MOHAMMED SHOEBUDDIN HABEEB :: 002521737

HOMEWORK 2 FoDS

TASK 1:

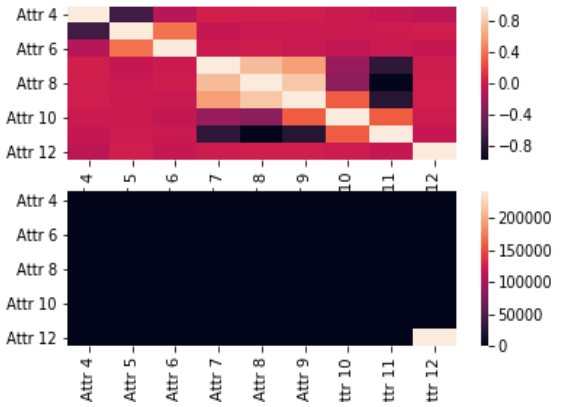
From the given data file, which has 13 columns of Attributes numbered from 0-12 and a nominal called attribute column 'Label.' Going ahead, I split the given ABT in:

- Quantitative Data has **9** columns starting from 'Attr 4' to 'Attr 12', all of which is Ratio Data. This file is then saved as a separate .csv file called *Quantitative.csv*
- Qualitative Data has **5** columns from 'Attr 0' to 'Attr 3' and 'Labels,' all of which is Nominal Data. This is saved in a separate .csv file called *Others.csv*

TASK 2:

	min	max	Std Dev	mean
Attr 4	3.26	1.97	1.46	-0.43
Attr 5	-2.56	8.88	3.49	4.10
Attr 6	-0.51	10.44	3.36	3.69
Attr 7	-2.61	2.33	**0.93**	**0.02**
Attr 8	-0.89	1.00	0.64	0.05
Attr 9	-0.89	1.00	0.64	0.05
Attr 10	-0.89	1.00	0.64	0.05
Attr 11	-0.89	1.00	0.64	0.05
Attr 12	-1631	1499	**491**	**16**

Data Summary Report



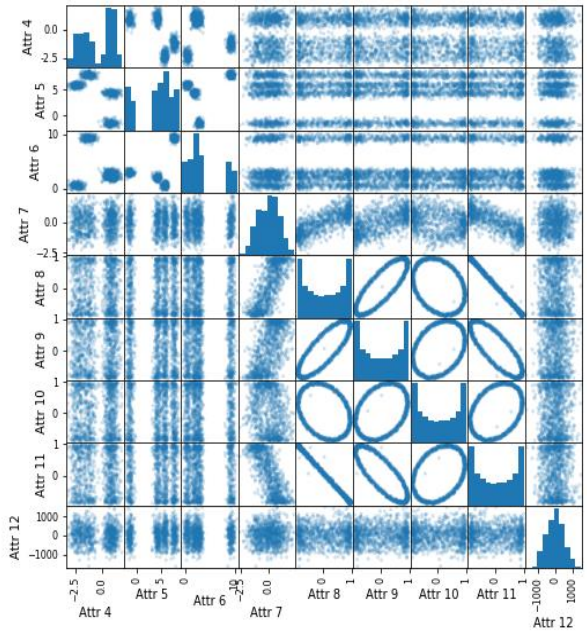
Correlation and Covariance Heat maps

We can see that they are considerably different. We look into heat maps to find a relation between the features, and the higher the values or lighter the cell, which maps two features is the stronger the relationship between the corresponding features. Given the data and features, both covariance and correlation are expected to give a similar trend for a relationship.

The advantage of the Correlation matrix over covariance is the use of scale. As we can see, the covariance matrix is severely affected by Attribute 12. Whereas for the correlation matrix, the scale reduces to [-1,1], which helps gives a better picture and strength of the relationship.

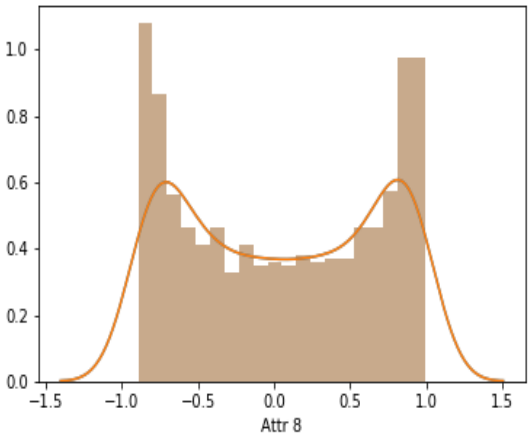
Major Observations :

1. Attribute 4 is a multimodal distribution where the mean and median are at the center of the trench. This majorly implies the presence of different target features as there are 2 big clusters.
2. Attribute 5 and Attribute 6 have major outliers after the 1st standard deviation. It would be preferable to break them into two histograms each. This is majorly from the cluster formed in SPLOMS, which also shows a good correlation.



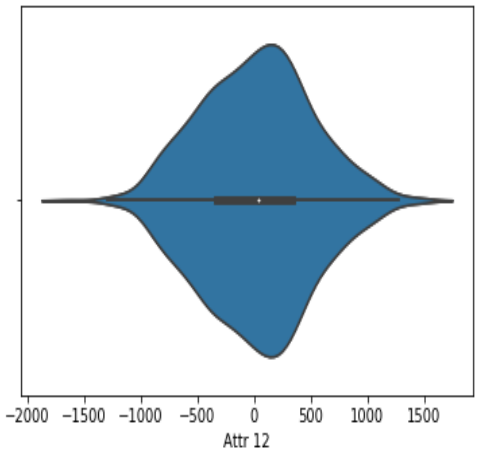
SPLOM for Original Data

3. Attributes 8 to 11 are not normal distributions as both their min and max within the boundaries of even 1st standard deviation. Another interesting trend that is evident from SPLOM is the cyclic nature of relationships amongst attributes 8 – 11. This is only possible if the feature has instances complementing each other, or we can say these features are just rearrangements of each other where the order of instances is changed. Given the cluster is strong bound we can replace all these columns with a single column. That is why all of them are highlighted in a similar color



Histogram for Attribute 8

4. Attribute 12: This is a normal distribution, but the major concern is the presence of outliers. This is inferred from the fact that standard deviation >> mean. From the violin plot, we see that domain(max-min) is too big compared to the Inter Quartile Range. We can try to clamp the values to get 95% values with 2nd standard deviation.



Violin Plot for Attribute 12

TASK 3

CLAMP Decision Making

For the given Data set Features, we have previously inferred the possible presence of outliers in Attributes 5 and 6 as their histograms show diverse multimodal plots. We can split them into 2 histograms, which is outside the scope of this assignment. Another alternative is clamping them right after 1st standard deviation on either side of the mean. But this could prove detrimental as there will be a loss of 32% of data. Apart from these 2 Features, there is a strong indication of the presence of outliers in Attribute 7 and Attribute 12 as the standard deviation is considerably bigger than the mean. We can deal with this by applying clamps after 1.5 IQR either side of our box plot, as recommended by the book. This method is applied to the first 8 Attributes to Attribute 11.

Upper Clamp = 3rd Quartile + IQR

Lower Clamp = 1st Quartile - IQR

Interestingly since the values for Attribute 12 have affected the scale by being extremely high, I have decided to take a different scale for Attribute 12. This method fixes the boundaries at 2 standard deviations on either side. This cannot be generalized and applied to other distributions because firstly, all other features are not normal distributions, and secondly, it gives similar values for either method.

Upper Clamp = Mean + 2*Standard Deviation

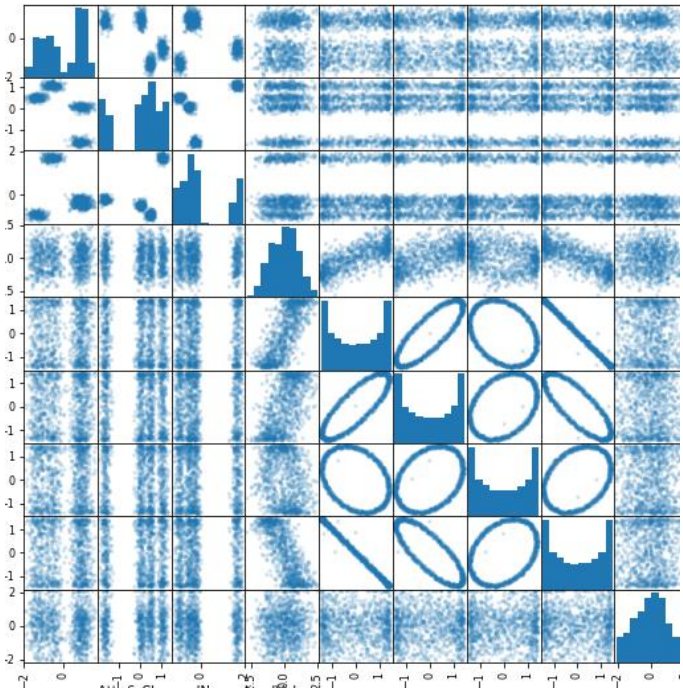
Lower Clamp = Mean – 2* standard deviation

NORMALIZATION and EFFECTS

Going ahead, we apply Z score Normalization for all the attributes to get them to a common scale. Major advantage of this scale is the rigidity it provides for outliers. For the sake of measurement, I am assuming the Attributes 8 -11 to be normal as the Z score application could be a challenge. I could have alternatively chosen Min-Max to accommodate the non normal values but we could as well replace Attr 8-11 with single attribute as they do not give any additional information about the data. The score which we get in the Z score is a major reason why I have stressed on taking this approach. This score gives the value of standard deviations from the mean. This is a better scale as there is little resistance for outliers in a min-max scale. We can compare the result after clamping and normalization below :

Summary Report after Clamping and Normalization

Clamp	Lower	Upper	Normal	max	min
Attr 4	-5.95	5.04	**Attr 4_CNV**	1.64	-1.93
Attr 5	-4.43	13.38	**Attr 5_CNV**	1.37	-1.91
Attr 6	-4.00	10.18	**Attr 6_CNV**	1.93	-1.25
Attr 7	-2.81	2.81	**Attr 7_CNV**	2.47	-2.82
Attr 8	-2.48	2.59	**Attr 8_CNV**	1.48	-1.48
Attr 9	-2.48	2.59	**Attr 9_CNV**	1.48	-1.48
Attr 10	-2.48	2.59	**Attr 10_CNV**	1.48	-1.48
Attr 11	-2.48	2.59	**Attr 11_CNV**	1.48	-1.48
Attr 12	-966.6	998.0	**Attr 12_CNV**	2.07	-2.06



SPLOM after Clamping and Normalizing

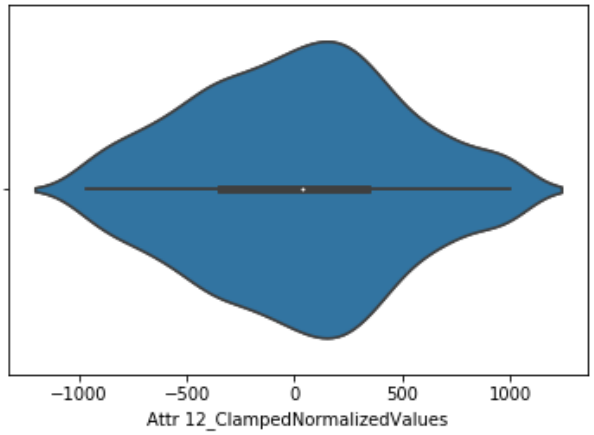
Major Observations :

1. After clustering, we can observe from the SPLOM that most of the attributes are broken down in smaller ranges. Attributes 4 is multimodal and is not much affected, and it gives further optimism to the fact that there are two target features in this Attribute and the distribution is leaning for the same.

2. Attributes 5 and 6 are heavily affected as there is scarce range after 1st standard distribution. This leads to clusters of the relationship instead of a trend in SPLOM and no concrete correlation could be established.

3. Attribute 7-11 were intact because Attribute 7 had a standard normal distribution, and Attribute 8 -11 were not particularly having outliers. The correlation between 8-11 is extremely strong and cyclic which shows that they supplement each other.

4. Attribute 12 was majorly impacted by clamping and normalization because of its huge range. When compared to the previous violin plot before clamping, we can observe a change in range, IQR and height of the density plot.



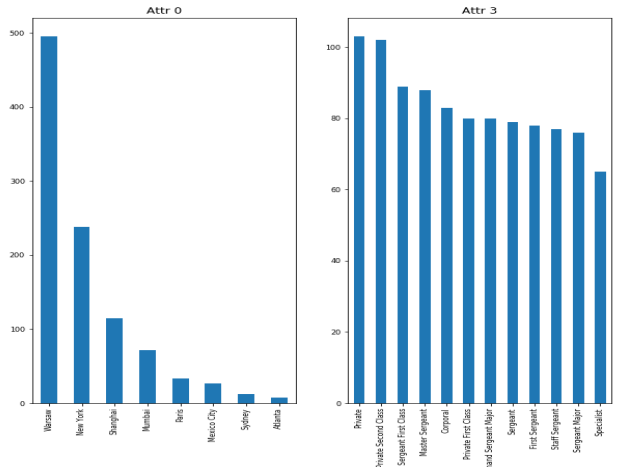
Attribute 12 after clamping and normalizing

Overall, we can conclude that this process has given us a better scale to correlate and compare or features. SPLOM matrix has more scattered features and has distinct boundaries.

TASK 4

Data Summary Report

	Card	Mode	Mode_F	Mode2	Mode2_F
Attr 0	8	Warsaw	495	New York	238
Attr 1	12	Red	417	Green	236
Attr 2	12	Purple	102	Lime	93
Attr 3	12	Private	103	Private Sec	102
Labels	3	X	340	Y	338



` *Ordering Attribute 0 and 3 by the total number of values*

Going ahead we focus on the categorical data. The file which has Qualitative data has 5 columns. The features Attr 0, Attr 1 and Attr 2 have nominal values, whereas Attr 3 has Ordinal values, which most likely represents the rank in a military regiment. Feature 'Labels' is a tertiary scale value which can range from X-Z.

Categorical data is typically visualized using Bar plots which gives insight in the cardinalities, mode, frequency and skewness of our data. We can also used stacked bar plots for representation but the major problem is the absence of binary data.

TASK *5*

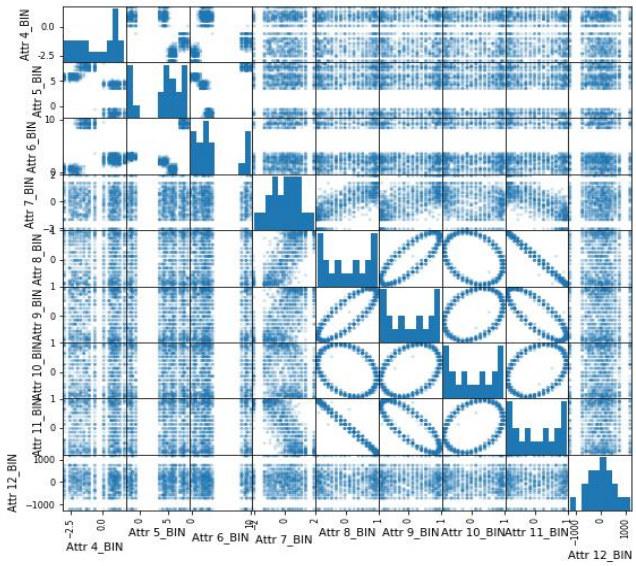
Equal Frequency binning is performed by allocating original data in 20 bins with 50 instances in each bin. For this process we first sort a given attribute and them group 50 instances by calling a function. The major aim here is to cluster similar data to reduce the variance in an attribute. With binning we account for minor errors by replacing the attribute with a central tendency which is called **quantization**. We create new attributes with average of each bin which can be used in the place of original data. If we think about it each bin has its own variance which when put together with other bins reduces the variance of the overall dataset. Below table shows how the min and max changes after binning process.

Change in original values boundaries after binning

	min	max		min	max
Attr 4	-3.26	1.97	**Attr 4_BIN**	-2.97	1.70
Attr 5	-2.56	8.88	**Attr 5_BIN**	-2.21	8.50
Attr 6	-0.51	10.44	**Attr 6_BIN**	-0.15	10.04
Attr 7	-2.61	2.33	**Attr 7_BIN**	-2.04	1.94
Attr 8	-0.89	1.00	**Attr 8_BIN**	-0.87	0.96
Attr 9	-0.89	1.00	**Attr 9_BIN**	-0.87	0.96
Attr 10	-0.89	1.00	**Attr 10_BIN**	-0.87	0.96
Attr 11	-0.89	1.00	**Attr 11_BIN**	-0.87	0.96
Attr 12	-1631	1499	**Attr 12_BIN**	-1216	1172

Binning helps make augmented decisions by deciding a threshold to work on and either deleting or merging bins which are below this threshold. This categorical split of continuous data helps understand the relationships in the dataset. We look into the SPLOM for the same.

If we observe the Scatter Plot Matrix closely we notice that there are clusters of data points together which are more distinct which implies lowering of variance. The bandwidth of the distributions is reduced as we replace all neighborhood values with central values. Also diagonal histograms represent the introduction of categorization error in our continuous values because of the rigid borders we generate. Our goal is to reduce this error to the minimum. We can represent this error using the difference between calculating the slope at that point using limits and differentiation.



SPLOM for Binned Attributes

Going ahead we can split the higher peaks in our values to smaller peaks and vice versa we can merge smaller peaks into a bigger peak.

REFERENCES

1. John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy, "Fundamentals of Machine Learning for Predictive Data Analytics", London.

2. Slides by Dr. Rafal Angryk on Fundamentals of Data Science, Georgia State University

3. Slides by Tan, Steinbach, Kumar on Introduction to Data Mining, Indiana State University

4. Slides by Brendan Blake Camp on Data Mining, Georgia State University

5. A visual introduction to Machine Learning - http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

6. Online resources include Wikipedia and Stack Overflow