*Real Data Processing Analysis Exercise using Python*

Name : Sruthi Pisipati

Program : Masters in Analytics

E-mail: spisipati2@student.gsu.edu

***Task -1) How many attributes do you have in each of the saved files?***

**Answer:** ﻿The number of Attributes in **Quantitative.csv** are **9** while the number of attributes in **Others.csv** are **5**

***Task-2) What information did you generate to 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 you made about quantitative attributes so far? Which of the generated results helped you made those observations?***

**Answer:** For Continuous features, we need the following information to be generated from the data, to be able to create a “Summary table in Data Quality report for Continuous features”.

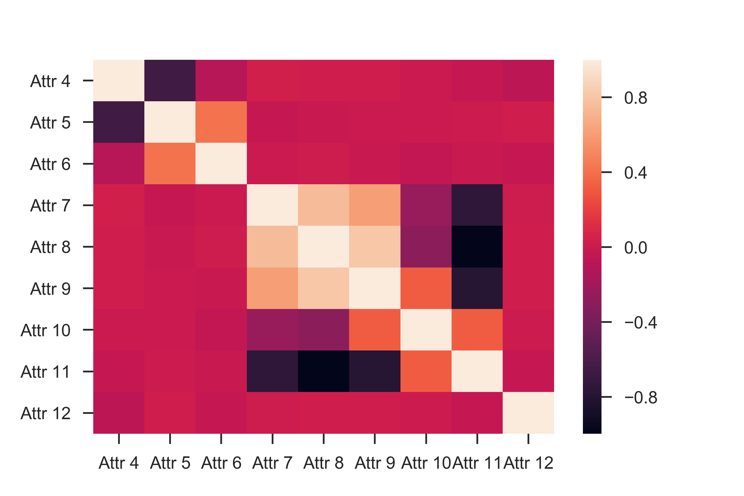
1. ﻿Count of records
2. ﻿Count of missing values
3. ﻿Percentage of missing values
4. ﻿Cardinality count of unique values
5. ﻿Minimum value
6. ﻿1st quarter value:
7. ﻿Median
8. ﻿3rd quarter value
9. ﻿Maximum value
10. ﻿Mean value
11. ﻿Standard deviation
12. ﻿Variance

**Heatmaps of covariance and correlation**

Ideally, the heat maps for covariance and correlation tables must convey the same insights. From the visualization perspective they look little different with respect to the colors, since the range of values is different for covariance & correlation. From the heatmap generated for covariance the insights cannot be clearly known, as the values are generated on a wide range. The values generated for correlation are normalized values of covariance. Hence, better insights are known from correlation heatmap. From the correlation heatmap it can be seen that Attr 4 & Attr 5, Attr 8 & Attr 11 have strong negative correlation. Attr 7 & Attr 8, Attr 7 & Attr 9 are close to having strong positive correlation. Apart from these, there are two clusters formed as shown in the correlation heatmap. All the black colored squares generated in the correlation heatmap, represent a strong negative correlation between the attributes. All the light cream-colored squares generated in the correlation heatmap indicate a strong positive correlation. Dark red colored squares indicate no correlation at all. For example, Attr 7 & Attr 4, Attr 7 & Attr 6 have no correlation.

**A picture containing object

Description automatically generated**

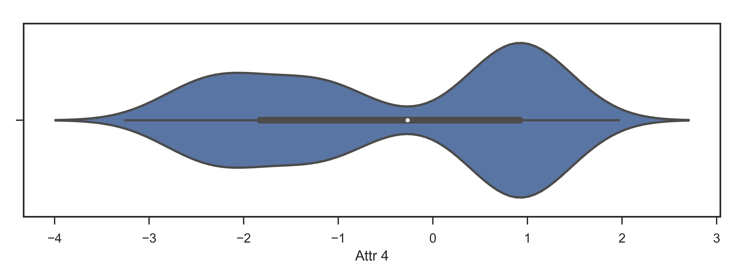
1. *Covariance Heatmap for raw data (Cov\_heatmap\_attributes.png)*

****

1. *Correlation Heatmap for raw data (Cor\_heatmap\_attributes.png)*

**Observations made about quantitative attributes**

From Violin & Histograms generated for quantitative attributes, it is known that Attr 4 has a bimodal distribution. Attr 5 also has a bimodal distribution with almost no data in the gap between the two peaks. Attr 6 follows a similar pattern as Attr 4 but the only difference being the high peak occurs first followed by low peak next. Attr 7 has a unimodal distribution which is evident from the histograms & violin plots. Attr 8, 9, 10 & 11 have two peaks, with data being uniformly distributed in between the two peaks. Attr 12 has a unimodal distribution which is close to normal distribution. All the above observations are made by observing the density of the distribution from violin plots & histograms. Few snippets of the findings are shown as below



1. *Violin plot of Attr4 showing Bimodal distribution.*(*Attr 4\_violin\_plot.png)*

A screenshot of a social media post

Description automatically generated

*Fig.4 Histogram of Attr 5 showing Bimodal distribution.*(*Attr 5\_histogram.png)*

A picture containing outdoor

Description automatically generated

*Fig.5 Violin plot of Attr 6 showing Bimodal distribution.*(*Attr 6\_violin\_plot.png)*

***A screenshot of a cell phone

Description automatically generated***

*Fig.6 Histogram of Attr 12 showing Unimodal distribution close to Normal distribution.*(*Attr 12\_histogram.png)*

***Task-3) Find Outliers - State clearly what method you picked and provide some short rationale for your choice if possible. Normalize the data-State clearly what method you picked and provide your rationale if possible. Discuss what you observed and try to provide explanations for the things you noticed from Box plots and SPLOMs.***

***Answer*:** In this task**,** to find outliers**,** the method of Z-score is chosen. This is because, Z-score method converts the entire data into a normal distribution with mean equal to 0 & standard deviation equal to 1. When all of the data gets converted into a normal distribution with mean equal to 0 and standard deviation equal to 1, it is easy to comprehend insights from the data & visualizing information for each attribute on a single scale varying from -3 to +3. To find outliers, the concept of outliers being 2 standard deviations away from mean is assumed & implemented. Hence, all the data points which are more than +2 & less than -2 (>2 & <-2) are considered as outliers & clamp transformation is done on them by replacing the values greater than +2 by +2 & the values less than -2 by -2. Therefore, minimum threshold here is -2 & maximum threshold is +2.

The next step of normalizing the data after applying the clamp transformation is not necessary here, since the data has already been transformed into a standard normal distribution.

**Observations made from box plots and SPLOMs**

Box plots have been generated both before applying z-score transformation & removal of outliers has been done and after the outliers have been removed as well. From the assumptions of outliers being greater than +2 or less than -2, it has been identified that there is 1 outlier in Attr 6, 32 outliers in Attr 7, 38 outliers in Attr 12. All these outliers are removed & box plots are generated again. It is observed that all the values are converted between the range -2 to +2 in the boxplots & the attributes do not have outliers anymore.

From the SPLOMs generated before & after the outliers removal & normalization, no much difference can be observed. The fact that Attr 8 & Attr 11 have strong negative correlation is reiterated through SPLOMs generated before & after outliers’ removal & normalization by z-score method.

A screenshot of a cell phone

Description automatically generated

*Fig.7 Boxplot of Attr 12 showing outliers before removing outliers Attr 12.png)*

A screenshot of a cell phone

Description automatically generated

*Fig.8 Boxplot of Attr 12 showing no outliers after removing outliers& normalizing data ( Attr 12\_Normalized.png)*

***Task-4) What information did you generate to create Summary table in Data Quality Report for Categorical features? What type of attributes do you have this file – “Others.csv”? What types of scales are applicable to each of them and why?***

**Answer:** For Continuous features, we need the following information to be generated from the data, to be able to create a “Summary table in Data Quality report for Continuous features”.

1. ﻿Count of records
2. ﻿Count of missing values
3. ﻿Percentage of missing values
4. ﻿Cardinality count of unique values
5. Mode value
6. Mode frequency
7. % Mode
8. 2nd Mode
9. 2nd Mode frequency
10. 2nd Mode %

The above information is generated for categorical features. The attributes which are in the file “Others.csv” are categorical features which are not quantifiable, but of type

“String”. The scales which can be applied to categorical feature cannot be mean, median, average etc. which are applied to the quantitative features. For categorical features we tend to know the frequency of each word or string. This can be evaluated by the concept of “Mode”. Mode is the term which is occurs the highest number of times in a dataset. Mode % is the percentage of the frequency with which the most frequently occurring word occurs.2nd Mode is the second highest occurring string in the dataset. 2nd Mode % is the percentage of the frequency with which the second most frequently occurring word occurs.

***Task-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.***

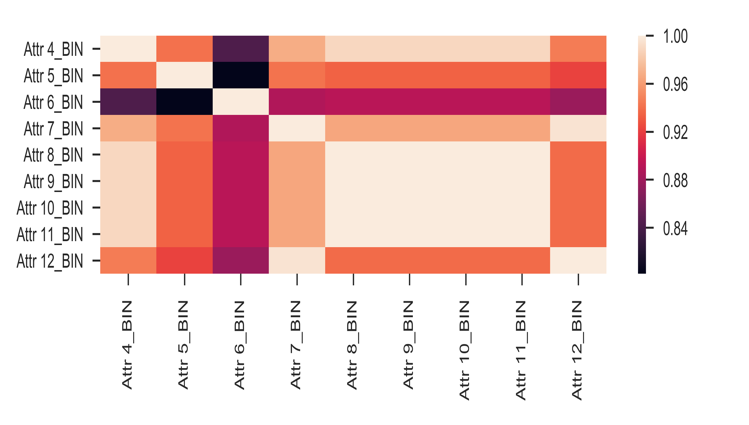
Answer: The original data can be compared with the binned version using the Correlation heatmap. A correlation heatmap has been generated for raw data which has been imported from “Quantitative.csv” & a correlation heatmap has been generated for the same. For creating binned version of data, all the values have been sorted & bins have been created for each attribute containing 1000 rows which gives 20 bins, with each bin size containing 50 values. Each value in a bin has been replaced with average of minimum & maximum value of that particular bin. Now correlation heat map has been generated for the binned version as well.

The insights obtained from “Cor\_heatmap\_Not\_Binned\_attributes.png” are very similar to what has been obtained from Task-2, since it is the same raw data. The observations obtained from “Cor\_heatmap\_Binned\_attributes.png” is different from what has been observed from raw data’s correlation heatmap. Long stripes of colors are seen rather than cluster formations

The following are the correlation heatmap figures which show marked difference between not binned & binned version of data.

A picture containing object

Description automatically generated

*Fig.9. Correlation Heatmap for raw data (Cor\_heatmap\_Binned\_attributes.png*



*Fig.10.Correlation Heatmap for binned attributes (Cor\_heatmap\_Not\_Binned\_attributes.png” ).*

The cause of strips of colors appearing in figure-10 is because of the frequency of same value appearing 50 times. The correlation heatmap in figure 10 shows that there is only positive correlation between the attributes , the negative correlation has been eliminated. While in figure 9 , there are both positive & negative correlation values. In figure-10 , we seen that strong correlation exists between Attr 4\_BIN & Attr 8\_BIN , Attr9\_BIN, Attr10\_BIN & Attr11\_BIN.