Real Data Pre-processing with Python

1. Task 1: For the given data file from our course website (dataPreP.csv): (1) open the file and read it to a Panda’s data frame, (2) identify attributes for which ALL quantitative scales from our lectures are applicable and (3) save it as Quantitative.csv file, then (4) save all remaining attributes in the Others.csv file. **How many attributes do you have in each of the saved files?**
2. Task 2: Read the data from your Quantitative.csv file to a new Pandas data frame and generate all info that you may need to create **Summary Table in Data Quality Report for Continuous Features**. Then show equal-width histograms (make sure to pick the right number of bins to show accurately the data distribution), and horizontal violin plots. Then move on to pair-wise analyses: show scatter plot matrix, covariance and correlation tables, and heat maps for both of these tables. **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?**
3. Task 3: For the data from the Quantitative.csv file – 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**. Save the results of your data transformations in a new QTransferred.csv file. In this file you need to store 2 new columns per each original attribute that you processed. The 1st column should be called OriginalAtrributeName\_ClampedValues, and 2nd column should be named OriginalAtrributeName\_ClampedNormalizedValues. Make sure to process all attributes from the Quantitative.csv file in this way.

1. Task 4: Read the data from Others.csv file to Panda’s data frame and generate all info that you may need to create **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?**
2. Task 5: Implement Equal-Frequency Binning on your own (you can use any sorting functions that are available in Python). For each of the attributes in the Quantitative.csv file perform your Equal- Frequency Binning (assume I want 50 instances in each bin) and generate a new attribute, that contains new values. These values need to be equal to the averages of the bins (i.e. (min. data value in the bin + max. data value in the bin)/2), to which the particular instances belongs to. Generate new file, QuantitativeBinned.csv, which will contain 2 times more columns than the Quantitative.csv, stored in the order where each original data column is followed by the column with the binned values. The names of the columns with binned values must had “\_BIN” suffix added at the end of each of them. **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**.