Executive Summary

Project Objective

This project aimed to apply a complete data mining pipeline—from **ETL to insights and storytelling**—on a company performance dataset. The goal was to uncover patterns, anomalies, and high-performing clusters to support **data-driven strategic decisions** such as investment prioritization, risk detection, and company evaluation.

TEAM MEMBERS & Contributions

- Mitchel_413 - Responsible for the Data Mining.
- Queen_897 - Responsible for Data Cleaning and Enrichment.
- Claire_470 - Responsible for Insight dashboard and Presentation deck.
- Kyra_619 - Responsible for the Data Mining.
- Esther_399 - Responsible for EDA.
- Julie_996 - Responsible for EDA.

Dataset Overview

The dataset consisted of company-level financial metrics across multiple years, including:

- Stock Price
- Revenue
- Profit Margin
- Market Capitalization
- Sector and cluster groupings

ETL and Data Preparation

The project began with:

- **Loading and cleaning the dataset** in Python (Pandas).
- **Handling missing values** and normalizing features like revenue and stock price.
- **Transforming the data** by adding Profit_Margin and Revenue_Growth.

Exploratory Data Analysis (EDA)

We conducted in-depth EDA to:

- Examine how data is distributed across companies and time periods using count plots for `Company` and `Date`.

- Understand the distribution shapes of key financial variables using boxplots and histograms to visualize revenue and profit variations.
- Test whether mean `Net_Income_Millions` significantly differs between companies using the ANOVA test.
- Identify which company pairs differ significantly using Welch's t-tests for all pairs (unequal variance).

Feature Engineering

New features were constructed to support advanced analysis:

- **Profit Margin Classification** (e.g., high vs. low performing).
- **Cluster Labels** (from K-means clustering).
- **Anomaly Labels** using Isolation Forest algorithm.
- Aggregated values per cluster: average profit margin, total revenue, etc.

These features were essential in surfacing deeper insights during modeling and visualization.

Modeling Insights

We applied:

- **K-Means Clustering** to group companies into performance-based clusters.
- **Isolation Forest** to detect unusual patterns in stock and revenue behavior.

Key findings:

- Cluster `0` and `2` showed **higher average profit margins**
- Anomalies were mostly in **lower-performing companies** or during early 2022
- Cluster `1` companies had lower margins but contributed **significant revenue**

Key Visual Insights (from Power BI Dashboard)

Based on the Power BI dashboard, we extracted the following insights:

1. **TechCorp consistently leads** in stock price growth over time, suggesting sustained performance and investment attractiveness.
2. **Cluster 2** companies have high profit margins, making them ideal for profit-driven investment strategies.
3. A total of **10 anomalies** were detected, many during Q1 2022, signaling possible market disruptions or outlier behavior.
4. **High Performing Companies** make up ~22% of the dataset, providing a strong focus segment for deeper portfolio analysis.
5. **Revenue per Cluster** shows that even lower-margin clusters (e.g., Cluster 1) can contribute significantly in absolute revenue—relevant for volume-based business models.

Business Implications

- **Investors** can use these clusters to identify high-potential companies and time investments based on trend signals.
- **Analysts** can flag and investigate anomalies for risk monitoring or fraud detection.
- **Executives** can focus on high-performing clusters for targeted expansion or acquisition.
- **Strategy teams** gain a visual tool to combine quantitative and categorical insights across time and sectors.