Title: Checking data frame for KMeans cluster.

1. Introduction

The main purpose of this data science project is to extract meaningful insights and knowledge from large volumes of data using scientific methods, statistical techniques, and computational tools.
like Data preprocessing which is Cleaning and transforming the collected data to make it suitable for analysis.
This step involves tasks such as handling missing values, dealing with outliers, normalizing data, and performing feature engineering as we did in this project.

2. Data Description

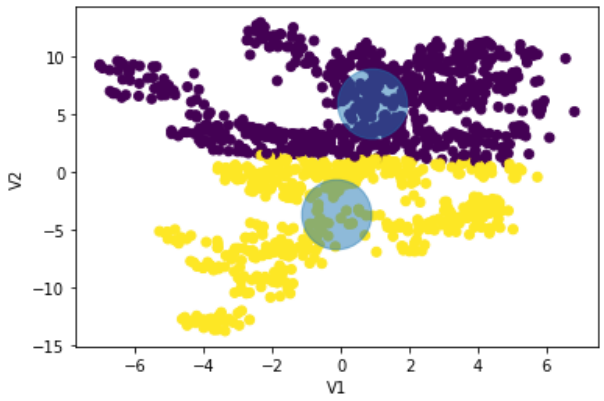
- After wrangling the data, we found that the dataset has **two** features.
- Each one has type **float.**
- contains **1372** data points.
- with **no null** values.
- **24** duplicated rows.
- After doing it we found that the means of the features are significantly different **0.433735** for V1 and **1.922353** for V2
- The standard deviations (**2.842763** for V1 and **5.869047** for V2) also differ, indicating potential differences in the variances of the features.

3. Methods

- Applying **normalization** (scaling the values to a specific range) can help address the issue of different scales and variances between the features. This preprocessing step is crucial for K-Means clustering as it is a distance-based algorithm that can be sensitive to feature scales.
- Once the necessary preprocessing steps are applied, K-Means clustering can be performed on the dataset to identify clusters based on the similarities between data points.
- In addition to the tasks, another crucial step in data preprocessing is data transformation. This involves converting raw data into a format that is more suitable for analysis.
- Another technique is dimensionality reduction, which involves reducing the number of features in a dataset while retaining as much relevant information as possible.

4. Results

- As a result, while the dataset appears suitable for K-Means clustering, it would be beneficial to perform normalization to ensure comparable contributions from both features and improve the quality of the clustering results.
- We found that we can divide the data frame into two clusters, one with the higher V2 and the other with the lower V2 as shown in the fig.



5. Recommendations

- I will recommend that following two experiments carried out on the data received it can be observed that outliers in V1 column fall between -8 to -6 and between 4 to 6 while for V2, we can say that the outliers fall between -12 to -15.
- This analysis can be used to determine the fake notes by carrying out more investigation on notes that fall within this range.
- With the increasing amount of data being generated every day, it is important to ensure that your analysis is based on the most current and accurate information available.
- This can help you identify new patterns and trends that may not be immediately apparent from your own analysis.

6. Conclusion

- The dataset has **two float** features with **1372** data points.
- The features are significantly different with mean **0.433735** for V1 and **1.922353** for V2 and variance **2.842763** for V1 and **5.869047** for V2.
- We used **normalization** to help us address the issue of different scales and variances between the features.
- We also used K-Means clustering to identify clusters based on the similarities between data points.
- The dataset appears **suitable** for K-Means clustering.
- The data frame has been divided into **two** clusters, one with the higher V2 and the other with the lower V2.
- It's better to stay up to date with the latest versions of the data and always have a fresh data source and collaborate with other experts in the field.