**Data Cleaning and Exploration Project**

Written Report

I. BUSINESS UNDERSTANDING We were given a Data frame which consisted of 14 columns and 1000 rows and were asked to explore the data. Python is the primary tool used for the data exploration. This assignment is primarily interested in my ability to understand the data. I’ve broken this report as per the ‘Task #s’ given to us in the problem statement.

# TASK 2

First, we need to develop a data quality report in order to assess the integrity of our data set.

It’s best to group the features based on how to they’re central tendency and variation can be statistically described. In this particular case, our data frame has 12 attributes (features). Of which, 3 are nominal, 2 are ordinal, and 9 are numeric; *please refer to the ‘Attribute Overview table’ in figure 1 for further details.* The ordinal and nominal features were grouped together and saved in a separate file, ‘Others.csv’. The same goes for the numeric features as well but in a separate file named, ‘Quantitative.csv’. In turn, the ‘Quantitative.csv’ file contains 9 attributes while the ‘Others.csv’ file contains 5 attributes.

Based on the summary table, the continuous features have no missing values and no irregular cardinality. The data set may have had outliers but to be safe, I will explore the data first then address these potential outliers afterwards.

After running histograms on the continuous features, I was able to confirm the distributions for each feature. It’s worth mentioning that after iterating through various bin sizes, 8 is the minimum number of bins needed to confirm these feature’s distributions and I used 15 bins per plot in order to better visually illustrate their distributions.

Two of the nine attributes have a uni-modal distribution; ‘Attr 7’ & ‘Attr 12’. Three of the nine attributes have a multi-modal distribution; ‘Attr 4’ to ‘Attr 6’. Four of the nine attributes have a U-shaped distribution; ‘Attr 8’ to ‘Attr 11’.

With the quality report confirming the initial integrity screening for our continuous features, we can move on to the pair-wise analysis where we look for insights on the relationships between these features.

A scatter plot was run on the continuous features because this tool can tell us whether there is a pair wise relationship between the variables? If there are relationships, what is the

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

nature of these relationships? Are there outliers in the data? And is there clustering by groups in the data? [1] The scatterplot showed that there are ten significant relationships amongst our variables. Of the ten, five showed to be a linear relationship. Of the five, only one showed a strong negative relationship. The other four showed to have a weak positive linear relationship. The five non-linear relationships have their data points related in a circular fashion.

The scatter plots showed that attributes 4, 5, & 6 have group clusters. Also, 6 of the plots in the scatter plot matrix showed

to have outliers. (see figure 2 for full recap on these finds)

From here, the covariance and correlation values were calculated for each variable relation and tabulated. These values were then visually represented via a Heat map.

The heat maps of the covariance and correlation values are different because the calculation for the covariance is heavily affected by potential outliers. In this case, attribute 12 showed to have a large covariance and in turn had the color spectrum have a large range. This impact would dwarf one’s ability to gage variation amongst covariance values visually on a heat map. If the outliers were taken out, we would see the heat maps being directly related to each other.

With that said, if we look at the tabulated covariance and correlation values directly and cross reference these with our observations made visually in the scatterplots, we’ll see that the correlation and covariance values reinforce our predictions on the relationships I had identified.

What is interesting though is that attribute 4/5 and attribute 5/6 showed to have a significant correlation as per there covariance and correlation values but after referencing their scatter plots you’d see that these features have clusters which don’t have a linear relationship. We can conclude that our correlation and covariance values were able to confirm that these clusters are linearly related, this is something that would be difficult to confirm had I just relied on the scatter plot matrices.

Attribute 12 actually shows to have high covariance values for every attribute it’s measured against, however, the associated correlation values show to have no correlation. Thus, it can be inferred that this attribute may have an outlier or the range for this attribute is much larger than the other attributes. In order to combat this, we should remove any outliers and normalize the data. Then re-preform the pair-wise analysis.

# TASK 3

First, I choose to use to the IQR method because this method can be universally used on all of our distributions. With that said, it may have been more optimal to use the STD method on some distributions but given our application the IQR method worked well enough to generate the insights within the dataset. After running this analysis and as just as I suspected, Attribute 12 had 5 outliers which were removed.

Second, I choose to use the min-max method for normalizing the filtered data frame because it’s a reliable method given that we now know our minimum and maximum values for each attribute aren’t affected by any outliers.

After which, I re-preformed the pair-wise analysis and found that the heat maps for the covariance values and the correlation values were directly related. I did notice that the covariance values were closer to zero but given the heat map is similar to the correlation heat map, the differentiation amongst these covariances is what we’re looking for. With this said, my normalized heat map results do agree with the relations I was identifying with the original data set.

# TASK 4

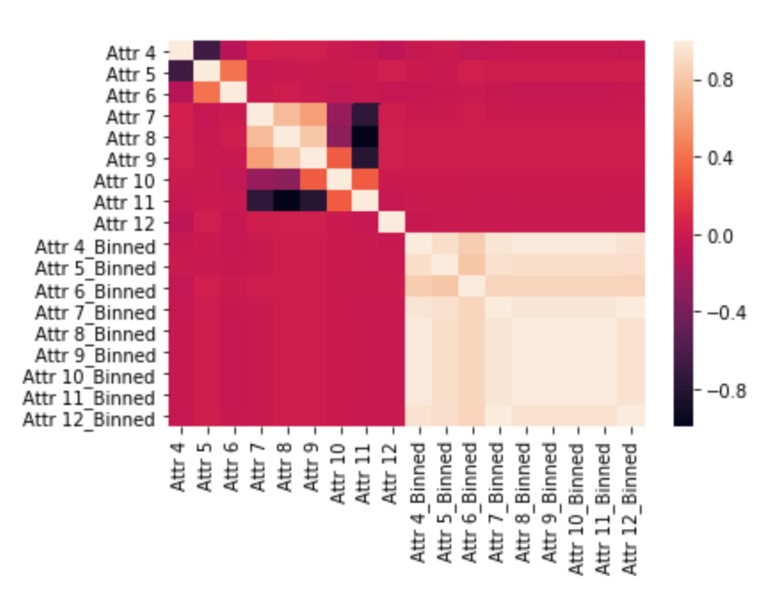
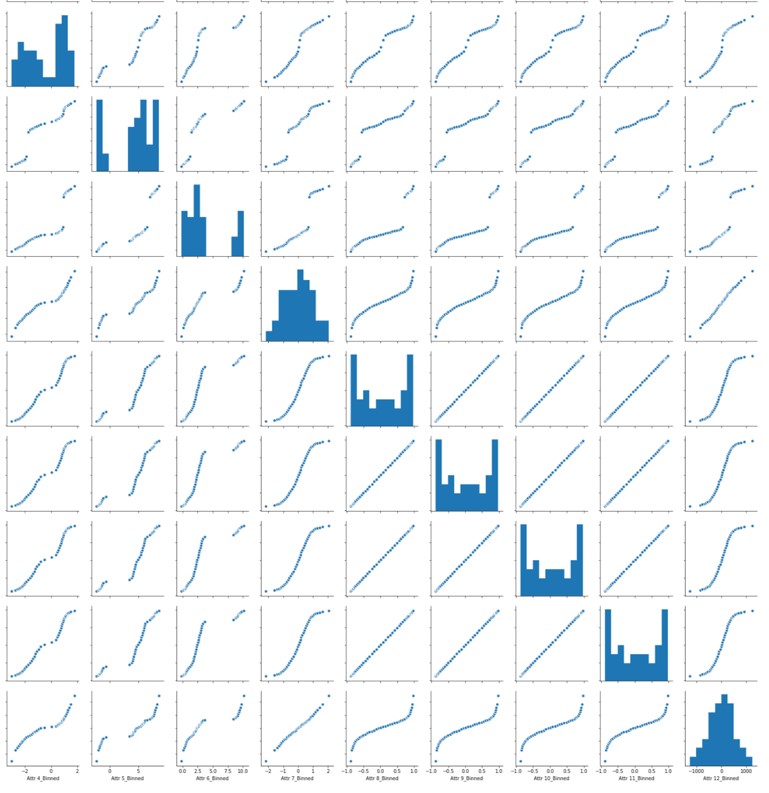
Moving on to the categorical features; to reiterate, there are five attributes. Three of which are nominal and two are ordinal. ‘Attr 0’ is nominal because it is an attribute consistent of city names which cannot be ranked. ‘Attr 1’ and ‘Attr 2’ are nominal as well because they are consistent of names of colors which also cannot be ranked. ‘Attr 3’ is ordinal because it consists of military ranks which can be ordered. ‘Attr 4’ can be interpreted as ordinal or nominal, it’s classification would depend on the business understanding of this attribute. But for this activity, I assumed it to be ordinal because it consisted of the letters X/Y/Z, which can be ordered as per the English alphabet.

# TASK 5

After generating the binned values for each continuous feature, I believe that the best way to compare this to the original data would be to overlay the SPLOM for the binned results on top of the SPLOM of the original data. [2] Reason being is that the binning allows us to smoothen the data. There are potentially an infinite number of variables that could be affecting the variation in the outcome of variable Y given variable X. Under this pretense, binning will create a scenario where we can separate signal from noise and potentially see the underlying relationship variable X may have on variable Y. In this particular case, identify if these features do, in fact, share a linear, quadratic, or maybe even an exponential relationship. If we overlay these two datasets on a scatter plot, we can visually/quickly see what these two variable’s relations may be.

We could even use this as an initial screening for identifying potential leads on data relations that may have a lot of noise. From here, we could run the covariance and correlation values for these specific plots and conclude on whether the variables share an underlaying relation or not.

For example, if you look at figure 1 you can see how the covariance and correlation values progressed as we ran through the data transformations. You can clearly see this additional binning analysis allowed us to potentially identify the underlaying relationship that I had originally concluded on. What’s interesting though that the regardless of the original relation being positive or negative, the binned analysis shows a positive relationship. I believe that perhaps this method is over fitting the relationship amongst the variables and in turn showing a positive relationship when in fact, there may be a negative relationship. Please see figures below.

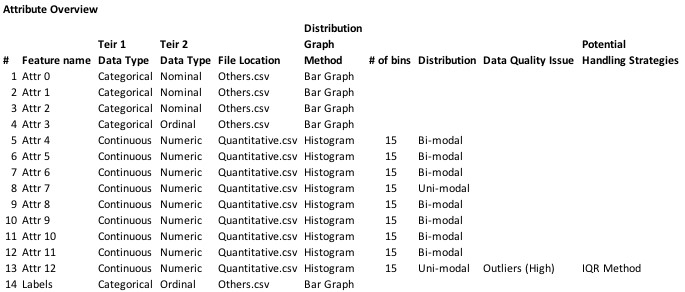


*References*

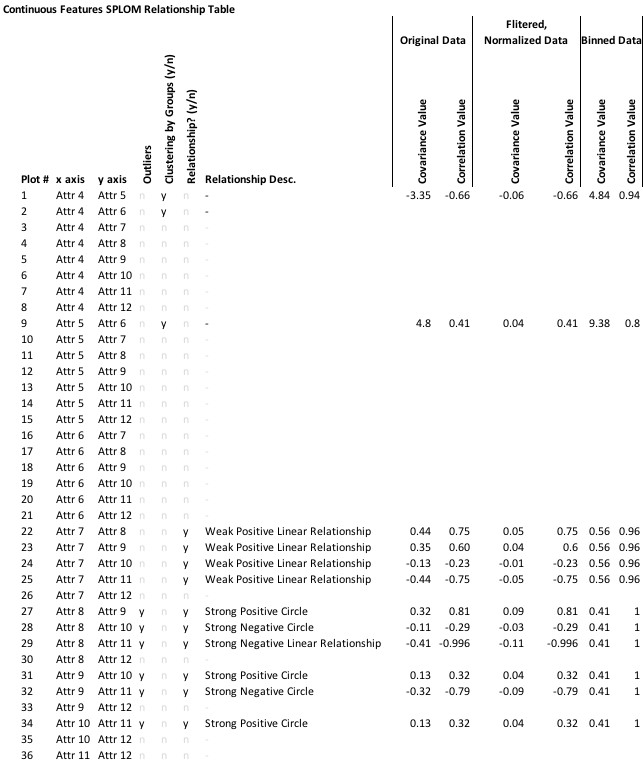
|  |  |  |
| --- | --- | --- |
| “Scatter Plot Matrix.” *1.3.3.26.11. Scatter Plot Matrix*, | | NIST, |
| www.itl.nist.gov/div898/handbook/eda/section3/scatplma.htm. |  |  |

[1]

[2] http://www.adass2014.org/presentations/P2-14.pdf



**Figure 1**



**Figure 2**