**Capstone Group Project: E-Commerce Sales Analytics**

1. **Project Overview**

This capstone project serves as the culmination of your learning journey in **Statistics for Data Analytics**. You will apply the complete range of statistical and analytical techniques covered in this course to a real-world **e-commerce dataset**.

The objective is to replicate the end-to-end workflow of a professional data analyst:

* Clean and prepare messy business data.
* Apply descriptive and inferential statistical techniques.
* Build predictive models using regression and time-series methods.
* Derive meaningful **business insights** and recommendations.

You will complete the project using **Python** and submit your work via **Git/GitHub**. At the end, you will not only demonstrate mastery of statistical concepts but also showcase your skills in reproducible analytics and professional reporting.

1. **Dataset Description**

The dataset (please find the dataset named “synthetic\_retail\_data.csv” from the portal) contains approximately **9,500 e-commerce transactions** recorded during 2023. Each record represents a customer purchase and includes details on products, pricing, discounts, customer demographics, marketing channels, and purchasing behavior.

**Key Variables**

* **InvoiceNo** – Unique transaction identifier
* **CustomerID** – Unique customer identifier
* **Date** – Date of purchase (2023)
* **ProductCategory** – Electronics, Clothing, Home, Beauty, Sports, Toys
* **Quantity** – Number of items purchased
* **UnitPrice** – Price per item ($)
* **DiscountApplied** – Discount percentage (0–55%)
* **ReviewRating** – Customer rating (1–5 stars, some missing values)
* **IsFirstPurchase** – Indicator for new vs. returning customers
* **MarketingChannel** – Source of acquisition (Email, Organic, Referral, Ads, Social)
* **Country** – Customer country (USA, UK, Germany, France, Canada, Australia)
* **TimeOnSite** – Time spent on the website before purchase (seconds, some missing values)
* **ShippingCost** – Shipping fee ($)
* **ItemsInCart** – Items added to cart
* **PreviousSpending** – Historical customer spending ($)
* **BrowsingSessions** – Website visits prior to purchase
* **TotalAmount** – Final transaction value including shipping ($)

**Notable Characteristics**

* Seasonal patterns in sales (peaks in May, July, November, and December).
* Category differences in pricing and popularity.
* Presence of missing values in *ReviewRating* and *TimeOnSite*.
* Approximately 2% of transactions contain extreme outliers in *Quantity* or *UnitPrice*.

1. **Project Requirements**
2. **Data Preparation**

* Import and examine the dataset.
* Identify missing values and apply appropriate treatment.
* Detect and address outliers in dataset.
* Prepare data for analysis and modeling.

1. **Descriptive Statistics & Exploratory Analysis**

* Compute summary statistics (mean, median, mode, variance, standard deviation, IQR).
* Generate distribution plots (histograms, boxplots, scatterplots).
* Create pivot-style summaries (e.g., revenue by product category, revenue by country).

1. **Probability & Hypothesis Testing**

* Estimate key probabilities (e.g., likelihood of 5-star review, probability of order value > $1,000).
* Conduct hypothesis tests:
  + **Two-Sample t-test**: Compare mean spending between first-time and returning customers.
  + **ANOVA**: Test whether average spending differs across countries.
  + **Chi-square test**: Assess association between marketing channel and customer review ratings.

1. **Confidence Intervals**

* Construct a 95% confidence interval for average daily revenue.
* Construct a 95% confidence interval for average customer review rating.

1. **Correlation & Regression Analysis**

* Create a correlation matrix of numeric variables.
* Fit a **multiple linear regression** model to predict *TotalAmount* using predictors from the dataset.
* Interpret coefficients and identify the strongest drivers of revenue.

1. **Time Series Analysis**

* Construct a time series of daily or monthly sales.
* Apply **moving averages** and **exponential smoothing** to forecast future sales.
* Identify seasonal peaks and business trends.

1. **Deliverables**

You are required to submit a complete project via **GitHub** containing by October 11, 2025:

1. **Jupyter Notebook / Python Scripts**
   * Clean, well-documented code.
   * Logical structure aligned with project requirements.
2. **README.md File**
   * Project overview.
   * Dataset description.
   * Methods applied.
   * Summary of findings and key business insights.
3. **Interpretive Commentary or Report** (within the external Microsoft word document)
   * Explanations of results.
   * Business implications.

1. **Grading Rubric**

* **Correctness of Analysis** – 16%
* **Code Quality & GitHub Submission** – 8%
* **Interpretation of Results** – 10%
* **Clarity of Documentation & Presentation** – 6%

1. **Instructor’s Note**

Upon submission, we will conduct a **live session** demonstrating the same analyses step-by-step in **Excel**. This session will allow you to cross-validate your Python results, strengthen your conceptual understanding, and focus on translating statistical findings into actionable **business recommendations**.

Good Luck !

Getnet B. (PhD Cand.)