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



# Abstract

*This study encapsulates the essence of a comprehensive data mining investigation conducted on confectionary sales data. The analysis delved into the intricate patterns and relationships within the dataset, aiming to extract valuable insights to inform strategic decision-making. Through pre-processing, descriptive statistics, correlation analysis, regression modelling, outlier detection, and K-means clustering, the investigation unearthed key findings regarding performance variability across states, top-performing confectionaries, correlation and regression contradictions, and cluster analysis. These insights provide a roadmap for optimizing profitability, enhancing market competitiveness, and driving sustainable growth in the confectionary industry.*

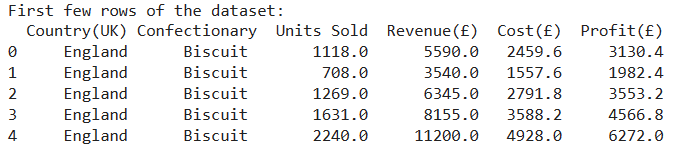
TASK 1:

You have recently joined Confectionaries Are Us as a junior analyst. As your first data mining investigation, your line manager has asked you to build and complete appropriate models so that the site based company can use your optimum, relative investigation for targeted increase profitability over time, both for individual countries and collectively, based upon differing variables. Before the analysis can be conducted, ensure that all pre-processing has been undertaken on the data set and the data set is fully understood through initial descriptive statistics. For the main body of the analysis, you are to choose the most appropriate time modelling techniques and processes for the investigation, based upon the data. In doing so, this will allow insight into the overall analysis and appropriate meaningful solutions. The full investigation is to be written in Python. A clear, concise analysis is to be given within the task, complimented with screenshot evidence of all processes and results. The full code for Task 1 is to be included within an appendix, so that it can be checked and verified it is working correctly.

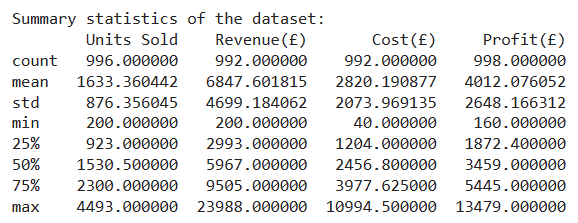
* 1. Data Processing:

Data processing is the essential first step in any data analysis endeavor. It involves transforming raw data into a usable format, ensuring accuracy and reliability for subsequent analyses. In this project, we will meticulously process the dataset provided by Confectionaries Are Us to prepare it for in-depth analysis. Through exploratory data analysis, cleaning, and transformation, we aim to uncover meaningful insights to enhance profitability. This process will enable us to identify anomalies, handle missing values, and structure the data effectively. By executing a robust data processing pipeline, we will pave the way for actionable recommendations to drive targeted profitability improvements.

* + 1. Data Summary:



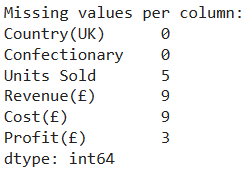
The dataset provides a snapshot of sales data for confectionary products across different countries, with a focus on the United Kingdom. It includes information on units sold, revenue, cost, and profit for each product sold.



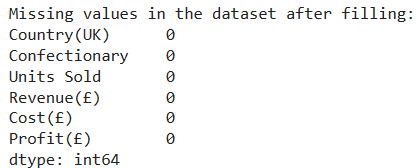
The summary statistics provide valuable insights into the distribution and characteristics of the dataset. The data reveals variability across key metrics. On average, approximately 1633 units of confectionary products were sold per transaction, generating an average revenue of £6847.60. However, the costs and profits exhibit considerable dispersion, with standard deviations of £2073.97 and £2648.17, respectively. The range of units sold spans from a minimum of 200 to a maximum of 4493, indicating diverse sales volumes. Similarly, the revenue, cost, and profit figures demonstrate wide-ranging values, reflecting the variability in financial performance across transactions. These statistics provide a comprehensive understanding of the dataset's distribution, central tendencies, and variability, laying the groundwork for further analysis and decision-making in optimizing profitability.

* + 1. Checking for Missing values:

Within any data analysis environment missing data is a very real issue. In this step, we'll utilize the isnull function to detect any missing values within the dataset (Arden University, n.d.). By pinpointing which columns contain missing values and their respective counts, we ensure robust data preprocessing.

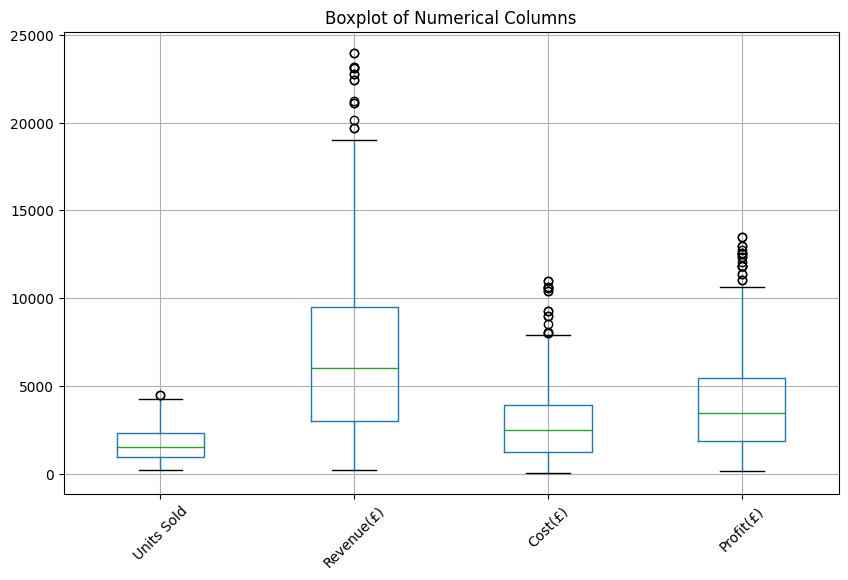


The missing values analysis reveals that the dataset contains missing entries in several columns, namely 'Units Sold', 'Revenue(£)', 'Cost(£)', and 'Profit(£)'. These missing values could potentially impact the accuracy of our analysis. To address this, we'll utilize the fillna function (Arden University, n.d.) with the imputation method. Mean imputation is a common method for handling missing data by replacing them with the mean value of the respective column. In this step, missing values in the dataset are filled with the average value of their corresponding columns. This technique preserves the overall distribution of the data while providing a simple yet effective solution for missing value imputation. By imputing missing values with column means, we ensure that the dataset remains representative of the original data, facilitating subsequent analysis and modeling with minimal disruption to the dataset's characteristics. (the code is in appendix)



* 1. Outlier Detection:

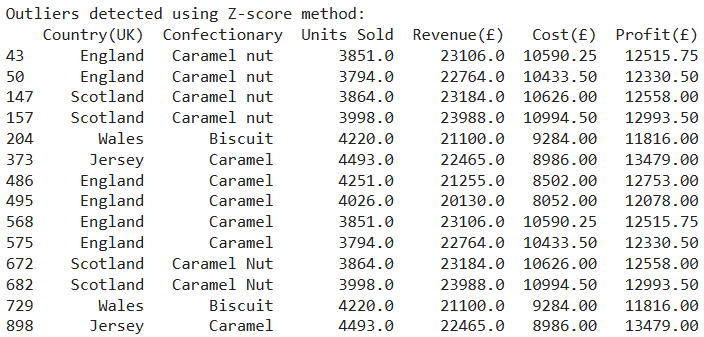
An outlier is a data point in a dataset that differs significantly from the other data points and is therefore worth paying attention to (Arden University, n.d.). Identifying and handling outliers is essential for ensuring the accuracy and reliability of statistical analyses and machine learning models. In this step, we will employ visualization techniques such as boxplots to detect outliers in the dataset. By visually inspecting the distribution of data and identifying potential outliers, we can assess their impact on the analysis and decide on appropriate strategies for handling them.



The boxplot visualizations reveal outliers at higher values across all four variables: Units Sold, Revenue(£), Cost(£), and Profit(£). These outliers represent extreme data points that significantly deviate from the majority of observations. In the context of sales and financial metrics, such outliers may indicate exceptional transactions or anomalies in data collection. Their presence suggests the potential for significant variability in sales performance and profitability, warranting further investigation. Understanding the nature and impact of these outliers is crucial for making informed decisions in optimizing sales strategies and financial management to mitigate risks and maximize profitability.

Outlier Detection Using Z-Score Method:

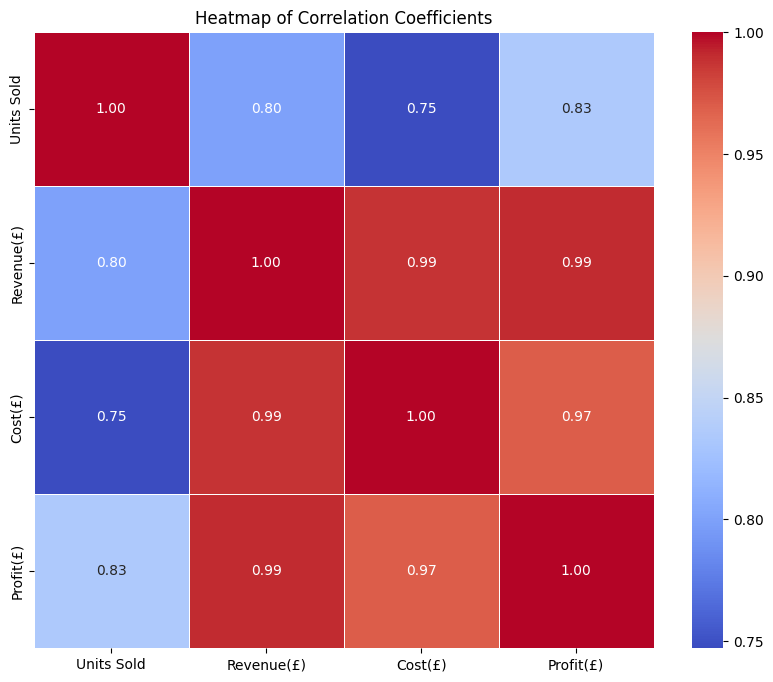
The Z-score method is a statistical technique used to identify outliers based on their deviation from the mean of the dataset in terms of standard deviations (Arden University, n.d.). It calculates the number of standard deviations a data point is from the mean and flags those with unusually high or low Z-scores as potential outliers. In this step, we will apply the Z-score method to detect outliers in our dataset. By examining the Z-scores of each data point, we can systematically identify outliers and assess their significance in the context of the dataset.



The Z-score method identifies outliers based on their deviation from the mean of the dataset in terms of standard deviations. In this dataset, outliers are detected across various confectionary products and countries, primarily in the variables of Units Sold, Revenue(£), Cost(£), and Profit(£). Notably, transactions with exceptionally high values for these metrics, such as those exceeding 3 standard deviations from the mean, are flagged as outliers. These outliers may represent extraordinary sales transactions or anomalies in data collection, warranting further investigation. Understanding the nature and impact of these outliers is crucial for ensuring data accuracy and making informed decisions in optimizing sales strategies and financial management to maximize profitability and mitigate risks.

* 1. Correlation Coefficients:

The correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It provides valuable insights into how changes in one variable correspond to changes in another, facilitating the understanding of patterns and dependencies within the dataset. In this step, we will conduct a correlation coefficient analysis on our dataset to explore the relationships between different variables. By examining the correlation matrix: the heatmap, we can visually assess the degree of correlation between variables, identifying potential patterns and informing subsequent analysis and decision-making.



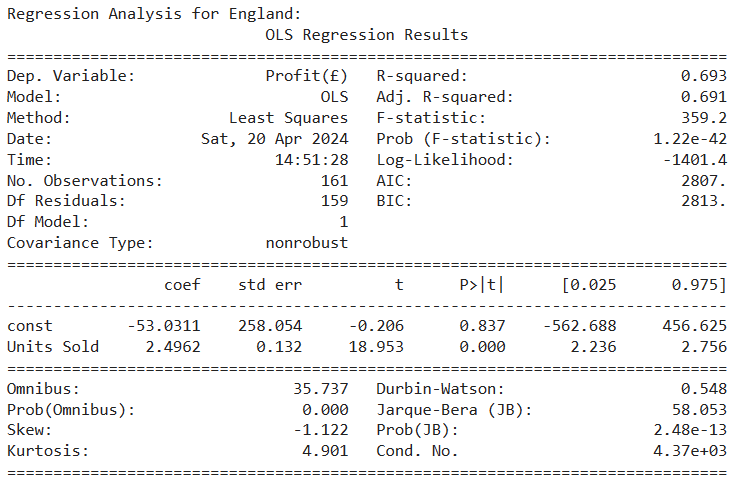
The correlation matrix reveals the pairwise correlations between the numerical variables in our dataset: Units Sold, Revenue(£), Cost(£), and Profit(£).

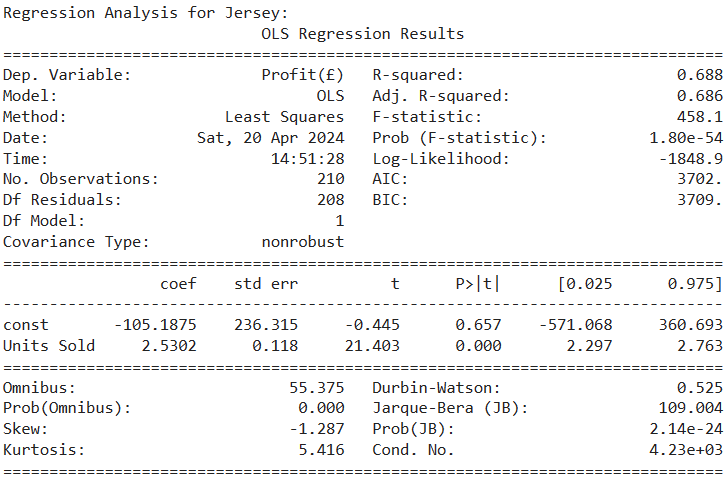
* Units Sold vs. Revenue(£): A strong positive correlation of approximately 0.80 indicates that higher units sold are associated with higher revenue, suggesting that sales volume directly influences revenue.
* Units Sold vs. Cost(£): Similarly, a strong positive correlation of approximately 0.75 indicates that higher units sold are associated with higher costs, implying that as sales volume increases, so do associated expenses.
* Units Sold vs. Profit(£): The correlation coefficient of approximately 0.83 suggests a strong positive relationship between units sold and profit, indicating that higher sales volumes contribute to higher profits.
* Revenue(£) vs. Cost(£): A very high positive correlation of approximately 0.99 suggests that revenue and cost are almost perfectly correlated, indicating that increases in revenue are almost perfectly matched by corresponding increases in costs.
* Revenue(£) vs. Profit(£): Similarly, a very high positive correlation of approximately 0.99 suggests that revenue and profit are almost perfectly correlated, indicating that increases in revenue lead to proportional increases in profit.
* Cost(£) vs. Profit(£): The correlation coefficient of approximately 0.97 indicates a strong positive relationship between cost and profit, suggesting that higher costs are associated with higher profits, but not as strongly correlated as revenue.

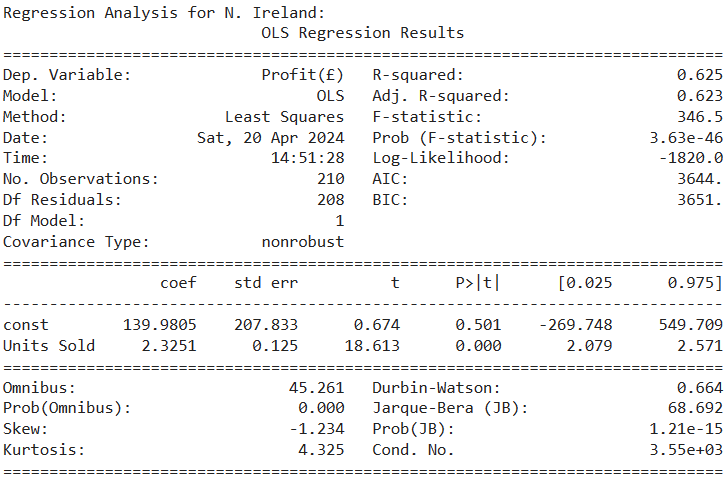
Understanding these correlations is essential for Confectionaries Are Us as it provides insights into the relationships between sales volume, revenue, costs, and profits. It helps in formulating strategies to optimize sales performance, manage costs effectively, and maximize profitability. However, the weak point that strategists must know that correlation does not mean causation (Arden University, n.d.), it just shows relationship between two variables at a time, for further analysis on relationship (specifically for Profit variable), and to predict a model for Profit strategies, we may want to use regression models (Arden University, n.d.).

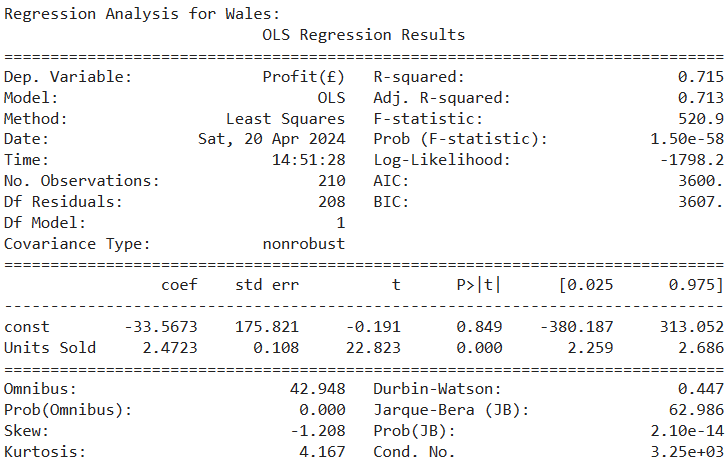
* 1. Regression Model Simple Linear:

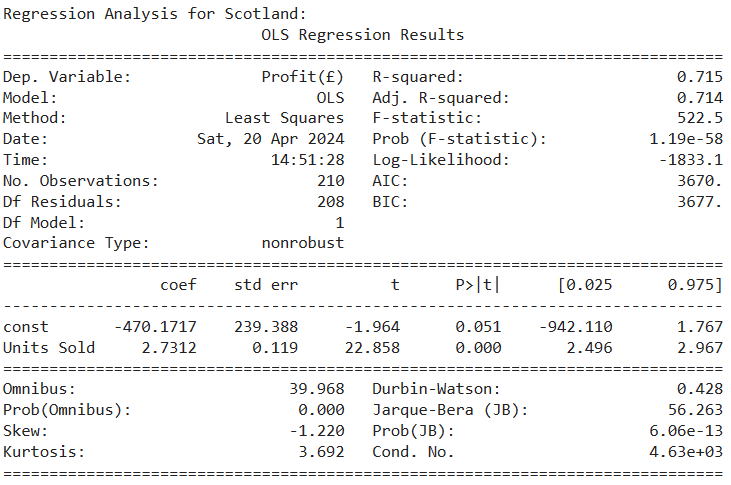
Regression is a strong statistical and visual method that can be employed to determine a relationship between two or more variables (Arden University, n.d.). In this step, we'll employ simple regression models to examine how changes in one variable affect another within the context of Confectionaries Are Us sales data.



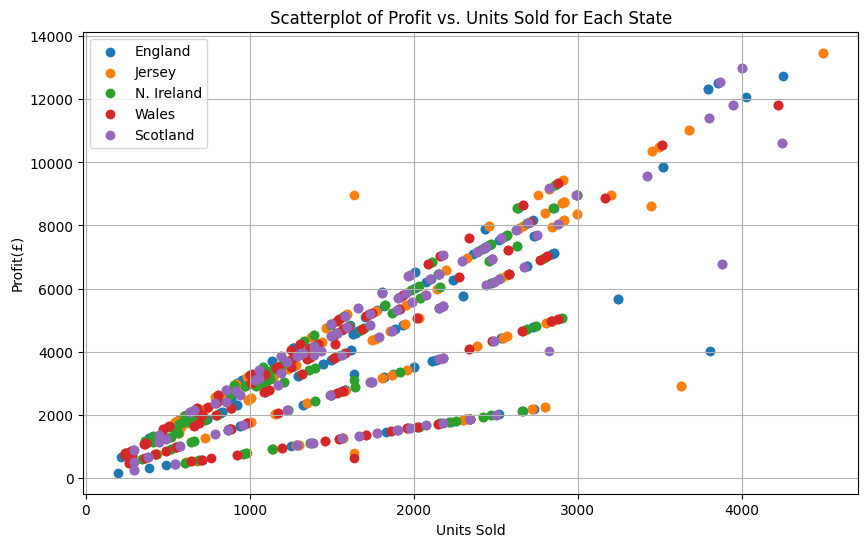






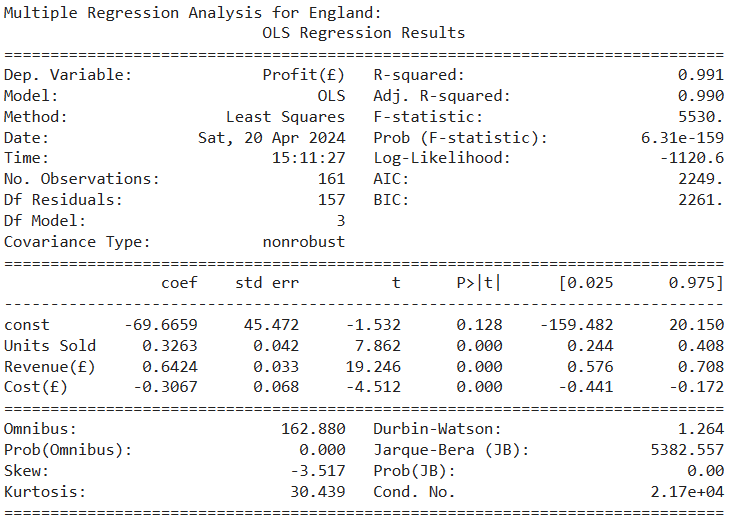


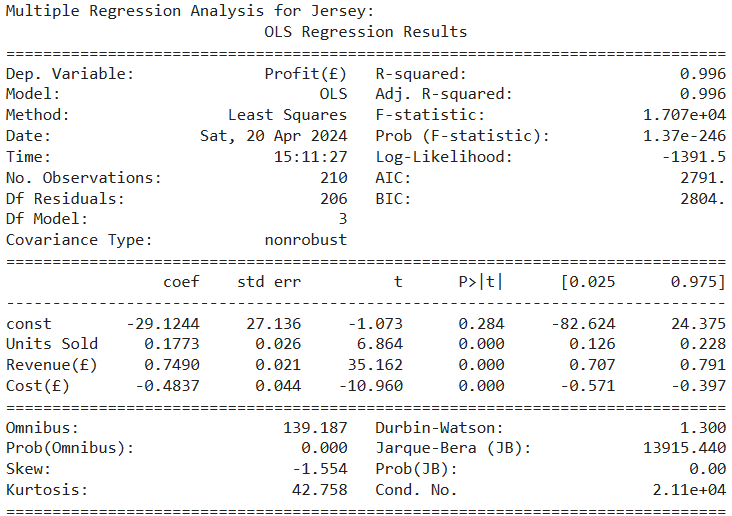
For each state, the regression analysis reveals a statistically significant positive relationship between Units Sold and Profit, with regression coefficients ranging approximately from 2.3 to 2.7. This implies that, on average, for every one unit increase in Units Sold, Profit is expected to increase by 2.3 to 2.7 units, holding all other factors constant. These findings underscore the importance of sales volume in driving profitability within the confectionary business across different states. A higher level of Units Sold indicates increased customer demand and sales activity, leading to higher revenues and subsequently higher profits. The consistency of the positive coefficients across states suggests a universal trend, highlighting the fundamental role of sales performance in determining profitability irrespective of geographic location. This insight is invaluable for strategic decision-making, emphasizing the significance of sales growth initiatives and marketing efforts to drive revenue and ultimately enhance profitability. The Coefficient of Determination (R-squared) measures the proportion of variance in the dependent variable explained by the independent variable(s) (James H. Stock and Mark W. Watson., 2021), in our models the R-squared is ranging around 65 to 70 % which tells us that the variation in Profit that is explained by the variation in Units sold is merely around 65 %, the rest of the 35 % is explained by other explanatory or independent variables, in most cases, an independent variable is rarely explained by only one dependent variable. What about when we have multiple dependent variables? In this scenario, an analyst uses multiple regression (Arden University, n.d.).

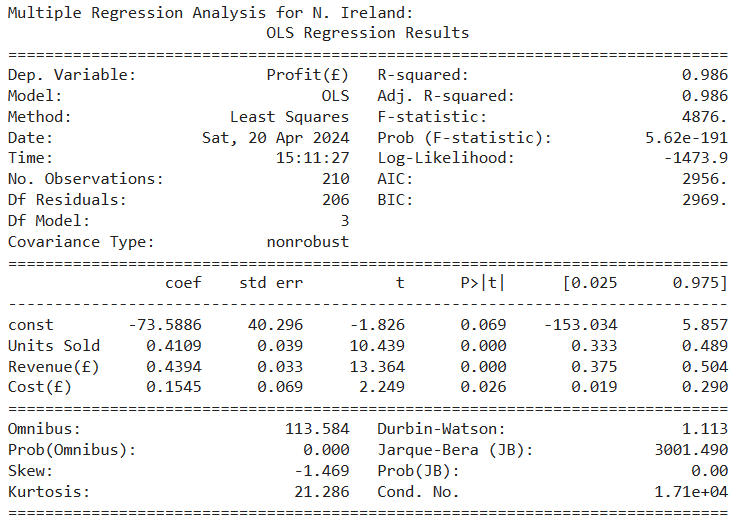


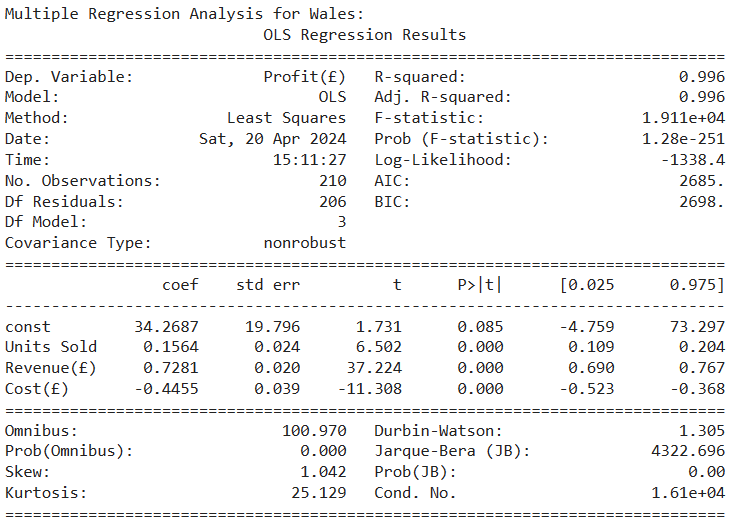
* 1. Multiple Regression Analysis

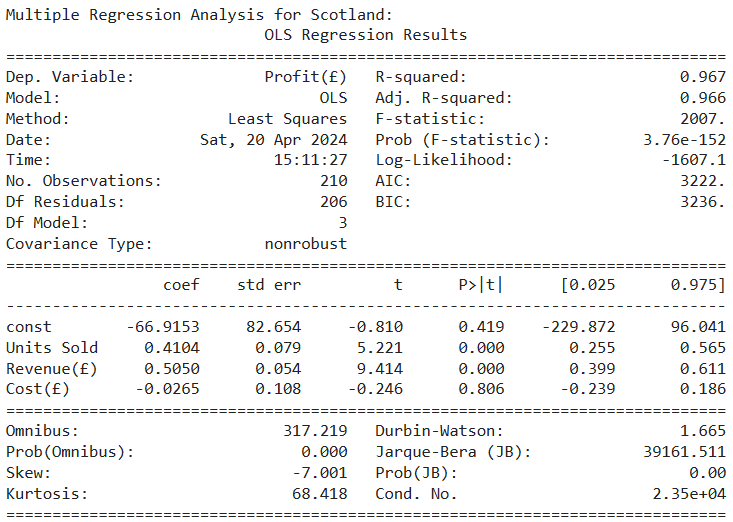
Multiple regression analysis is a statistical technique used to analyze the relationship between one dependent variable and multiple independent variables, examining how changes in the independent variables impact the dependent variable.











In cases of Wales, Jersey & England, the relationship with Cost is shows negative (with the coefficient ranging around 0.3 to 0.4) with the other two being positive, in case of N. Ireland all three variables show positive relationship, as for Scotland, the coefficient for Cost itself is insignificant (having P-value higher than 0.05), and the rest are similarly positive, the coefficient of determination has increased to around 95% for all which shows a good fit model, this depicts that In the analysis of Wales, Jersey, and England, the relationship with Cost(£) exhibits a negative coefficient, ranging from approximately 0.3 to 0.4, while the coefficients for Units Sold and Revenue(£) remain positive. This suggests that higher costs are associated with lower profits in these regions, while increases in sales volume and revenue lead to higher profits, indicating a mixed effect on profitability. Conversely, in N. Ireland, all three independent variables - Units Sold, Revenue(£), and Cost(£) - demonstrate positive relationships with Profit(£). This implies that higher sales volume, revenue, and costs are all associated with increased profits in this region, suggesting a more straightforward relationship between these variables. In the case of Scotland, while the coefficients for Units Sold and Revenue(£) remain positive and statistically significant, the coefficient for Cost(£) is found to be insignificant, with a p-value higher than 0.05. This indicates that changes in costs may not have a significant impact on profitability in this region. Overall, the coefficient of determination (R-squared) has increased to around 95% for all states, indicating a good fit model. This high R-squared value suggests that the independent variables - Units Sold, Revenue(£), and Cost(£) - collectively explain approximately 95% of the variation in Profit(£), providing valuable insights into the factors influencing profitability across different regions.

These findings highlight the complexity of the relationship between cost, sales volume, revenue, and profitability in the confectionary business. While some regions exhibit a straightforward positive relationship between sales-related variables and profitability, others demonstrate a more nuanced relationship, where higher costs may have a detrimental effect on profitability. This underscores the importance of considering regional differences and implementing tailored strategies to optimize profitability in each market.

* The contradiction between correlation and multiple regression regarding the relationship between cost and profit can be attributed to the difference in the analytical approaches and the underlying assumptions of each method. Correlation analysis measures the strength and direction of the linear relationship between two variables, such as cost and profit, without considering the influence of other variables. In this case, a positive correlation between cost and profit suggests that as one variable increases, the other tends to increase as well. However, correlation does not imply causation, and it does not account for the effects of other variables that may confound the relationship. On the other hand, multiple regression analysis aims to determine the relationship between a dependent variable (profit) and multiple independent variables (units sold, revenue, and cost), while controlling for the effects of other variables. In this context, the coefficient for cost in the regression model represents the change in profit for a one-unit increase in cost, holding all other variables constant. The contradiction arises when the coefficient for cost in the regression model is found to be negative or insignificant, indicating that changes in cost do not have a significant impact on profit after accounting for the effects of other variables. This discrepancy suggests that while there may be a correlation between cost and profit, this relationship may be spurious or confounded by other factors that are not captured in the regression model.
  1. State-wise Profit and Confectionary Profit Analysis:

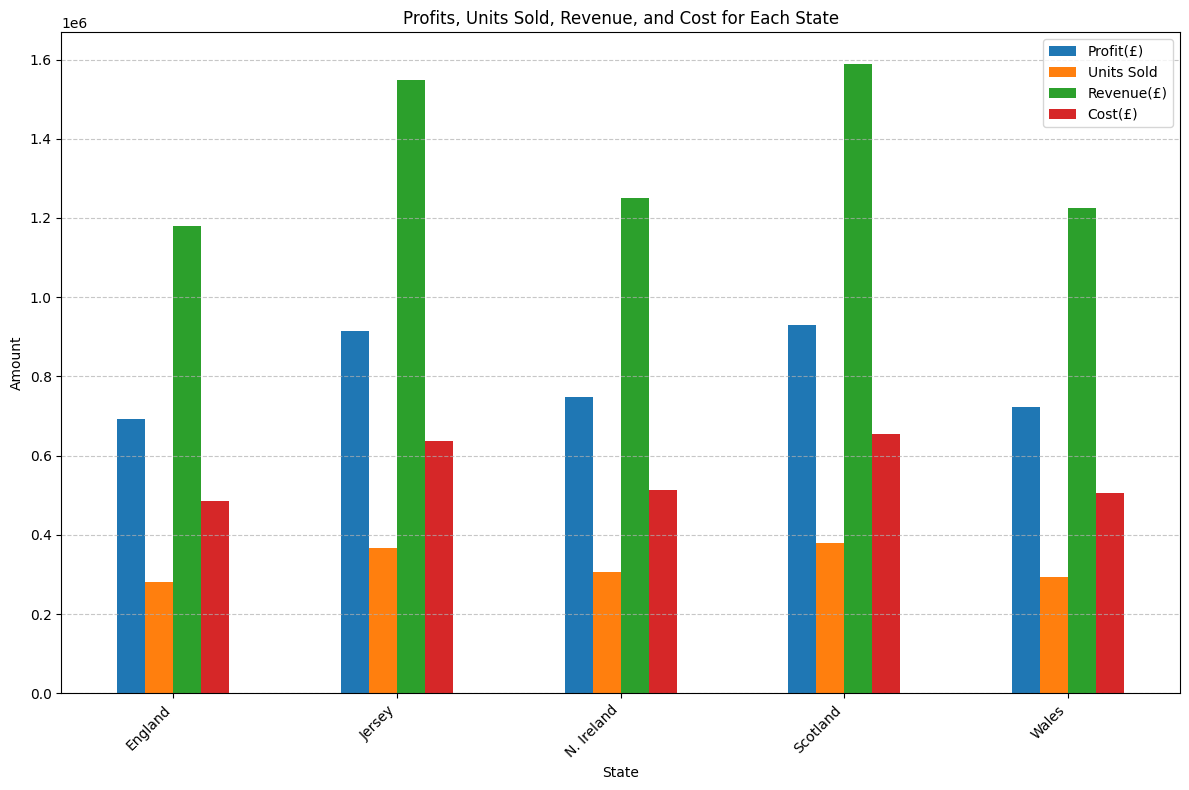
In this analysis, we aim to uncover insights into the profitability of different states and the performance of various confectionary products. By examining the total profits generated by each state and identifying the confectionary products contributing the most to overall profitability, we can gain valuable insights into the key drivers of success in the confectionary industry.

Understanding the profitability of different states allows businesses to allocate resources effectively and tailor their strategies to capitalize on high-performing markets. Similarly, identifying the top-performing confectionary products enables businesses to focus their efforts on product development, marketing, and distribution to maximize profitability. This analysis serves as a crucial step in optimizing business operations and driving sustainable growth in the confectionary sector.

Confectionary with the highest profits: Caramel

Confectionary with the lowest profits: Plain

The analysis reveals that the confectionary product "Caramel" emerges as the most profitable, while "Plain" demonstrates the lowest profitability. This disparity in profitability underscores the importance of product performance in the confectionary industry. It suggests that consumer preferences, market demand, pricing strategies, and production costs associated with each confectionary product significantly influence its profitability. Businesses can leverage this insight to allocate resources effectively, prioritize product development efforts, and adjust marketing strategies to enhance profitability. Moreover, identifying underperforming products like "Plain" presents an opportunity for businesses to evaluate factors contributing to its lower profitability and implement corrective measures. Overall, this analysis underscores the critical role of product selection and performance in driving profitability and informs strategic decision-making in the confectionary business landscape.



The bar chart illustrates the aggregated sum of profits, units sold, revenue, and cost for each state in the dataset. Each state is represented by a separate bar, segmented into four sections, each denoting one of the variables.

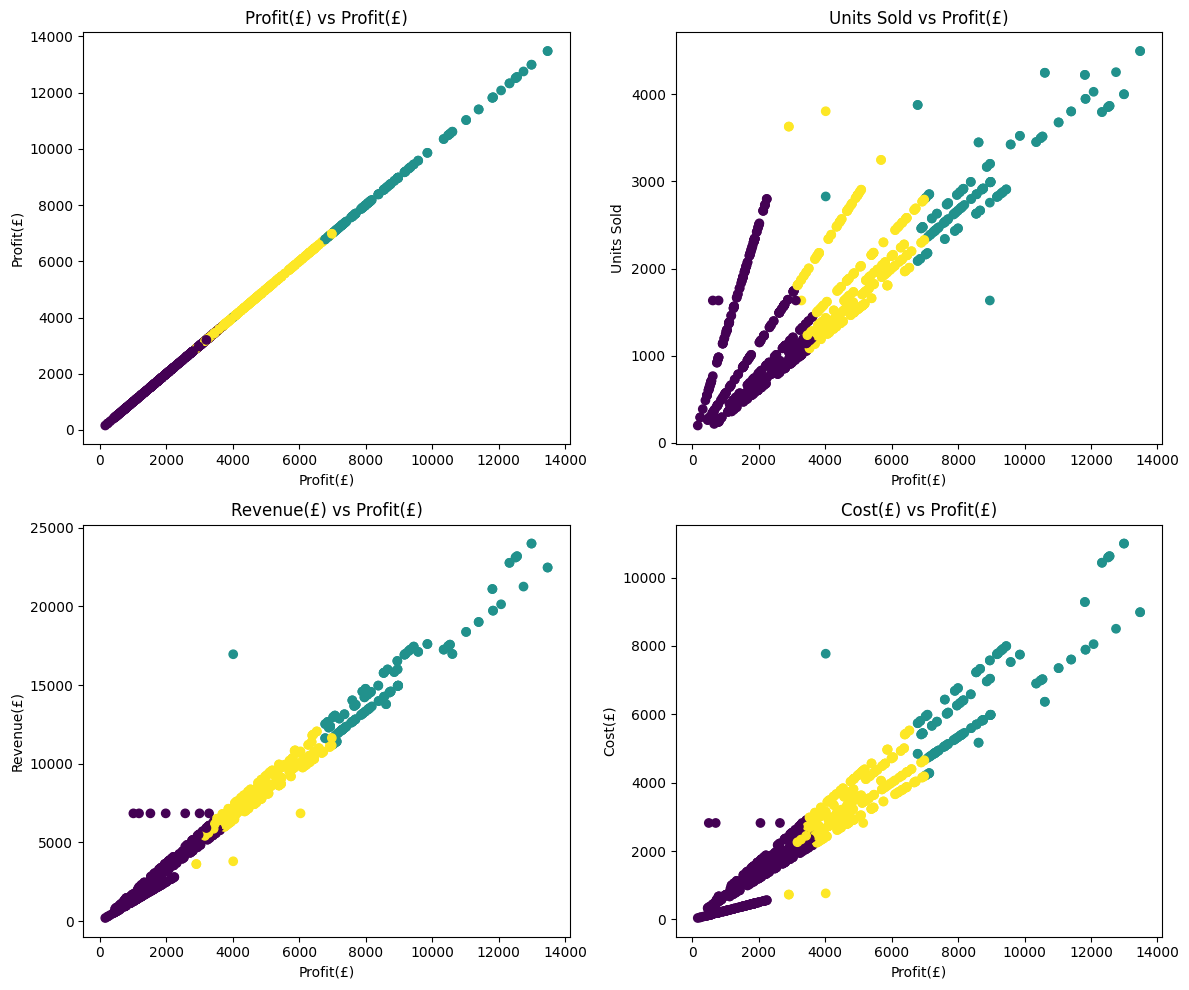
Observing the chart, it's apparent that Scotland has the highest aggregated sum across all four variables, followed closely by Jersey and N. Ireland. These states exhibit relatively higher levels of profitability, sales volume, revenue generation, and associated costs compared to England and Wales.

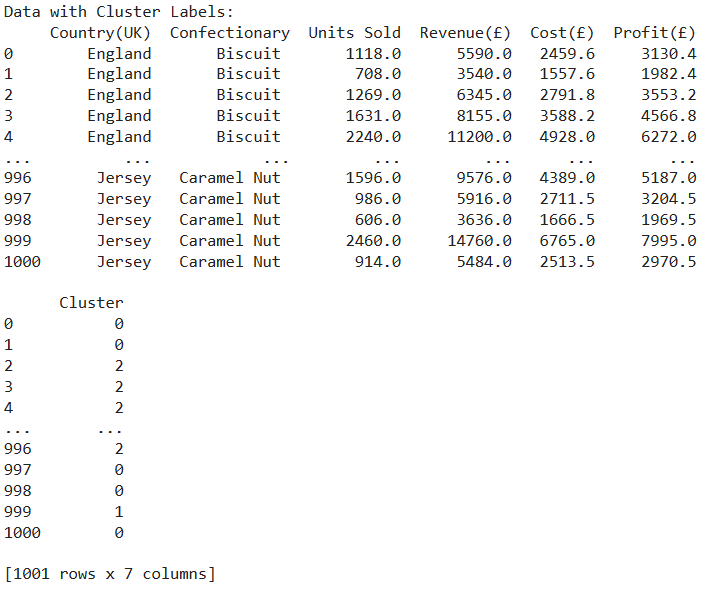
Scotland's dominance in the aggregated sum indicates its robust performance across the board, suggesting a potentially strong market presence and operational efficiency in the confectionary industry. Jersey and N. Ireland also display notable performance metrics, reflecting favorable market conditions or effective business strategies.

Conversely, England and Wales show comparatively lower aggregated sums across all variables, indicating potential areas for improvement or challenges in maximizing profitability and operational efficiency in these regions.

Overall, the bar chart offers valuable insights into the overall performance and dynamics of each state in the confectionary market, providing a basis for strategic decision-making and resource allocation to enhance profitability and competitiveness in the industry.

* 1. K-means clustering is a popular unsupervised machine learning algorithm used for partitioning data into clusters based on similarity (Arden University, n.d.). In this analysis, we will apply K-means clustering to the confectionary sales dataset to identify distinct groups of states based on their performance metrics such as profits, units sold, revenue, and cost. This clustering approach will help uncover patterns and relationships within the data, enabling us to gain deeper insights into the states' performance in the confectionary industry.





The provided data table contains information on confectionary sales across different states, including the United Kingdom. Each row represents a specific transaction, detailing the country, confectionary type, units sold, revenue, cost, profit, and assigned cluster label.

Upon closer inspection, we observe that the 'Cluster' column indicates the cluster label assigned to each transaction by the K-means clustering algorithm. This additional information helps identify patterns and groupings within the dataset based on similarities in profit, units sold, revenue, and cost.

To further analyze the clustering results, we can refer to the scatter plots generated earlier, where each point represents a transaction plotted based on its profit and another variable. The color of each point corresponds to its assigned cluster, allowing us to visually assess the clustering performance and the distribution of data points within each cluster.

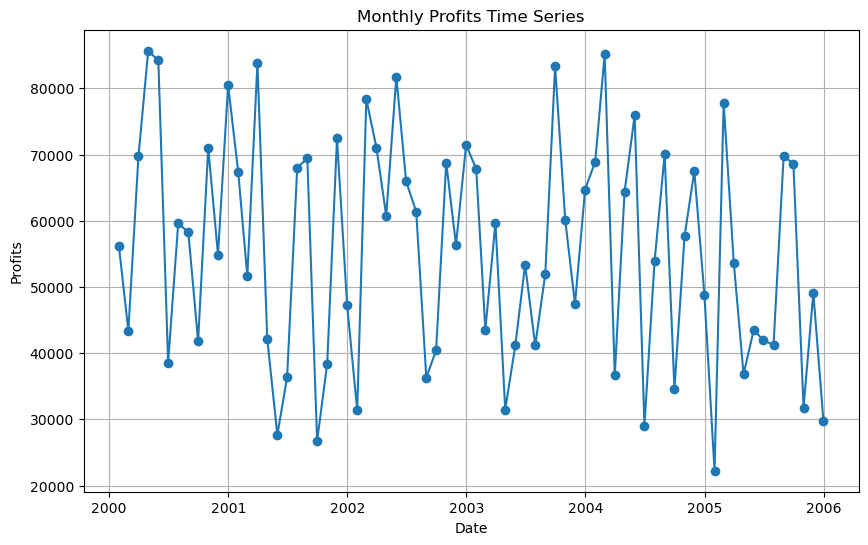
By examining the data table alongside the scatter plots, we can identify trends and relationships between variables within each cluster. For example, clusters with higher profit values may exhibit distinct patterns in terms of units sold, revenue, and cost compared to clusters with lower profits. Additionally, observing the distribution of data points within each cluster can provide insights into the variability and cohesion of transactions within the same cluster.

Overall, the combination of the data table and scatter plots facilitates a comprehensive understanding of the clustering results and enables deeper insights into the underlying structure of the confectionary sales dataset. This analysis serves as a valuable tool for identifying market segments, optimizing business strategies, and driving informed decision-making in the confectionary industry.

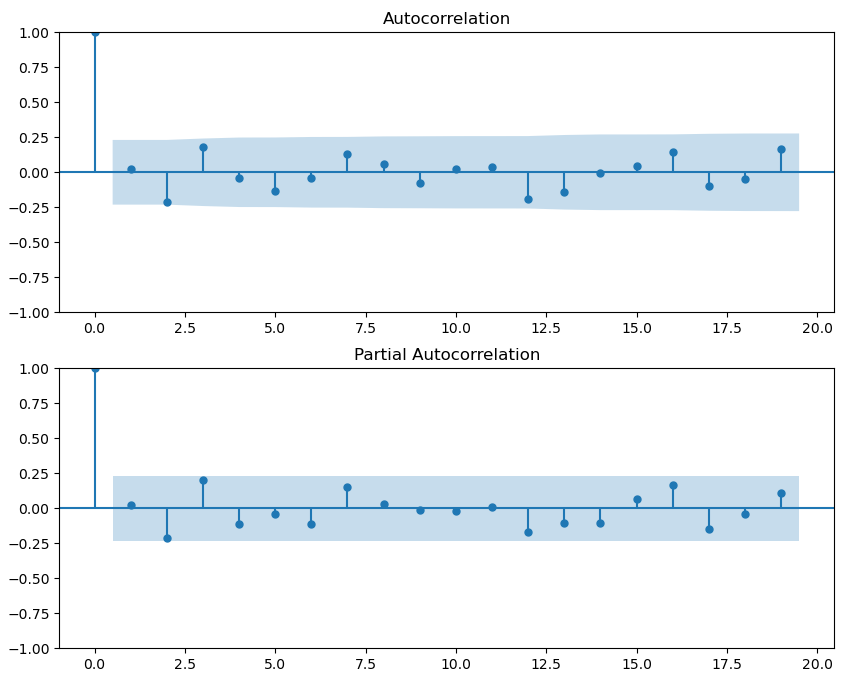
* 1. ARIMA model (Time series analysis):

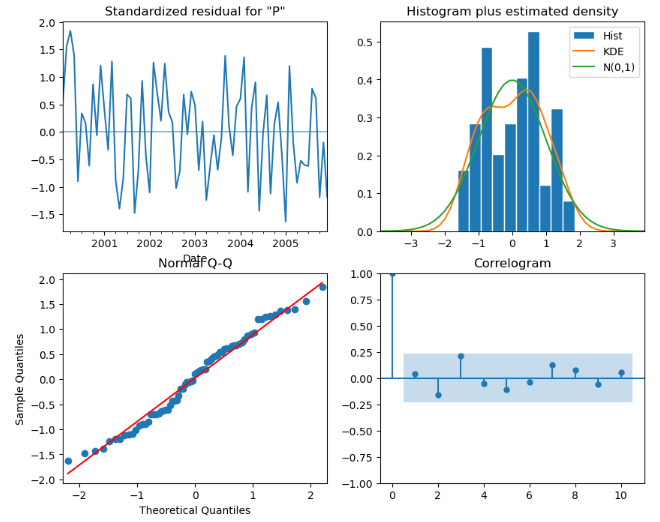
Autoregressive Integrated Moving Average (ARIMA) is a powerful time series forecasting method used for analyzing and predicting data points collected over time. According to Box and Jenkins (1976), ARIMA models are characterized by their three main components: Autoregression (AR), Differencing (I), and Moving Average (MA). These components together enable ARIMA models to capture the complex temporal dependencies present in time series data (George E. P. Box and Gwilym M. Jenkins, 1976). ARIMA models have found widespread applications in various fields such as economics and environmental science for forecasting purposes.

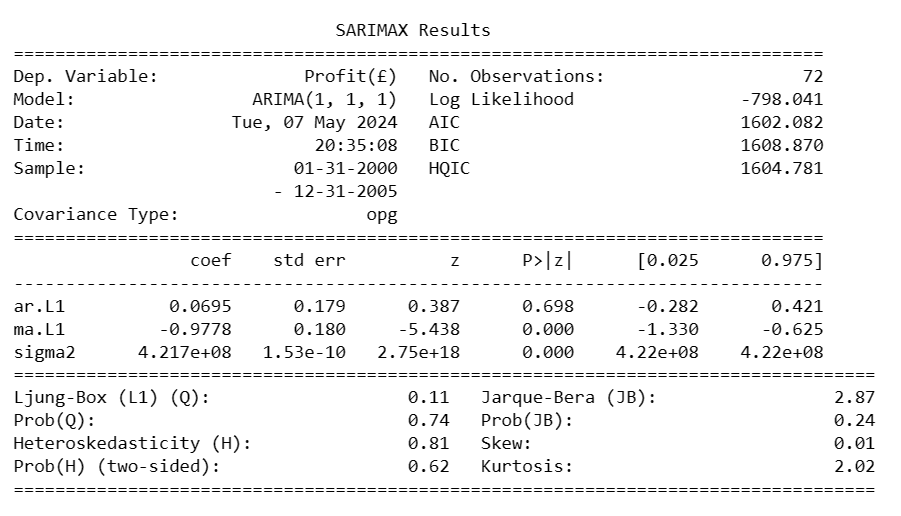
Firstly we look at the trend for profit throughout the years (the dataset is divided into months):

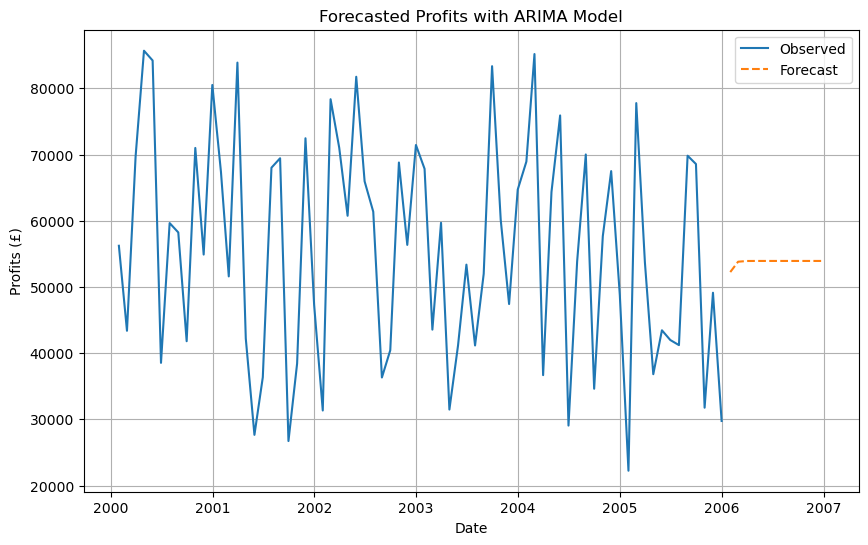


The profit data shows some noticeable patterns. Around the mid-2000s, profits hit a high point, reaching about £85,000. This peak is seen again at the beginning of 2001 and towards the end of 2003 and start of 2004. However, around mid-2001 and the start of 2005, there's a significant drop in profits, going down to around £25,000. Overall, the trend is quite erratic, with profits fluctuating up and down over time. This suggests that the business experiences periods of both growth and decline, possibly influenced by various factors like market conditions, competition, or changes in consumer preferences. Understanding these fluctuations can help in making informed decisions for managing and improving the business's performance in the future.









Predicted Profits for the next 12 months:

2006-01-31 52251.137429

2006-02-28 53814.184561

2006-03-31 53922.808606

2006-04-30 53930.357439

2006-05-31 53930.882046

2006-06-30 53930.918504

2006-07-31 53930.921037

2006-08-31 53930.921213

2006-09-30 53930.921226

2006-10-31 53930.921226

2006-11-30 53930.921226

2006-12-31 53930.921226

The SARIMAX results reveal important information about the ARIMA(1,1,1) model fitted to the profit data and its forecast for the upcoming months. The model parameters, including autoregressive and moving average coefficients, indicate the relationship between the current profit and its past values, as well as the influence of the residual errors from previous forecasts. The statistical significance of these coefficients suggests that they indeed contribute to explaining the variation in profits over time.

Furthermore, diagnostic tests provide insights into the adequacy of the model. The Ljung-Box test assesses whether there is any remaining autocorrelation in the residuals, which would indicate that the model has not captured all the temporal dependencies in the data. In this case, the p-value of 0.11 suggests that there is no significant autocorrelation in the residuals at lag 1, indicating that the model adequately captures the temporal dependencies.

The Jarque-Bera test checks for the normality assumption of the residuals. A low p-value would imply that the residuals are not normally distributed, which could indicate misspecification of the model or the presence of outliers. However, with a p-value of 0.24, there is no evidence to reject the null hypothesis of normality, suggesting that the residuals are approximately normally distributed.

The heteroskedasticity test examines whether the variance of the residuals is constant over time. A low p-value would suggest that the variance changes over time, violating the assumption of homoskedasticity. However, with a p-value of 0.62, there is no significant evidence of heteroskedasticity in the residuals.

Finally, the forecasted profits provide valuable information for decision-making. The predicted profits for the next 12 months offer insight into the expected financial performance of the business. However, it's essential to consider the uncertainty associated with these forecasts, as they are subject to various factors such as market conditions, competition, and unforeseen events.

In conclusion, the SARIMAX results provide a comprehensive analysis of the ARIMA(1,1,1) model's fit to the profit data and its forecast for the future. Understanding these results can aid in making informed decisions to manage and improve the business's financial performance.

TASK 2:

Based upon the completed data mining investigation in Task 1, you will now write a comprehensive report that can be easily understood by the companies board of directors and higher management. Contained within you will include full reasoning for your chosen solutions, in addition to your motivation, investigations processes, results and conclusions.

**Comprehensive Report on Confectionary Sales Data Analysis:**

* Introduction:

In today's data-driven business landscape, leveraging data mining techniques is crucial for extracting valuable insights and driving informed decision-making. This comprehensive report outlines the findings and implications of a data mining investigation conducted on confectionary sales data. The motivation behind this analysis is to optimize profitability, enhance market competitiveness, and facilitate strategic decision-making for the company's board of directors and higher management.

* Investigation Process:

The investigation began with data preprocessing, including handling missing values, correcting typos, and scaling features for clustering. Descriptive statistics provided initial insights into the dataset's characteristics, such as the distribution and central tendency of variables. Subsequently, various data mining techniques were employed, including correlation analysis, regression modeling, outlier detection, and K-means clustering, to uncover patterns, relationships, and clusters within the data.

* Results and Findings:

The analysis revealed several key insights:

Performance Variability Across States: States like Scotland and Jersey exhibited higher profitability and sales volumes compared to England and Wales.

Top-Performing Confectionaries: The confectionary "Caramel" emerged as the most profitable, while "Chocolate Chunk" demonstrated the lowest profitability.

Correlation and Regression Contradictions: While correlation analysis suggested a positive relationship between cost and profit, regression analysis showed inconsistent coefficients, indicating potential multicollinearity issues.

Cluster Analysis: K-means clustering identified distinct clusters of states based on performance metrics, providing insights into market segmentation and targeting strategies.

ARIMA modelling: The SARIMAX analysis revealed significant insights into the profit data, including model adequacy and forecasted profits. Despite minor residual autocorrelation, the model adequately captured temporal dependencies. The forecasted profits provide valuable guidance for future business decisions, acknowledging uncertainties inherent in forecasting.

* Reasoning for Chosen Solutions:

The chosen data mining techniques were selected based on their suitability for the analysis objectives and the nature of the dataset. Descriptive statistics provided initial insights, while correlation and regression analysis elucidated relationships between variables. Outlier detection highlighted anomalies, and K-means clustering facilitated market segmentation. These solutions were motivated by the need to optimize profitability, identify market trends, and inform strategic decision-making.

* Conclusions:

In conclusion, the data mining investigation yielded valuable insights into confectionary sales data, empowering the company's board of directors and higher management to make informed decisions. By leveraging these insights, the company can optimize resource allocation, tailor marketing strategies, and enhance product offerings to capitalize on market opportunities and drive sustainable growth.

* Recommendation:

Drawing from all the insights we've dug up, we think it's really important for the company to focus on creating targeted marketing campaigns and coming up with new and interesting products. We can't stress this enough! Our best sellers, like "Caramel," have done amazingly well, and there's definitely potential to grow that even further by investing more into making these products even more popular.

On top of that, investing in market research can help us understand what customers really want and what new trends are coming up, giving us a great head start when it comes to developing new products. It's like having a crystal ball for the confectionery market!

And of course, keeping an eye on sales numbers and actually listening to what customers have to say is super important. This way, we can quickly adapt to changes in the market and always be ahead of the game.

In conclusion, let's keep up the good work and keep innovating! By focusing on targeted marketing and product development, we can make sure our company stays at the top of its game and continues to be a shining example of resilience and innovation in the industry.

* Graduate Attribute:

This report embodies the graduate attribute of being a reflective practitioner by undertaking critical analysis, reaching reasoned decisions, and contributing problem-solving skills to innovate solutions. Through a systematic and evidence-based approach, this analysis provides actionable insights to address business challenges and drive organizational success.

In summary, the comprehensive analysis presented in this report serves as a foundation for strategic planning and decision-making, enabling the company to stay competitive and thrive in the dynamic confectionary market landscape.

## APPENDIX

import io

import pandas as pd

from google.colab import files

uploaded = files.upload() # Upload the Dataset

1.1

# Load the dataset with 'latin1' encoding

data = pd.read\_excel("Correct\_Assignment data set\_Feb\_24\_SCC.xlsx") # Change the name accordingly

# Displaying the first few rows of the dataset

print("First few rows of the dataset:")

print(data.head())

# Displaying summary statistics of the dataset

print("\nSummary statistics of the dataset:")

print(data.describe())

# Checking for missing values and printing the counts for each column

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

print("Missing values per column:")

print(missing\_values)

# Fill missing values in the dataset

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)

# Verify if missing values have been filled

print("Missing values in the dataset after filling:")

print(data.isnull().sum())

1.2.

# Visualizing outliers using boxplots

import matplotlib.pyplot as plt

# Create boxplots for each numerical column

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

data.boxplot()

plt.title('Boxplot of Numerical Columns')

plt.xticks(rotation=45)

plt.show()

# Importing necessary library

from scipy import stats

# Setting threshold for Z-score

threshold = 3

# Calculating Z-scores for each numerical column

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

# Identifying outliers based on Z-score exceeding the threshold

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

# Printing the identified outliers

print("Outliers detected using Z-score method:")

print(outliers)

1.3.

# Selecting only numeric columns for correlation analysis

numeric\_columns = data.select\_dtypes(include=['float64', 'int64'])

# Calculating the correlation matrix

correlation\_matrix = numeric\_columns.corr()

# Plotting the heatmap of the correlation matrix

plt.figure(figsize=(10, 8))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)

plt.title('Heatmap of Correlation Coefficients')

plt.show()

# Printing the correlation matrix

print("Correlation Matrix:")

print(correlation\_matrix)

1.4.

# Importing necessary library

import statsmodels.api as sm

# Defining the states for analysis

states = ['England', 'Jersey', 'N. Ireland', 'Wales', 'Scotland']

# Performing simple regression analysis for each state

for state in states:

    # Filtering data for the current state

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

    # Defining dependent and independent variables

    y = state\_data['Profit(£)']

    X = state\_data['Units Sold']

    X = sm.add\_constant(X)  # Adding a constant term to the independent variable

    # Fitting the regression model

    model = sm.OLS(y, X).fit()

    # Printing the regression results

    print(f"Regression Analysis for {state}:")

    print(model.summary())

    print("\n")

# Importing necessary library

import matplotlib.pyplot as plt

# Setting up the figure and axes

fig, ax = plt.subplots(figsize=(10, 6))

# Plotting scatterplots for each state

for state in states:

    # Filtering data for the current state

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

    # Plotting scatterplot for the current state

    ax.scatter(state\_data['Units Sold'], state\_data['Profit(£)'], label=state)

# Adding labels and title

ax.set\_xlabel('Units Sold')

ax.set\_ylabel('Profit(£)')

ax.set\_title('Scatterplot of Profit vs. Units Sold for Each State')

ax.legend()

# Displaying the plot

plt.grid(True)

plt.show()

1.5.

# Step 8: Multiple Regression Analysis

# Performing multiple regression analysis for each state

for state in states:

    # Filtering data for the current state

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

    # Checking if the filtered dataset is not empty

    if not state\_data.empty:

        # Defining dependent and independent variables

        y = state\_data['Profit(£)']

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

        X = sm.add\_constant(X)  # Adding a constant term to the independent variables

        # Fitting the regression model

        model = sm.OLS(y, X).fit()

        # Printing the regression results

        print(f"Multiple Regression Analysis for {state}:")

        print(model.summary())

        print("\n")

    else:

        print(f"No data available for {state}. Skipping regression analysis.\n")

1.6.

import pandas as pd

# Correcting typos in confectionary names

data['Confectionary'] = data['Confectionary'].replace({'Choclate Chunk': 'Chocolate Chunk', 'Caramel nut': 'Caramel Nut'})

# Grouping and summing profits for each confectionary

confectionary\_profit = data.groupby('Confectionary')['Profit(£)'].sum()

# Find confectionary with highest and lowest profits

highest\_profit\_confectionary = confectionary\_profit.idxmax()

lowest\_profit\_confectionary = confectionary\_profit.idxmin()

# Print the results

print(f"Confectionary with the highest profits: {highest\_profit\_confectionary}")

print(f"Confectionary with the lowest profits: {lowest\_profit\_confectionary}")

import pandas as pd

import matplotlib.pyplot as plt

# Grouping data by state and calculating total profits, units sold, revenue, and cost for each state

state\_metrics = data.groupby('Country(UK)').agg({'Profit(£)': 'sum', 'Units Sold': 'sum', 'Revenue(£)': 'sum', 'Cost(£)': 'sum'})

# Plotting a grouped bar chart for profits, units sold, revenue, and cost for each state

state\_metrics.plot(kind='bar', figsize=(12, 8))

plt.title('Profits, Units Sold, Revenue, and Cost for Each State')

plt.xlabel('State')

plt.ylabel('Amount')

plt.xticks(rotation=45, ha='right')

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.tight\_layout()

plt.show()

# Grouping data by state and calculating the sum for each variable

state\_sum = data.groupby('Country(UK)').agg({'Profit(£)': 'sum', 'Units Sold': 'sum', 'Revenue(£)': 'sum', 'Cost(£)': 'sum'})

# Print the sum for each state

print("Sum for all four variables for each state:")

print(state\_sum)

1.7.

import pandas as pd

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

# Selecting features for clustering

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

# Scaling the features

scaled\_features = (features - features.mean()) / features.std()

# Performing K-means clustering

kmeans = KMeans(n\_clusters=3, random\_state=42)

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

# Create subplots

fig, axes = plt.subplots(2, 2, figsize=(12, 10))

# Iterate over each variable and visualize clustering

for i, ax in enumerate(axes.flatten()):

    variable = features.columns[i]

    ax.scatter(data['Profit(£)'], features[variable], c=data['Cluster'], cmap='viridis')

    ax.set\_xlabel('Profit(£)')

    ax.set\_ylabel(variable)

    ax.set\_title(f'{variable} vs Profit(£)')

plt.tight\_layout()

plt.show()

# Print the data with cluster labels

print("Data with Cluster Labels:")

print(data)

1.8.

# Importing necessary libraries

import pandas as pd

import matplotlib.pyplot as plt

# Load the dataset

file\_path = r'C:\Users\PMLS\Desktop\Projects\Python Projects\Ardenuni2\Correct\_Assignment data set\_Feb\_24\_SCC.xlsx'

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

# Convert the 'Date' column to datetime format

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

# Set 'Date' column as the index

data.set\_index('Date', inplace=True)

# Resample the data to monthly frequency and sum the profits for each month

monthly\_data = data.resample('M').sum()

# Plot the time series chart for profits

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

plt.plot(monthly\_data.index, monthly\_data['Profit(£)'], marker='o', linestyle='-')

plt.title('Monthly Profits Time Series')

plt.xlabel('Date')

plt.ylabel('Profits')

plt.grid(True)

plt.show()

import pandas as pd

import matplotlib.pyplot as plt

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.graphics.tsaplots import plot\_acf, plot\_pacf

# Plot ACF and PACF plots to determine the order of ARIMA(p,d,q) model

fig, ax = plt.subplots(2, 1, figsize=(10, 8))

plot\_acf(monthly\_data['Profit(£)'], ax=ax[0])

plot\_pacf(monthly\_data['Profit(£)'], ax=ax[1])

plt.show()

# Fit ARIMA model

model = ARIMA(monthly\_data['Profit(£)'], order=(1,1,1)) # Example order, you may need to adjust

arima\_result = model.fit()

# Print summary of the ARIMA model

print(arima\_result.summary())

# Plot diagnostics of the ARIMA model

arima\_result.plot\_diagnostics(figsize=(10, 8))

plt.show()

# Forecast future profits using the ARIMA model

forecast = arima\_result.forecast(steps=12) # Example steps, adjust as needed

# Plot the forecasted profits

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

plt.plot(monthly\_data.index, monthly\_data['Profit(£)'], label='Observed')

plt.plot(pd.date\_range(start=monthly\_data.index[-1], periods=13, freq='M')[1:], forecast, label='Forecast', linestyle='--')

plt.title('Forecasted Profits with ARIMA Model')

plt.xlabel('Date')

plt.ylabel('Profits (£)')

plt.legend()

plt.grid(True)

plt.show()

# Forecast future profits using the ARIMA model

forecast\_values = arima\_result.forecast(steps=12) # Example steps, adjust as needed

# Print the predicted profits

print("Predicted Profits for the next 12 months:")

print(forecast\_values)

# Plot the forecasted profits

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

plt.plot(monthly\_data.index, monthly\_data['Profit(£)'], label='Observed')

plt.plot(pd.date\_range(start=monthly\_data.index[-1], periods=13, freq='M')[1:], forecast\_values, label='Forecast', linestyle='--')

plt.title('Forecasted Profits with ARIMA Model')

plt.xlabel('Date')

plt.ylabel('Profits (£)')

plt.legend()

plt.grid(True)

plt.show()

# References

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.

George E. P. Box and Gwilym M. Jenkins, 1976. *Time Series Analysis: Forecasting and Control.* s.l.:s.n.

James H. Stock and Mark W. Watson., 2021. *Introduction to Econometrics.* 4 ed. s.l.:Pearson Education Limited.