|  |
| --- |
|  |
| COM6013 |
| Data Mining Level 6 |
|  |
|  |
|  |
|  |



# Abstract

*This study delves into a comprehensive analysis of sales data from* Confectionaries Are Us*, aimed at optimizing profitability across various regions and over time. The dataset encompasses key metrics such as units sold, revenue, cost, and profit, spanning multiple years and countries. Through the application of diverse data mining techniques including regression analysis, clustering, and time series forecasting using ARIMA models, along with meticulous data preprocessing and visualization, we unearthed valuable insights into sales patterns, customer preferences, and profitability drivers. By leveraging these insights, businesses can tailor marketing strategies, anticipate market trends, and make informed decisions to enhance profitability and sustain competitiveness in the evolving market landscape.*

Table of Contents

**Abstract**2

**Introduction4**

**Task 15-23**

**Task 124-25**

**Appendix26-42**

1. **Introduction**

Welcome to the Python Data Mining Mini Project! We will dig into a dataset provided by *Confectionaries Are Us* to find ways to increase their profits. I, as a new junior analyst, have the job of exploring the data to find ways to make the company more profitable, both on a country level and as a whole.

* 1. Goal:

The main aim of this project is to use different models and techniques to find useful insights from the dataset. Before we start analysing, we need to get the dataset ready and understand its characteristics using some basic statistics. This will be the starting point for our analysis, making sure that all our findings are based on a strong understanding of the data.

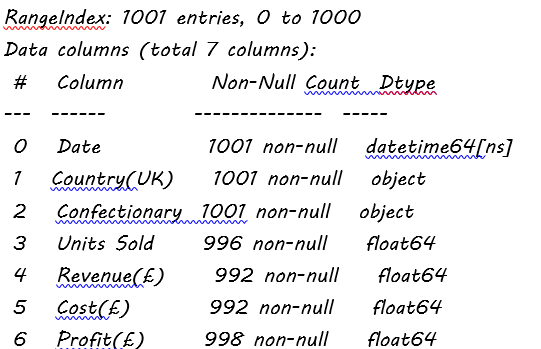
* 1. Plan:
* Prepare the dataset by dealing with missing data, changing data types, and making other needed adjustments.
* Look into the data to find any patterns, trends, and connections.
* Use time-based methods to predict profits over time, for each country and overall.
* Choose the best methods for modeling based on the dataset and our goals.
* Show the analysis clearly with evidence from screenshots of the processes and results.
* **Include all the code for Task 1 in an appendix so that others can check and repeat the analysis.**
  1. Approach:

We will do the analysis completely in Python, using libraries like pandas, NumPy, and scikit-learn for working with the data and creating models. We'll document each step carefully to make sure that all our findings are transparent and can be repeated by others.

* 1. Wrap-up:

By the end of this project, we want to give *Confectionaries Are Us* useful insights and suggestions to help them make better decisions and boost their profits. This data mining investigation will be important for the company's future plans, helping them navigate the changing world of confectionaries with confidence and foresight.

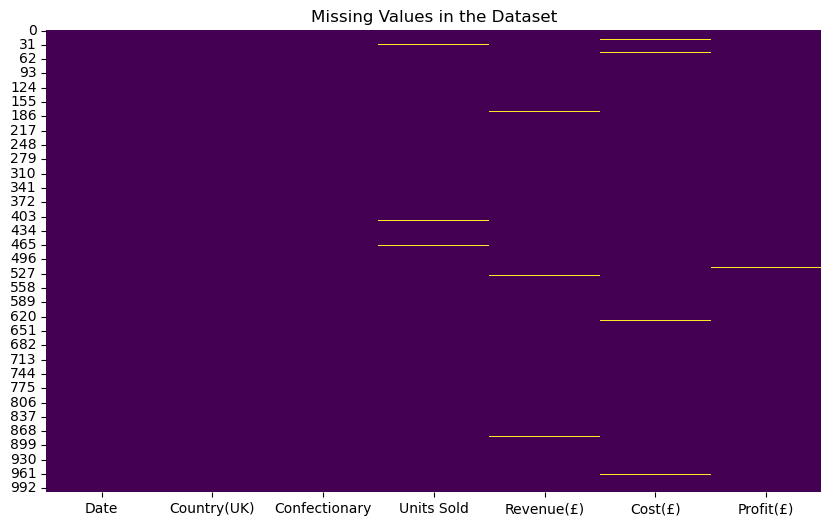
1. **Task 1**
   1. Intro: Embarking on my journey as a junior analyst at *Confectionaries Are Us*, I am tasked with my inaugural data mining project. The primary objective? To decipher the intricacies of the company's dataset and devise strategies to bolster profitability, both locally and globally.
   2. Data Summary:



* 1. Data Preparation: Handling Missing Data:

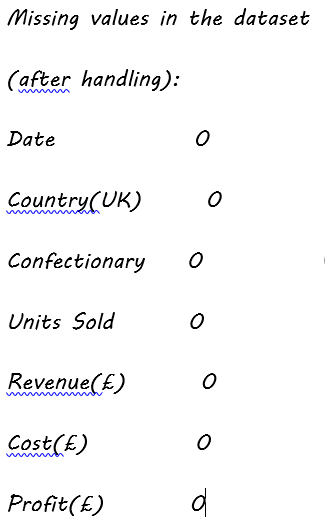
Within any data analysis environment missing data is a very real issue (Arden University, n.d.), addressing missing data is a critical aspect of any data analysis endeavor, including our data mining project at Confectionaries Are Us. Missing data can significantly impact the accuracy and reliability of our findings, necessitating careful handling to maintain the integrity of our analysis. In this section, we'll explore various techniques for managing missing data.

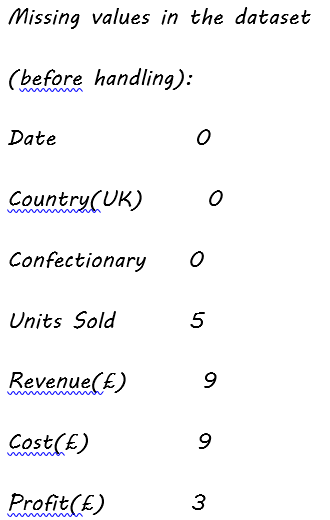
Chart 2.3



The presence of horizontal lines in the missing values chart for various variables indicates the occurrence of missing values within those respective columns. Each line represents a distinct segment where data is absent. In the case of "Units Sold," "Revenue," and "Cost," three, three, and four horizontal lines, respectively, suggest the presence of multiple intervals where values are missing. Conversely, the solitary line in the "Profit" column denotes only one segment with missing data. Understanding the distribution of missing values is crucial for subsequent analysis and decision-making processes. It prompts us to consider strategies for handling missing data, such as imputation or exclusion, to ensure the integrity and reliability of our findings.

In this section, we'll focus on employing imputation as a technique to handle missing data in our dataset. Imputation involves replacing missing values with estimated values based on the available data.





Missing values in the dataset

(before handling):

Date 0

Country(UK) 0

Confectionary 0

Units Sold 5

Revenue(£) 9

Cost(£) 9

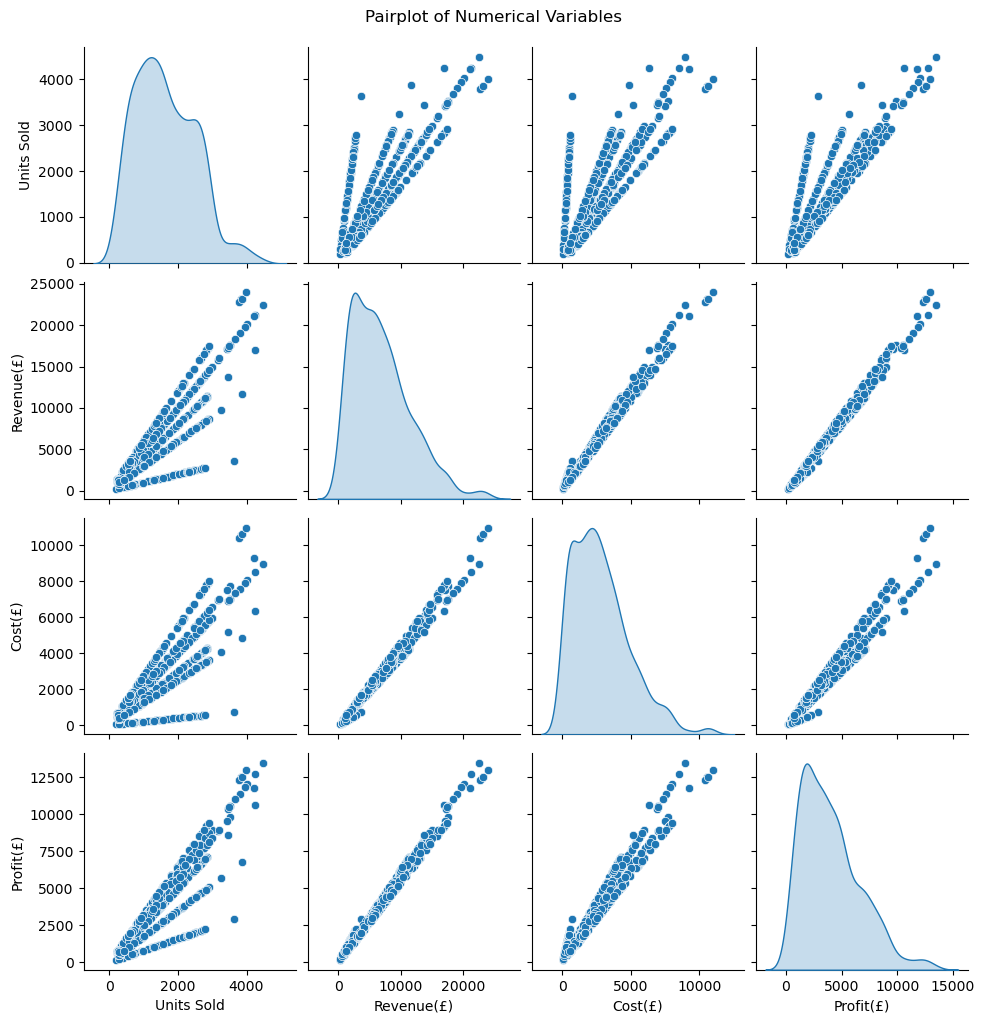
Profit(£) 3

* 1. Data Preparation: Exploring Relationships: Pairplot Analysis of the Dataset:

The pairplot is a comprehensive visualization tool that enables the exploration of relationships between variables in a dataset through scatterplots and distribution plots. Let's delve into its insights.

Chart 2.4.0 shows strong positive relationships between all four variables, suggesting a coherent pattern in their behaviours. Additionally, the right-skewed distribution of each variable indicates that higher values occur less frequently, with the majority of data concentrated towards the lower end of the scale. This insight informs our understanding of the dataset's overall characteristics and potential trends.

Chart 2.4

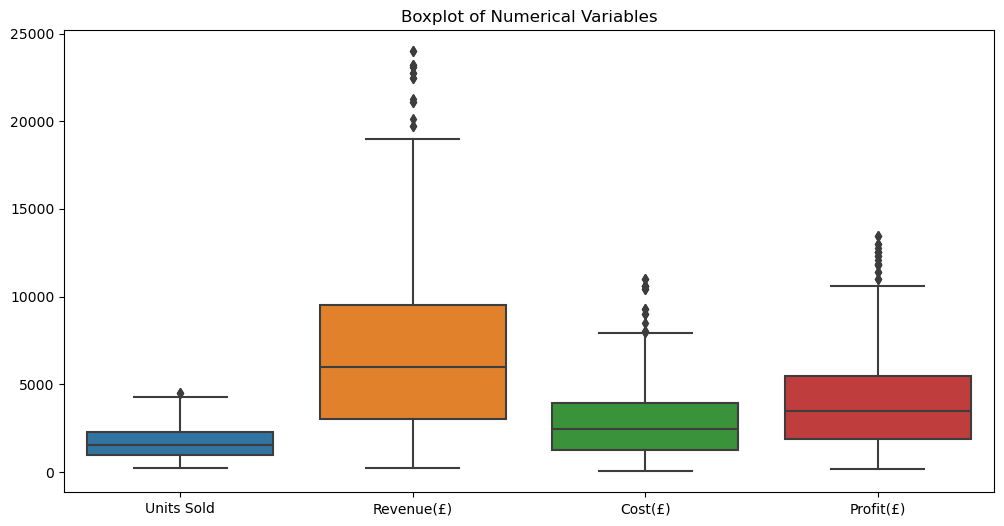


* 1. Data Preparation: Detecting outliers:

Identifying outliers is crucial in data preparation to ensure the quality and integrity of our analysis. An outlier is a data point in a dataset that differs significantly from the other data points and is therefore

worth paying attention to (Arden University, n.d.). In this step, we'll examine the distribution of numerical variables to pinpoint any data points that deviate significantly from the norm.

Chart 2.5



The boxplot (Chart 2.5.0) analysis revealed the presence of outliers in all variables. For multivariate outliers, we can check the scatterplots drawn in Chart 2.4.0. The most common statistical method to look for outliers is by using the interquartile range (IQR) and z scores (Arden University, n.d.).

The z-score of a data point describes how far the data point is from the mean, in relation to the standard deviation. In other words if a data point has a z-score of +2, that means that this data point is 2 standard deviations higher than the mean. (For example if the mean is 50 and the standard deviation is 10, then the data point of 70 would have a z-score of +2) (Arden University, n.d.).

Let’s find out outliers using Z score method:



* 1. Regression: Simple Linear Regression Analysis:

Regression is a strong statistical and visual method that can be employed to determine a relationship (Arden University, n.d.). In this section, we'll perform a simple linear regression analysis to model the relationship between the "Units Sold" and "Profit(£)" variables. Simple linear regression is a fundamental technique for predicting a continuous outcome based on a single independent variable.

Simple Linear Regression Results:

R-squared: 0.72

Intercept: -102.46

Coefficient: 2.5

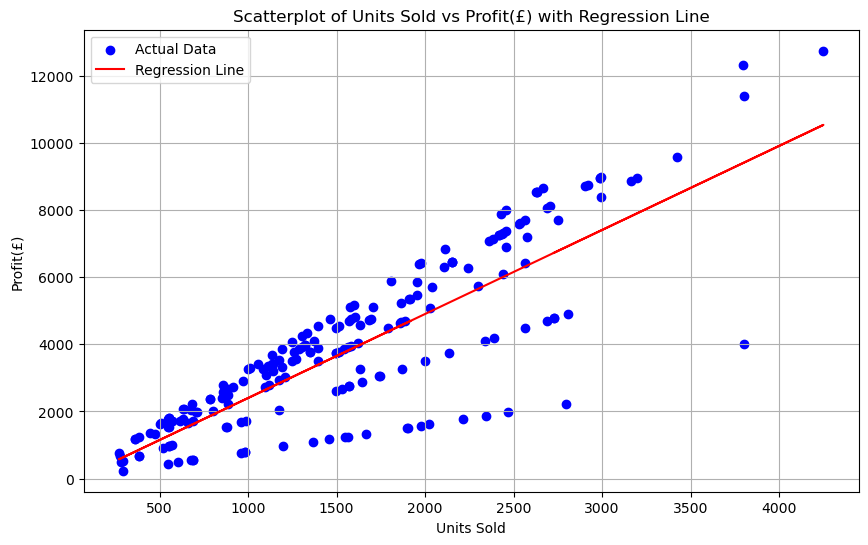
The results of the simple linear regression analysis provide insights into the relationship between "Units Sold" and "Profit(£)" for Confectionaries Are Us. With an R-squared value of approximately 0.72, we can infer that around 72% of the variation in "Profit(£)" can be explained by the linear relationship with "Units Sold." This suggests that "Units Sold" is a significant predictor of "Profit(£)," albeit other factors not included in the model may also influence profitability. (see the next section)

The coefficient of approximately 2.5 indicates that, on average, for each additional unit sold, the predicted "Profit(£)" increases by approximately £2.5. This implies that there is a positive linear relationship between the number of units sold and the profitability of *Confectionaries Are Us* products. However, it's essential to consider practical implications and potential limitations of this relationship.

The intercept of approximately -102.46 may not have a straightforward interpretation in this context, as it suggests a negative "Profit(£)" when "Units Sold" is zero. This scenario is unlikely in a business setting, where selling zero units would result in zero revenue and, therefore, zero profitability. Thus, the intercept value should be interpreted cautiously, and its practical significance may be limited.

Overall, while the simple linear regression model provides valuable insights into the relationship between "Units Sold" and "Profit(£)," it's essential to consider other variables and potential confounding factors that may influence profitability. Additionally, further analysis and validation of the model's assumptions and predictive performance are warranted to ensure its reliability and applicability in decision-making processes for *Confectionaries Are Us.*

Chart 2.6



* 1. Regression: Multiple Regression Analysis

In this section, we'll conduct a multiple regression analysis to examine the combined influence of "Units Sold," "Revenue(£)," and "Cost(£)" on the profitability of *Confectionaries Are Us*. Multiple regression allows us to assess the unique contribution of each variable to the prediction of profitability.

Multiple Regression Results:

R-squared: 0.98

Intercept: -21.01

Coefficients: [0.26, 0.66, -0.32]

**Coefficients:** The coefficients represent the change in the dependent variable ("Profit(£)") for a one-unit change in each independent variable, holding all other variables constant. The coefficients for "Units Sold," "Revenue(£)," and "Cost(£)" are approximately 0.26, 0.66, and -0.32, respectively. This implies that:

For each additional unit sold, the predicted "Profit(£)" increases by approximately £0.26.

For each additional unit of revenue generated, the predicted "Profit(£)" increases by approximately £0.66.

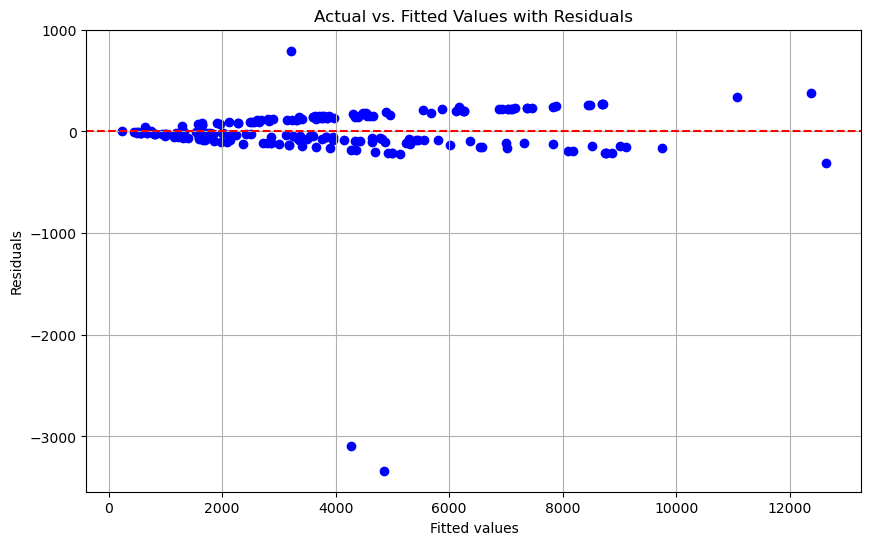
For each additional unit of cost incurred, the predicted "Profit(£)" decreases by approximately £0.32.

The multiple regression analysis reveals a highly significant relationship between "Profit(£)" and the combination of "Units Sold," "Revenue(£)," and "Cost(£)" for *Confectionaries Are Us.* The exceptionally high R-squared value of approximately 0.98 indicates that the chosen independent variables collectively account for 98% of the variability in profitability, underscoring the robustness of the model in explaining variations in "Profit(£)." This suggests that "Units Sold," "Revenue(£)," and "Cost(£)" are strong predictors of profitability for the company.

The positive coefficient for "Units Sold" implies that increasing the number of units sold leads to a corresponding increase in profitability, albeit at a relatively modest rate. Similarly, the positive coefficient for "Revenue(£)" indicates that higher revenue generation positively impacts profitability, with each additional unit of revenue contributing more significantly to profitability compared to "Units Sold." Conversely, the negative coefficient for "Cost(£)" suggests that higher costs incurred have a detrimental effect on profitability, with each additional unit of cost leading to a reduction in predicted profitability.

This information can inform strategic decision-making processes aimed at improving the company's financial performance and sustaining long-term growth and success.

Chart 2.7



* 1. Correlation Analysis:

In this section, we'll conduct correlation analysis to explore the relationships between variables for Confectionaries Are Us. A correlation matrix will be generated to quantify the strength and direction of linear relationships between pairs of variables. It's essential to note that correlation does not imply causation (Arden University, n.d.).

Correlation Matrix:

Units Sold Revenue(£) Cost(£) Profit(£)

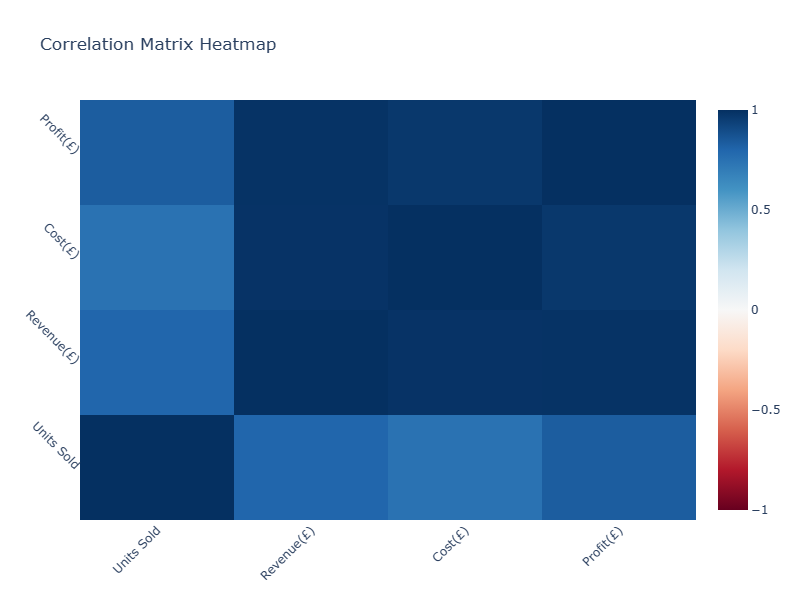
Units Sold 1.000000 0.799822 0.747128 0.834562

Revenue(£) 0.799822 1.000000 0.987456 0.990139

Cost(£) 0.747128 0.987456 1.000000 0.969486

Profit(£) 0.834562 0.990139 0.969486 1.000000

Chart 2.8



Units Sold and Revenue(£): The correlation coefficient between "Units Sold" and "Revenue(£)" is approximately 0.80, indicating a strong positive correlation. This suggests that there is a significant relationship between the number of units sold and the revenue generated by *Confectionaries Are Us*. As the number of units sold increases, the revenue tends to increase as well. This correlation is expected, as higher sales volumes typically lead to higher revenues for a company.

Units Sold and Cost(£): The correlation coefficient between "Units Sold" and "Cost(£)" is approximately 0.75, indicating a strong positive correlation. This implies that there is a notable relationship between the number of units sold and the cost incurred by the company. It suggests that as the number of units sold increases, the associated costs, such as production or operational expenses, also tend to increase. This correlation highlights the importance of managing costs effectively to maximize profitability.

Units Sold and Profit(£): The correlation coefficient between "Units Sold" and "Profit(£)" is approximately 0.83, indicating a strong positive correlation. This suggests that there is a significant relationship between the number of units sold and the profitability of *Confectionaries Are Us*. As the number of units sold increases, the company's profit tends to increase as well. This correlation underscores the fundamental principle that higher sales volumes contribute to higher profits for the company.

Revenue(£) and Cost(£): The correlation coefficient between "Revenue(£)" and "Cost(£)" is approximately 0.99, indicating a very strong positive correlation. This implies that there is an extremely tight relationship between the revenue generated by the company and the associated costs. It suggests that as revenue increases, the costs also tend to increase proportionally. This correlation emphasizes the importance of cost management to ensure that revenue growth translates into improved profitability.

Revenue(£) and Profit(£): The correlation coefficient between "Revenue(£)" and "Profit(£)" is approximately 0.99, indicating a very strong positive correlation. This implies that there is an exceptionally close relationship between the revenue generated by the company and its profitability. It suggests that as revenue increases, the company's profit also tends to increase nearly in tandem. This correlation highlights the critical role that revenue growth plays in driving overall profitability for *Confectionaries Are Us*.

The correlation coefficient between "Cost(£)" and "Profit(£)" is approximately 0.97, indicating a very strong positive correlation. However, in the multiple regression analysis, if "Cost(£)" had a negative coefficient while predicting "Profit(£)," it suggests that while cost and profit are positively correlated, the relationship is not as straightforward when considering other variables.

In the regression context, a negative coefficient for "Cost(£)" might imply that while increasing costs could lead to higher revenues (as indicated by the positive correlation between "Revenue(£)" and "Cost(£)"), it might also lead to diminishing returns in terms of profit, thus showing a negative impact on profitability.

* 1. Regional Preferences for Confectionaries: Analysis of Sales Data across UK Regions:

Understanding regional preferences for different confectionaries is crucial for Confectionaries Are Us to tailor their marketing and production strategies effectively. By analyzing sales data across various regions in the UK, we aim to identify which confectionaries are the most popular in each region. This analysis will provide valuable insights into consumer preferences, allowing the company to optimize its product offerings and marketing efforts to meet the unique demands of each market segment. We will examine sales data for six confectionary types: Biscuit, Biscuit Nut, Plain, Caramel Nut, Chocolate Chunk, and Caramel, across regions such as England, Scotland, Wales, Northern Ireland, and Jersey. Let's delve into the data to uncover regional trends in confectionary sales.

Confectionary with the highest sales in each region:

Country(UK) Confectionary Units Sold

2 England Caramel 185144

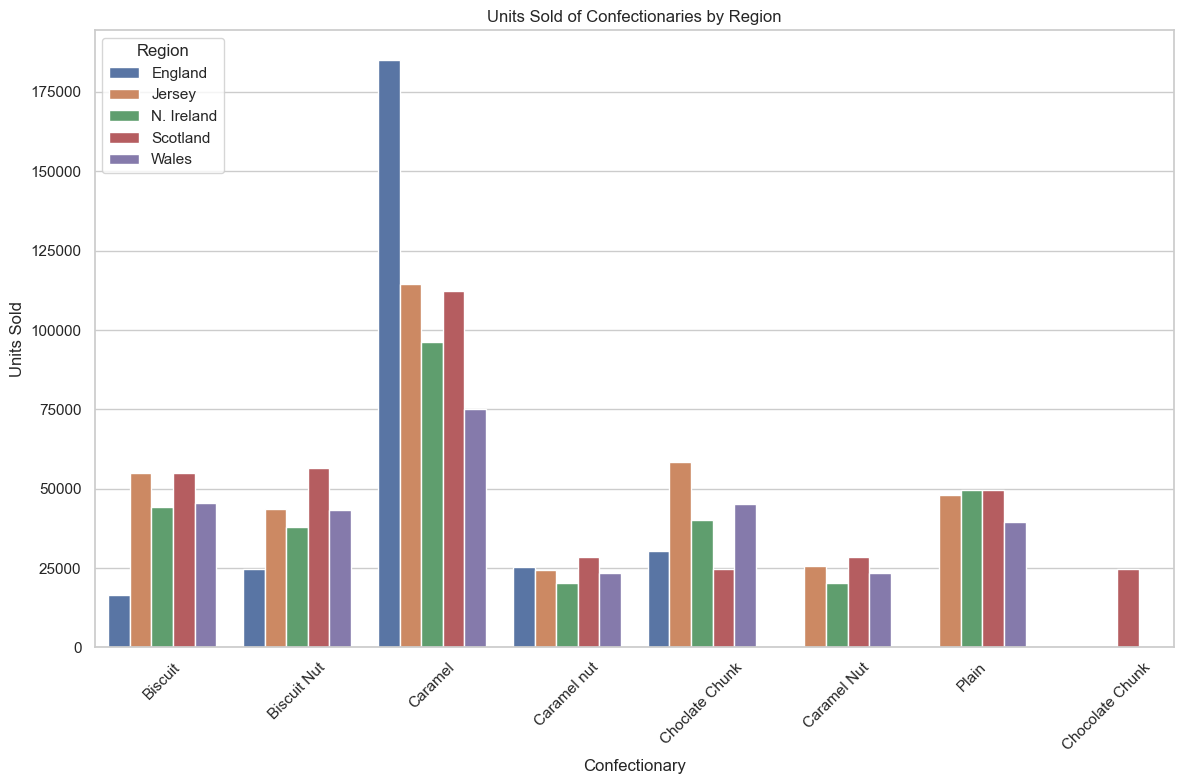
7 Jersey Caramel 114498

14 N. Ireland Caramel 96101

21 Scotland Caramel 112194

29 Wales Caramel 74976

Chart 2.9



England: In England, the confectionary with the highest sales is "Caramel," with a total of 185,144 units sold. This indicates a strong preference for caramel-flavored confectionaries among consumers in England. The popularity of caramel-based products may be attributed to various factors such as taste preferences, marketing strategies, and cultural influences.

Jersey: Similarly, in Jersey, "Caramel" emerges as the top-selling confectionary, with a total of 114,498 units sold. Despite being a smaller region compared to England, Jersey demonstrates a significant demand for caramel-flavored treats.

Northern Ireland: In Northern Ireland, "Caramel" also dominates as the confectionary with the highest sales, with a total of 96,101 units sold.

Scotland: In Scotland, "Caramel" emerges once again as the top-selling confectionary, with a total of 112,194 units sold.

Wales: Lastly, similar to other regions, in Wales, "Caramel" emerges as the confectionary with the highest sales, with a total of 74,976 units sold.

For Confectionaries Are Us, the analysis of regional preferences for confectionaries across various regions in the UK provides valuable insights that can inform strategic decision-making and drive business growth. The consistent dominance of "Caramel" as the top-selling confectionary in England, Jersey, Northern Ireland, Scotland, and Wales underscores its widespread popularity among consumers, highlighting the universal appeal of certain flavors in the confectionary market.

* 1. Enhanced Data Mining Analysis: Advanced Techniques and Insights

In this phase, we'll apply advanced data mining techniques—including clustering, decision trees, and association analysis—to delve deeper into the sales data of *Confectionaries Are Us*. These methods will provide additional insights and support more sophisticated analysis for strategic decision-making.

* + 1. K-means specifically attempts to put the data into the number of clusters the analysist tells it to, in an unsupervised (Arden University, n.d.). In the context of *Confectionaries Are Us*, we'll employ K-means clustering to segment sales data into clusters characterized by similar patterns of 'Units Sold', 'Revenue(£)', 'Cost(£)', and 'Profit(£)'. This technique enables us to uncover underlying structures within the data and identify meaningful insights for strategic decision-making.

Units Sold Revenue(£) Cost(£) Profit(£) Cluster

0 1118.0 5590.0 2459.6 3130.4 0

1 708.0 3540.0 1557.6 1982.4 1

2 1269.0 6345.0 2791.8 3553.2 0

3 1631.0 8155.0 3588.2 4566.8 0

4 2240.0 11200.0 4928.0 6272.0 0

In the clustering results, each row represents a sales transaction, with columns indicating the 'Units Sold', 'Revenue(£)', 'Cost(£)', 'Profit(£)', and the assigned cluster label. For instance, the first row shows that 1118 units were sold, generating £5590 in revenue, with a cost of £2459.6 and a profit of £3130.4, belonging to Cluster 0. Similarly, the second row represents a different sales transaction with 708 units sold, resulting in £3540 revenue, £1557.6 cost, and £1982.4 profit, assigned to Cluster 1. Clustering allows us to group similar sales transactions together, enabling us to identify distinct patterns or segments within the sales data. These clusters can provide valuable insights into customer behavior, sales trends, and profitability, guiding strategic decision-making processes for *Confectionaries Are Us.*

* + 1. Decision Tree Analysis for Sales Data Segmentation: In this section, we'll utilize decision tree analysis to further dissect the sales data of Confectionaries Are Us. Decision trees are powerful tools for classification and regression tasks, offering transparent insights into the factors influencing outcomes. By constructing a decision tree based on features such as 'Units Sold', 'Revenue(£)', 'Cost(£)', and 'Profit(£)', we aim to uncover key decision pathways. When a decision tree predicts numerical values, it is known as a Regression Tree (Arden University, n.d.).

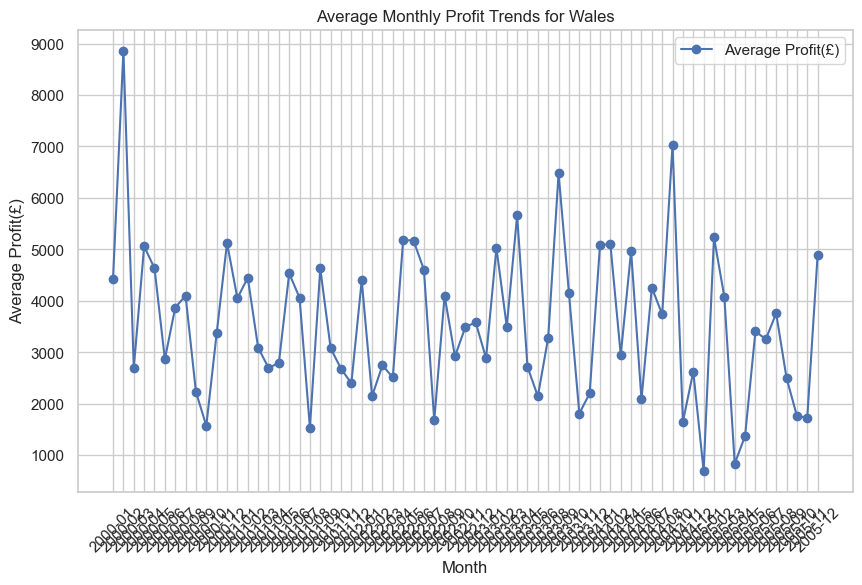
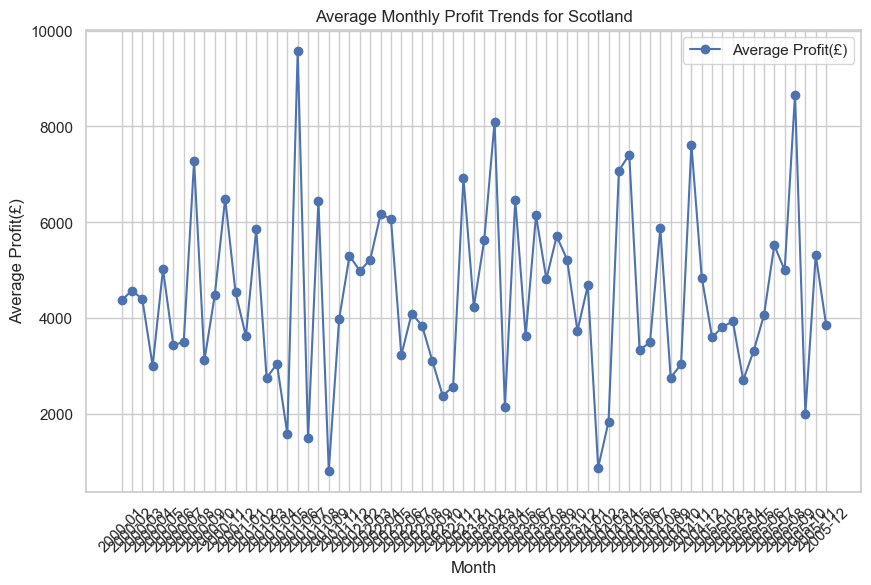
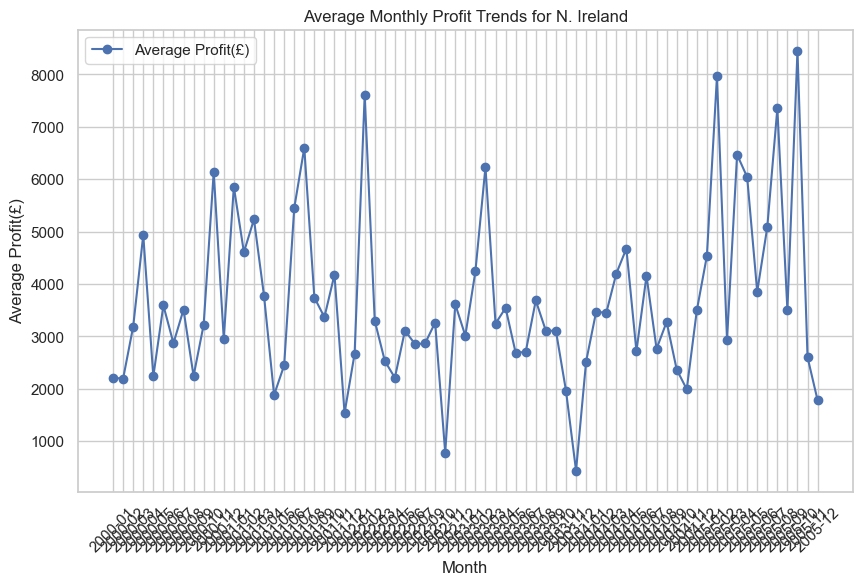
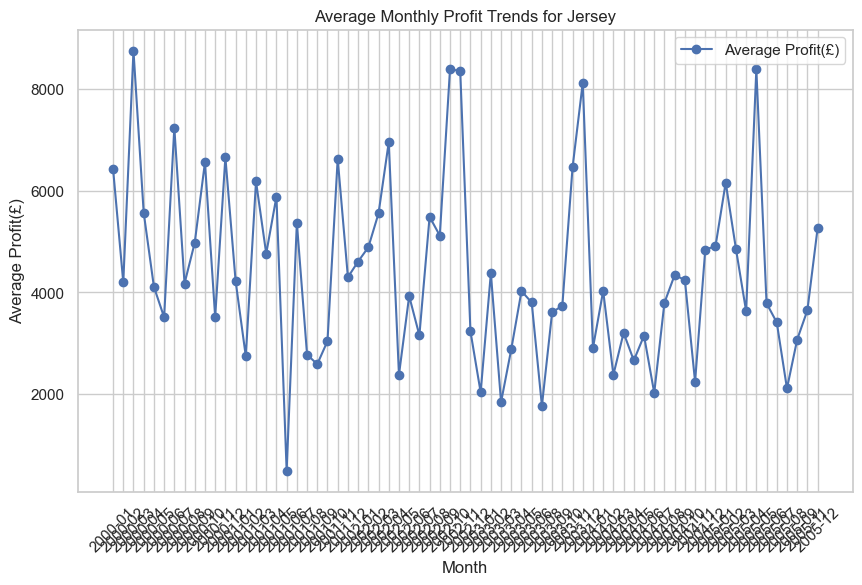
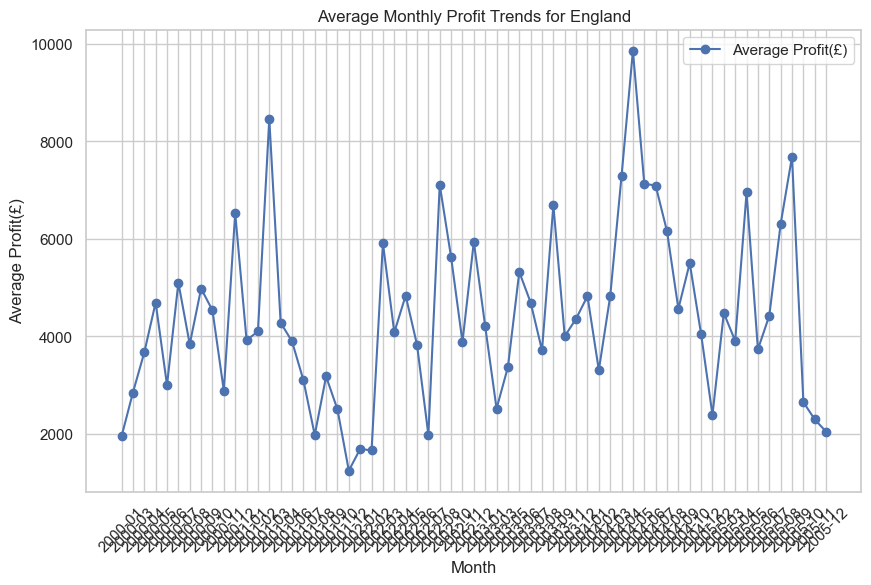
Accuracy of Decision Tree Classifier: 1.00

An accuracy of 1.00 indicates that the decision tree classifier achieved perfect accuracy on the testing data, meaning it correctly classified all instances into their respective clusters. While achieving perfect accuracy may seem promising, it's essential to ensure the model's performance is not overfitting to the training data. Nonetheless, the high accuracy suggests that the decision tree algorithm effectively segmented the sales data based on the provided features, contributing valuable insights for *Confectionaries Are Us.*

* 1. Comprehensive Analysis of Profitability Trends and Forecasting:

In this comprehensive analysis, we examine profitability trends over time for each country and utilize ARIMA modeling to forecast future profits, aiding strategic decision-making.

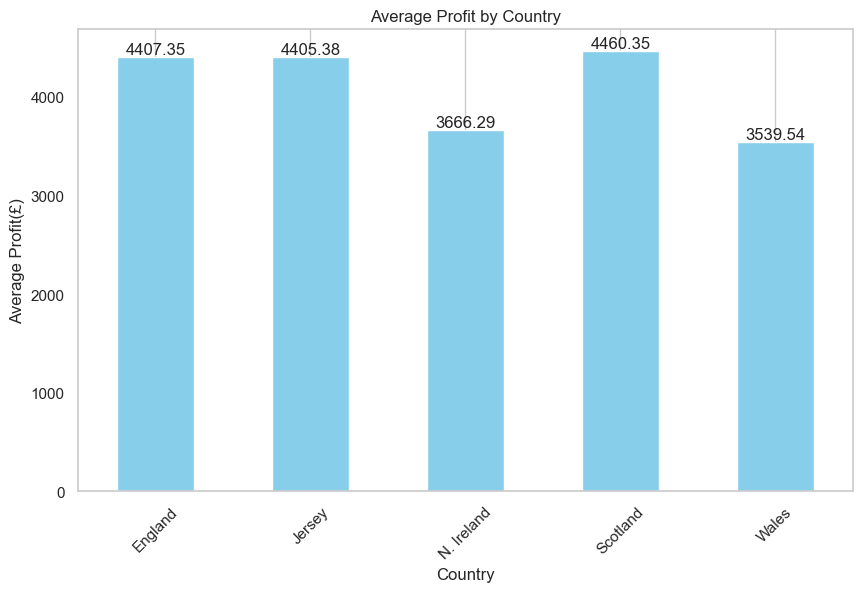
Charts 2.11.1



These charts average monthly profit trends for each country, aiming to identify profitability patterns over time.

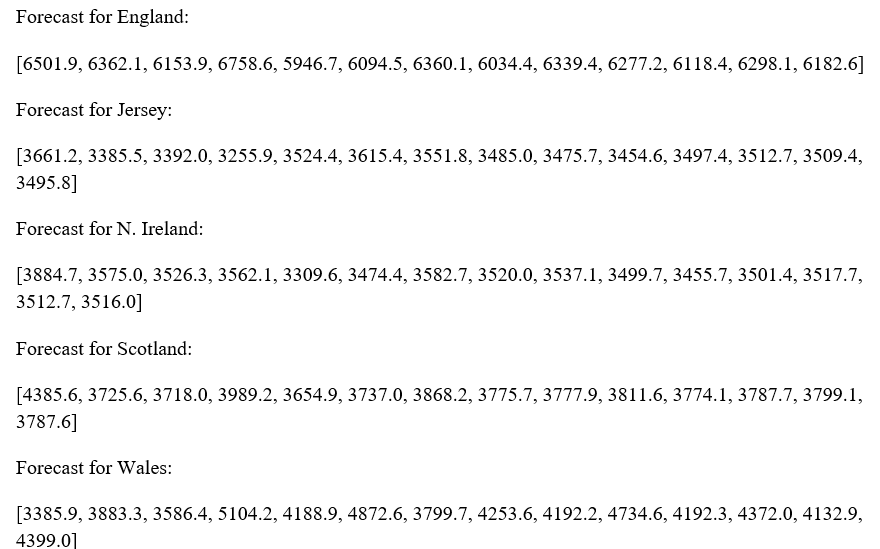
We visualize the average profits of each country to identify the country with the highest profitability over the entire period.

Chart 2.11.2



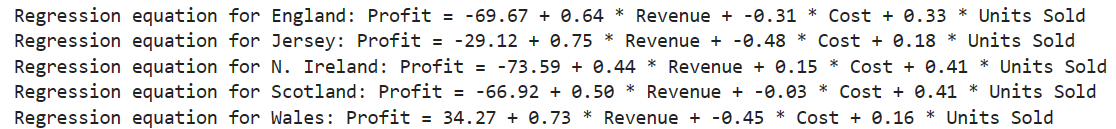
The average profit across countries varies, with Scotland having the highest average profit, followed closely by England and Jersey.

We employ ARIMA (AutoRegressive Integrated Moving Average) modeling to forecast future profit trends for each country, aiding strategic planning and decision-making. ARIMA (AutoRegressive Integrated Moving Average) is a popular time series forecasting method that models the relationship between a series of observations and its lagged values, as well as the difference between consecutive observations. It combines autoregression (AR), differencing (I), and moving average (MA) components to capture various patterns and trends in the data (Hyndman, R. J., & Athanasopoulos, G., 2018).



The forecast results represent predicted future profit values for each country using ARIMA modeling. Each forecast provides a sequence of predicted profit values for successive time periods, rounded off to 1 decimal unit. These forecasts offer insights into expected profitability trends, aiding decision-making and strategic planning for *Confectionaries Are Us* across different regions.

* 1. Generate regression equations predicting profit based on other variables for each country:



In case of England, strategists should focus more on Revenue generation through units sold, as the cost affect negatively the Profit. For Jersey, the situation is almost same as England. For N. Ireland, either of the three variables may be focused on to increase Profit as per the predictive measures. For Scotland and Wales the case is almost the same as England as well.

1. **Task 2**

This report presents the findings and insights derived from the data mining investigation conducted for *Confectionaries Are Us*. Through the application of various data mining techniques, including regression analysis, clustering, and time series forecasting, we aimed to uncover patterns and trends in the company's sales data to inform strategic decision-making.

The data mining investigation conducted for *Confectionaries Are Us* represents a critical role in understanding and optimizing the company's sales dynamics. By the usage of advanced analytical techniques, we aimed to extract applicable insights from the wealth of sales data available to the company. Our approach encompassed a multifaceted analysis, starting with exploratory data analysis (EDA) to uncover underlying patterns and relationships.

During the exploratory data analysis phase, we delved into the key variables driving sales performance, including revenue, cost, units sold, and profit. Visualizations such as scatter plots and pair plots facilitated the identification of trends, outliers, and potential areas for further investigation. This initial phase set the stage for more sophisticated analyses aimed at uncovering deeper insights.

Regression analysis emerged as a powerful tool for modeling the complex relationships between sales variables and profitability. We developed regression equations tailored to each country, clearly explaining the primary drivers of profit and quantifying their impact on the company's bottom line. These regression models serve as valuable predictive tools, enabling the company to forecast profit outcomes based on changes in revenue, cost, and units sold.

Going on with Regression analysis, we used correlation analysis to further look into the relationships between the variables that we earlier saw in scatterplots plotted inside the pairplots

Additionally, clustering analysis provided further insights by segmenting countries based on shared sales characteristics. This segmentation approach allows the company to tailor its marketing strategies and operational tactics to the unique needs of each market segment, thereby maximizing sales potential and optimizing resource allocation.

Furthermore, time series forecasting techniques, particularly ARIMA models, facilitated the generation of future profit forecasts for each country. These forecasts enable the company to anticipate sales trends, identify emerging opportunities, and proactively adjust its strategies to capitalize on market dynamics.

For our visualizations, we carefully selected graphs and plots that effectively communicated complex relationships and patterns in the data. Scatter plots, pair plots, and box plots were chosen for their ability to reveal correlations, distributions, and outliers intuitively. We ensured color schemes were chosen thoughtfully, prioritizing clarity and accessibility. Bright, contrasting colors were used for emphasis, while muted tones helped differentiate categories without overwhelming the viewer.

Moving forward, we recommend strategists make benefit of the insights gleaned from our analysis to inform targeted marketing strategies and product development initiatives. Additionally, proactive forecasting of sales trends using ARIMA models enables strategists to anticipate market shifts and capitalize on emerging opportunities. Continued investment in data-driven decision-making and agile adaptation to market dynamics will be crucial for sustained growth and competitiveness.

In conclusion, the data mining investigation has yielded actionable insights that can drive strategic decision-making and foster sustained growth for *Confectionaries Are Us*. By harnessing the power of advanced analytics, the company is well-positioned to optimize its sales performance, enhance profitability, and maintain a competitive edge in the market. Continued analysis and adaptation will be essential to navigate evolving market conditions and sustain long-term success.

Appendix

**2.2**

import pandas as pd

# Set the file path

file\_path = r"C:\Users\PMLS\Desktop\Projects\Python Projects\ا\data-set-assignment.xlsx" # Change the directory accordingly

# Load the dataset into a pandas DataFrame

data = pd.read\_excel(file\_path)

# Display summary of the dataset

print(data.info())

print(data.describe())

**2.3 & 2.4**

# Import necessary libraries

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Set the file path

file\_path = r"C:\Users\PMLS\Desktop\Projects\Python Projects\ا\data-set-assignment.xlsx" # Change the directory accordingly

# Load the dataset into a pandas DataFrame

data = pd.read\_excel(file\_path)

# Display summary of the dataset

print(data.info())

print(data.describe())

# Visualize missing values

plt.figure(figsize=(10, 6))

sns.heatmap(data.isnull(), cmap='viridis', cbar=False)

plt.title('Missing Values in the Dataset')

plt.show()

# Visualize distribution of numerical variables

plt.figure(figsize=(12, 8))

sns.pairplot(data.dropna(), diag\_kind='kde')

plt.suptitle('Pairplot of Numerical Variables', y=1.02)

plt.show()

# Check for missing values in the dataset

missing\_values = data.isnull().sum()

# Print the result

print("Missing values in the dataset:")

print(missing\_values)

# Imputation: Fill missing values with mean

data['Units Sold'].fillna(data['Units Sold'].mean(), inplace=True)

data['Revenue(£)'].fillna(data['Revenue(£)'].mean(), inplace=True)

data['Cost(£)'].fillna(data['Cost(£)'].mean(), inplace=True)

data['Profit(£)'].fillna(data['Profit(£)'].mean(), inplace=True)

# Exclusion: Drop rows with missing values

data.dropna(inplace=True)

# Check for missing values in the dataset

missing\_values = data.isnull().sum()

# Print the result

print("Missing values in the dataset:")

print(missing\_values)

**2.5.**

# Handle missing values

data.dropna(inplace=True)

# Check for duplicate rows

duplicate\_rows = data[data.duplicated()]

if not duplicate\_rows.empty:

print("Duplicate rows found. Removing duplicates...")

data.drop\_duplicates(inplace=True)

else:

print("No duplicate rows found.")

# Check for outliers

plt.figure(figsize=(12, 6))

sns.boxplot(data=data[['Units Sold', 'Revenue(£)', 'Cost(£)', 'Profit(£)']])

plt.title('Boxplot of Numerical Variables')

plt.show()

from scipy.stats import zscore

# Calculate z-scores for numerical variables

z\_scores = zscore(data[['Units Sold', 'Revenue(£)', 'Cost(£)', 'Profit(£)']])

# Define threshold for outlier detection (e.g., z-score > 3)

threshold = 3

# Identify outliers

outliers = data[(z\_scores > threshold).any(axis=1)]

# Print the outliers

print("Outliers identified using z-score method:")

print(outliers)

# Print outliers as a formatted table using Pandas

print("Outliers identified using z-score method:")

print(outliers.to\_string(index=False))

**2.6.**

# Importing necessary libraries

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error, r2\_score

# Splitting the data into training and testing sets

X = data[['Units Sold']]

y = data['Profit(£)']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Fitting the linear regression model

model = LinearRegression()

model.fit(X\_train, y\_train)

# Predicting on the test set

y\_pred = model.predict(X\_test)

# Evaluating the model

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

# Printing the results rounded to 2 decimal places

print("Simple Linear Regression Results:")

print("Mean Squared Error:", round(mse, 2))

print("R-squared:", round(r2, 2))

print("Intercept:", round(model.intercept\_, 2))

print("Coefficient:", round(model.coef\_[0], 2))

import statsmodels.api as sm

X = sm.add\_constant(data[['Units Sold']])

y = data['Profit(£)']

# Fit the multiple linear regression model

model = sm.OLS(y, X).fit()

# Get the p-values

p\_values = model.pvalues

# Print the p-values

print("P-values for Multiple Linear Regression:")

print(p\_values)

# Plotting the scatterplot and regression line

plt.figure(figsize=(10, 6))

plt.scatter(X\_test, y\_test, color='blue', label='Actual Data')

plt.plot(X\_test, model.predict(X\_test), color='red', label='Regression Line')

plt.xlabel('Units Sold')

plt.ylabel('Profit(£)')

plt.title('Scatterplot of Units Sold vs Profit(£) with Regression Line')

plt.legend()

plt.grid(True)

plt.show()

**2.7.**

# Splitting the data into training and testing sets

X = data[['Units Sold', 'Revenue(£)', 'Cost(£)']]

y = data['Profit(£)']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Fitting the multiple regression model

model = LinearRegression()

model.fit(X\_train, y\_train)

# Predicting on the test set

y\_pred = model.predict(X\_test)

# Evaluating the model

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

# Printing the results

print("Multiple Regression Results:")

print("R-squared:", round(r2, 2))

print("Intercept:", round(model.intercept\_, 2))

print("Coefficients:", [round(coef, 2) for coef in model.coef\_])

import statsmodels.api as sm

X = sm.add\_constant(data[['Units Sold', 'Revenue(£)', 'Cost(£)']])

y = data['Profit(£)']

# Fit the multiple linear regression model

model = sm.OLS(y, X).fit()

# Get the p-values

p\_values = model.pvalues

# Print the p-values

print("P-values for Multiple Linear Regression:")

print(p\_values)

# Calculating residuals

residuals = y\_test - y\_pred

# Plotting actual vs. fitted values with residuals

plt.figure(figsize=(10, 6))

plt.scatter(y\_pred, residuals, color='blue')

plt.axhline(y=0, color='red', linestyle='--')

plt.xlabel('Fitted values')

plt.ylabel('Residuals')

plt.title('Actual vs. Fitted Values with Residuals')

plt.grid(True)

plt.show()

**2.8.**

# Calculating correlation matrix

correlation\_matrix = data[['Units Sold', 'Revenue(£)', 'Cost(£)', 'Profit(£)']].corr()

# Displaying correlation matrix

print("Correlation Matrix:")

print(correlation\_matrix)

import plotly.graph\_objects as go

# Create heatmap figure

fig = go.Figure(data=go.Heatmap(

z=correlation\_matrix.values,

x=correlation\_matrix.columns,

y=correlation\_matrix.index,

colorscale='RdBu',

zmin=-1, zmax=1,

hoverongaps = False))

# Customize layout

fig.update\_layout(

title='Correlation Matrix Heatmap',

xaxis=dict(tickangle=-45),

yaxis=dict(tickangle=45),

width=800,

height=600

)

# Show plot

fig.show()

**2.9.**

# Group the data by Country(UK) and Confectionary, and calculate total units sold

sales\_by\_country\_confectionary = data.groupby(['Country(UK)', 'Confectionary'])['Units Sold'].sum().reset\_index()

# Find the confectionary with the highest sales in each region

highest\_sales\_by\_country = sales\_by\_country\_confectionary.loc[sales\_by\_country\_confectionary.groupby('Country(UK)')['Units Sold'].idxmax()]

# Print the results

print("Confectionary with the highest sales in each region:")

print(highest\_sales\_by\_country)

import seaborn as sns

# Set the style of seaborn

sns.set(style="whitegrid")

# Plot clustered bar chart

plt.figure(figsize=(12, 8))

sns.barplot(data=sales\_by\_country\_confectionary, x='Confectionary', y='Units Sold', hue='Country(UK)')

plt.title('Units Sold of Confectionaries by Region')

plt.xlabel('Confectionary')

plt.ylabel('Units Sold')

plt.xticks(rotation=45)

plt.legend(title='Region')

plt.tight\_layout()

plt.show()

**2.10.**

# Import necessary libraries

from sklearn.cluster import KMeans

# Prepare data for KNN clustering

X = data[['Units Sold', 'Revenue(£)', 'Cost(£)', 'Profit(£)']]

# Perform KNN clustering

kmeans = KMeans(n\_clusters=3)

data['Cluster'] = kmeans.fit\_predict(X)

# Print the clustering results

print(data[['Units Sold', 'Revenue(£)', 'Cost(£)', 'Profit(£)', 'Cluster']].head())

# Import necessary libraries

from sklearn.tree import DecisionTreeClassifier

from sklearn.model\_selection import train\_test\_split

# Prepare data for decision tree analysis

X = data[['Units Sold', 'Revenue(£)', 'Cost(£)', 'Profit(£)']]

y = data['Cluster'] # Assuming 'Cluster' column represents the target variable

# Split data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Initialize and train the decision tree classifier

clf = DecisionTreeClassifier()

clf.fit(X\_train, y\_train)

# Evaluate the trained model

accuracy = clf.score(X\_test, y\_test)

print("Accuracy of Decision Tree Classifier: {:.2f}".format(accuracy))

**2.11.**

import pandas as pd

import matplotlib.pyplot as plt

# Convert Date column to datetime format

data['Date'] = pd.to\_datetime(data['Date'])

# Group data by country and month, and calculate average profit

monthly\_profit = data.groupby([data['Country(UK)'], data['Date'].dt.to\_period('M')])['Profit(£)'].mean().reset\_index()

# Separate data for each country

countries = monthly\_profit['Country(UK)'].unique()

country\_data = {country: monthly\_profit[monthly\_profit['Country(UK)'] == country] for country in countries}

# Visualize average profit trends for each country over monthly intervals

for country, country\_df in country\_data.items():

plt.figure(figsize=(10, 6))

plt.plot(country\_df['Date'].astype(str), country\_df['Profit(£)'], marker='o', label='Average Profit(£)')

plt.title(f'Average Monthly Profit Trends for {country}')

plt.xlabel('Month')

plt.ylabel('Average Profit(£)')

plt.xticks(rotation=45)

plt.legend()

plt.grid(True)

plt.show()

import matplotlib.pyplot as plt

# Calculate average profit for each country

average\_profit\_by\_country = monthly\_profit.groupby('Country(UK)')['Profit(£)'].mean()

# Plot bar chart

plt.figure(figsize=(10, 6))

average\_profit\_by\_country.plot(kind='bar', color='skyblue')

plt.title('Average Profit by Country')

plt.xlabel('Country')

plt.ylabel('Average Profit(£)')

plt.xticks(rotation=45)

plt.grid(axis='y')

# Add annotations

for i, value in enumerate(average\_profit\_by\_country):

plt.text(i, value, f'{value:.2f}', ha='center', va='bottom')

plt.show()

print("Average Profit by Country:")

print(average\_profit\_by\_country)

**2.11.**

from statsmodels.tsa.arima.model import ARIMA

import numpy as np

# Function to fit ARIMA model and make forecast

def fit\_arima(country\_data):

# Split data into train and test sets (80% train, 20% test)

train\_size = int(len(country\_data) \* 0.8)

train, test = country\_data[:train\_size], country\_data[train\_size:]

# Fit ARIMA model

model = ARIMA(train, order=(5,1,0))

model\_fit = model.fit()

# Make forecast

forecast = model\_fit.forecast(steps=len(test))

return forecast

# Perform ARIMA modeling for each country

forecast\_results = {}

for country, country\_df in country\_data.items():

# Remove missing values and convert to numpy array

country\_df = country\_df.dropna()

country\_data\_np = np.array(country\_df['Profit(£)'])

# Fit ARIMA model and make forecast

forecast = fit\_arima(country\_data\_np)

forecast\_results[country] = forecast

# Print forecast results rounded off to 1 decimal unit

for country, forecast in forecast\_results.items():

print(f"Forecast for {country}:")

rounded\_forecast = [round(val, 1) for val in forecast]

print(rounded\_forecast)

print()

**2.12.**

from sklearn.linear\_model import LinearRegression

# Iterate over countries

for country in countries:

# Select data for the country

country\_data = data[data['Country(UK)'] == country]

# Prepare independent variables (revenue, cost, units sold)

X = country\_data[['Revenue(£)', 'Cost(£)', 'Units Sold']]

y = country\_data['Profit(£)']

# Fit linear regression model

model = LinearRegression().fit(X, y)

# Print regression equation

print(f"Regression equation for {country}: Profit = {model.intercept\_:.2f} + "

f"{model.coef\_[0]:.2f} \* Revenue + {model.coef\_[1]:.2f} \* Cost + "

f"{model.coef\_[2]:.2f} \* Units Sold")

# References

Arden University, n.d. *COM6013 | LESSON 5 | Classification.* s.l.:s.n.

Arden University, n.d. *COM6013 | LESSON 6 | Clustering.* s.l.:s.n.

Arden University, n.d. *COM6013 | LESSON 7 | Regression.* s.l.:s.n.

Arden University, n.d. *COM6013 | LESSON 8 | Outlier detection.* s.l.:s.n.

Arden University, n.d. *COM6013 Data Mining Lesson 4, Data Preparation.* s.l.:s.n.

Hyndman, R. J., & Athanasopoulos, G., 2018. *Forecasting: principles and practice.* s.l.:OTexts.