FORTUNE 500 Report

I. INTRODUCTION

The Fortune magazine produces an annual list of the top 500 largest US companies ranked based on their revenue for their respective fiscal years. The Fortune data store lists make it easy to research and acquire the information needed to evaluate our business landscape. In this assignment, a **grading scale system** for Fortune 500 companies are developed with the help of a hybrid approach of clustering techniques.
As of now, companies are ranked by their annual revenues for their respective fiscal years. Hence, this analysis and solution can be very helpful for effective financial targeting.

This study also focuses on research questions and hypothesis such as (profit does not affect the ranking of the company, Effects of Rank with respect to Profits and Revenues, assets do not affect the ranking of the company, and Effects of Rank with respect to the Number of Employees).The possible obstacles incorporated in the data can be the **Time period** since a decade is deemed as a sufficient duration for doing financial analysis, and however this data is limited for a year 2017. Also, **multicollinearity** can be expected to have in the dataset **[1],** and there might be high chances of outliers and **skewness [2]** in the dataset.

II. METHODOLOGY

This section confers about summary of the data, its technical description, and reproducible methodology for data exploration, data cleaning, feature engineering, modelling, and analysis.

In the dataset, there are 500 records of the companies with 23 features. The features include Rank, Title, Website, Employees, Sector, Industry, Headquarter location, Headquarter address, Headquarter city, Headquarter state, Headquarter zip, Headquarter telephone, CEO, CEO title, Address, Ticker, Full name, Revenues, Rev change, Profits, Profit change, Assets and Total share equity.
There are no null values and duplicate records in the dataset. As far as distribution of data is concerned, features like Total equity shares, Profits, Revenues, and Assets are right skewed as shown in Fig 1.

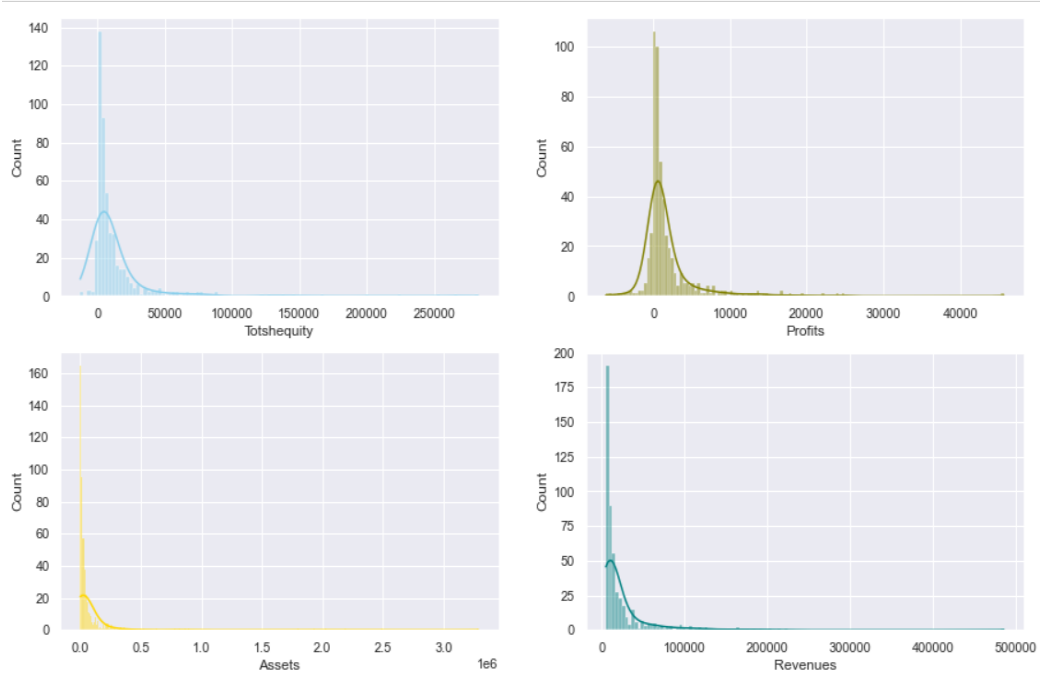

Fig.1 Dist. of Revenue, Profit, Total share equity And Assets -- **[3]**

Since, Data can generate effective results if normalization is applied as it standardizes attributes with equal weight so that redundant or noise data can be removed, which eventually enhances the accuracy of the result. To ensure that the model capabilities are not affected, data transformation is required. **[4]**

To begin with data cleaning **[5]** , unnecessary columns (Website, Ceo, Ceo-title, Fullname) has been removed from the dataset that contains insignificant information **[6].**

To assess the relationship between the financial performances with headquarter location, the top 10 state with high profits has been considered along with number of companies present in that location. These 10 companies suggests that profits vary irrespective of the headquarter location. It goes same with other key financial performance metric like revenue, assets. [Ref :- Fig 2]

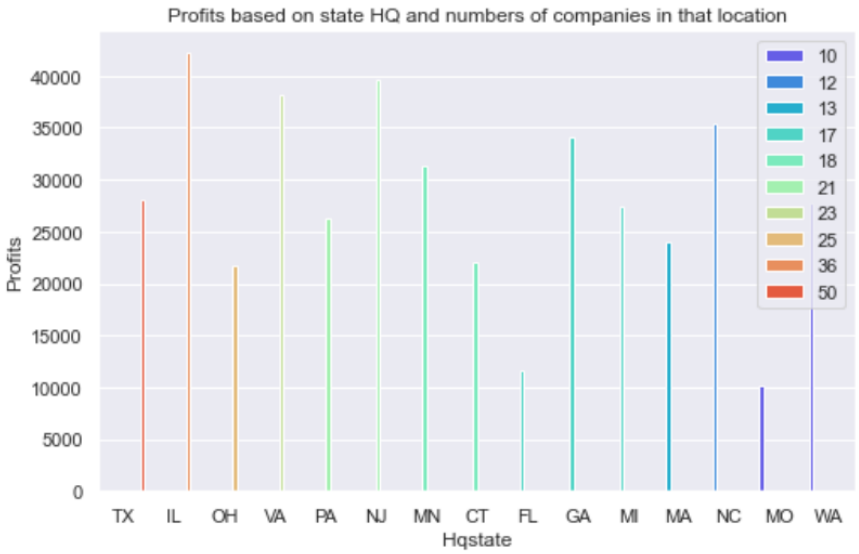

Fig. 2. Bar chart of top 10 HQstate with Max profit along with no. of companies. – **[7]**

Hence, all the features from the datasets which signifies address/location of the headquarters of the companies has been removed. Then, Exploratory data analysis is used to discover patterns in the data like identifying the sectors or industry that are having high Revenue , profits, assets through bar plots as shown in Fig. 3(a) and Fig. 3(b)

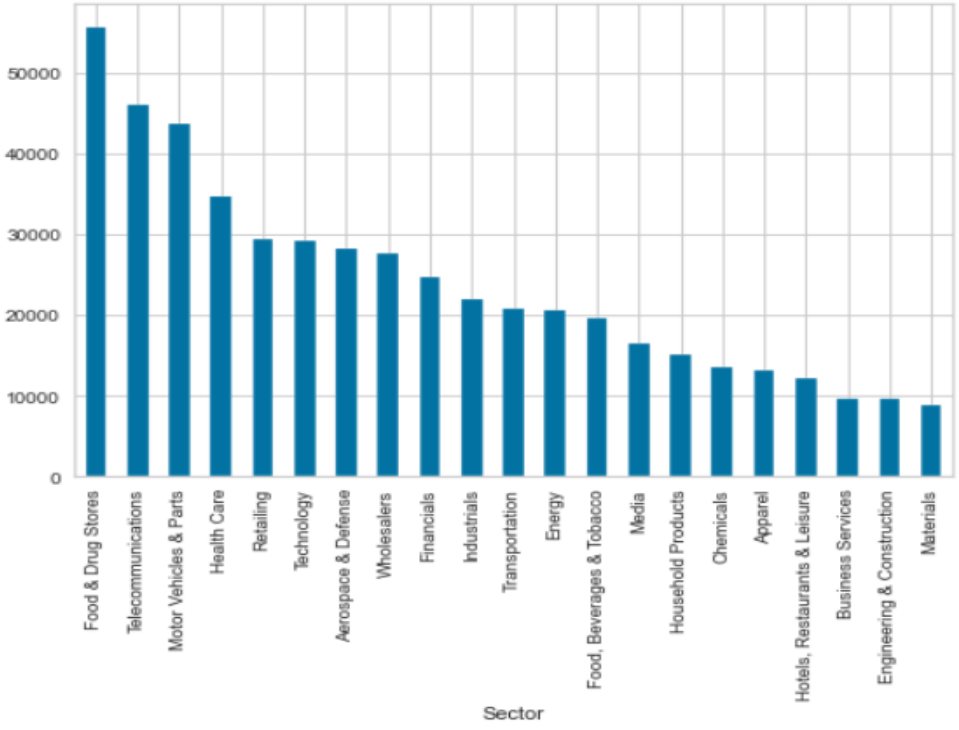

Fig. 3(a). Revenue based on the sector ***[8]***

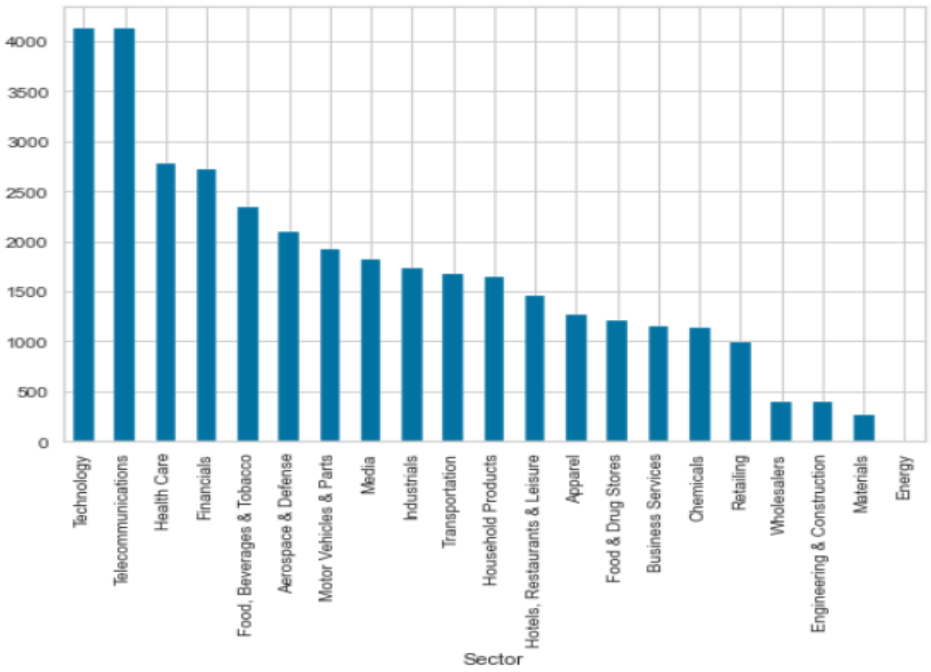

*Fig 3(b) Profit based on the sector **[9]***

Scatter plots are identifying the relation between Revenue and profits (Ref. Fig. 4), with Pearson's correlation of 60 % which is considered as strong relationship. However, correlation is weak between Revenue and Assets.[Ref Fig 5]

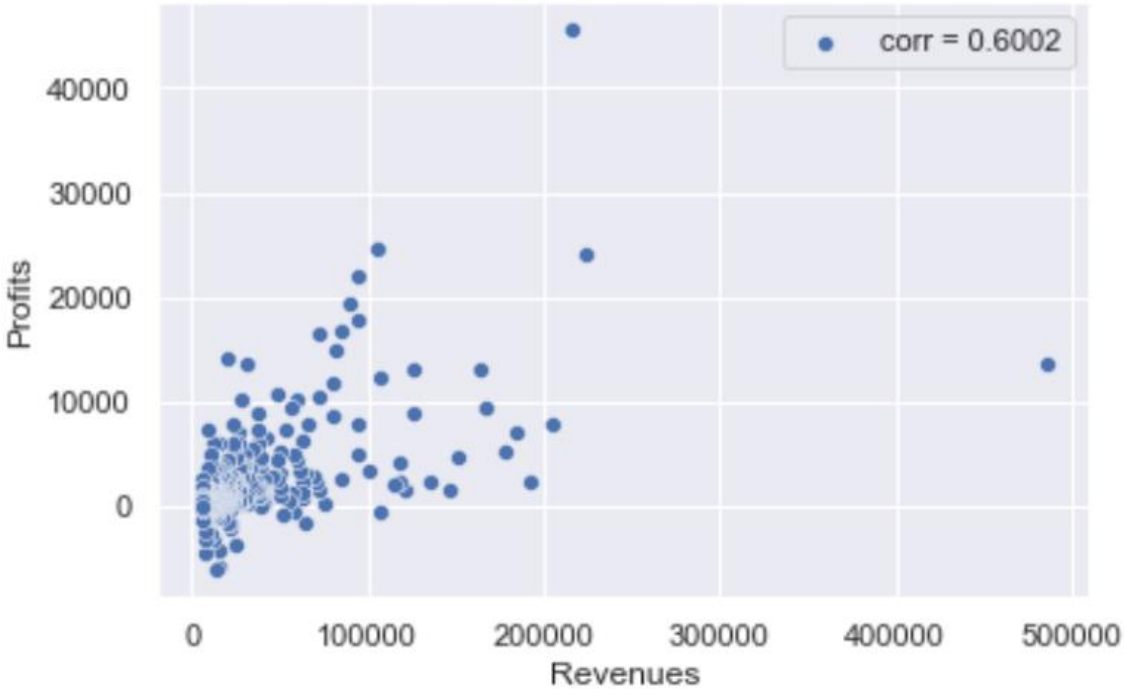

Fig.4 Scatter plots of Revenue and Profit

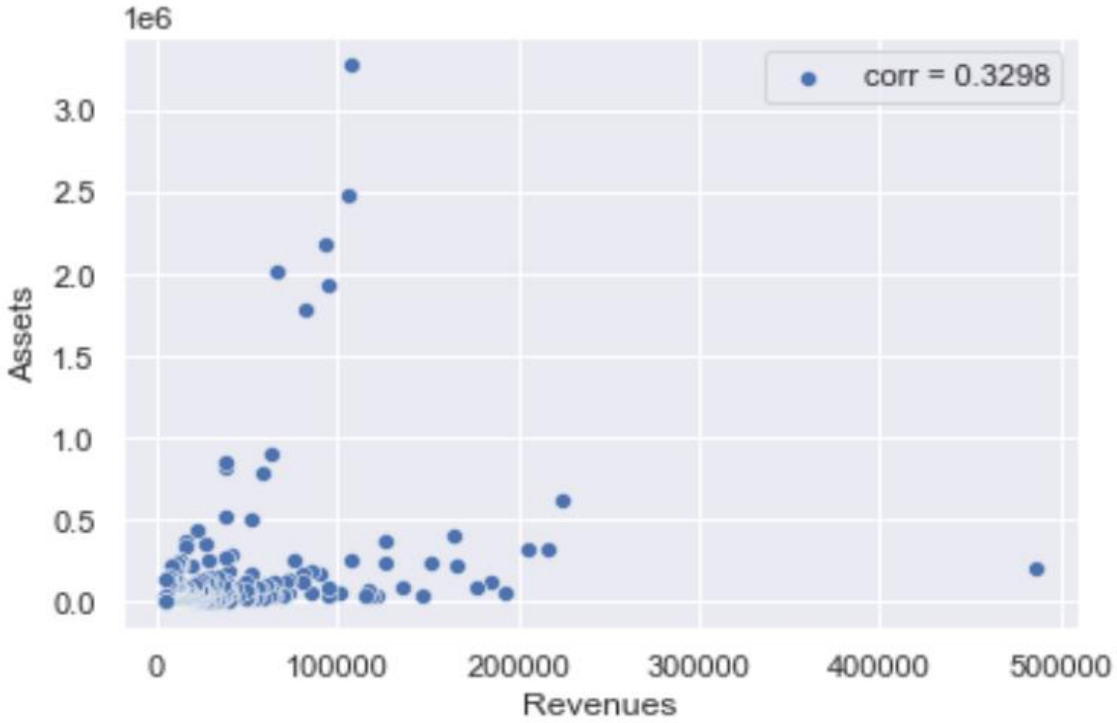

Fig.5 Scatter plots of Revenue and Assets

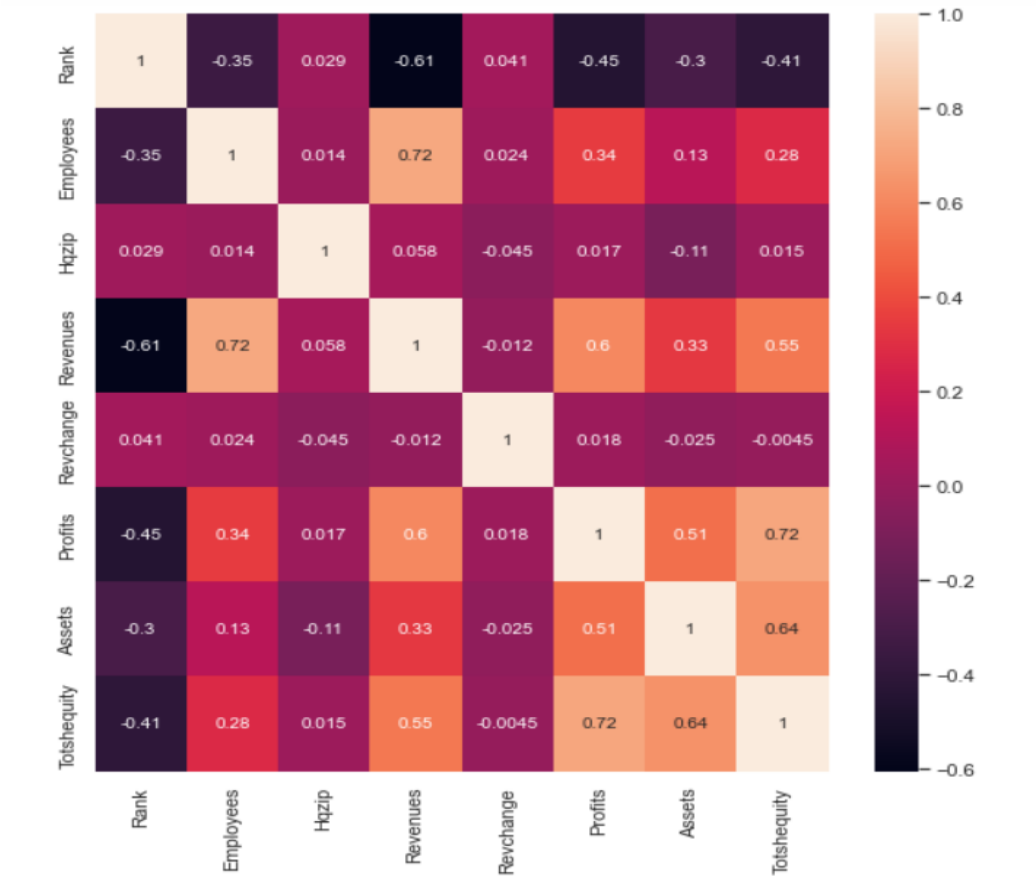

Fig.6 Heat Map [10]

Correlation between different features can be identified through Heat map. As per the Fig 6 , there is a strong correlation between Revenue and Employees, Revenue and Profit.

Based on Fig 3(a) and 3(b), Financial performance of the companies varies with the sectors and type of Industries. Hence, these two attributes will play huge role in identifying the similar clusters. However, these categorical columns need to be encoded since ML models require numeric input variables. Therefore, application of binary encoding and one hot encoding are required as a part of feature engineering process.**[11]**

Up Next, K- Means, hierarchical clustering and DB scan clustering are used to obtain similarities in the data points with different hyperparameter and evaluated the result based on Silhouette score.

III. ML MODELS RESULTS

K-Means clustering Model

The first phase of the analysis is characterized by identifying the number of clusters that group similar data points. This can be done with the help of Elbow method **[12]** as refer in Fig 7.

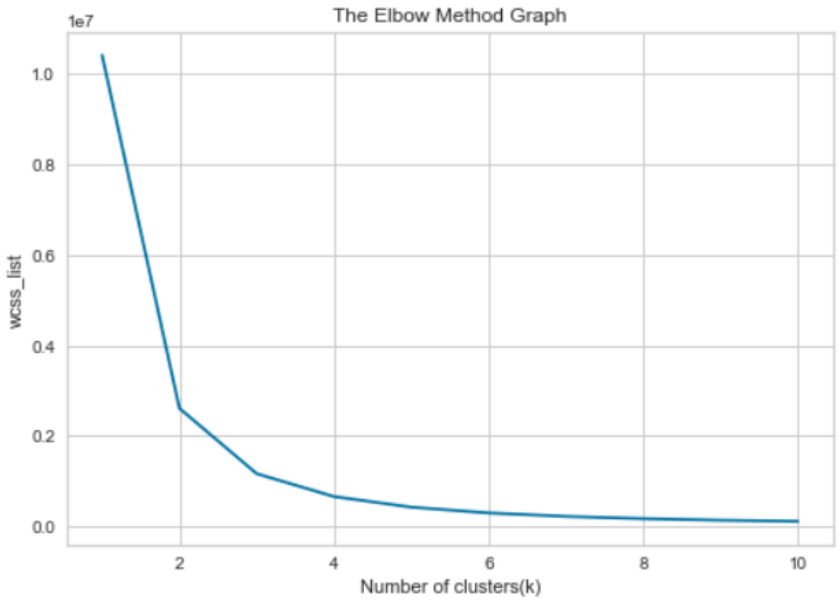

Fig 7 Elbow method for Kmeans w.r.t WSS [13]

With the help of the Elbow, the WSS becomes normal after k = 3 . This point will be considered as a value of k. Visualization of all 3 clusters in 2D and in 3D frames can be shown in Fig 8. and Fig 9 respectively.

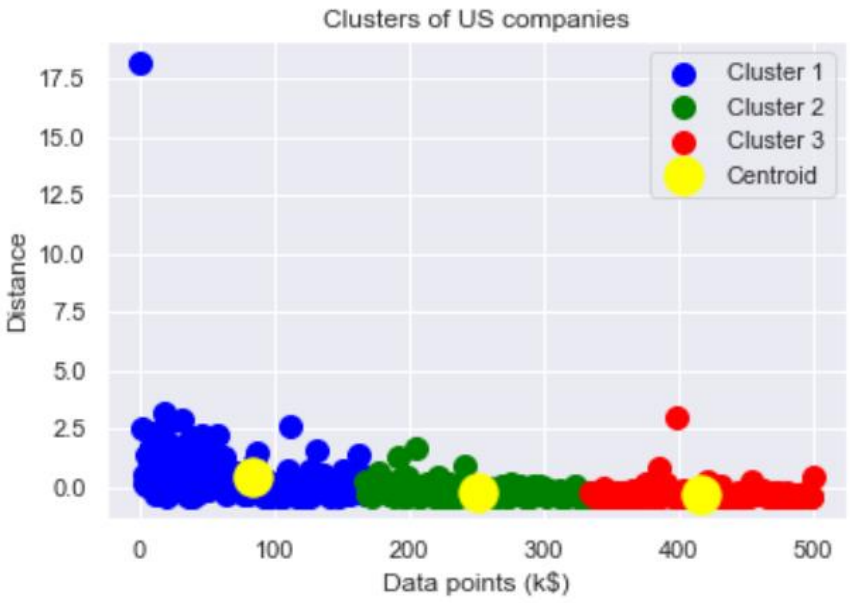

Fig 8. 2D clustering using K means

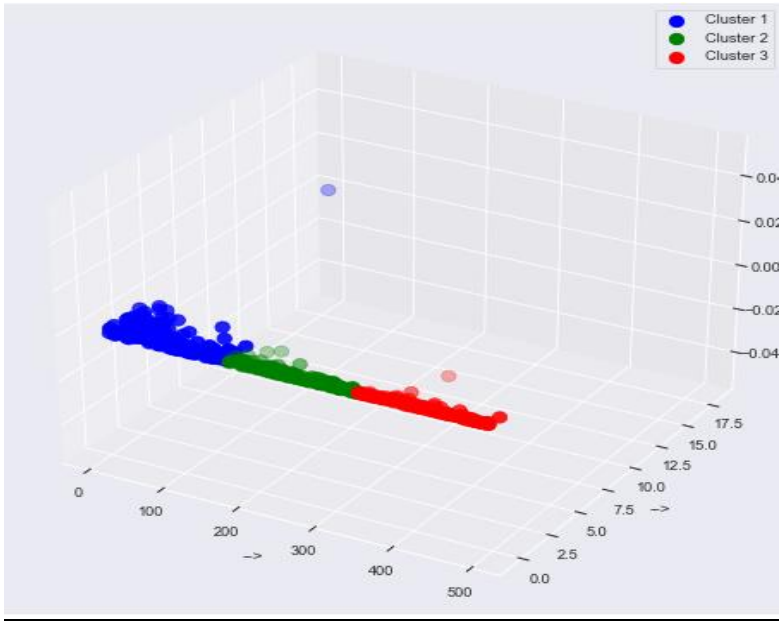

Fig 9. 3D clustering using K means

One of the common challenges with K-Means is that it gives biased result while dealing with categorical columns in the dataset. As, prediction is giving clusters values as 0,1,2 based on ranking of the Fortune 500 standings since it follows Euclidean distance [14] . Hence, K- mode is the best approach to deal as per our result of the prediction.[17]

However, silhouette scores for K- Means model are giving reasonable values based on changing hyperparameter.

Table 1: Performance Comparison of K Means with different K values

Value of k	silhouette score
2	0.6243
3	0.5861
4	0.5657
5	0.5520
6	0.5427

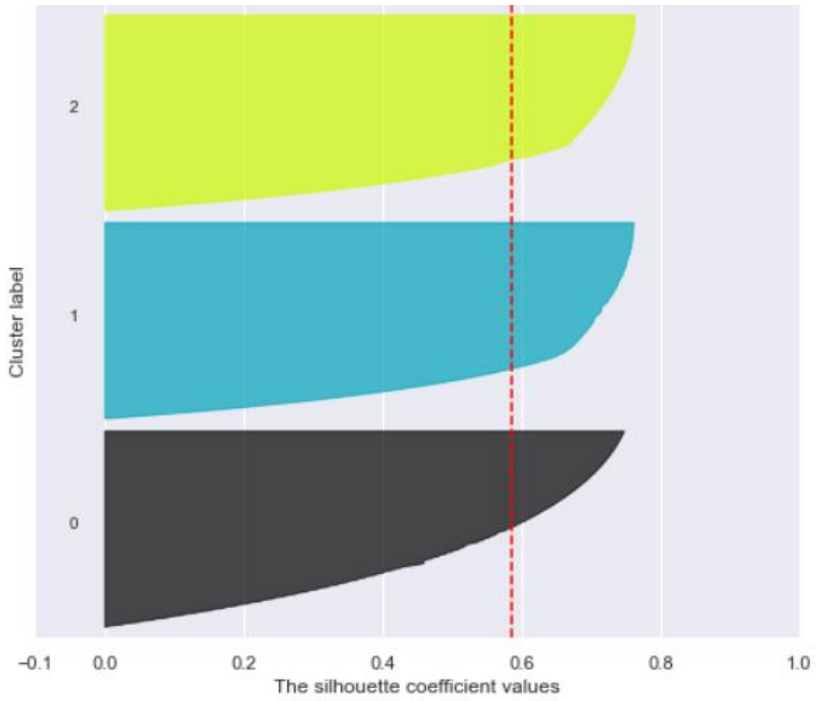

Fig 10 : Silhouette plot for K=3

K-Mode

The second phase of my analysis is characterized is by a **variation of k-means** known as **k-mode**, which is suitable for categorical data. It uses a distance measure which mixes the hamming distance for categorical features and the Euclidean distance for numeric features. Also, result of the prediction is unbiased [17]

TABLE 2. MODEL PREDICTION SCORE

Value of k	K-Means	K- Mode	Hierarchical
2	0.6243	0.63	0.6166
3	0.5861	0.68	–
4	0.5657	0.56	–

Hierarchical Clustering

The third phase of my analysis is characterized by hierarchical clustering. Since it works well for small data sets.

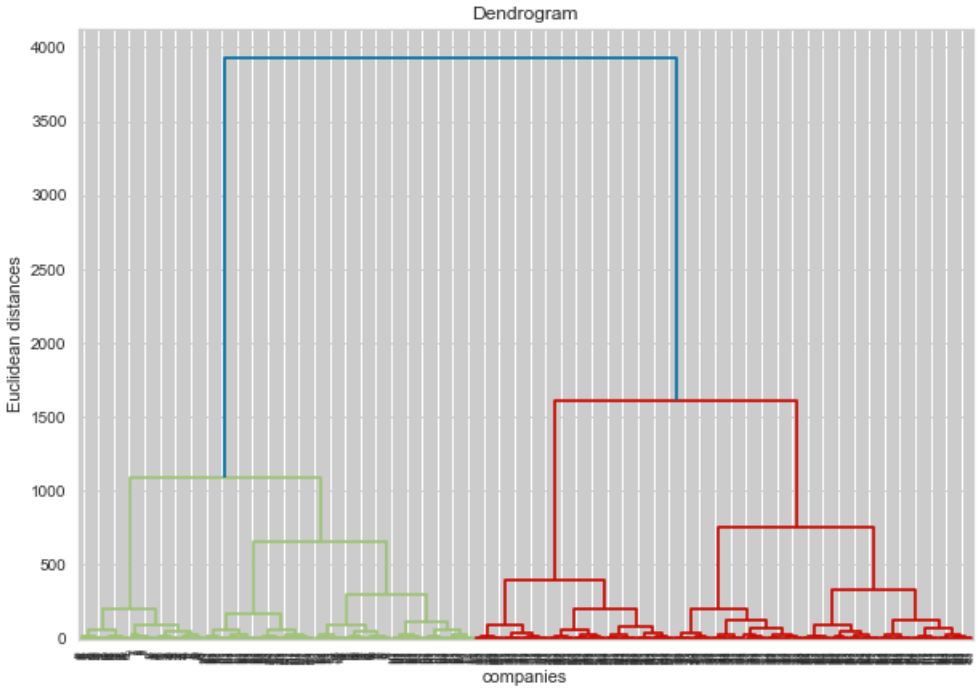

Fig 11 :- Dendrogram of Hierarchical clustering [15],

From the dendrogram, two distinct clusters can be observed. The performance of the model can be again examined with Silhouette Score of 0.6022.

Considering hyperparameters K =3, K-Mode outperformed other clustering model because of its special features to deal with various categorical columns in the dataset.

IV. DISCUSSION

This phase aimed to identify patterns of the clusters in the dataset. The outcome of the clustering was three different company segments.

Clusters 0: - The companies having high revenue, Profit, and employees count. These are considered best companies in Fortune 500 as per analysis. This segment covers more Finance, and Technologies

industries.

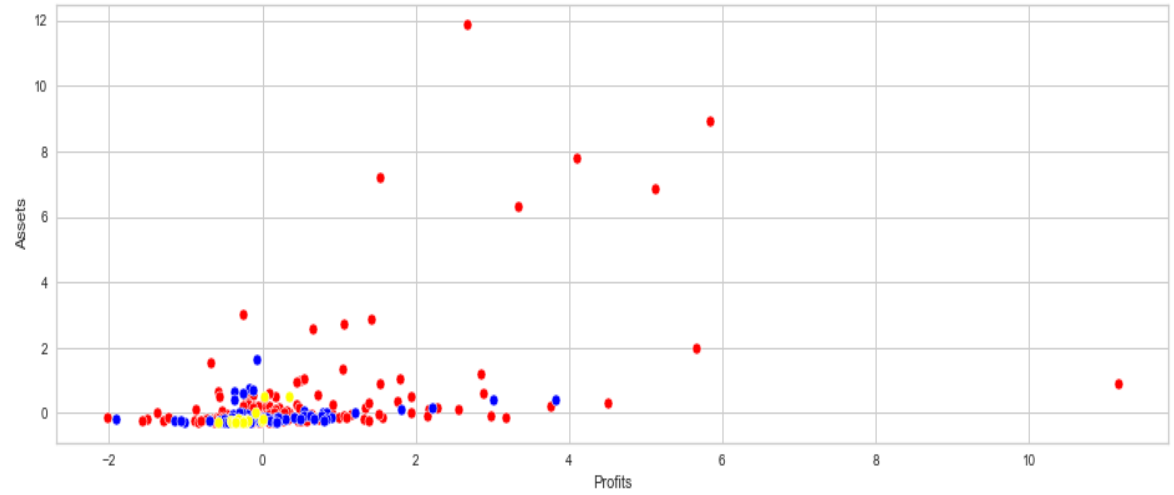

Fig 12 :- After standardization Revenue vs Profit (Cluster 0 – Red /1 – Blue /2- Yellow)

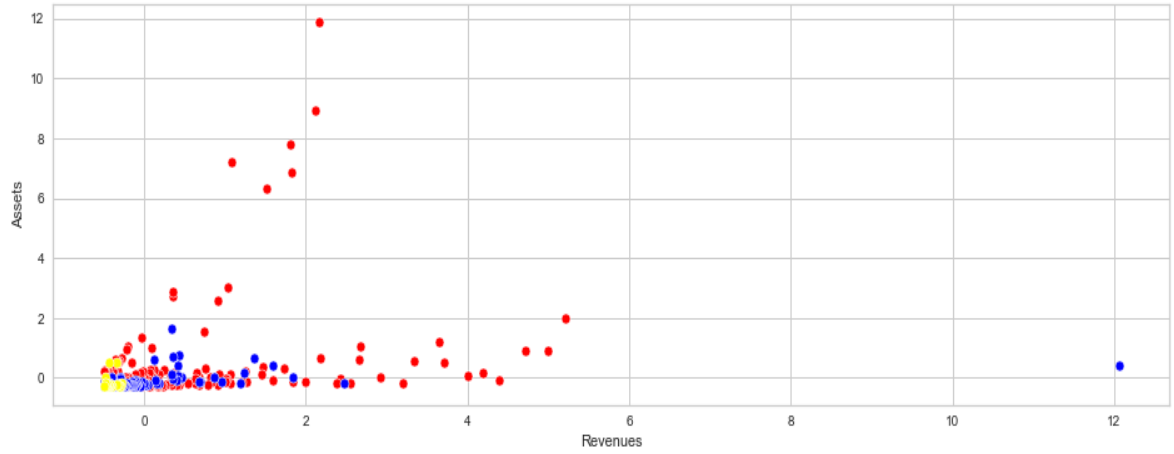

Fig 13 :- After standardization Revenue vs Assets (Cluster 0 – Red /1 – Blue /2- Yellow)

Cluster 1 :- In this cluster, The patterns of the revenue, profit, and employee count are lesser as compared to cluster 0, with huge numbers of companies surrounding in the lower range. These can be considered good companies in Fortune 500 standings as per evidence of our analytics.

Cluster 2:- At last, The patterns of the revenue and Profit are worse as compared other two clusters.

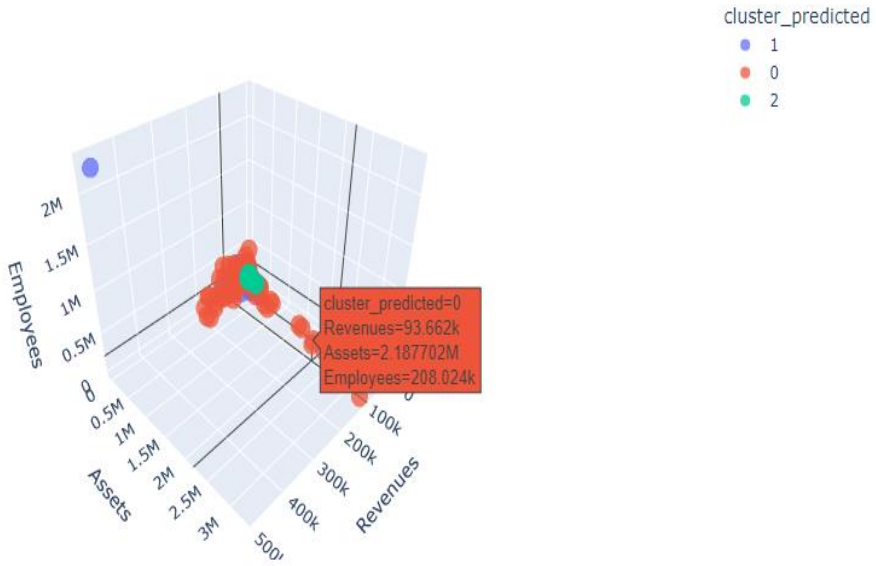

Fig 13 :- comparison of clusters based on Employee count, Revenue , Assets

Fig 13 will provide more background information for the companies' segmentation based on key features like (Revenue, Assets, and Employees count of the company).

According to previous research, the approaches of Fortune 500 companies are ranked by their annual revenues for their respective fiscal years. This approach though presented challenges like the risk of missing a potential financial paradigm, since the data used did not bring out the entire nature of the financial targeting.**[16]**

However, this study has identified companies' segments based on clustering algorithms on different attributes in the dataset. The key attribute of revenue was identified, and the patterns also suggest that Fortune 500 standings depend on key factors like the type of industry and sector, Employee count, and Profit.

This cluster segment of the company is more effective and dynamic since it has used a clustering algorithm, and the features used for the process are the best in the data.

Hence, The Companies can come up with better and more effective strategies using updated financial data and it will help to reduce the risk of missing potential information.

This study will give the companies a wider view to develop more data-oriented companies segments with more key parameters as compared to the use of only one or two factors.

V. REFERENCES

[1] ADAM HAYES, "*Multicollinearity":* https://www.investopedia.com/terms/m/multicollinearity.asp

[2] Jessica Sam Wong, " *Fortune 500 CEOs"*: https://www.stern.nyu.edu/sites/default/files/assets/documents/Jessica%20Wong_Thesis_NYUNY%20Honors%202018.pdf

[3] Jupyter notebook Fortune500. Section: Data Exploration. Link :- Fortune 500 - Jupyter Notebook

[4] Abhijeet Sahu, Zeyu Mao, Katherine Davis, Ana E. Goulart. "*Data Processing and Model Selection for Machine Learning-based Network Intrusion Detection"* - https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9101394

[5] Jupyter notebook Fortune500. Section: Data Cleaning. Link:- Fortune 500 - Jupyter Notebook

[6] *Minorities in leadership and financial performance of their Fortune 500 companies*. - Damineh Mycroft https://core.ac.uk/download/pdf/288853665.pdf

[7] Jupyter notebook Fortune500. Section: - Data Cleaning. Link:- Fortune 500 - Jupyter Notebook

[8] Jupyter notebook Fortune500. Section: EDA. Link: - Fortune 500 - Jupyter Notebook

[9] Jupyter notebook Fortune500. Section: EDA. Link: - Fortune 500 - Jupyter Notebook

[10] Jupyter notebook Fortune500. Section: EDA. Link: - HeatMap Fortune 500 - Jupyter Notebook

[11] Jupyter notebook Fortune500. Section: Feature Engineering. Link: - HeatMap Fortune 500 - Jupyter Notebook

[12] Fan Liu and Yong Deng: "*Determine the Number of Unknown Targets in Open World Based on Elbow Method* - https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8957623

[13] Jupyter notebook Fortune500. Section: KMeans Elbow Method - Fortune 500 - Jupyter Notebook

[14] Jupyter notebook Fortune500. Section: KMeans biased y_prediction - Fortune 500 - Jupyter Notebook

[15] Jupyter notebook Fortune500. Section: - Hierarchical clustering - Fortune 500 - Jupyter Notebook

[16] Dr. LAWRENCE MUCHEMI – "*A CLUSTERING APPROACH TO MARKET SEGMENTATION USING INTEGRATED BUSINESS DATA"* - http://erepository.uonbi.ac.ke/bitstream/handle/11295/160755/Makara%20I_A%20Clustering%20Approach%20to%20Market%20Segmentation%20Using%20Integrated%20Business%20Data.pdf?sequence=1&isAllowed=y

[17] Jupyter notebook Fortune500. Section: - Hierarchical clustering - Fortune 500 - Jupyter Notebook