UNSUPERVISED LEARNING FOR INVESTMENT RISK ASSESSMENT

University of Colorado **Boulder**

MS-DS, peculiar.d@colorado.edu

1. INTRODUCTION

- Difficult for investors to accurately assess the risk

2. PROBLEM TO SOLVE

To simplify means of risk assessment

with K-Means unsupervised learning

3. PROJECT DATA

3.1 DATA SOURCE

- 5 years daily closing prices from Yahoo Finance
- Stocks chosen: AAPL, GOOGL, JPM, META, MSFT, TSLA, WMT, XOM
- SPY S&P 500 = Market Index

```
raw_data.head()
```

Date	AAPL	GOOGL	JPM	META	MSFT	SPY	TSLA	WMT	XOM
2020-07-31	103.174988	73.953972	84.504875	252.285950	196.417145	304.028687	95.384003	40.069988	33.220543
2020-08-03	105.774734	73.696022	84.032700	250.585266	207.463821	306.142395	99.000000	40.039024	33.354755
2020-08-04	106.481102	73.225838	83.551765	248.466904	204.350098	307.324829	99.133331	40.763618	34.317909
2020-08-05	106.867065	73.513611	85.003304	247.760773	204.014740	309.233612	99.001335	40.196945	34.617886
2020-08-06	110.595596	74.798904	85.029541	263.832581	207.281799	311.300690	99.305336	40.054504	34.452110

```
raw_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1256 entries, 2020-07-31 to 2025-07-31
Data columns (total 9 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   AAPL    1256 non-null   float64
 1   GOOGL   1256 non-null   float64
 2   JPM     1256 non-null   float64
 3   META    1256 non-null   float64
 4   MSFT    1256 non-null   float64
 5   SPY     1256 non-null   float64
 6   TSLA    1256 non-null   float64
 7   WMT     1256 non-null   float64
 8   XOM     1256 non-null   float64
dtypes: float64(9)
memory usage: 98.1 KB
```

3.2 DATA DESCRIPTION

- no missing values or null values
- META highest standard deviations indicating greater price volatility

```
raw_data.describe()
```

	AAPL	GOOGL	JPM	META	MSFT	SPY	TSLA	WMT	XOM
count	1256.000000	1256.000000	1256.000000	1256.000000	1256.000000	1256.000000	1256.000000	1256.000000	1256.000000
mean	167.905860	130.604268	156.695646	350.255804	318.396989	443.649150	243.691259	56.147963	83.071187
std	36.273604	31.327491	50.751819	161.194474	82.331232	83.726914	67.992640	17.786710	27.638863
min	103.174988	70.049385	81.024651	88.424896	192.454895	301.618561	91.625999	37.783627	25.414991
25%	140.274353	104.766687	122.733952	235.781498	247.085953	383.762642	196.527500	44.447901	53.794540
50%	165.716248	131.136780	139.942612	312.087860	300.394058	421.030548	237.358337	47.602816	96.145061
75%	191.494617	153.929337	190.125843	484.064819	400.295288	510.787209	283.197487	59.664686	105.906893
max	258.103729	205.893341	299.630005	773.440002	533.500000	637.099976	479.859985	104.266106	120.995163

3.2 DATA DESCRIPTION

- stocks have strong positive correlations
- SPY exhibits high correlations = reliable market indicator

```
raw_data.corr()
```

	AAPL	GOOGL	JPM	META	MSFT	SPY	TSLA	WMT	XOM
AAPL	1.000000	0.872149	0.831888	0.759897	0.899263	0.918245	0.494653	0.831585	0.771971
GOOGL	0.872149	1.000000	0.882249	0.850855	0.922173	0.942725	0.560523	0.790185	0.588163
JPM	0.831888	0.882249	1.000000	0.931153	0.896804	0.967957	0.448948	0.946752	0.615042
META	0.759897	0.850855	0.931153	1.000000	0.864529	0.910558	0.367877	0.899927	0.419617
MSFT	0.899263	0.922173	0.896804	0.864529	1.000000	0.956914	0.390478	0.828455	0.719200
SPY	0.918245	0.942725	0.967957	0.910558	0.956914	1.000000	0.484034	0.915494	0.689340
TSLA	0.494653	0.560523	0.448948	0.367877	0.390478	0.484034	1.000000	0.422705	0.132739
WMT	0.831585	0.790185	0.946752	0.899927	0.828455	0.915494	0.422705	1.000000	0.601955
XOM	0.771971	0.588163	0.615042	0.419617	0.719200	0.689340	0.132739	0.601955	1.000000

4. DATA CLEANING

HANDLING MISSING VALUES

- Mean Imputation
- SimpleImputer(strategy="mean")

5. FEATURE SELECTION

β

Beta

$$\beta = \frac{Cov(r_s,r_m)}{Var(r_m)}$$

σ

Volatility

$$\sigma_{annual} = \sigma_{daily} \times \sqrt{252}$$

MDD

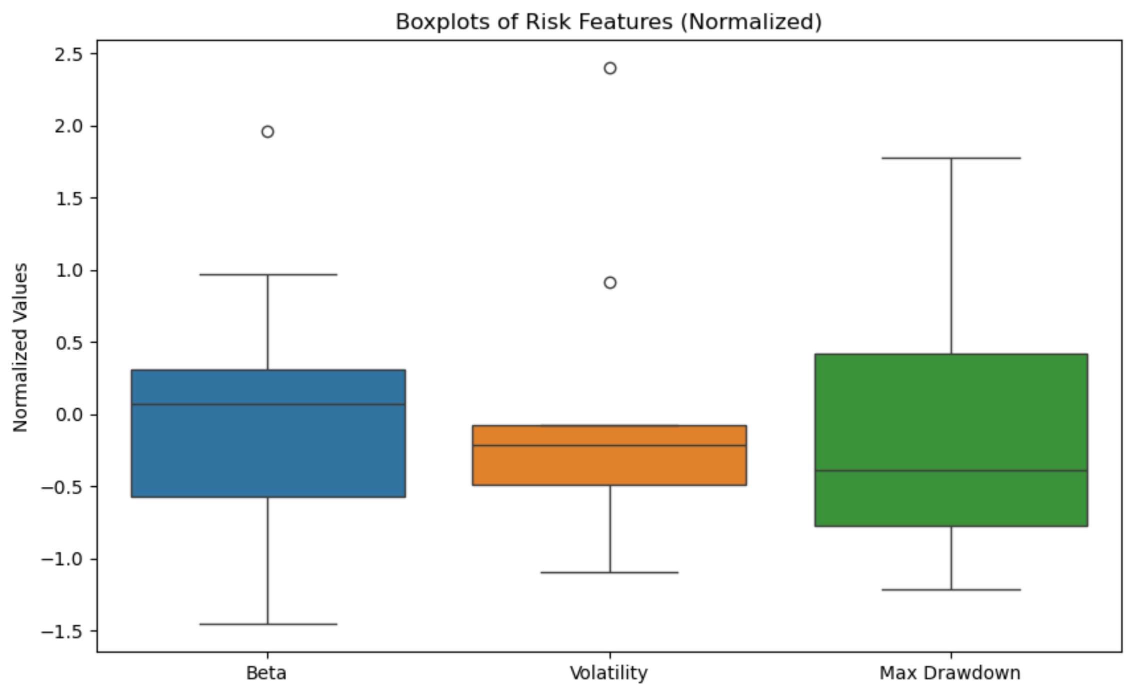
Maximum Drawdown

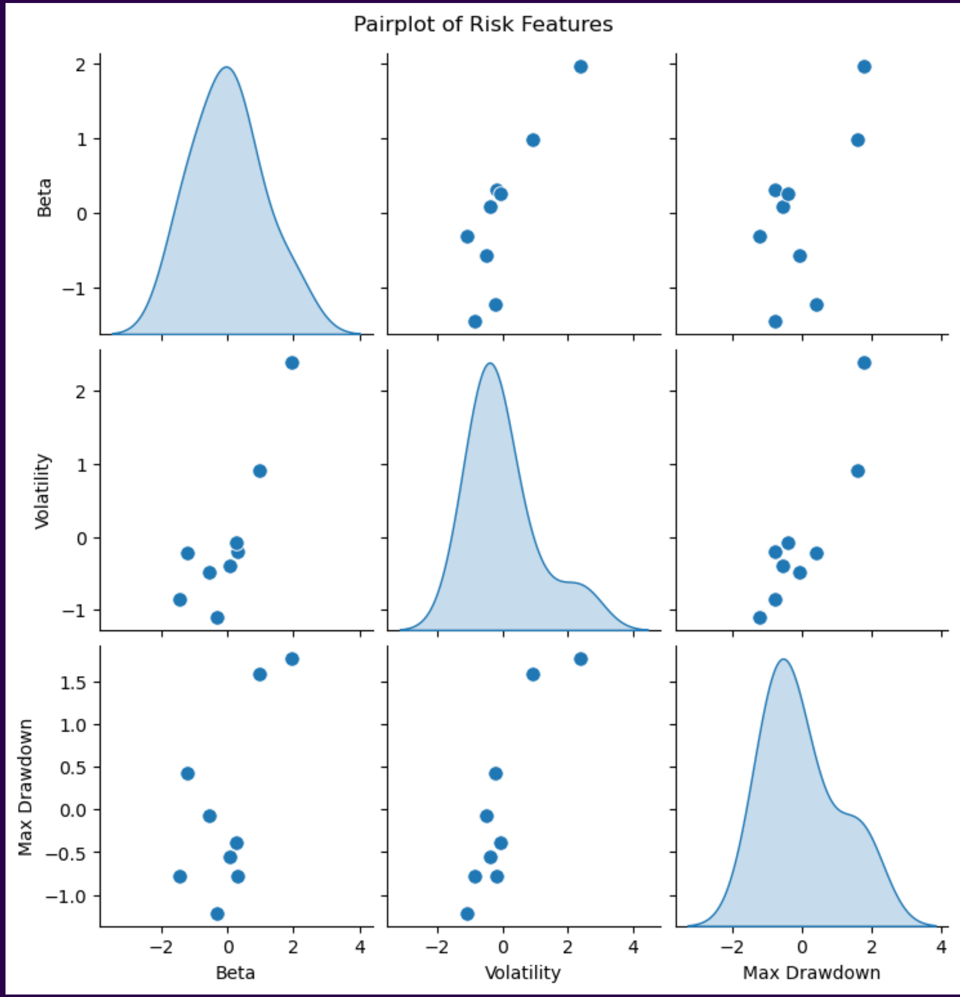
$$MDD = \frac{MaximumValue - MinimumValue}{MaximumValue}$$

6. DATA PREPROCESSING

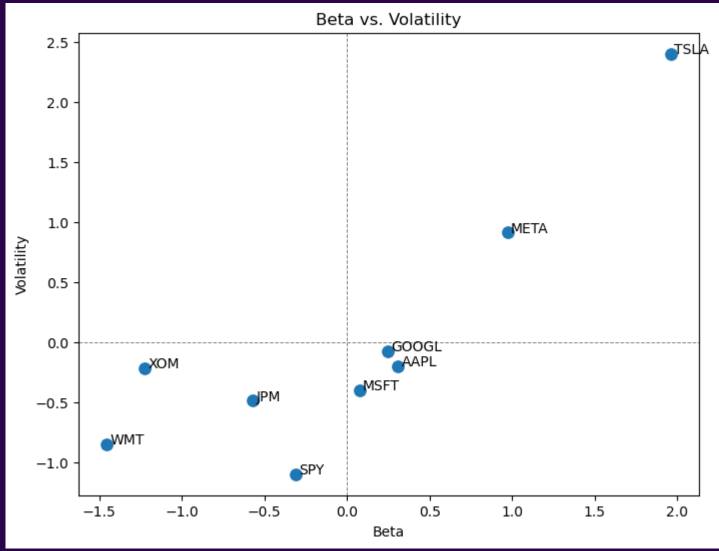
- Convert Raw Prices to Percentage Returns and drop Null
  - `percentage_returns = data.pct_change().dropna()`
- Computing the Features
- Normalization
  - `StandardScaler()`

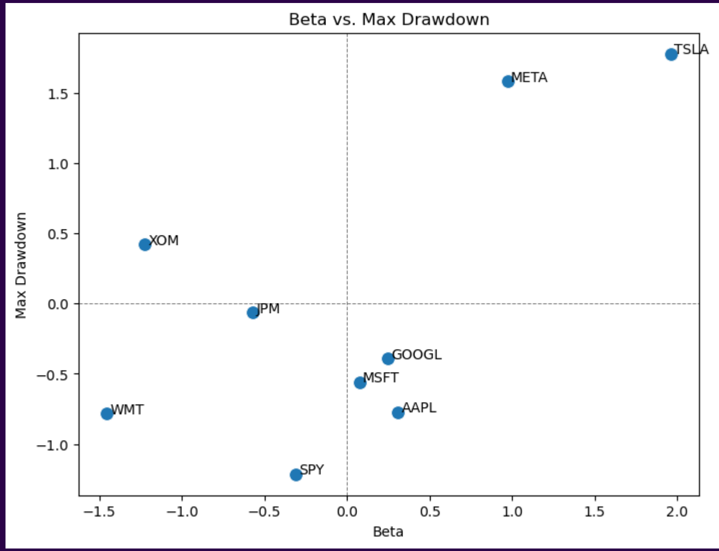
7. EXPLORATORY DATA ANALYSIS

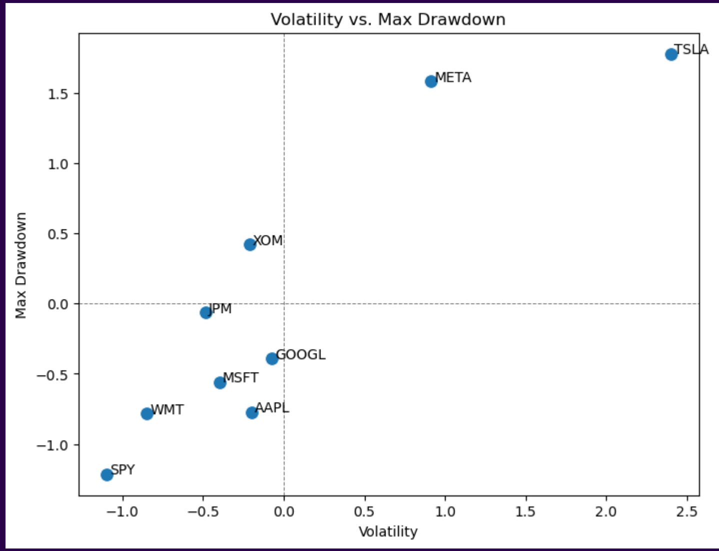




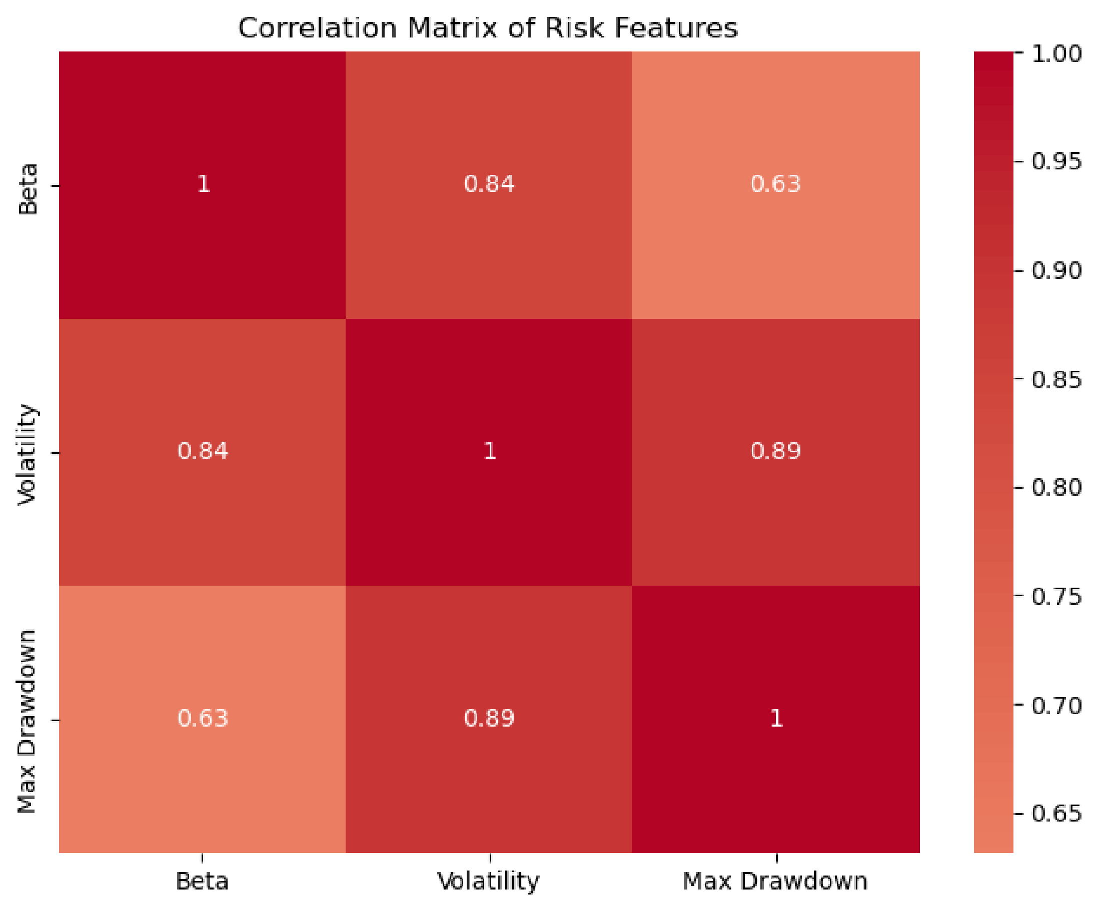
7. EXPLORATORY DATA ANALYSIS







7. EXPLORATORY DATA ANALYSIS



PRINCIPAL COMPONENT ANALYSIS (PCA)

```
Explained Variance Ratio:
[0.86174555 0.12367929 0.01457516]

Principal Components Dataframe:
            PC1       PC2       PC3
AAPL  -0.387376  0.755482  0.110087
MSFT  -0.518628  0.450794 -0.066348
TSLA   3.558718  0.140403  0.288992
GOOGL -0.130835  0.446530  0.043829
JPM   -0.648667 -0.355784 -0.139417
XOM   -0.568823 -1.183178  0.078094
WMT   -1.765195 -0.517076  0.256833
META   1.991097 -0.377443 -0.421015
SPY   -1.530291  0.640272 -0.151055
```

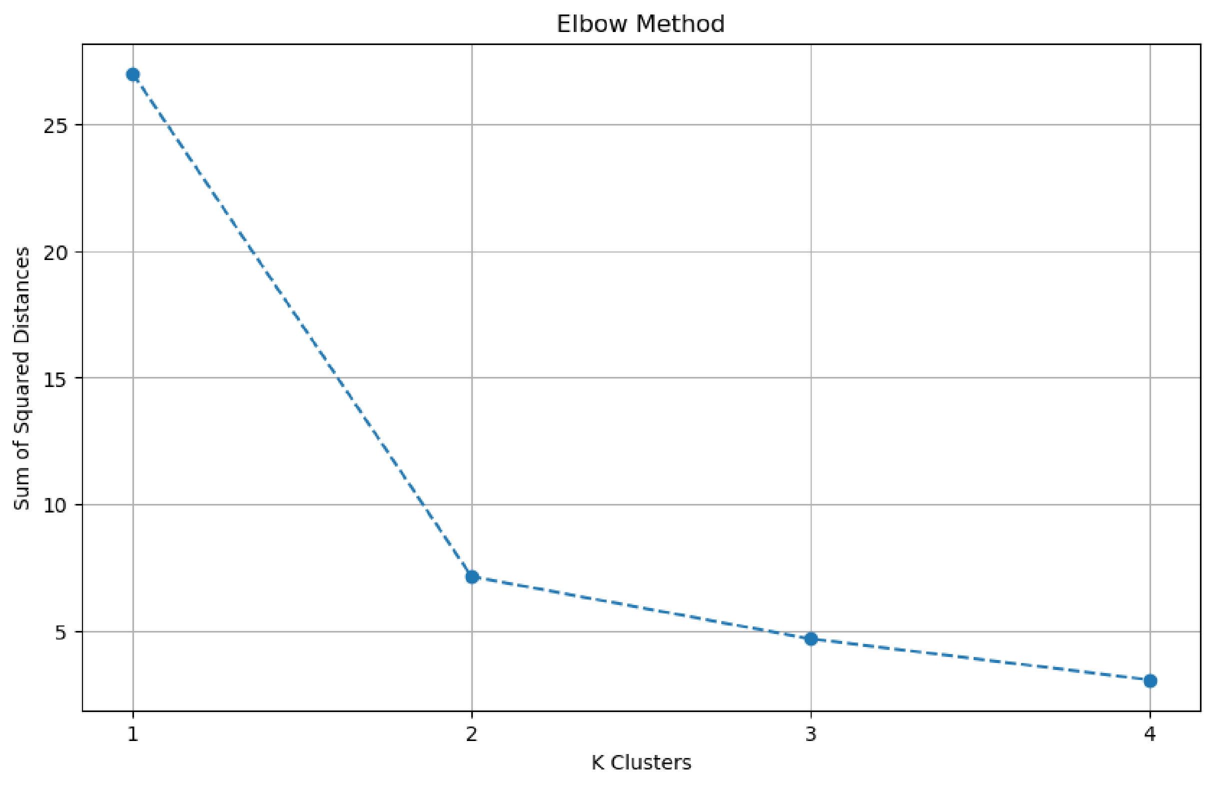
8. MODELLING

8.1 K-MEANS MODEL

- simplicity
- scalability
- interpretable

8.2 THE ELBOW METHOD

- 3 clusters instead of 2
- Low, Medium, and High-risk

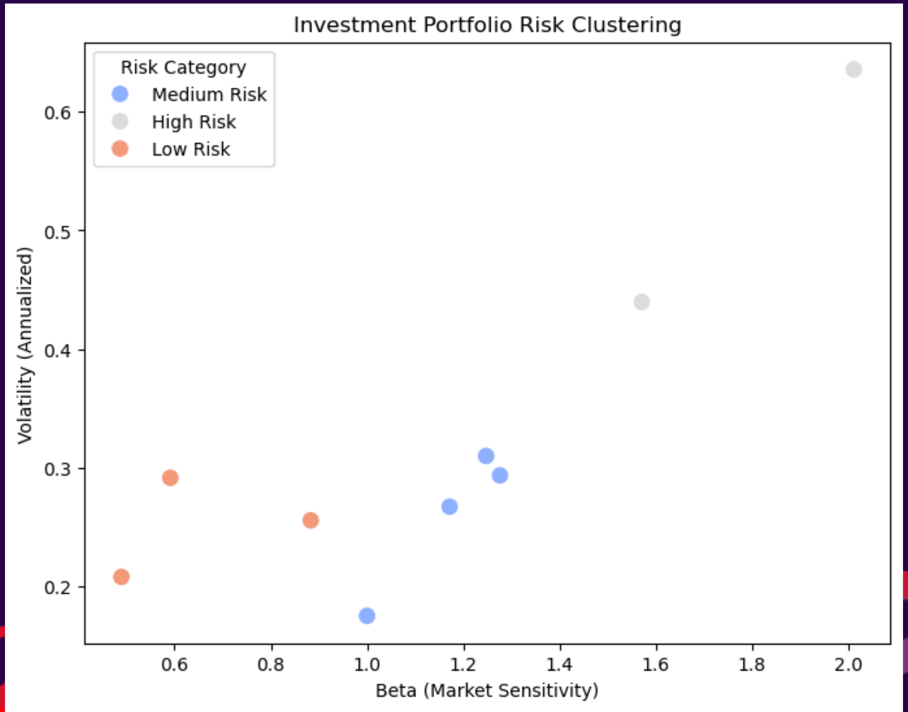


8. MODELLING

8.3 FINAL K-MEANS CLUSTERING

- KMeans(n_clusters=3, random_state=42, n_init=10)
- n_clusters=3 represent Low Risk, Medium Risk, and High Risk.
- random_state=42 to ensure consistent results across multiple runs.
- n_init=10 run K-Means 10 times to reduces the risk of poor clustering.

```
######################################################################
######### K-means Risk Assessment with and without PCA ###############
######################################################################
       Risk Cluster  Risk Cluster PCA Risk Category
AAPL              2                 2   Medium Risk
MSFT              2                 2   Medium Risk
TSLA              1                 1     High Risk
GOOGL             2                 2   Medium Risk
JPM               0                 0      Low Risk
XOM               0                 0      Low Risk
WMT               0                 0      Low Risk
META              1                 1     High Risk
SPY               2                 2   Medium Risk
######################################################################
```



8. MODELLING

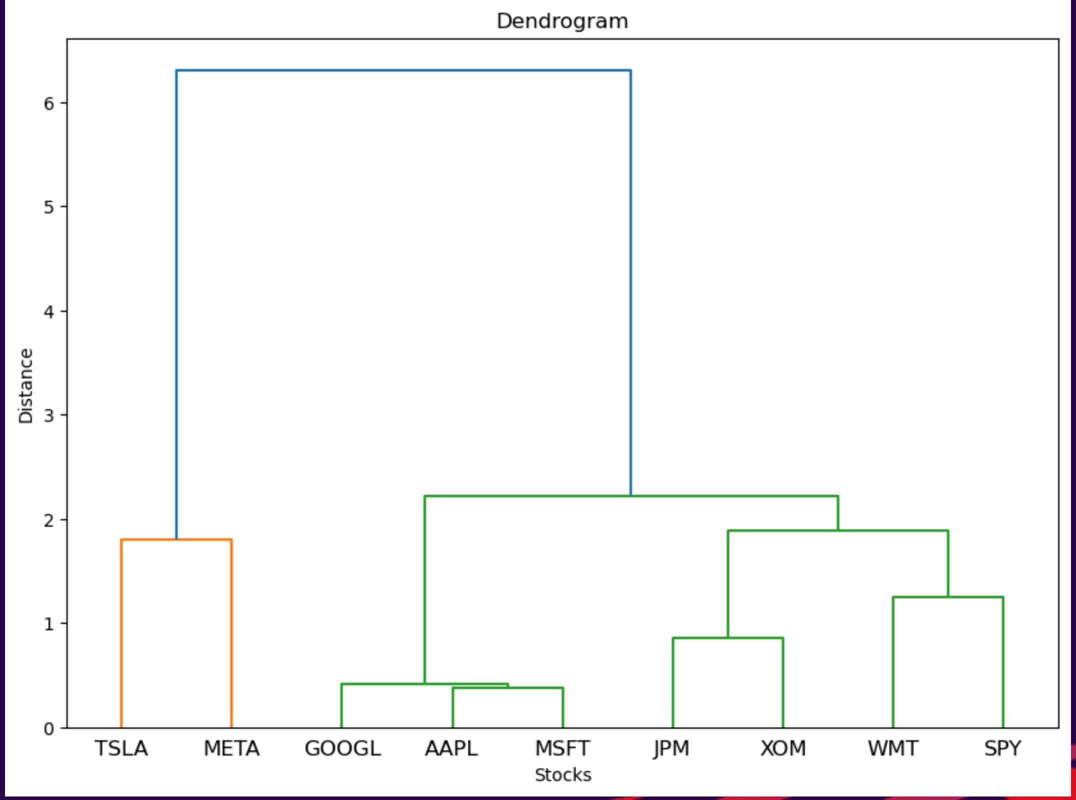
8.4 HIERARCHICAL CLUSTERING MODEL

- n_clusters not specified
- Result: only two clusters

	Beta	Volatility	Max Drawdown	Hierarchical Label	Risk Level
AAPL	0.310311	-0.197959	-0.772950	0	Low Risk
MSFT	0.075435	-0.398340	-0.558773	0	Low Risk
TSLA	1.961124	2.400308	1.777701	1	High Risk
GOOGL	0.245398	-0.073583	-0.390888	0	Low Risk
JPM	-0.572193	-0.485203	-0.062940	0	Low Risk
XOM	-1.227125	-0.213245	0.422210	0	Low Risk
WMT	-1.455297	-0.847456	-0.783054	0	Low Risk
META	0.971996	0.912514	1.583265	1	High Risk
SPY	-0.309650	-1.097037	-1.214571	0	Low Risk

8. MODELLING

8.4 HIERARCHICAL CLUSTERING MODEL



8. MODELLING

8.4 HIERARCHICAL CLUSTERING MODEL

- n_clusters = 3
- Result: same as K-Means

	Beta	Volatility	Max Drawdown	Hierarchical Label	Risk Level
AAPL	0.310311	-0.197959	-0.772950	2	Medium Risk
MSFT	0.075435	-0.398340	-0.558773	2	Medium Risk
TSLA	1.961124	2.400308	1.777701	1	High Risk
GOOGL	0.245398	-0.073583	-0.390888	2	Medium Risk
JPM	-0.572193	-0.485203	-0.062940	0	Low Risk
XOM	-1.227125	-0.213245	0.422210	0	Low Risk
WMT	-1.455297	-0.847456	-0.783054	0	Low Risk
META	0.971996	0.912514	1.583265	1	High Risk
SPY	-0.309650	-1.097037	-1.214571	0	Low Risk

9. RESULT AND ANALYSIS

EVALUATION METRICS

Morningstar Risk Score	K-Means Risk Category
High, Very High, and Extreme	High Risk
Medium	Medium Risk
Low	Low Risk

EVALUATION ANALYSIS

Ticker	Morningstar	K-Means (5-Year Dataset)
AAPL	Medium	Medium
GOOGL	Medium	Medium
JPM	NA	Medium
META	High	High
MSFT	Medium	Medium
TSLA	Very High	High
WMT	Medium	Low
XOM	High	Medium

9. RESULT AND ANALYSIS

NORMALIZED VS PCA	same K-means results
K-MEANS VS HIERARCHICAL	produce the same groupings

10. DISCUSSION AND CONCLUSION

10.1 CHALLENEGS

- Limited Download from Data Source
- Fine-Tune K-Means
- Optimal Number of Clusters
- Alternative Approach

10.2 KEY TAKEAWAY

Percentage Changes

To prevent raw prices distorting features calculations

Standardization

Prevent larger numerical values from dominating

Qualitative & Quantitative Data

Risk-Adjusted Performance Score

Defination of Risk

Market risk is about more than volatility

10.3 CONCLUSION

- Successfully applied K-Means clustering
- Risk-adjusted performance
    - Sharpe or Sortino ratios
- Incorperate qualitative metrics

REFERENCES

- Datrics AI. (n.d.). K-Means Clustering in Banking: Applications & Examples. Retrieved from https://www.datrics.ai/articles/how-k-means-clustering-is-transforming-the-banking-sector
- Stock classification using k-means clustering (n.d.). Medium. Retrieved from https://medium.com/@facujallia/stock-classification-using-k-means-clustering-8441f75363de
- Li, B., Tao, R., Li, M., & Sharma, K. (2022). Identification of Enterprise Financial Risk Based on Clustering Algorithm. *Computational Intelligence and Neuroscience*, *2022*, Article ID 1086944. https://doi.org/10.1155/2022/1086945
- Zhu, Y., & Liu, Q. (2021). Early Warning of Financial Risk Based on K-Means Clustering Algorithm. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/349851488_Early_Warning_of_Financial_Risk_Based_on_K-Means_Clustering_Algorithm
- Morningstar, Inc. (n.d.). Morningstar's risk ratings explained. Retrieved from https://global.morningstar.com/en-gb/personal-finance/morningstar-s-risk-ratings-explained
- Portfolios Lab (n.d.). Example of AAPL Risk-Adjusted Performance. Retrieved from https://portfolioslab.com/symbol/AAPL
- Tipranks (n.d.). Example of AAPL's Risk Factors. Retrieved from https://www.tipranks.com/stocks/aapl/risk-factors
- Infront. (n.d.). Infront Advanced Analytics Solution for Wealth Managers, Analysts & Brokers. Retrieved from https://www.infront.co/global/en/solutions/modules/data-display-and-analytics/analytics.html

GITHUB REPOSITORY LINK

- https://github.com/peculiardatabits/DTSA-5510-Final-Project