**Table of Contents**

Contents

[INTRODUCTION 1](#_Toc149883928)

[DATA CLEANING 2](#_Toc149883929)

[Visualizations and Data Analysis 3](#_Toc149883930)

[Investigation and Findings: 6](#_Toc149883931)

[Conclusion: 7](#_Toc149883932)

# INTRODUCTION

This financial dataset contains 16 key metrics for 100 public companies randomly sampled from over 3,500, including revenue, assets, working capital, and cash per share. It covers 5 years of webscraped filings data with the target variable indicating yearly stock price movement. The cleaned sample allows modeling the relationship between financials and stock performance.

**DATA DEMONSTRATION**

I selected the following 15 financial indicator columns:

Revenue, Cost of Revenue, Gross Profit, Investments, Net Income, Receivables, Total current assets, Tax assets, Total assets, Total current liabilities, Other Assets, returnOnAssets, Net Income per Share, Cash per Share, Working Capital and Class as my target varialbe

**Data Selection Method**:

I added a "Random" column and used the RANDBETWEEN formula to generate a random number for each row:

**=RANDBETWEEN(1,16586)**

Sorted the data by the Random column to shuffle the rows into a random order and kept the first 100 rows/companies, then deleted the remaining rows.

This resulted in a final cleaned dataset with 100 randomly selected companies over 5 years and 15 financial indicators plus the Class target variable.

The final cleaned data has 100 rows and 16 columns. The steps taken allow for a randomized, manageable sample size and a relevant subset of predictors and the target for modeling. The data has been checked for inconsistencies, formatted properly, and structured optimally for analysis.

**Possible Ethical Implication**

1. Using imperfect data could lead to inaccurate or misleading conclusions. The data was webscraped and may contain errors or inconsistencies that could skew results. Steps should be taken to verify data quality and note any limitations.

2. Biases may be baked into the data that lead to unfair outcomes. For example, the companies represented may not proportionally represent different industries, locations, size of companies etc. Samples should aim to represent population diversity where possible. Models built on imbalanced data may disadvantage certain groups.

In general, when using data for analysis, it is important to consider:

* How accurate and complete is the data? Are there incorrect or missing values that could distort findings?
* Does the data fairly represent the intended population? Or does it reflect biases towards certain groups?
* How will model results and insights be used? Could they negatively impact or exclude certain groups unfairly?
* Is the data being used ethically and responsibly to generate insights for social good?

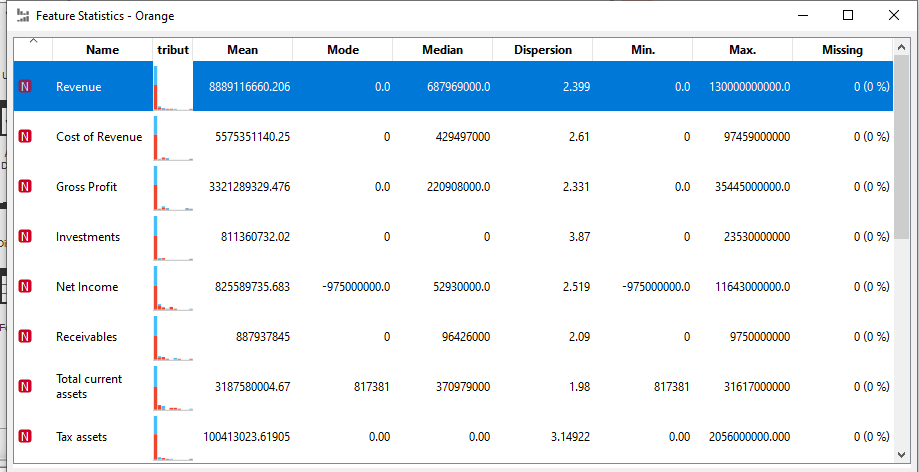
# 

# DATA CLEANING

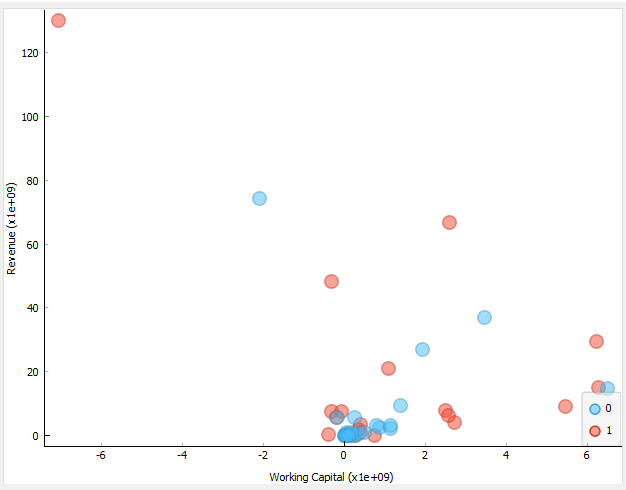
Preliminary data exploration in Orange revealed the presence of missing values across multiple attributes in the dataset. To handle these missing data points, the Preprocess widget was utilized to filter out rows containing any missing values, thus removing incomplete instances listwise. The Remove Missing Values function in preprocessor widget was applied with the default settings, which eliminates objects with NA values on a perinstance basis. After running the preprocessing widget, the resulting data matrix was reduced from the original sample size down to 63 total complete instances.

# Visualizations and Data Analysis

**Summary Statistics**

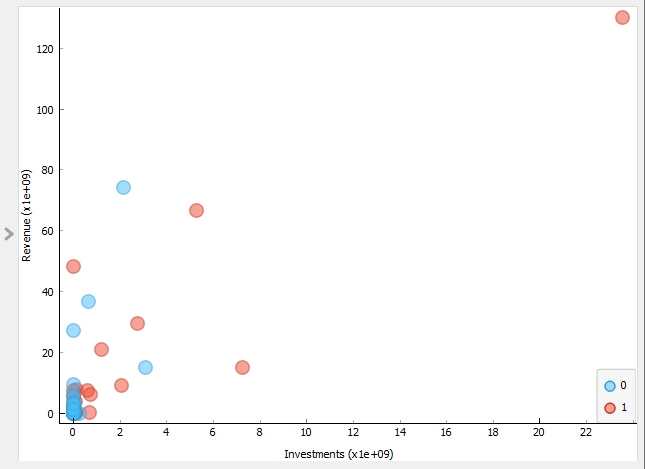


Generated summary statistics in R to inspect features. High maximum revenue relative to the mean indicates potential outliers. Additional missing value checks showed no remaining NA values. Summary displays distributional statistics including minimum, maximum, mean, and mode for each predictor. This initial exploratory data analysis provides insights into the central tendency and spread of key variables.



I explore the data more by plotting a scatter plot of working capital against Revenue and labeled it based on class (1=stock increase during the year, 0=stock decreases during the year) to understand if growth of stock depends on the working capital. One could see on the graph the presence of an outlier which means there are stock which very low working capital but very high revenue while other stocks are distributed randomly which shows the working capital does not really lead to high revenue.

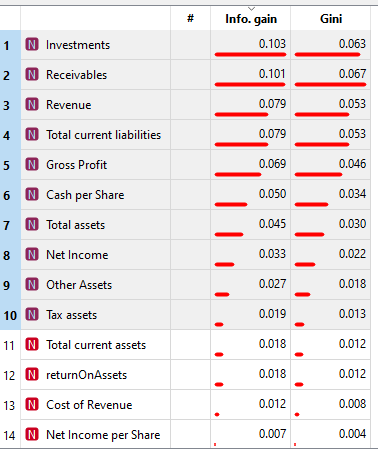
Which leads to checking if investment might play an important role when it comes to revenue generated by each stock.



The scatter plot shows stocks with very high investments have high revenues, while lower investments correspond to lower revenues. Specifically, stocks with small investments rarely yield high revenues, as depicted by the cluster of points in the bottom left. The positive correlation indicates companies that invest more tend to have higher revenues.

# Investigation and Findings:

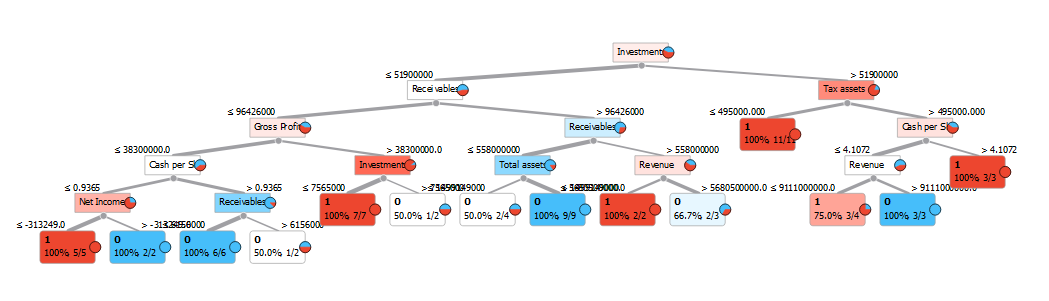
**Feature Selection:**



The top 10 financial indicators ranked by information gain and Gini decrease for predicting the target Class are: Investments, Receivables e.t.c. This ranking shows the relative importance of each predictor in modeling Class, with Investments having the highest variable importance score of 0.103 based on information gain.

**Decision Tree**:

We construct a decision tree model to better understand our data and check the possibility to classify the future performance of a stock by looking at the chosen financial indicators.

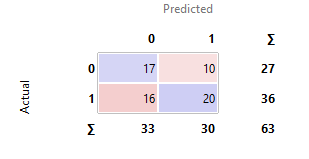


From the decision tree, investment is the best classifier applicable to our model and split into working capital and tax assets which shows that investments greater than 51900000 and have tax asset less than 495000 have highest probability of increasing during the year

**Accuracy Test Score**:



Using the Test and Score function to calculate the accuracy of our model which result to 58% accuracy. This shows more data is needed or there is probability of diversity and bias in the data resulting to very low accuracy score. A Confusion matrix is then constructed to see the values predicted correctly and wrongly



The confusion matrix indicates 17 incorrect predictions and 20 correct predictions compared to the actual target values. This provides the number of false negatives and true positives for assessing model performance.

# Conclusion:

In this analysis, decision trees modeled stock price changes using key financial metrics. The optimal tree utilized Investment amount and Tax Assets as top splitters, achieving 58% test accuracy. Results provide a baseline modeling approach and reveal correlations between higher investments, lower taxes, and increased probability of price growth. Further feature engineering and algorithms could enhance accuracy in predicting price fluctuations from financial filings data.