Trade & Ahead

Unsupervised Learning

Mona Desai

Date:02/25/2023

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- K-Means Clustering
- Hierarchical Clustering
- Appendix

Executive Summary

- **Conclusion, Actionable insights**
  - Performing K-Means & Hierarchical clustering provided with 8 clusters.
  - Cluster 2 in K-Means has the most companies; cluster 3 has most Energy companies
  - Cluster 1 has most companies in it and Cluster 2 has most Energy company using Hierarchical Clustering
  - Volatility and ROE on Cluster 6 in K-Means has a higher Volatility Range; Volatility and ROE on Cluster 4 is highest on Hierarchical
  - PE ratio on Cluster 7 in K-Means is highest ; PE ratio on Cluster 5 in is highest on Hierarchical
  - The P/E ratio of the Energy Sector is the most followed by Information Technology,Real Estate and Health Care
  - The percentage changed in the stock price in 13 weeks are the most in Healthcare Industries, ~9.5% followed by the sector Consumer Staples, ~8.7% and Information Technology, ~7.4%
  - Energy Sector has the Highest Volatility followed by Materials,Information Technology, Consumer Discretionary and Health Care
  - Both clustering techniques, Hierarchical and K-Means, providing similar number of clusters, and very comparable.
  - Median EPS of largest group in both K-Means & Hierarchical clustering is same and is ~2.5.

Executive Summary

- **Recommendations**
  - By K-Means, clusters 4 and 5 followed by 1 and 2 has less volatility. These stocks would be safe investments
  - By K-Means, Cluster 6 has the highest, meaning most risky for Risk Management
  - Consumer Discretionary, Consumer Staples, Financials, Energy and Information Technology sector stocks have the best EPS - Performance of this stocks are higher than other sectors
  - Following sectors seems to lower performance - Telecommunication, Utility, Real Estate
  - Providing data of Price change for longer period would be more useful in this analysis

Business Problem Overview and Solution Approach

- **Problem**
  1. Analyze given list of stocks and its financial data by using Clustering methodology
  2. Identify and classify the stocks based on insights gleaned from their characteristics
  3. Group them or build clusters by their attributes

- **Solution approach / methodology**
  - Exploratory Data Analysis
    - Overview of dataset, Univariate analysis, Bivariate analysis
  - Scaling the Numerical columns
  - K-means clustering
    - Elbow plot, Silhouette scores, Cluster Profiling,
  - Hierarchical clustering
    - Cophenetic Correlation, Linkage, Dendrograms, Cluster Profiling

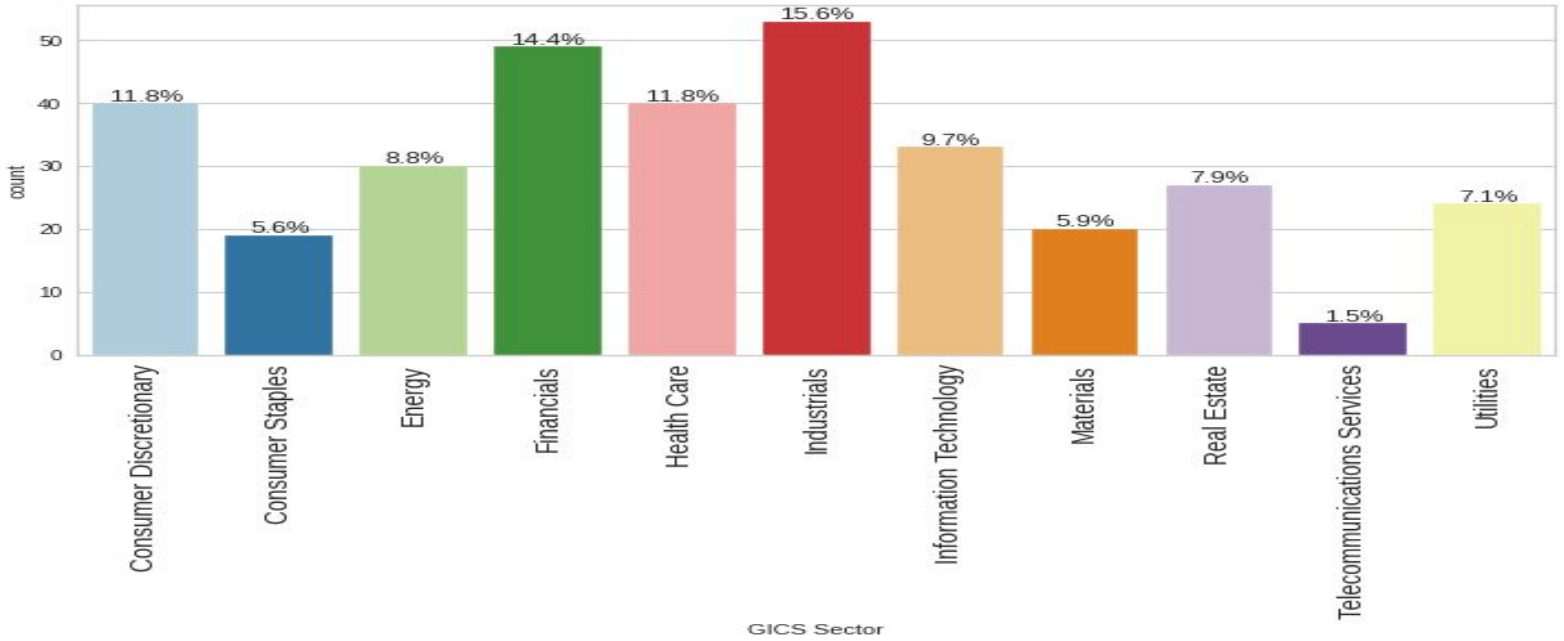
EDA Results

- **Key results from EDA**
  - Data contains 340 unique rows and 15 columns
  - There are 11 Numerical columns and 4 categorical columns
  - Statistical summary suggests,
    - Mean of the volatility is ~1.53 of all stocks
    - A lot of variations among all the numerical columns
    - There are stocks with the min price of $4.5 and max ~$1275
    - Industrial sectors has the most stocks (53/340)
    - Maximum stocks are from Oil & Gas Exploration & Production sub Industry
    - Earning Per Share varies from $-61 to $ 50, with 50% of the stocks has ~$ 2.9 Earning Per share
  - **Univariate Analysis**
    - Columns Estimated shares outstanding, P/E Ratio, Cash Ratio, ROE, Volatility, Current Price are all shows distributed right skewed
    - Columns Price Change is normally distributed

Link to Appendix slide on data background check

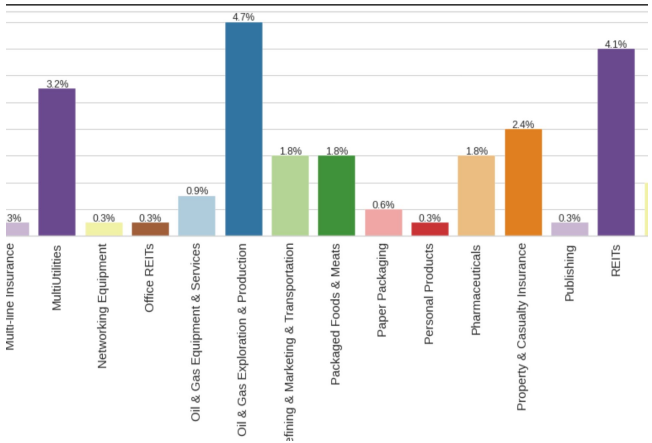
EDA Results continued..

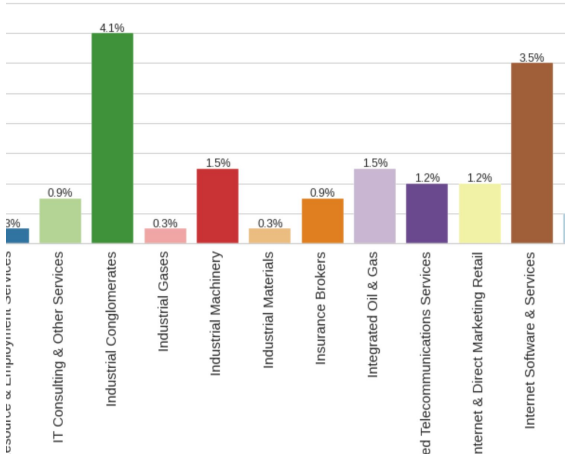
- From GICS sector Barchart, the most, 15.6% companies belongs to Industrial Sector followed by Financials, Consumer Discretionary and Health Care. Sector Information Technology has ~10% companies from the dataset.

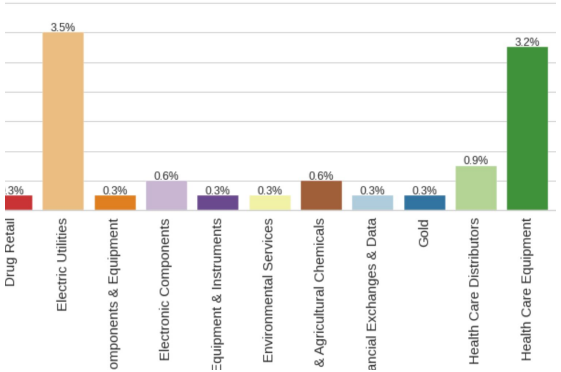


EDA Results continued..

- 4.7% (The most) companies are from Oil and Gas Exploration and Production, 4.1% are from Retailers and Industrial Conglomerates, 3.5% Internet Software and Services and Electric Utilities , 3.2% MultiUtilities and Health Care Equipment GICS Sub Industry



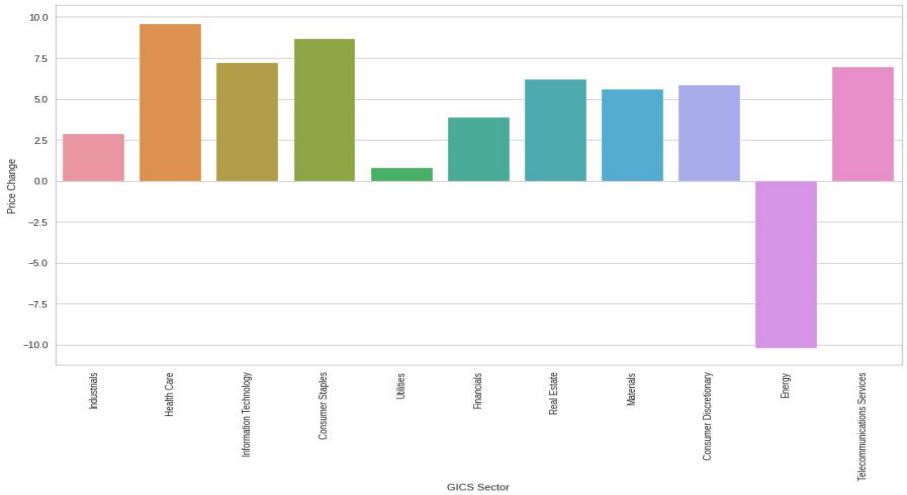




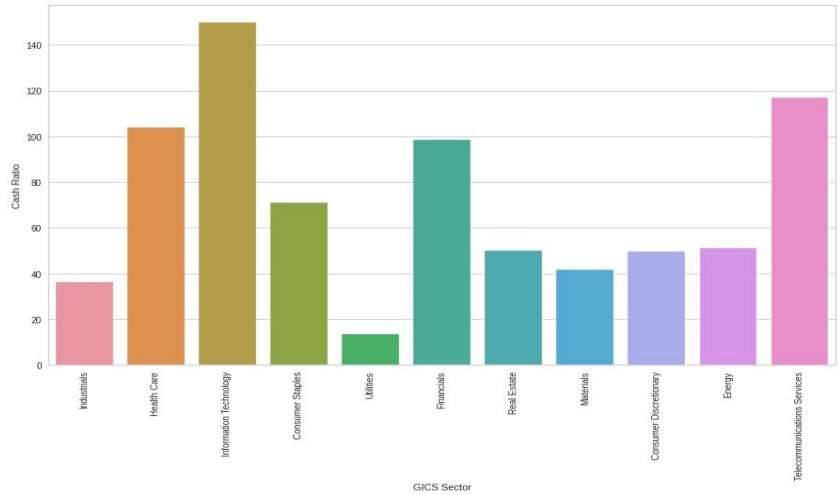
EDA Results continued..

- **Bivariate Analysis**

- The percentage changed in the stock price in 13 weeks are the most in Healthcare Industries, ~9.5% followed by the sector Consumer Staples, ~8.7% and Information Technology, ~7.4%
- The percentage changed -ve in the stock price among Energy sector, ~-ve 10.1%



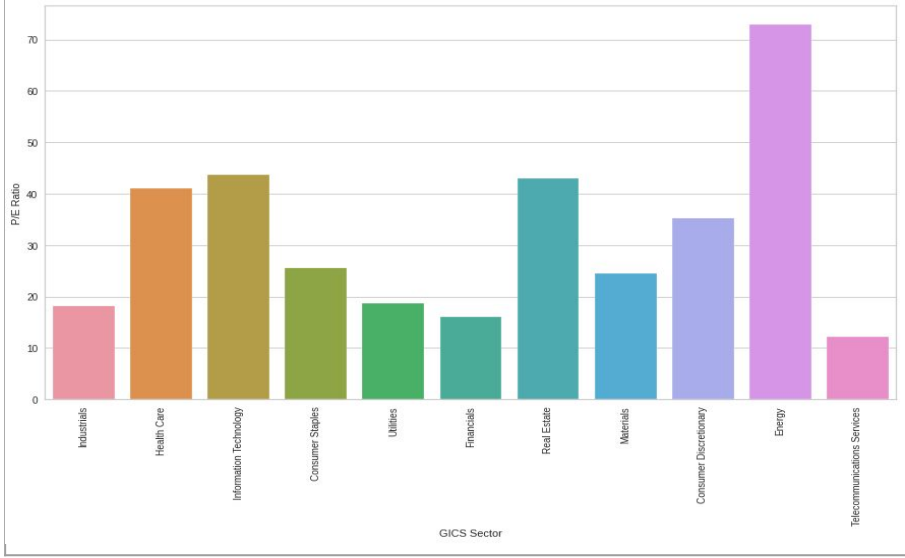
- Information Technology has the most Cash Ratio followed by Telecommunication Industries and Financial sector
- Utility sector has the least cash in hand



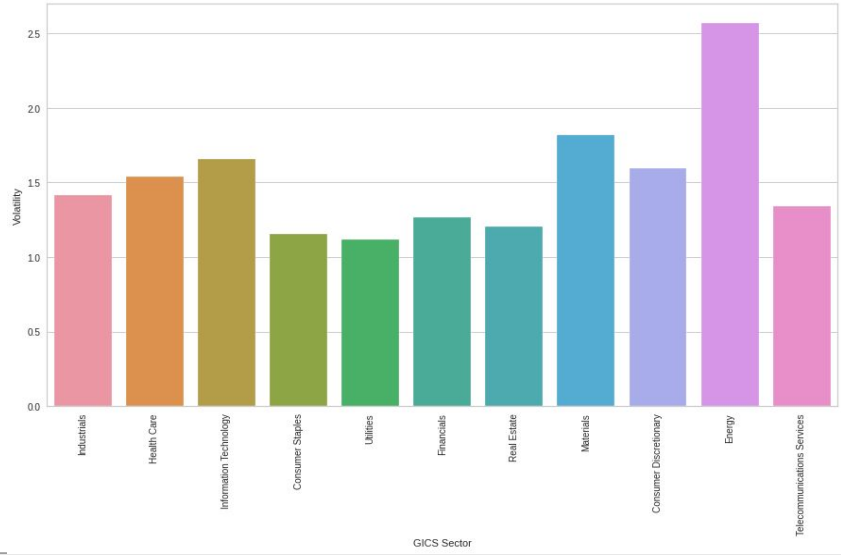
EDA Results continued..

- **Bivariate Analysis**

- The P/E ratio of the Energy Sector is the most followed by Information Technology,Real Estate and Health Care
- Telecommunication service has the least P/E Ratio



Energy Sector has the Highest Volatility followed by Materials,Information Technology, Consumer Discretionary and Health Care



Data Preprocessing

- **Duplicate value check**
  - There is not any duplicate value existing in the dataset
- **Missing value treatment**
  - The Dataset has no missing values
- **Outlier check**
  - All numerical columns has outliers, however we will not treat them as they are proper values
- **Data preparation for modeling**
  - Scaled all numerical columns before clustering, created a dataframe for all scaled numerical columns.
  - Used the scaled value data frame to create clusters
  - Selecting the best value of K using elbow method and Silhouette scores

K-Means Clustering Summary

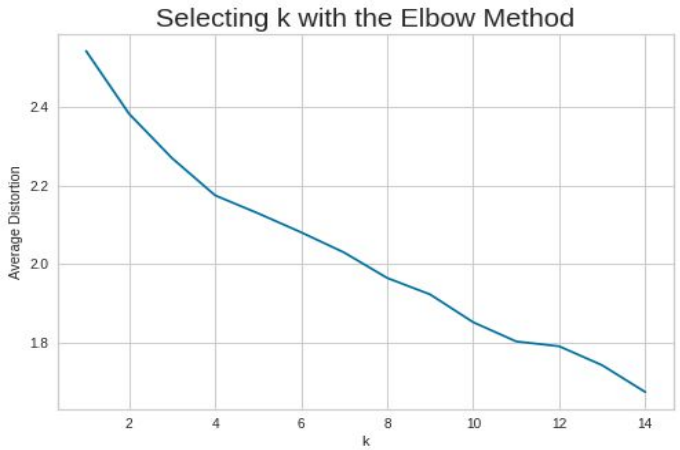
- **Optimal Number of clusters using K-Means**
  - The appropriate value of k from the elbow curve seems to be 4, 8 or 11.
  - We chose to take 8 as the appropriate no. of clusters as the silhouette score is high enough and there is knick at 8 in the elbow curve.
  - There seems to be a knick at 4 and 11 as well but silhouette score is not high enough compare to 8 centers
  - There are also more variations among clusters for K=8.
- **Cluster Profiling**
  - Cluster 2 has the most 264 companies, Cluster 3 has 27, cluster 0 has 20 and cluster 5 has 11 companies
  - 50 companies from Industrial sectors, 45 Financial, 32 Consumer Discretionary are the major sectors in Cluster 2
  - Cluster 3 has 21 companies from Energy sector
  - 8 companies from Information Technology and 5 from Health Care sectors in cluster 0

Hierarchical Clustering Summary

- **Optimal Number of clusters using Hierarchical Clustering**
  - The cophenetic correlation is highest (0.94) for average linkage with Euclidean Distance
  - I move ahead with the “Ward” linkage even though the Cophenetic Coefficient is 0.71 as the variability in the clusters were not enough in linkage method.
  - 8 appears to be the appropriate number of clusters from the dendrogram for ward linkage.
- **Cluster Profiling**
  - Cluster 1 has 285 companies, Cluster 2 has 22, cluster 0 has 12 and cluster 3 has 9 companies.
  - 52 companies from Industrial sector, 44 from financial, 35 from Consumer Discretionary, 34 from Health Care sector and 27 from Information sector in Cluster 1.
  - 20 companies from Energy sector and 1 from Information Technology in Cluster 2

K-Means Clustering Technique

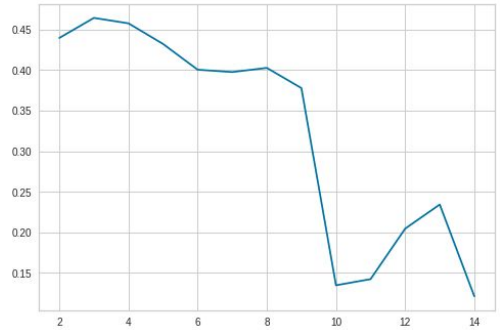
- **Application of K-Means Clustering**
  - We chose to take 8 as the appropriate no. of clusters as the silhouette score is high enough and there is knick at 8 in the elbow curve.
- **Observations using Elbow Curve along with visuals**
  - The appropriate value of k from the elbow curve seems to be 4,8 or 11.



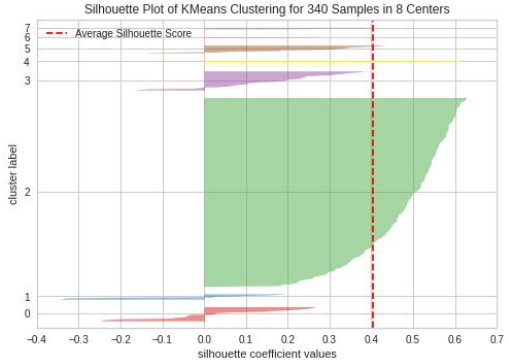
K-Means Clustering Technique continued..

- Observations from Silhouette scores for different number of clusters
  - Silhouette scores; 0.46 for 3 clusters is a good value of k
  - I chose clusters 8 for the Silhouette score 0.4 for more variation

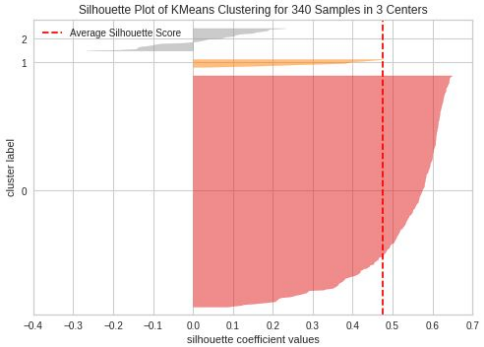
Silhouette Score



8 clusters



3 clusters



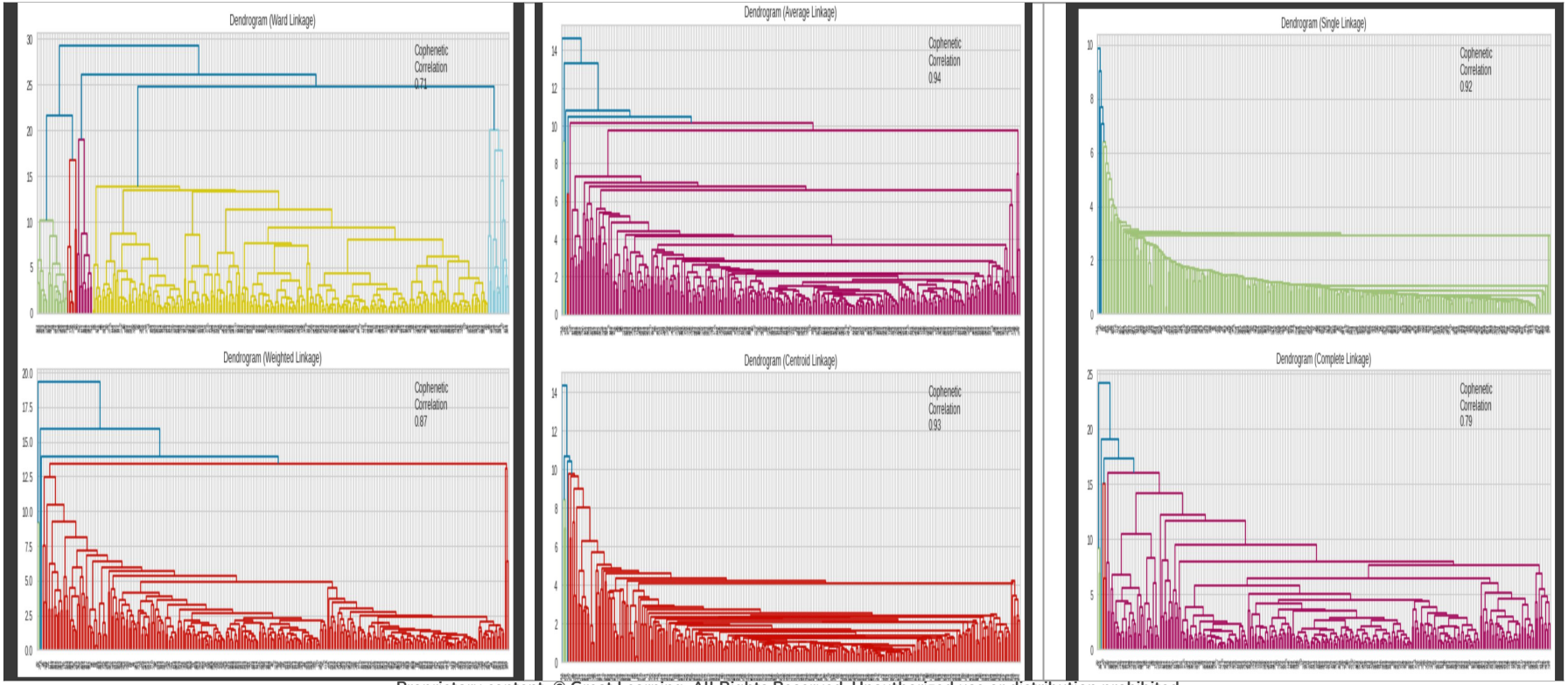
Hierarchical Clustering Technique

- **Application of Hierarchical Clustering**
  - I chose the “Ward” linkage even though the Cophenetic Coefficient is 0.71 which is less than “average” linkage as the variability in the clusters were not enough in “average” linkage method.
  - 8 appears to be the appropriate number of clusters from the dendrogram for ward linkage.
- **Observations using different linkage methods**
  - The Cophenetic Coefficient 0.94 is the highest for the average Linkage

	Linkage	Cophenetic Coefficient
4	ward	0.710118
1	complete	0.787328
5	weighted	0.869378
0	single	0.923227
3	centroid	0.931401
2	average	0.942254

Hierarchical Clustering Technique continued..

- Dendrograms for linkage methods used and their observations



Hierarchical Clustering Technique continued..

- Observations from Cophenetic correlation for different combinations of distance and metrics

```
Cophenetic correlation for Euclidean distance and single linkage is 0.9232271494002922.
Cophenetic correlation for Euclidean distance and complete linkage is 0.7873280186580672.
Cophenetic correlation for Euclidean distance and average linkage is 0.9422540609560814.
Cophenetic correlation for Euclidean distance and weighted linkage is 0.8693784298129404.
Cophenetic correlation for Chebyshev distance and single linkage is 0.9062538164750717.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.598891419111242.
Cophenetic correlation for Chebyshev distance and average linkage is 0.9338265528030499.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.9127355892367.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.925919553052459.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.7925307202850002.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.9247324030159736.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.8708317490180428.
Cophenetic correlation for Cityblock distance and single linkage is 0.9334186366528574.
Cophenetic correlation for Cityblock distance and complete linkage is 0.7375328863205818.
Cophenetic correlation for Cityblock distance and average linkage is 0.9302145048594667.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.731045513520281.
********************************************************************************
Highest cophenetic correlation is 0.9422540609560814, which is obtained with Euclidean distance and average linkage.
```

K-Means vs Hierarchical Clustering

- Comparison of clusters obtained from K-Means and Hierarchical Clustering on various parameters
  - Time to execute - Very similar between both models.
  - Distinct Clusters - Both technique had same number of clusters
  - Observations - Cluster 2 & 3 in K-Means, Cluster 1 & 2 in Hierarchical had most
  - Number of clusters - 8 in both of the techniques.