**HANOI SCHOOL OF BUSINESS AND MANAGEMENT**

Vietnam National University, Hanoi

PROJECT REPORT

SUBJECT: OVERVIEW OF DATA SCIENCE

Topic:

**ANALYSIS OF BUSINESS PERFORMANCE MANAGEMENT OF SOFT DRINK ENTERPRISES THROUGH COCA-COLA**

**SUPERVISOR**:

Dr. Emmanuel Plan

Mr. Nguyễn Huy Anh



PRESENTATION DATE: 27/12/2024

**STUDENT GROUP RESPONSIBLE:** GROUP 01

**CLASS:** MET04 (Management of enterprise and technology)

***Hanoi, December 27, 2024***

**ACKNOWLEDGMENTS**

The research team would like to express our sincere thanks to all those who have supported us in the process of implementing the topic "Analysis of business performance management of soft drink enterprises, through the main company Coca Cola".

In particular, the research team would like to express our deepest and most sincere gratitude to Dr. Emmanuel Plan and Mr. Nguyen Huy Anh - the lecturers who directly guided and instructed us wholeheartedly, as well as provided us with the most favorable conditions to complete this scientific research project.

Although the research team has made great efforts to complete the topic as completely as possible, the research project is inevitably flawed. The research team respectfully hopes to receive sympathy and comments from teachers and students so that we can gain experience for future research.

Sincerely thank you!

***Hanoi, December 27, 2024.***

***Research Group 01***

**SELF-EVALUATION TABLE FOR GROUP 01**

**(Evaluation Scale: 10)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Full name | Student ID | Assigned tasks | Score |
| 1 | Phạm Hồng Nhung (**Leader**) | 22080070 | Overall project management and strategic decision -making | 100% |
| 2 | Nguyễn Thị Thùy Dung | 22080021 | Search and synthesize information. Prepare Report and PowerPoint for Presentation. | 100% |
| 3 | Nguyễn Hữu Minh Hoàng | 22080034 | Search for information about Coca Cola and Analyze output data. | 100% |
| 4 | Trần Phúc Hưng | 22080037 | Data visualization of Stock data (linear and logistic implementation). | 100% |
| 5 | Nguyễn Thương Huyền | 22080042 | Data visualization of Human Resources data (linear and logistic implementation). | 100% |
| 6 | Lê Mai Linh | 22080047 | Search and synthesize information. Prepare Report and PowerPoint for Presentation. | 100% |
| 7 | Trần Hương Thảo | 22080083 | Search for information about Coca Cola and Analyze output data. | 100% |
| 8 | Nguyễn Duy Tùng | 22080092 | Search for information about Coca Cola and Analyze output data. Prepare Appendices. | 100% |

*Hanoi, December 27, 2024*

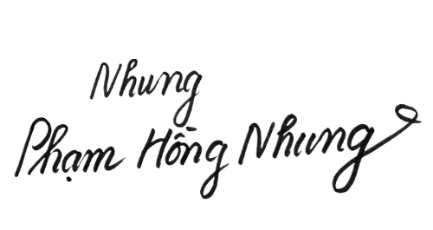
 **Group Leader**

TABLE OF CONTENTS

[PROLOGUE 1](#_Toc185583938)

[CHAPTER 1: INTRODUCTION 2](#_Toc185583939)

[CHAPTER 2: GENERAL STUDIES ON COCA-COLA'S BUSINESS PERFORMANCE MANAGEMENT 2](#_Toc185583940)

[CHAPTER 3: DATA ANALYSIS 3](#_Toc185583941)

[1.STOCK 3](#_Toc185583942)

[2.HUMAN RESOURCE DATASET 10](#_Toc185583943)

[CHAPTER 4: CONCLUSION 15](#_Toc185583944)

[SUMMARY OF KEY FINDINGS 15](#_Toc185583945)

# PROLOGUE

The research topic is to evaluate the business management performance of soft drink enterprises, through the main company, Coca Cola. The research focuses on the key performance indicators, management strategies, and factors affecting the sustainable success of Coca-Cola in the context of fierce market competition.

The reason why we choose this topic is to know how this company manages performance will help to better understand the successful business strategies applied by Coca-Cola. So that we can provide some valuable lessons for other businesses in the same field. Moreover, Coca-Cola has experienced many challenges and opportunities, from changing consumer habits to increasing competition from emerging brands. These factors highlight the importance of performance management in maintaining competitive position and sustainable development.

The purpose of this study is to identify the influencing factors as well as performance management methods in enterprises to maintain market position and optimize profits. Furthermore, through the study, the research team hopes to draw out business strategies that can be applied to other enterprises in the beverage industry as well as similar fields, to improve management efficiency and promote sustainable development.

# CHAPTER 1: INTRODUCTION

As consumers increasingly prioritize healthy and sustainable products, competition intensifies among companies like Coca-Cola and PepsiCo in areas such as pricing, brand promotion, and product offerings. Coca-Cola (CCE) excels in innovative promotions and boasts a portfolio of over 4,300 products, while PepsiCo generates billions annually with 22 major brands. This study evaluates its business performance through financial indicators, particularly stock metrics, and human resource factors, using both qualitative and quantitative methods. Qualitative analysis identifies strategies affecting performance, while quantitative approaches apply algorithms like linear regression and logistic regression to assess data. The research aims to determine how financial ratios reflect performance and which human factors influence sustainability and growth.

# CHAPTER 2: GENERAL STUDIES ON COCA-COLA'S BUSINESS PERFORMANCE MANAGEMENT

A diagram of a business performance management

Description automatically generatedCCE leads the carbonated soft drink market with over 400 brands in 200 countries. By leveraging strategies like the "glocal" model and Balanced Scorecard (BSC), this company can optimize costs and boost efficiency. Culturally tailored marketing increased revenue by 8% in Asia-Pacific, while joint ventures in China expanded local production. With a 24% profit margin and a 0.48 debt/equity ratio, it demonstrates financial stability and innovation. Its human resource strategies, including group interviews, internal selection, and the 70:20:10 training model, enhance employee motivation, efficiency, and morale, while competitive compensation and engagement improve satisfaction and reduce turnover. Globally, CCE sustains success by leveraging its $61 billion brand value, advanced supply chains, and healthier products, while addressing challenges through digital transformation, consumer behavior analysis, and smart manufacturing. Besides, Decision Support Systems (DSS) further aid strategic decision-making by simplifying complexities and aligning actions with organizational goals.

# 

# CHAPTER 3: DATA ANALYSIS

1. **STOCK**
   1. **INTRODUCTION**

In this notebook you will implement the regression algorithm: Linear Regression:

**Dataset description:** The dataset used in this project is related to Coca-Cola stock prices and focuses on building a Linear Regression model to predict future closing prices. Key features in the dataset include the opening price as the input variable and the closing price as the target variable.

**a. Data Pre-Processing**

A table of numbers and symbols

Description automatically generatedStep 1: We will check the data types for data cleaning purposes.

A screenshot of a computer

Description automatically generated

Step 2: The values ​​in these datasets are not null and have no missing values

Step 3: Datasets have 2 values ​​that need to be removed: Dividends and Stock Splits because the values ​​in this column do not change

In this data preprocessing step, we will check the data to ensure that the data is clean

Summary: After checking the values ​​in the data set, all the values ​​are correct and there are no null or missing values.

A close up of a sign

Description automatically generatedObserving the dataframe, the Date column will not be necessary to process in building a machine learning model and is in the form of an object, we can remove it

In addition, there are 2 other variables, Dividends and Stocks Splits, whose values ​​remain constant for almost a period of time. This could be a noise or extraneous variable, so it is not necessary and we can remove it.

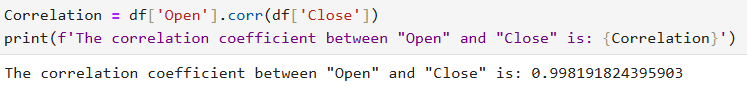
A collage of graphs

Description automatically generated

After preparing the data, we can plot how the target (Close value) relates to the input (Open value). In this dataset, we're using the closing price to predict future values, and the opening price as our input.

Data points with constant volume (i.e., those along a horizontal line on the volume chart) likely have a weak connection to other factors. The chart we will be using is the second one on top.

In this dataset, we will choose the variable "Close" as the target variable with the purpose of predicting the closing price of coca cola company.

****

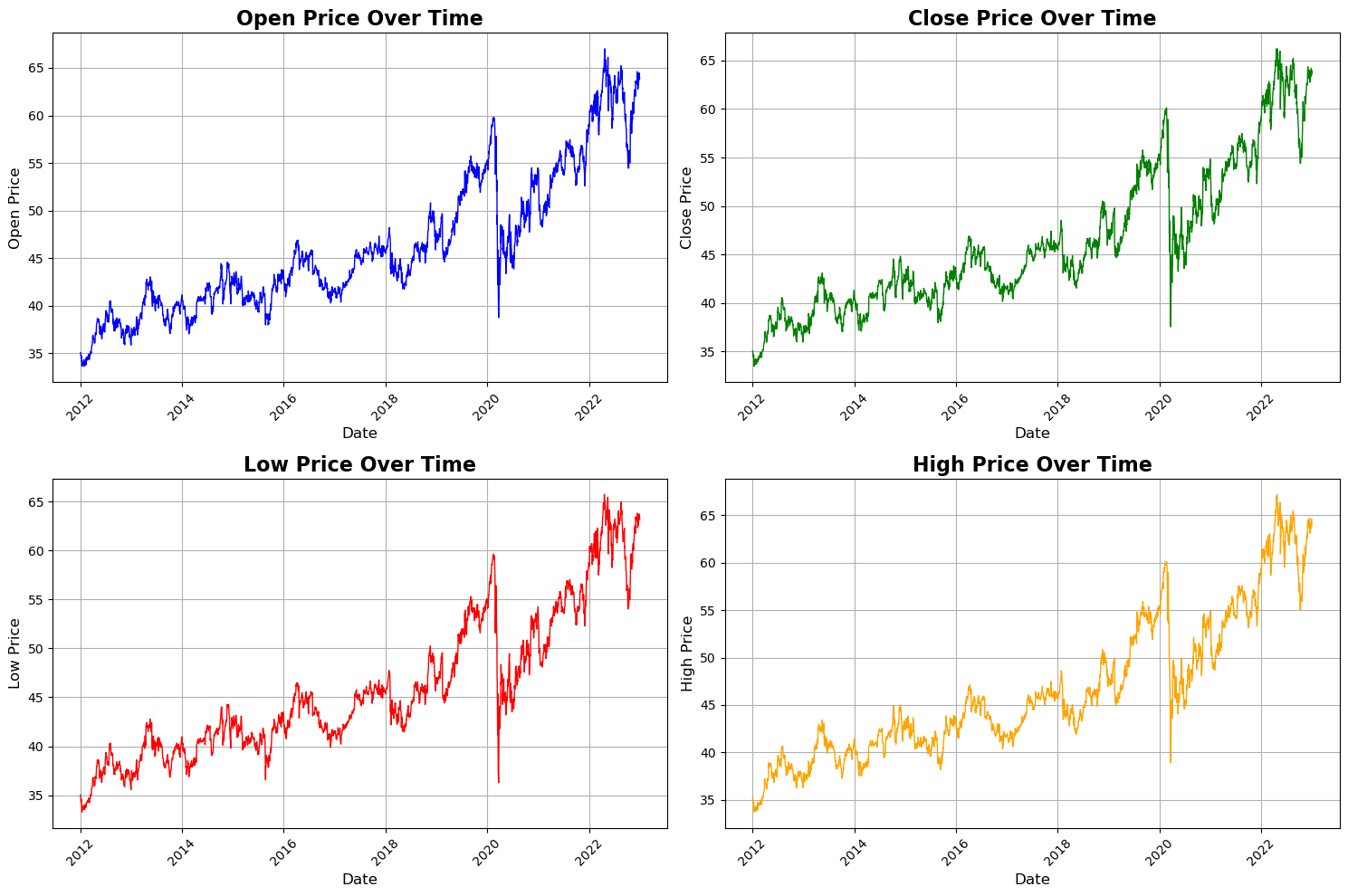
The correlation coefficient between "Open" and "Close" is: 0.998191824395903

We can check the correlation between Close and Open variables through Correlation. Their correlation index is 0.998, which shows that these two variables have a strong correlation with each other.

A blue line graph with numbers

Description automatically generated

This KDE chart helps to identify the peaks and valleys of the Close variable, this chart helps us to understand the distribution and frequency of the Close value.



This OHLC chart shows the company's stock price over time. Looking at the chart, we can see how each value fluctuates over time, but if we compare it as a whole in machine learning, these values ​​can help us choose input data and evaluate model performance.

**b. Building the model**

After preprocessing the data, we will proceed to build the model.

A blue line with red line

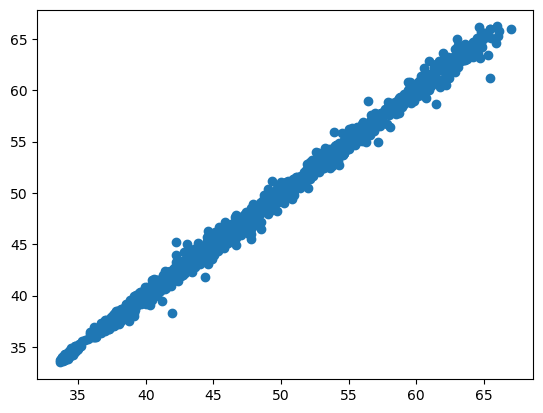
Description automatically generated

Chart A Chart B

After preprocessing the data, we can see that this is a visual graph of 2 variables, the target variable Close and the input variable Open. We proceed to use the linear regression machine learning model to calculate the slope and intercept coefficients.

[Chart A] The cofficient(s) a is(are) [[0.99860318]]

The intercept b is [0.06809476]

[Chart B] The cofficient(s) a is(are) [[0.99860318]]

The intercept b is [0.06809476]

After calculating the 2 coefficients, we get the red regression line as shown above.

We proceed to divide the test data file into 4 variables X\_test, X\_train, y\_test, y\_train and then conduct training to produce y\_test\_pred and y\_train\_pred

**c. Make Prediction**

Next, we proceed to predict y\_test\_pred and y\_train\_pred.After building the model, we will proceed to make prediction.

R2\_train is: 0.9963111485528755

R2\_test is: 0.9966582066837604

MSE\_train is: 0.1948712018929935

MSE\_test is: 0.19385699150272606

Evaluating the model on both the Test set and the Training set using the R2 and MSE indices, we see that these two indices are also close to each other, proving that the model predicts well.

A screenshot of a graph

Description automatically generated

Next, we calculate Residual using the formula

Residual test = y\_test - y\_test\_pred

Residual train = y\_train - y\_train\_pred

Visually on the chart, we see that the data of Residual test is in the range (-2;2) and Residual train is also in the range (-2;2) and the frequency of the 2 Residual is around 1200 and 250. We conclude that the Residual results of both datasets evaluate this model quite well and effectively.

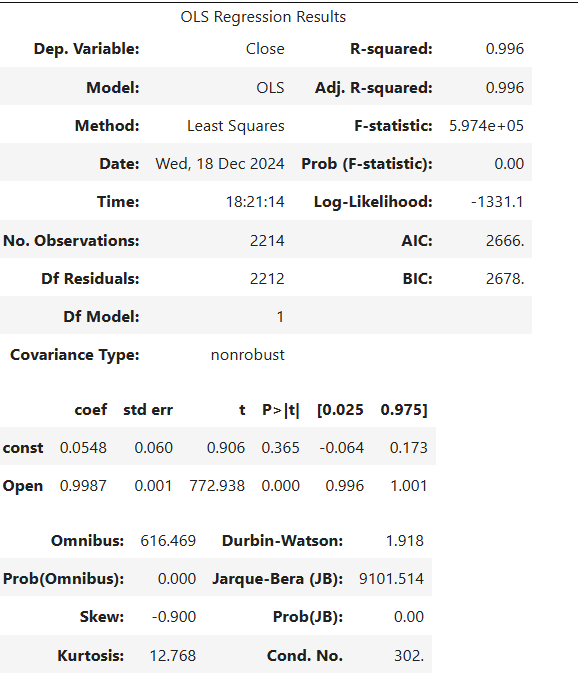
A comparison of blue dots

Description automatically generated

durbin\_watson 1.9126791959757588

A screenshot of a computer

Description automatically generated Notes [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. The R-squared and Adj R\_squared values ​​are very high at 0.996, indicating that this model solution captures 99.6% of the Close price fluctuations in the market. The Open variable's numerical recovery coefficient is 0.9987, selecting the Close price characteristic relationship. The Open variable's P value is 0.00, which is very small, indicating that the Open variable is statistically significant in the model.



Notes [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. Errors range from -2 to 2 with frequencies around 1200 and 250 for both data sets, showing that the model errors are quite well distributed and the predictions do not deviate much from the actual values. Durbin-Waston is 1.91, which is approximately equal to 2 indicating that there is no significant auto-correlation in the model error. It can be seen that the linear regression model does not violate the assumption of residual correlation, helping to ensure accurate predictions.

1. **HUMAN RESOURCE DATASET**
   1. **INTRODUCTION**

In this notebook you will implement the classification algorithm: Logistic Regression:

Dataset description: With massive enterprises like Coca-Cola, HR data is typically confidential and will not be public. Therefore, our team has chosen a publicly accessible HR dataset for analysis to provide general insights applicable to Coca-Cola and align with the topic of enterprise business performance.

This is a list of sections: Data Preprocessing, building the model, make prediction

1. **Data Preprocessing**

Step 1: We will check the data types for data cleaning purposes.

Step 2: In this dataset we will choose the variable "Attrition" as the target variable with the purpose of predicting the number of employees who have left.

Step 3: We will deal with binary problems

Step 4: First, we need to encode Attrition to binary

Step 5: Next, we will separate X (feature) and y (target variable)

Step 6: Drop irrelevant or constant columns

Step 7: After separating X and y, we can identify the numberic feature and the categorical feature

Step 8: Next, we need to standardize the data and One-hot encoding for the categorical variable

Step 9: Let us visualize the pairlot

**A graph of a number of blue and orange dots

Description automatically generated with medium confidence** To predict employee attrition, we first encode the "Attrition" variable into binary form as the target variable. Next, we separate the dataset into X (features) and y (target), dropping any irrelevant or constant columns. Afterward, we distinguish between numerical and categorical features. We standardize the numerical data and apply One-Hot Encoding for the categorical variables. Finally, we can visualize the relationships between features using a pairplot for better insight into the data.

1. **Building the model**

Step 1: After preprocessing the data, we will proceed to build the model.

**A screenshot of a computer code

Description automatically generated** Step 2: Proceed to check the classes in the target variable

**A graph with blue rectangular bars

Description automatically generated with medium confidence**

**A graph of age distribution

Description automatically generated**The dataset shows the imbalance between the classes in the "Attrition" variable, with fewer employees leaving compared to those who stayed. To address this, we use two key techniques to ensure that the data is balanced. First, during the train-test split, we apply stratify=y, which preserves the proportion of each class in both the training and test sets, ensuring consistent class representation. Second, we will set class\_weight='balanced' in our model, which adjusts the importance of each class, giving more weight to the minority class and helping the model treat both classes equally.

This chart shows the relationship between "Attrition" and “Age” in the dataset. It emphasize how age affects whether an employee stays or leaves the company.

**A graph showing a graph of business travel

Description automatically generated with medium confidence**

Step 3: Split the test sets to train and build a logistic regression model

With this graph we split the test sets to train and build the logistic regression model. There will be 3 set: “Travel\_rarely”,”Travel\_frequently” and “Non\_travel”. This chart shows the connection between the target variable and the categorical variable

1. **Make Prediction**

Step 1: After building the model, we will proceed to make prediction.

Step 2:Next, we will evaluate accuracy, and construct the confusion matrix as before.

The accuracy is 75.17006802721087 %.

The confusion matrix is [[190 57]

[ 16 31]]

The model's accuracy is 75.17%. This means the model correctly predicted 75.17% of the samples in the test set.

**A screenshot of a graph

Description automatically generated** About confusion matrix:

The value of AUC is 0.7986906710310966

After building the model, we continue to make predictions. Then we will evaluate accuracy, and construct the confusion matrix. The accuracy is 75.17006802721087 percent which means the model correctly predicted 75.17% of the samples in the test set. . The confusion matrix is [[190 57]

[ 16 31]]

190 is the number of correctly predicted class 0 samples (True Negatives). 31 is the number of correctly predicted class 1 samples (True Positives). 57 is the number of class 0 samples that are wrongly predicted to be class 1 (False Positive ). 16 is the number of class 1 samples that are wrongly predicted to be class 0 (False Negatives). The value of AUC is 0.7986906710310966 which mean our model has a good level of performance in distinguishing between classes.

**A graph of a positive rate

Description automatically generated**

The value of AUC is 0.7986906710310966 show that the model performs well. The ROC curve evaluates the performance of a classificaiton model, with TPR on the y-axis and FPR on the x-axis. This allows HR to take proactive measures, such as improving engagement, offering targeted rêtntion strategies, addressing specific concerns to reduce turnover.

# CHAPTER 4: CONCLUSION

**SUMMARY OF KEY FINDINGS**

The analysis of HR and stock data using Logistic Regression and Linear Regression models provided valuable insights into Coca-Cola's business performance. The HR model achieved an accuracy of 75.17% and an AUC of 0.7987, identifying key human factors like age and travel frequency as significant predictors of employee attrition. It shows that employee engagement and tailored retention strategies are crucial for sustaining business performance. Similarly, the stock index shows a strong 0.9987 correlation between opening and closing prices, with R-squared values of 0.996 explaining 99.6% of price fluctuations. Therefore, financial ratios, particularly stock-related metrics have effectively reflected the company's market performance and provide actionable insights for strategic planning.

In conclusion, financial ratios like stock price correlations offer a clear reflection of the company’s business performance by capturing market trends and dynamics. While human factors such as employee age, travel frequency, and engagement significantly impact the company’s ability to sustain and grow its performance. These results underscore the importance of data-driven strategies in both HR management and financial decision-making to ensure continued success and competitiveness.