UNBIASED APPROACH TO PE PORTFOLIO DIVERSIFICATION

AGENDA

1. Business Problem / Data Problem
2. Stakeholder Identification
3. Process Workflow
4. Dataset Characteristics
5. Exploratory Data Analysis
6. Unsupervised Learning Process
7. K-means Findings
8. Reporting For Non-technical Stakeholders
9. Improvements For Future
10. Appendix

BUSINESS PROBLEM / DATA PROBLEM

BUSINESS PROBLEM

- You have been tasked with finding out if it is possible for our fund to have an unbiased approach to portfolio diversification, and discovery of target assets. Traditional methods of new VC investment include:
    - Allocation into specific companies
    - Network investing
    - Partnership deals
    - Traditional forms of research

DATA PROBLEM

- Data containing historical information on private companies including:
    - Company funding history
    - Company disclosed valuation history
    - Company characteristics ie(industry)
    - Company investor information
    - Company operating status (IPO, acquired, operating, not in operations)

STAKEHOLDER IDENTIFICATION

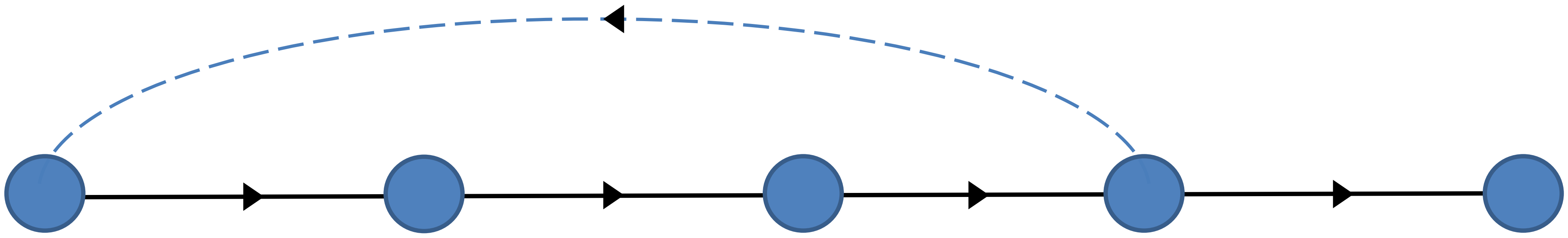
STRATEGY PLANNERS

… who are coming up with new investment strategies

INVESTMENT COMMITTEE

… who require insights into potential target assets

PROCESS WORKFLOW



Collection

- Dataset sourced from Kaggle containing data extracted from Crunchbase API, an industry standard database in the PEVC industry

Data Cleaning

- Dataset was cleaned
  - Unnecessary columns dropped
  - Nulls rows removed
  - Ensured 0 duplicates
  - Data types formatted for analysis

Exploration

- Statical and visual exploratory data analysis was conducted to gain a better understanding of our dataset

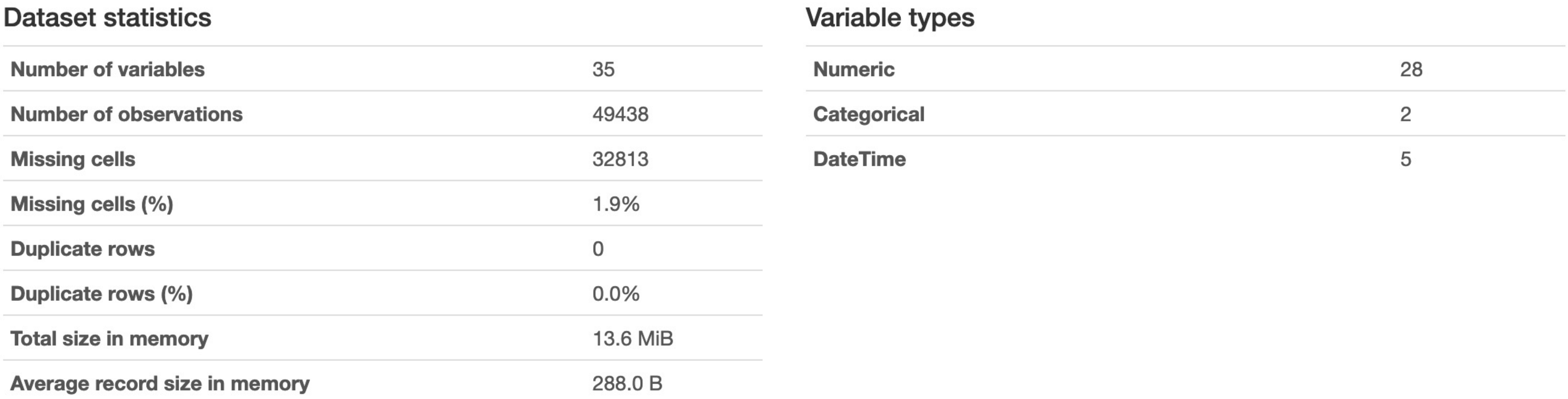
Model Building

- Unsupervised Learning Process (explained in greater detail in following slides)

Reporting

- Reporting of findings in PowerPoint and Tableau for non-technical stakeholders
- Jupyter notebook presented for technical stakeholders

DATASET CHARACTERISTICS



Dataset statistics

Number of variables	35
Number of observations	49438
Missing cells	32813
Missing cells (%)	1.9%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	13.6 MiB
Average record size in memory	288.0 B

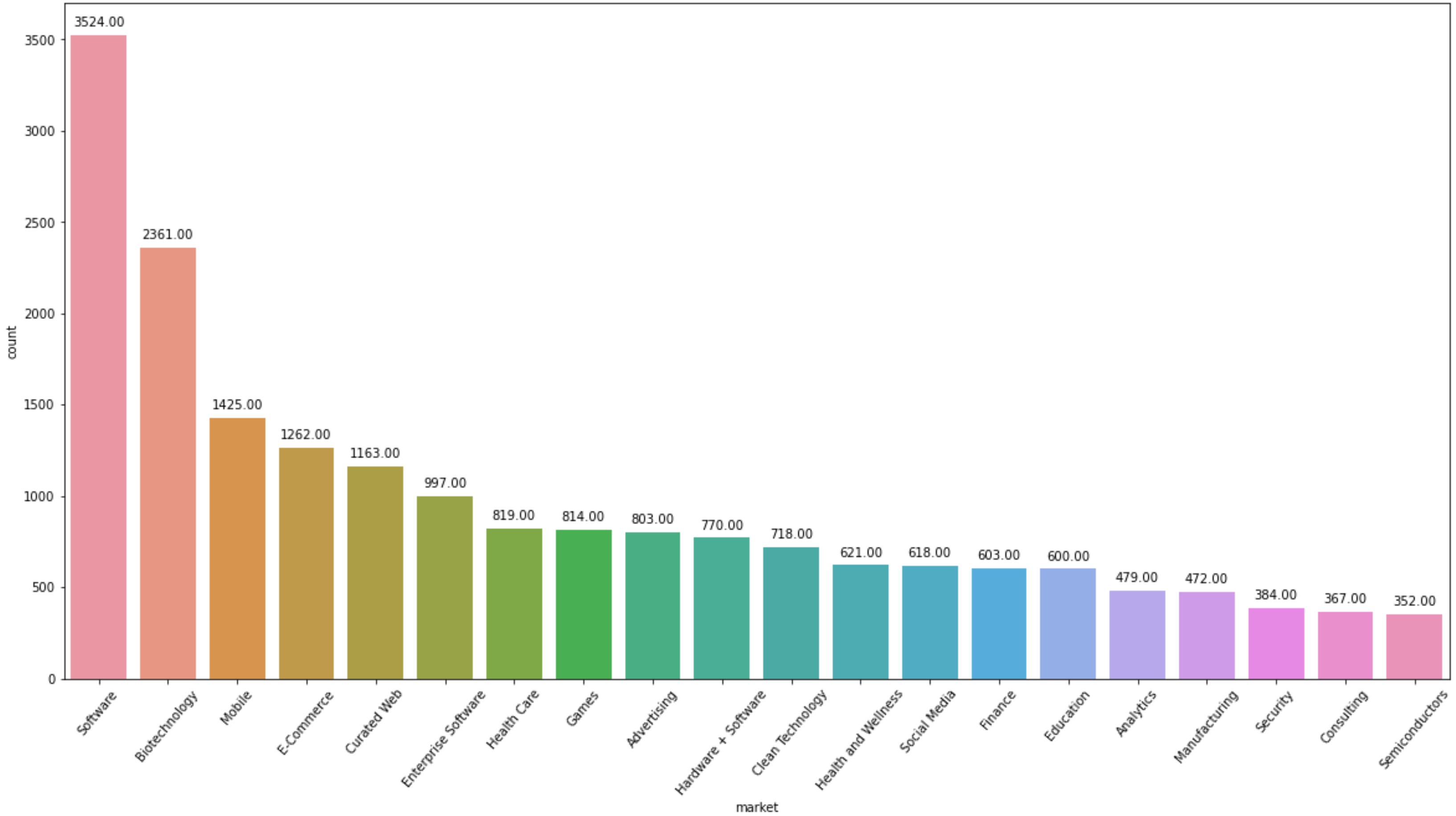
Variable types

Numeric	28
Categorical	2
DateTime	5

- Data available for download on my GitHub: https://github.com/adireksa/iod/raw/main/Projects/Mini%20Project%203/investments_VC.csv

EXPLORATORY DATA ANALYSIS

- Software companies were made up the highest count in the dataset

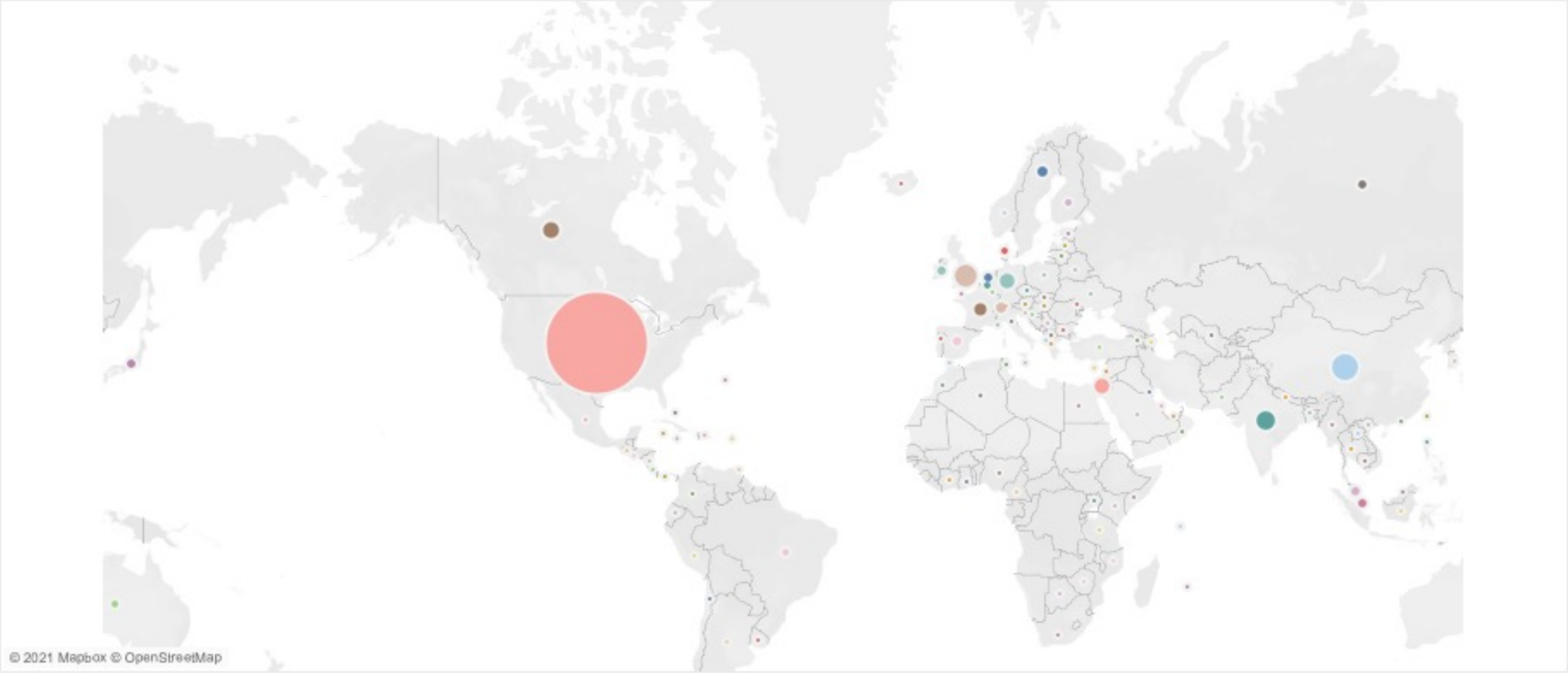


Source: Kaggle

EXPLORATORY DATA ANALYSIS

- USA incorporated companies also accounted for the highest count in the dataset

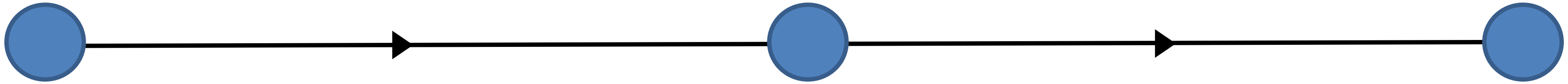
- USA incorporated companies accounted for the largest deal transaction value in the dataset



Source: Kaggle

UNSUPERVISED LEARNING APPROACH TO CLUSTERING PRIVATE COMPANIES

UNSUPERVISED LEARNING PROCESS



PCA

- Principal component analysis was conducted to extract features, and pick optimum number of features (pca inputs)
- This is done while preserving the most important structure or relationships between the variables observed in the data.
- It is a method that uses simple matrix operations from linear algebra and statistics to calculate a projection of the original data into the same number or fewer dimensions.

Elbow Method for optimal k in KMeans

- A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k (# of centroids).
- It is an empirical method to find out the best value of k. it picks up the range of values and takes the best among them. It calculates the sum of the square of the points and calculates the average distance.

K-Means Clustering

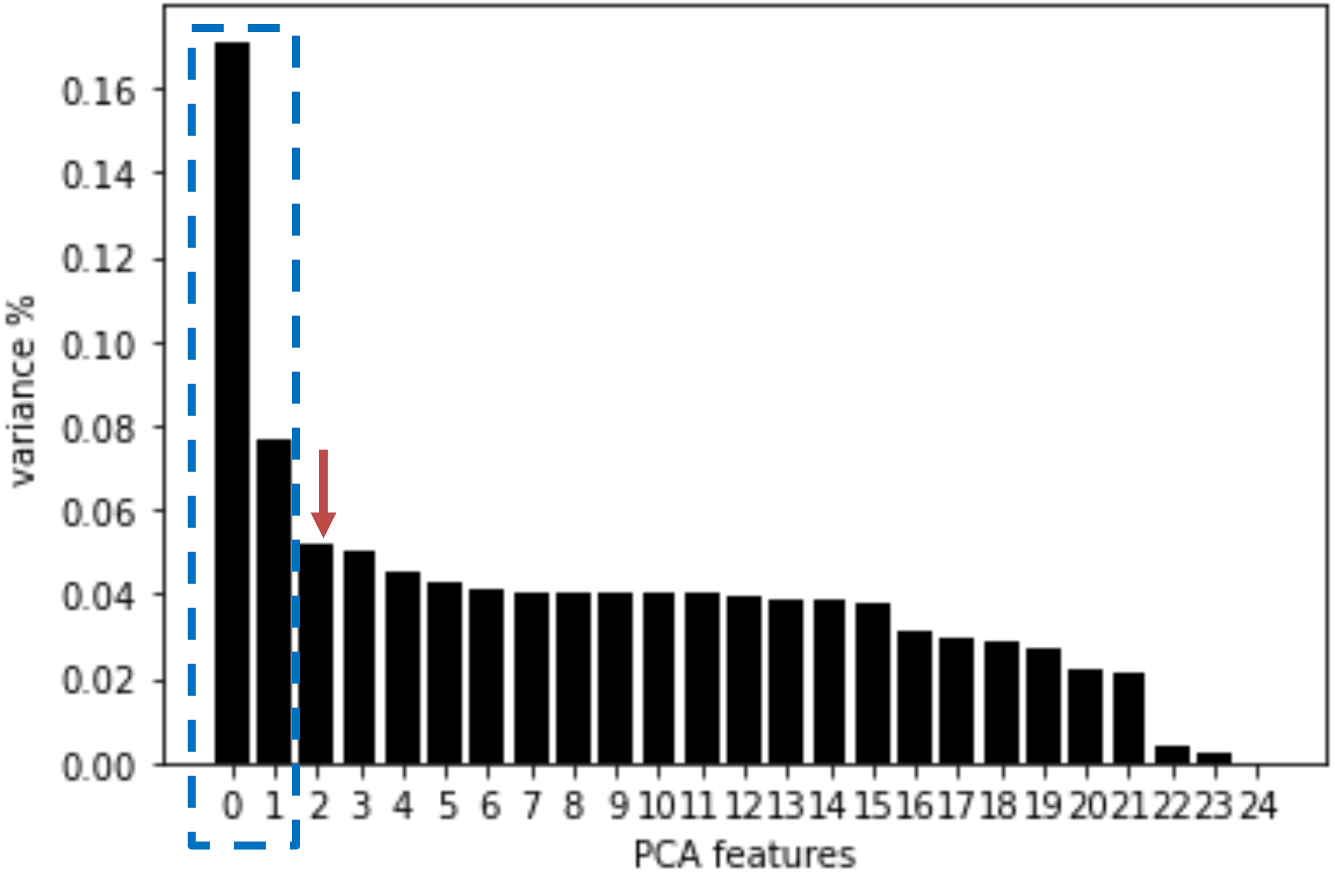
- A K-means clustering algorithm tries to group similar items in the form of clusters. The number of groups is represented by K.
- It finds the similarity between the items and groups them into the clusters. K-means clustering algorithm works in three steps.
  1. Select k Values (elbow method)
  2. Initialize centroids
  3. Select groups

Notes: Centroid: A centroid is the imaginary or real location representing the center of the cluster

Source: Machinelearningmastery.com

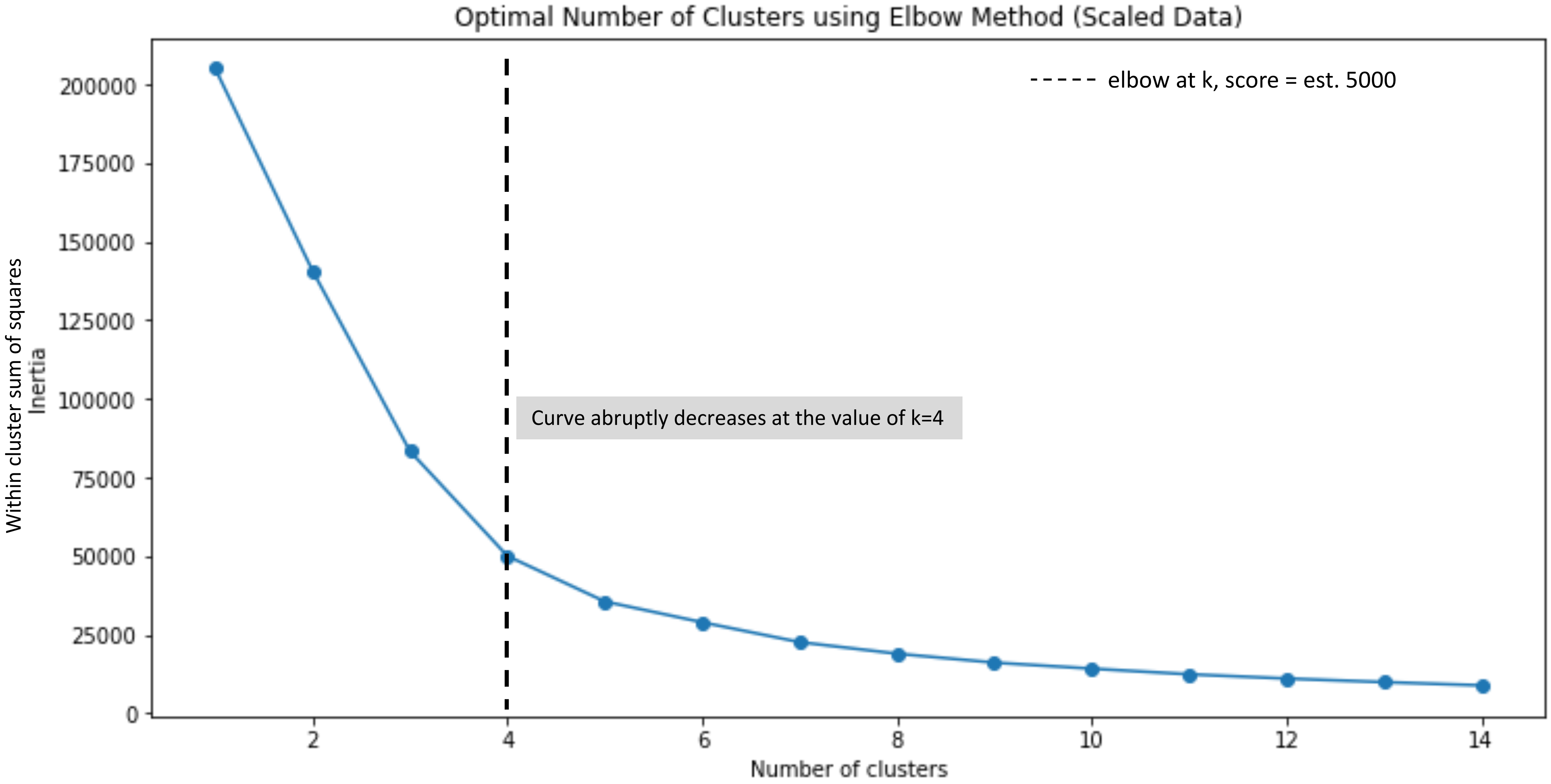
PRINCIPAL COMPONENT ANALYSIS – FEATURE EXTRACTION

- 2 PCA component feature inputs were seen to explain the most variance in the dataset



ELBOW METHOD – OPTIMAL K

- Optimal value of K is 4



- - - - - elbow at k, score = est. 5000

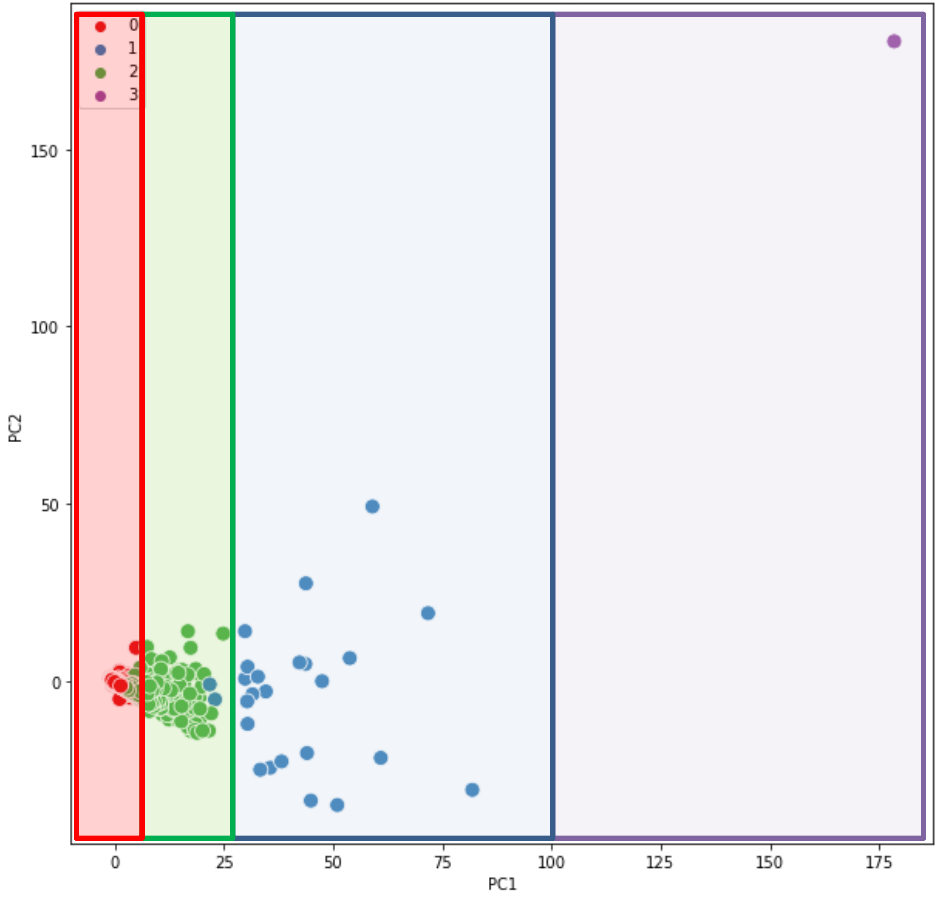
Curve abruptly decreases at the value of k=4

K MEANS – FINDINGS

- The clusters were unbalanced using this approach as shown in value counts below

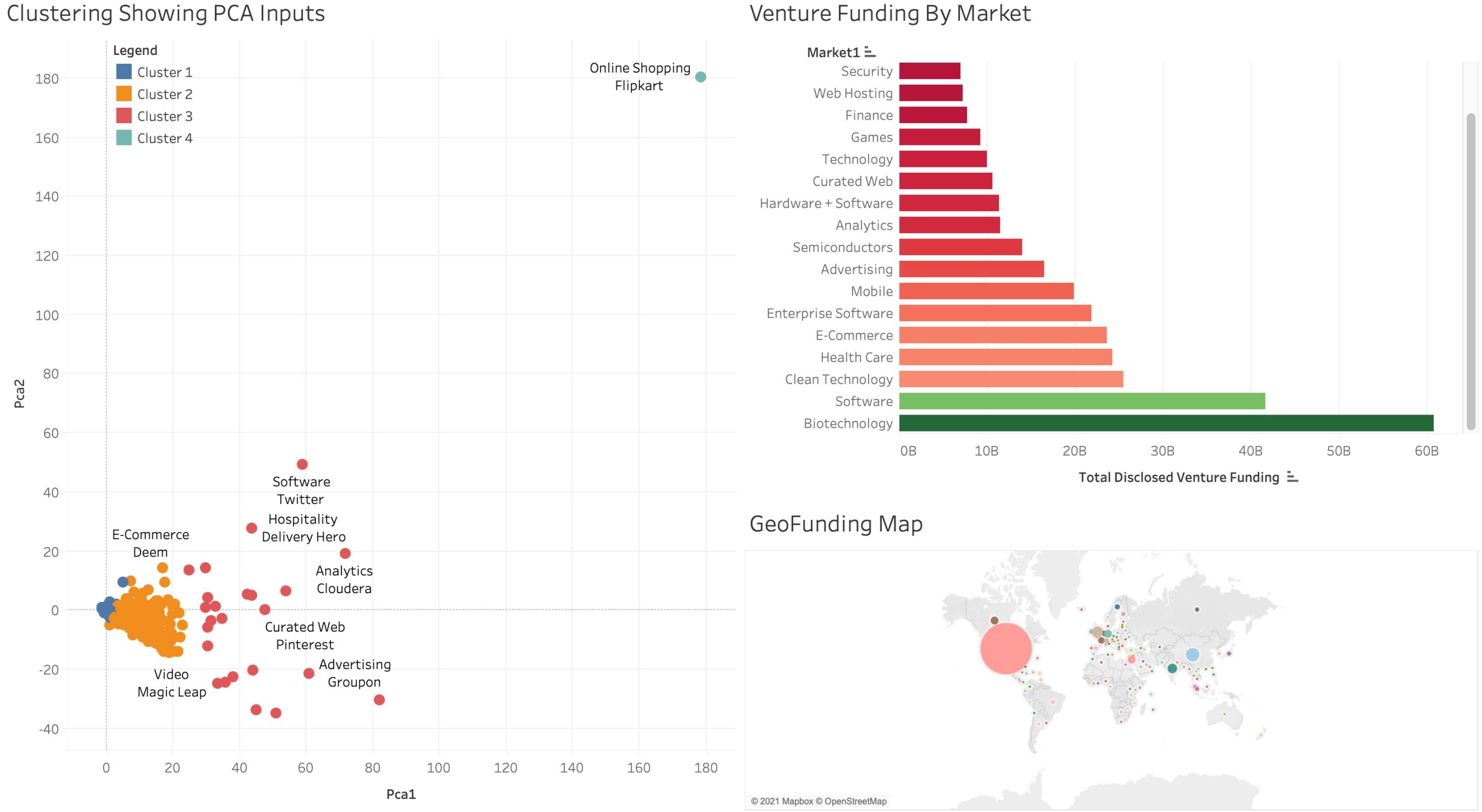
Cluster ID	Count of Cluster
0	1,410
1	31,686
2	47
3	4

KMeans Clusters Derived from Original Dataset with PCA



REPORTING FOR NON-TECHNICAL STAKEHOLDERS

- Our findings were published onto a Tableau dashboard for non-technical stakeholders to use as a discovery tool for potential target assets



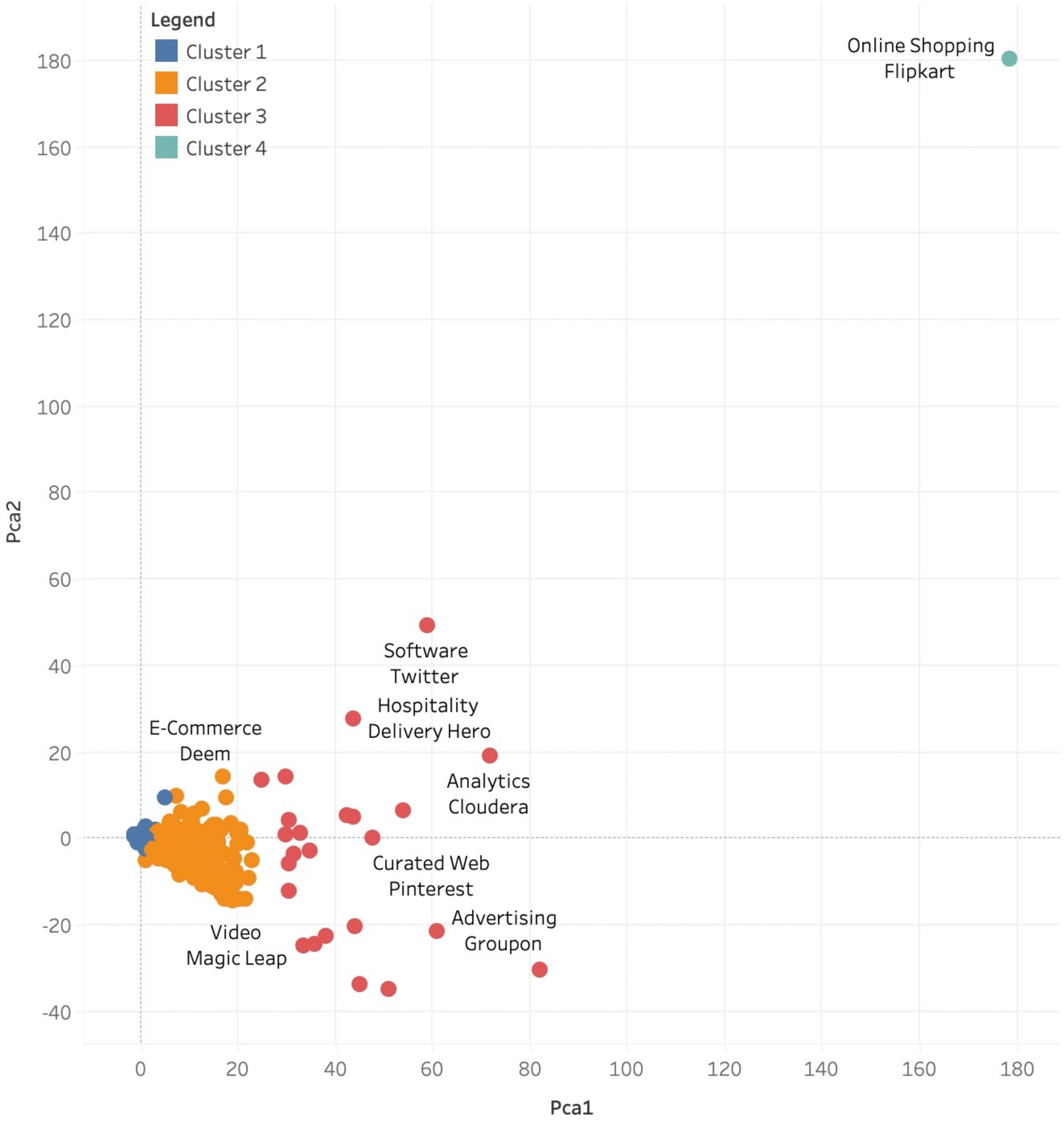
Clustering Showing PCA Inputs

Venture Funding By Market

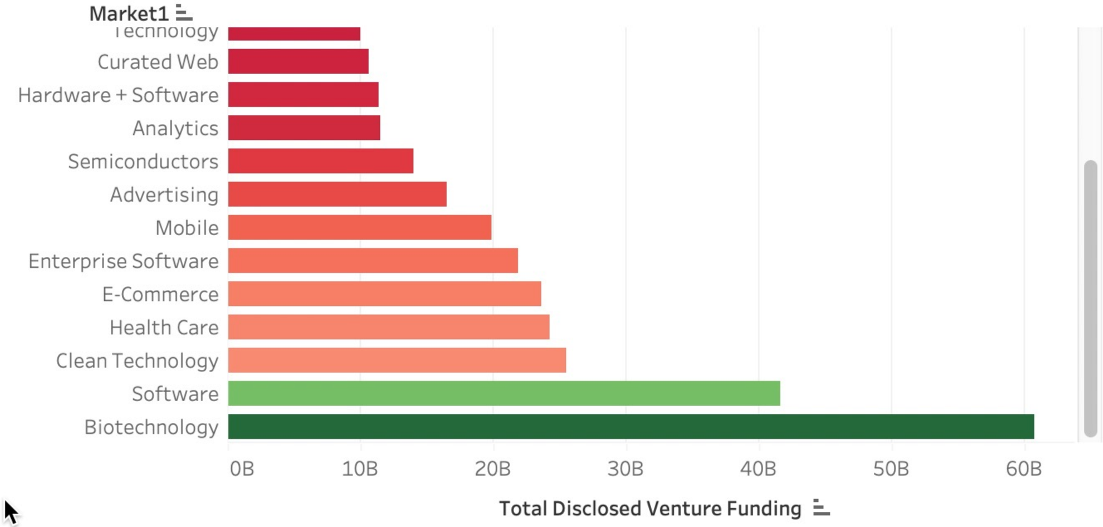
GeoFunding Map

- Dashboard available for access by Tableau account holders until 15-Sep-2021: Link to dashboard

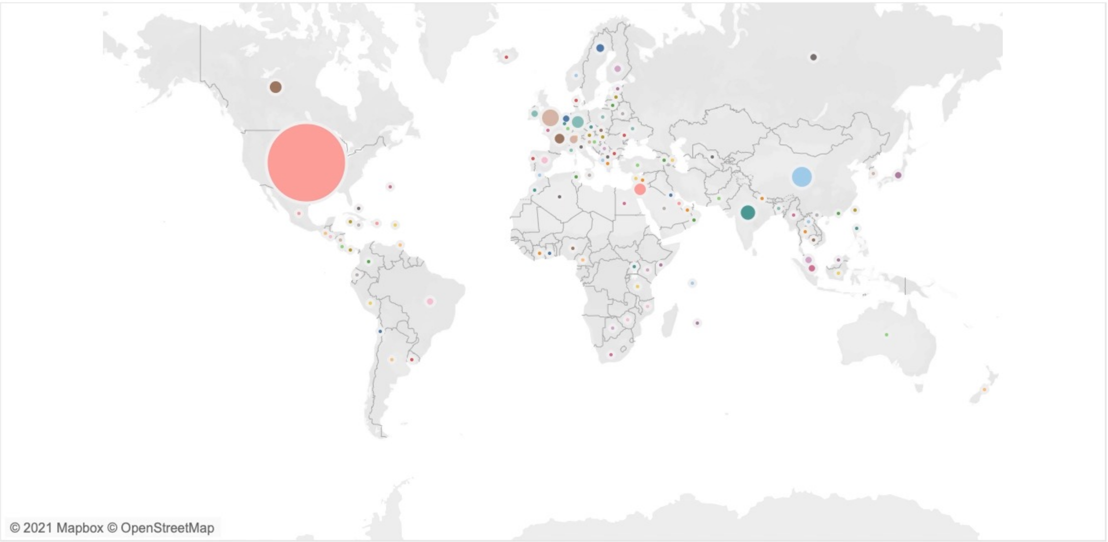
Clustering Showing PCA Inputs



Venture Funding By Market



GeoFunding Map



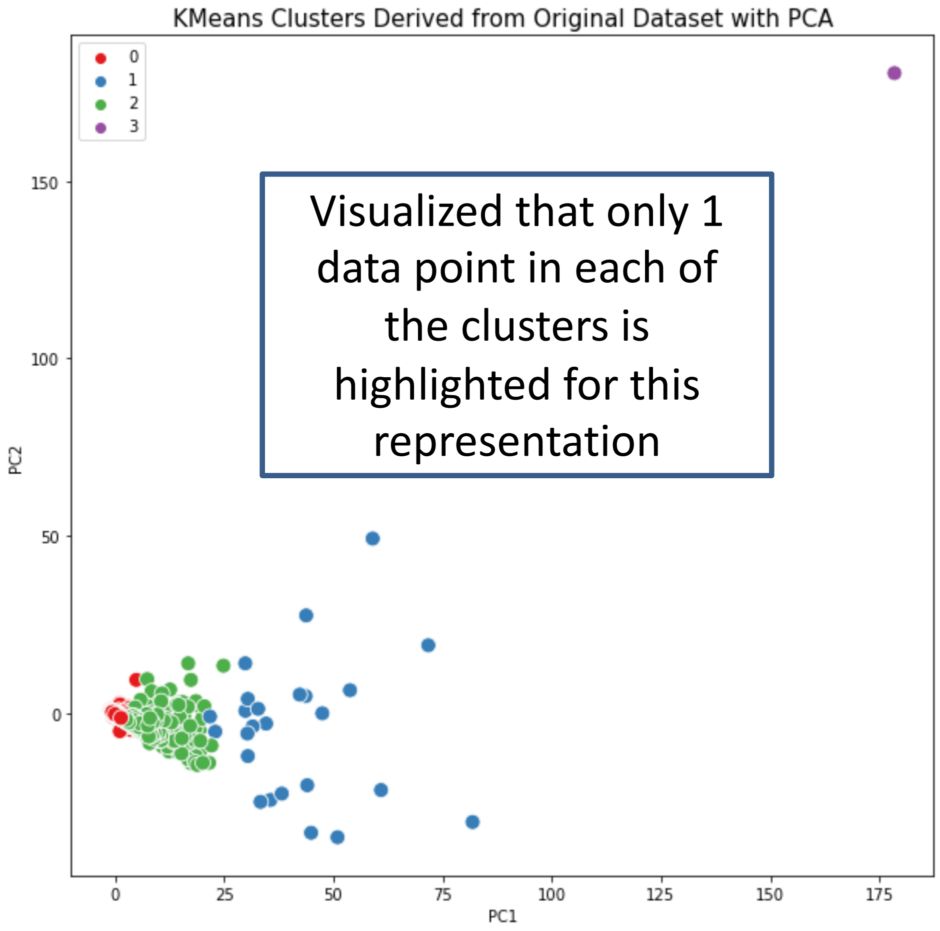
IMPROVEMENTS FOR FUTURE

- Create more features from feature engineering.
- Identified clusters without PCA feature extraction for comparison
- Compare other clustering tools such as DBSCAN, SVM
- Gain more insight using “semi-supervised learning”
    1. Deploy classification to identify company success/failure, by looking at operating status
    2. Clustering dataset

APPENDIX

K MEANS CLUSTERS – GRAPHICAL REPRESENTATION (1 OF 4)

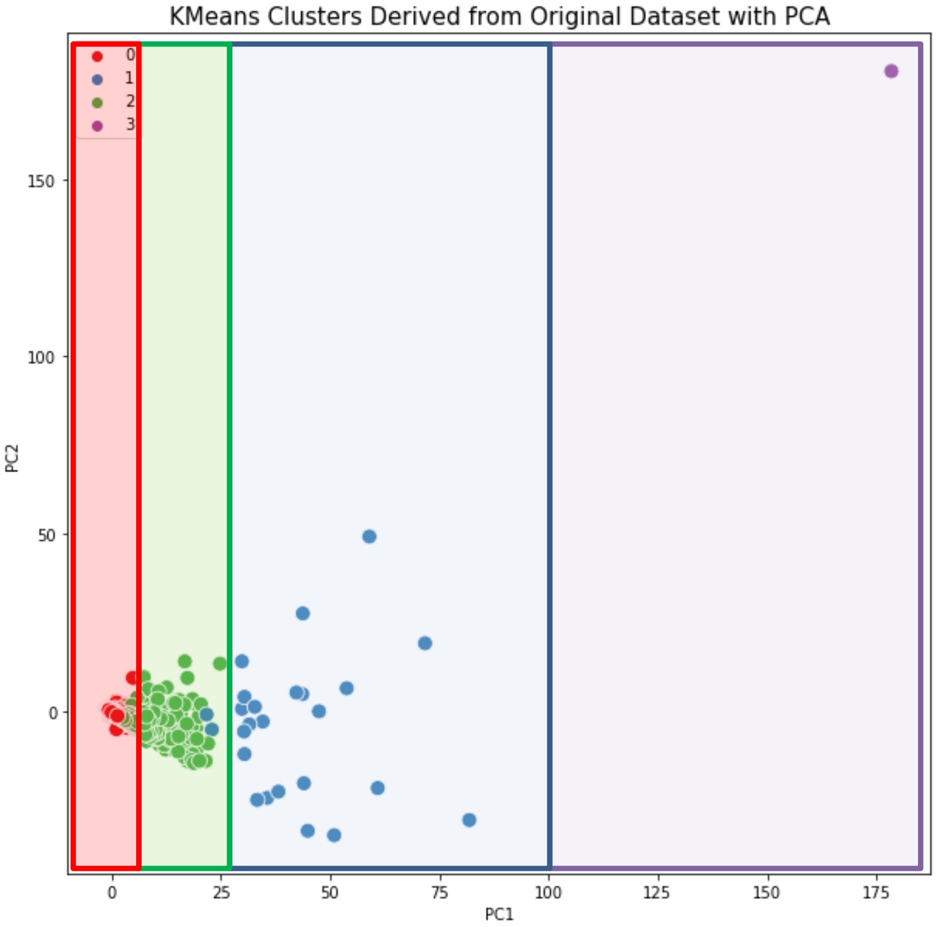
- K initialized. In this case, k=4. and randomly generated within the dataset



Visualized that only 1 data point in each of the clusters is highlighted for this representation

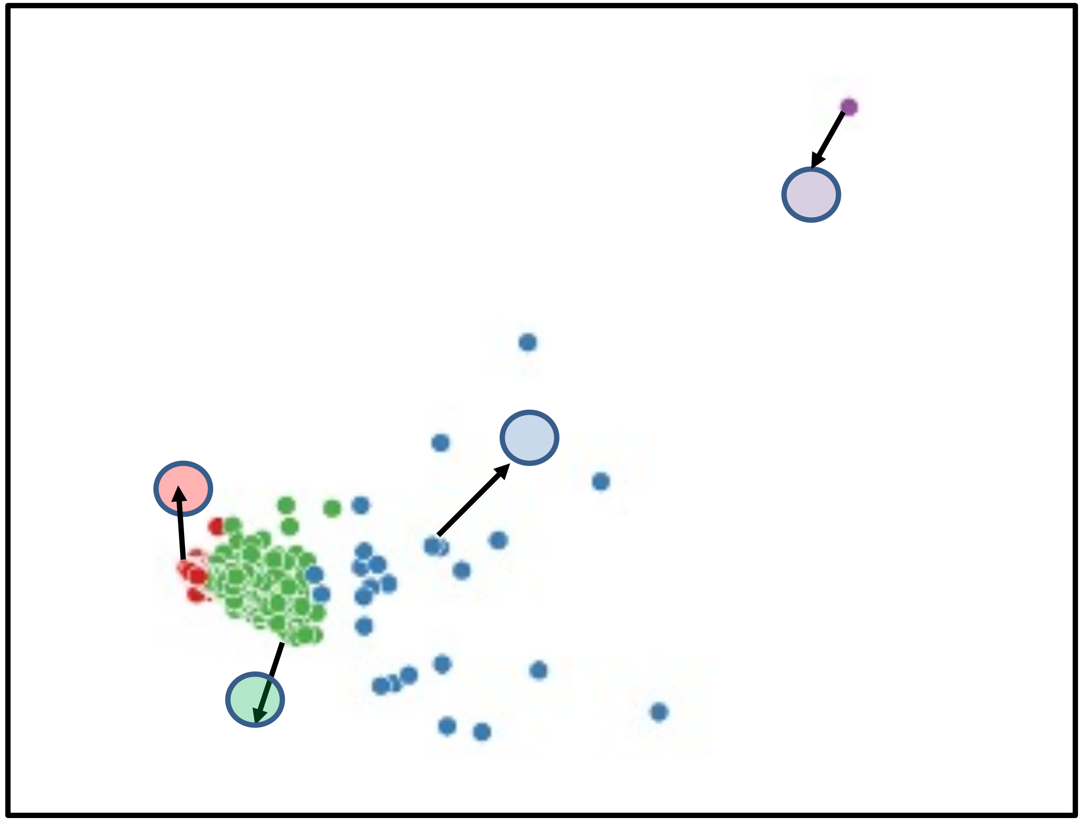
K MEANS CLUSTERS – GRAPHICAL REPRESENTATION (2 OF 4)

- k clusters are created by associating every observation with the nearest mean. The partitions here represent a partition of a plane into regions close to each of a given set of objects.



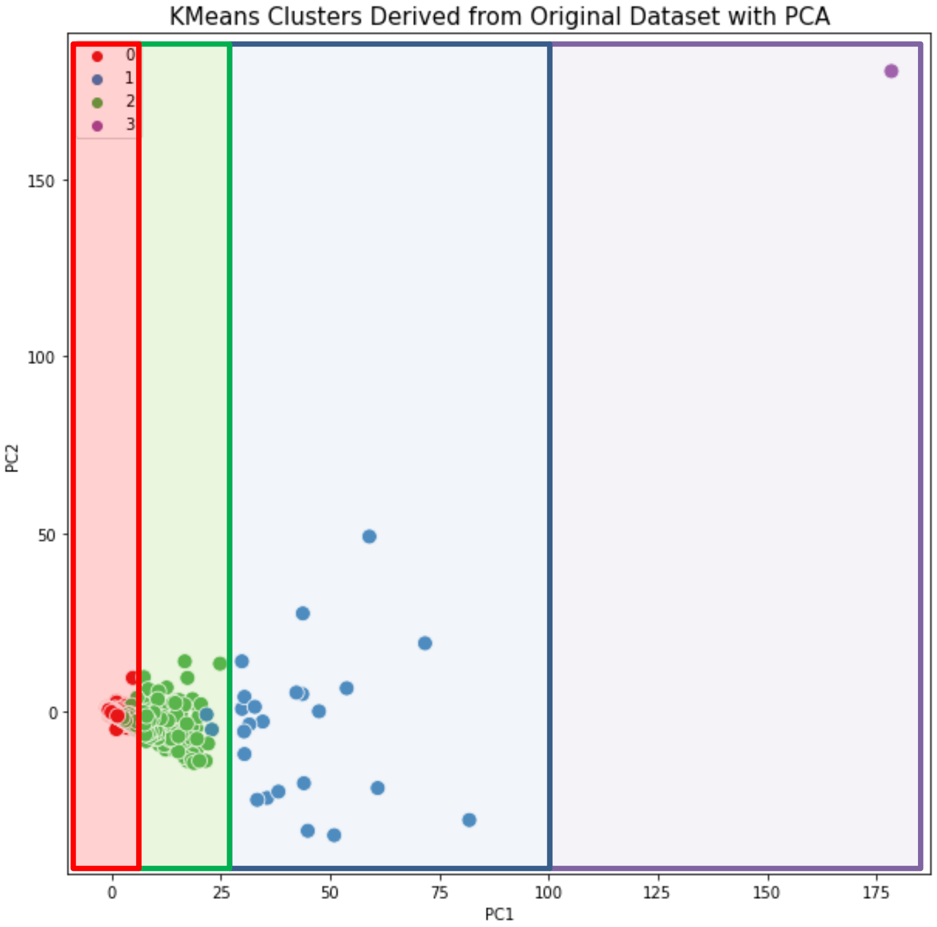
K MEANS CLUSTERS – GRAPHICAL REPRESENTATION (3 OF 4)

- The centroid moves around and the centroid of each of the new k clusters becomes the new k mean



K MEANS CLUSTERS – GRAPHICAL REPRESENTATION (4 OF 4)

- Steps 2 and 3 are repeated until a convergence has been reached
- The objective of k-means clustering is to partition the data set into k clusters, such that each cluster is as “tight” as possible.



Thank you