Applying K-Means and Divisive Hierarchical Clustering Algorithms to Hepatitis C Data

Joe Griffin
josephgriffin@lewisu.edu

Alex Biskis
alexandrasbiskis@lewisu.edu

Victoria Griffin
victoriakgriffin@lewisu.edu

DATA-51000-002, Fall 1
Data Mining and Analytics
Lewis University

I. INTRODUCTION

In the course of data mining education, it is necessary to learn the most common algorithms used in the profession. These algorithms aid in organizing data, solving complex problems, and in the end, attaining the answers that are sought after. One such category, called clustering, and the steps taken to prepare the data for clustering, will be discussed throughout the entirety of this analysis.

For this paper, a dataset from the website Kaggle.com. With the parameter that was given stating "a larger set with a decent number of attributes" and through confirmation by the professor, the group was able to narrow the search down to an interesting topic pertaining to Hepatitis C in blood donations. This dataset was chosen for its size, interesting attributes, and origin; UC Irivine's machine learning repository, which is sponsored by the National Scientific Foundation. It is also relevant to mention that the group members are focusing on Bioinformatics/Computational Biology as a concentration and felt most comfortable with this particular data.

As a concept, clustering can aid in the process of grouping similar observations within a dataset that has many dimensions (i.e. variables/attributes/features). It is important to note that as the number of dimensions grows this may result in a significant increase in clustering runtime, decrease in overall algorithmic efficiency, observations looking equidistant, and patterns within the dataset becoming increasingly more difficult to detect. As such, methodologies that are geared towards reducing the dimensionality of a dataset, while still retaining a majority of the relevant information in the variable, may be needed to carry out the clustering process . There are many types of clustering algorithms available. For simplicity, the two clustering algorithms used in this analysis were K-means and DBSCAN. These two algorithms are exceptional when utilized with large data sets and are very commonly used for a variety of analyses in the field.

The Hepatitis C prediction data set was ultimately chosen for its potential in real-world applications. With the possibility of finding novel patterns and associations, this work could be used at blood banks for quickly finding patients that may not have realized they had Hepatitis C. Hepatitis C is a viral infection in the liver that is caused by the hepatitis C virus (HCV). The infection occurs from blood-to-blood contact with an infected individual. The infection can either be acute or chronic. All infections are initially classified as acute. Sometimes the human body is able to naturally clear the infection. Other times, the virus can mutate and evade the human immune response. If the virus manages to remain in the human body for more than 6 months, the case is then referred to as a chronic infection. In 2019, there were an estimated 57,500 acute cases of Hepatitis C in the US. Many people do not know they have a Hepatitis C infection as it is common to not display any symptoms. Between the years of 2013-2016, it is estimated that there were 2.4 million people in the US living with Hepatitis C. Lastly, in 2018, over 15,000 death certificates listed Hepatitis C as either the primary or contributing cause of death. Due to many people not knowing they have the infection, this number is estimated to be quite conservative (CDC).

The future sections of this report cover different areas of the analysis. Within section II, the dataset is described more in detail with information ranging from the types of variables, total missing values observed, distributions of variables, and so on. Section III explains all the steps taken during the analysis and the tools utilized to carry it out. In section IV the group dives into the results from the clustering algorithms and discusses the findings. Section V wraps up the analysis with final thoughts and future recommendations/implications.

II. DATA DESCRIPTION

The size of the Hepatitis C prediction data set is approximately 46 kilobytes. The dataset is relatively new, with a collection date of June 2020. The data contains exactly 615 rows (615 blood donors) and 14 different attributes, for a total of 8,610 data points . Of those 14 attributes, 13 are numerical and the last is a string as seen in Table 1. Overall, there are 31 data points, or 0.36% of all data points, that are missing. The 31 data points that are missing are spread across 26 different rows, or 4.23% of all rows, and the total percent missing per variable, along with other statistical descriptions, are summarized in Table 2 in Appendix A. Variables 'Feature 1' and 'Category' were excluded from the clustering analysis. The variable distributions are summarized in Figure 1 in Appendix A with histograms and QQ plots. Although all variables were not normally distributed, the top six variables that were seen to have the highest skewness were summarized.

Table 1: Summary of Hepatitis C Dataset

Attribute	Type	Example Value	Description
Feature 1	Nominal (primary key)	513	Record identifier
Category	Ordinal (0, 1, 2, 3)	0	0=Blood donor, 1= Hepatitis, 2=Fibrosis, 3=Cirrhosis
Age	Numeric (integer)	23	Reported age of donor
Sex	Binary (string)	m or f	m=male, f=female
ALB	Numeric (real)	23.0	Level of albumin in blood
ALP	Numeric (real)	102.9	Level of alkaline phosphatase in blood
ALT	Numeric (real)	3.7	Level of alanine transaminase in blood
AST	Numeric (real)	95.4	Level of aspartate aminotransferase in blood
BIL	Numeric (real)	30.0	Level of bilirubin in blood
CHE	Numeric (real)	9.64	Level of cholinesterase in blood
CHOL	Numeric (real)	4.62	Level of cholesterol in blood
CREA	Numeric (real)	60.5	Level of creatinine in blood
GGT	Numeric (real)	169.8	Level of gamma-glutamyl transferase in blood
PROT	Numeric (real)	72.3	Level of protein in blood

III. METHODOLOGY

Initial dataset investigation was the preliminary step taken to understand what kind of dataset was at hand. Python version 3.9.8 was utilized through the analysis, along with various libraries such as Pandas version 1.5.0, Numpy version 1.23.3, Matplotlib version 3.6.0, Statsmodels version 0.14.0, and Scikit Learn (sklearn) version 1.1.2. The Pandas data analysis library was utilized to gain a surface level background of the data types and amounts of missing values present in the dataset. Then, Pandas was used once again for a statistical description that included mean, standard deviation, min, max, and quartiles to obtain a high level overview of the dataset. From there, numerical variables were assessed visually for distributions through the implementation of Matplotlib histograms and Statsmodels QQ plots.

Dataset cleaning and preparation was the next step in the analysis. The variable “Feature 1” was removed through Pandas functionality due to it being redundant indexing information. The dataset variable “Category”, indicating whether or not the subject has Hepatitis C, was removed in a similar fashion as clustering is mainly used in data analytics to uncover groups that are not already known. Next, using Pandas, the data column indicating the sex of subjects was encoded with the numerical values ‘1’ and ‘2’ for ‘m’ and ‘f’ respectively.

As mentioned, a small subset of the dataset contained missing values which poses an issue to the clustering algorithms. To combat this, three separate methodologies were investigated to see how well each would work. The first method used was exploring the individual variables to see how much each contributed to the total missing values. The general rule is to remove variables that are missing 10% or more of their respective data, however in this case the caveat was not satisfied across the board, so this method was abandoned. The second method explored was removing the individual observations that had at least one missing value. Using Pandas, a total of 26 rows, or 4.23% of total rows, were removed. The last method investigated was imputing missing values with the respective variable mean that the missing value was located in using the Sklearn SimpleImputer function. Finally, variables were normalized for both of the missing value corrected datasets using the Sklearn StandardScaler function.

Given the large number of features in the dataset, principal component analysis (PCA) was applied through the Sklearn PCA function to reduce overall dataset dimensionality. Scree plots were generated to identify the proportion of variance explained by each principal component using the Sklearn PCA function and the Matplotlib library. Prior to applying the clustering algorithms, the ‘Elbow Method’ and Silhouette scores were utilized to identify how many clusters would be used in each respective algorithm. Finally, the clustering algorithms were applied using the KMeans and DBSCAN functions from the Sklearn library. A methods schema is seen in Fig 2.

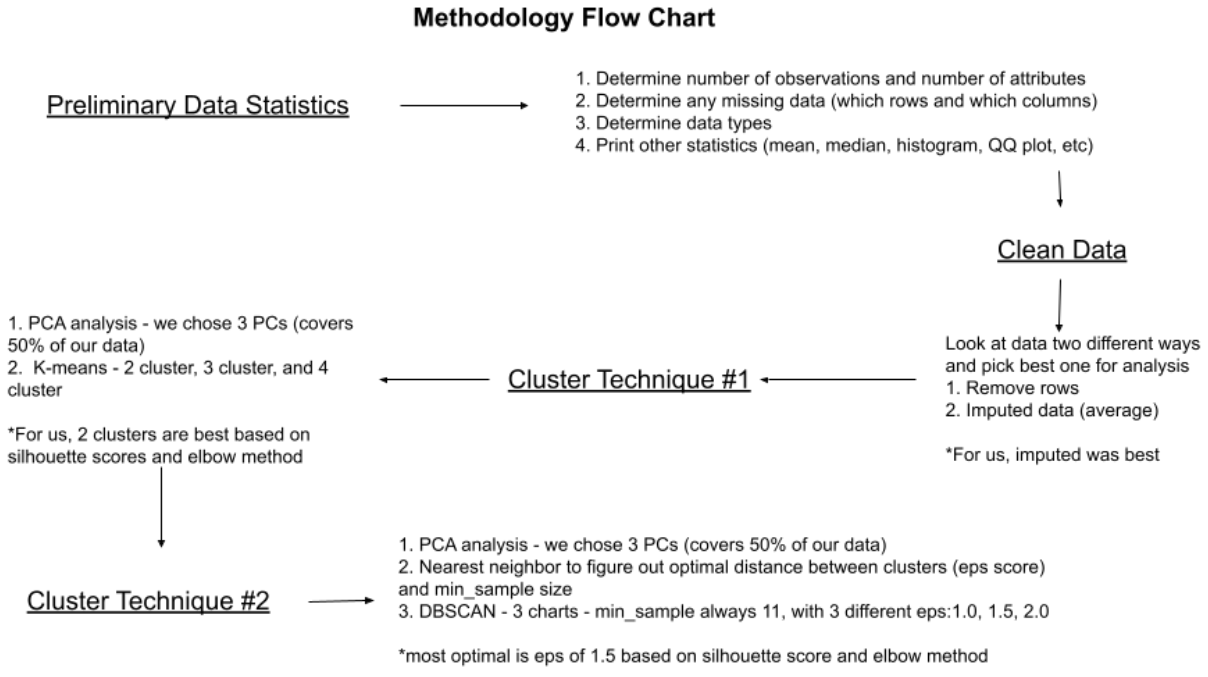

Figure 2: Methods Schema Layout

IV. RESULTS AND DISCUSSION

The values from the PCA Scree plots were nearly identical for both of the modified datasets. As seen in Fig. 3, principal components one to ten from both varied datasets had almost identical amounts of explained variance at each respective principal component. For that reason, the 'Removed Rows Dataset' was abandoned, and all clustering analyses were performed on the 'Imputed Dataset' due to the fact that it had more observations. After careful analysis, it was decided that the first three Principal Components would be used, which accounts for approximately 50% of explained variance in the data, for visualization purposes. We didn't want to use more than three, as we felt it necessary to minimize our dimensions to a reasonable amount.

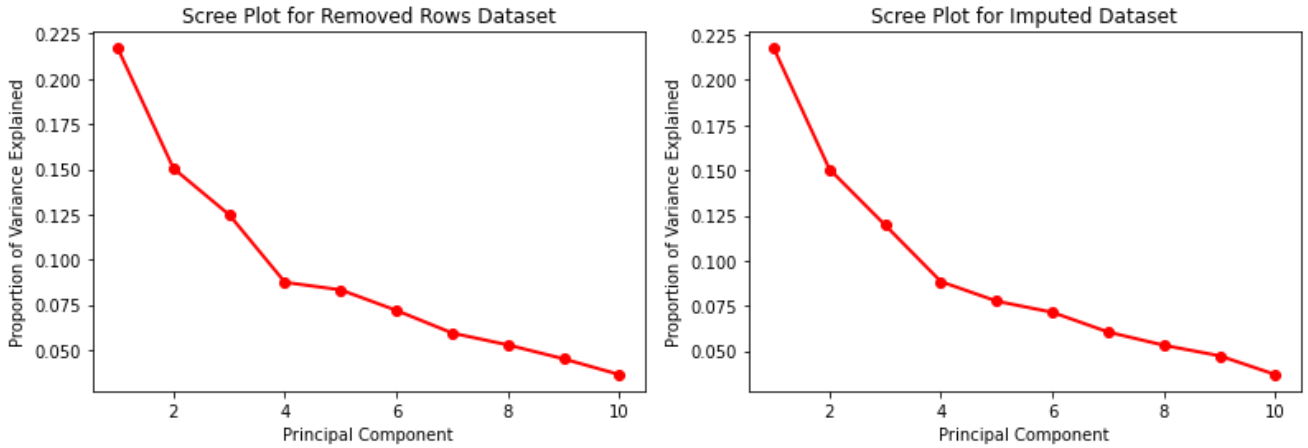

Fig 3: Scree Plots results of Principal Component Analysis for the two varied datasets.

Understanding how many clusters should be used in clustering algorithms is a crucial step to an efficient analysis. Specifically, KMeans has the option to select the number of clusters that one thinks would be most suitable. To find the most optimal number of K clusters for KMeans, methods such as the 'Elbow Method' as seen in Fig. 4, and Silhouette scores, as seen in Table 3.

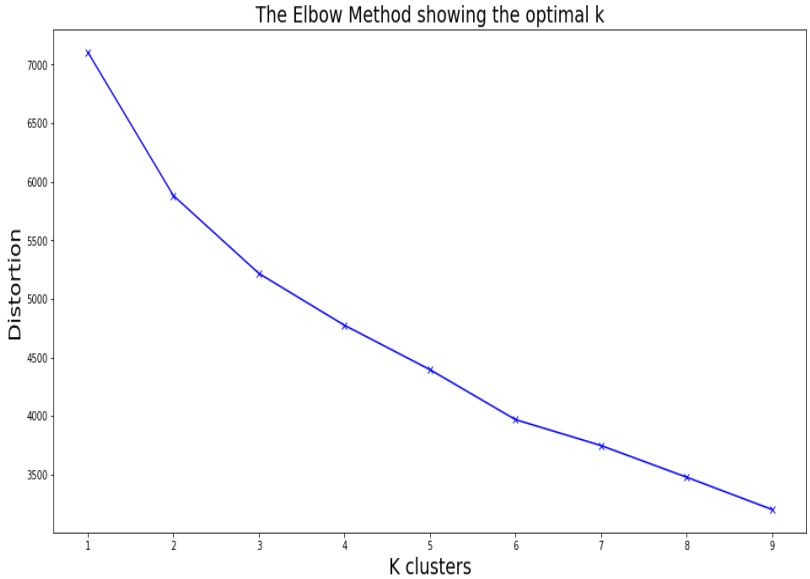

Fig 4: Elbow Method Plot showing the optimal K value for KMeans

Table 3: Silhouette Scores for KMeans

Number of Clusters	Silhouette Score
2	0.702
3	0.625
4	0.589

It was found that two clusters, with a silhouette score of 0.702, was the most optimal. Nevertheless, the KMeans algorithm was applied with two, three, and four clusters to visualize the differences as seen in Fig. 5. These two clusters are significant when it comes to the breakdown of the data. Within Fig. 5, plot A, the blue cluster represents patients that do not have Hepatitis C, while the green cluster signifies patients with Hepatitis C. The reasoning behind this is because within a blood donation clinic, it is expected to see more healthy individuals than not, which is represented by the highly dense blue cluster.

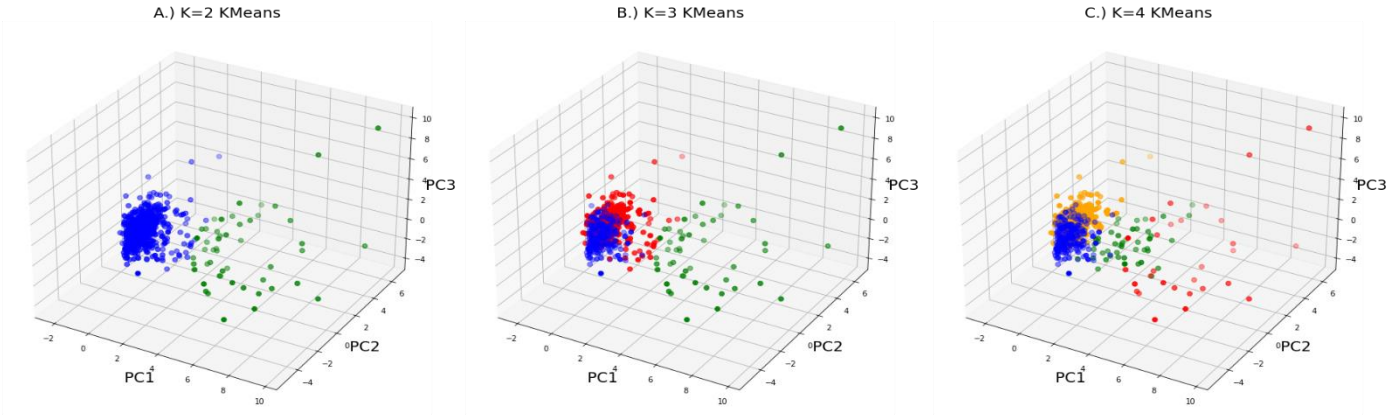

Fig 5: 3-dimensional plot of the KMeans Algorithm after applying PCA to the imputed dataset. PC: Principal Component

Similarly, for DBSCAN, although the option to select the number of clusters preemptively is absent, there are other parameters that need to be estimated such as eps and min_sample. Eps value represents the maximum distance between two samples for one to be classified as in-neighborhood of the other, and min_sample determines how many data points in a neighborhood are needed for a point to be considered as a core point [source]. After reviewing the 'Elbow Plot' constructed through NearestNeighbors, as seen in Fig 6, and silhouette scores, the optimal eps was 1.5 (we tested 1.0, 1.5, and 2.0 as reference), and the best min_sample was 11 (this number we did not change when comparing). The explanation of the separation of the two clusters for DBSCAN similarly matches the explanation given above for KMeans. The close-knit, dense group seen in Fig 7. represents the non-Hepatitis C patients, and the less dense, more sporadic cluster represents the Hepatitis C group of patients (with a few outliers - extreme cases). There is overlap seen with these two clusters, and we take this to mean that when it comes to individuals and health, it is often hard to make an accurate diagnosis, especially when looking at a large number of attributes. These patients could be considered "borderline" Hepatitis C, or at different stages of their Hepatitis C infection.

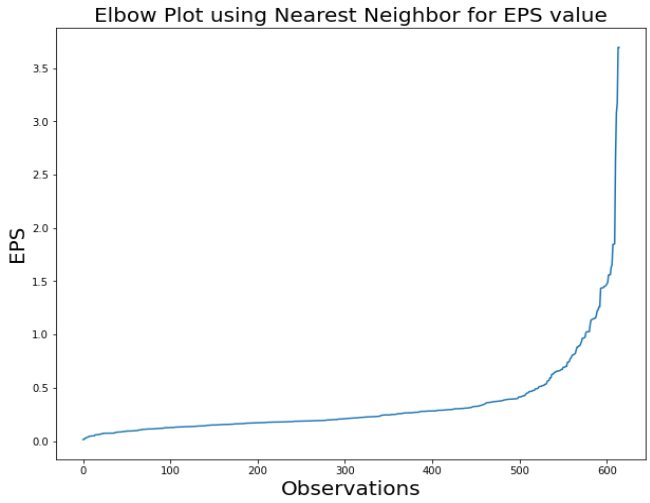

Figure 6: Nearest Neighbor Elbow Plot for EPS

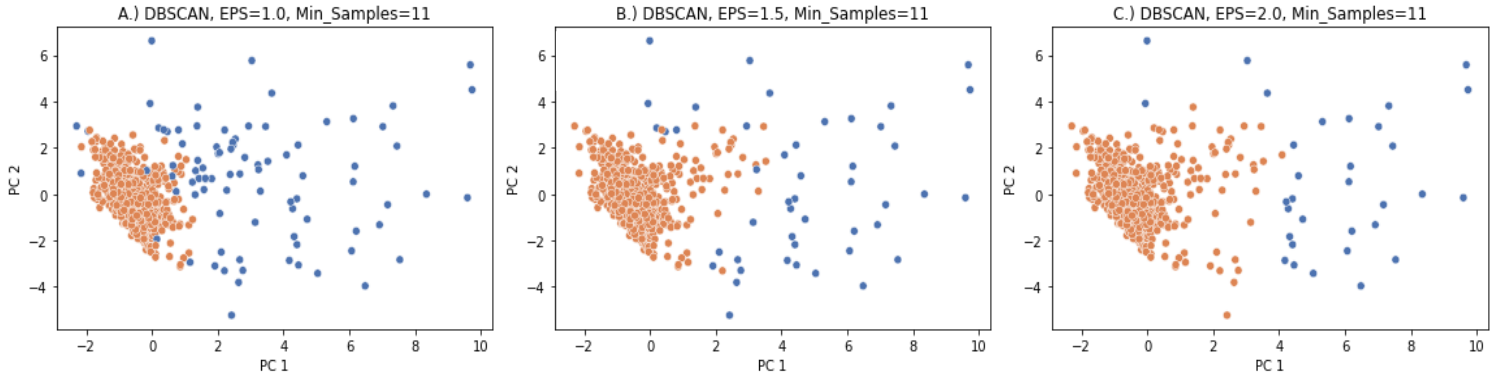

Figure 7: 2-dimensional Plots of DBSCAN Algorithm after applying PCA to the imputed dataset. PC: Principal Component

V. CONCLUSION

The purpose of this assignment was to choose a quality data set in which to apply two different clustering algorithms. Then as analysis, perform a detailed compare and contrast of the two clustering techniques, describe the overall significance of the clusters, and any other important findings.

We first performed a preliminary data analysis before anything else was done. This included careful attention to missing data, what data types existed, how many attributes and how they specifically related to the data, basic statistical analysis, and more. From there we cleaned the data in two ways, row deletion and imputation (replacing missing data with the average), and found the best representation of the data was imputed. Moving on, we did analysis with PCA to minimize the number of dimensions we would be working with. Lastly we performed our two clustering techniques, KMeans and DBSCAN with the cleaned and pruned data. Once we had performed the two different clustering techniques we took those results and with thoughtful reflection, we considered what those results meant.

In summary, the two clustering algorithms performed quite well, and we are confident that our clustering algorithms showed a clear separation of non-Hepatitis C subjects from the subjects with Hepatitis C. When applying this to the real-world, given that most people with Hepatitis C have no symptoms, this can be a very powerful tool to help identify those with the infection in clinical settings (e.g. blood banks).

REFERENCES

APPENDIX A

Table 2: Statistical Description of the Dataset

Attribute	Percent Missing	Count	Mean	Standard Deviation	Min/Max
Feature 1	0.0	615	N/A	N/A	N/A
Category	0.0	615	N/A	N/A	N/A
Age	0.0	615	47.408	10.055	19/77
Sex	0.0	615	N/A	N/A	N/A
ALB	0.162	614	41.620	5.780	14.9/82.2
ALP	2.926	597	68.293	26.028	11.3/416.6
ALT	0.162	614	28.451	25.469	0.9/325.3
AST	0.0	615	34.786	33.091	10.6/324.0
BIL	0.0	615	11.396	19.673	0.8/254.0
CHE	0.0	615	8.197	2.205	1.42/16.41
CHOL	1.626	605	5.368	1.132	1.43/9.67
CREA	0.0	615	81.287	49.756	8.0/1079.1
GGT	0.0	615	39.533	54.661	4.5/650.9
PROT	0.162	614	72.044	5.402	44.8/90.0

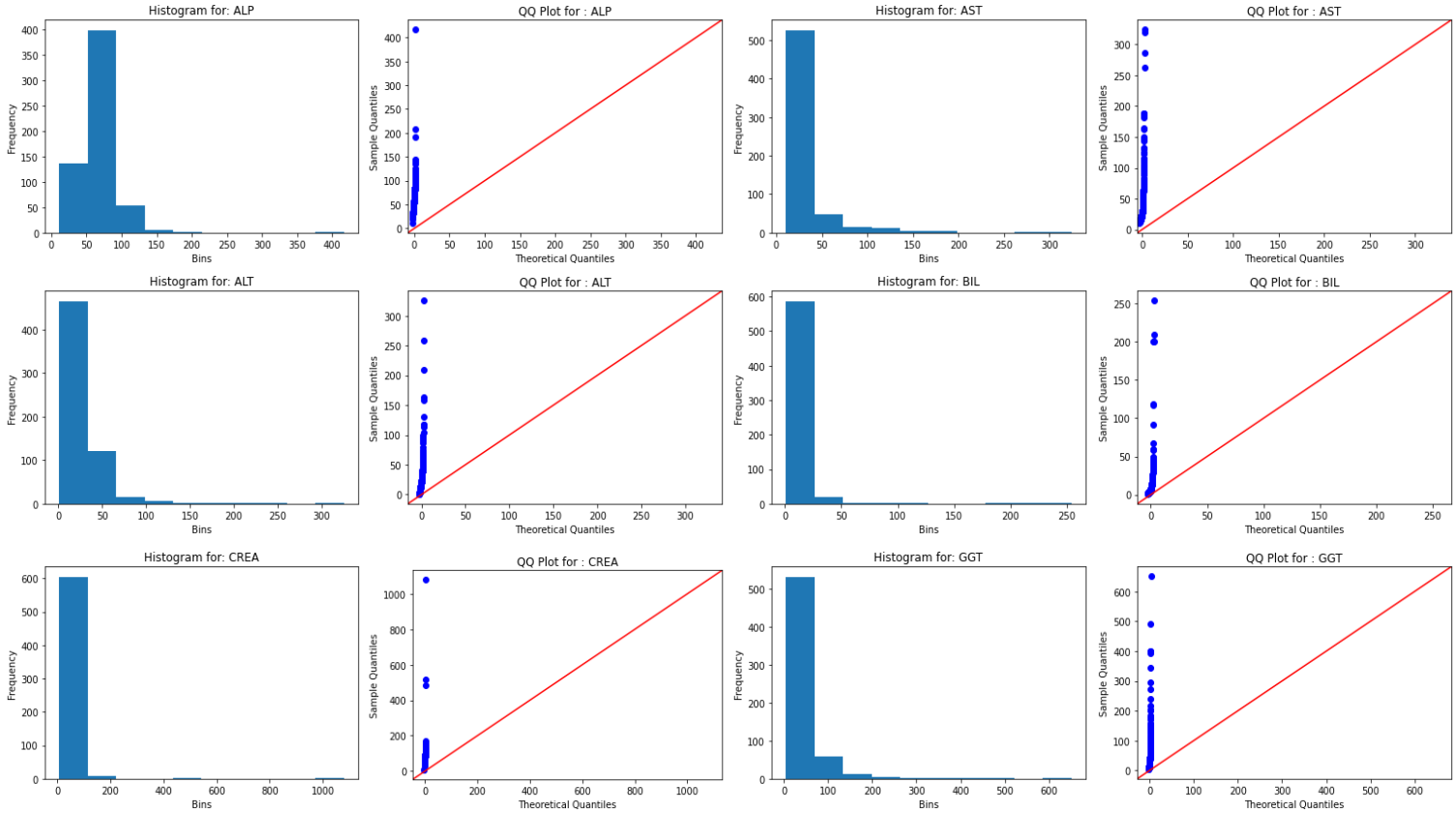

Fig 1: Histograms and QQ Plots for variables seen to have a high skewness.