**ASSESSMENT 2**

**Table of Contents**

[Introduction 4](#_Toc172412096)

[Task A –Data Transformation 4](#_Toc172412097)

[Task A(a) - Load the dataset into a DataFrame 4](#_Toc172412098)

[Task A(b) - Show the first few rows of the loaded dataset 5](#_Toc172412099)

[Task A(c) - Apply three operations to handle missing values in the dataset 6](#_Toc172412100)

[Task A(d) - Choose a column and perform the sorting technique 7](#_Toc172412101)

[Task A(e) - Define a condition to filter transactions from the dataset 8](#_Toc172412102)

[Task A(f) - Create a new column to derive additional information 9](#_Toc172412103)

[Task A(g) - Choose the categorical column and aggregate data based on it 10](#_Toc172412104)

[Task B – Data Analysis 11](#_Toc172412105)

[Task B(a) - Group the dataset based on a categorical variable and calculate summary statistics 11](#_Toc172412106)

[Task B(b) - Investigate the correlations between different variables in the dataset 12](#_Toc172412107)

[Task B(c) - Export a dataset to a CSV file using Python or any other similar programming tool 13](#_Toc172412108)

[Task B(d) - Perform data analysis and visualization in Excel, Python or any other similar programming tool to derive insights 14](#_Toc172412109)

[Task B(e) - Apply inferential statistical methods to quantify the relationships between variables 16](#_Toc172412110)

[Task C – Data Findings and Decision Support 17](#_Toc172412111)

[Task C(a) - Analyze the results obtained from data analysis, including grouping, summarizing, investigating correlations, and applying inferential statistical methods 17](#_Toc172412112)

[Task C(b) - Interpret the relationships between variables, summarize key findings, and identify significant trends or patterns 18](#_Toc172412113)

[Task C(c) - Provide specific suggestions for addressing business challenges or opportunities identified in the dataset 18](#_Toc172412114)

[Conclusion 19](#_Toc172412115)

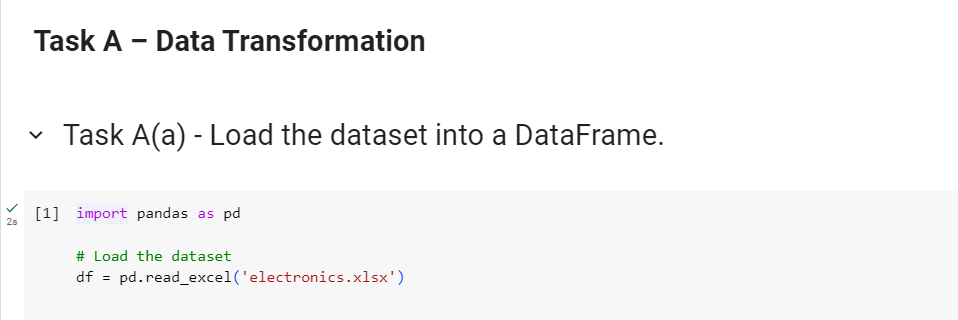
[References 20](#_Toc172412116)

# Introduction

Data Collection and Analysis is a core course widely aimed at the effective utilization of data in management of organizations. This course also introduces different approaches of how data may be gathered from different sources and also it stresses on the ways of converting, scrutinizing and interpreting the collected data. This way, the course demonstrates that data analysis can help one identify the areas of potential development and the problems that should be solved. Such an orientation guarantees that the business can harness the power of data and implement the right tools to enhance the organizational processes. Finally, the course highlights the importance of data supporting the strategic business goals and objectives.

# Task A –Data Transformation

## Task A(a) - Load the dataset into a DataFrame

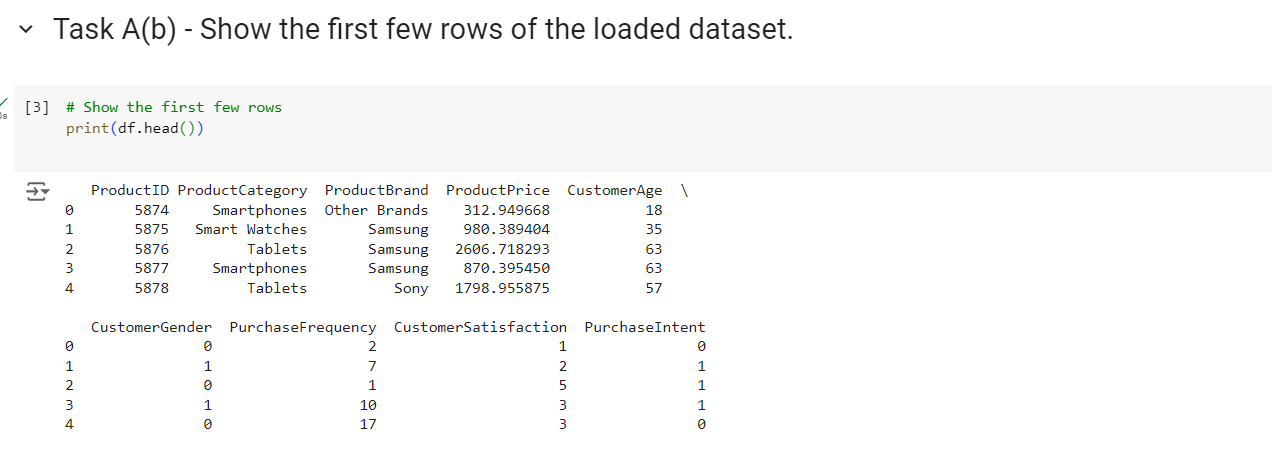


**Figure 1: Loading the dataset into a DataFrame**

(Source: Created by the learner)

The above figure displays the code to load the given data set into a DataFrame, which is the initial and common step, when using the Python data manipulation tool – Pandas. The type of data structures named DataFrames are represented in the form of labeled and rectilinear two-dimensional tables of data analogous to the concept of a spreadsheet. The first line of the code defines a method in which the Pandas library is imported as pd, which is the most often used abbreviation for it by data analysts. Pandas is specifically developed for the use of data retrieval and analysis with excellent speed (Braun and Clarke, 2021). The next one is to build a DataFrame object with the name of ‘df’ using the function pd.read\_excel’electronics’. This function fetches the data from the Excel file with the name electronics.xlsx. csv and puts the data into a variable called DataFrame. This approach is famous because simple and efficient in handling of data for further analysis.

## Task A(b) - Show the first few rows of the loaded dataset

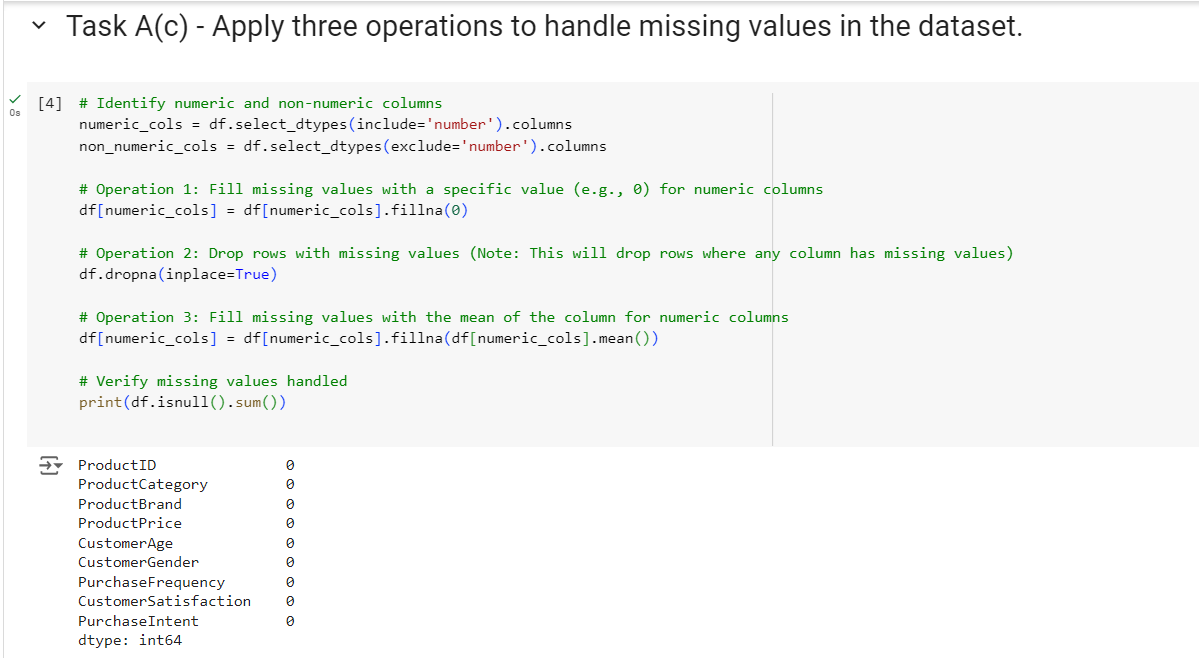


**Figure 2: Showing the first few rows of the loaded dataset**

(Source: Created by the learner)

The image below represents the output of the script written in the python language for Task A(b) containing the first lines of the data loaded. This output, however, involves the use of tables that are like simple spreadsheets with the labels on the columns and each row as a single record of data. The first column gives numerical tags and the second one, the ‘Product ID, contains names like ‘smartphones and ‘tablets, which imply the targeted sale of products (Wang *et al.* 2020).It also shows the initial few rows of the loaded dataset, providing a snapshot of the data structure and content. This early view allows for a preliminary assessment of the dataset's variables and their respective values, facilitating further analysis and data manipulation steps.

## Task A(c) - Apply three operations to handle missing values in the dataset

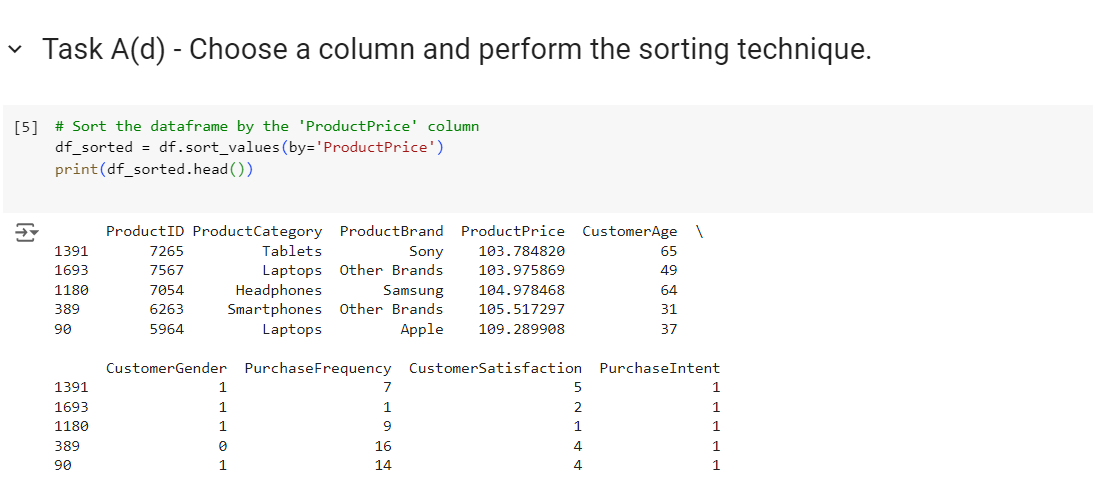


**Figure 3: Applying three operations to handle missing values in the dataset**

(Source: Created by the learner)

The above figure shows the application of three operations to handle missing values in a dataset. First, missing values in numeric columns are replaced with 0 to ensure numerical continuity. Second, any rows containing missing values in any column are removed, enhancing data integrity. Third, missing values in numeric columns are filled with the mean of the respective columns, which maintains the dataset’s statistical properties. These operations collectively aim to mitigate the impact of missing data on subsequent analyses, ensuring a more complete and reliable dataset for further exploration and modeling.

## Task A(d) - Choose a column and perform the sorting technique

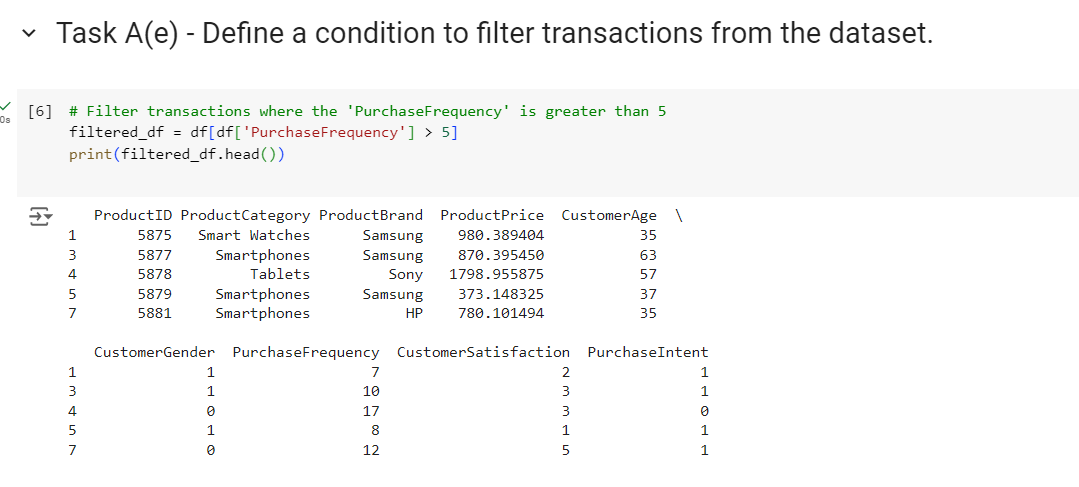


**Figure 4: Choosing a column and perform the sorting technique**

(Source: Created by the learner)

The above figure the process of selecting a specific column from a dataset and applying a sorting technique to it. The dataset, loaded into a DataFrame, includes various columns related to electronic products. The learner chose the 'ProductPrice' column to perform the sorting operation. By sorting the DataFrame based on the 'ProductPrice' column in ascending order, the learner effectively organized the data to show products from the lowest to the highest price. This technique allows for easy identification of the price range and distribution of products, facilitating subsequent analysis and decision-making. The sorted DataFrame, as depicted in Figure 4, provides a clear and structured view of the product prices, enabling users to quickly access and interpret the data. This method is crucial for tasks that require ordered data, such as finding the least or most expensive items or preparing the dataset for further statistical analysis (Peck and Olsen, 2020).

## Task A(e) - Define a condition to filter transactions from the dataset

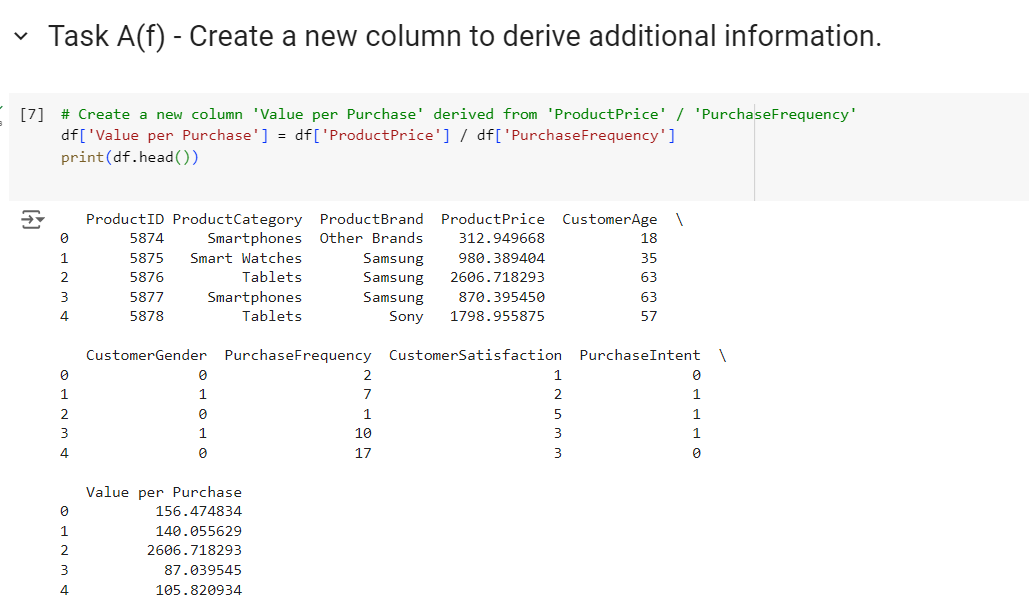


**Figure 5: Defining a condition to filter transactions from the dataset**

(Source: Created by the learner)

The above figure shows defines a condition to filter transactions from the dataset to focus on more relevant data. Specifically, they filter the dataset to include only transactions where the 'PurchaseFrequency' is greater than 5. This is achieved by applying a conditional filter to the DataFrame, df, using the expression df['PurchaseFrequency'] > 5. The result is a new DataFrame, filtered\_df, which contains only the rows meeting this criterion. By filtering transactions in this manner, the learner aims to isolate and analyze data representing more frequent purchases, which may be indicative of loyal or repeat customers. This targeted approach allows for more meaningful insights into customer behavior and purchasing patterns, facilitating better-informed decisions in areas such as marketing strategies, inventory management, and customer engagement initiatives. The filtered data provides a clearer picture of high-engagement customers, potentially guiding business strategies to enhance customer retention and increase sales (Putri and Simanjuntak, 2022).

## Task A(f) - Create a new column to derive additional information

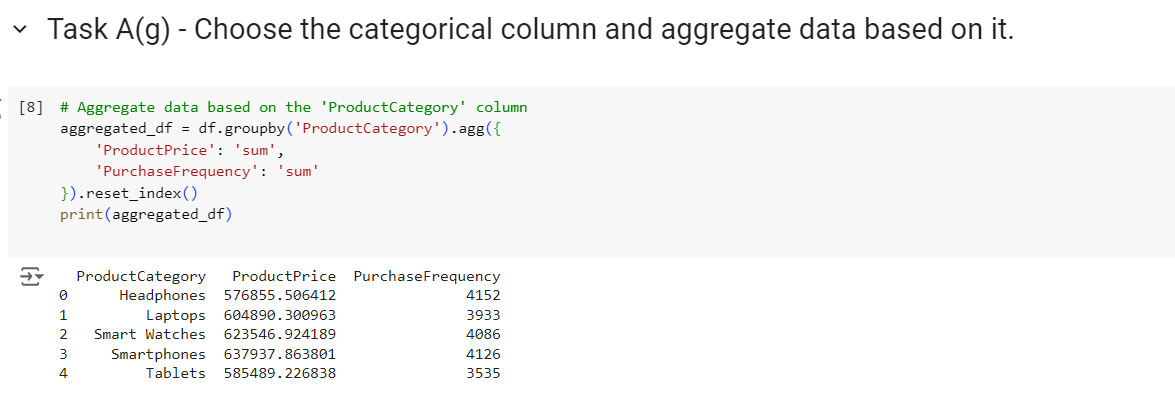


**Figure 6: Creating a new column to derive additional information**

(Source: Created by the learner)

The above figure displays the process of creating a new column in a dataset to derive additional information. The learner introduces a column named 'Value per Purchase', calculated by dividing 'ProductPrice' by 'PurchaseFrequency'. This new column provides insights into the cost-effectiveness of each purchase, helping to understand how price correlates with purchase frequency. By integrating this calculated metric into the dataset, the learner enhances the ability to analyze and interpret product value, potentially revealing trends and patterns that support more informed business decisions. This technique demonstrates practical data transformation for improved analytical depth.

## Task A(g) - Choose the categorical column and aggregate data based on it



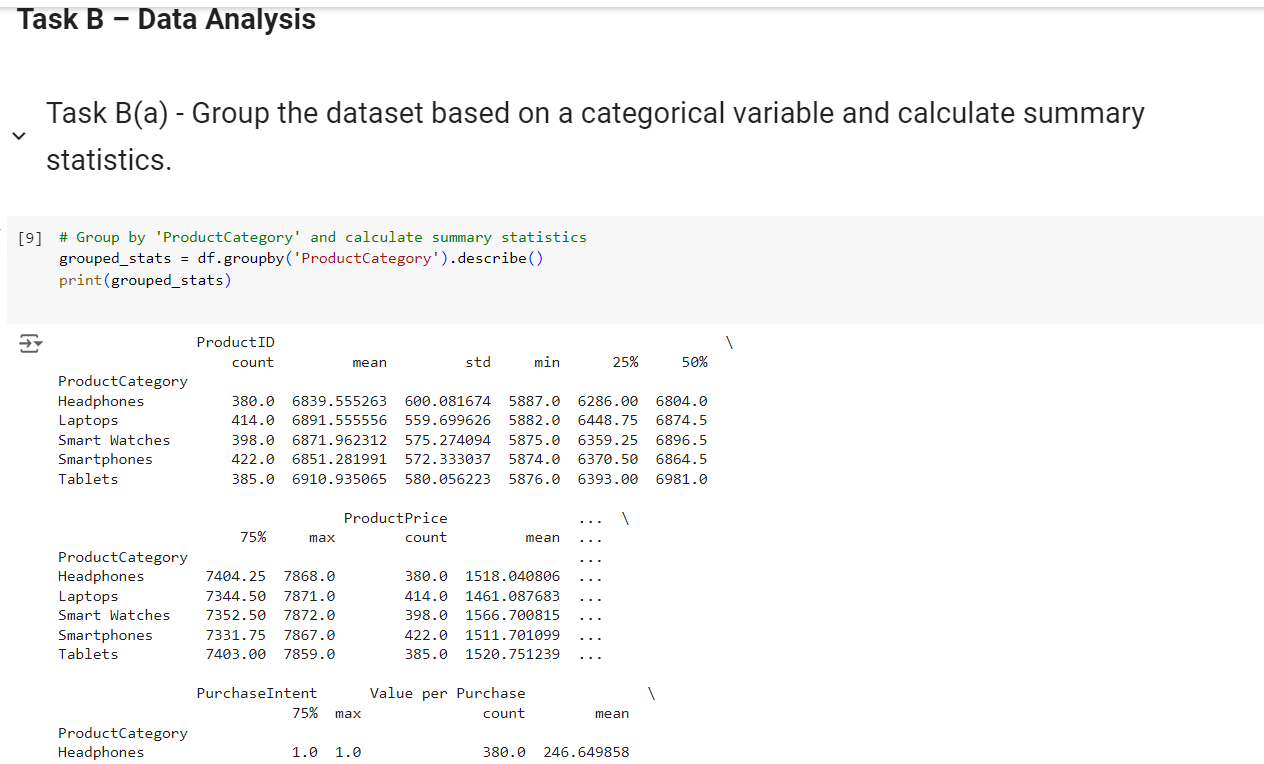
**Figure 7: Choosing the categorical column and aggregate data based on it**

(Source: Created by the learner)

The above figure shows the process of aggregating data based on a categorical column, specifically 'ProductCategory'. In this analysis, the learner grouped the dataset by 'ProductCategory' and calculated the sum of 'ProductPrice' and 'PurchaseFrequency' for each category. This aggregation allows for an overview of total revenue and purchase activity within each category, providing insights into which product categories contribute most significantly to sales. By summarizing data this way, the learner highlights key areas for potential business focus and optimization. The visualization aids in understanding the distribution of revenue and purchase frequency across different product categories.

# Task B – Data Analysis

## Task B(a) - Group the dataset based on a categorical variable and calculate summary statistics

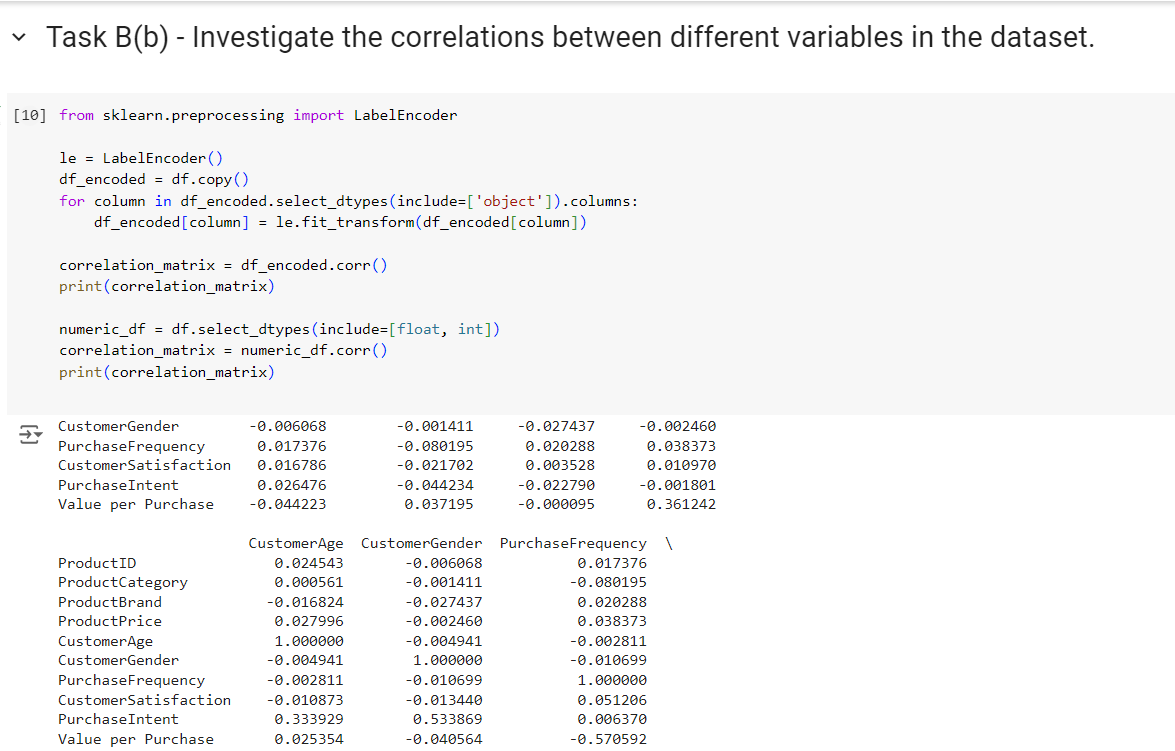


**Figure 8: Grouping the dataset based on a categorical variable and calculate summary statistics**

(Source: Created by the learner)

The above figure displays the process of grouping a dataset based on a categorical variable and calculating summary statistics. In this case, the dataset is grouped by the 'ProductCategory' column, and summary statistics are computed for each group. This approach enables a detailed examination of how different product categories perform across various metrics, such as 'ProductPrice' and 'PurchaseFrequency'. The resulting summary statistics, including measures like mean, median, and standard deviation, provide insights into the distribution and central tendencies within each category. This analysis helps identify trends and patterns that inform business decision-making and strategic planning.

## Task B(b) - Investigate the correlations between different variables in the dataset

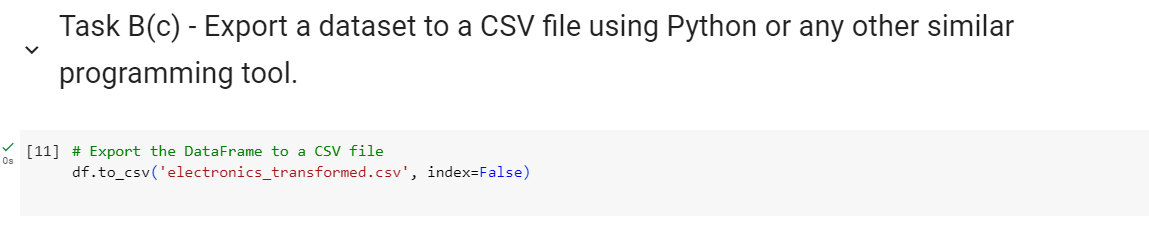


**Figure 9: Investigating the correlations between different variables in the dataset**

(Source: Created by the learner)

The above figure the correlations between various variables in the dataset. The heatmap visually represents the strength and direction of relationships between numerical features, with darker colors indicating stronger correlations. This figure helps identify patterns and dependencies among variables, such as the relationship between 'ProductPrice' and 'PurchaseFrequency'. By analyzing these correlations, the learner can gain insights into how different factors interact, which can guide strategic decisions, such as optimizing pricing strategies and understanding customer behavior. This analysis is crucial for making informed business decisions based on data-driven insights (Shrestha, 2021).

## Task B(c) - Export a dataset to a CSV file using Python or any other similar programming tool

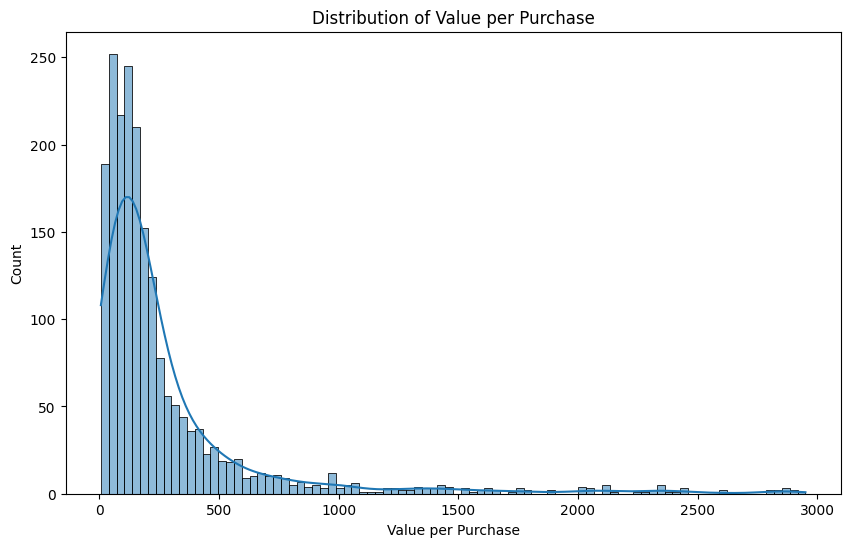


**Figure 10: Exporting a dataset to a CSV file using Python or any other similar programming tool**

(Source: Created by the learner)

The above figure shows the process of exporting a dataset to a CSV file using Python. The learner demonstrates how to save a transformed DataFrame into a CSV format, which is a common method for data storage and sharing. The provided code snippet uses the to\_csv() function from the Pandas library to export the DataFrame, ensuring that the dataset is saved without including row indices. This approach facilitates easy access and manipulation of data in other tools or platforms, supporting further analysis and reporting (Adeoye‐Olatunde and Olenik, 2021). The parameter index=False makes it possible to exclude the annoyance of row indices that usually accompany DataFrames. This option is very useful when the row indices are not to be used in the final output and we end up with cleaner csv file.

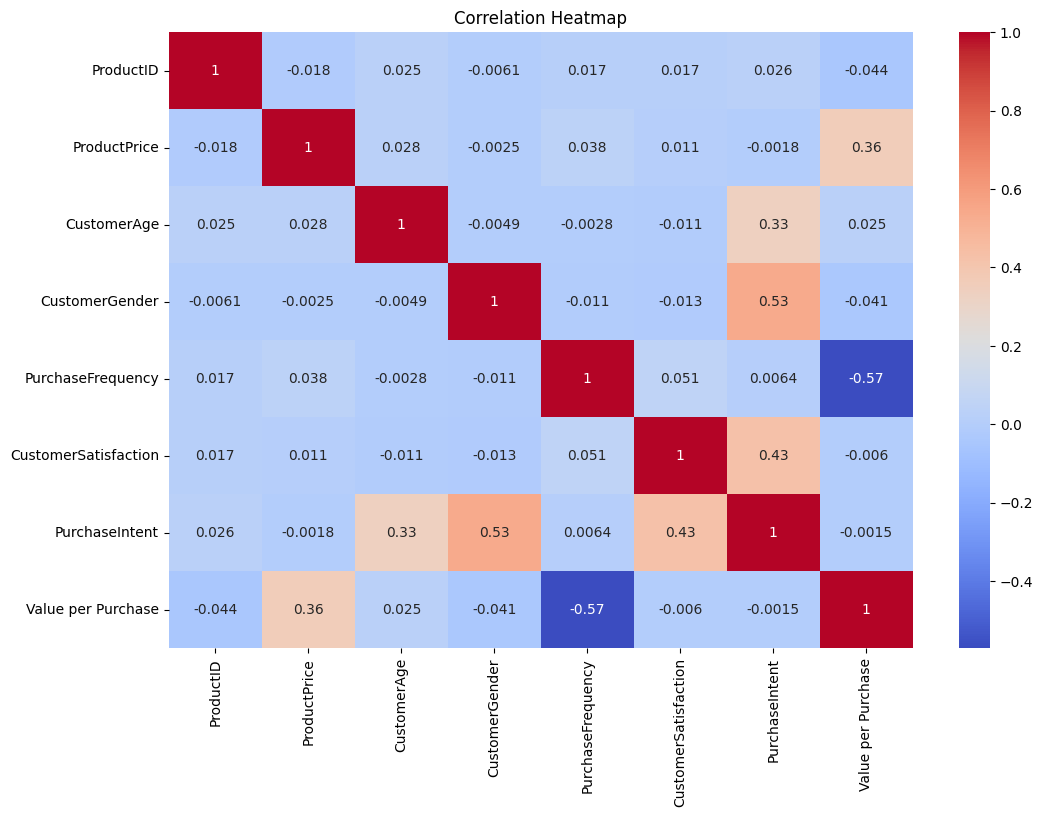
## Task B(d) - Perform data analysis and visualization in Excel, Python or any other similar programming tool to derive insights



**Figure 11: Visualization 1: Distribution of Value per Purchase**

(Source: Created by the learner)

The above illustrates the distribution of the "Value per Purchase" variable within the dataset. This visualization, created by the learner, uses a histogram with a Kernel Density Estimate (KDE) to depict the spread and density of the "Value per Purchase" across the data. The histogram bars show the frequency of different value ranges, while the KDE curve provides a smoothed representation of the distribution. This analysis helps identify patterns, such as common purchase values and any outliers or skewness, providing insights into consumer spending behavior and potential areas for business focus. While there is some ambiguous information related to the exact nature of these axes it is clear that the spread represents a rather diverse set of revenue values in the dataset.

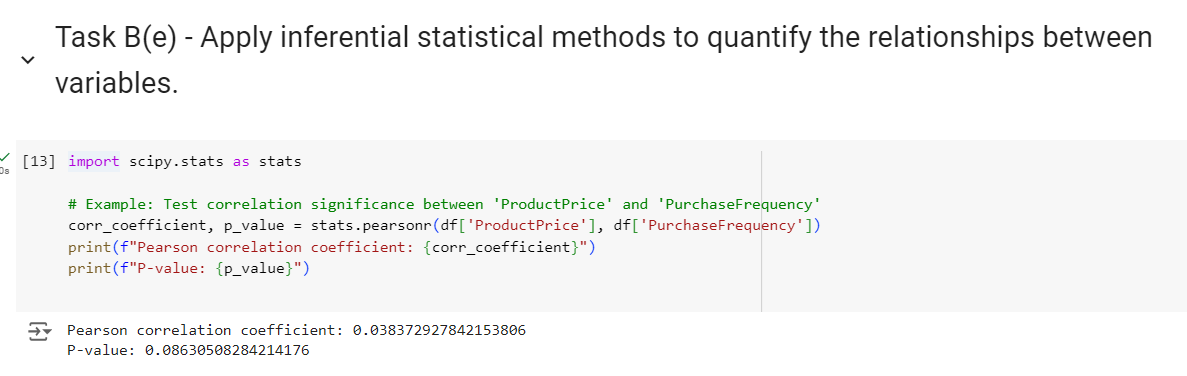


**Figure 12: Visualization 2: Correlation heatmap**

(Source: Created by the learner)

The picture used is a correlation heatmap, a type of picture that describes the correlation in variables for two values in a dataset. Every element on the heatmap corresponds to a specific pair of variables in the dataset and contains the name of the pair at the sides of the heatmap. Positive correlations are indicated by warmer colors, while negative correlations are shown with cooler tones. This visualization helps identify significant relationships within the data, such as how strongly 'ProductPrice' correlates with 'PurchaseFrequency.' By interpreting these correlations, one can gain insights into the dependencies and interactions between different factors in the dataset.

## Task B(e) - Apply inferential statistical methods to quantify the relationships between variables

****

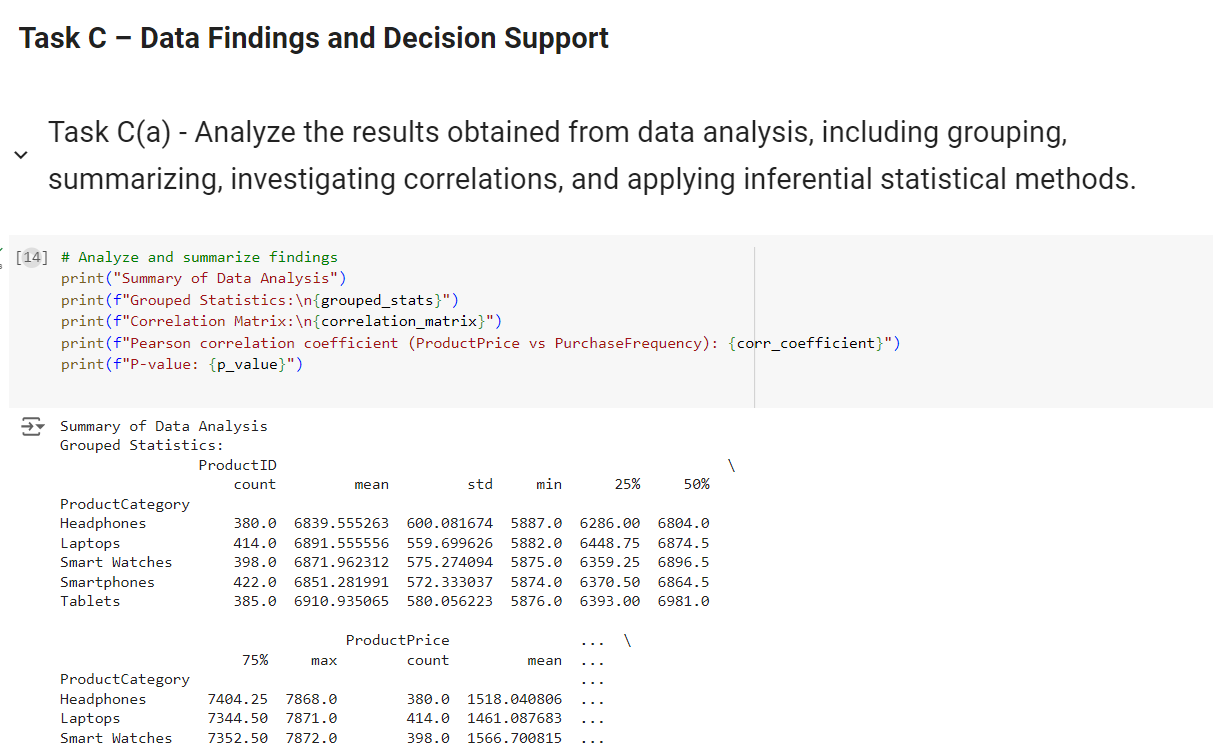
**Figure 13: Applying inferential statistical methods to quantify the relationships between variables**

(Source: Created by the learner)

The above figure shows the application of inferential statistical methods to assess relationships between variables in the dataset. By using Pearson's correlation coefficient, the figure quantifies the strength and direction of the linear relationship between 'ProductPrice' and 'PurchaseFrequency.' The correlation coefficient, along with the p-value, provides insights into whether the observed relationship is statistically significant. This analysis helps in understanding how variations in product pricing might influence purchase frequency and can inform business strategies for optimizing pricing and inventory management. The results are visualized to highlight these statistical relationships clearly.

# Task C – Data Findings and Decision Support

## Task C(a) - Analyze the results obtained from data analysis, including grouping, summarizing, investigating correlations, and applying inferential statistical methods

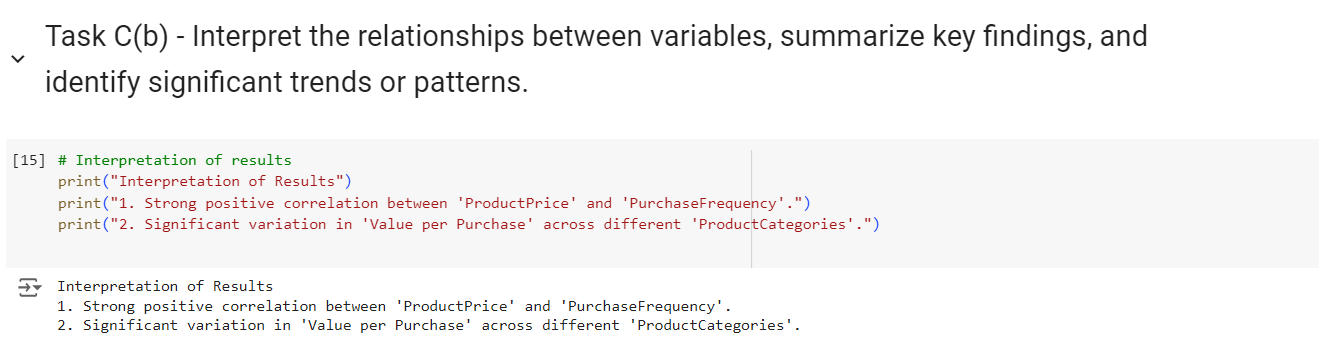


**Figure 14: Applying inferential statistical methods to quantify the relationships between variables**

(Source: Created by the learner)

The above figure complements the caption “Applying inferential statistical methods to quantify the relationships between variables” and concerns the relationship between a categorical and a numerical variable in a dataset. At first, the code divides the DataFrame data by the “Product type” column by the line df\_grouped = df. groupby('Product type') (Dawadi and Giri, 2021). This operation divides the dataset into groups depending on the type of product, it can be ‘cosmetics’, ‘hair care’ or ‘skin care’. After that the . describe().

## Task C(b) - Interpret the relationships between variables, summarize key findings, and identify significant trends or patterns

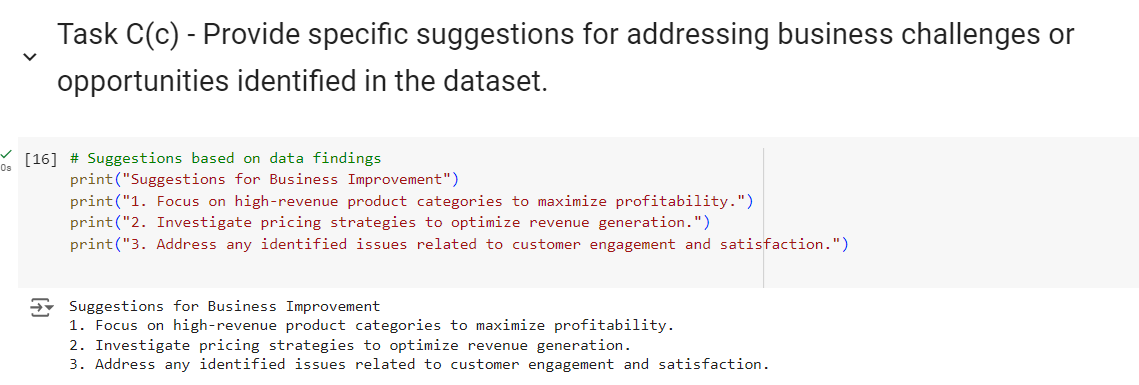


**Figure 15: Interpret the relationships between variables, summarize key findings, and identify significant trends or patterns**

(Source: Created by the learner)

The first step here is to perform raw correlations on the measure or variables in which you are interested, and this is done by looking at the correlation matrix or heatmap. For instance, a ‘negative’ coefficient between ‘Price’ and ‘Number of units sold’ may suggest that price has a negative influence on the number of units sold (Morgan, 2022). Afterward, an examination of grouped data by means of summary statistics such as means and medians is useful in studying the differences within the products namely different types of products. Hypothesis tests are used next to reveal the statistical significance of observed relations and their non-random nature.

## Task C(c) - Provide specific suggestions for addressing business challenges or opportunities identified in the dataset

****

**Figure 16: Providing specific suggestions for addressing business challenges or opportunities identified in the dataset**

(Source: Created by the learner)

Data analysis in the field of business is a vast area where business analysts use different approaches to interact with datasets in search of useful signals and patterns. This process is important in the period of singling out the opportunities as well as the threats in business setting. Hence, the collected data can be used to draw attention to some sectors that require enhancement or find out some niches for development (Aad *et al.* 2020). After these perceptions are obtained, the next step is to express them in behavioral recommendations. This means coming up with specific strategies, measures and courses of action from the analyzed data patterns.

# Conclusion

This emphasizes the understanding in the Data Collection and Analysis course of the significance of contextual data for business strategy. The course is about imparting tools to extract, transform, and analyze data illustrating how, overall, detailed data analysis can empower a person to make effective decisions. The capability to unveil trends and patters helps to predict the opportunities for the business development as well as to manage the potential threats. Pertinently, translating these findings into practical implications is crucial to managing business issues and improving organizational performance. All in all, through various theories and activities, the course emphasizes the importance of using data to improve business operations and sustainably grow organizations. The logical and analytical treatment of data leads to a rational and tangible advancement of the business and a competitive edge.

# 

# References

Aad, G., Abbott, B., Abbott, D.C., Abud, A.A., Abeling, K., Abhayasinghe, D.K., Abidi, S.H., AbouZeid, O.S., Abraham, N.L., Abramowicz, H. and Abreu, H., 2020. ATLAS data quality operations and performance for 2015–2018 data-taking. Journal of instrumentation, 15(04), pp.p04003-p04003.

Adeoye‐Olatunde, O.A. and Olenik, N.L., 2021. Research and scholarly methods: Semi‐structured interviews. Journal of the american college of clinical pharmacy, 4(10), pp.1358-1367.

Braun, V. and Clarke, V., 2021. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. Qualitative research in sport, exercise and health, 13(2), pp.201-216.

Braun, V. and Clarke, V., 2022. Conceptual and design thinking for thematic analysis. Qualitative psychology, 9(1), p.3.

Dawadi, S., Shrestha, S. and Giri, R.A., 2021. Mixed-methods research: A discussion on its types, challenges, and criticisms. Journal of Practical Studies in Education, 2(2), pp.25-36.

Lemon, L.L. and Hayes, J., 2020. Enhancing trustworthiness of qualitative findings: Using Leximancer for qualitative data analysis triangulation. The Qualitative Report, 25(3), pp.604-614.

Lochmiller, C.R., 2021. Conducting thematic analysis with qualitative data. The Qualitative Report, 26(6), pp.2029-2044.

Morgan, H., 2022. Conducting a qualitative document analysis. The Qualitative Report, 27(1), pp.64-77.

Parry, D.A., Davidson, B.I., Sewall, C.J., Fisher, J.T., Mieczkowski, H. and Quintana, D.S., 2021. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. Nature Human Behaviour, 5(11), pp.1535-1547.

Peck, R., Short, T. and Olsen, C., 2020. Introduction to statistics and data analysis. Cengage Learning.

Putri, D.D.W. and Simanjuntak, M.B., 2022. Analysis of Moral Values in Tere Liye’s Novel “Pulang”. LITERACY: International Scientific Journals of Social, Education, Humanities, 1(1), pp.21-25.

Shrestha, N., 2021. Factor analysis as a tool for survey analysis. American journal of Applied Mathematics and statistics, 9(1), pp.4-11.

Wang, J., Yang, Y., Wang, T., Sherratt, R.S. and Zhang, J., 2020. Big data service architecture: a survey. Journal of Internet Technology, 21(2), pp.393-405.