Implementation of K-means Clustering in Banknote Authentication Dataset

Chissanu Kittipakorn

64011728

a. Dataset Overview

This dataset contains 1,372 data that has 2 columns which are V1 (Variance of Wavelet Transformed Image) and V2 (Skewness of Wavelet Transformed Image). The data range from negative to positive floats which I think the positive means that the banknote is likely to be authentic

b. Data Processing and Exploratory Data Analysis

Use Pandas and Matplotlib to extract the content and visualize for better understanding as well as check if the data contain invalid inputs.

```
df = pd.read_csv('banknote_data.csv')
df
```

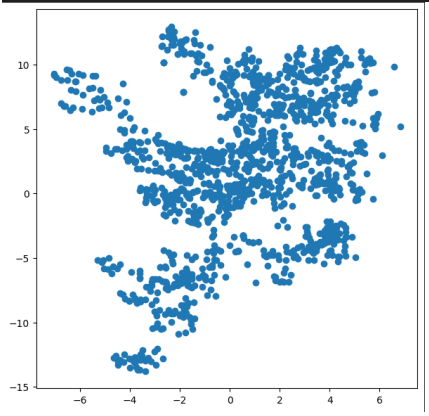
✓ 0.1s

	V1	V2
0	3.62160	8.66610
1	4.54590	8.16740
2	3.86600	-2.63830
3	3.45660	9.52280
4	0.32924	-4.45520
...	...	...
1367	0.40614	1.34920
1368	-1.38870	-4.87730
1369	-3.75030	-13.45860
1370	-3.56370	-8.38270
1371	-2.54190	-0.65804

```
df.describe()
```

✓ 0.0s

	V1	V2
count	1372.000000	1372.000000
mean	0.433735	1.922353
std	2.842763	5.869047
min	-7.042100	-13.773100
25%	-1.773000	-1.708200
50%	0.496180	2.319650
75%	2.821475	6.814625
max	6.824800	12.951600

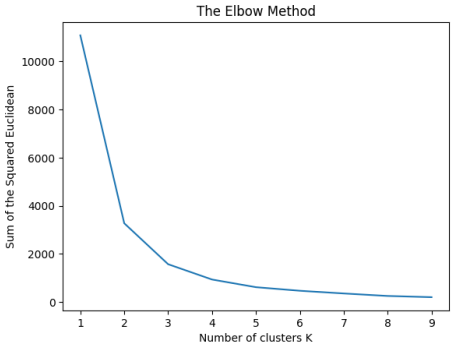



```
df.isnull().sum()
```

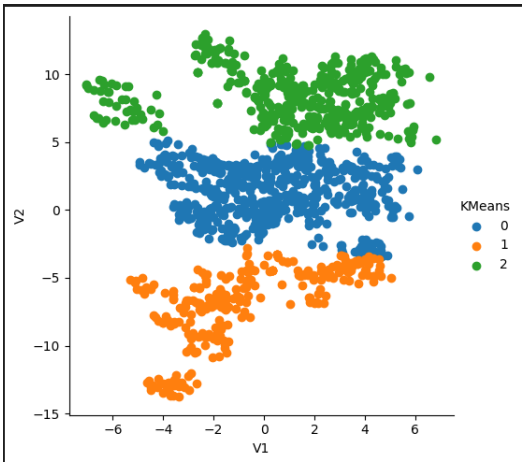
✓ 0.0s

```
V1      0
V2      0
dtype: int64
```

c. Training Models



Use Elbow method to find best number of clusters for this case I think either 2 or 3 would be the best choice and I will go with 3 number of clusters



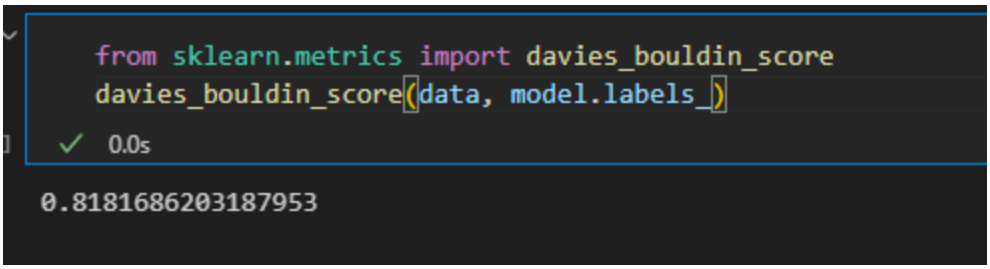
I use seaborn to make a group of clusters and the result are this image.

d. Model Evaluation

I use the Silhouette score to calculate with 3 clusters and the reason the percentage is low because I think there are other factor that affect the prediction due to bad data

```
KMeans
KMeans(n_clusters=3)
```

```
silhouette_score(data, model.labels_)
✓ 0.0s
0.43362017621560883
```



For the second evaluation use Davies boudin score which gives better results.

e. Conclusion

After testing with different number of clusters I found out that 2 clusters give the best result which increase from mine about 1-5% which is significant. I also learned that it is quite hard to train a model with limited number of features and unclean data.