Data Pre-processing

Name: Li Sun

I. Task 1

There are 9 attributes with data type continuous in Quantitative.csv file and 4 attributes with data type categorical in Others.csv file. The details are shown in Table 1.1.

Table 1.1 Attributes Data Type

Attribute	Data Type	File
Attr 0	Categorical	Others.csv
Attr 1	Categorical	Others.csv
Attr 2	Categorical	Others.csv
Attr 3	Categorical	Others.csv
Attr 4	Continuous	Quantitative.csv
Attr 5	Continuous	Quantitative.csv
Attr 6	Continuous	Quantitative.csv
Attr 7	Continuous	Quantitative.csv
Attr 8	Continuous	Quantitative.csv
Attr 9	Continuous	Quantitative.csv
Attr 10	Continuous	Quantitative.csv
Attr 11	Continuous	Quantitative.csv
Attr 12	Continuous	Quantitative.csv

II. Task 2

A. Data Quality Report

The Summary Table in Data Quality Report for Continuous Feature are shown below. Due to space limitations, the table is split into two parts as table 2.1.

In the summary table, the continuous features include attribute size, missing number, cardinality, mean, standard deviation, variance, min value, 1st quantile value, median, 3rd quantile value and max value.

Table 2.1 Data Quality Report for Continuous Features

Feature	Size	% Miss	Card.	Mean	Std. Dev.	Variance
Attr 4	1000	0	1000	-0.429	1.4634	2.1417
Attr 5	1000	0	1000	4.0967	3.4886	12.1705
Attr 6	1000	0	1000	3.6872	3.3611	11.2975
Attr 7	1000	0	1000	0.0227	0.9346	0.8735
Attr 8	1000	0	1000	0.0504	0.6376	0.4065
Attr 9	1000	0	1000	0.0504	0.6376	0.4065
Attr 10	1000	0	1000	0.0504	0.6376	0.4065
Attr 11	1000	0	1000	0.0504	0.6376	0.4065
Attr 12	1000	0	1000	15.6814	491.1578	241236.0224

Feature	Min	Q1	Median	Q3	Max
Attr 4	-3.256	-1.8270	-0.2661	0.9191	1.9698
Attr 5	-2.5600	2.2493	5.0226	6.7005	8.8836
Attr 6	-0.5099	1.3195	2.4627	4.8628	10.4448
Attr 7	-2.6124	-0.7039	0.0695	0.7008	2.3334
Attr 8	-0.8927	-0.5750	0.0449	0.6916	0.9950
Attr 9	-0.8927	-0.5750	0.0449	0.6916	0.9950
Attr 10	-0.8927	-0.5750	0.0449	0.6916	0.9950
Attr 11	-0.8927	-0.5750	0.0449	0.6916	0.9950
Attr 12	-1631.2834	-337.6814	32.8913	335.3138	1499.2532

B. Equal-width Histogram

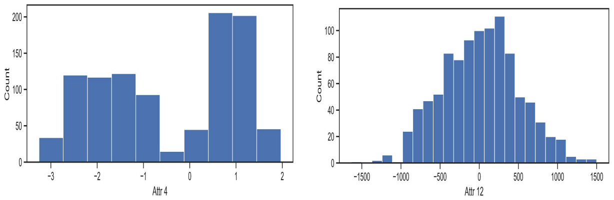
The Freedman–Diaconis rule was used to find the optimal bin number. The bin-width is set to ℏ=2*IQR*n^(−1/3), and the number of bins is (max−min)/ℏ, where n is the number of observations, max and min is the max and min value of an attribute. [1]. The bin numbers for each attribute are listed in the Table 2.2

Table 2.2 Bin Number in Each Attribute

Attribute	Bin Number
Attr 4	10
Attr 5	13
Attr 6	16
Attr 7	18
Attr 8	8
Attr 9	8
Attr 10	8
Attr 11	8
Attr 12	24

Figure 2.1 is an example of equal-width histograms of Attr 4 (top) and Attr 12 (bottom).

Figure 2.1 Equal-width Histograms of Attr 4 (left) and Attr 12 (right)

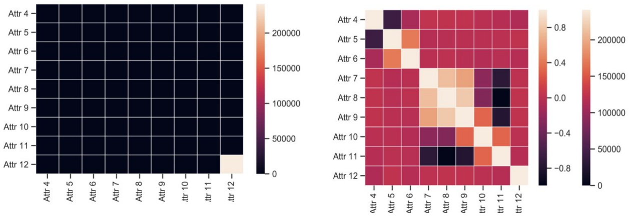


C. Heatmaps analysis

The heatmaps of covariance and correlation table are shown in Figure 2.2.

It is obvious that the heatmaps of covariance and correlation table are different. The reason for making this difference is because the covariance and correlation have different range. As a result, their corresponding heatmap scales are different too. In our case, the range of the variance is [-44.87, 241236.02], and the range of the correlation is [-1, 1].

Figure 2.2 Heatmap of the covariance (left) and correlation (right) tables



In the heatmap of covariance, Attr 12 has extremely huge covariance value which is higher than 200,000, while other attributes has covariance close to zero. That is why there is only one non-black unit block in the entire heatmap. So the heatmap of variance in this case does not provide very few useful information about how the two attributes change together.

On the contrary, the heatmap of correlation illustrates the relationship between attributes efficiently. From observation of the heatmap, we can find that Attr 4 and Attr 5 are negative correlated, all of Attr 7, Attr 8, Attr 9 are strongly negatively correlated to Attr 12. At the same time, Attr 7 and Attr 8, Attr 8 and Attr 9 are positively correlated to each other.

In fact, the correlation is a normalized version of covariance. The correlation range of [-1,1] makes it more interpretable than the unbounded covariance.

D. Observation

All 9 continuous attributes have the same size, and none has missing value and repeated number. The range of majority attributes is less than 15, except Attr 12 whose range is around 3000.

Attr 8, Attr 9, Attr 10 and Attr 11 have exact the same characteristics of features. But their correlation values do not equal to 1, which implies that these 4 attributes have the same data but with different order.

Both equal-width histograms and violin histograms exhibit the same characteristics that Attr 7 and Attr 12 are unimodal distribution, and Attr 4, Attr 5 and Attr 6 are multimodal distribution.

III. Task 3

A. Find outliers

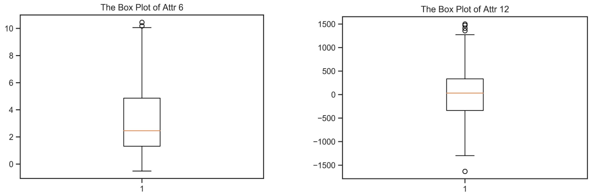
The values located at least 1.5*IQR above Q3, or 1.5*IQR below Q1 are defined as outliers. There are three reasons to utilize the above outlier identification method. First, the majority of the attributes are not normal distributed. Second, the data size is 1000 for each attribute, so data sorting is not expensive in our case. Last, we can conveniently use boxplot to check the identification correction.

Here 7 outliers are identified which are listed in Table 3.1. Their corresponding boxplots are shown in Figure 3.1.

Table 3.1 Outliers

Attribute	Index of Outlier	Value of Outlier
Attr 6	394	10.187496
	879	10.444894
Attr 12	390	1359.629602
	513	1472.698870
	736	1410.167016
	800	-1631.283364
	999	1499.253199

Figure 3.1 The Boxplots of Attr 6 and Attr 12



B. *Normalize the data*

Based on the observation of Task 2, we know that most of the attributes do not follow normal distribution or exponential distribution, and they are not associated with signal processing either, so the Min-Max normalization is selected and performed. In this normalization, all feature values are linearly converted into range [0,1], while all relative differences and proportions between the features are perfectly preserved.

C. *Observation and interpretation*

1. Box plots comparison and interpretation

The ranges of boxplots before and after normalization change. All others are almost the same.

After transformation clamping and normalization of data, the boxplots are implemented. Compared with the plots from Task 2, the shapes of the plots of normalized attributes are keeping similar, except the position of the outliers. The normalized plots for Attr 6 and Attr 8 are displayed in Figure 3.2. The details of the outliers are listed in Table 3.2.

Due to the algorithms of clamp transformation and normalization, all the original outlier values are assigned to 1 or -1. Just like discussed above, Min-Max normalization preserves the relative differences and proportions between features, all these new outlier values are still located beyond the range [Q1-1.5*IQR, Q3+1.5*IQR], and they are outliers again with values 1 or -1.

Figure 3.2 Boxplots of Normalized Attr 6 and Attr12

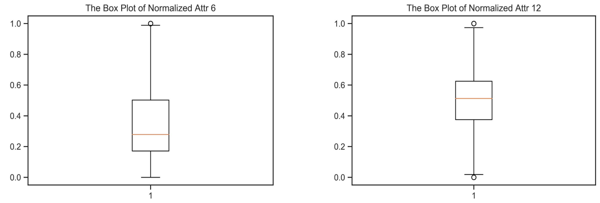


Table 3.2 Outliers

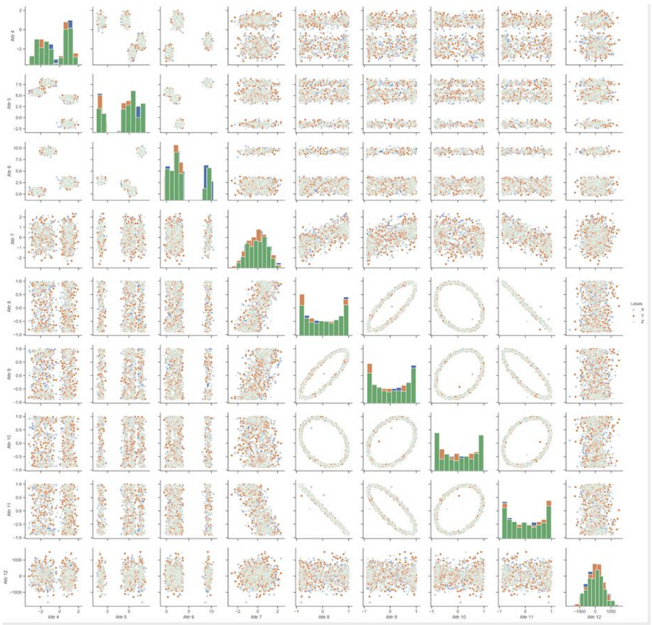
Attribute	Index of Outlier	Value of Outlier
Attr 6	394	1
	879	1
Attr 12	390	1
	513	1
	736	1
	800	-1
	999	1

2. SPLOMs comparison and interpretation

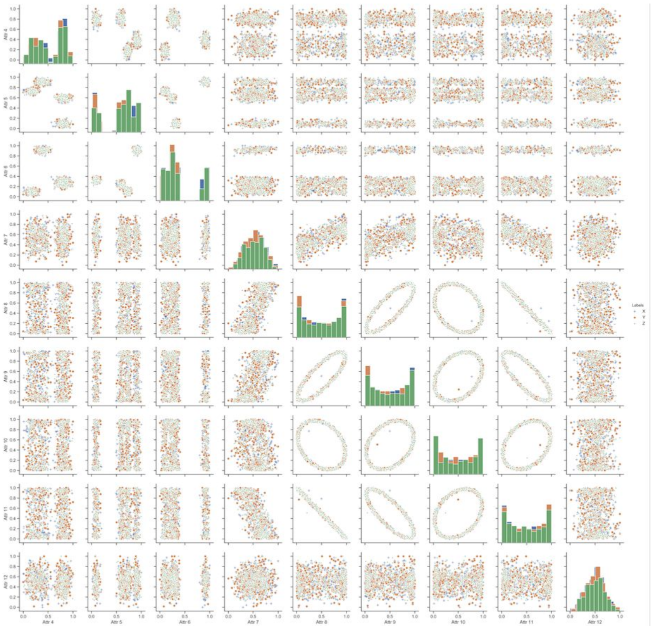
Since the Min-Max normalization performs linear scaling, all differences between values for the features are maintained, the SPLOMS before and after normalization are almost the same. The SPLOMS before and after normalization are displayed in Figure 3.3.

Figure 3.3 SPLOMS before and after data normalization

SPLOM of original data



SPLOM of normalized data



IV. Task 4

A. Data Quality Report

The summary table of Data Quality Report for Categorical Features is listed in Table 4. Due to space limitations, the table is split into two parts as table 4.1.

Table 4.1 Data Quality Report for Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq
Attr 0	1000	0	8	Warsaw	495
Attr 1	1000	0	12	Red	417
Attr 2	1000	0	12	Purple	102
Attr 3	1000	0	12	Private	103

Feature	Mode %	2nd Mode	2nd Mode Freq	2nd Mode %
Attr 0	0.495	New York	238	0.238
Attr 1	0.417	Green	236	0 236
Attr 2	0.102	Lime	93	0.093
Attr 3	0.103	Private Second Class	102	0.102

B. Observations

All the data in this file are categorical. Nominal scale is applicable to Attr 0, Attr 1 and Attr 2, which are city or color labels. They are mutually exclusive and have no numerical significance. Ordinal scale is applicable to Attr 3, which records the military ranks. The values of Attr 3 are rank-ordered with no measurable intervals.

V. Task 5

A. Equal Frequency binning implementation

Equal frequency binning with smoothing by bin mean are implemented. A segment of the final file combining original and binned values are shown in Table 5.

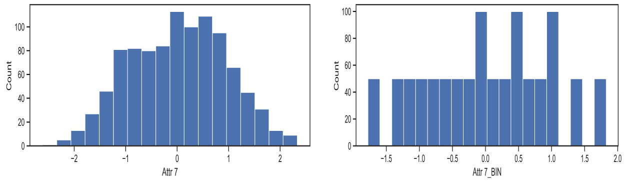
Table 5 Segmented Final File

	Attr 4	Attr 4_BIN		Attr 12	Attr 12_BIN
0	-1.4083953	-1.4539103	…	218.530030	239.885174
1	-1.0046708	-1.0272617	…	597.228645	582.341583
2	-2.6189992	-2.5339306	…	377.771683	376.065812
3	-2.0357268	-1.9500671	…	262.408992	239.885174
4	-1.1491732	-1.0272617	…	-148.911921	-139.849211
…	…	…	…	…	…

B. Comparison with original data and interpretation

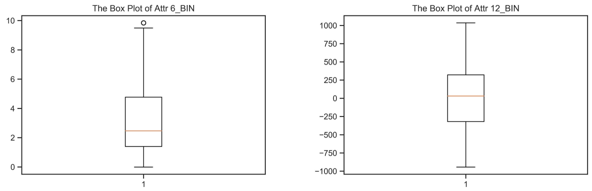
The technique of equal frequency binning with smoothing by bin mean has impacted original data in two ways. First, it works as a discretization method for data reduction. The data quality reports show the cardinality numbers are reduced from 1000 to 20. The reason is that the original data are binned to a fixed 20 with equal value in each bin. Second, It works for data smoothing. The minimum and maximum values in each attribute are smoothed by mean, which leads to a smaller range for each binned attribute. Figure 5.1, an example of equal-width and equal-frequency histograms of Attr 7 , illustrates the smooth effect visually.

Figure 5.1 Equal Width Histogram (left) and Equal Frequency Histogram (right) of Attr 7



Since binning does not use class information and is therefore an unsupervised technique, it is sensitive to the user-specified number of bins, as well as the presence of outliers[2]. In order to examine the influence on outliers, binned Attr 6 and Attr 12 are identified and compared with the original ones. The result is that the outliers are changed dramatically after binning. The boxplots of binned Attr 6 and Attr 12, which are displayed in Figure 5.2, present the changes.

Figure 5.2 Boxplots of Binned Attr 6 (left) and 12(right)



As expected, the 5 outliers in original Attr 12 drop into the regular range, and there is no more outlier in the binned data. However, the outliers in Attr 6 surge up from 2 to 50! The entire 50 instances in the bin with the maximum binned values are outliers now. If looking back to the Figure 3.1, we will find the boxplot of original Attr 6 has a much longer whisker on outlier side, and 2 outliers are relatively far from the median, it is very likely that the outliers will drag its bin value outward and make its entire elements in the bin become outliers.

VI. Reference

[1] Calculating optimal number of bins in a histogram

https://stats.stackexchange.com/questions/798/calculating-optimal-number-of-bins-in-a-histogram

[2] Data Preprocessing

http://mercury.webster.edu/aleshunas/Support%20Materials/Data_preprocessing.pdf