

PROJECT 4

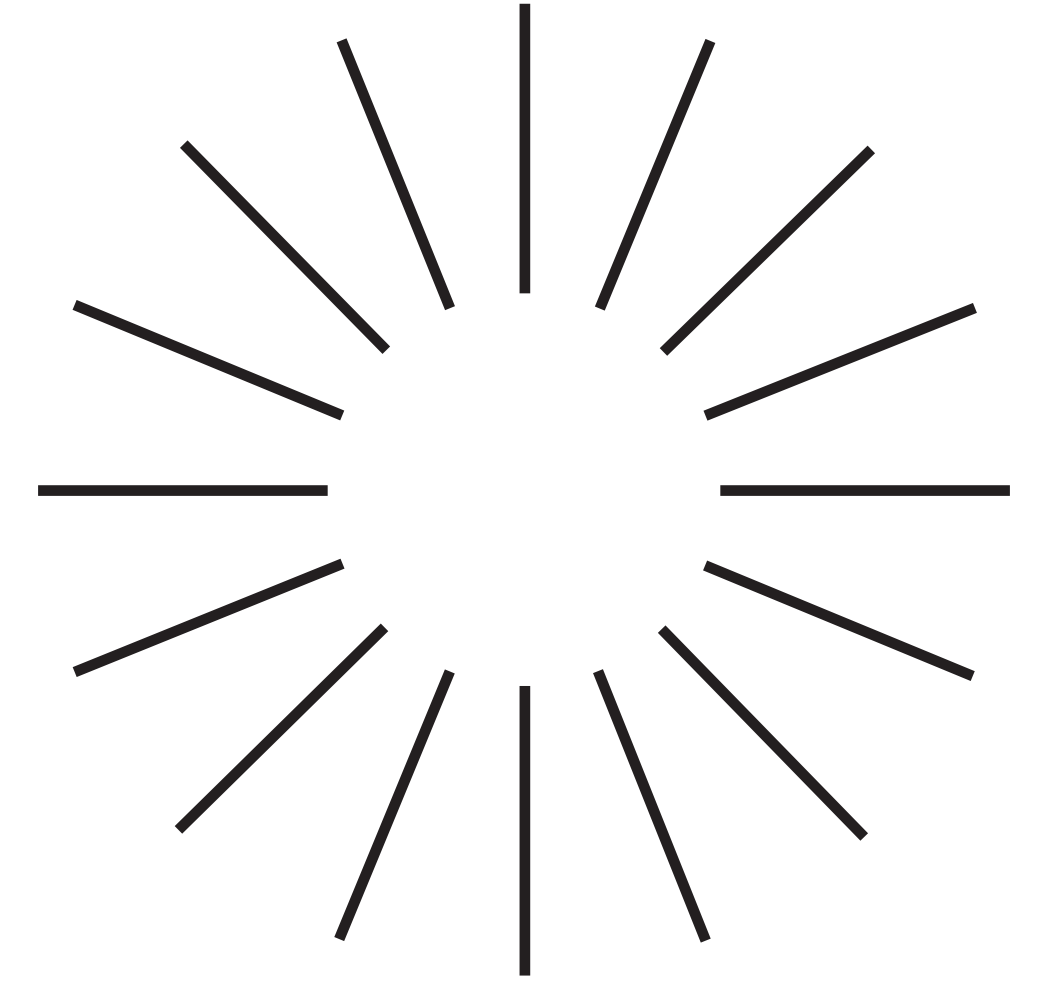
NAVINA SENTHIL

LUIS ULLOA

GENERAL ASSEMBLY
CLUSTERING MODELS
USING **KMEANS** & **DBSCAN**
NOVEMBER 1, 2024

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.

MODEL FITTING

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

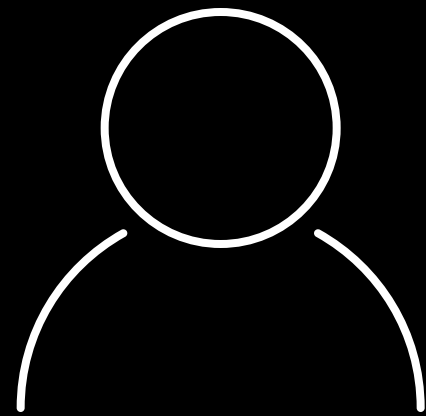
DATA EXPLORATION

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

FINDINGS

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

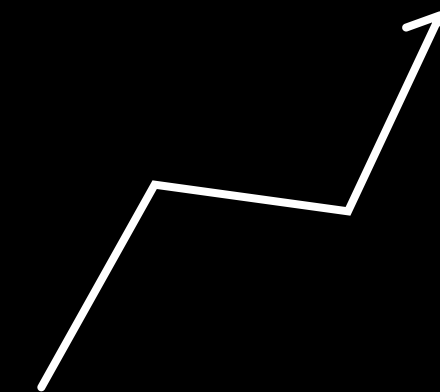
DATASETS



USERS

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

Source: Kaggle



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



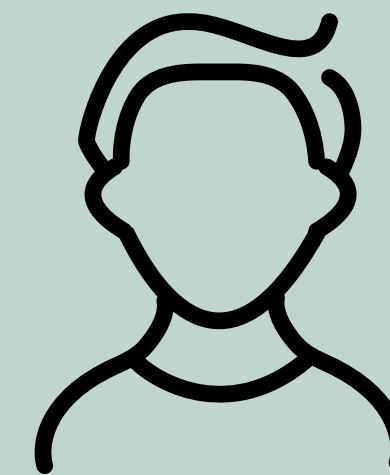
LOW-RISK INVESTOR

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



MEDIUM-RISK INVESTOR

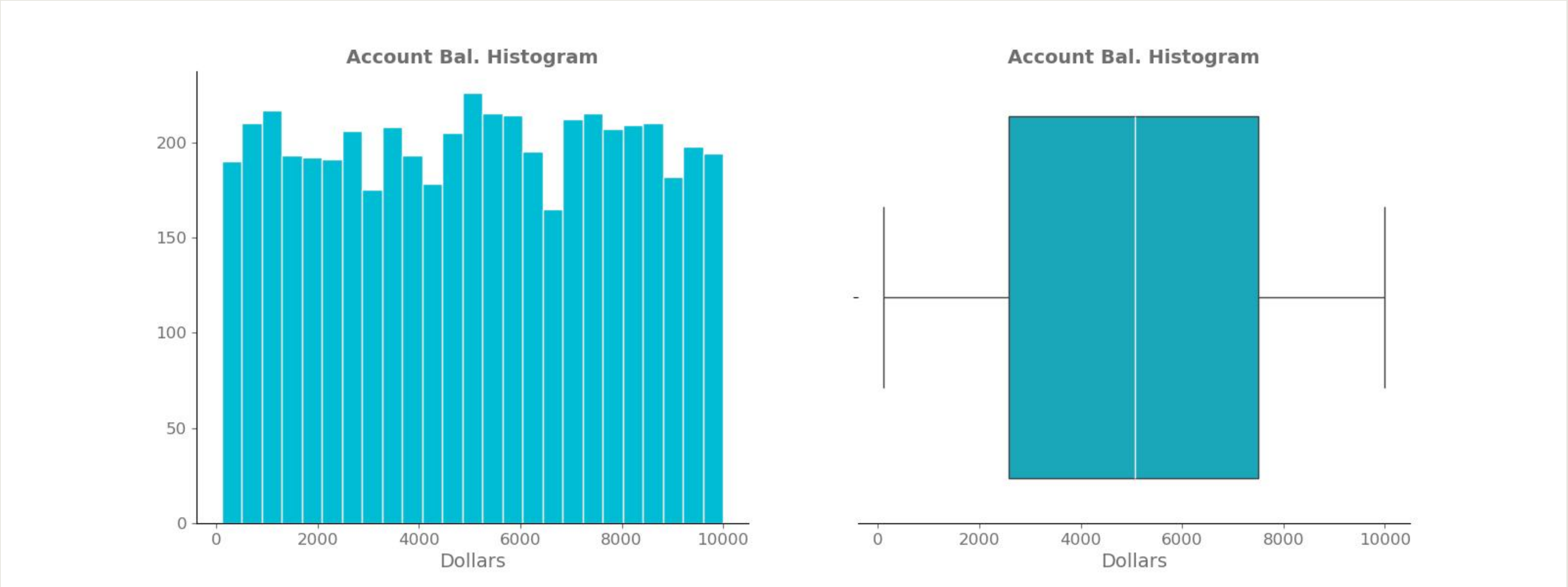
Goal: Balances risk and reward, seeking moderate growth while being open to fluctuations for better returns.



HIGH-RISK INVESTOR

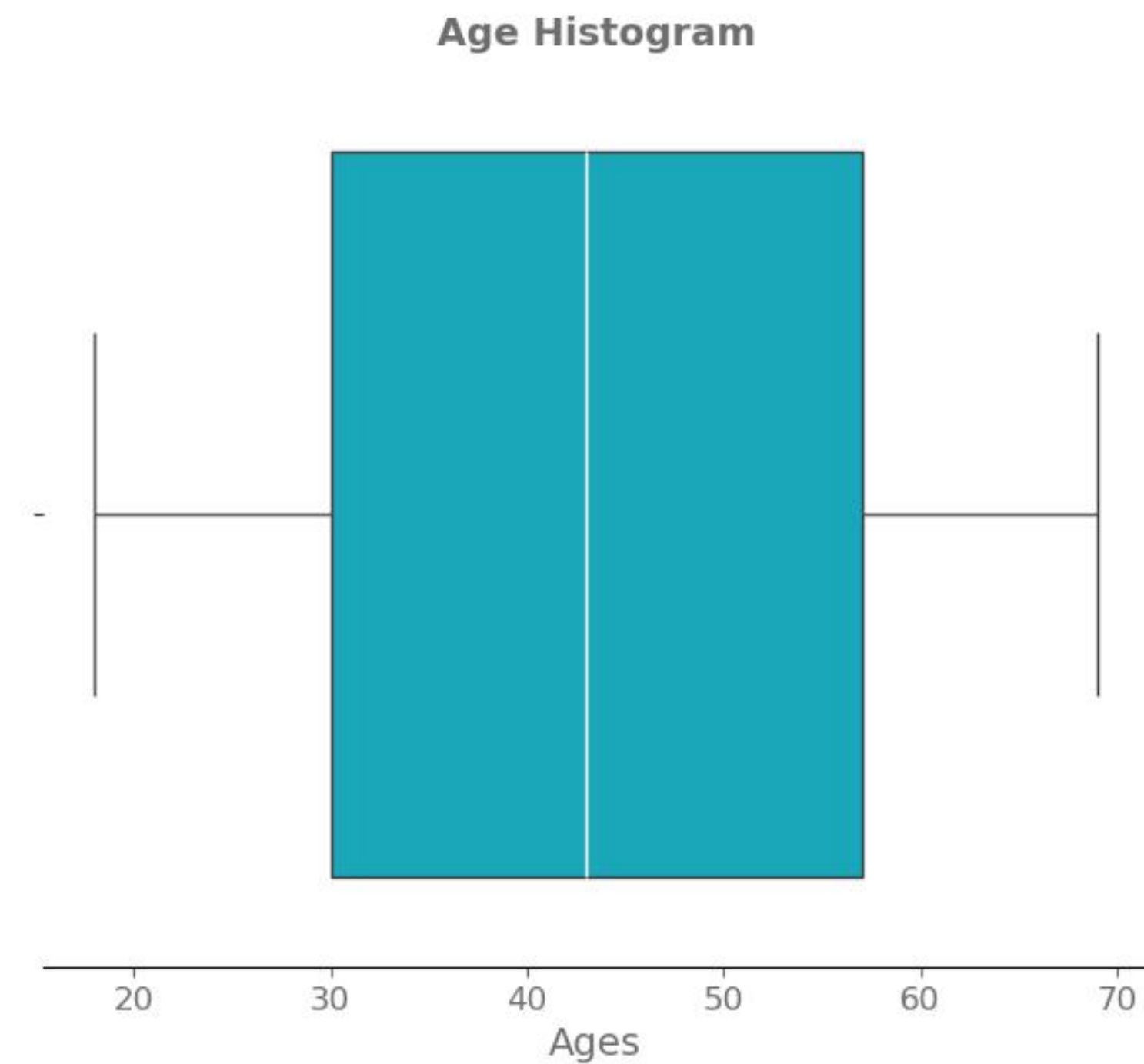
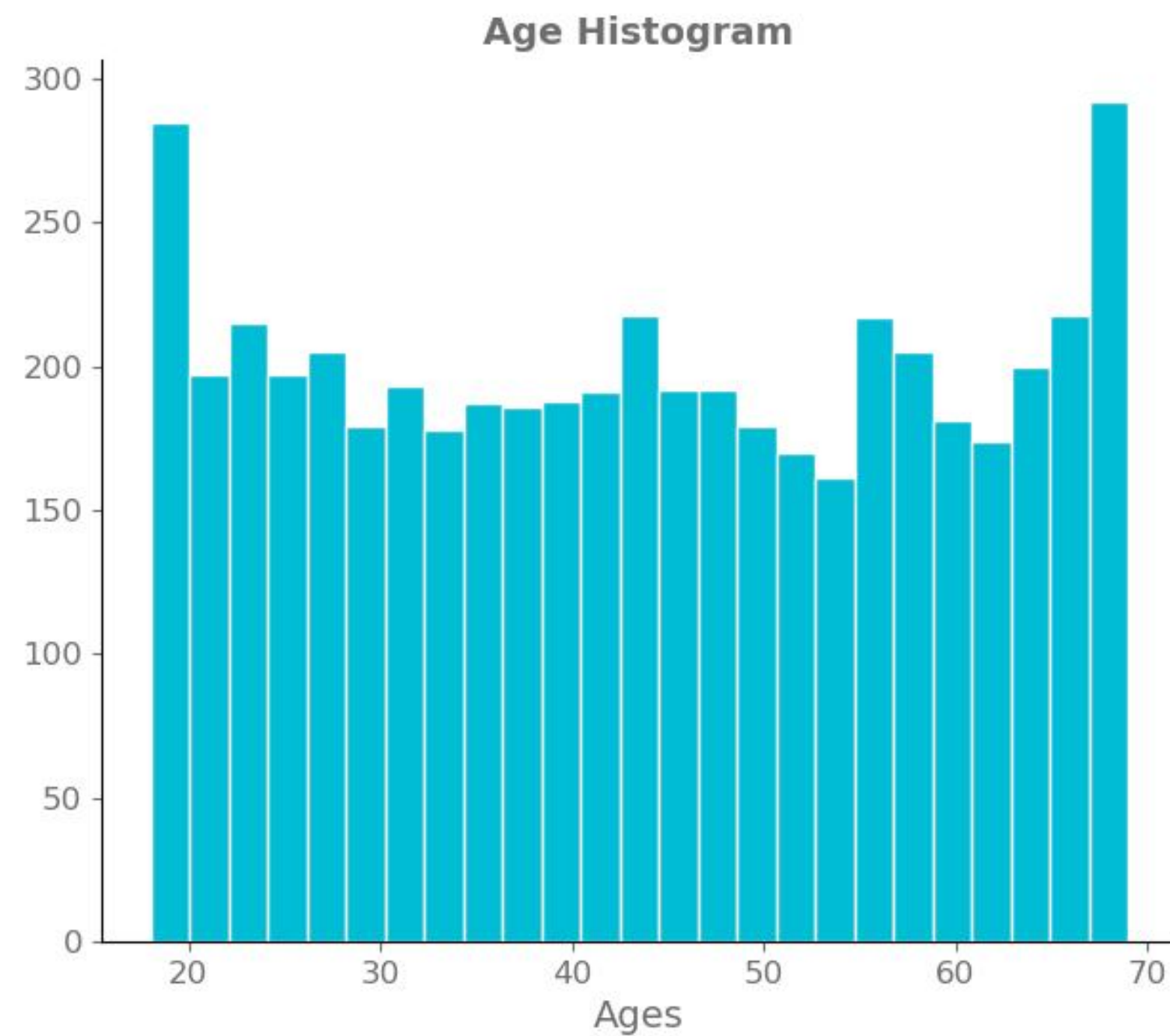
Goal: Growth-driven and comfortable with high market swings, aiming for significant returns despite potential risks.

DISTRIBUTION OF ACCOUNT BALANCES



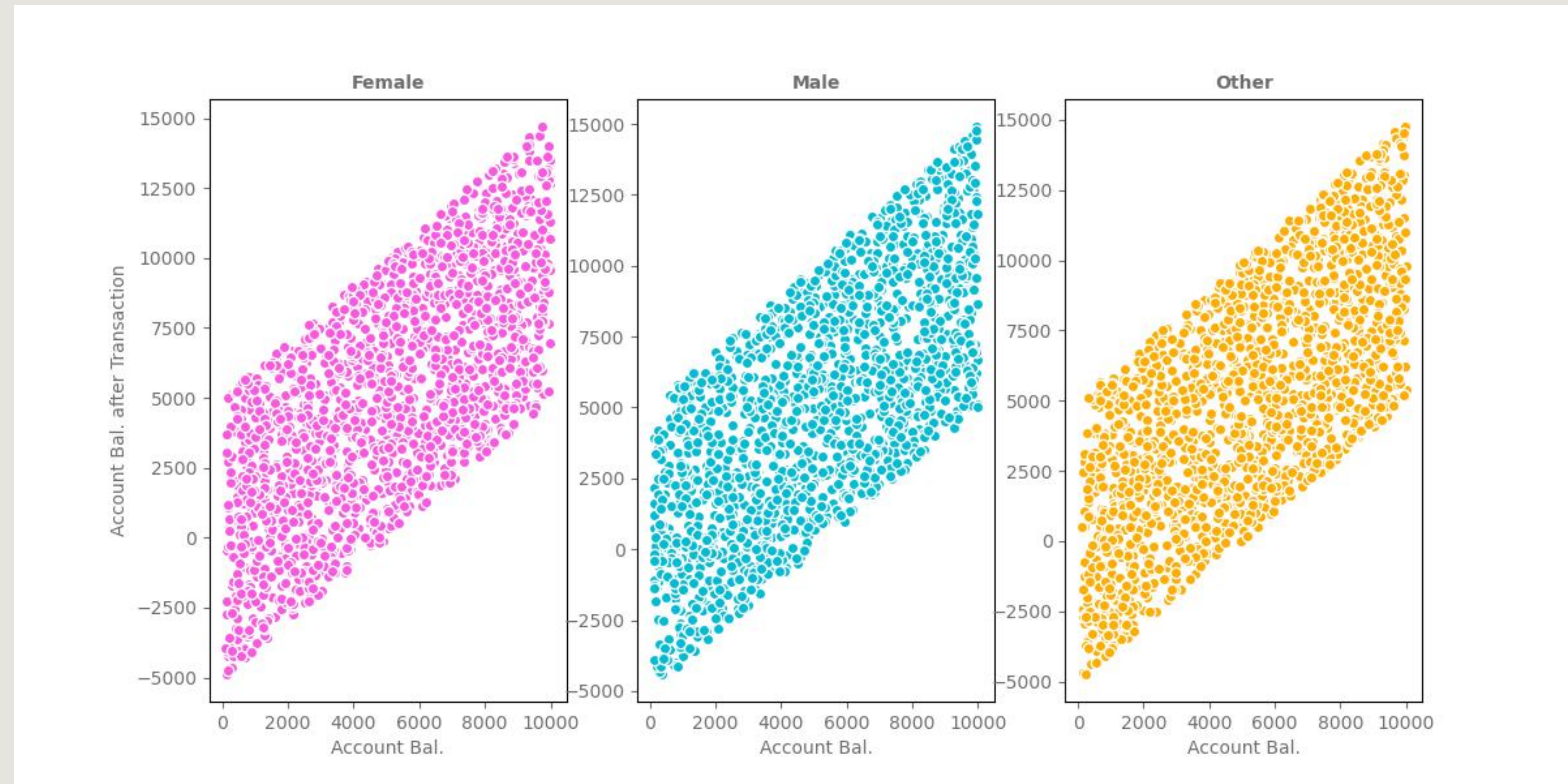
UNIFORM DATA DISTRIBUTION

DISTRIBUTION OF AGE



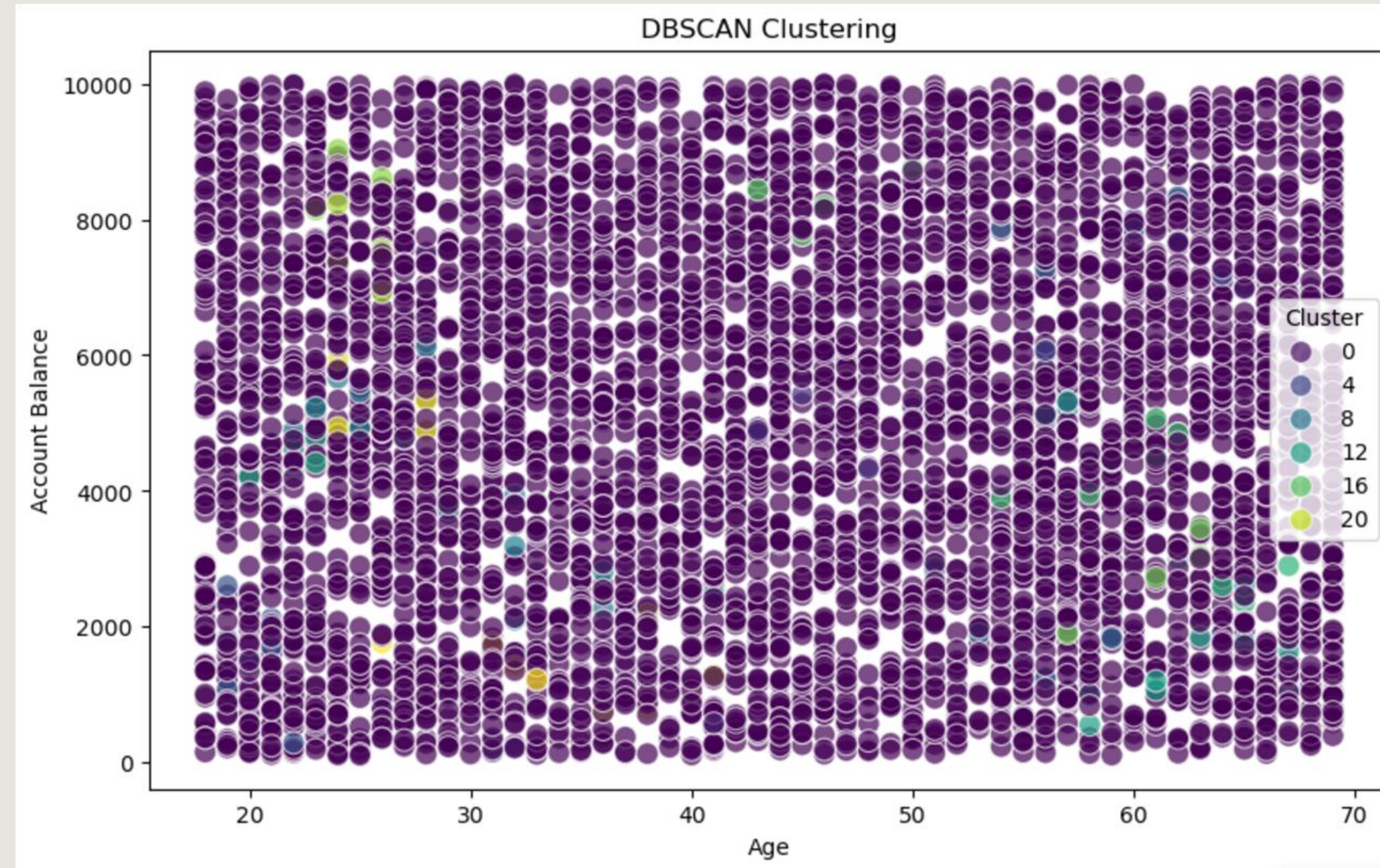
UNIFORM DATA DISTRIBUTION

GENDER VS ACCOUNT BAL.



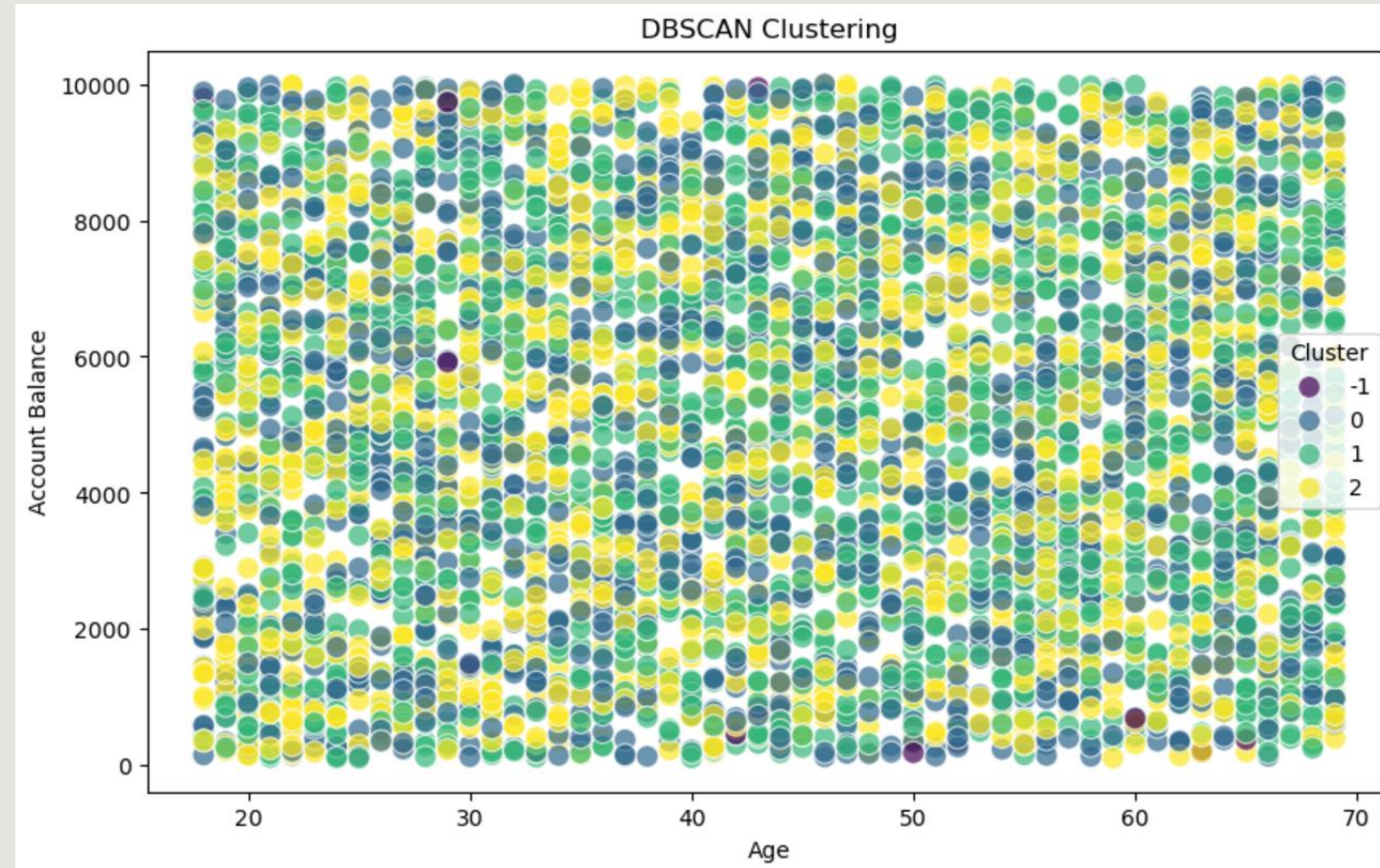
VERY LITTLE CORRELATION BETWEEN GENDER AND
ACCOUNT BALANCE

USER DATA USING DBSCAN



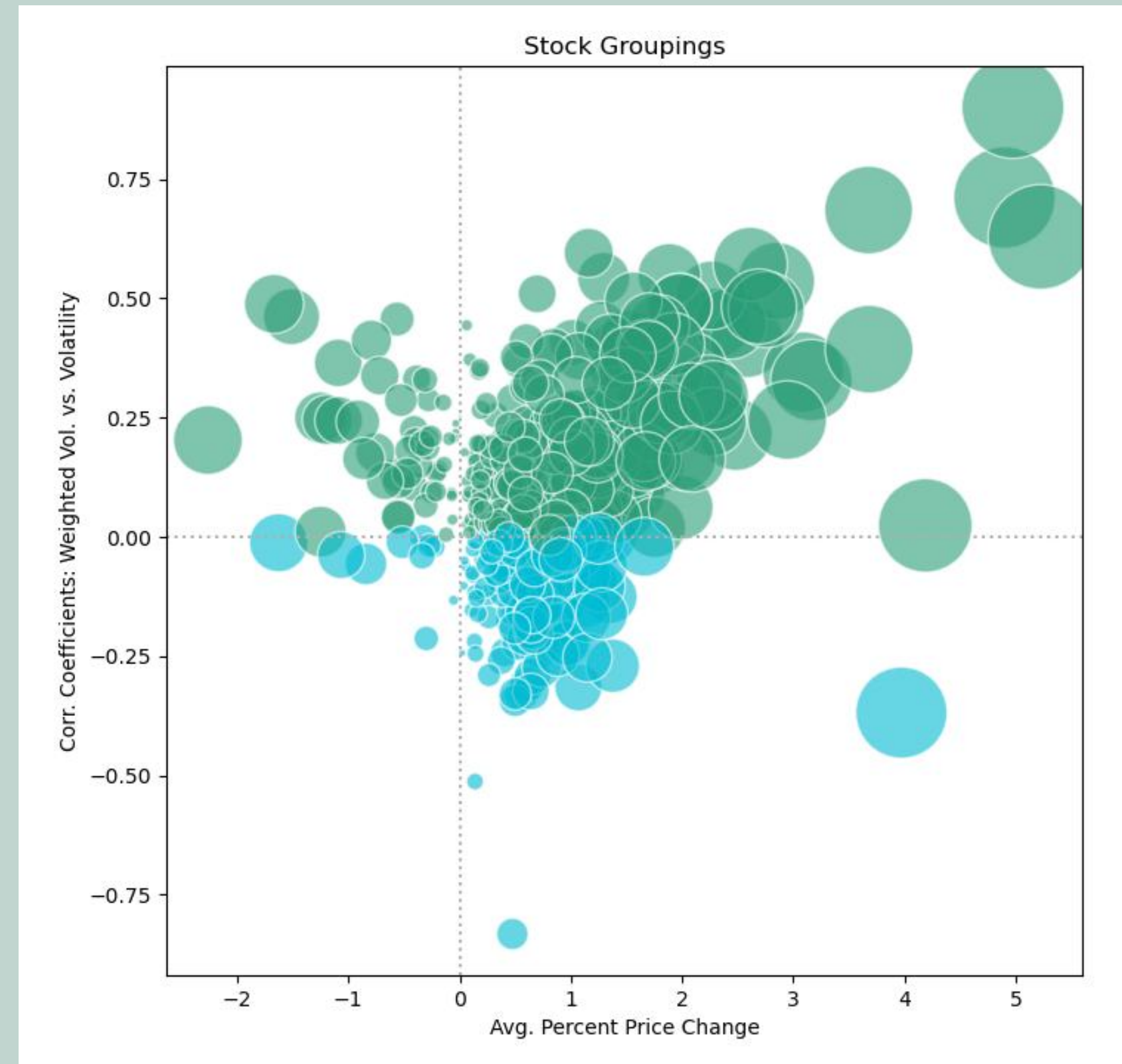
NO VIABLE CLUSTERS TO WORK WITH

USER DATA USING DBSCAN



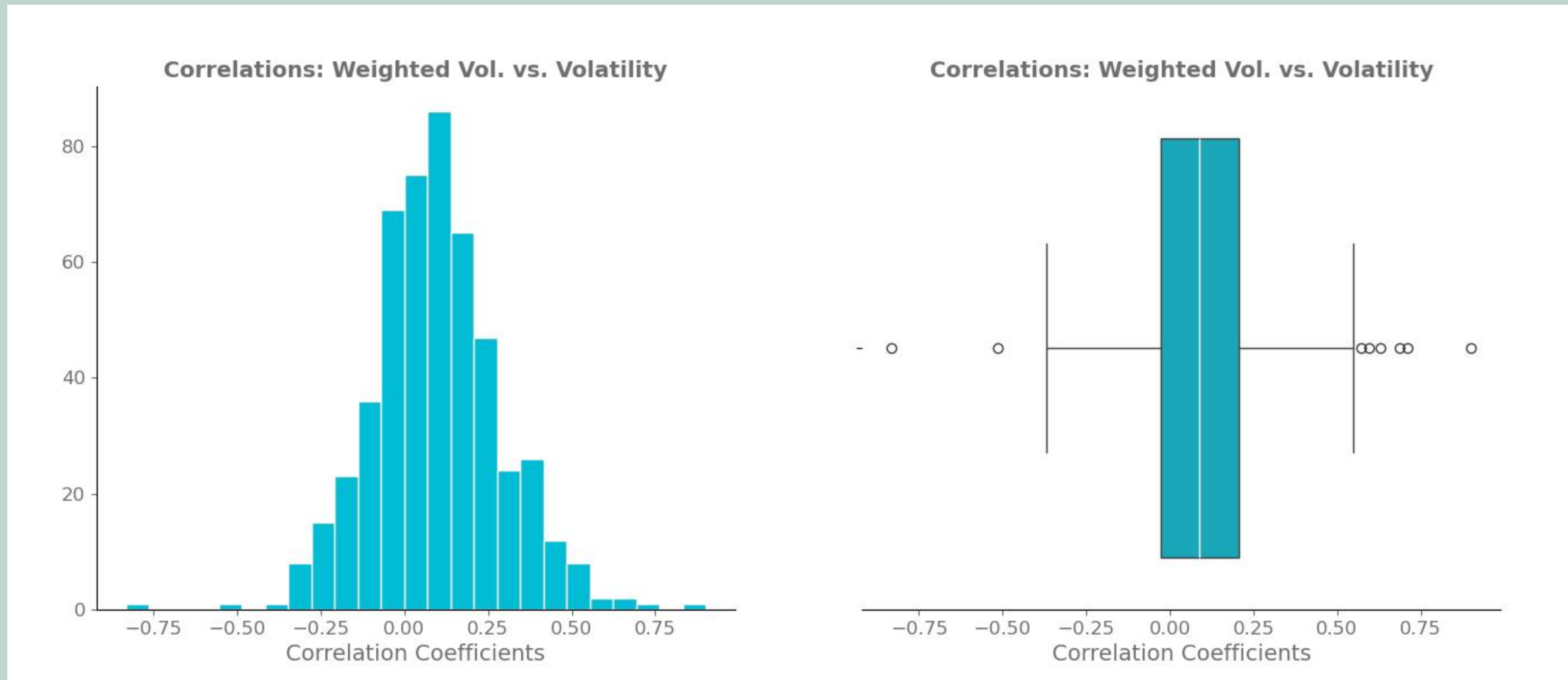
ONCE AGAIN, NO VIABLE CLUSTERS TO WORK WITH

CORRELATION: WEIGHTED VOLUME VS VOLATILITY



BLUE BUBBLES: LOW VOLATILITY
GREEN BUBBLES: HIGH VOLATILITY

CORRELATION: WEIGHTED VOLUME VS VOLATILITY



NORMAL DISTRIBUTION, VERY FEW OUTLIERS

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 recommendations