Clustering Companies by Financial Health Using JPMorgan Chase QuantChallenge 2023 Dataset

PROJECT 2 – UNSUPERVISED LEARNING IN FINANCE
SHRUTHI YENAMAGANDLA

Problem Statement:

- THE GOAL OF THIS PROJECT IS TO GROUP COMPANIES BASED ON THEIR FINANCIAL PERFORMANCE.
- USED CLUSTERING (UNSUPERVISED LEARNING) TO IDENTIFY PATTERNS IN HOW COMPANIES BEHAVE FINANCIALLY.
- DATA COMES FROM THE 2023 JPMORGAN QUANT CHALLENGE AND INCLUDES FEATURES LIKE EQUITY, LIABILITIES, REVENUE, DIVIDENDS, ETC.
- THIS PROJECT FOCUSES ON COMPANY-LEVEL INSIGHTS, NOT STOCK PRICES.

Dataset Overview

15,000 ROWS

12 FEATURES

→ Source: JPMorgan Chase Quant Challenge 2023

→ Features Used:
- liabilities
- equity
- total_assets
- current_assets
- current_liabilities
- total_revenue
- net_income
- dividend
- shares_outstanding

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15000 entries, 0 to 14999
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Date                 15000 non-null  object
 1   Stock                15000 non-null  object
 2   liabilities          15000 non-null  float64
 3   equity               15000 non-null  float64
 4   total_assets         15000 non-null  float64
 5   current_assets       15000 non-null  float64
 6   current_liabilities  15000 non-null  float64
 7   total_revenue        15000 non-null  float64
 8   net_income           15000 non-null  float64
 9   dividend             15000 non-null  float64
 10  shares_outstanding   15000 non-null  int64
 11  price                15000 non-null  float64
dtypes: float64(9), int64(1), object(2)
```

Objective

→ COMPANIES BASED ON FINANCIAL PERFORMANCE BY USING UNSUPERVISED LEARNING TO SUPPORT INVESTMENT ANALYSIS.

WHILE THE DATASET IS ORGANIZED AROUND INDIVIDUAL STOCKS, THE FEATURES REFLECT THE FINANCIAL CONDITION OF UNDERLYING COMPANIES.

Data Preprocessing

Scaling:
- Scaling with StandardScaler
- Distribution is normal – preserve zero mean
- Mean of 0 and SD of 1

Feature Selection:
- Selected only the features that would cluster companies/stocks based on financial health & performance
- liabilites, equity, total_assets, current_assets, current_liabilities, total_revenue, net_income, dividend, and shares_outstanding

Model Selection Rationale

- K-Means Clustering
  - Fast, good for compact clusters
- Hierarchical Clustering
  - Captures structure, good for exploring data

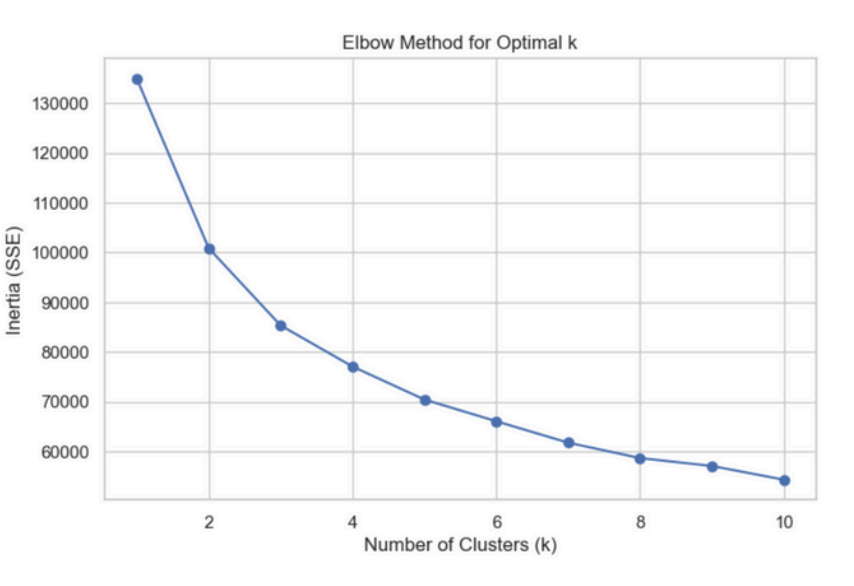
Implementation Approach

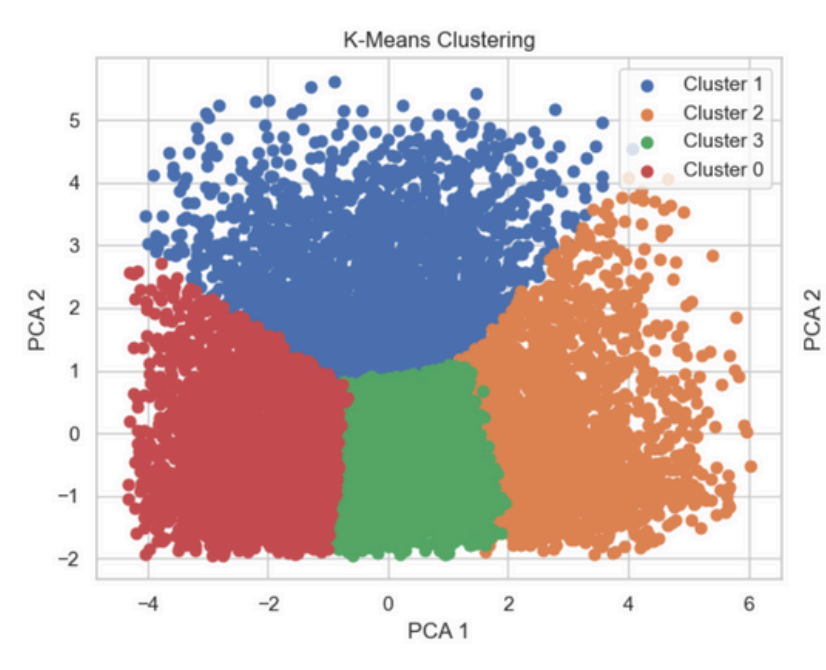
- Used Elbow method for optimal k
- PCA used for 2D visualization
- Silhouette score used to evaluate clustering quality

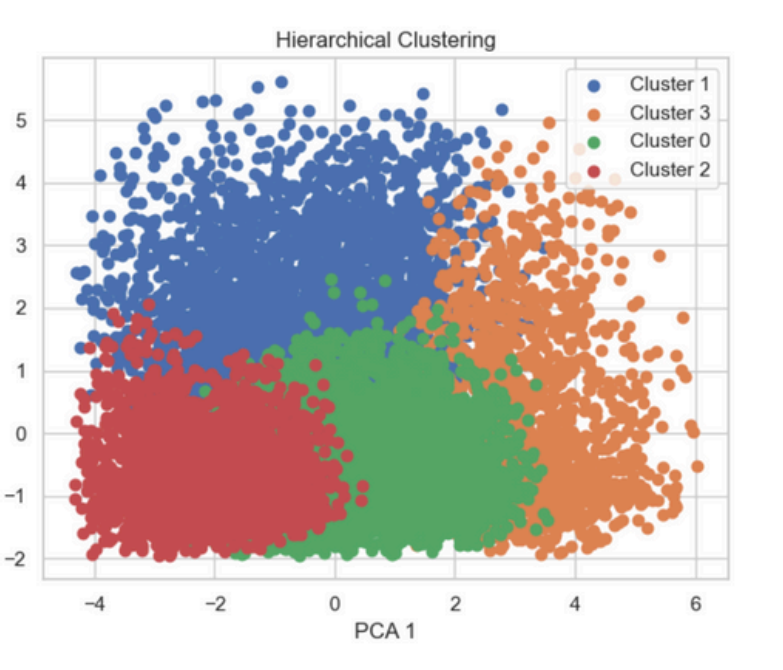
Key Findings

- PCA plots
- Cluster Summary Table
  - Cluster 1 had the strongest financials
  - Cluster 2 had the weakest
- Silhouette Scores:
  - K-Means: 0.168
  - Hierarchical: 0.127

Visualizations







Model Performance Metrics

hierarchical_cluster	liabilities	equity	total_assets	current_assets	current_liabilities	total_revenue	net_income	dividend	shares_outstanding
0	605.043777	585.171609	1190.215386	633.855606	660.835718	494.261314	-67.335028	5.678993	1.047988e+06
1	492.572283	527.916273	1020.488555	534.621129	555.910062	717.134727	218.764025	55.129956	1.049094e+06
2	335.348976	343.553484	678.902460	343.458725	357.791344	524.039264	-74.223652	4.906725	1.054778e+06
3	791.277593	807.309345	1598.586937	1068.924726	1149.771700	535.376841	6.628353	17.351257	1.052632e+06

Interpretation:

NOTE: Assuming all features weigh equally
Cluster 0 (Red): Performs Okay
Cluster 1 (Blue): Performs Best
Cluster 2 (Green): Performs Worst
Cluster 3 (Orange): Performs Good

Best -> Worst: **Blue -> Orange -> Red -> Green**

```
# 2. Silhouette Score (How "good" the clusters are)
from sklearn.metrics import silhouette_score

# K-Means Silhouette
kmeans_silhouette = silhouette_score(X_train_set_scaled, training_set['cluster'])
print(f"K-Means Silhouette Score: {kmeans_silhouette}")

# Hierarchical Silhouette
hc_silhouette = silhouette_score(X_train_set_scaled, training_set['hierarchical_cluster'])
print(f"Hierarchical Silhouette Score: {hc_silhouette}")
```

```
K-Means Silhouette Score: 0.1681088062988105
Hierarchical Silhouette Score: 0.12725926833984824
```

Summary and Challenges Faced

- Applied unsupervised learning techniques - KMeans and Hierarchical Clustering
- Identified four distinct groups of companies
- Grouping was validated using PCA visualizations, Cluster Summaries, and Silhouette Scores
- Can serve as a foundation for further investment analysis

Challenges:
- Choosing k
- "Elbow point" was not very obvious (k=3 or k=4)
- Low Silhouette Scores - not a strong separation

Future Improvements

- Use Dimensionality Reduction to improve separation
- Try Different Clustering Models
- Evaluate with More Metrics
- Transition to Supervised Learning
  - Use clustered labels as input for future price prediction models, combining both clustering and regression