An Analytical Way to Get More Customers After COVID-19

Tan Li Tung
Electrical and Electronics Department
Universiti Teknologi PETRONAS
li_17002803@utp.edu.my
+6011-10868940

An Analytical Way to Get More Customers After COVID-19 is an exploratory analysis to find insights in the current customer database. This document will visualize the data using plots and graphs. Techniques like Simple Linear Regression, Multiple Linear Regression, Logistic Regression, Naive Bayes, Linear Discriminant Analysis and Quadratic Discriminant Analysis will be carried out for predictive analysis. This report would include the use of Principal Component Analysis and Singular Vector Decomposition as the unsupervised learning on the country dataset.

1 Introduction to The Study

Due to the recent COVID-19 outbreak, lots of businesses lost their customers. The businesses need to prepare for the opening of their business so that they can serve their customers better. They also need to acquire new customers and retain the existing customers so that the business can survive the pandemic. In this study, a dataset contains 500 records of customers dataset is used for analysis. The dataset includes the average session length, time on app, time on website, length of membership and also yearly amount spent by the customers. The goal of the study, the yearly amount spent by the customers is set to be the target variable for prediction. The **relationship of the other variables and the target variable should be examining** and **policies to keep the customer are to be proposed.** All the codes are run in Google Colab which can be found at https://colab.research.google.com/drive/1Bj4sz-yz61W897xDiZyAqkVPlTCtKkCz?usp=sharing

2 Method and Algorithm Used

2.1 Exploratory Data Analysis (EDA)

Feature Selection. Before the data is used further, the data needs to be pre-processed. The Email, Address and Avatar attributes are dropped as it does not affect the yearly amount spent of a customer.

Description of Data. Some of the information can be gotten from the description of data and it is shown as below in Figure 1.

	avg_session_length	time_on_app	time_on_website	length_of_membership	yearly_amount_spent
count	500.000000	500.000000	500.000000	500.000000	500.000000
mean	33.053194	12.052488	37.060445	3.533462	499.314038
std	0.992563	0.994216	1.010489	0.999278	79.314782
min	29.532429	8.508152	33.913847	0.269901	256.670582
25%	32.341822	11.388153	36.349257	2.930450	445.038277
50%	33.082008	11.983231	37.069367	3.533975	498.887875
75%	33.711985	12.753850	37.716432	4.126502	549.313828
max	36.139662	15.126994	40.005182	6.922689	765.518462

Figure 1: Description of Data

In Figure 1, it is shown than this dataset has no null values and it can be considered clean. Therefore, the pro-processing of data becomes easier. The minimum and maximum values in the dataset is also reasonable which means that there are no outliers.

Distribution of Attributes and Target Variable.

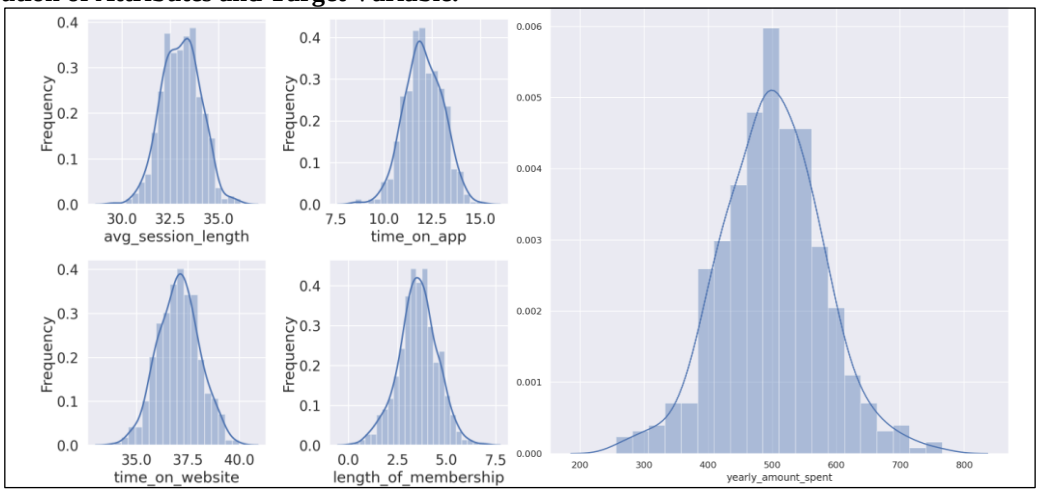


Figure 2: Distribution of Attributes and Target Variable Numbering and Running Heads

As shown in Figure 1, all the attributes and target variable are almost normally distributed. This makes the data more explainable.

2.2 Simple Linear Regression (SLR)

The business is interested to determine whether each of the attributes are highly related to the yearly amount spent. This can be achieved by using SLR. The relationship between each attribute and the target variable are plotted in Figure 3 below.

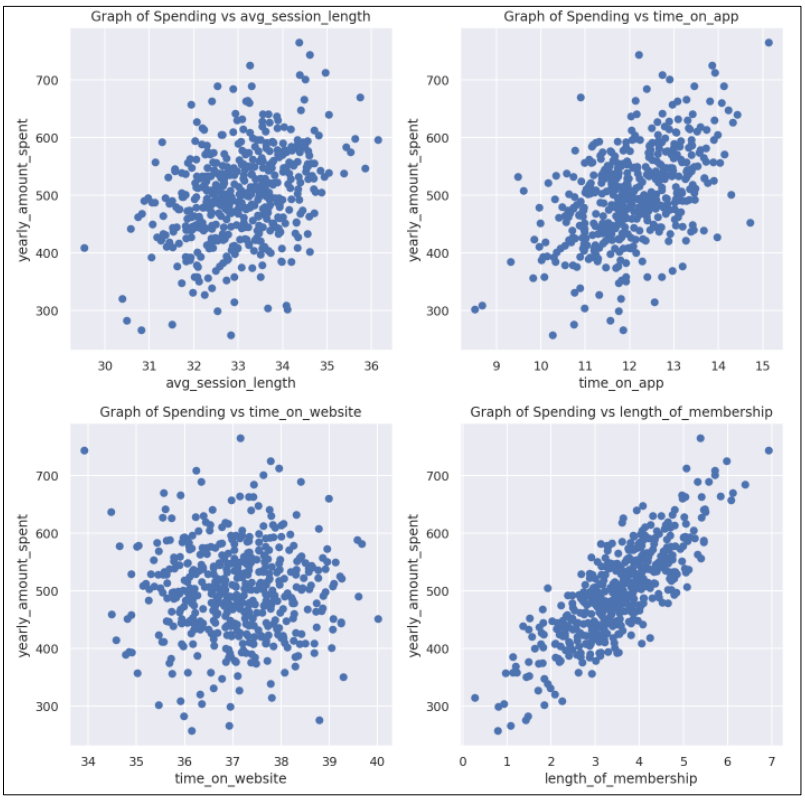


Figure 3: Relationship Between Attributes and Target Variable

In Figure 3, it is clearly shown that the length of membership has a linear relationship with the yearly amount spent. This means that the longer the membership, the more the member spent. However, we still need to test the hypothesis to know that whether the the target variable depends on the attributes. The result of SLR on each attribute are shown below.

```
=======length_of_membership=======

b0 =  272.3997860603549
b1 =  64.21868431527953

============ANOVA============

SST:  3139126.45168154          || SSE:  1084203.3584881509     || SSR:  2054923.093193388
MSR:  2054923.093193388         || MSE:  2177.11517768705       || F-test:  943.8743132444292
Sigma:  46.65956683989951       || R2:  0.6546162204114542

=======length_of_membership=======

T-test
---------
T calculated value 30.72253754565904 > T in table 1.9647389829672648
Hence Ho is rejected, yearly_amount_spent depend on length_of_membership
The confidence interval of b1 => [60.1118316.4 , 68.3255386.4]

F-test
---------
F calculated Value 943.8743132444292, > F in table 3.860199271191399
Hence Ho is rejected, yearly_amount_spent depend on length_of_membership
```

Figure 4(a)

```
=======avg_session_length=======

T-test
---------
T calculated value 8.476510436294737 > T in table 1.9647389829672648
Hence Ho is rejected, yearly_amount_spent depend on avg_session_length
The confidence interval of b1 => [21.7978876.4 , 34.9516546.4]

F-test
---------
F calculated Value 71.8512291766136, > F in table 3.860199271191399
Hence Ho is rejected, yearly_amount_spent depend on avg_session_length
```

Figure 4(b)

```
=======time_on_app=======

T-test
---------
T calculated value 12.86101802904204 > T in table 1.9647389829672648
Hence Ho is rejected, yearly_amount_spent depend on time_on_app
The confidence interval of b1 => [33.7490966.4 , 45.9198876.4]

F-test
---------
F calculated Value 165.40578474334444, > F in table 3.860199271191399
Hence Ho is rejected, yearly_amount_spent depend on time_on_app
```

Figure 4(c)

```
=======time_on_website=======

T-test
---------
Failed to reject Ho, there is No dependencis between yearly_amount_spend and time_on_website

F-test
---------
Failed to reject Ho, there is No dependencies between yearly_amount_spent and time_on_website
```

Figure 4(d)

Figure 4: SLR Results on Each Attribute

As the result of SLR suggest, the length of membership is 65.46% linearly related to the yearly amount spent. The average session length and time on app is not highly related to the yearly amount spent but both attributes are still statistically significant as shown in Figure 4(b) and 4(c). The yearly amount spent has no dependencies on the time on website as shown in Figure 4(d). Through SLR, we found out that the customers need to be retained so that they spent more. Since the yearly amount spent does not depends on the time on website, the business can consider the switch the website service to the app as most customers spent their time on app.

Although the SLR can give a prediction accuracy up to 65.46% through the equation:

*(yearly amount spent) = 272.400 + 64.219 * (length of membership)*

However, the accuracy is not that high to make accurate predictions. Therefore, there is a need to perform other techniques to improve the accuracy of the predictions.

2.3 Multiple Linear Regression (MLR)

MLR is like SLR but it considered all the attributes. Through MLR, the results are as below.

```
b0 =  -1051.594254996944
b1 =  25.734271083505504
b2 =  38.709153813578176
b3 =  0.43673882831422906
b4 =  61.577323749790814


============ANOVA============
SST:  3139126.45168154          || SSE:  49235.51262222303      || SSR:  3089890.939059321
MSR:  772472.7347648302         || MSE:  99.46568206509703      || F:  7766.223673601033        || R2:  0.9843155370195906
```

Figure 5: Result of MLR

In the result above, we can see that the R^2 score is 98.43%. This indicates that the MLR model can perform predictions better. The equation can be written as:

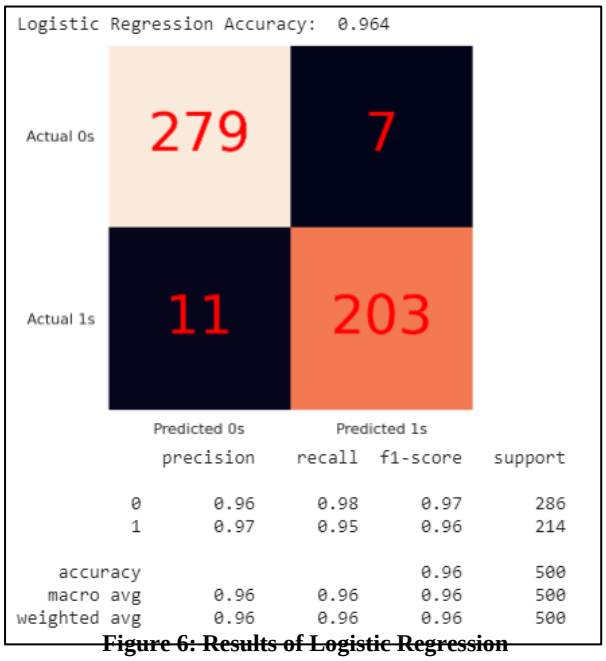
*(yearly amount spent) = -1051.59 +25.73 * (average session length) + 38.71 * (time on app) + 0.44 * (time on website) + 61.58 * (length of membership)*

The MLR model also have a lower mean squared error (MSE) and mean squared regression (MSR) as compared to the SLR model using only the length of membership. Both SLR and MLR models can predict the yearly amount spent of the customers. The prediction can be generalized into categories to make it easier to predict.

2.4 Logistic Regression

Logistic Regression is used to solve a classification problem. Therefore, the target variable needs to be changed to classes before logistic regression can be used. In the study, the target variable is split into 2 categories which are 0 (low) and 1 (high) where 0 are the yearly amount spent less than the midpoint of yearly amount spent while 1 represent yearly amount spent that are at least the midpoint of the yearly amount spent.

The logistic regression predicts the class of a new data based on the probabilities of it belongs to a class. The results are shown as below. The function used to determine the probability of a data belonging to a particular class is called a logistic function.



	precision	recall	f1-score	support
0	0.96	0.98	0.97	286
1	0.97	0.95	0.96	214
accuracy			0.96	500
macro avg	0.96	0.96	0.96	500
weighted avg	0.96	0.96	0.96	500

Figure 6: Results of Logistic Regression

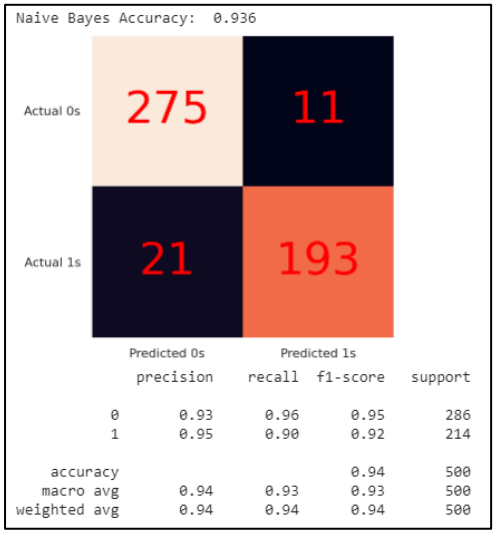
As seen in Figure 6, the model correctly predicts 482 data out of 500 data. The model provides an accuracy of 96.4% which is high enough to be used for predictions. However, more models still can be used to predict the class of yearly amount spent from the attributes.

The Logistic Regression can be performed easily using the ScikitLearn library. By default, the Logistic Regression model in ScikitLearn used the logistic function (solver algorithm) of Limited-memory Broyden–Fletcher–Goldfarb–Shanno Algorithm (L-BFGS).

The L-BFGS algorithm approximate the Hessian matrix using updates specified by gradient evaluations. This algorithm is chosen as compared to others as it is efficient, and it consume lesser amount of memory.

2.5 Naive Bayes Classifier

Naive Bayes is a classifier working based on the Bayes' Theorem. The algorithm assumes that the attributes are independent of each other. In real life, this is nearly impossible, but the Naive Bayes classifier works surprisingly well. The model can predict the class with a high accuracy.



	precision	recall	f1-score	support
0	0.93	0.96	0.95	286
1	0.95	0.90	0.92	214
accuracy			0.94	500
macro avg	0.94	0.93	0.93	500
weighted avg	0.94	0.94	0.94	500

Figure 7: Results of Naive Bayes Classifier

As shown in Figure 7, the Naive Bayes model correctly predict 486 data out of 500 data. In other words, the Naive Bayes model predicts the output with an accuracy of 93.6%. This model has a high accuracy, but it is not as good as the Logistic Regression model.

This is because the model assumes that all the attributes are normally (Gaussian) distributed. However, as shown in Section 2.1, Figure 2, the distribution is not exact normal. This is the reason that the Naive Bayes Classifier has a lower accuracy as compared to the Logistic Regression that uses a logistic function to calculate probabilities.

Naive Bayes classifier can be constructed easily as it does not require a lot of training. Due to the assumption made earlier, the model is insensitive towards the irrelevant attributes, which makes the model accurate.

2.6 Linear Discriminant Analysis (LDA)

LDA is a method used to reduce the dimension of the dataset. At the same time, it also has the ability to perform classification. The general idea of LDA is to find a line that can best separate the classes.

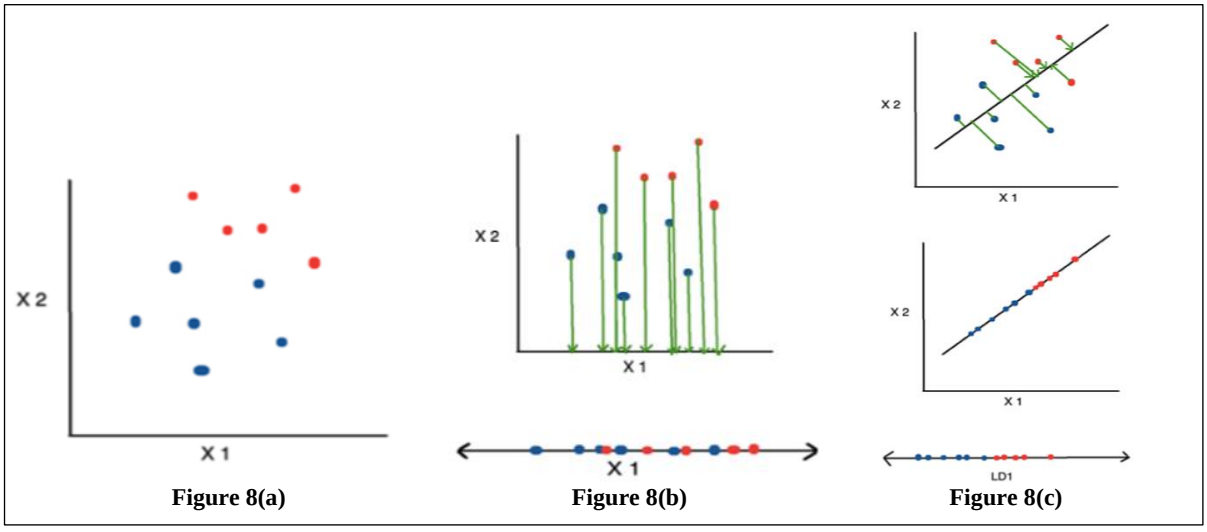


Figure 8(a) **Figure 8(b)** **Figure 8(c)**

Figure 8(a): Sample Dataset, 8(b): Data Projected to x-axis, 8(c): Data Projected to LD1

Consider a dataset as shown in Figure 8(a). To differentiate the classes, the data can be projected to the a-axis as shown in Figure 8(b). However, there are lots of overlapping in the output, therefore a linear line with the best separability needs to be found. The line in Figure 8(c) has the best separability in this case. The data are projected to the linear line LD1 which the data points are found to be easily separable.

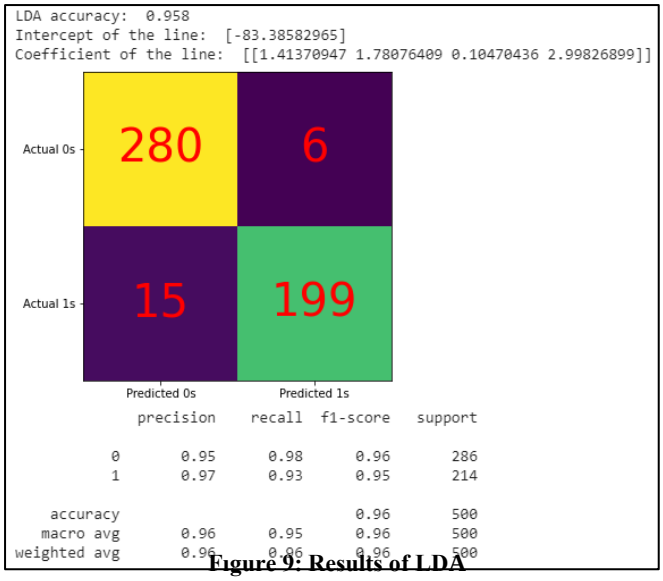


Figure 9: Results of LDA

Base on the confusion matrix shown in Figure 9, the LDA model yields an accuracy of 95.8% and the linear line can be observed through the intercept and coefficient which can be written as:

*LD1 = -83.39 + 1.41 * (average session length) + 1.78 * (time on app) + 0.1 * (time on website) + 2.98 * (length of membership)*

The line is found to have a greater lean towards the length of membership, which shows that length of membership contains most information that can be the most significant attribute for the classification. However, the line that separate the classes may not always be a linear line. Therefore, QDA can be carried out to examine different relationship between the attributes and target variable.

2.7 Quadratic Discriminant Analysis (QDA)

QDA shares the same concept as LDA. However, the line that can fit make the best separability might not always be a linear line. Therefore, QDA finds a line in quadratic that have the best separability of data.

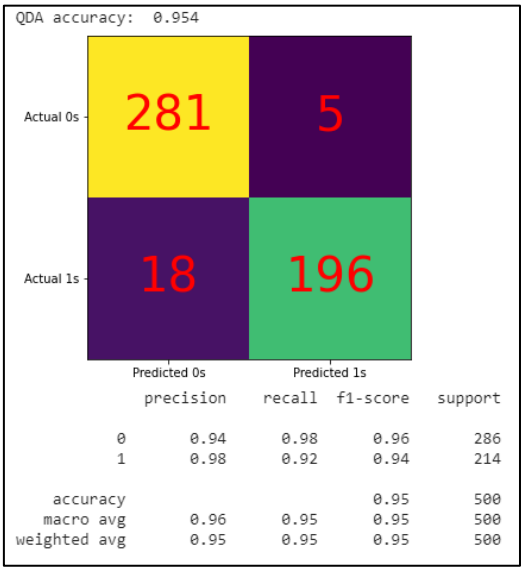


Figure 10: Results of QDA

In Figure 10, it is observed that the QDA model provides an accuracy of 95.4%. The accuracy is lower than the LDA model. This happens due to the data is related to the target variable linearly instead of quadratically.

2.8 Conclusion of Supervised Learning

As 6 techniques are performed with 2 regression and 4 classification, the business can now get some insights from the analysis. The customers generally spent more when they have a longer length of membership. Therefore, the business can send out gifts to their existing customer to retain them longer. The analysis also suggests that the time on website does not affect the yearly amount spent. Therefore, the business can considering move the website fully to the app. The MLR model can be used to predict the yearly amount spent of the customers and the Logistic Regression model can be used to predict classes of the customers. If a customer is predicted to be in the high yearly amount spent class, the business can consider upgrading their membership so that they can retain the customers by rewarding them.

3 Unsupervised Learning on Country Dataset

3.1 Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that preserve most of the information. The principal components (PC) represents the direction of the data that explains the maximum amount of variance. PC_n always have a greater variance than the $PC_{(n+1)}$, for instance, PC1 has a greater variance than PC2.

In the country dataset, there are 167 records from 167 different countries. Therefore, the country column is dropped as it does not play a significant role in the clustering in PCA.

```
Explained variation per principal component: [0.94792307 0.05207346]
```

	principal component 1	principal component 2
0	-19815.128018	-1651.124720
1	-11338.938992	1493.469028
2	-8926.352589	3264.684102
3	-14652.465937	-867.795270
4	896.142032	1898.478151
...	...	...
162	-17180.967679	-2487.153835
163	-100.445973	-832.252802
164	-17202.017189	-222.889872
165	-17209.340213	-229.815901
166	-17978.421643	-1163.124631

167 rows × 2 columns

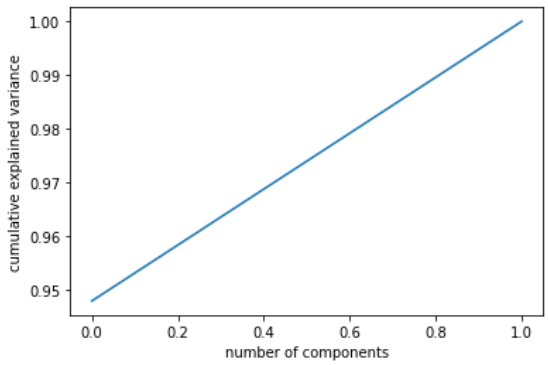


Figure 11: Output of PCA

As shown in Figure 11, when the number of components reaches 1, the cumulative explained variance also reaches 1. This means that the PCA has a high performance. The result also shows that 94.79% of information is hold by PC1 and 5.20% information is hold by PC2. This means that if a 4-dimension data is projected into 2-dimension data, there are only 0.01% of information loss. This means that the PCA works perfectly fine on the country dataset.

3.2 Singular Vector Decomposition (SVD)

Every table data can be represented by a matrix. With the help of SVD, any matrix can be decomposed into 3 matrices.

$$\left(A \right) = \left(U \right) \bullet \left(\sum \right) \bullet \left(V \right)^T$$

$n \times p \qquad n \times \ell \qquad \ell \times \ell \qquad \ell \times p$

Figure 12: SVD Concept

A is the matrix that needs to be decomposed

U is the left singular vector matrix

$\sum$ is the singular values

V^T (Transpose of V) is the right singular vector matrix.

SVD can be used to decompose the country data. The result is shown as below.

```
U:  [[-3.64683441e-03 -7.74480085e-03 -1.04434457e-01 ... -2.48363921e-02
  -8.01744598e-02 -1.09311396e-01]
 [-2.35129587e-02 -4.14338548e-02 -5.44062005e-02 ... -1.25510417e-01
  -9.17577217e-03 -6.51725205e-04]
 [-2.92612753e-02 -6.17536695e-02 -4.70183049e-02 ... -3.32352600e-02
  -3.60771176e-02 -2.25951311e-02]
 ...
 [-9.82048538e-03 -2.37501795e-02 -9.98635426e-02 ...  9.63875510e-01
  -7.25818218e-03 -2.86053793e-03]
 [-9.80280116e-03 -2.36688518e-02 -8.40085846e-02 ... -6.38754995e-03
   9.73517253e-01 -1.76789981e-02]
 [-7.93092898e-03 -1.26762558e-02 -9.53267358e-02 ... -1.47461437e-03
  -1.99469044e-02  9.63731047e-01]]

S:  [4.33547092e+05 8.15635234e+04 1.01650158e+03 4.49350195e+02
 2.98137776e+02 1.49777287e+02 1.09392439e+02 2.67184658e+01
 1.01698190e+01]

V-transpose:  [[-2.50831759e-04 -1.08749581e-03 -1.44528920e-04 -9.61447024e-04
  -7.57832441e-01 -1.03418990e-04 -1.47130086e-03 -3.91099845e-05
  -6.52445878e-01]
 [-1.42343122e-03 -1.87282758e-03 -1.52840786e-05 -1.56359599e-03
  -6.52437369e-01 -5.87778767e-04 -2.26985596e-03 -1.01012013e-04
   7.57833751e-01]
 [-5.98081282e-01 -3.04674782e-01 -5.96081698e-02 -4.54358343e-01
   3.63015323e-03 -8.21201539e-02 -5.75755112e-01 -3.49910571e-02
  -1.48253017e-03]
 [ 7.48305494e-01 -4.23764717e-01 -1.39937343e-02 -4.61773617e-01
   1.31901098e-03  8.08137727e-02 -2.00061107e-01  2.17853708e-02
   7.18528398e-06]
 [-2.70127785e-01 -4.67716466e-01  7.61938887e-02 -3.28123369e-01
  -2.05598611e-04  8.22020281e-02  7.66761365e-01  1.10273683e-02
  -1.53886402e-04]
 [-3.87771708e-02  5.93145158e-01 -7.68671319e-02 -5.28266407e-01
  -6.84145353e-04  5.97790154e-01  6.55347330e-02  7.26107737e-03
   3.73414511e-04]
 [ 8.18451925e-02  3.89225049e-01 -4.23580737e-02 -4.34677446e-01
  -1.97039351e-04 -7.88975139e-01  1.69061278e-01 -1.73439513e-03
  -5.75113270e-05]
 [-3.68791247e-03  7.42593354e-02  9.91243000e-01 -6.81570659e-02
   7.05483965e-05  2.67748430e-03 -8.37309916e-02 -1.56782637e-02
  -1.34110209e-04]
 [ 3.39212798e-02 -1.26222344e-03 -1.34187680e-02  2.07351314e-04
  -1.07555234e-05  1.12194923e-02  2.57662659e-02 -9.98938371e-01
   4.21993972e-06]]
-0.11582609270422234
```

Figure 13: The U, S, V Matrix After Decomposition

As shown in Figure 13, the country dataset is decomposed into 3 components. In the S matrix, we can see that there are sigma points. These point shows that how close all the points are close to that axis. In our case, the sigma points are as below.

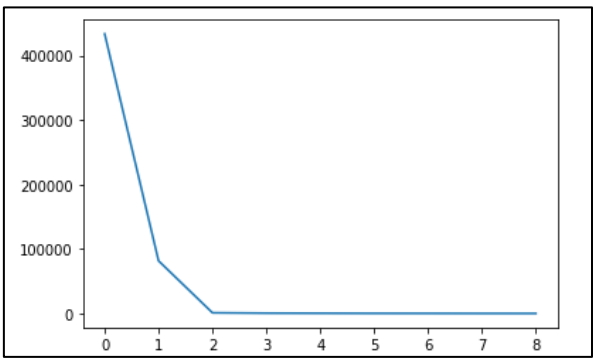


Figure 12: Sigma Points

3.3 Conclusion of Unsupervised Learning

There are 2 most popular algorithm used for unsupervised learning, which are PCA and SVD. Both algorithms are performed on the country dataset and the clusters can be shown after performing them. PCA and SVD are generally helpful in performing clustering that can find the common pattern in the dataset. Although the patterns can be deduced, but it still requires experts to decide whether the pattern are useful or not.