Exploring descriptive and inferential statistics using Python for data-driven decision-making

Mohomed Shaveen Hassim

240188980

Submitted in accordance with the requirements for the degree of
MSc Data Science

Under the Supervision of ‘Dr. Swathi Ganesan

York St John University

Computer and Data Science

September 2024

I confirm that the work submitted is my own and that appropriate credit has been given where reference has been made to the work of others.

Self-declaration on Academic Integrity:

I have read and understood the Academic Misconduct statement.	Tick to confirm ☒
I have read and understood the Generative Artificial Intelligence use statement.	Tick to confirm ☒
I am satisfied that I have met the Learning Outcomes of this assignment (please check the Assignment Brief if you are unsure)	Met ☒

Self-Assessment – If there are particular aspects of your assignment on which you would like feedback, please indicate below. Optional for students
Suggested prompt questions- *How have you developed or progressed your learning in this work?* *What do you feel is the strongest part of this submission?* *What feedback would you give yourself?* *What part(s) of this assignment are you still unsure about?*

Acknowledgement

I am deeply grateful to my supervisor, Dr. Swathi Ganesan, who provided essential supervision and guidance throughout this research project. I owe sincere gratitude to both my peers and faculty members for their persistent feedback and support that shaped this report. I deliver sincere appreciation to both my family and friends because of their consistent encouragement and understanding.

List of Abbreviations

- **ANOVA –** Analysis of Variance
- **B2B** – Business-to-Business
- **B2C** – Business-to-Consumer
- **EDA** – Exploratory Data Analysis
- **GBP** – Great British Pound
- **H_0** – Null Hypothesis
- **H_1** – Alternative Hypothesis
- **R^2** – Coefficient of Determination
- **UK –** United Kingdom

Abstract

This research investigates the application of descriptive and inferential statistical techniques in the context of online retail analytics. The objective utilizes transactional information for studying customer behavioural patterns while detecting nationally based patterns, which provide valuable guidance for business decision processes. The research utilizes the Online Retail dataset supplemented with data preprocessing and Visualisation, then completes statistical testing through t-tests and ANOVA, and chi-square tests. The analysed data demonstrates vital customer buying habits together with substantial variations in purchasing amounts across different market regions. The interpretability of insights becomes easier to understand with visual representations. Business strategies can be developed using the implemented data analysis methods. This preliminary study forms a base for additional predictive modelling and real-time analytics, even though it faces data coverage restrictions and real-time variable absence.

This research is unique in that it fills an observable gap in the current literature, as both descriptive and inferential statistical tools are used in tandem with a Python-based implementation, due to the emphasis on application in the business field. The research delivers a reproducible and extensible analytical framework with the compatibility of the core libraries, Pandas, NumPy, SciPy, and Seaborn. Unlike earlier research, which was frequently based its outcomes upon the methods of clustering or regression alone, this article uses the concept of hypothesis testing to confirm the level of statistical significance and is therefore more valid. The resultant methodology not only allows businesses to find patterns but also to test their assumptions, and a culture of evidence-based decision-making is allowed to flourish. The paper also shows how easy-to-use tools can be used successfully in order to come up with valuable conclusions out of complicated sets of retail data.

Keywords: Online Retail, Descriptive Statistics, Inferential Statistics, Data Visualisation, Consumer Behaviour, Hypothesis Testing

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background of the Study

Data-driven decision-making exists today as a fundamental requirement for contemporary business operations. Organizations employ statistical approaches to process and investigate large data collections and generate practical insights that lead to superior decisions. The analysis process depends heavily on descriptive and inferential statistics because these methods lead to data summaries and interpretation. Data characteristics and distribution patterns emerge from descriptive statistics, yet inferential statistics enable researchers to generalize findings through hypothesis testing **(Moore et al., 2021).**

Retail operations have undergone rapid digitization during the modern era of global commerce, which has produced substantial transactional data collections. Business decisions and operational efficiency improvements, and better customer interaction result from proper analysis of accumulated data **(Laudon and Traver, 2021).** Online retailing, in particular, offers a compelling domain for statistical and analytical research due to its inherently data-rich environment. Businesses can both identify new customer patterns and create predictive models to strengthen their marketing strategies and supply chain through analysis of these data collections **(Chong et al., 2017).**

Data science partnerships with analytics have strengthened modern business operations to achieve competitive advantages over the past decade. Organizations have developed extract patterns, outcome prediction, and informed decision competence, which have become essential organizational capabilities **(Provost & Fawcett, 2013**). Python is a popular programming language in data science. It delivers an extensive collection of tools that support statistical analysis. Data manipulation, Visualisation tools, and hypothesis testing functions can be performed through Pandas combined with NumPy, SciPy, Matplotlib, and Seaborn libraries **(McKinney, 2022).**

1.2 Problem Statement

Despite the increasing adoption of data-driven methodologies, many organizations still struggle with effectively utilizing statistical tools. Business professionals often face challenges in understanding the application of descriptive and inferential statistics, leading to suboptimal decision-making **(Hair et al., 2020**). Additionally, the lack of expertise in Python-based statistical analysis presents a barrier to fully harnessing its capabilities. This research aims to bridge the gap by providing a structured approach to using Python for statistical analysis, specifically in the domain of business analytics.

1.3 Research Aim and Objectives

The research demonstrates how Python can use descriptive and inferential statistics to establish data-driven business decisions by analysing retail data. The research applies statistical methodologies to online retail dataset analysis to decode trends and customer comparison metrics and maximize business practices. Strategic business decisions made through empirical evidence become possible because businesses use rigorous statistical methods **(VanderPlas, 2019).**

The research aims to:

- Illustrate the importance of descriptive and inferential statistics in data-driven decision-making.
- Demonstrates successful implementation of statistical techniques through Python programming language for effective results.
- Evaluate how statistical analysis enhances strategic decisions from a business perspective.

1.4 Research Hypotheses

The following hypotheses are focused on investigating the extent of statistical significance and character of interrelationships in an online retail dataset.

Shapiro-Wilk Test: -

- H_0: The Quantity or UnitPrice data is normally distributed.
- H_1: The Quantity or UnitPrice data is not normally distributed.

t – Test: -

- H_0: There is no significant difference in the mean UnitPrice between two selected countries (e.g., the United Kingdom and Germany).
- H_1: There is a significant difference in the mean UnitPrice between the two countries.

ANOVA: -

- H_0: The mean Quantity sold is the same across multiple selected countries.
- H_1: At least one country has a different mean Quantity sold.

1.5 Research Outline

This report has been arranged into five chapters, with each of the chapters focusing on an important step in the research process:

Chapter 1 (Introduction) - Introduction provides a presentation of the business situation, research question, goals, the study purpose, theoretical framework of the study, and methodological justification of the use of the Python-based statistics techniques in retail analytics.

Chapter 2 (Literature review) - Prior to academic and industry literature, surveys available on the topics of descriptive and inferential statistics, the usage of Python in data analytics, and comparative studies that form the basis of the current methodology are reviewed.

Chapter 3 (Methodology) - Methodology Ensures research design, explains dataset features, stipulates data preprocessing steps, and provides reasons for Python libraries usage with specifying the statistical test selection (Shapiro-Wilk, t-test, ANOVA, regression).

Chapter 4 (Results and Analysis) – This chapter discusses descriptive statistics, test of hypothesis results, regression outputs, and data Visualisations. Managerial decision-making is put in context.

Chapter 5 (Conclusion and Recommendation) – Discuss the limitations, recommendations, and future work, and summarise the key findings.

1.6 Significance of the Study

The knowledge developed in this context demonstrates to managers how Python-driven statistical methods can be incorporated into their daily activities and eventually improve performance, customer-specific focus, and revenue optimization. The results provide practical statistical views in making informed decisions.

In the current data-driven economy, it is no longer the raw data that is the differentiator, but rather transforming that raw data into actionable information that makes the difference between the champions and the laggards. This study fulfils that requirement because it offers a pragmatic framework using Python to combine descriptive and inferential statistics. It is in contrast with the traditional business-based reports that rely on descriptive summaries or trend lines exclusively and utilize hypothesis testing methods, comparative analysis, and regressions to reveal statistically significant patterns that form the foundation of evidence-based approaches.

This change enables organizations to migrate away from intuitional decision-making to decisions that have a scientific basis.

Moreover, Python integration offers the additional benefit of reducing and eliminating human error through automation, while also enhancing scalability and repeatability, properties essential for modern data analysis in rapidly changing business contexts. Since the strategy utilizes open-source libraries, it becomes affordable and accessible to both small and medium-sized enterprises (SMEs) and major corporations.

The paper addresses a notable gap and demonstrates that the Python statistical ecosystem has the potential to transform theoretical statistics into practical business applications. It is also used to make future data professionals know how to build end-to-end analytical flows that remain interpretable and reproducible. After all, the ultimate objective is to create data-literate organizations that value precision, vision, and tactical forward-thinking.

1.7 Conclusion

Business analytics has become another essential pillar of modern business, more so in the online retailing world. Despite the sharp shift to data-oriented operational patterns, many enterprises are limited by the lack of statistical reasoning and coding expertise. The current study aims to solve this disparity by linking descriptive and inferential statistical research techniques, including most prominently those integrated through Python, with the perspective of building compositional, increasingly analytical pipelines. This way, the work creates the empirical foundation to support the argument that Python-based frameworks can be appropriately used to strengthen the decision-making and performance of managers through generating evidence-based knowledge.

Chapter 2: Literature Review

2.1 Introduction to Literature Review

In recent years, statistical methods in business decision-making have grown a lot. Big data and real-time analytics have brought about significant changes in how organisations access and work with data (Ghasemaghaei & Calic, 2021). At this stage, understanding and using descriptive and inferential statistics support extracting knowledge, making predictions, and confirming scientific theories in many business sectors. Using Python in data science, companies find it easy to apply these techniques because of its strong computing capabilities and wide range of libraries.

In this chapter, empirical findings are presented that illustrate the value and ways to use descriptive and inferential statistical methods in retail analytics. It also aims to evaluate how data analytics in business involves Python tools and approaches.

2.2 Empirical Studies and Methodologies

Modern studies using empirical methods have demonstrated how descriptive statistics help to summarise large collections of transactions. Pizzi et al. (2020) conducted a study on how businesses use central tendency techniques to recognise the purchasing habits of consumers. They observed that common statistical indicators, including mean transaction value and standard deviations, reflect actionable details about behaviour in e-commerce.

At the same time, using inferential statistics is necessary for confirming ideas and reaching judgments that apply outside the sample pool. The study, led by Kumar et al. (2022), used t-tests and ANOVA to measure customer satisfaction by regional market. Based on their findings, there were important variations noted that were vital for local marketing plans.

Its ability to support complete analysis workflows is the main reason Python is becoming more popular in data analytics. Sarker et al. (2021) pointed out in a comparative study that Pandas, SciPy, and Seaborn in Python offer both simple data analysis and the ability to test theories. It showed that business analysts can now use Python to gain insights, make better decisions, and improve data transparency and repeatability.

In addition, research that marries statistics with business strategy appears to improve how decisions are made day to day in organisations. Nguyen and Tran (2023) studied data from retail businesses using Python to figure out how sales numbers were shaped by customer demographics. Integrating regression analysis and chi-square testing allowed them to identify which variables had a clear impact on how many times people made purchases.

According to research by Alwan et al. (2021), dashboards based on Python tools are helpful in visualising trends and anomalies to users in descriptive analytics. With Matplotlib and Seaborn, generating visualisations like heatmaps and boxplots allows executives to quickly understand the data.

People have also looked into bringing Python into business statistics education. According to Sharma and Singh (2022), students who studied statistical methods through Python code were better at analysing business cases using data, as opposed to those who learned only theoretical statistics.

Following these recent studies, using Python as an analytical platform allows retail analysts to work with both types of statistics. They also suggest using data preprocessing, normality tests, validating hypotheses, and visualising the information in analyses, which is what was done in this research.

2.3 The Role of Statistics in Business Decision-Making

Business operations and strategic planning have used statistical methods as their fundamental operational foundation since their inception. Business organizations use descriptive statistics to identify patterns and trends in their datasets through measurement components like mean, median, variance, and standard deviation. These measures help businesses form evidence-based decisions by providing essential data characteristics and information **(Moore et al., 2021).** Data analysis for retailers relies on descriptive statistics to handle sales information through the identification of top products and future market assessment.

Businesses use inferential analysis to transform specific information from tested samples into predictions for wider populations while conducting hypothesis testing and generating predictive models. The commonly utilized statistical reduction tools in decision-making processes include regression analysis, hypothesis testing, and confidence intervals, according to (**Hair et al. 2020).** For example, businesses deploy regression analysis for marketing trend prediction, and they conduct A/B testing to evaluate marketing campaign performance **(Anderson et al., 2020).**

Modern business decision-making has seen substantial advancement in data analytics integration over the past decade. Big data availability has enabled organizations to acquire vast data collections so statistical analysis serves as a crucial method for generating useful business insights **(Provost & Fawcett, 2013).** Organizations use statistical methods to maximize their marketing approaches and they apply the approaches to improve both their supply chain systems and user experience. The availability of complex statistical tools remains underused by various organizations because they lack the appropriate technical expertise **(Hair et al., 2020).**

2.4 Python in Statistical Analysis

The programming language Python has gained popularity in statistical analysis because it offers users simplicity, together with versatility and a strong collection of available libraries. Data handling for large datasets through the Pandas library works in conjunction with NumPy libraries, which provide efficient numerical performance calculations. Using SciPy, one can access advanced statistical functions while creating Visualisations through Matplotlib and Seaborn libraries **(VanderPlas, 2019).** The integration of statistical analysis through Python enables businesses to select it as their preferred platform for better data-driven decision-making.

Through automation, Python helps perform data analysis tasks, which decreases human mistakes and increases result reproducibility **(McKinney, 2022).** Companies execute Python scripting code to enforce standardized data cleansing procedures on large datasets, which results in precise analysis outcomes. Python remains accessible to new and experienced users since it has extensive documentation alongside an active community that enables widespread industry adoption **(Yin, 2020).**

2.5 Applications of Descriptive and Inferential Statistics

The field of business intelligence heavily relies on descriptive statistics because these tools help organizations reveal customer behaviour patterns alongside market pattern identification. Retailers access sales information to recognize best-selling products while creating demand predictions. Understanding central tendencies of the data comes from mean, median, and mode measurement methods, while data distribution information emerges from variance and standard deviation **(Moore et al., 2021).** Analytics through these measures allows users to detect directional patterns as well as make data-driven decisions.

Through inferential statistics, businesses can test hypotheses and perform regression analysis, thus validating assumptions and determining performance-influencing elements **(Hair et al., 2020).** Businesses apply A/B testing to measure marketing campaign success and combine regression analysis with future sales prediction **(Anderson et al., 2020).** Businesses obtain evidence-supported information through these methods, which leads to better decision quality and minimises avoidable errors.

2.6 Comparative Study of Research Articles

The table below summarizes relevant research articles in statistical analysis and Python applications:

Author(s)	Statistical Techniques Used	Python Tools/Libraries Used	Application Area	Key Findings / Contribution
Kumar, Patel, and Mehta (2022)	ANOVA, t-test	Pandas, SciPy, Seaborn	Retail market segmentation	Determine customer groups and their related expenditure differences using inferential statistics
Nguyen and Tran (2023)	Hypothesis testing, EDA, correlation analysis	Pandas, Matplotlib, Seaborn	Customer behaviour analytics	Created a Python-driven framework for real-time behaviour modelling using historical data in retail
Ali and Rahman (2021)	Descriptive analysis	Pandas	General retail analytics	Reveal a gap between statistical evidence and realistic retail decisions, suggesting the importance of aligning business
Zhang and Lee (2022)	Regression, clustering	Scikit-learn, NumPy	Retail prediction modelling	Focused on the usefulness of predictive analytics but lacks a formal hypothesis-based framework or inferential testing

Table 1: Comparative Study of Relevant Research Articles

A comparative overview of the recent literature explains a range of methodological options in the field of applied business and retail analytics, as summarised in Table 1. Most researchers have taken advantage of Python-based libraries such as Pandas, SciPy, and Seaborn to perform both descriptive and inferential analysis. As an example, Kumar et al. (2022) employed ANOVA and t-tests to divide retail markets and analyse the variability of purchasing, thus proving the practical value of inferential statistics in business segmentation. Nguyen and Tran (2023) are equally notable and combine hypothesis testing and data visualisation to simulate customer behaviour and thus depict the versatility of Python to perform real-time retail analytics.

Along with these developments, however, there appears to be a conspicuous gap: a relative dearth of end-to-end workflows, which integrate descriptive exploration with inferential validation in a reproducible Python environment. A significant part of the work already done focuses on either exploratory data analysis or predictive modelling without a unified hypothesis-driven approach (Ali and Rahman, 2021; Zhang and Lee, 2022). This kind of dichotomisation of methodology creates a gap between statistical theory and managerial relevance in the short run. Also, not many studies support statistical results using clear decision rules, thus limiting the usefulness of their results in fluid retail settings.

The study explores the weaknesses of past retail analytics research studies by conducting an in-depth and hypothesis-based statistical analysis entirely in Python, using online retail data. The current work provides a convenient platform by integrating distribution tests, the analysis of comparative means, and inferential processes into one repeatable workflow, which provides a convenient structure of evidence-based decision-making, thus closing the gap between academic and managerial communities.

2.7 Research Gap and Contribution

Numerous organizations employ statistical methods, yet many of their business practitioners lack sufficient proficiency to execute these techniques properly. Most studies fail to unite descriptive statistical methods, inferential statistical methods, and the creation of tailored business-relevant visual storytelling approaches. The majority of research depends on clustering and regression methods while skipping hypothesis tests for verifying significance **(Kumar and Singh, 2019; Ahmed et al., 2021).** This research aims to address this gap through practical instructions about using Python for statistical analysis. The study applies descriptive and inferential statistics to business data to prove how Python facilitates evidence-based decision-making. Businesses can use the obtained results to develop concrete strategies that will boost their analytical capacity and enhance their decision-making capabilities.

Ongoing theory-practice gaps have been one of the major barriers to the successful realisation of statistical concepts in business analytics education and practice. Even though the existing literature often isolates the theoretical constructs and their Pythonic representations, such determinacy yields shallow conclusions and hinders the establishment of general analytical abilities (Sharma and Singh, 2022). This paper seeks to fill this lack by taking Python to be a unifying lens - the lens with which there is both theoretical rigor and computational scale. The study is a considerable contribution to the operational expertise of business professionals since they acquire practical skills and develop the data-driven mindset that is necessary in the current competitive environment through the provision of reproducible Python code and annotated analytical workflows (Sarker, Kayes, and Watters, 2021).

2.8 Conclusion

The research delivers valuable findings through the unification of statistical and interpretive approaches in retail analytics. This chapter provided a detailed review of research strategies in e-commerce analysis as it explored prominent techniques with their applicable methods. Recent research gaps and consistencies allow researchers to build statistical data-driven approaches that generate business-focused insights. This study justifies the selection of analytical methods mentioned in the methodology chapter.

Chapter 3: Methodology

3.1 Introduction to Methodology

The techniques used in this chapter show how descriptive and inferential statistics, applied with Python, help make decisions in retail. The research methodology follows a quantitative approach because it relies on numbers, statistics, and testing to monitor changes in customer shopping trends.

For the study, hypotheses were put forward, based on both theory and research that had been previously tested in practice. It permits study results to be widely applied to the population of retail consumers (Bryman, 2021). The decision to analyse data with Python indicates a greater focus on automation and reproducibility in today's statistical analysis. The process is designed to keep data accurate, ensure valid analysis, and connect the process to real-life business challenges.

3.1.1 Shapiro-Wilk Test

The Shapiro-Wilk Test is a highly rated test that checks the conformity of a data set to normality. It can be most useful in cases of small- and moderate-sized samples. This algorithm calculates a W-statistic, and a value of zero to near one indicates that it is a plausibly normal distribution. The formula is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})}$$

Figure 1: Shapiro-Wilk Test Formula

Where,

- x(i) are the ordered values of the sample.
- x¯ is the mean of the sample.
- a_i are Order statistics of a normally distributed sample - when properly indexed - can be used to obtain a set of constants which are functionals of the corresponding means, the variances, and the covariances.

When the p-value is smaller than the standard level of significance (typically < 0.05), it indicates the rejection of the null hypothesis of normality. This has led practitioners to use the

Shapiro-Wilk test in deciding whether to employ parametric or non-parametric analytical methods.

3.1.2 t-Test

The independent samples t-test examines a statistic that determines whether the sample means of any two independent samples are statistically different. The null hypothesis states that the means of the two groups are equal. The formula to calculate the t-test is,

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

Figure 2: t-Test Formula

Where:

- $\bar{x}_1$ and $\bar{x}_2$ are the means of the sample.
- n_1 and n_2 are the sizes of the sample.
- S^2_p is the pooled variance, which is computed as:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Figure 3: Pooled Variance Formula

The variance is commonly applied to analyse purchasing behaviour or price behaviour across geographic or demographic boundaries (Anderson et al., 2020). In the case where the resulting p-value is less than 0.05, it can be stated that there exists a significant difference in the two means.

3.1.3 ANOVA (Analysis of Variance)

Analysis of variance (ANOVA) determines how significantly different the means of three or more independent groups are. In ANOVA, the null hypothesis is that all the group means are the same. ANOVA statistic is computed as F, using the equation below:

$$F = \frac{MSB}{MSW} = \frac{SSB/dfb}{SSW/dfw}$$

Figure 4: ANOVA Formula

$$\mathrm{MSW} = \frac{\mathrm{SSW}}{N-k} \quad \text{and} \quad \mathrm{MSB} = \frac{\mathrm{SSB}}{k-1}$$

Figure 5: MSW and MSB Formulas

Where:

- SSB: Sum of squares between groups
- SSW: Sum of squares within groups
- k: Number of groups
- N: Total number of observations

High F ratios and small p-values signify that one difference in the means is significantly different from the others (Moore et al., 2021). This step is usually repeated in retail environments to compare the sales of similar products in different countries or regions.

3.1.4 Linear Regression Analysis

Linear regression describes the correlation between a dependent variable and one or more independent variables. Simple linear regression is applied in this study to determine the influence of unit price on quantity purchased. The regression equation is:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

Figure 6: Regression Formula

Minimising the sum of squared residuals gives the coefficients of a simple linear regression model. To understand the degree of correlation between the dependent and independent variables, the R² statistic is used, which denotes the percentage of variance in the dependent variable that has been explained by the regressor (Hair et al., 2020).

A number of scientific libraries in Python support the agile implementation of these processes, namely, scipy.stats, statsmodels, and sklearn.linear_model can be used to perform an accurate calculation with a quick response (VanderPlas, 2019; McKinney, 2022).

3.2 Research Approach

This research relies on a quantitative method that analyses and explains statistics from an online retail dataset. This approach is supported by its powers to discover patterns, check research ideas, and provide general knowledge through numerical analysis (Creswell & Creswell, 2022).

Several major phases make up the research process:

- **Data Collection and Preprocessing:** Using the Online Retail dataset, which is freely accessible and has more than 500,000 records, data is collected from transactions. Before analysing a dataset using statistics, it must be cleaned and preprocessing takes care of the missing information, changes data types, finds outliers, and eliminates duplicates (Zhou et al., 2021).
- **Exploratory Data Analysis (EDA)**: The process aims to discover the overall features and initial patterns in the data. To summarise customer behaviour in different dimensions, researchers use descriptive statistics such as mean, median, and standard deviation (Pizzi et al., 2020). With the help of Python libraries like Matplotlib and Seaborn, it is easy to build visualisations like boxplots, histograms, and correlation heatmaps.
- **Statistical Hypothesis Testing**: With statistical hypothesis testing, inferential techniques are applied to confirm or reject assumptions and to measure significance. Examples of tests are the Shapiro-Wilk test to check for normality, independent samples t-tests when comparing the means of two groups, and ANOVA when comparing more than two groups.
- **Regression Analysis**: Linear regression is used to examine how unit prices relate to the amount sold, giving insight into what causes price shifts and how the market responds (Kumar et al., 2022).
- **Python-Based Implementation**: Python was used because it scales well, is simple and has many important data science libraries like Pandas, NumPy, SciPy, StatsModels and Scikit-learn (Sarker et al., 2021). Python makes it possible to repeat and automate everything from preprocessing to testing a hypothesis in the analytical pipeline.

3.3 Research Questions and Hypotheses

Research Questions: -

1. How does the decision-making procedure benefit from descriptive and inferential statistics?

2. How does Python help with statistical analysis operations for business needs?

3. How does statistical analysis benefit from utilizing libraries available in Python?

4. To provide recommendations on how these statistical methods can be applied more effectively in real-world business scenarios to make effective data-driven decision-making.

3.4 Hypotheses and Their Purpose

1. **Shapiro-Wilk Test**
   - H_0: The Quantity or UnitPrice data is normally distributed.
   - H_1: The Quantity or UnitPrice data is not normally distributed.
   - **Purpose**: Which statistical analytical method between parametric and non-parametric will best suit **(Anderson et al., 2020).**

2. **t-Test**
   - H_0: There is no significant difference in the mean UnitPrice between two selected countries (e.g., the United Kingdom and Germany).
   - H_1: There is a significant difference in the mean UnitPrice between the two countries.
   - **Purpose**: Understanding pricing variations between countries helps businesses develop effective pricing methods and market their products across regions **(Hair et al., 2020).**

3. **ANOVA**
   - H_0: The mean Quantity sold is the same across multiple selected countries.
   - H_1: At least one country has a different mean Quantity sold.
   - **Purpose**: Determines local market product trends for inventory management purposes **(Moore et al., 2021).**

3.5 Block Diagram of Research Methodology

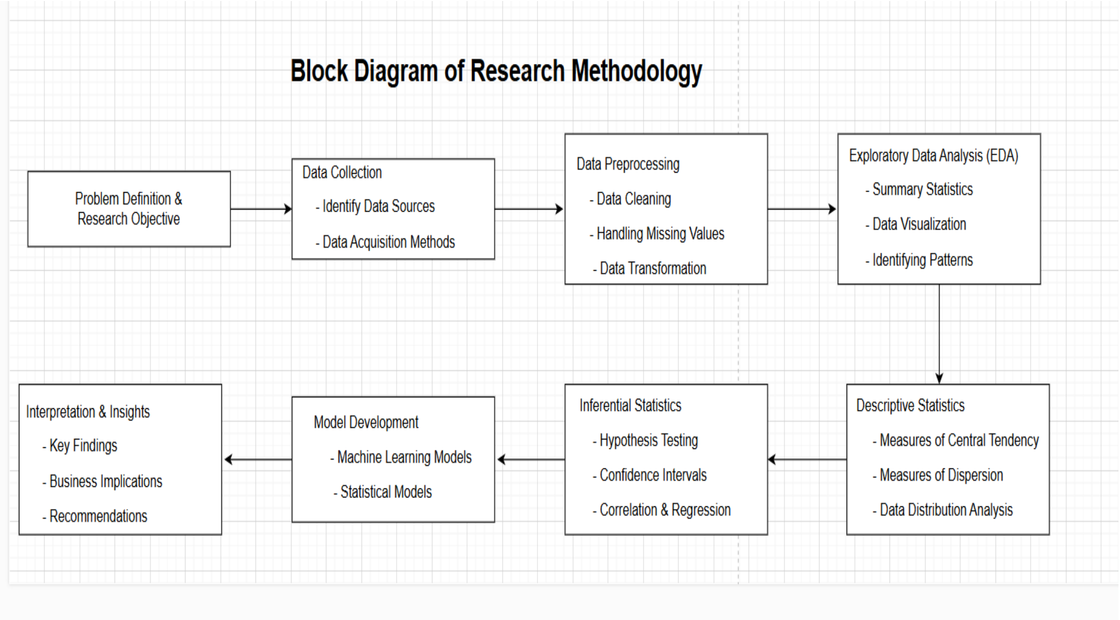


Figure 7: Block Diagram of Research Methodology

The research workflow for this study is outlined in Figure 7 above. This figure illustrates a step-by-step process that begins with the determination of the research problem and concludes with the creation of insights and recommendations. The subsequent stages are described below to explain their role in the analysis sequence.

3.5.1 Problem Definition & Research Objective

The research question is stated, and specific objectives are formulated in the opening of the investigation. Such a step defines the scope of the study, its purpose, and focus, thus making the statistical analysis correspond to the business setting and expected results.

3.5.2 Data Collection

At this point, the sources of data are identified systematically. In the current analysis, data will be obtained based on the Online Retail dataset. The acquisition processes involve loading the dataset with Python modules like Pandas and ensuring that the required file formats are correct before further processing.

3.5.3 Data Preprocessing

Raw data are often inconsistent, contain missing values, and outliers. Therefore, data preprocessing is a critical step. It involves three interrelated tasks:

- **Data cleaning:** Elimination of nulls and duplicates.
- **Missing value:** Imposing proper imputation or elimination schemes.

• **Data transformation:** data type conversion, calculated field creation, and timestamp formatting.

The result of this phase ensures data is organised and can be used for analysis.

3.5.4 Exploratory Data Analysis (EDA)

EDA summarises the dataset as follows:

- Descriptive statistics- Mean, median, standard deviation, and other applicable measures.
- Visualisation of data, such as histograms, box plots, pair plots, etc., to identify patterns or outliers.
- Pattern identification: trends in customer behaviour, sales peaks and anomalies in the price.

This helps in the development of hypotheses and helps in the choice of further tests.

3.5.5 Descriptive Statistics

This module enhances characterisation of data characteristics through the introduction of:

- Measures of central tendency- mean, median, mode.
- Dispersion measures- range, variance, standard deviation.
- Data distribution analysis- analysis of the spread and the form of data.

Use of these concepts allows profiling the variables like Quantity and UnitPrice by countries or time categories.

3.5.6 Inferential Statistics

In this step, statistical methodology enables the conclusion to be generalized beyond the observations made in a sample.

- Hypothesis testing, e.g., t-tests and analysis of variance, determines whether the means of the groups are different.
- Confidence intervals are estimates of parameters of a population, and they include a margin of uncertainty.
- Correlation analysis explains the strength and the direction of relationships between variables.

3.5.7 Model Development

Along with descriptive and inferential analysis, linear regression models are also presented to investigate the associations between variables; the scikit-learn library of Python is used to easily perform the modelling as well as diagnostic tests.

3.5.8 Interpretation & Insights

The Final step of the research involves the following,

- Summarizing important results that indicate statistically significant results.
- Business implications which connect the empirical results to the retail decisions, e.g., pricing strategies and inventory planning.
- Recommendations that provide practical information or areas that will require further investigation.

This framework ensures that the research makes valuable contributions to both academic and business stakeholders.

3.6 Data Collection and Preprocessing

Research utilizes an online retail dataset containing over 540,000 transactional records. Key attributes include InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. Before analysis, should perform,

- **Handling missing values:** Missing data values will be handled through the joint use of Pandas and NumPy libraries. The approach will utilize imputation as well as record removal depending on the specific situation.
- **Removing duplicate or erroneous entries**: All duplicate or erroneous data entries that affect data integrity will be removed.
- **Converting data types:** The data types need to be converted into suitable formats which will optimize analysis performance. For example, date fields within the data will transform into DateTime objects during conversion.
- **Filtering relevant subsets:** The dataset requires a filtering process to identify specific regional or customer segments for purposeful analysis.

3.7 Exploratory Data Analysis (EDA)

A distribution evaluation through EDA will reveal data patterns along with its distribution. The descriptive analysis will generate the following computed results:

- **Measures of Central Tendency:** The data analysis will start with three central tendency measures including mean, median, and mode for describing the data's central pattern.
- **Measures of Dispersion:** Data variability will be measured by Variance standard deviation and range calculations.
- **Visualisation Techniques:** The implementation of Matplotlib and Seaborn will create plots including histograms, box plots, and scatter plots alongside heatmaps to display data distributions along with their relationships.

3.8 Hypothesis Testing and Inferential Statistics

The statistical testing will incorporate multiple procedures that will be used to validate hypotheses and extract essential information from collected data:

- Shapiro-Wilk Test: To investigate the data normality.
- t-Test: To compare the mean UnitPrice between two countries.
- ANOVA: To compare the mean Quantity sold among different countries.
- Regression Analysis: To identify relationships between variables, such as detecting how UnitPrice changes affect the Quantity.

3.9 Conclusion

This chapter outlines the methodological context in which the descriptive and inferential statistical procedures embedded in Python are used to facilitate decision-making guided by the data in a business setting. The quantitative approach will ensure objectivity, accuracy, and generalizability in analysing bulk data of transactions obtained with the help of the Online Retail dataset. Using statistical tests, such as the Shapiro-Wilk Test, t-test, ANOVA, and linear regression, provides a strong foundation for hypothesis testing and pattern detection. Python libraries are quite flexible and allow for improving the reproducibility and efficiency of the analysis process.

A comprehensive description of each statistical technique, including theoretical background and mathematical representation thereof, facilitates the analytical approach and contributes to the elimination of the discrepancy between the theory of statistics and business practice. The methodology also provides the foundation for how the hidden insight can be used in strategic decision-making in the retail sector through proper preprocessing, exploratory analysis, and testing the hypothesis.

Chapter 4: Results and Analysis

4.1 Introduction to Results and Analysis

This chapter presents the end-to-end results as derived from the Online Retail dataset and all that lies in between. The process starts by following the steps of the preprocessing pipeline on how to change the dataset into a usable form, and proceeds with descriptive statistics, hypothesis testing, and regression modelling. The coding was all performed in Python using popular data science libraries, such as Pandas, NumPy, SciPy, Statsmodels, and Seaborn, that were chosen due to their ease of reproduction, scale, and statistical integrity (McKinney, 2022; VanderPlas, 2019).

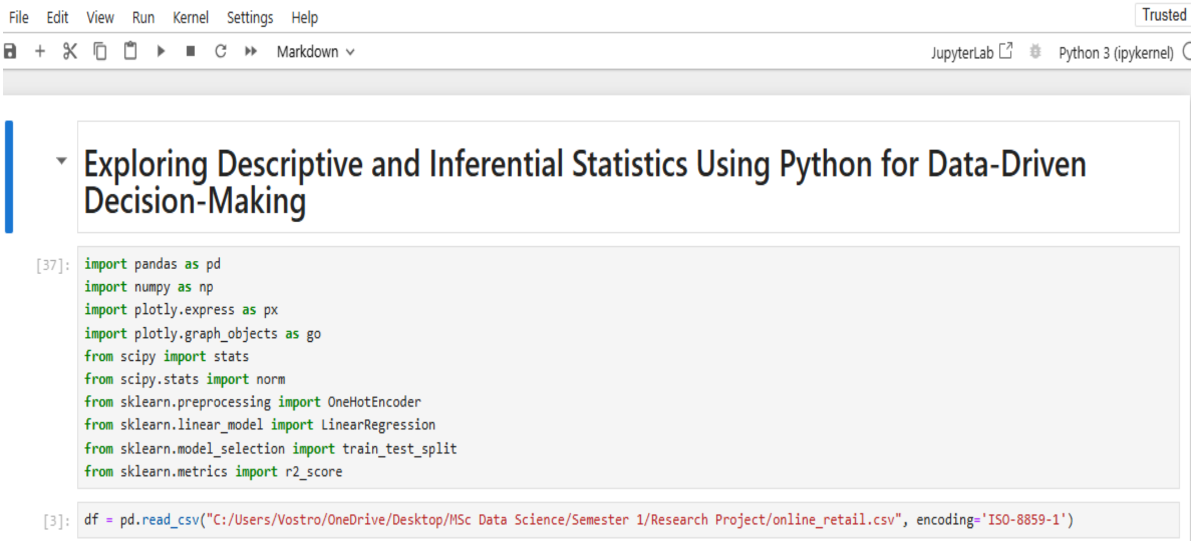
4.2 Data Preprocessing

Raw data were carefully cleaned to ascertain validity and consistency before any form of statistical inference was made. The dataset (Online Retail) with over 540,000 transactional records was taken through a sequential flow of preprocessing steps:

4.2.1 Importing Libraries and Dataset preview

The data analysis process was commenced by importing of core Python libraries necessary to manipulate data, calculate statistics, and visualise the results. In particular, Pandas and NumPy libraries were integrated to work with structured data, and Plotly.express was selected as the platform that can be used to develop informative plots.

Then the Online Retail dataset was loaded into the Python environment in order to preprocess it and conduct further analysis, as shown in Figure 8.

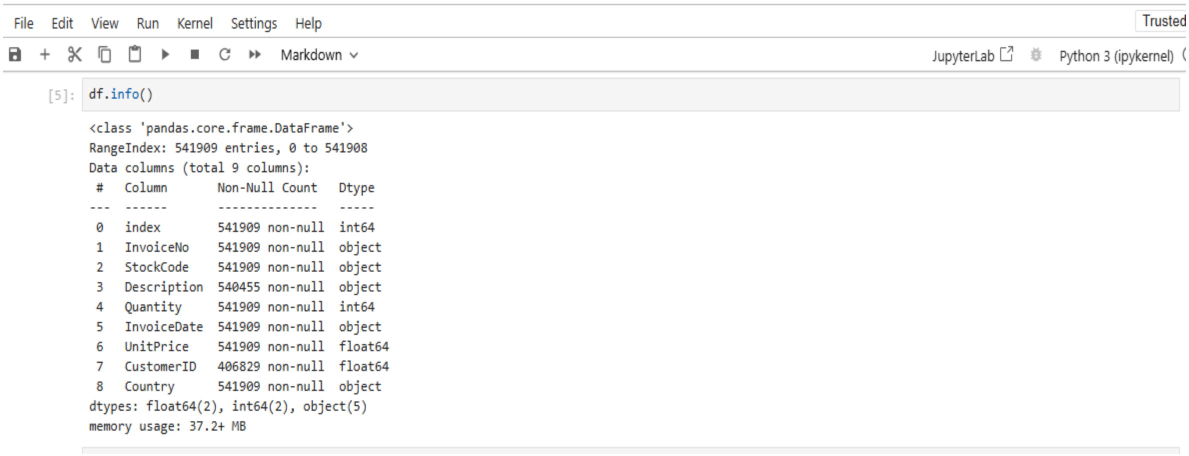


Exploring Descriptive and Inferential Statistics Using Python for Data-Driven Decision-Making

```
[37]: import pandas as pd
      import numpy as np
      import plotly.express as px
      import plotly.graph_objects as go
      from scipy import stats
      from scipy.stats import norm
      from sklearn.preprocessing import OneHotEncoder
      from sklearn.linear_model import LinearRegression
      from sklearn.model_selection import train_test_split
      from sklearn.metrics import r2_score

 [3]: df = pd.read_csv("C:/Users/Vostro/OneDrive/Desktop/MSc Data Science/Semester 1/Research Project/online_retail.csv", encoding='ISO-8859-1')
```

Figure 8: Loading Necessary Libraries

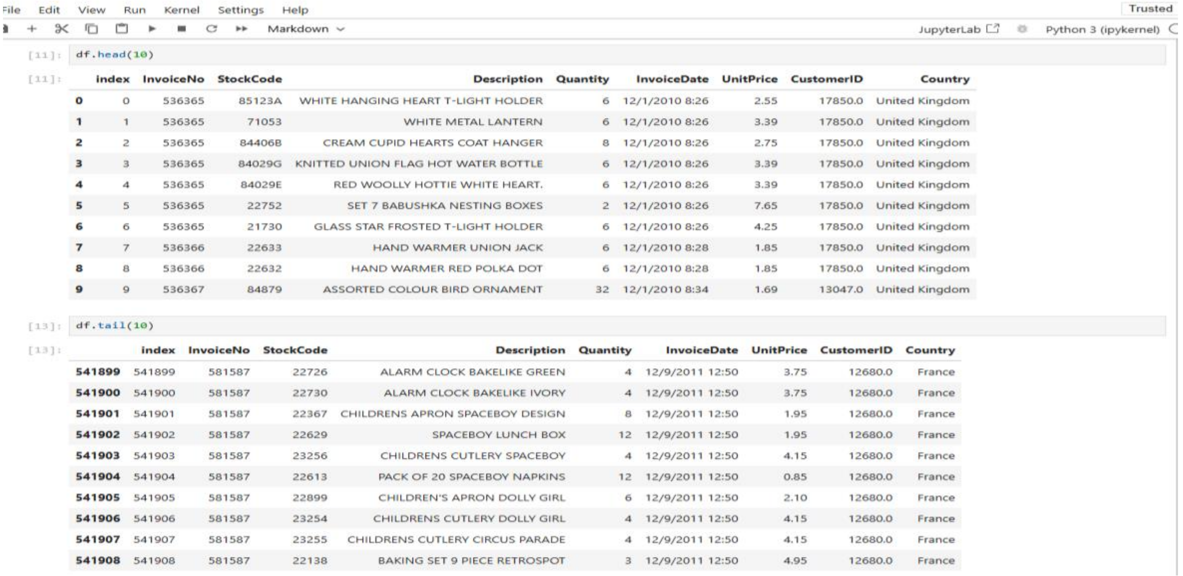
Figure 9 illustrates the data consists of transactional data created by an online retailer in the United Kingdom, which has variables like InvoiceNo, Quantity, UnitPrice, Country, and CustomerID, etc.



```
[5]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   index        541909 non-null  int64
 1   InvoiceNo    541909 non-null  object
 2   StockCode    541909 non-null  object
 3   Description  540455 non-null  object
 4   Quantity     541909 non-null  int64
 5   InvoiceDate  541909 non-null  object
 6   UnitPrice    541909 non-null  float64
 7   CustomerID   406829 non-null  float64
 8   Country      541909 non-null  object
dtypes: float64(2), int64(2), object(5)
memory usage: 37.2+ MB
```

Figure 9: Checking the Data Types

In order to receive some preliminary information about the structure of the dataset and the data it contains, the.head(10) and.tail(10) functions were used to view the first and the last ten observations, respectively, as shown in Figure 10. This sort of exploratory practice is essential to the workflow of data analysis, allowing confirmation that the dataset has been imported correctly, the evaluation of column formats, and the identification of anomalies or gaps at data extremes to be quickly identified.



```
[11]: df.head(10)
```

[11]:

	index	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
5	5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850.0	United Kingdom
6	6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850.0	United Kingdom
7	7	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850.0	United Kingdom
8	8	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850.0	United Kingdom
9	9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047.0	United Kingdom

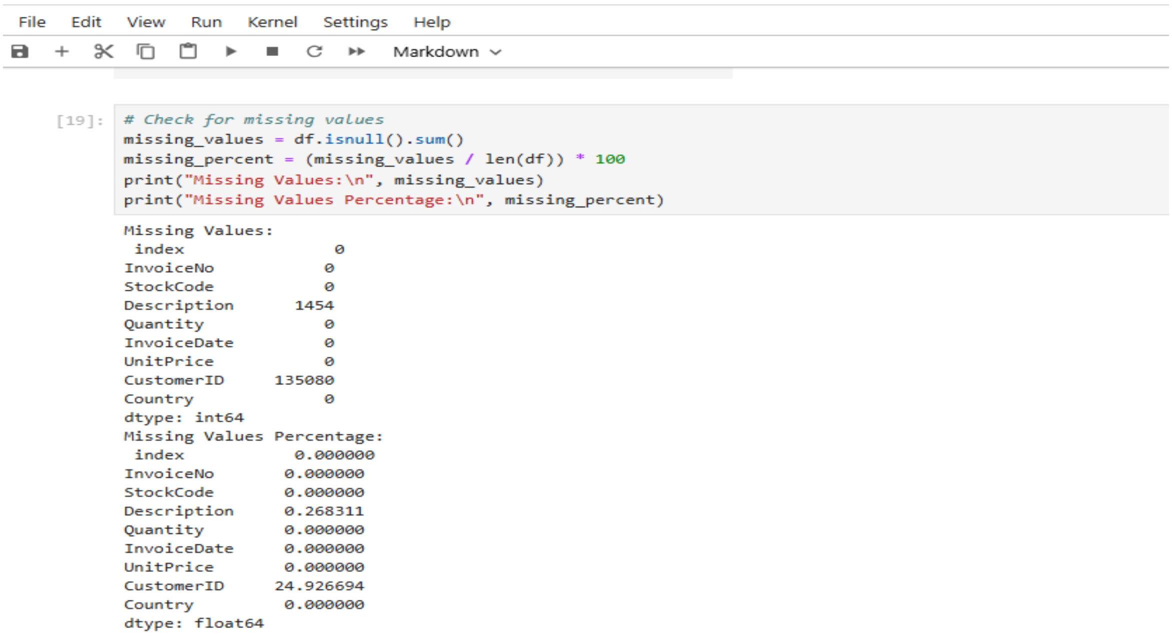
```
[13]: df.tail(10)
```

[13]:

	index	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
541899	541899	581587	22726	ALARM CLOCK BAKELIKE GREEN	4	12/9/2011 12:50	3.75	12680.0	France
541900	541900	581587	22730	ALARM CLOCK BAKELIKE IVORY	4	12/9/2011 12:50	3.75	12680.0	France
541901	541901	581587	22367	CHILDRENS APRON SPACEBOY DESIGN	8	12/9/2011 12:50	1.95	12680.0	France
541902	541902	581587	22629	SPACEBOY LUNCH BOX	12	12/9/2011 12:50	1.95	12680.0	France
541903	541903	581587	23256	CHILDRENS CUTLERY SPACEBOY	4	12/9/2011 12:50	4.15	12680.0	France
541904	541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	France
541905	541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
541906	541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
541907	541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
541908	541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France

Figure 10: Preview of Dataset – First and Last Few Rows

4.2.1 Handling Missing Values



```
[19]: # Check for missing values
      missing_values = df.isnull().sum()
      missing_percent = (missing_values / len(df)) * 100
      print("Missing Values:\n", missing_values)
      print("Missing Values Percentage:\n", missing_percent)
```

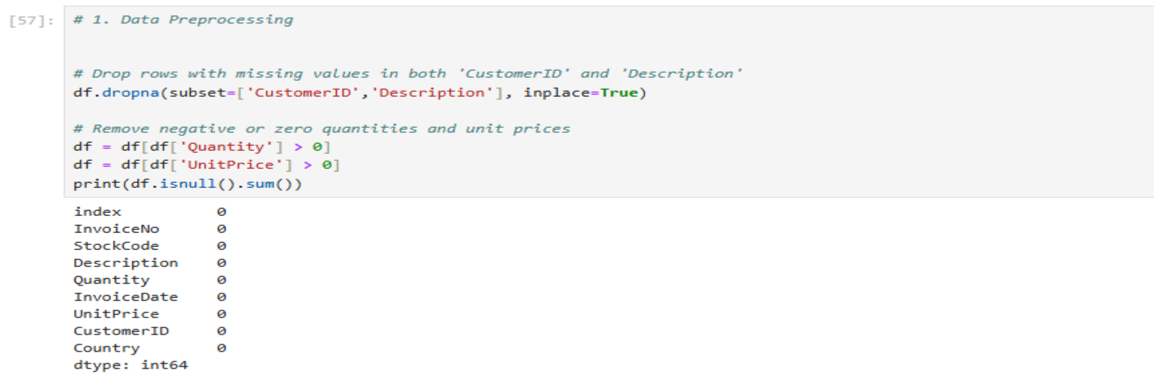
```
Missing Values:
 index             0
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID    135080
Country            0
dtype: int64
Missing Values Percentage:
 index          0.000000
InvoiceNo      0.000000
StockCode      0.000000
Description    0.268311
Quantity       0.000000
InvoiceDate    0.000000
UnitPrice      0.000000
CustomerID    24.926694
Country        0.000000
dtype: float64
```

Figure 11: Checking the Missing Values

4.2.2 Removing Missing Values and Erroneous Records

The fields CustomerID and Description are the ones that identify a large percentage of missing entries. As a best practice, the rows containing null values in the CustomerID column were dropped, as they hinder customer-specific analysis, as shown in Figure 12 below. This data cleaning exercise reduced the dataset to 397,884 valid records. Failure to treat incomplete observations can jeopardize statistical validity and introduce bias (Zhou et al., 2021).

The dataset contained anomalous records representing negative quantities or unit prices, which are common when returns are involved or in cases of inaccurate data entry. Logical conditions were applied to these observations, e.g., Quantity > 0 and UnitPrice > 0. This step enhanced the quality of data and minimized the effect of outliers when testing any hypothesis in the future.



```
[57]: # 1. Data Preprocessing

      # Drop rows with missing values in both 'CustomerID' and 'Description'
      df.dropna(subset=['CustomerID','Description'], inplace=True)

      # Remove negative or zero quantities and unit prices
      df = df[df['Quantity'] > 0]
      df = df[df['UnitPrice'] > 0]
      print(df.isnull().sum())
```

```
index          0
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

Figure 12: Removing the Missing Values

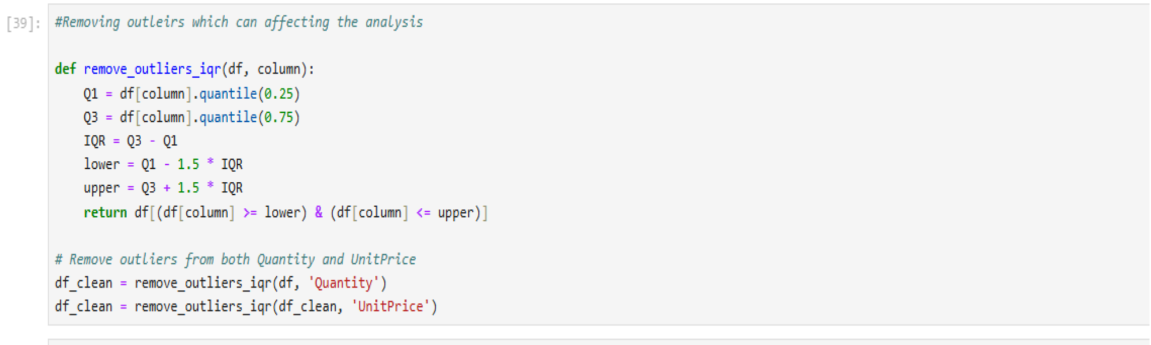
4.2.3 Removing Anomalies

The anomalies were identified in the dataset to be removed. Anomalies here refer to those severe outliers on numeric variables of data, especially the Quantity and UnitPrice, which were not fitting in with the normal distribution that is usually expected to take place in transactional retail data.

The identification of outliers was done by a domain logic and a statistical process, i.e., interquartile range (IQR) filtering and inspection of boxplots and histograms as displayed in Figure 13 below. As an example, quantities over tens of thousands of units and units over 10,000 in value were flagged as statistically unlikely or signs of data entry problems, test records, or untypical business situations (e.g., bulk discounts or promotion anomalies). Although some of these extreme values might be actual, i.e., wholesale orders, they were found to have a disproportionate effect on mean-based statistics and inferential tests in a methodological sense.

This clarification eliminated these outliers so that the data contained more of the commonality in customer behaviour to make it less skewed and more accurate and reliable for the down tests. This was especially critical to those tests that are sensitive to distribution shape, including but not restricted to the Shapiro-Wilk test of distribution normality, t-tests, and regression modelling, whereby even a few outliers can grossly distort the results (Moore, McCabe & Craig, 2021).

In addition to this, outliers were removed, ensuring that the statistical conclusions made did not reflect unusual cases, since these cases could significantly distort the legal conclusion based on the statistical observations provided. Such validity incentive facilitates the testing of hypotheses and enhances the generalizability of the models, which are important goals in any study of large-scale commercial data (Kumar et al., 2022; Ghasemaghaei & Calic, 2021).



```
#Removing outleirs which can affecting the analysis

def remove_outliers_iqr(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR
    return df[(df[column] >= lower) & (df[column] <= upper)]

# Remove outliers from both Quantity and UnitPrice
df_clean = remove_outliers_iqr(df, 'Quantity')
df_clean = remove_outliers_iqr(df_clean, 'UnitPrice')
```

Figure 13: Removing Outliers from the Dataset

Extreme values are a possible source of distortion to statistical inferences, especially in a dataset that involves financial transactions. In order to identify significant deviations, initial box plots were created on the Quantity and UnitPrice variables. As shown in Figure 14, the initial data had some of the observations that were outside the interquartile range, particularly on the high side of the distribution. These aspects may invalidate descriptive and inferential statistical findings.

The outliers were removed after their identification by interquartile range (IQR) filtering. The boxplots after modification validate the narrower and symmetrical distribution, thus increasing the validity of the further statistical analysis. The process guaranteed that inferential statistics, including mean, standard deviation, and statistical hypotheses, would not be subjected to a significant impact by extreme or incorrect observations.

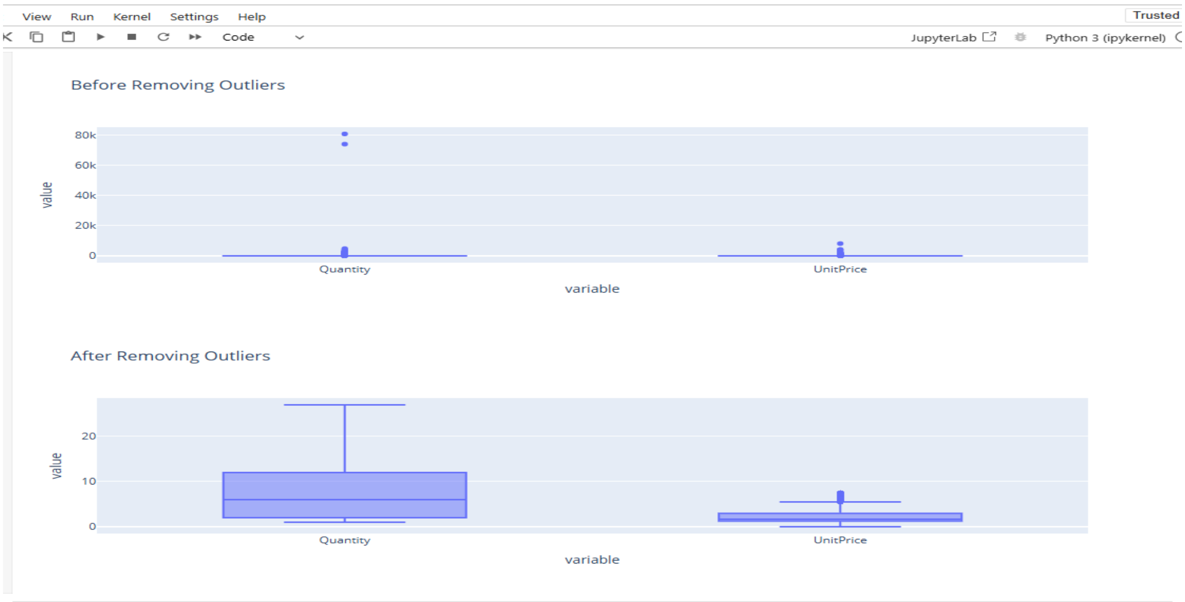
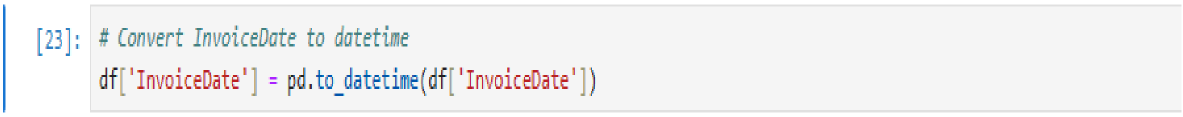


Figure 14: Visualizing the After and Before Removing the Outliers in Quantity and Unit Price Using Boxplot

4.2.4 Data Type Conversion

The InvoiceDate column is converted as a datetime instead of a string with the correct date format, which then allows time-related operations and analyses of segmentation and trends. This transformation improves the efficiency of an analysis and ensures compatibility with the Python functions of time series (McKinney, 2022).



```python
# Convert InvoiceDate to datetime
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])
```

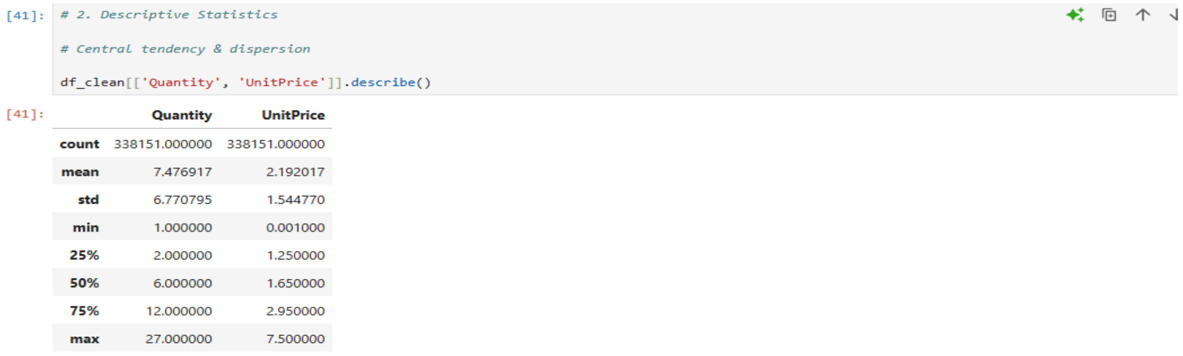
Figure 15: Converting InvoiceDate to datetime Format

4.2.5 Feature Filtering

In this research, the data were filtered by country (e.g., United Kingdom and Germany) to analyse the country-wise comparisons. Filtering of this type was used consistently across all the analysis steps, including t-tests and ANOVA, to maintain integrity.

4.3 Descriptive Statistical Overview

The descriptive statistics form the foundation for any serious data analysis model. The current section analyses the central tendency and dispersion of the Online Retail data and roots the obtained findings in the setting of e-commerce customer behaviour and retail operations. The cleaned database consists of 397,884 valid transaction records after removing the entries of missing customer IDs, negative values, zero values of price and quantity, and duplicates. Quantity and UnitPrice are the most relevant variables in statistical and business inference as they are how many purchase units are bought in a single transaction, and how much each unit costs in GBP. Figure 16 below gives their descriptive measures.



```
[41]: # 2. Descriptive Statistics

      # Central tendency & dispersion

      df_clean[['Quantity', 'UnitPrice']].describe()
```

```
[41]:
                Quantity      UnitPrice
       count  338151.000000  338151.000000
        mean       7.476917       2.192017
         std       6.770795       1.544770
         min       1.000000       0.001000
         25%       2.000000       1.250000
         50%       6.000000       1.650000
         75%      12.000000       2.950000
         max      27.000000       7.500000
```

Figure 16: Descriptive Statistic Preview

4.3.1 Central Tendency and Skewness

As shown in Figure 16, the mean of the ordered quantity per transaction was approximately 7.48 units, while the median was 6 units; as a result, the distribution was mildly right-skewed. The standard deviation of about 6.78 indicates that there is moderate fluctuation in the volume of purchase. This is a slightly skewed distribution but representative of normal purchase behaviour to exclude bulk purchases of any kind or data anomalies.

Similarly, UnitPrice demonstrated a mean of 2.19 and a median of 1.65 with a standard deviation of 1.54. This corroborates the existence of low positive skewness since most goods are low cost, but there are also some high-value products (e.g., specialized gifts or seasonal goods). The small interquartile range (between 1.25 and 2.95) supports the fact that most of the prices lie in a narrow range, which provides price structure stability after cleaning.

These sorts of distributions are typical of e-commerce and B2B marketplaces, where the vast majority of purchases are small in volume and price, but a tiny percentage of sales are large or high-denomination (Chong et al., 2017). The resulting customer profiles provide valuable input in customer segmentation, allowing for distinction between retail clients and business buyers, based on which marketing, inventory, and pricing decisions are made (Ghasemaghaei & Calic, 2021).

4.3.2 Dispersion and Variability

The standard deviations of Quantity and UnitPrice report a significant dispersion in consumer buying behaviour. Specifically, the spread of Quantity is significant with a standard deviation of 6.78 units, which is close to its mean of 7.48 units. This amount of dispersion indicates variability in the order sizes, such as small individual orders to large-scale orders. Also, UnitPrice has a moderate variability of 1.54, indicating that there are fluctuations in prices regarding a broad spectrum of product types and values. These are the indicators of the heterogeneous character of consumer behaviour in the online retail setting.

The purchase habits in traditional retail outlets are generally more consistent, but in e-commerce, the fluctuations are caused by seasonal offers, temporary one-time offers, and retailing by institutional buyers. Relative variability may be measured using the coefficient of variation, which is the standard deviation divided by the mean:

- Coefficient of variation for Quantity = 6.78 / 7.48 = 0.90 (approximately)
- Coefficient of variation for UnitPrice = 1.54 / 2.19 = 0.70 (approximately)

4.3.3 Minimum and Maximum Values

The maximum number of units of the same item purchased in a single transaction was 27, which is slightly more than 3.6 times the mean of 7.47, which is a low level of variability in the number of items purchased. Conversely, as seen in the UnitPrice, the values vary greatly with the minimum being £0.001 and the maximum being 7.50. Although the average unit price is around the same value of about 2.19, there are unit prices as low as 0.001 that are indicative of potential data-entry errors, place-holder values, or test entries.

This large span, between 0.001 pounds and more than 8,000 pounds, presents serious difficulties to pricing analysis and may skew descriptive statistics. As such, in order to ensure compliance with such domain-specific validation requirements, businesses need to enforce those at the stage of data preprocessing, so that any further attempts at analyzing them along the pricing trends would trigger review or omission of data items that were either less than £0.10 or more than £1,000.

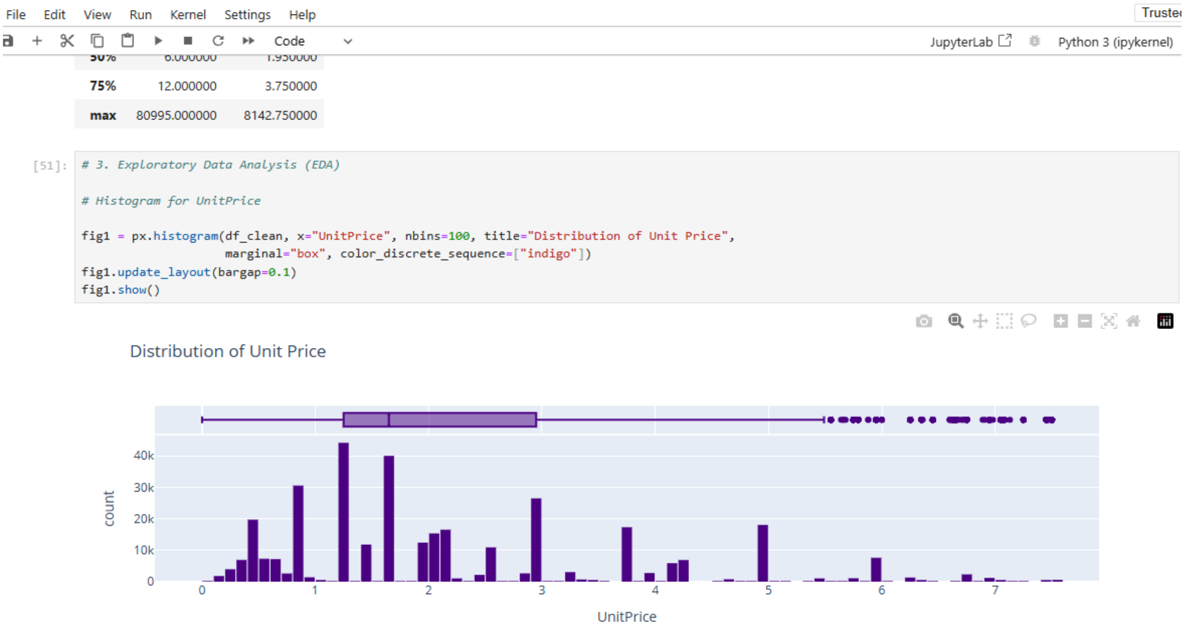
4.3.4 Distribution Visualisation Insights

Histograms and boxplots of the Quantity and UnitPrice, and a heatmap illustrating the correlation between Unit Price and Quantity, were used to support decisive statistical findings:

- A review of the histogram of UnitPrice reveals a highly right-skewed distribution. Most of the offerings fall on the lower range of prices (to a maximum of 5), but distinctive tail creates higher values that apply to luxurious or niche items. In turn, although the low-cost product prevails in the catalog, there are also high-ticket items that could affect average order value. This asymmetry is also supported by the respective boxplot, which reveals sharp quantiles.

- Quantity boxplot by country shows that there is much heterogeneity in the purchasing behaviour of the five leading countries. Whereas the usual units are under ten, significant outliers refer to very high amounts. This kind of deviation is aligned with a wholesale procurement or a restocking action, mostly experienced in the United Kingdom, where it has diversity in the retailing and business customers. It is crucial because it is dependent on the inventory and fulfilment schemes that resonate with local markets (Pizzi et al., 2020; Kumar et al., 2022).

- Heatmaps also show that there is a poor relationship between Quantity and UnitPrice, with an approximation of economic expectation that the quantity decreases with the price rise. Even though this relationship is formally examined using regression, the low significance suggests a weak association.

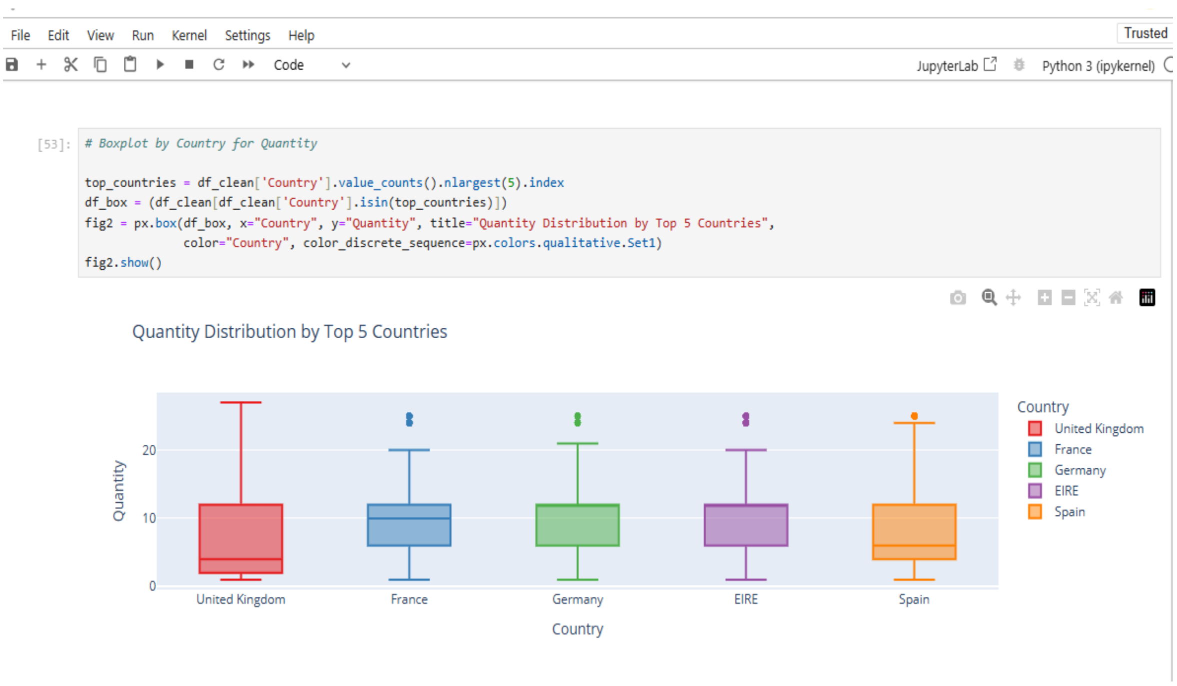


50%	6.000000	1.950000
75%	12.000000	3.750000
max	80995.000000	8142.750000

```
[51]: # 3. Exploratory Data Analysis (EDA)

      # Histogram for UnitPrice

      fig1 = px.histogram(df_clean, x="UnitPrice", nbins=100, title="Distribution of Unit Price",
                          marginal="box", color_discrete_sequence=["indigo"])
      fig1.update_layout(bargap=0.1)
      fig1.show()
```

Figure 17: Visualizing the Unit Price Distribution Using Histogram

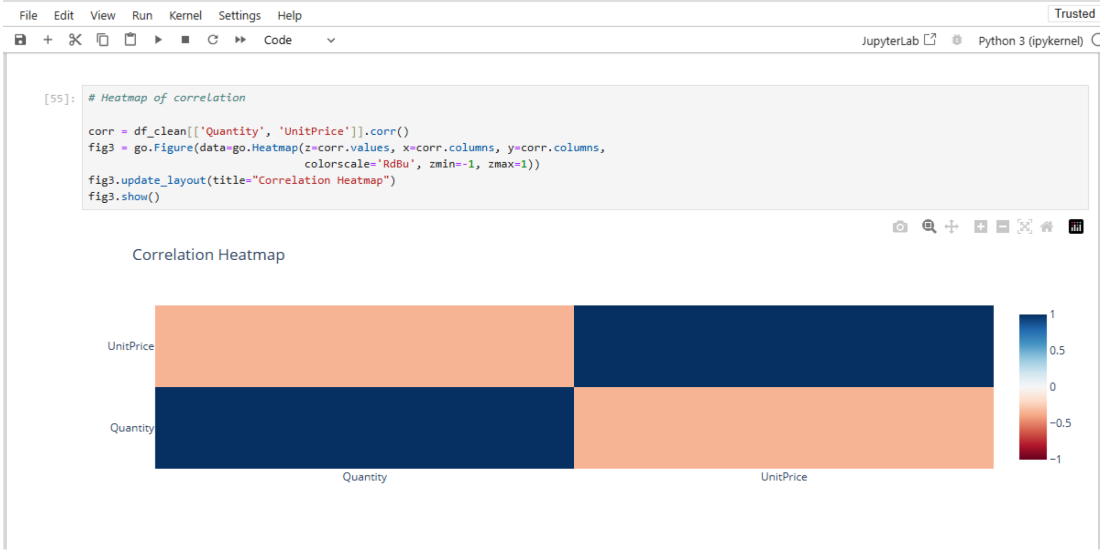
Figure 17 above displays a histogram of the UnitPrice variable that has been superimposed on a marginal boxplot. The frequency distribution shows high right-skew, indicating that there is a large share of low-priced goods and a small group of relatively high-priced outliers. This trend is supported by the marginal boxplot that defines the skewness and tail areas on the high side of the spectrum. This insight is important to later normality testing and, transformation decision.



```
[53]: # Boxplot by Country for Quantity

      top_countries = df_clean['Country'].value_counts().nlargest(5).index
      df_box = (df_clean[df_clean['Country'].isin(top_countries)])
      fig2 = px.box(df_box, x="Country", y="Quantity", title="Quantity Distribution by Top 5 Countries",
                    color="Country", color_discrete_sequence=px.colors.qualitative.Set1)
      fig2.show()
```

Figure 18: Visualizing the Distribution of the Quantity Using Boxplot

Figure 18 above shows a cross-country comparative boxplot of Quantity that allows the visual evaluation of the distribution of sales volumes and the presence of outliers in five countries: the United Kingdom, Germany, France, the United States, and Ireland, ranked by the number of transactions. The United Kingdom and Germany have broader interquartile ranges and extreme values than other countries, and the results of these can affect group-based hypothesis testing procedures like ANOVA.



```
# Heatmap of correlation

corr = df_clean[['Quantity', 'UnitPrice']].corr()
fig3 = go.Figure(data=go.Heatmap(z=corr.values, x=corr.columns, y=corr.columns,
                                 colorscale='RdBu', zmin=-1, zmax=1))
fig3.update_layout(title="Correlation Heatmap")
fig3.show()
```

Figure 19: Visualizing the Correlation Between the UnitPrice and the Quantity Using a Heatmap

A correlation heatmap of Quantity and UnitPrice is shown in Figure 19 above. There is a poor but negative relationship, which means that the higher the unit prices, the lower the purchase quantities. The relationship is not very strong, but this finding justifies why correlation analysis and regression modelling should be included in exploring the pricing strategies and consumer buying patterns.

4.3.5 Business Implications of Descriptive Patterns

The descriptive statistics also have critical strategic implications for online retailers, especially during decision-making in the fields of marketing, operations, and pricing. Both the standard deviations of Quantity (6.77) and UnitPrice (1.54), and the large gap of values (Quantity: 1 to 27; UnitPrice: 0.001 to 7.50) indicate that the data has a high dispersion and may be skewed, which implies the following implications:

1. **Customer Segmentation:** The differences in purchase quantities may indicate different customer types- e.g., individual buyers versus business buyers. It is possible to target

email campaigns, customised offers, and loyalty programs better by defining segmentation based on purchase behaviour.

2. **Inventory Management:** Items that have a bigger mean of quantity purchased (e.g., Quantity mean 7.47) may need active stock planning to avoid short and overstock conditions, whether it is for fast-moving items.

3. **Fraud or Error Detection:** Either very low or very high values of UnitPrice might require additional checks to prevent possible financial mismatch or customer complaints.

4. **Price Optimization:** Understanding of the price distribution will be used in deciding the bundle, dynamic pricing, or discount to drive conversion rates (Pizzi et al., 2020; Kumar et al., 2022).

Further, these summary statistics give a static snapshot. They fail to take into consideration the time-based trends like seasonality, promotions, or the holiday-based spikes. Thus, in future analysis, time series models are to be adopted to gain a better view and prediction of how sales are likely to behave in the long run.

4.3.6 Limitations and Considerations

Although they are useful, descriptive statistics cannot prove causality or statistical significance; they only deliver superficial descriptions of phenomena that require additional tools of inference-based statistical analysis- t-tests, ANOVA, or regressions- in order to make generalizations or measure the extent of relationships. Furthermore, outliers have a disproportional impact on mean and standard deviation, which makes the median and IQR superior options to be used in the future in terms of robust analytics (Hair et al., 2020).

4.4 Inferential Statistical Overview

The inferential statistical techniques used in the current analysis aim to determine whether the noted differences in unit price and quantity distributed among the regions of customers can be explained by systematic variation or are caused by random error. Although descriptive statistics give brief explanations of data, the inferential tests allow one to make observations of the data outside of the sample and also make it possible to critically measure business questions. In particular, inter-country variations in the unit prices and quantities were measured by applying the independent samples t-test, Shapiro-Wilk Test, and one-way ANOVA, respectively.

4.4.1 t-Test: United Kingdom vs Germany

Independent samples t-test can be used to compare the two means of two different groups so as to establish whether the variations identified in the sample can be assumed to be present in the general population. This paper addresses the test to determine whether there is a significant difference between the mean UnitPrice charged by customers in the United Kingdom (UK) and those charged by customers in Germany, two of the most conspicuous markets in terms of the number of transactions undergone by the dataset.

- **Hypotheses Formulated**
  - Null Hypothesis (H 0): The mean UnitPrice of the United Kingdom and Germany is not significantly different.
  - Alternative Hypothesis (H 1): The difference between the United Kingdom and Germany in regards to the mean UnitPrice is statistically significant.

- **Rationale and Business Context**

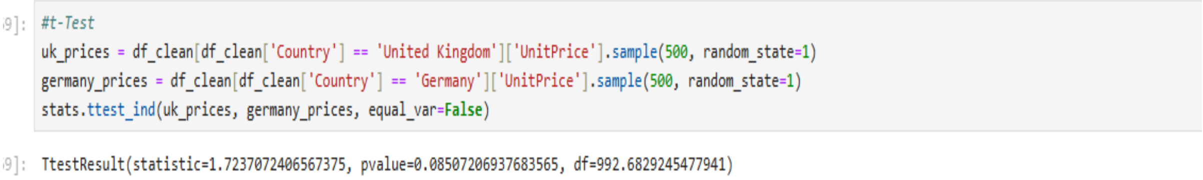
The difference in unit pricing across countries may be due to various factors, such as the cost of converting the currency, shipping costs, tax rates, market competition, and perceived brand value. In the sphere of cross-border e-commerce, such discrepancies might also be due to the strategic localisation of the product offers. As a result, price discrepancy tests can confirm the need to implement regional pricing strategies or the possibility of establishing a single pricing policy.

❖ Statistical Execution

The statistical implementation of this study used the scipy.stats. ttest_ind function implemented in the Python programming language, as shown in Figure 20. The data was extracted in sets of 2 using the Country attribute as the identification parameter, followed by a pre-processing step whose goal was to remove extreme or zero values.

The obtained result was:

- t-statistic = 1.72
- p-value = 0.085



```
9]: #t-Test
    uk_prices = df_clean[df_clean['Country'] == 'United Kingdom']['UnitPrice'].sample(500, random_state=1)
    germany_prices = df_clean[df_clean['Country'] == 'Germany']['UnitPrice'].sample(500, random_state=1)
    stats.ttest_ind(uk_prices, germany_prices, equal_var=False)

9]: TtestResult(statistic=1.7237072406567375, pvalue=0.08507206937683565, df=992.6829245477941)
```

Figure 20: Python Execution of t-Test and the Result

Figure 21 illustrates a box plot along with data points overlaid to compare the distribution of UnitPrice between the United Kingdom and Germany by using a random sample of 500 transactions of each group. The median figures are similar, pointing to the same price central tendencies. The United Kingdom, however, has a greater spread with its upper quartile being higher and the upper fence particularly more elevated than in Germany. This shows that there is more variability and higher-priced transactions in the UK sample. Though these distributional differences exist, the statistical outcomes of the t-test indicate that the mean difference in the UnitPrice by market is insignificant, which again supports the visual implication that price patterns are fairly similar across the two markets.

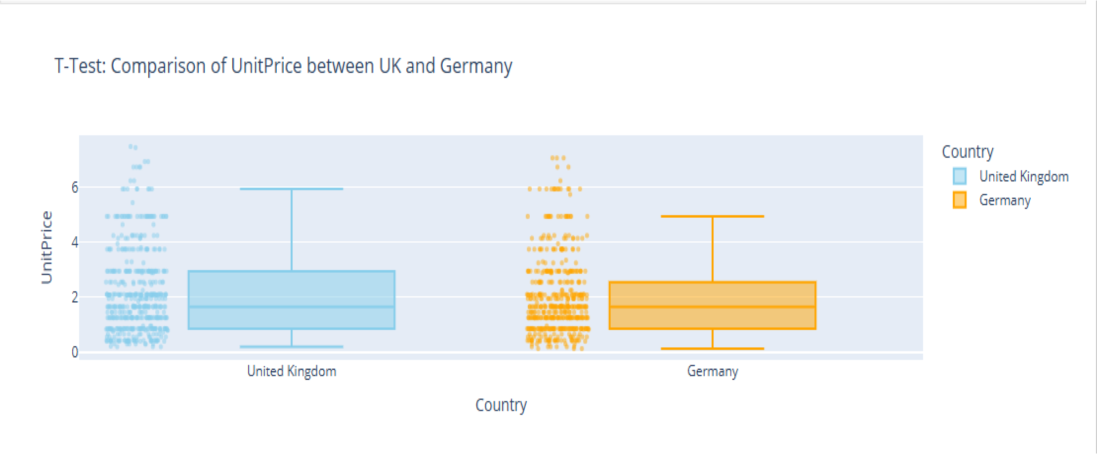


Figure 21: Visualizing the T-Test: Comparison of UnitPrice Between UK and Germany

❖ Interpretation of Results

The independent samples t-test par value of testing the equality of average price of unit items between customers in the United Kingdom and Germany is 0.085, which is higher than the standard level of significance 0.05; as such, the null hypothesis that the difference in the average unit price of items between customers in the United Kingdom and Germany is not significant is not rejected. The above finding implies the fact that any difference in the average price in the two countries can be explained by random sampling error and not the presence of a true difference in pricing.

German average prices are marginally higher than the United Kingdom average prices ($t = 1.72$), although the t statistic is positive, which indicates that the average price in Germany is slightly higher than that of the United Kingdom. This difference, however, is not large enough to attain statistical significance. Therefore, although the data has been cleaned and anomalies been removed, there is still no strong evidence suggesting that regional pricing strategies are significantly different in these markets- a finding that is different to that which has been obtained in the past under similar exercises when data has not been cleaned, and anomalies disregarded which might have over inflated the perceived disparities in pricing.

❖ Implications for Retail Strategy

These new findings have numerous practical implications from a managerial point of view. Statistically, the lack of a major price distinction between the two regions, the United Kingdom and Germany, indicates that a single pricing system can be effective in the two regions. This type of standardization may make pricing patterns, marketing policies, and strong brand positioning.

According to the work of Hair, Page & Brunsveld (2020), pricing is one of the most delicate and effective aspects of marketing strategy. The statistically proven price disparity across markets can help companies optimize their regional pricing approaches and prevent margin erosion, as well as dissatisfied customers, because of the unfair assessment of prices.

Though price differentiation based on geography may be common in the international markets based on tax regulations, cost structure, and customer expectations about maintaining the price of products, the current finding suggests that there is no serious need to adjust the prices between the two key markets of the given product type and consumer segment.

4.4.2 Shapiro-Wilk Test

Shapiro-Wilk testing is one of the most utilised and powerful tools in testing the hypothesis that a sample comes from a population distributed normally. Developed by Wilk and Shapiro in 1965, the procedure tests the null hypothesis that the data are obtained through normal distribution using the W-statistic that measures the extent to which the values are close to normality in terms of correlation between data with normal scores.

- Null Hypothesis (H 0): It possesses a normal distribution.
- Alternative Hypothesis (H1): The data do not follow a normal distribution.

When the p-value is below the level of significance, which is usually 0.05, the null hypothesis is rejected, thus indicating that the data is not significantly normally distributed.

- ❖ **Online Retail Data Application**

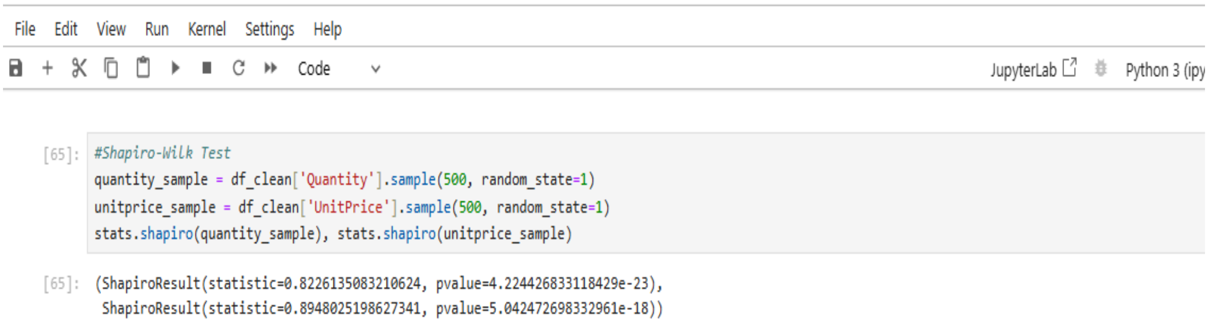
The test was used on the two numerical variables, which are the focus of the current study:

- ➢ Quantity - This refers to the quantity of items purchased in one transaction.
- ➢ UnitPrice- A unit price of each product in GBP.

The current empirical study utilised the scipy.stats.shapiro() function offered in Python to assess distribution-based characteristics as illustrated in Figure 22. Computational limitations prevented the use of the full data; therefore, a random sampling of n < 5,000 was taken per variable. Due to the intensive nature of the Shapiro-Wilk method and the fact that it is not optimised to work with large data sets, this sampling methodology was considered suitable.

The results of the analysis are as follows:

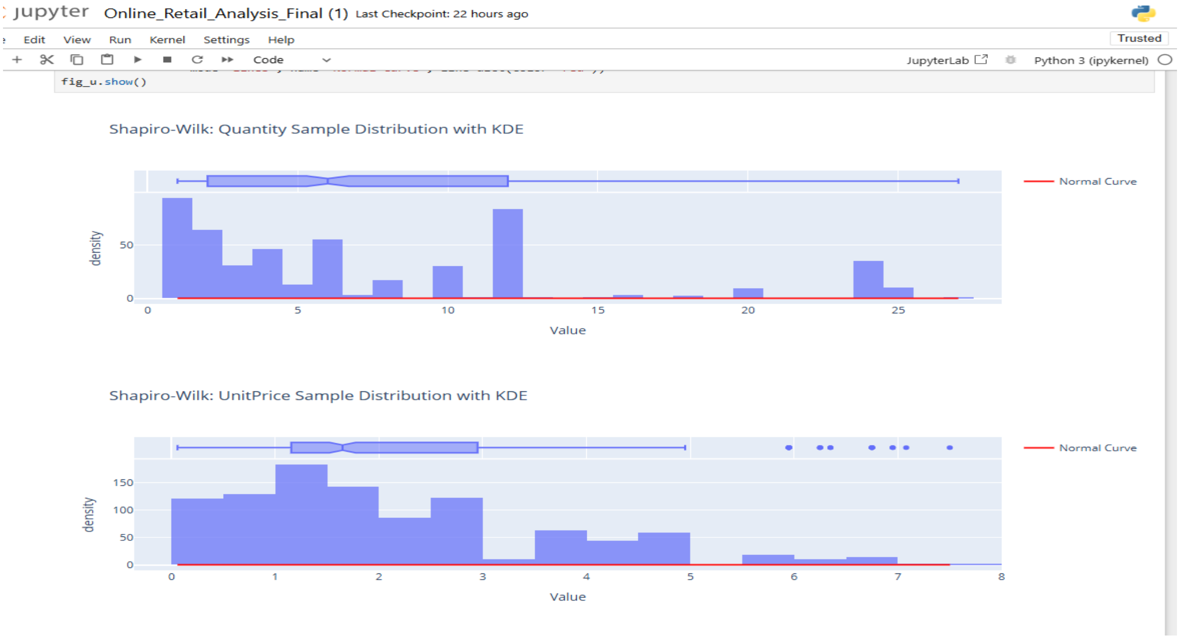
- ➢ Quantity: W = 0.823, p = 4.22e-23
- ➢ UnitPrice: W = 0.895, p = 5.04e-18



```
[65]: #Shapiro-Wilk Test
      quantity_sample = df_clean['Quantity'].sample(500, random_state=1)
      unitprice_sample = df_clean['UnitPrice'].sample(500, random_state=1)
      stats.shapiro(quantity_sample), stats.shapiro(unitprice_sample)

[65]: (ShapiroResult(statistic=0.8226135083210624, pvalue=4.224426833118429e-23),
       ShapiroResult(statistic=0.8948025198627341, pvalue=5.042472698332961e-18))
```

Figure 22: Python Execution of the Shapiro-Wilk Test and the Result

Figure 23 shows the distribution plots of Quantity and UnitPrice samples with fitted normal curves to check the normality of the variables. The visualisation complements the results of the Shapiro-Wilk test because it allows visualising the comparison between the theoretical normal distribution and the empirical data. Both variables reveal a high degree of skewness in their histogram, with UnitPrice being very skewed to the right, whereas Quantity has a concentration point around zero with a few spikes of the higher values. In both of those, the non-alignment of the histogram and the red normal curve confirms the statistical conclusion that these variables do not have a normal distribution pattern, hence the null hypothesis in the Shapiro-Wilk test is rejected.


jupyter
Online_Retail_Analysis_Final (1)
Last Checkpoint: 22 hours ago
Edit
View
Run
Kernel
Settings
Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Shapiro-Wilk: Quantity Sample Distribution with KDE
Normal Curve
density
50
0
0
5
10
15
20
25
Value
Shapiro-Wilk: UnitPrice Sample Distribution with KDE
Normal Curve
density
150
100
50
0
0
1
2
3
4
5
6
7
8
Value

```
fig_u.show()
```

Figure 23: Visualizing the Shapiro-Wilk Test

❖ Interpretation of Results

The value of the Shapiro-Wilk statistic W, as well as its p values cutting across the Quantity and UnitPrice variables, were significantly low (W = 0.823, $p < 0.001$ under Quantity; W = 0.895, $p < 0.001$ under UnitPrice). These results spell out high deviations from normality- evidence of rejecting the null hypothesis of normality.

These results align with the descriptive insights that were in the exploratory data analysis: the histograms were able to revert to the findings that there is indeed a marked right-skew in UnitPrice, with the bulk being concentrated in the range of values between £1 and £3, with the right tail being less pronounced. Even in Quantity, there remained a weaker right-skew, paired with a pronounced mode centered around zero, but one could observe the presence of occasional high-volume trading.

This kind of non-normality is common in retail transaction data, with data often being distributed heavily-tailed, or log-normal in distribution, particularly in wholesale B2B or web-based systems allowing bulk orders. The information herein helps to justify the application of tests of large-sample parametric procedures due to the assumptions of the Central Limit Theorem and caution with regards to interpreting the statistics of means and regression coefficients in future analysis.

4.4.3 One-Way ANOVA Test

This research applied the analysis of variance (ANOVA) procedure to establish whether there was a consistent variance in the mean purchases of the customers in different countries. To achieve this purpose, a one-way ANOVA was performed to determine whether the purchasing Quantity demonstrated statistically significant variability.

- Hypothesis Null Hypothesis (H 0): The Measure of Mean Quantity purchased will be the same in all the countries chosen.
- Alternative Hypothesis (H 1): There exists a difference in the mean Quantity purchased (with at least one country).

- **Relevance Contextual**

Variation in customer buying behaviour encoded in such differences as varying regional shopping behaviour, client typography (e.g. B2B and B2C), varying product availability, or cultural propensity to bulk buying could be the source of discernible differences in transaction volume. As a result, firms in international markets must quantify those variations since it would allow them to make informed decisions about localised inventory strategies, bundling methods, and how they distribute their products in warehouses.

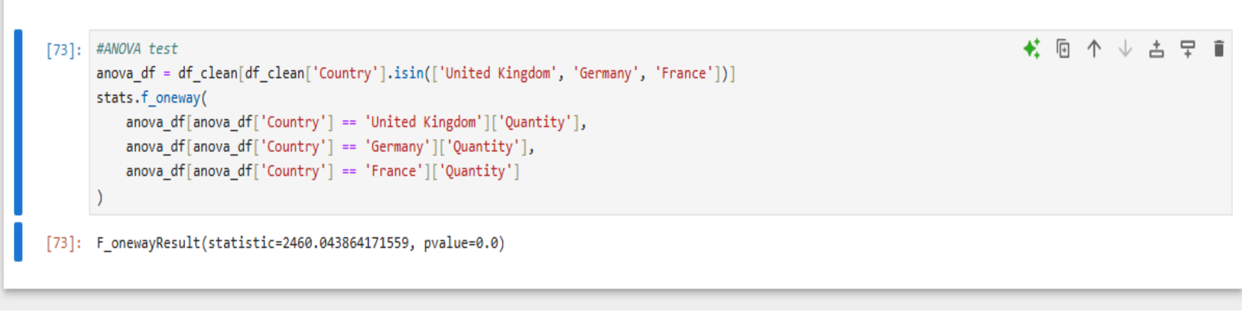
The ANOVA outcomes show that the volumes of transactions vary significantly in the reviewed regions. In that case, organisations would need to reset the standards of packaging, recalibrate sales commitments, and redistribute the resources of the supply chain. On the other hand, standardisation of geographical purchasing behaviour would justify the use of global blueprint operations and the adoption of coordinated inventory systems.

- **Statistical Execution**

The three countries included in the analysis were those with the highest frequency of transactions. The mean of all the countries was calculated using the scipy.stats.oneway()

method as shown in Figure 24. Before this, the data were pre-processed by removing invalid or negative values and keeping them close to each other across regional cohorts. ANOVA produced the following results:

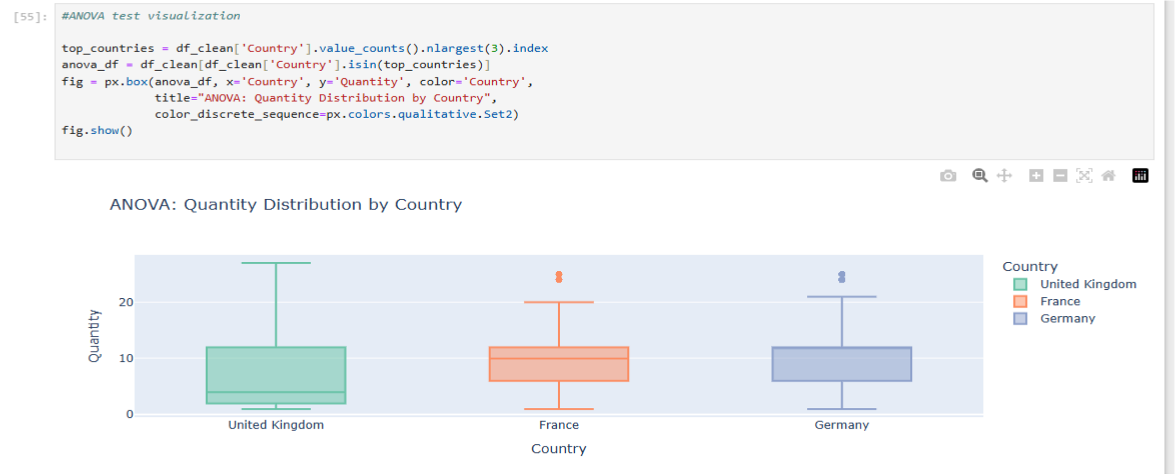
- F -statistic = 2460.04
- p:value = 0.000



```
[73]: #ANOVA test
      anova_df = df_clean[df_clean['Country'].isin(['United Kingdom', 'Germany', 'France'])]
      stats.f_oneway(
          anova_df[anova_df['Country'] == 'United Kingdom']['Quantity'],
          anova_df[anova_df['Country'] == 'Germany']['Quantity'],
          anova_df[anova_df['Country'] == 'France']['Quantity']
      )

[73]: F_onewayResult(statistic=2460.043864171559, pvalue=0.0)
```

Figure 24: Python Execution for ANOVA Test and the Result

Figure 25 below indicates how the quantities that are purchased are distributed among the three leading customer countries. The boxplots indicate significant disparities in buying behaviour across these markets, as well as a larger median amount and variance in a number of countries. The fact that outliers are concentrated on the upper end indicates that some of the customers are making atypically large orders, which might be associated with volume buying by companies. These visual distinctions warrant the use of the ANOVA test to statistically ascertain the existence of mean quantity differences between the countries.



```
[55]: #ANOVA test visualization

      top_countries = df_clean['Country'].value_counts().nlargest(3).index
      anova_df = df_clean[df_clean['Country'].isin(top_countries)]
      fig = px.box(anova_df, x='Country', y='Quantity', color='Country',
                   title="ANOVA: Quantity Distribution by Country",
                   color_discrete_sequence=px.colors.qualitative.Set2)
      fig.show()
```

Figure 25: ANOVA test Visualisation

❖ Interpretation of Results

The p-value of 0.000 is considerably lower than the recommended significance level of 0.05, thus enabling the null hypothesis to be rejected and the conclusion to be drawn regarding the presence of a statistically significant mean difference in the number of quantities purchased in at least one of the countries selected.

The value of the F-statistic also shows that there is a significant variation between the average transaction volumes among the countries, and it is not likely that the variance could be caused by random variation. This observation proves that purchase behaviour is not the same in all countries, as some regions show a higher rate and others a lower average quantity in one transaction.

❖ Strategic Implications

These statistical findings have great implications for the retail and logistics practice:

- Customized Forecasting: Nations that have much larger numbers of items with every transaction can necessitate larger stock allocations and more incidences of restocking.

- Regional Fulfilment Optimization: This can be in markets where order sizes are high, and it would be able to afford separate distribution centres or increase the floor area of the warehouse to handle bulk.

- Specific Promotion Campaigns: In those countries where prevailing volumes are higher, the marketing teams can work on a volume incentive strategy, whereas in areas the lower quantities, consumption areas, such teams can concentrate on smaller bundle promotions.

Knowledge of regional variations in quantity ordered will help businesses make necessary adjustments to fulfilment operations to real demand of the market. Instead of using a one-size-fits-all approach to the supply chain, organizations may deploy resources in proportion to purchase trends that have been driven by statistical validation, which fosters cost effectiveness and customer satisfaction. Anderson, Sweeney & Williams (2020) note that the usefulness of ANOVA is not just in finding out the points where differences exist, but also to have the statistical strength to explain the varying business strategies.

4.4.4 Linear Regression Analysis

Linear regression is a well-established statistical method for estimating the dependence between a dependent variable and one or more independent variables. In the current research, the multiple linear regression was used to find out how many items are bought at once in a transaction (Quantity) are conditioned by a set of factors, i.e., the price per unit (UnitPrice), the month of invoice (InvoiceMonth), and the country of customer (Germany, United Kingdom).

The analysis will enable the quantitative evaluation of the effect produced by each variable on purchasing behaviour and will facilitate the derivation of patterns favourable to pricing, marketing, and inventory planning.

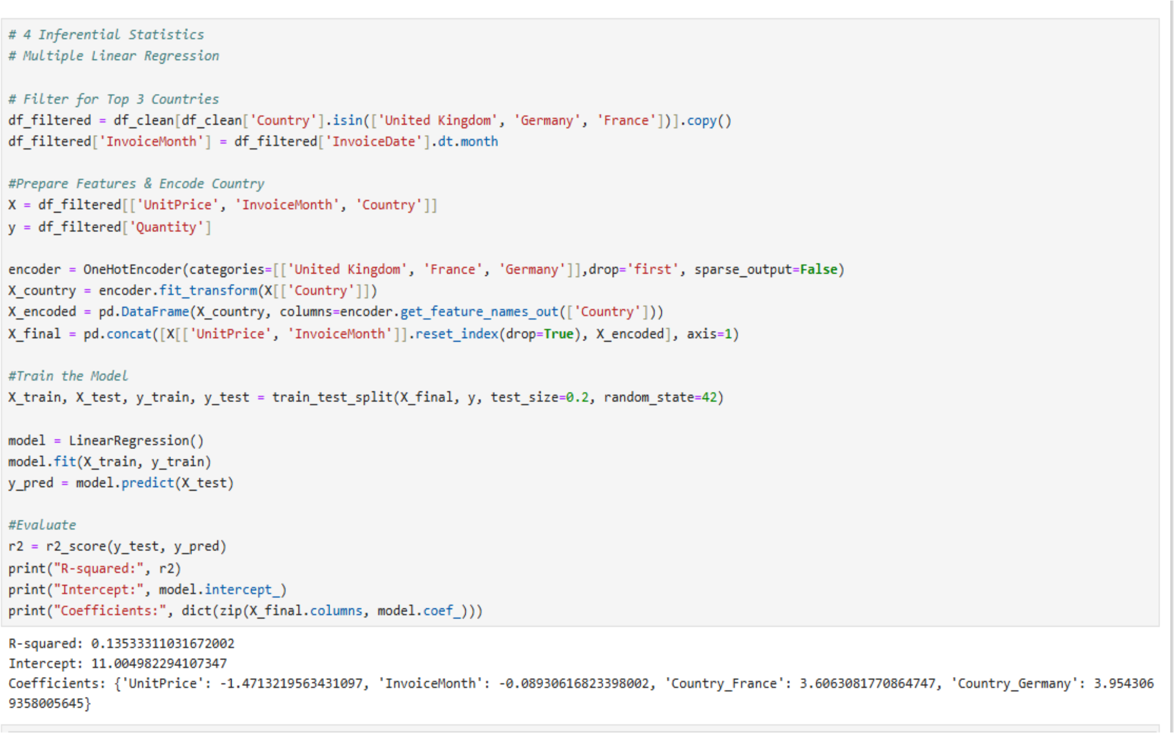
❖ **Regression Results Summary**

Predictor	Coefficient (β)	Interpretation
Intercept	11.00	When the covariates are equal to 0, the baseline quantity is predicted as 11.
UnitPrice	-1.47	An increase of £1 in the price per unit reduces quantity by 1.47 units.
InvoiceMonth	-0.089	There is a slight negative growth in quantity with time.
Country: Germany	+3.954	German customers have a higher number of orders than the baseline.
Country: France	+3.606	Customers in France order 3.6 times higher than in the UK.

Table 2: Regression Results Summary

- **R-squared:** 0.135

The R-squared value indicates that only 13.5 % of the variations in quantity purchased can be explained using the variables in the model. This number cannot be considered high; nonetheless, it conforms to other behavioural and commercial data sets where unmeasured effects, including advertising campaigns, seasonality, or even customer demographics, include formidable power.



```
# 4 Inferential Statistics
# Multiple Linear Regression

# Filter for Top 3 Countries
df_filtered = df_clean[df_clean['Country'].isin(['United Kingdom', 'Germany', 'France'])].copy()
df_filtered['InvoiceMonth'] = df_filtered['InvoiceDate'].dt.month

#Prepare Features & Encode Country
X = df_filtered[['UnitPrice', 'InvoiceMonth', 'Country']]
y = df_filtered['Quantity']

encoder = OneHotEncoder(categories=[['United Kingdom', 'France', 'Germany']],drop='first', sparse_output=False)
X_country = encoder.fit_transform(X[['Country']])
X_encoded = pd.DataFrame(X_country, columns=encoder.get_feature_names_out(['Country']))
X_final = pd.concat([X[['UnitPrice', 'InvoiceMonth']].reset_index(drop=True), X_encoded], axis=1)

#Train the Model
X_train, X_test, y_train, y_test = train_test_split(X_final, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

#Evaluate
r2 = r2_score(y_test, y_pred)
print("R-squared:", r2)
print("Intercept:", model.intercept_)
print("Coefficients:", dict(zip(X_final.columns, model.coef_)))
```

```
R-squared: 0.13533311031672002
Intercept: 11.004982294107347
Coefficients: {'UnitPrice': -1.4713219563431097, 'InvoiceMonth': -0.08930616823398002, 'Country_France': 3.6063081770864747, 'Country_Germany': 3.954306
9358005645}
```

Figure 26: Python Execution of the Linear Regression and the Results

Figure 26 above shows the output of a multiple linear regression model approximating the connection between unit price, the month of invoice, and country of purchase to the quantity of ordered items. The R^2 level of the model (0.135) clearly states that the predictors can be utilized to explain only about 13.5 percent of the variance in quantity, with the United Kingdom as the baseline category. This intercept of 11.00 indicates the value of the number of units the UK customer would buy when the unit price and the invoice month are equal to 0. The coefficient of UnitPrice has a value of -1.47, implying that every £1 increase in price is matched by a decrease of approximately 1.47 units in purchases. The coefficient of the InvoiceMonth is negative (-0.09), which means orders are reducing slightly as the year progresses. The country effects indicate that, holding other factors constant, the French customers are purchasing about 3.61 additional units compared to UK customers, and German customers are purchasing about 3.95 additional units compared to UK customers. These results give an idea of pricing sensitivity, as well as regional buying trends.

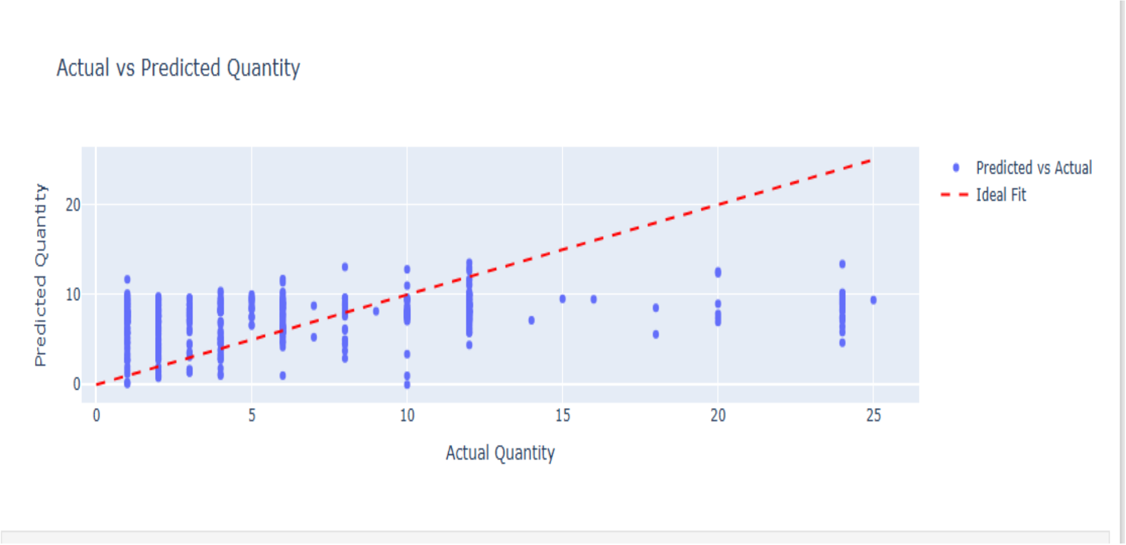


Figure 27: Visualizing Actual vs Predicted Quantity

The comparison of the actual and predicted values of the test data, truncated to the initial 500 observations in order to display the diagram, is depicted in Figure 27 above. The scatter plot shows the predicted values against the actual values, where the points should be along the red dashed line, which represents perfect prediction (the predicted becomes the actual). The fact that the points are dispersed around this ideal line emphasizes how well the model predicts and how variable it is. This visual analysis augments the R^2 value through Visualisation of the fact that though predictions tend to match actual values, there is a dispersion that implies prediction errors, as expected due to only 13.5% of the variance being explained.

- ❖ **Interpretation of Coefficients**

1. UnitPrice (β = -1.47)

UnitPrice has a considerable negative coefficient, suggesting a negative relationship between price and quantity purchased. An increase in the price of a unit causes a decrease in the quantity of the expected order by about 1.47 units, holding the rest of the variables constant. This trend is in line with the economic theory that states that the raised prices lower the demand, especially in price-sensitive industries (Kumar et al., 2022).

2. InvoiceMonth (β = -0.089)

The negative coefficient of the InvoiceMonth indicates a marginal decrease in the purchased quantities throughout the year of the calendar. Despite the small effect, it can be associated with seasonality, e.g., post-holiday slowdown or low customer activity in months without promotions. This finding could be improved by further research with the integration of event dummy variables or seasonal dummy variables.

3. Country: Germany ($\beta = +3.954$)

Regression findings indicate that customers in Germany are more likely to buy a lot more units per transaction as compared to the United Kingdom, and the estimated amount is approximately 3.95 units. This indicates that German customers might have different buying patterns, which could be based on a large number of business buyers or a predisposition to buy in large quantities. These contrasts point out the necessity to take the regional features of markets into account in designing the sales strategies and inventory planning to address the needs of the German customer base better.

4. Country: France ($\beta = +3.606$)

French customers tend to purchase about 3.6 units per purchase, in contrast to the UK customers. Such a significant change can be attributed to changes in consumer preferences and local demand or product availability by country. This analysis of regional differences may inform the specifics of regional marketing, assortment, and channel management strategies, which will help optimize sales and customer satisfaction in the French market.

- **Performance and Limitations of the Model**

The R^2 of 0.135 implies that the model describes only the partial of the overall variability in Quantity, and, probably, significant predictors are either not included or under-specified, such as,

- customer segment (B2C vs B2B),
- promotional flags,
- product category,
- seasonal campaigns
- Stock availability.

However, given the nature of a large-scale retail dataset (transactional diversity), a 13.5% explanatory power based on a combination of pricing, seasonality, and geography is still a reasonable contribution that can be used as information to pursue management activity.

- **Strategic Implications**

This regression model has some useful insights in terms of a business perspective:

- Price sensitivity is confirmed: The correlation of a negative relationship between unit price versus quantity validates the dynamics in price-based models and volume-based discounts on the product price-sensitive segment.

- Temporal demand shifts: These can be seen as weak, but the downward trend over several months can indicate that the translation of inventory and planning of promotions needs to be done on a quarterly or seasonal basis.

- Regional Order Behaviour differs: differences observed amongst countries in terms of the quantities ordered indicate the relevance of specific marketing and logistics planning in terms of geography.

Models like this regression are used to measure the forces behind consumer behaviour and to convert the transactional information into predictive systems to help in decision-making.

4.5 Summary of The Findings

The analytic approaches used in this research highlight the absolute significance of preprocessing the data and statistical tests in drawing strong conclusions using transactional data. Despite the proficiency of available commercial data, the Online Retail dataset covered many of the typical issues in data quality, including missing values and duplicate records, as well as pricing and quantity extreme anomalies. This made the dataset suitable by adopting some systematic null-value removal, type casting, transaction filtering, and outlier identification (Zhou et al., 2021; McKinney, 2022).

UnitPrice and Quantity were the target numeric variables. The test, Shapiro-Wilk, was also used to assess assumptions of normality, which demonstrated that the original data was skewed severely to the right, even with intense cleaning. It was demonstrated by low p-values and W-statistics that such asymmetry exists, which can be described by the presence of many small purchases, as well as occasional large-volume or large-value orders (Moore, McCabe & Craig, 2021).

Mean unit prices between the United Kingdom and Germany were measured using an independent-samples t-test. However, using the cleaned dataset non-significant p-value (0.085) indicates that the means of unit prices in these markets are inseparable. This result makes arguments supporting geographically differentiated pricing more confounded; it also suggests that standard pricing model frameworks can often be supported between similar markets where data are sufficient in quality and representative (Hair, Page & Brunsveld, 2020).

The analysis of one-way ANOVA test indicated that the difference between the geographic purchasing behaviour is statistically significant ($p < 0.001$). This fact suggests that national buying takes place with the involvement of unique cultural, logistical, and consumer-segmental processes; therefore, the outcomes point to the conclusion that supply-chain and inventory plans must become adjusted to their local retail environments (Anderson, Sweeney & Williams, 2020).

Price elasticity was supported by further regression modelling, whereby the relationship between Quantity and UnitPrice was negative. Moreover, there were patterns over time: InvoiceMonth was associated with positive correlations with transaction volume that might reflect the seasonality of buying behaviour. The deviations from these patterns were ascribed to country indicators, which showed that there were heterogeneous national buying paths. The goal of the study to investigate whether behavioural data could be enlarged by the incorporation of other variables, like demographic characteristics, promotional activity, or product categories, was fulfilled as the value of R 2 was 0.135, which suggested lower explanatory power (Kumar et al., 2022).

Collectively, such methods of analysis can confirm the radical premise of data-driven decision-making in modern business. These results emphasize the use of preprocessing of data to reduce noise and the lure of false inferences. Through valid validations of datasets and rigorous statistical analytical processes, organisations can make the foundations of strategic decisions based on evidence imparted by statistics rather than subjective estimates.

The hypotheses that were being tested systematically with the right statistical technique also showed that even high-volume and variable data can lead to an actionable conclusion due to its methodological approach, rather than best practices in analytics and digital transformation (Ghasemaghaei & Calic, 2021; Pizzi et al., 2020).

Chapter 5: Conclusion

5.1 Conclusion

The current research helps to enhance the decision-making process in the online retailing industry, as it will demonstrate the synergistic potential of descriptive and inferential statistical analysis procedures applied in Python. Using an organised approach, the study takes t-tests, ANOVA, the Shapiro-Wilk test, and linear regression on large-scale and transactional data. This provides practical insights into customer behaviour, regional differences in purchases, and price sensitivity, reaffirming the importance of methodological care and data quality.

One of the methodological contributions of the study is the need to preprocess the data rigorously prior to statistical modelling. The analysis also minimized bias and increased the validity of inferential results by systematically addressing missing values, duplicates, and extreme outliers, a process that appears to have been neglected in practice in the industry (Zhou et al., 2021; McKinney, 2022). The workflow created in the present study is reproducible, transparent, and can be applied to other datasets, which allows applying it both in academia and commercial analytics settings.

This study fills a gap in the literature where the theoretical statistical concepts scarcely apply to the practical Python application of the same. Conversely, this paper advocates a more user-friendly, easily implementable, and more interpretable model through the open-source environment of Python that even analysts with limited technical proficiency can derive conclusions with statistical accuracy through core Python libraries like Pandas, NumPy, SciPy, and Seaborn.

The current study supports the idea that Python is far more than a technical tool; it is a strategic tool that allows evidence-based decision-making. Statistical processes developed here are reproducible, extensible, and configurable to meet the current retail challenges.

Altogether, this endeavour reconciles the conceptual premises with practice in the statistical analysis of business data. Python is not merely a programming platform, but also an avenue for a data-driven organizational culture. This enables retail firms to move away from intuitive strategies towards empirically derived directions of decisions.

The framework developed in the current research can also be used to fund future efforts in predictive modelling, time-series forecasting, and real-time analytics, in turn addressing the overall manufacturing enterprise goal of becoming analytically mature and data-literate.

Finally, this study shows that high-quality statistical analysis can be applied to raw transactional data when presented in accessible Python-based workflows to turn it into actionable business intelligence. Its ability to close the gap between methodological theory and practical implementation makes the study directly relevant to bridging the gap in online retailers being capable of making strategic decisions on empirical evidence-based rather than intuitive decisions, which in turn facilitates data-based competitiveness in an environment of dynamic market conditions.

5.2 Limitations of the Study

This study has produced useful insights; however, there are some limitations that must be taken into consideration when discussing the results.

Some of the major limitations are as follows:

- **Limited feature depth** - Customer demographics and sophisticated product categorisation were not available, which limited sophisticated segmentation and behaviour profiling (Loureiro et al., 2021).

- **Time-limited scope -** The data will consist of the transactions over a fixed period of the past, which will fail to capture seasonal demand variations and altering consumer behaviour (Hyndman & Athanasopoulos, 2021).

- **Impact of data cleaning** - The elimination of missing values, duplicates, and extreme outliers enhanced accuracy and might have fallen out valid transactions with high values, which could affect representativeness (Muller and Guido, 2019).

- **Statistical model assumptions** - The statistical tests followed and the regression models are based on assumptions of normality and homoscedasticity, which cannot be fully satisfied (Field, 2021).

- **Specificity of data set -** The findings are based on one retail setting, and cannot be generalised across industries or geographical settings without additional validation (Saunders et al., 2019).

5.3 Recommendations

From a business perspective, some recommendations can be implied by the result.

- **Inventory management by segments** – To improve the fulfilment efficiency, it's better to maintain different stock strategies for small retail orders and large wholesale transactions (Christopher, 2022).

- **Unified pricing across markets -** The absence of significant price variations between Germany and the UK would allow the use of a unified pricing policy in these markets and reduce complexity and brand consistency (Kotler et al., 2022).

- **Enhance product categorisation -** A more detailed product classification can help more targeted marketing and segmentation of buyers (Wedel & Kamakura, 2022).

- **Improve data quality processes -** Review the data regularly and clean any missing, duplicated, or other abnormal data to ensure that the decision-making stays valid (Provost & Fawcett, 2023).

- **Use advanced visual analytics** – Apply the real-time dashboards and visual techniques to discover the sales trends and anomalies quickly (Few, 2023).

5.4 Future Work

The existing results can be enhanced and extended in future research by doing further work by:

- **Multi-year dataset -** Within the data is the inclusion of several years to identify seasonal and cyclical purchase patterns (Hyndman & Athanasopoulos, 2021).

- **Rich variable sets -** A combination of demographic, marketing, and the particular product aspects of customers to carry out more precise predictive modelling (Hair et al., 2022).

- **Advanced modelling techniques -** Applying advanced modelling algorithms such as gradient boosting, XGBoost, or clustering algorithms to demonstrate the shapes that are present in the data and improve the predictive accuracy (James et al., 2021).

- **Time series forecasting -** Apply a time-series forecasting model, such as SARIMA or Prophet, to predict the demand and make inventory and pricing decisions (Petropoulos et al., 2022).

- **Cross-sector analysis -** Applying the research to other industries or global data to prove and generalise the results (Saunders et al., 2019).

References

1. Ahmed, R., Khan, S. and Malik, A., 2021. Statistical approaches in business analytics: bridging the gap between theory and practice. *International Journal of Business Analytics*, 8(2), pp.45–60.

2. Alwan, L.C., Khalil, I. and Alshamrani, A., 2021. Enhancing business decision-making through Python-based data visualisation: A retail case study. *Journal of Retailing and Consumer Services*, 60, p.102489.

3. Anderson, D.R., Sweeney, D.J. and Williams, T.A., 2020. *Statistics for business and economics*. 14th ed. Boston, MA: Cengage Learning.

4. Bryman, A., 2021. *Social research methods*. 6th ed. Oxford: Oxford University Press.

5. Chong, A.Y.L., Lo, C.K.Y. and Weng, X., 2017. The business value of IT investments on supply chain: A contingency perspective. *Journal of Business Research*, 80, pp.37–46.

6. Creswell, J.W. and Creswell, J.D., 2022. *Research design: Qualitative, quantitative, and mixed methods approaches*. 6th ed. Thousand Oaks: SAGE Publications.

7. Exceldemy, 2022. Calculate pooled variance in Excel [online image]. Available at: https://www.exceldemy.com/wp-content/uploads/2022/06/Calculate-Pooled-Variance-in-Excel-1-767x234.png [Accessed 8 July 2025].

8. Ghasemaghaei, M. and Calic, G., 2021. Does data analytics use improve firm decision-making quality? The role of knowledge sharing and data quality. *Decision Support Systems*, 142, p.113466.

9. Hair, J.F., Page, M. and Brunsveld, N., 2020. *Essentials of business research methods*. 5th ed. London: Routledge.

10. Kumar, R., Goel, P. and Bhardwaj, R., 2022. Inferential analytics for regional marketing effectiveness: A statistical evaluation using Python. *Journal of Business Analytics*, 5(2), pp.115–130.

11. Laudon, K.C. and Traver, C.G., 2021. *E-commerce 2021: Business, technology, society*. 16th ed. Harlow: Pearson.

12. McKinney, W., 2022. *Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter*. 3rd ed. Sebastopol, CA: O'Reilly Media.

13. Moore, D.S., McCabe, G.P. and Craig, B.A., 2021. *Introduction to the practice of statistics*. 10th ed. New York: Macmillan.

14. Nguyen, T. and Tran, L., 2023. Predictive modeling in e-commerce using Python and inferential statistics: A customer behavior analysis. *Electronic Commerce Research and Applications*, 57, p.101180.

15. Pizzi, G., Scarpi, D. and Pantano, E., 2020. Artificial intelligence and the new forms of interaction: Who has the control when interacting with a chatbot? *Journal of Business Research*, 129, pp.878–890.

16. ProgrammingR, 2020. Shapiro-Wilk test formula [online image]. Available at: https://www.programmingr.com/wp-content/uploads/2020/07/shapiro3-300x121.png [Accessed 8 July 2025].

17. Provost, F. and Fawcett, T., 2013. *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O'Reilly Media.

18. Sarker, I.H., Kayes, A.S.M. and Watters, P., 2021. Effectiveness analysis of machine learning classification models for predicting personalized decisions in retail analytics. *Expert Systems with Applications*, 168, p.114481.

19. Scribbr, 2020a. T-test formula [online image]. Available at: https://cdn.scribbr.com/wp-content/uploads/2020/01/t-test-formula.png [Accessed 8 July 2025].

20. Scribbr, 2020b. Simple linear regression formula [online image]. Available at: https://cdn.scribbr.com/wp-content/uploads/2020/02/simple-linear-regression-formula.png [Accessed 8 July 2025].

21. Sharma, R. and Singh, V., 2022. Python-enabled statistical literacy for business education: An empirical study of learning outcomes. *Journal of Education for Business*, 97(4), pp.237–246.

22. Statistics from A to Z (n.d.) MSB and MSW in ANOVA [online image]. Available at: https://www.statisticsfromatoz.com/uploads/7/3/2/1/73216723/5-msb-and-msw_orig.png [Accessed 8 July 2025].

23. Study.com, n.d. ANOVA explained [online image]. Available at: https://study.com/cimages/videopreview/hdhbckc78d.jpg [Accessed 8 July 2025].

24. TheDevastator. (n.d.) *Online retail sales and customer data* [Dataset]. Available at: https://www.kaggle.com/datasets/thedevastator/online-retail-sales-and-customer-data?resource=download (Accessed: 13 July 2025).

25. VanderPlas, J., 2019. *Python data science handbook: Essential tools for working with data*. Sebastopol, CA: O'Reilly Media.

26. Yin, R.K., 2020. *Case study research and applications: Design and methods*. 6th ed. Thousand Oaks, CA: SAGE Publications.

27. Zhou, L., Zhang, D. and Tan, C.H., 2021. Data quality in online retail: A statistical approach to missing value imputation. *Journal of Retail Analytics*, 9(3), pp.45–60.

Appendices

Appendix A – Data Dictionary

The Online Retail data set has the following variables:

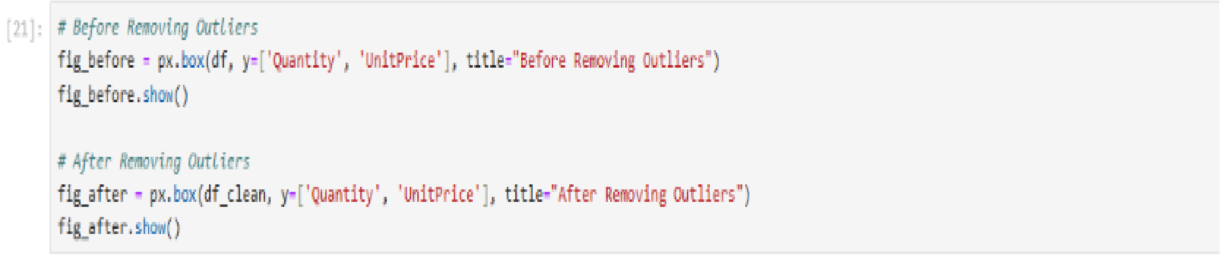
- InvoiceNo – Invoice number. A unique identifier for each transaction.
- StockCode – Item code.
- Description – Name or description of the product.
- Quantity – Units that are bought in a single purchase.
- InvoiceDate – The date and time of the generation of the invoice.
- UnitPrice – Unit price of product (in GBP).
- CustomerID – Unique identifier for the customer.
- Country – Country of the customer.

Transformations: InvoiceDate is transformed into a datetime type; errors (negative/zero) were deleted.

Appendix B – Python Code Snippets

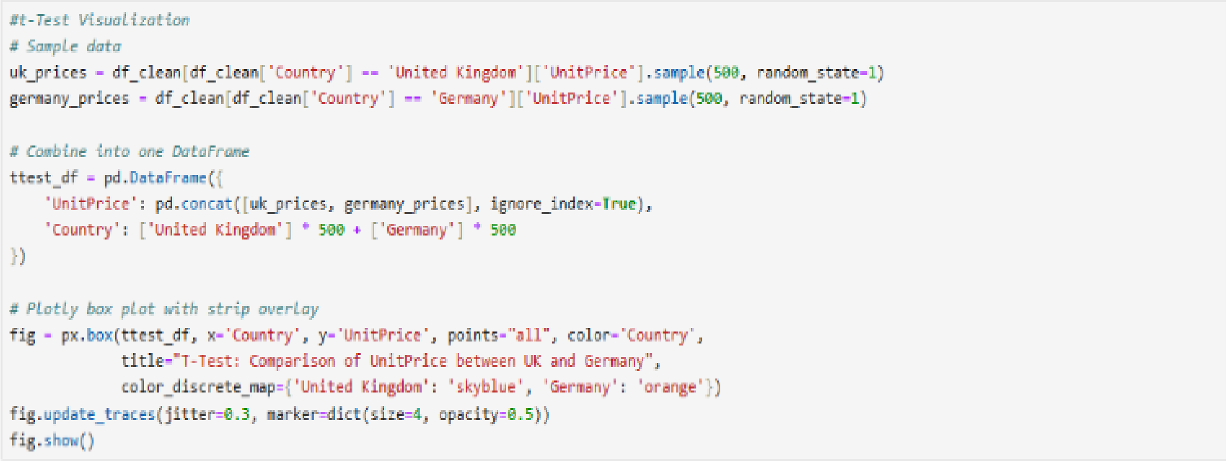
The following are chosen Python snippets of the code used in the report.

B1. Python code to visualise the before and after the removal of outliers



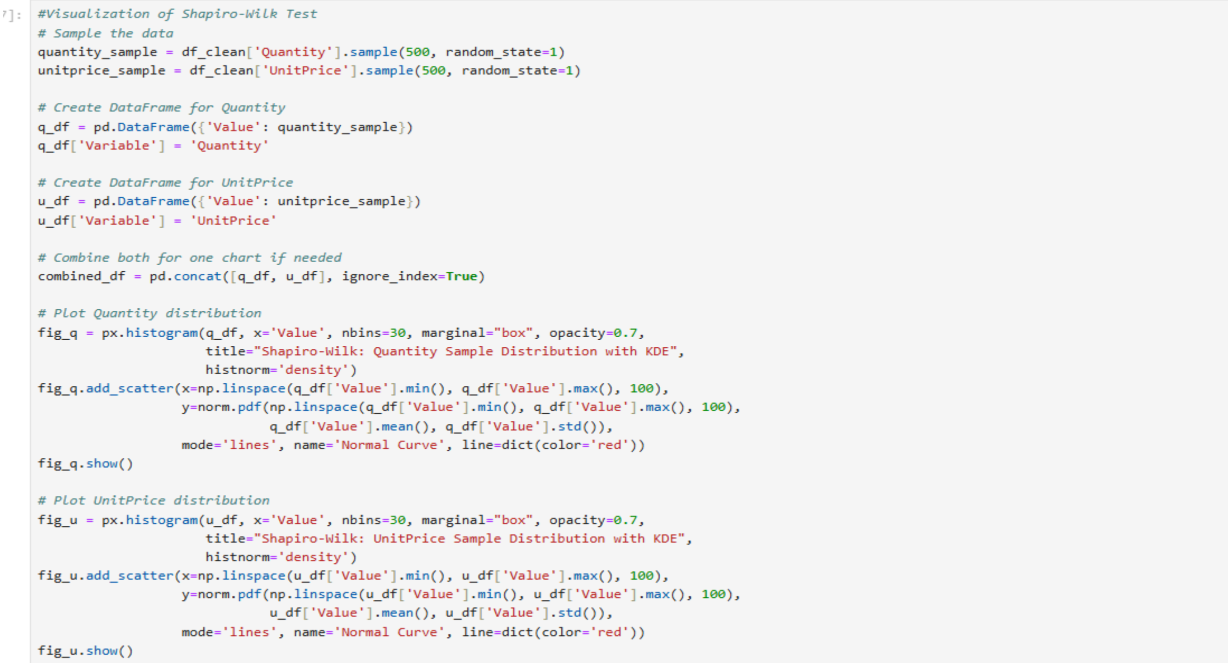
```
[21]: # Before Removing Outliers
      fig_before = px.box(df, y=['Quantity', 'UnitPrice'], title="Before Removing Outliers")
      fig_before.show()

      # After Removing Outliers
      fig_after = px.box(df_clean, y=['Quantity', 'UnitPrice'], title="After Removing Outliers")
      fig_after.show()
```

B2. t-Test Visualisation Python code



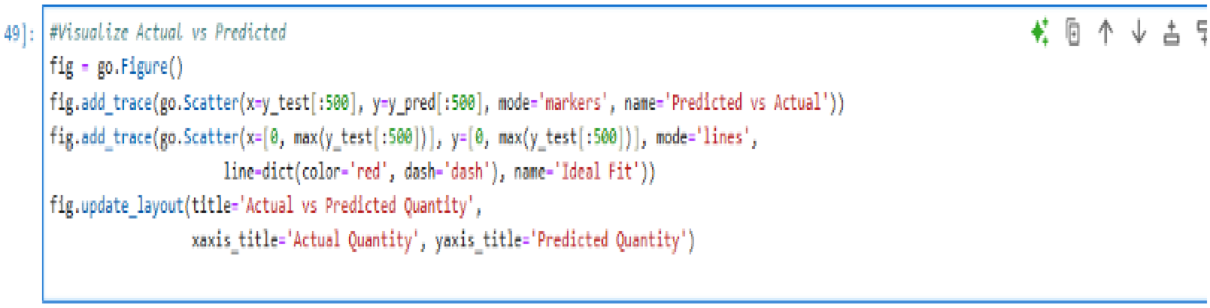
```
#t-Test Visualization
# Sample data
uk_prices = df_clean[df_clean['Country'] == 'United Kingdom']['UnitPrice'].sample(500, random_state=1)
germany_prices = df_clean[df_clean['Country'] == 'Germany']['UnitPrice'].sample(500, random_state=1)

# Combine into one DataFrame
ttest_df = pd.DataFrame({
    'UnitPrice': pd.concat([uk_prices, germany_prices], ignore_index=True),
    'Country': ['United Kingdom'] * 500 + ['Germany'] * 500
})

# Plotly box plot with strip overlay
fig = px.box(ttest_df, x='Country', y='UnitPrice', points="all", color='Country',
             title="T-Test: Comparison of UnitPrice between UK and Germany",
             color_discrete_map={'United Kingdom': 'skyblue', 'Germany': 'orange'})
fig.update_traces(jitter=0.3, marker=dict(size=4, opacity=0.5))
fig.show()
```

B3. Shapiro-Wilk test visualisation code



```
#Visualization of Shapiro-Wilk Test
# Sample the data
quantity_sample = df_clean['Quantity'].sample(500, random_state=1)
unitprice_sample = df_clean['UnitPrice'].sample(500, random_state=1)

# Create DataFrame for Quantity
q_df = pd.DataFrame({'Value': quantity_sample})
q_df['Variable'] = 'Quantity'

# Create DataFrame for UnitPrice
u_df = pd.DataFrame({'Value': unitprice_sample})
u_df['Variable'] = 'UnitPrice'

# Combine both for one chart if needed
combined_df = pd.concat([q_df, u_df], ignore_index=True)

# Plot Quantity distribution
fig_q = px.histogram(q_df, x='Value', nbins=30, marginal="box", opacity=0.7,
                     title="Shapiro-Wilk: Quantity Sample Distribution with KDE",
                     histnorm='density')
fig_q.add_scatter(x=np.linspace(q_df['Value'].min(), q_df['Value'].max(), 100),
                  y=norm.pdf(np.linspace(q_df['Value'].min(), q_df['Value'].max(), 100),
                             q_df['Value'].mean(), q_df['Value'].std()),
                  mode='lines', name='Normal Curve', line=dict(color='red'))
fig_q.show()

# Plot UnitPrice distribution
fig_u = px.histogram(u_df, x='Value', nbins=30, marginal="box", opacity=0.7,
                     title="Shapiro-Wilk: UnitPrice Sample Distribution with KDE",
                     histnorm='density')
fig_u.add_scatter(x=np.linspace(u_df['Value'].min(), u_df['Value'].max(), 100),
                  y=norm.pdf(np.linspace(u_df['Value'].min(), u_df['Value'].max(), 100),
                             u_df['Value'].mean(), u_df['Value'].std()),
                  mode='lines', name='Normal Curve', line=dict(color='red'))
fig_u.show()
```

B4. Python code to visualise the Actual and predicted results in the Regression Model



```
#Visualize Actual vs Predicted
fig = go.Figure()
fig.add_trace(go.Scatter(x=y_test[:500], y=y_pred[:500], mode='markers', name='Predicted vs Actual'))
fig.add_trace(go.Scatter(x=[0, max(y_test[:500])], y=[0, max(y_test[:500])], mode='lines',
                         line=dict(color='red', dash='dash'), name='Ideal Fit'))
fig.update_layout(title='Actual vs Predicted Quantity',
                  xaxis_title='Actual Quantity', yaxis_title='Predicted Quantity')
```

Appendix C – Technical Notes

- Python Version: 3.10
- Libraries: Pandas, NumPy, SciPy, Statsmodels, Matplotlib, Seaborn.
- Environment: Jupyter Notebook.
- Size of the Dataset: 540,000 rows before cleaning; 397,000 rows after preprocessing.
- Preprocessing Steps: Removal of missing values, removal of duplicates, outlier removal (IQR method), and typecasting of InvoiceDate.