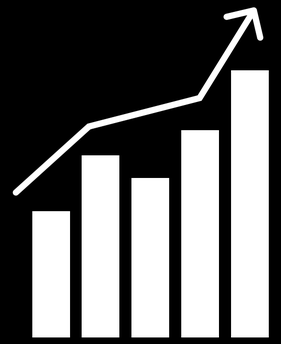
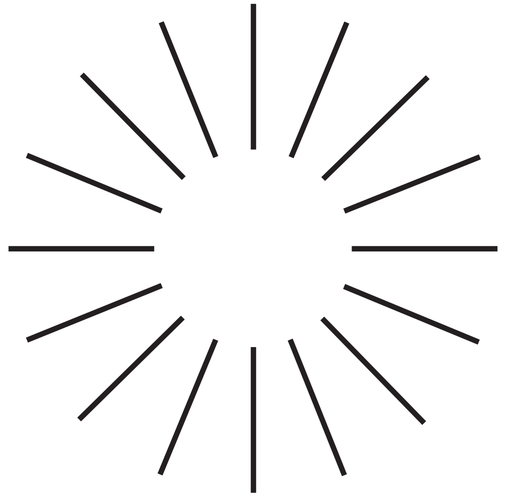
PROJECT 4



NAVINA SENTHIL LUIS ULLOA

NOVEMBER 1, 2024
CLUSTERING MODELS
USING KMEANS & DBSCAN
HONORABLE MENTION: ADITYA MAHAJAN

AGENDA



— Problem Statement
— Project Process
— Datasets
— User Personas
— EDA
— Modeling
— Conclusion

“Can we cluster stocks based on multiple financial metrics—including volatility, market capitalization, historical performance, and liquidity—to generate **tailored investment strategies** for investors with different risk profiles?”

PROJECT PROCESS

DATA COLLECTION

Gather relevant user and stock data from APIs and datasets for analysis.

DATA EXPLORATION

Examine patterns, distributions, and relationships within the data to inform modeling decisions.

MODEL FITTING

Implement clustering algorithms to classify stocks based on features relevant to investment strategies.

FINDINGS

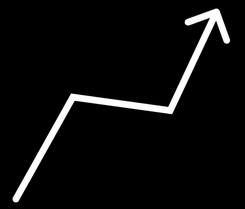
Assess model outcomes to draw insights and create personalized investment strategies.

DATASETS

USERS

Captures key demographic and financial details — like age, account balance, and credit usage — to help define distinct investor profiles.

Source: GitHub



STOCKS

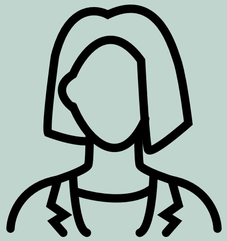
Tracks financial metrics like price volatility, market cap, and trading volume for clustering stocks based on investment risk levels.

Source: Yahoo Finance

USER PERSONAS

LOW-RISK INVESTOR

Goal: Focused on preserving capital with stable, low-risk investments that yield consistent returns.



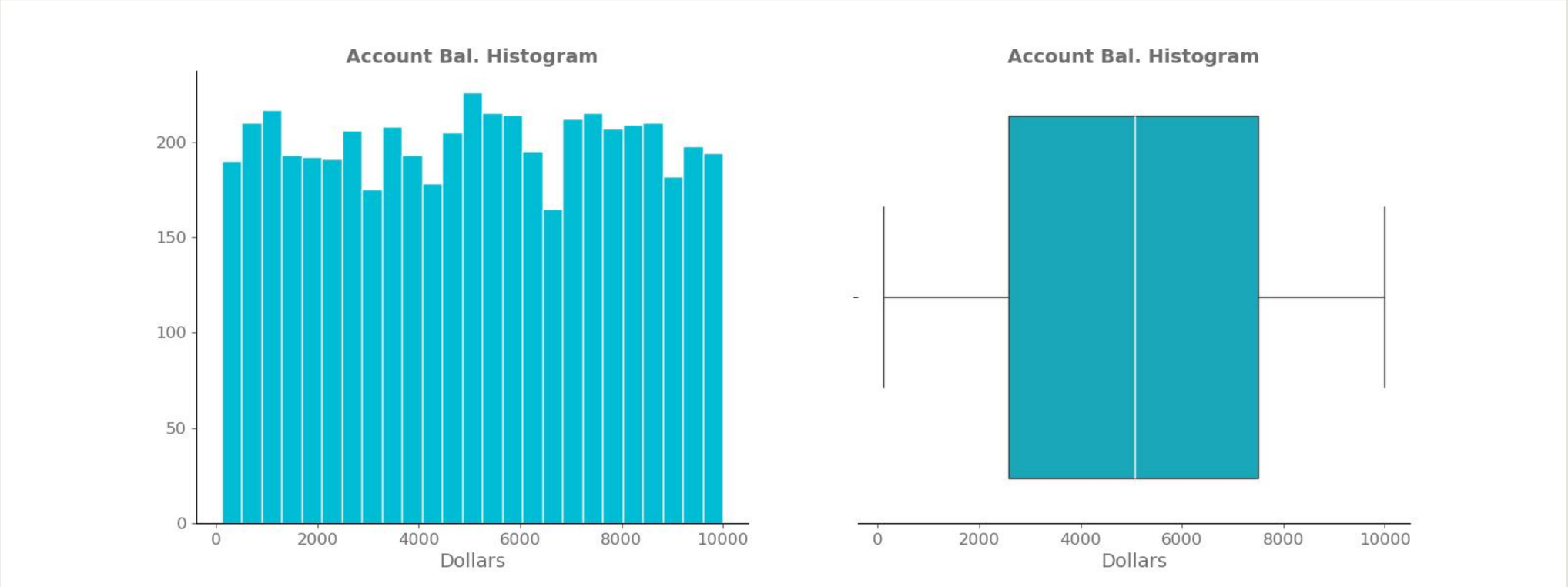
MEDIUM-RISK INVESTOR

Goal: Balances risk and reward, seeking moderate growth while being open to fluctuations for better returns.

HIGH-RISK INVESTOR

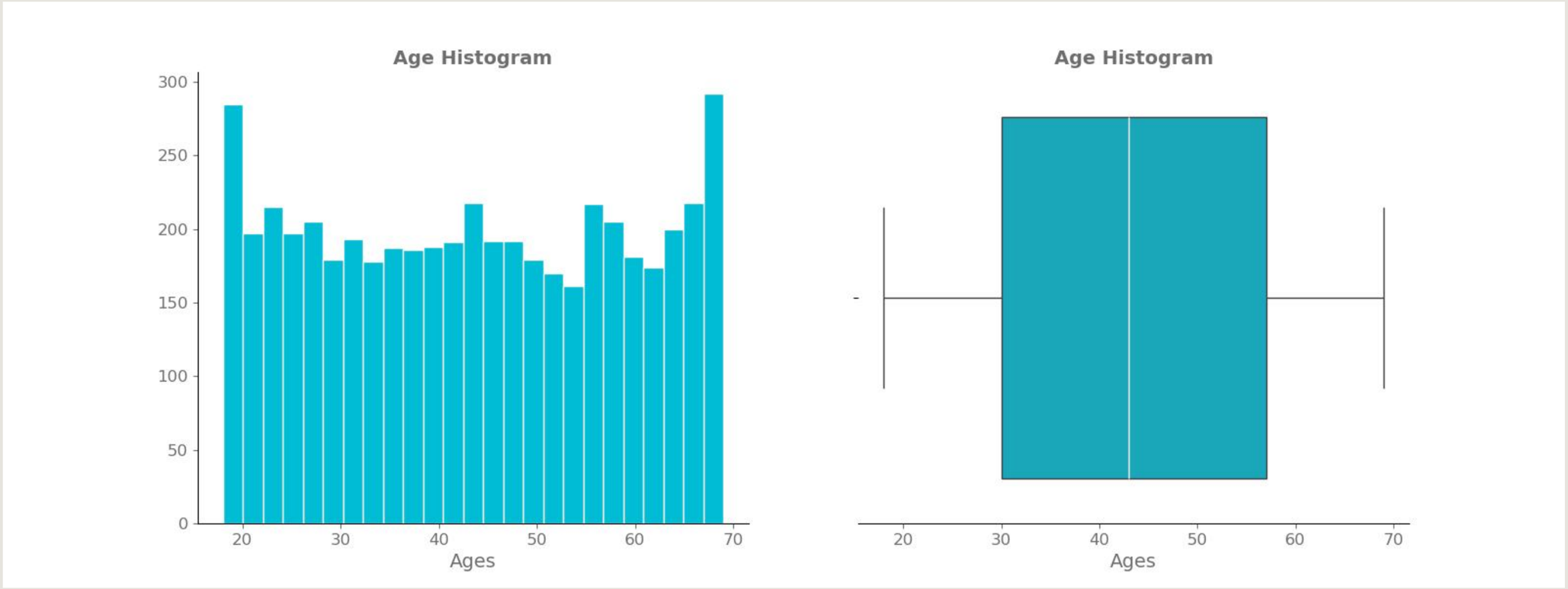
Goal: Growth-driven and comfortable with high market swings, aiming for significant returns despite potential risks.

DISTRIBUTION OF ACCOUNT BALANCES



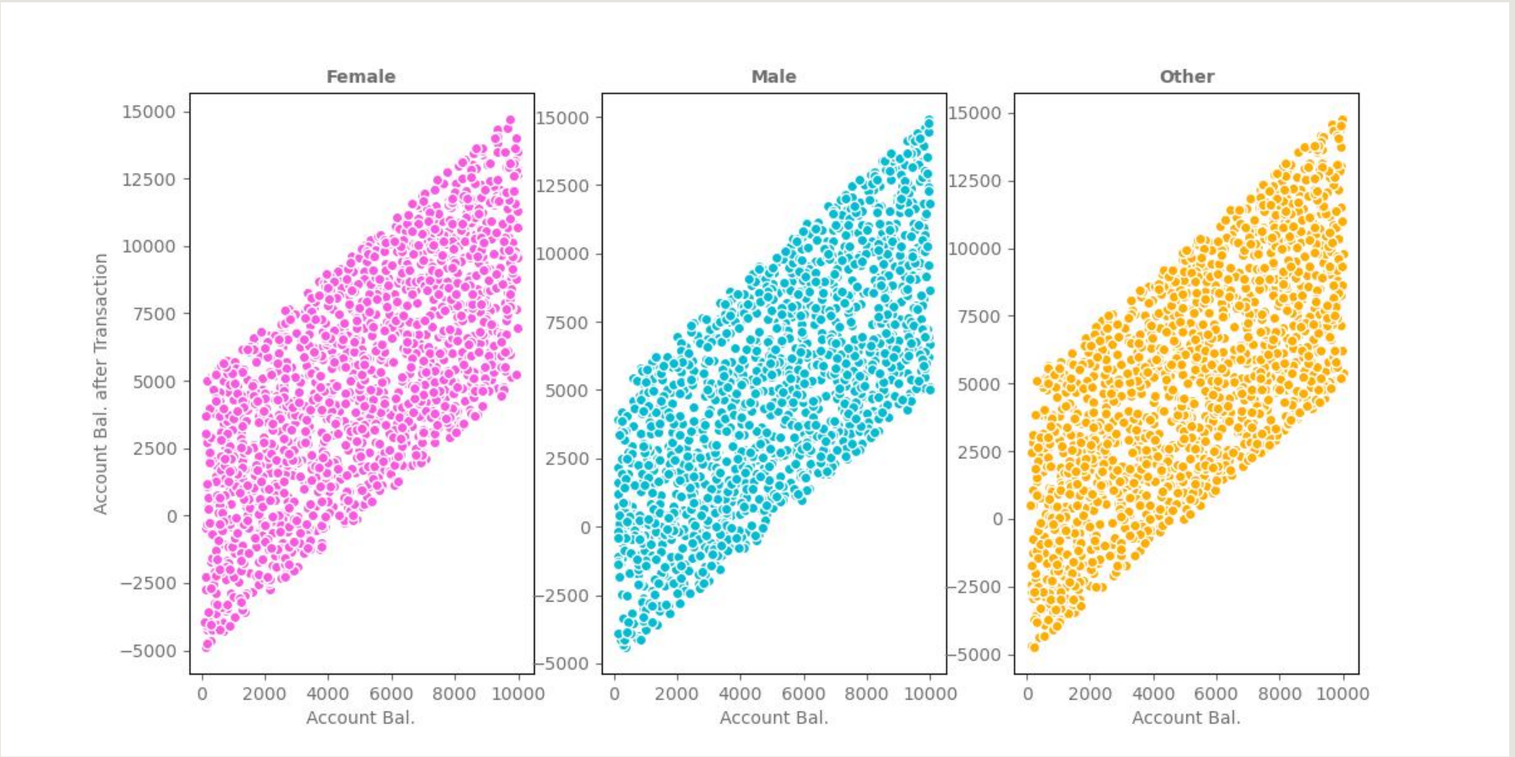
UNIFORM DATA DISTRIBUTION

DISTRIBUTION OF AGE



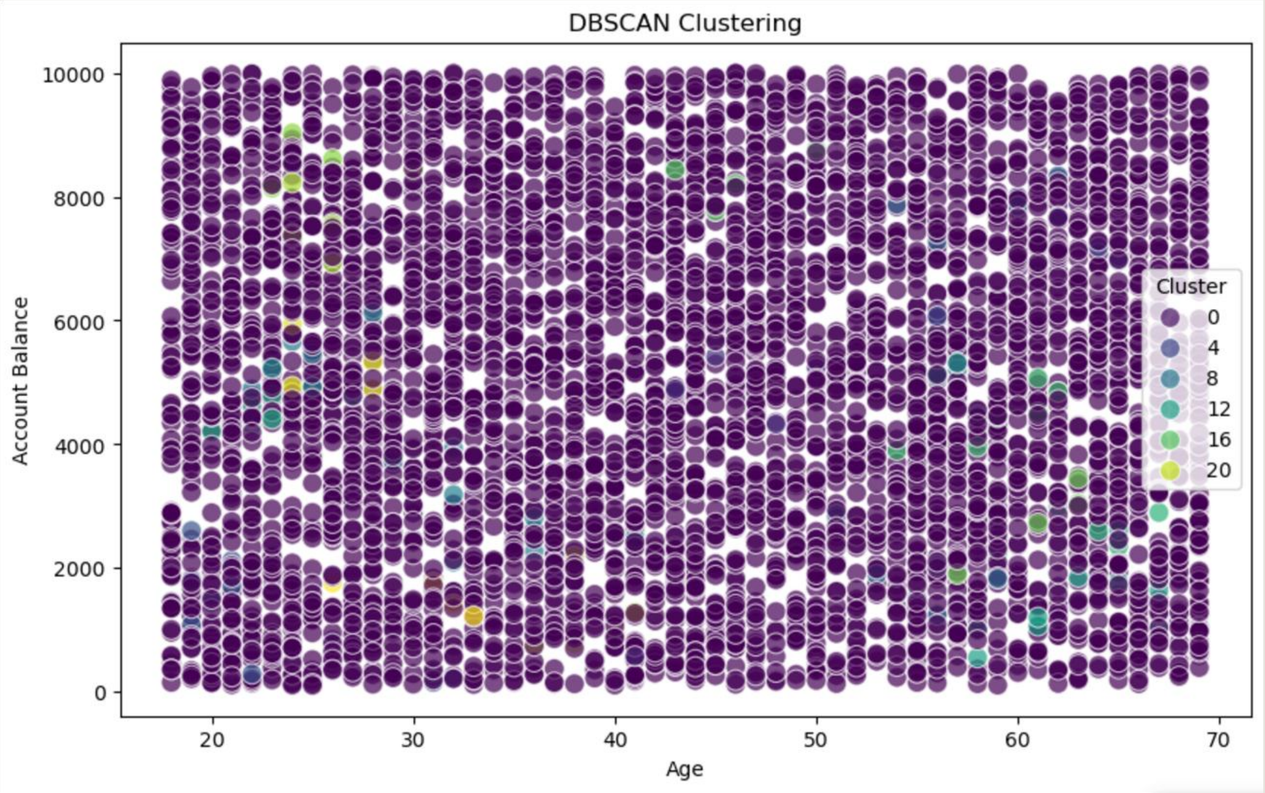
UNIFORM DATA DISTRIBUTION

GENDER VS ACCOUNT BAL.



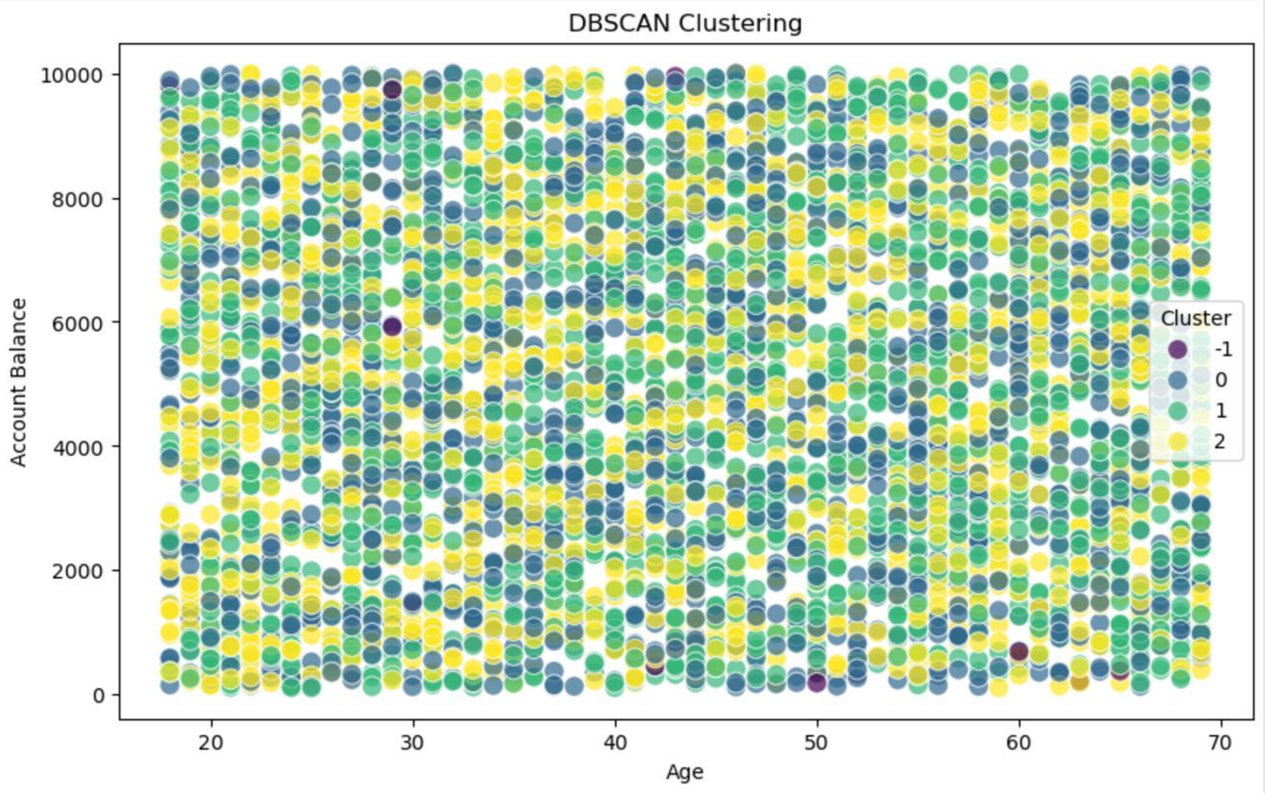
IDENTICAL CORRELATIONS BETWEEN GENDER AND ACCOUNT BALANCE

USER DATA USING DBSCAN



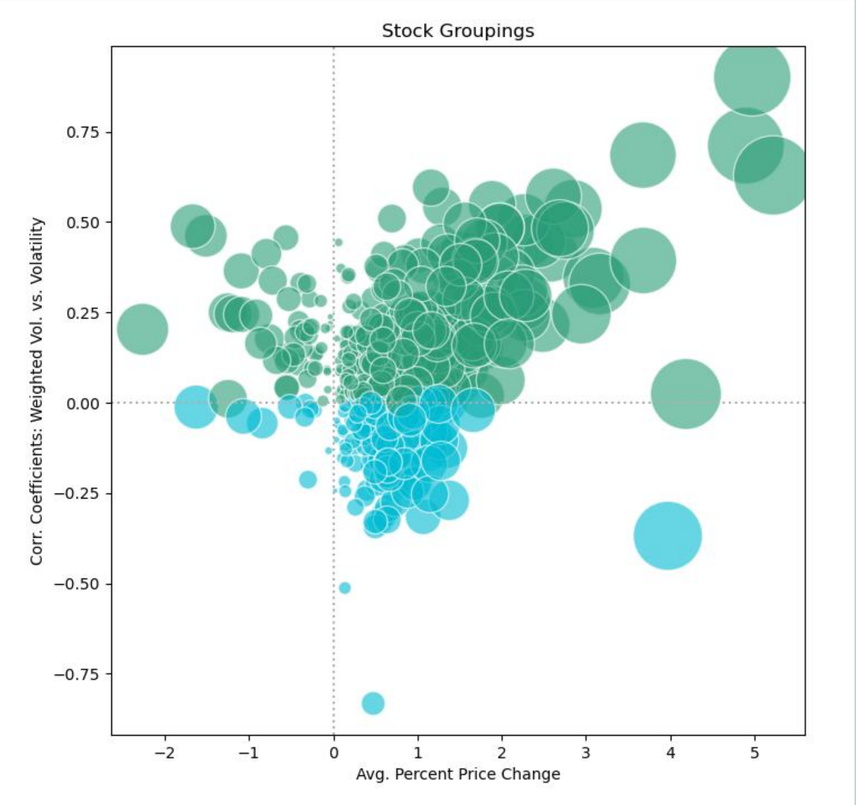
NO VIABLE CLUSTERS TO WORK WITH

USER DATA USING DBSCAN



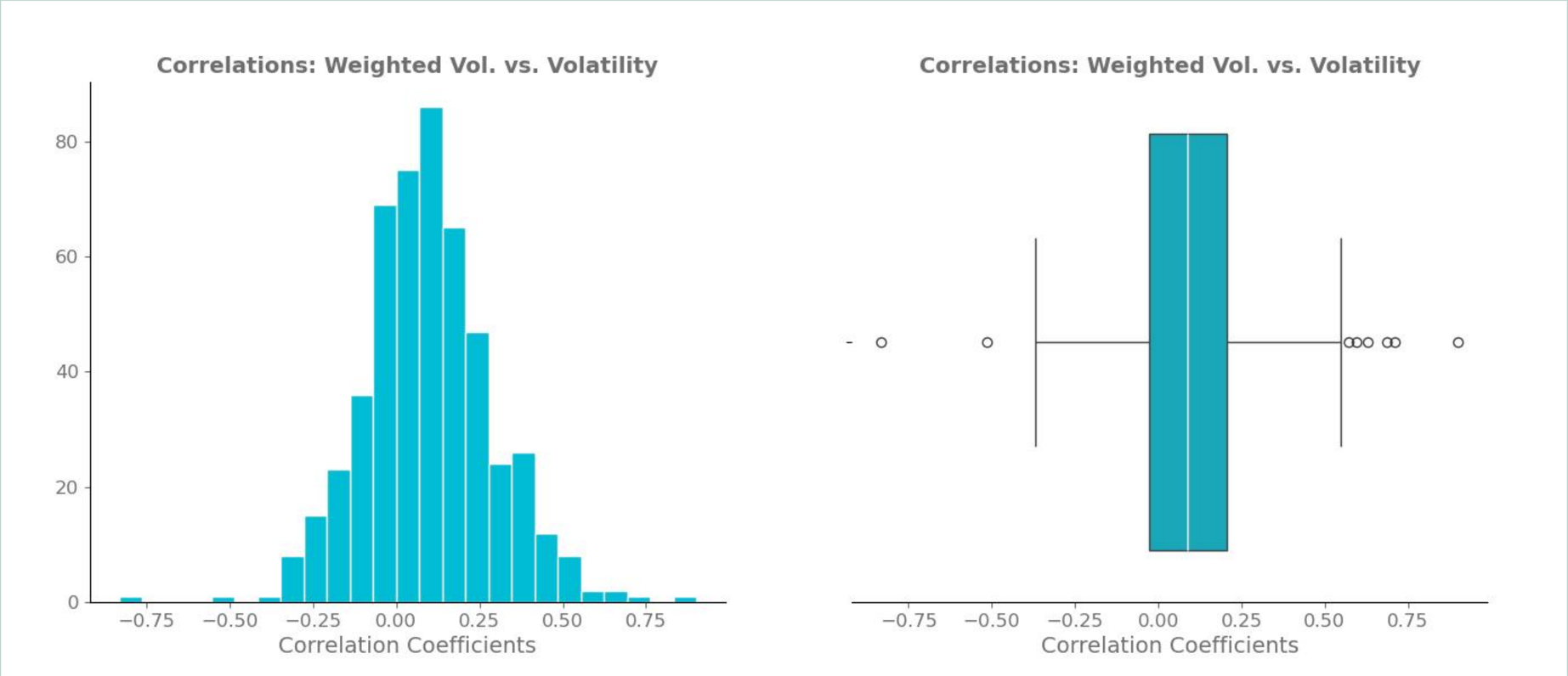
ONCE AGAIN, NO VIABLE CLUSTERS TO WORK WITH

STOCK GROUPINGS BY VOLATILITY



BLUE BUBBLES: LOW VOLATILITY
GREEN BUBBLES: HIGH VOLATILITY
BUBBLE SIZE: PERCENTAGE CHANGE IN PRICE OVER TIME

CORRELATION: WEIGHTED VOLUME VS VOLATILITY



NORMAL DISTRIBUTION, VERY FEW OUTLIERS

SUMMARY TABLE

Characteristics	Clusters	Inertia	Silhouette	Remarks
KMeans, all numeric features	11	61,844	0.266	DF col: Cluster A
KMeans, three engineered features	11	12,193	0.377	DF col: Cluster B
KMeans, best two engineered features	11	3,287	0.533	DF col: Cluster C
DBSCAN, all numeric features	5	NA	0.800	DF col: Cluster D
DBSCAN, three engineered features	6	NA	0.842	DF col: Cluster E
DBSCAN, three engineered features	7	NA	0.824	DF col: Cluster F

BEST PERFORMING MODEL BASED ON SILHOUETTE SCORE
IS DBSCAN WITH 6 CLUSTERS

FINDINGS

1

Clear cluster distinction for high, medium and low risk stock data

2

Unable to find clear insights from user data

3

How might we be able to improve on this project?

MOVING FORWARD

- Different dataset for user data in order to define clear clusters
- Expand stocks and pull more data in order to allow users to explore beyond S&P’s 500
- Create Streamlit app in order to allow users to input preferences and get reccomendations

THANK YOU!