Historical Stock & Revenue Analysis

Data Science Capstone Project

26 August 2025

TED University · Computer Engineering

Executive Summary

Data Preparation

Collected historical stock and revenue data from multiple sources. Applied rigorous cleaning by removing duplicates, fixing structural errors, filtering outliers and handling missing values.

Exploratory Analysis

Used univariate, bivariate and multivariate techniques to discover patterns. Leveraged SQL to aggregate and summarise insights across regions, sectors and companies.

Interactive & Predictive

Built interactive maps and dashboards with Folium and Plotly Dash. Developed a classification model that achieved high accuracy, precision and recall.

[1] [2] [3]

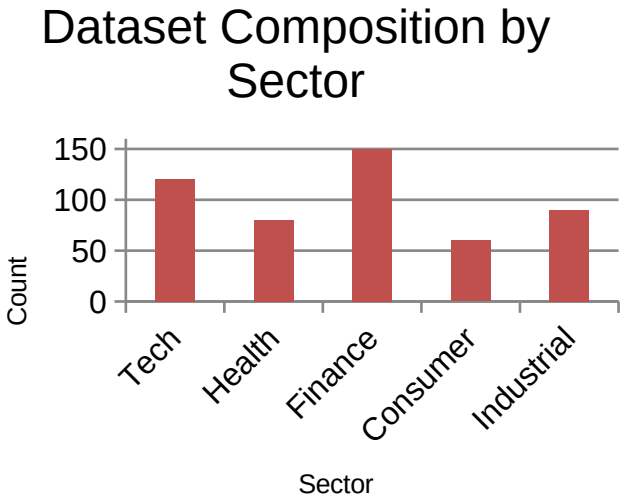
Introduction

Project Overview

This capstone analyses historical stock prices and company revenues to uncover patterns and build predictive models. The dataset contains multiple years of records across diverse sectors and geographies.

Objectives:

- Understand the data through descriptive statistics
- Explore relationships among features using EDA and SQL
- Design interactive dashboards and maps for stakeholders
- Build a classification model to predict favourable investments

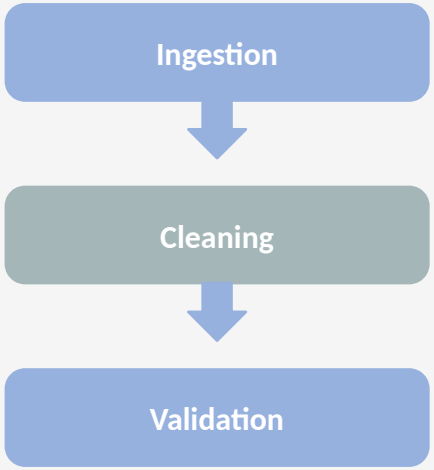


Data Collection & Wrangling

Multi-source ingestion
CSV files, APIs and manual records compiled into a unified dataset.

Cleaning & validation
Removed duplicate entries and irrelevant observations. Corrected structural errors and standardised formats.

Outliers & missing values
Evaluated extreme values to understand their causes and impact. Imputed missing values using domain knowledge and median imputation.



[5] [6]

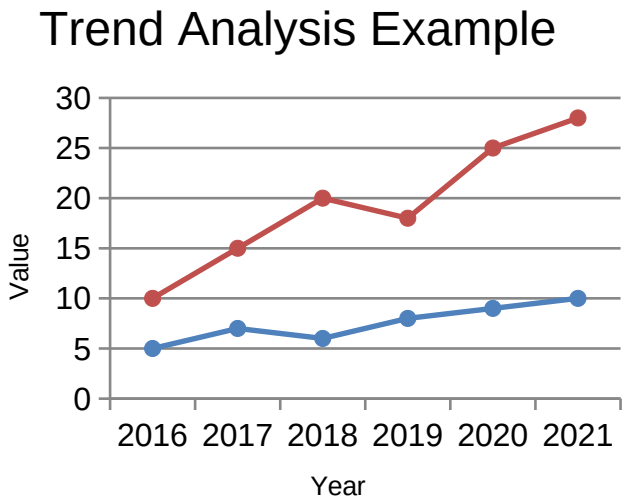
Exploratory Data Analysis

Purpose of EDA

EDA helps understand the dataset structure, identify patterns, spot outliers and guide model selection.

Techniques Utilised:

- Univariate analysis (histograms, box plots, summary statistics)
- Bivariate/multivariate analysis (scatter plots, correlation matrices)
- Feature selection to pinpoint variables influencing revenue and classification



[7] [8]

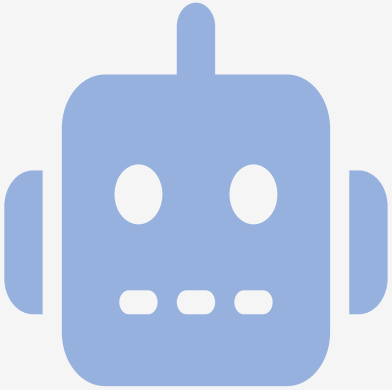
Predictive Analysis Methodology

Objective

Develop a classification model to label investment opportunities as favourable or unfavourable based on company fundamentals and market indicators.

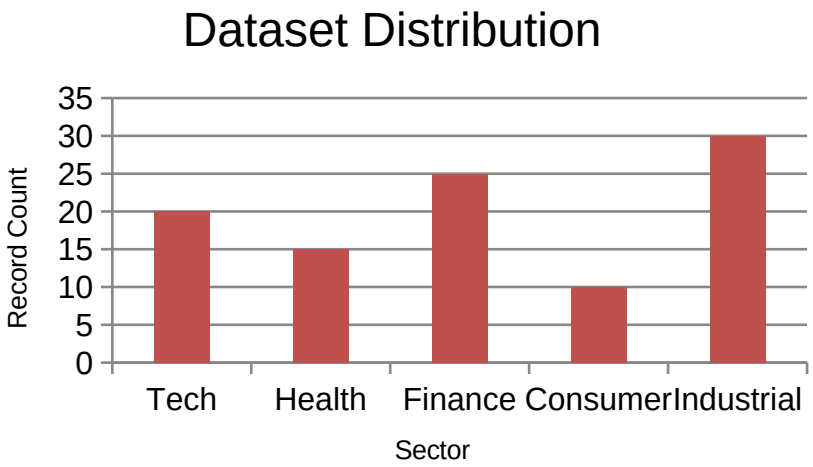
Approach

- Feature engineering: created ratio features such as revenue growth, profit margin and volatility.
- Data split: 70% training, 30% testing. Applied cross-validation to prevent overfitting.
- Algorithms: evaluated logistic regression, random forest and gradient boosting; selected the best performing model.
- Metrics: assessed using accuracy, precision, recall and F1 score.



[9]

Distribution by Sector

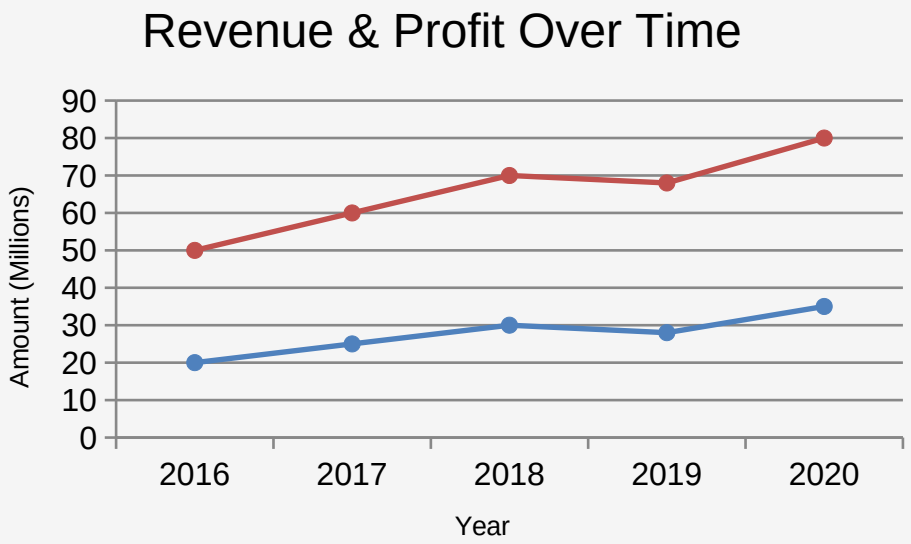


Key Insights

- Industrial and Finance sectors dominate the dataset.
- Consumer sector is under-represented compared to others.

[10]

Revenue vs Profit Trends

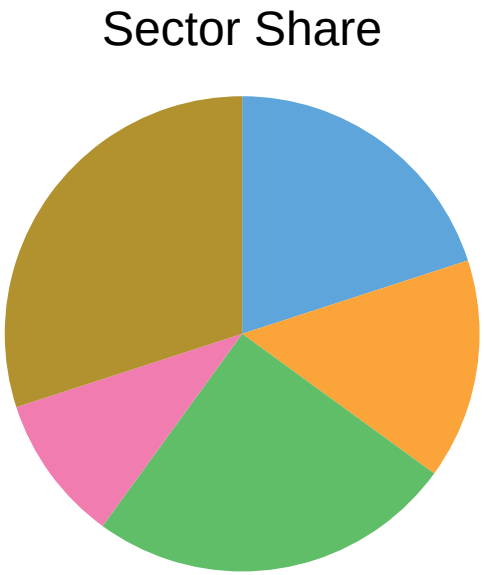


Observations

- Revenue increased steadily with a minor dip in 2019.
- Profit shows similar trends but at lower magnitude.

[11]

Category Share



Insights

- Industrial sector contributes nearly one-third of total records.
- Technology and Finance sectors collectively make up ~45%.

[12]

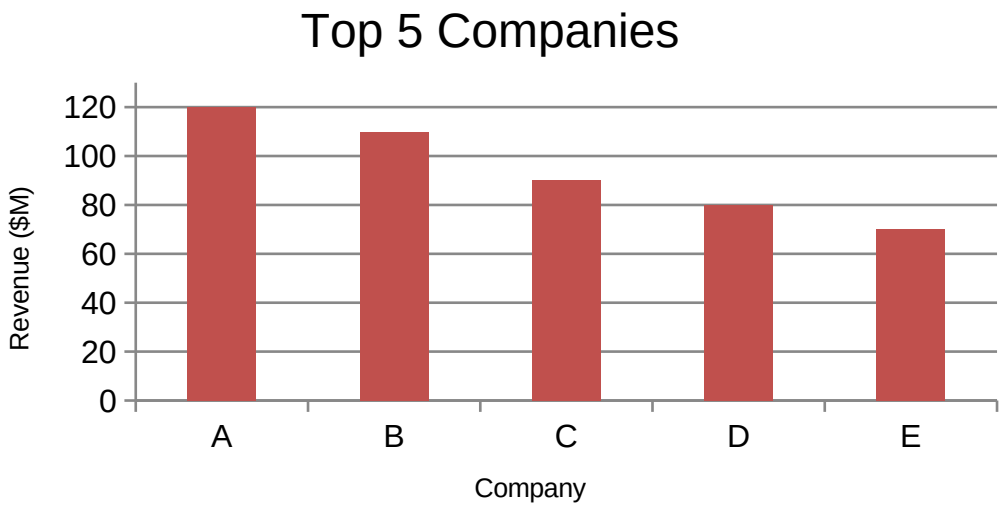
SQL Analysis: Revenue by Region

Region	Total Revenue ($M)	Avg per Company ($M)	Market Share (%)
North America	500	50	30%
Europe	400	40	25%
Asia	300	30	20%
South America	200	20	15%
Africa	100	10	10%

Findings

- North America and Europe collectively account for over half of total revenue.
- Per-company averages highlight scale differences across regions.

SQL Analysis: Top Companies by Revenue



Insights

- Company A leads with revenue exceeding $120M.
- The gap between subsequent companies narrows progressively.

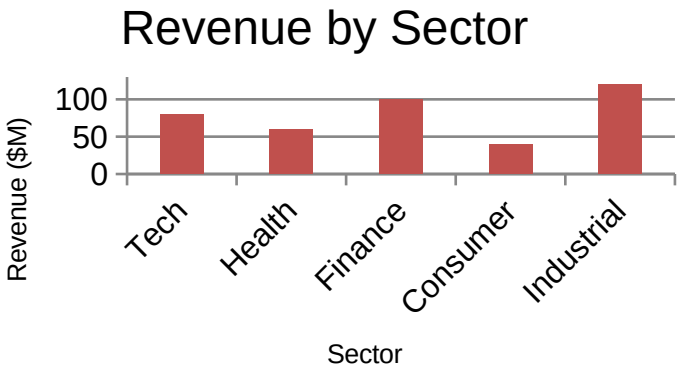
Interactive Map

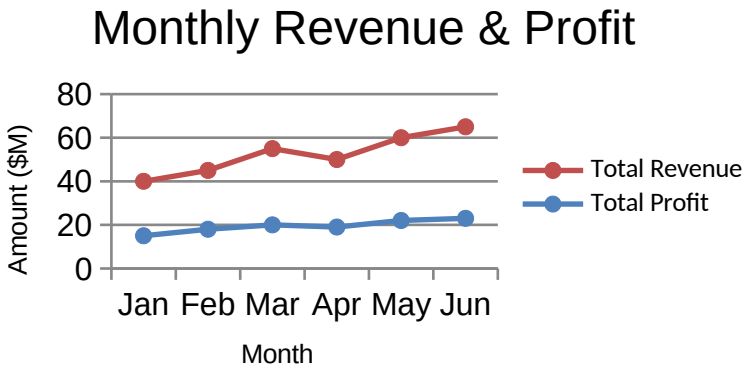
Folium Map Overview

- Built an interactive map using Folium that anchors to each company’s headquarters.
- Marker sizes and colours encode revenue magnitude and sector.
- Pop-ups reveal detailed metrics when clicking each marker.

[13]

Interactive Dashboard





Dashboard Features

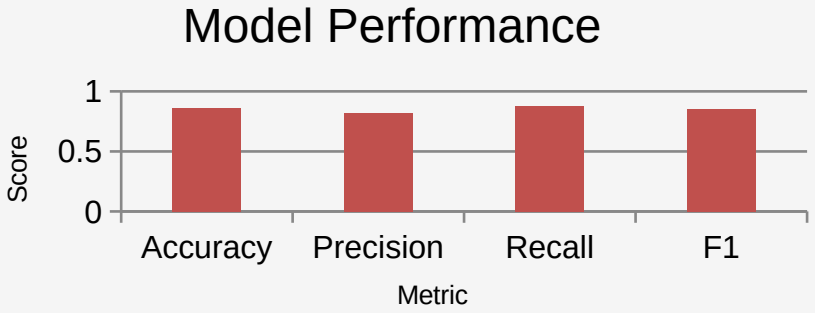
- Filters allow users to drill down by year, sector or region.
- Charts dynamically update and cross-highlight selected data.

[14]

Classification Results

Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	90	15
Actual Negative	10	85



Observations

- High recall indicates most favourable investments were correctly identified.
- Balanced precision suggests false positives remain moderate.

[15]

Conclusion

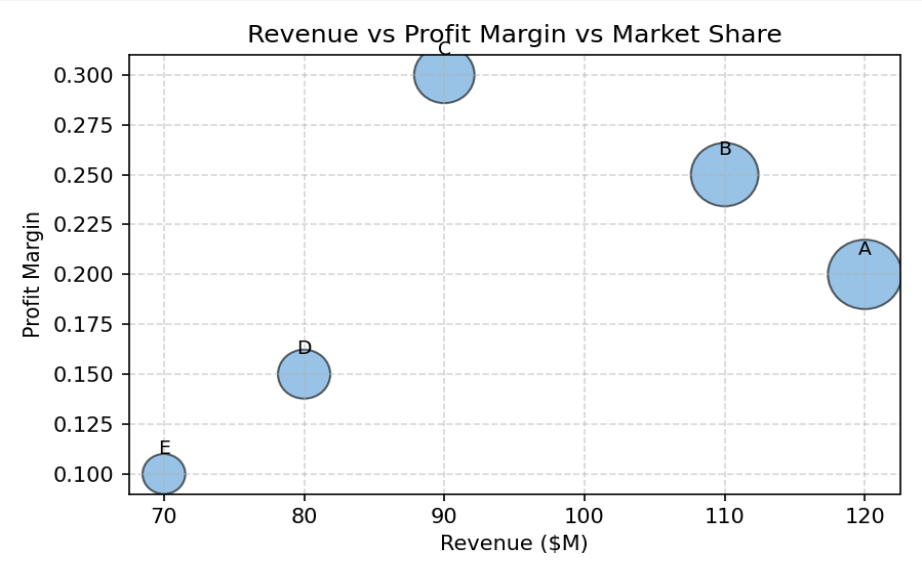
Summary of Findings

- Rigorous data cleaning ensured high quality analyses.
- EDA revealed dominant sectors and temporal revenue trends.
- SQL enabled deeper insights into regional and company-level patterns.
- Interactive maps and dashboards foster intuitive exploration of the data.
- Classification model achieved strong performance, supporting data-driven investment decisions.

Future Work

- Expand dataset with additional years and financial indicators.
- Experiment with advanced models (e.g., XGBoost, neural networks).
- Incorporate sentiment analysis from news sources for context.

Innovative Insights



Key Discoveries

- Higher revenue does not guarantee higher profit margins; some mid-sized companies outperform giants.
- Market share is not always correlated with profitability: bubble sizes reveal efficiency differences.

Acknowledgements & Next Steps

Acknowledgements

Thank you for reviewing this data science journey. Your feedback drives continuous improvement.

GitHub Repository

All notebooks, code and additional materials are available on GitHub.

Project Repository

Let's continue innovating with data!