide explanations for the things you noticed.

Ans – First by 'comparing the gaps' identified if there are any outliers:

If the gap between the 3rd quartile and the maximum value is noticeably larger than the gap between the median and the 3rd quartile, this suggests that the maximum value is unusual and is likely to be an outlier. And similar for the other (left/minimum) side. But again here as we are not able to compare the gaps as attributes are on different scale. Hence tried to normalize the gap by dividing by Inter Quartile Range.

Below is the normalized difference in gaps:

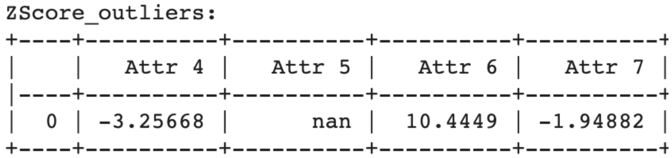


```
+-------------------------+-------------------------+
|   right_outlier_measure |   left_outlier_measure  |
+-------------------------+-------------------------|
|              -0.0490405 |              -0.0478125 |
|               0.113491  |               0.457425  |
|               0.898016  |               0.193684  |
|               0.712849  |               0.808044  |
|              -0.271083  |              -0.238688  |
|              -0.271083  |              -0.238688  |
|              -0.271083  |              -0.238688  |
|              -0.271083  |              -0.238688  |
|               1.28012   |               1.37152   |
+-------------------------+-------------------------+
```

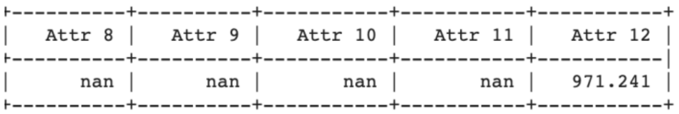
So value 0 suggests no outlier. The more away from the zero the more possibility of outlier. Hence, we notice more number of outliers would be in Attr 12, Attr 6, Attr 7 and Attr 5 respectively.

Next, to identify the outliers value, used ZScore method. As there we are not aware of the domain of the data, the threshold of ZScore allow as to be flexible with the boundary for outliers.

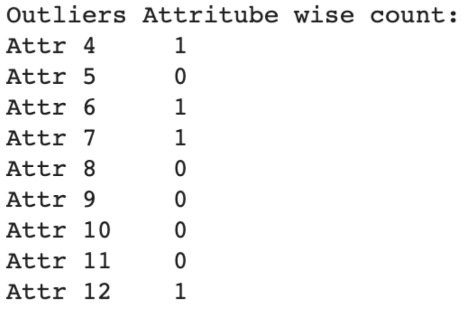
Outliers identified with <u>threshold = 1.92</u>



```
ZScore_outliers:
+----+----------+----------+----------+----------+
|    |   Attr 4 |   Attr 5 |   Attr 6 |   Attr 7 |
|----+----------+----------+----------+----------+
|  0 | -3.25668 |      nan |  10.4449 | -1.94882 |
+----+----------+----------+----------+----------+
```



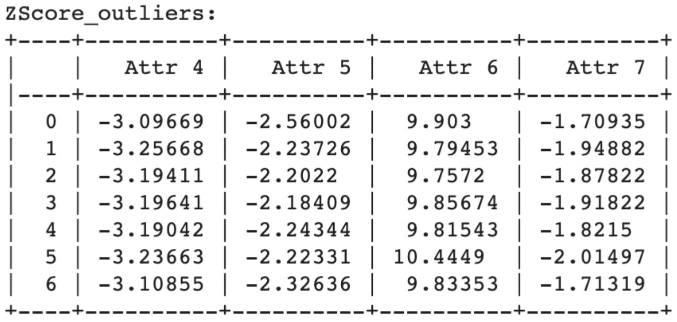
```
+----------+----------+-----------+-----------+-----------+
|   Attr 8 |   Attr 9 |   Attr 10 |   Attr 11 |   Attr 12 |
+----------+----------+-----------+-----------+-----------|
|      nan |      nan |       nan |       nan |   971.241 |
+----------+----------+-----------+-----------+-----------+
```



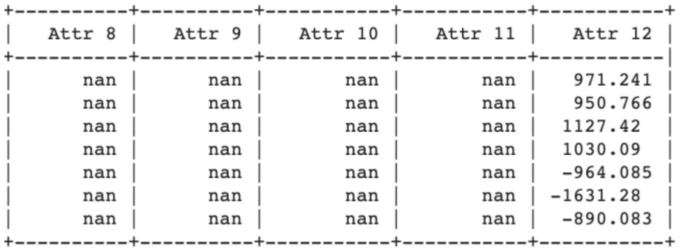
```
Outliers Attritube wise count:
Attr 4      1
Attr 5      0
Attr 6      1
Attr 7      1
Attr 8      0
Attr 9      0
Attr 10     0
Attr 11     0
Attr 12     1
```

Although threshold 1.92 seems to be good, checking-

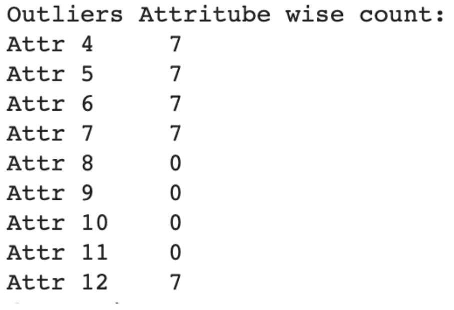
Outliers identified with threshold =1.8



```
ZScore_outliers:
+----+----------+----------+----------+----------+
|    |   Attr 4 |   Attr 5 |   Attr 6 |   Attr 7 |
|----+----------+----------+----------+----------+
|  0 | -3.09669 | -2.56002 |   9.903  | -1.70935 |
|  1 | -3.25668 | -2.23726 |   9.79453 | -1.94882 |
|  2 | -3.19411 | -2.2022  |   9.7572  | -1.87822 |
|  3 | -3.19641 | -2.18409 |   9.85674 | -1.91822 |
|  4 | -3.19042 | -2.24344 |   9.81543 | -1.8215  |
|  5 | -3.23663 | -2.22331 |  10.4449  | -2.01497 |
|  6 | -3.10855 | -2.32636 |   9.83353 | -1.71319 |
+----+----------+----------+----------+----------+
```



```
+----------+----------+-----------+-----------+-----------+
|   Attr 8 |   Attr 9 |   Attr 10 |   Attr 11 |   Attr 12 |
+----------+----------+-----------+-----------+-----------|
|      nan |      nan |       nan |       nan |   971.241 |
|      nan |      nan |       nan |       nan |   950.766 |
|      nan |      nan |       nan |       nan |  1127.42  |
|      nan |      nan |       nan |       nan |  1030.09  |
|      nan |      nan |       nan |       nan |  -964.085 |
|      nan |      nan |       nan |       nan | -1631.28  |
|      nan |      nan |       nan |       nan |  -890.083 |
+----------+----------+-----------+-----------+-----------+
```

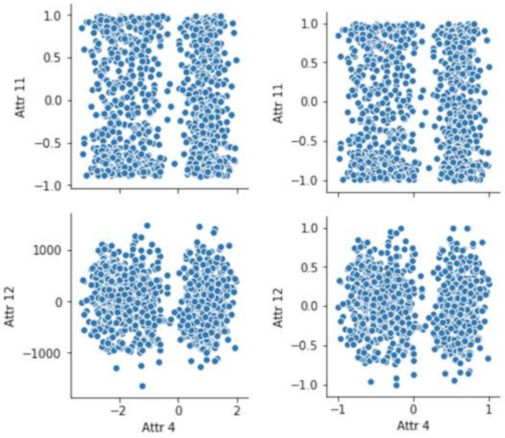


```
Outliers Attritube wise count:
Attr 4      7
Attr 5      7
Attr 6      7
Attr 7      7
Attr 8      0
Attr 9      0
Attr 10     0
Attr 11     0
Attr 12     7
```

For normalization picked 'Range based Normalization'. Range based normalization gets impacted by the outliers but here we have already removed the outliers by clampT . The other method of normalization 'Standardization' is not suitable here as for Standardization the data needs to be normally distributed.

SPLOM comparison:

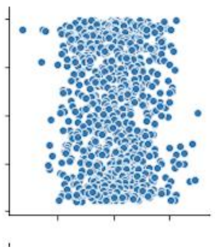
First of all, as the data is normalized we can observe the scale is uniform along all attributes in the later SPLOM.



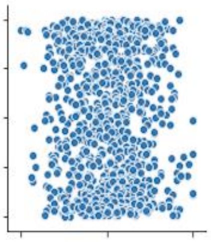
Second, the datapoint spread due to normalization:

Attr 11 VS Attr 12

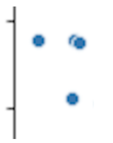
Before



After



Third the outlier removal. Zooming the above:



Que 4 – Summary Table in Data Quality Report for Categorical Features. What type of attributes do you have in this file? What types of scales are applicable to each of them and why?

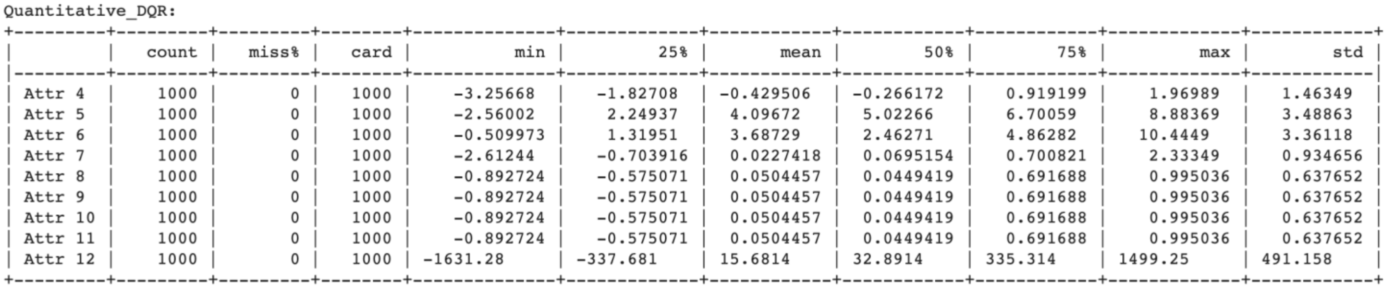
Ans – *Table 2 Data Quality Report for Categorical (at last)*
Attr 0, Attr 1, Attr 2, Labels : Nominal

Attr 3 : Ordinal

Que 5 – How would you compare the original data with the binned version? Discuss what you observed, and try to provide explanations for the things you noticed.

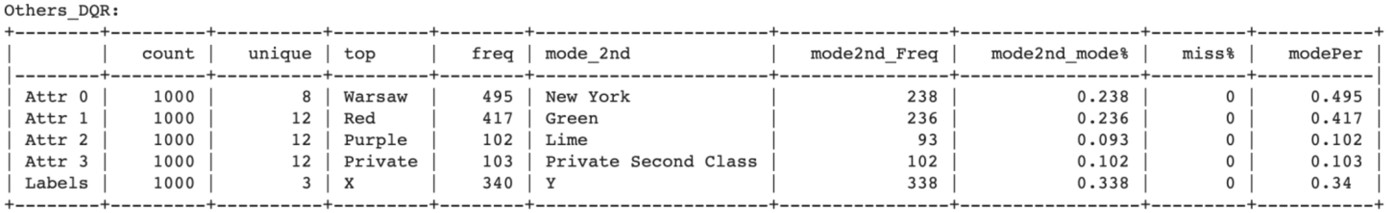
Ans – We can see the range of bin varies, when going along the course of actual sorted data. Sometimes there is only one value covered by a bin. So the values covered by bin ranges from 50 to 1. By this approach we are combing buckets having very few values and splitting the ones having large numbers of value.

Table 1 Data Quality Report for Continuous



```
Quantitative_DQR:
+----------+---------+---------+--------+-------------+------------+------------+------------+------------+-------------+------------+
|          |   count |   miss% |   card |         min |        25% |       mean |        50% |        75% |         max |        std |
|----------+---------+---------+--------+-------------+------------+------------+------------+------------+-------------+------------|
| Attr 4   |    1000 |       0 |   1000 |   -3.25668  |  -1.82708  | -0.429506  | -0.266172  |   0.919199 |    1.96989  |   1.46349  |
| Attr 5   |    1000 |       0 |   1000 |   -2.56002  |   2.24937  |  4.09672   |  5.02266   |   6.70059  |    8.88369  |   3.48863  |
| Attr 6   |    1000 |       0 |   1000 |  -0.509973  |   1.31951  |  3.68729   |  2.46271   |   4.86282  |   10.4449   |   3.36118  |
| Attr 7   |    1000 |       0 |   1000 |   -2.61244  |  -0.703916 |  0.0227418 |  0.0695154 |   0.700821 |    2.33349  |   0.934656 |
| Attr 8   |    1000 |       0 |   1000 |  -0.892724  |  -0.575071 |  0.0504457 |  0.0449419 |   0.691688 |    0.995036 |   0.637652 |
| Attr 9   |    1000 |       0 |   1000 |  -0.892724  |  -0.575071 |  0.0504457 |  0.0449419 |   0.691688 |    0.995036 |   0.637652 |
| Attr 10  |    1000 |       0 |   1000 |  -0.892724  |  -0.575071 |  0.0504457 |  0.0449419 |   0.691688 |    0.995036 |   0.637652 |
| Attr 11  |    1000 |       0 |   1000 |  -0.892724  |  -0.575071 |  0.0504457 |  0.0449419 |   0.691688 |    0.995036 |   0.637652 |
| Attr 12  |    1000 |       0 |   1000 | -1631.28    | -337.681   | 15.6814    | 32.8914    | 335.314    | 1499.25     | 491.158    |
+----------+---------+---------+--------+-------------+------------+------------+------------+------------+-------------+------------+
```

Table 2 Data Quality Report for Categorical



```
Others_DQR:
+--------+---------+----------+---------+--------+----------------------+---------------+----------------+---------+-----------+
|        |   count |   unique | top     |   freq | mode_2nd             |  mode2nd_Freq |  mode2nd_mode% |   miss% |   modePer |
|--------+---------+----------+---------+--------+----------------------+---------------+----------------+---------+-----------|
| Attr 0 |    1000 |        8 | Warsaw  |    495 | New York             |           238 |          0.238 |       0 |     0.495 |
| Attr 1 |    1000 |       12 | Red     |    417 | Green                |           236 |          0.236 |       0 |     0.417 |
| Attr 2 |    1000 |       12 | Purple  |    102 | Lime                 |            93 |          0.093 |       0 |     0.102 |
| Attr 3 |    1000 |       12 | Private |    103 | Private Second Class |           102 |          0.102 |       0 |     0.103 |
| Labels |    1000 |        3 | X       |    340 | Y                    |           338 |          0.338 |       0 |     0.34  |
+--------+---------+----------+---------+--------+----------------------+---------------+----------------+---------+-----------+
```