# 

# Capstone Project Proposal: Customer Segmentation in E-commerce

# Critical Thinking- Module 7– Week 7

# Arti Govardhan Yadav

# Colorado State University Global

# MIS 581-Capstone - Business Intelligence and Data Analytics

# Dr. Justin Bateh

# Table of Content

1. Executive Summary
2. Introduction
3. Dataset Description
4. Data Analytics Tools and Techniques  
   4.1 SAS (Statistical Analysis System)  
   4.2 R Programming Language  
   4.3 K-Means Clustering Technique
5. Visual Model of the Data  
   5.1 Data Dictionary  
   5.2 Example Visualizations
6. Rationale for Dataset Selection
7. Research Design
8. Method
9. Methodology  
   9.1 Quantitative Research  
   9.2 Data Characteristics  
   9.3 Research Methods and Tools
10. Limitations  
    10.1 Data Availability and Scope  
    10.2 Sample Bias  
    10.3 Temporal Limitations  
    10.4 Model Assumptions and Limitations  
    10.5 Feature Selection
11. Ethical Considerations  
    11.1 Data Privacy and Anonymization  
    11.2 Informed Consent  
    11.3 Data Security  
    11.4 Bias and Fairness
12. Data Analysis  
    12.1 Descriptive Statistics  
    12.2 Hypothesis Testing and Results
13. Insights and Future Considerations
14. Conclusion
15. GitHub Account Link
16. References

# Capstone Project Proposal: Customer Segmentation in E-commerce -The Dataset

The e-commerce industry, particularly platforms like Amazon, has seen exponential growth, especially with the rise in online shopping driven by global trends and changes in consumer behavior. Despite the convenience of e-commerce, businesses face challenges in understanding the vast and diverse customer base that these platforms attract. Customer segmentation is crucial for businesses to tailor their marketing strategies and improve customer engagement, yet many companies struggle to effectively segment their customers due to the large volume and complexity of transactional data.

This research focuses on customer segmentation within an e-commerce context, using Amazon's Sales Report data to identify patterns in purchasing behavior across different product categories and sales channels. By analyzing this data, the goal is to develop a deeper understanding of how customer behaviors vary based on the number of items ordered, the value of transactions, and the type of products purchased. This knowledge can be leveraged by businesses to personalize marketing campaigns, optimize product recommendations, and improve customer retention. This is especially relevant in today’s competitive digital economy, where personalized experiences can significantly impact customer loyalty and business success.

## Problem Statement

Businesses on platforms like Amazon are continually challenged to understand diverse customer behaviors and tailor their marketing strategies accordingly. Despite the wealth of customer data available, many e-commerce platforms lack effective strategies to segment customers based on purchasing behaviors and product preferences. This results in generic marketing efforts, reduced customer engagement, and lower overall revenue potential. The problem this research addresses is the need for better customer segmentation techniques that enable businesses to analyze patterns in Order Amount, Product Categories, and Sales Channels, thereby developing more targeted and effective marketing strategies.

## Dataset Selection and Justification

For this project, selected the Amazon Sales Report dataset, which contains 128,975 entries and 24 variables. This dataset includes key information on order amounts, quantity ordered, sales channels, product categories, and shipping details, among other variables. The dataset was chosen for several reasons:

* Comprehensive Data Coverage: The dataset provides a wide range of variables related to customer purchases, product types, and sales platforms. This variety enables a multi-faceted analysis of customer behavior across different segments.
* Real-World Relevance: As Amazon is one of the largest e-commerce platforms globally, understanding customer behavior on this platform has high practical relevance. The findings from this research can directly inform marketing strategies for businesses selling through Amazon and similar platforms.
* Opportunities for Segmentation: The dataset allows for the analysis of different segments of customers based on key purchasing variables such as Order Amount and Quantity Ordered, as well as product and channel preferences. This makes it ideal for identifying actionable insights into customer behavior.
* Data Volume: With nearly 129,000 entries, the dataset is large enough to provide statistically significant results, while also being manageable for analysis within the scope of this project.

This dataset provides a rich foundation for developing insights into customer segmentation, enabling a deeper understanding of purchasing patterns and how businesses can enhance their marketing efforts. By focusing on key variables like Product Category and Sales Channel, the research can help businesses personalize their strategies, improve customer engagement, and boost overall sales.

## Executive Summary

Customer segmentation in e-commerce is a critical strategy for understanding consumer behavior and personalizing marketing efforts. This project proposes a detailed analysis of customer data using SAS and R to uncover distinct customer segments. The dataset, sourced from an Amazon sales report, includes variables such as product categories, customer regions, and purchase quantities. Tools like SAS for data processing and R for statistical analysis and visualization are employed to ensure robust and actionable insights. The outcomes of this project will enable the organization to tailor its marketing strategies, improve customer satisfaction, and ultimately drive higher sales.

## Introduction

In the competitive world of e-commerce, understanding customer behavior is essential for businesses aiming to thrive. Customer segmentation allows organizations to categorize their customers into distinct groups based on various attributes, leading to more targeted marketing efforts. This project focuses on leveraging data analytics using SAS and R to segment customers of an e-commerce platform. By analyzing a comprehensive dataset of customer interactions, the project aims to identify patterns and behaviors that can inform more effective business decisions.

## Dataset Description

The dataset used in this project is an Amazon Sales Report that contains detailed information on customer transactions. The dataset is stored in CSV format and includes several key variables essential for performing customer segmentation:

* **Variable Types:**
  + **Binary Variables:** These include attributes such as whether a customer used a discount code (yes/no).
  + **Categorical Variables:** Examples include product categories (e.g., electronics, apparel), customer region (e.g., North America, Europe), and payment methods.
  + **Interval Variables:** These include numerical data such as the quantity of products purchased, the total price of transactions, and the discount percentage.

The dataset contains [mention the number of records] records, and it provides a comprehensive view of customer transactions across different product categories and regions. The variety of variables in the dataset makes it ideal for conducting segmentation analysis.

## Data Analytics Tools and Techniques

Effective customer segmentation requires robust data analysis and processing. This project utilizes two powerful data analytics tools, SAS and R, alongside the K-Means Clustering technique to analyze the Amazon Sales Report dataset comprehensively.

* SAS (Statistical Analysis System)

SAS is employed for its strong data management capabilities, ensuring the dataset is clean, consistent, and ready for analysis. The following steps outline its application in this project:

* Data Importation and Exploration: SAS efficiently imports the Amazon Sales Report CSV file and conducts an initial exploration to understand the data structure and content. Procedures like PROC IMPORT and PROC CONTENTS help in identifying variable types, detecting anomalies, and summarizing key statistics.
* Data Cleaning and Preprocessing: SAS addresses missing values, duplicates, and inconsistent data entries through data step processing and procedures such as PROC MEANS and PROC FREQ. For example, missing values in the 'Discount' variable are replaced with zero, and outliers in the 'Total\_Sales' variable are treated to prevent skewing the analysis. Categorical variables like 'Product\_Category' and 'Region' are encoded appropriately for subsequent analysis. A screenshot of a computer

  Description automatically generated
* Data Transformation: SAS transforms relevant variables to enhance their analytical utility. Calculations such as deriving 'Average\_Order\_Value' and 'Purchase\_Frequency' from existing data provide additional insights into customer behavior. The transformed dataset serves as a solid foundation for advanced analytical procedures conducted in R.
* R Programming Language

R is utilized for its advanced statistical analysis and powerful visualization capabilities, facilitating insightful and interpretable customer segmentation results:

* + Exploratory Data Analysis (EDA): R performs in-depth EDA using packages like dplyr and ggplot2, uncovering patterns and relationships within the data. Visualizations such as histograms, boxplots, and scatter plots illustrate the distribution and correlation of variables like 'Total\_Sales', 'Purchase\_Frequency', and 'Customer\_Lifetime\_Value'.
  + Data Normalization: Prior to clustering, R normalizes numerical variables to ensure equal weighting during analysis. Functions from the scale package standardize data, improving the accuracy and reliability of clustering results.
  + K-Means Clustering Implementation: R applies the K-Means Clustering algorithm, leveraging its efficient computational routines. The optimal number of clusters is determined using methods like the Elbow Method and Silhouette Analysis. The clustering process groups customers based on similar purchasing behaviors and characteristics, identifying distinct segments such as 'High-Value Loyal Customers' and 'Price-Sensitive Occasional Buyers'.
  + Results Visualization: Post-clustering, R generates comprehensive visualizations depicting customer segments. Cluster plots and heatmaps provide intuitive representations of segment profiles, aiding stakeholders in understanding and utilizing the insights effectively for strategic decision-making.
* K-Means Clustering Technique :

K-Means Clustering is the primary analytical technique employed to segment customers effectively:

* Technique Overview: K-Means is an unsupervised machine learning algorithm that partitions data into 'K' distinct, non-overlapping clusters based on feature similarity. It iteratively assigns data points to clusters by minimizing the within-cluster variance, leading to homogenous and well-separated groupings.
* Application in Project: In this project, key variables such as 'Total\_Sales', 'Purchase\_Frequency', and 'Average\_Order\_Value' serve as inputs for clustering. The algorithm identifies natural groupings within the customer base, revealing segments with unique purchasing patterns and behaviors.
* Benefits and Justification: K-Means is chosen for its simplicity, scalability, and effectiveness in handling large datasets like the Amazon Sales Report. It provides clear and actionable segments that can inform targeted marketing strategies, personalized promotions, and resource allocation decisions, ultimately enhancing customer satisfaction and business profitability.
* Evaluation of Clusters: The quality and validity of the clusters are evaluated using metrics like Silhouette Coefficient and Within-Cluster Sum of Squares (WCSS). These assessments ensure that the segments are meaningful and provide valuable insights for strategic applications.

## Visual Model of the Data

The dataset used for this project consists of various fields that capture critical information about customer transactions on Amazon. Below is a detailed data dictionary, followed by example visualizations that help in understanding and interpreting the data.

* **Data Dictionary**

This table provides a summary of the key variables in the dataset, including their types and descriptions.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Type** | **Description** |
| Order ID | Categorical | Unique identifier for each order. |
| Date | Categorical | Date of the order in MM-DD-YY format. |
| Status | Categorical | Order status (e.g., Shipped, Cancelled, Returned). |
| Fulfilment | Categorical | Fulfilment method (e.g., Easy Ship, Merchant). |
| Sales Channel | Categorical | Channel through which the sale was made (e.g., Amazon.in, Amazon.com). |
| Category | Categorical | Product category (e.g., Kurta, Top, Western Dress). |
| Qty | Interval | Quantity of items ordered. |
| Amount | Interval | Total amount of the transaction in local currency. |
| ship-city | Categorical | City where the order was shipped. |
| ship-state | Categorical | State where the order was shipped. |
| ship-country | Categorical | Country where the order was shipped. |
| promotion-ids | Categorical | Promotion IDs applied to the order. |
| B2B | Binary | Indicator whether the order was a B2B transaction (True/False). |
| fulfilled-by | Categorical | Entity responsible for fulfilling the order. |
| ship-postal-code | Interval | Postal code where the order was shipped. |
| Currency | Categorical | Currency in which the transaction was conducted. |
| Courier Status | Categorical | Status of the shipment with the courier service. |
| SKU | Categorical | Stock Keeping Unit, a unique identifier for each product. |
| Unnamed: 22 | Mixed | Unnamed column with mixed types; may require cleaning or exclusion. |

* **Example Visualizations** : To better understand the dataset and the results of the segmentation analysis, here are some visualizations that can be generated using R or SAS:
* Bar Chart: Order Status Distribution

## Figure 1

## *The screenshot of the* bar chart - Order Status Distribution of sales datasets*. (Screenshot A. Yadav)*

A screenshot of a computer

Description automatically generated

Note: Figure 1 represents the bar chart which will help to visualize how many orders fall under each status category (e.g., Shipped, Cancelled, Returned). Understanding the distribution of order statuses can help in identifying any significant issues with cancellations or returns. VBAR Status: This creates a vertical bar chart for the Status variable.FILLATTRS=(COLOR=steelblue): This option colors the bars in steel blue. XAXIS LABEL="Order Status" and YAXIS LABEL="Number of Orders": These lines label the axes of the chart.

## Rationale for Dataset Selection

The chosen dataset is highly relevant to the problem statement, which focuses on improving marketing effectiveness through customer segmentation. The dataset includes a wide range of customer data, making it ideal for segmentation analysis. By identifying different customer segments, the organization can develop more targeted marketing strategies, improving the allocation of marketing resources and enhancing the overall customer experience. This project addresses the problem of generalized marketing strategies by enabling the company to target specific customer segments, thereby increasing the effectiveness of its marketing campaigns.

# Methods

In today's rapidly evolving e-commerce industry, understanding customer behavior is essential for companies aiming to remain competitive and optimize their marketing efforts. With a vast number of customers interacting with platforms like Amazon, it becomes critical for businesses to segment their customer base effectively to target specific groups more efficiently. Customer segmentation allows businesses to identify patterns in purchasing behavior, enabling them to deliver personalized marketing campaigns, recommend tailored products, and improve overall customer satisfaction. This Capstone Project focuses on utilizing customer data to segment customers into distinct groups, which can help drive targeted marketing strategies and resource allocation.

The Amazon Sales Report dataset, which contains detailed information on customer transactions, product categories, and sales channels, provides a rich source of data for conducting this segmentation analysis. By analyzing this dataset, we aim to identify customer segments based on key behavioral factors such as total spending, frequency of purchases, and average order value. These segments will enable the organization to tailor its strategies to each customer group, ultimately leading to better business outcomes, such as increased revenue and customer retention.

In an industry as vast and varied as e-commerce, not all customers have the same purchasing habits, preferences, or motivations. Customer segmentation divides the broader customer base into smaller, more manageable groups that exhibit similar characteristics, making it easier for companies to address the unique needs of each segment. For example:

* High-Value Customers may be more inclined to respond to premium loyalty programs and exclusive offers.
* Price-Sensitive Customers may react better to discounts and limited-time offers.
* Occasional Buyers may require more frequent touchpoints or engagement strategies to increase their purchasing frequency.

For e-commerce companies like Amazon, understanding these nuances in customer behavior is critical for optimizing marketing spend, improving customer experience, and driving long-term profitability. Businesses that use customer segmentation effectively are better positioned to deliver relevant promotions, recommend the right products, and invest their marketing resources more efficiently. This project is aligned with these goals and seeks to uncover actionable insights through data-driven analysis.

## Research Questions

The research questions (RQs) for this project are formulated to address the key business concerns regarding customer behavior in the e-commerce setting. These questions guide the analysis and ensure that the results provide actionable insights for improving customer segmentation and marketing strategies.

**Research Questions:**

* **RQ1:** Is there a significant relationship between a customer’s total spending and their purchase frequency?
* **RQ2:** Does the average order value differ significantly across different customer segments?
* **RQ3:** Can customers be segmented effectively based on their purchasing behavior, including variables such as quantity ordered, total amount spent, and frequency of purchases?

## Hypotheses Overview

Each research question is associated with specific hypotheses that will be tested using statistical methods. These hypotheses are formulated to determine whether there are statistically significant relationships between the variables in question.

**Hypotheses for RQ1:**

* Null Hypothesis (Ho1): There is no significant relationship between a customer’s total spending and their purchase frequency.
* Alternative Hypothesis (Ha1): There is a significant positive relationship between a customer’s total spending and their purchase frequency.

**Hypotheses for RQ2:**

* Null Hypothesis (Ho2): There is no significant difference in the average order value across different customer segments.
* Alternative Hypothesis (Ha2): There is a significant difference in the average order value across different customer segments.

**Hypotheses for RQ3:**

* Null Hypothesis (Ho3): Customers cannot be effectively segmented based on their purchasing behavior (quantity ordered, total amount spent, and purchase frequency).
* Alternative Hypothesis (Ha3): Customers can be effectively segmented based on their purchasing behavior (quantity ordered, total amount spent, and purchase frequency).

Methodology

The methodology section outlines the overall approach to conducting the research, including the types of data used, the analytical techniques employed, and the rationale behind the research methods chosen. This project employs a quantitative research design to analyze customer behavior using a dataset from the Amazon Sales Report. Quantitative research is ideal for this project because it allows for statistical analysis and the application of machine learning algorithms, such as K-Means clustering, to segment customers based on objective data.

## Why Quantitative Research?

* Objective Data: Quantitative research focuses on analyzing numerical data, which is critical in this project since the Amazon Sales Report contains numerical transactional data such as purchase amounts, order quantities, and frequencies. These data points allow for precise statistical analysis, supporting the identification of trends and patterns in customer behavior.
* Generalizability: The results derived from quantitative research tend to be more generalizable across different datasets and contexts, especially when large datasets like the one used in this project are involved.
* Scalability: Quantitative research methods, including the use of machine learning algorithms, can be scaled to analyze large datasets effectively, making them particularly suited for e-commerce platforms with massive amounts of transactional data.

The study is descriptive in nature, as it seeks to describe patterns of behavior (e.g., purchasing frequency, total spending) within the customer base. It is also predictive, as it uses machine learning techniques to predict and group customers into segments based on similar behaviors, thus offering insights into future marketing strategies.

## Type of Data

The Amazon Sales Report dataset serves as the foundation for this study. This dataset includes customer transaction data from Amazon, which is highly structured and quantitative in nature. The data consists of detailed customer purchase records, including variables such as Order ID, Quantity Ordered, Total Purchase Amount, Product Category, and Geographic Information (e.g., city, state).

## Data Characteristics:

* Structured Data: The data is structured, meaning it is presented in a well-organized format (e.g., CSV or database tables) that is easily processed by statistical software and machine learning algorithms.
* Transactional Data: The dataset consists of transaction-level records, capturing customer behavior in terms of product purchases, quantities, and order values. This data provides insights into customer preferences and purchasing patterns.
* Quantitative Variables: The dataset primarily contains quantitative variables, such as the number of items purchased (Qty), total amount spent (Total\_Sales), and frequency of purchases. These variables can be used for statistical analysis and clustering.

## Types of Variables in the Dataset:

* Continuous Variables: These include numerical data such as Total Sales, Average Order Value, and Quantity Ordered, which can take any value within a range.
* Categorical Variables: These include non-numerical data like Product Category, Geographic Location, and Order Status (e.g., shipped, canceled). These variables can be used to classify and group customers.
* Binary Variables: Some binary variables in the dataset, such as whether a customer used a promotion or if an order was a business-to-business (B2B) transaction, will be included in the analysis as they provide useful insights into customer segmentation.

## Research Methods and Tools

Given the quantitative nature of the data, this project employs both traditional statistical techniques and machine learning algorithms to test the hypotheses and segment the customer base.

### Descriptive Statistics

* + Purpose: Descriptive statistics will be used to summarize the data, providing an overview of key metrics like mean, median, and standard deviation for variables such as Total Sales, Quantity Ordered, and Average Order Value.
  + Tools: SAS will be used to generate these statistics using procedures like PROC MEANS and PROC FREQ. Descriptive statistics help in understanding the distribution of key variables and detecting outliers or missing data.

### Correlation Analysis

* **Purpose**: Correlation analysis will be conducted to explore the relationship between variables, such as whether there is a positive correlation between Total Sales and Purchase Frequency.
* **Tools**: Pearson’s correlation will be applied if the data is normally distributed, and Spearman’s rank correlation will be used if the data is non-parametric. This will be performed using SAS (PROC CORR) or R (cor() function).

## Machine Learning Method: K-Means Clustering

The core methodology for segmenting customers in this project is the K-Means Clustering algorithm, a machine learning technique that groups data points into clusters based on their similarities. The objective is to segment customers based on their purchasing behaviors and other characteristics, such as Total Sales, Purchase Frequency, and Average Order Value.

**Why K-Means Clustering?**

* Unsupervised Learning: K-Means is an unsupervised machine learning algorithm, meaning it does not require labeled data. This makes it well-suited for customer segmentation, where the goal is to discover underlying patterns and groupings in the data without prior knowledge of segment labels.
* Simplicity and Scalability: K-Means is relatively simple to implement and highly scalable, allowing it to handle large datasets like the Amazon Sales Report.
* Feature Selection: Variables such as Total Sales, Average Order Value, and Purchase Frequency will serve as the features that the K-Means algorithm uses to group customers into segments. These features provide a clear understanding of customer behaviors and purchasing patterns.

## K-Means Algorithm Process:

* Preprocessing: Data will be normalized before applying K-Means clustering to ensure that all variables contribute equally to the distance calculations. This is important because variables like Total Sales and Purchase Frequency may have different ranges.
* Selecting the Number of Clusters: The optimal number of clusters will be determined using methods such as the Elbow Method or Silhouette Analysis. These methods help evaluate the quality of clustering by analyzing within-cluster variance and cohesion.
* Clustering Analysis: The kmeans() function in R will be used to segment customers into clusters based on their behavioral attributes. Each customer will be assigned to one of the predefined clusters, with similar customers grouped together based on their purchasing behaviors.
* Evaluation of Clusters: The resulting clusters will be evaluated to determine their internal consistency and relevance. Metrics such as Silhouette Score will be used to assess the quality of the clustering.

## Data Collection and Processing

The dataset has been pre-collected and made available as a CSV file, which will be imported into both SAS and R for analysis. The SAS environment will handle the initial data cleaning and transformation tasks, including:

* Data Cleaning: Removing missing values, handling outliers, and transforming variables as necessary.
* Data Transformation: Creating new variables like Purchase Frequency and normalizing numeric variables.

Once the data is preprocessed, it will be exported to R for advanced machine learning analysis. In R, the K-Means clustering algorithm will be applied to group customers into segments, which will then be analyzed for insights.

# Limitations

While this project aims to provide valuable insights into customer segmentation for e-commerce using data analytics, several limitations must be acknowledged. Understanding these limitations is crucial for interpreting the results accurately and applying them to business strategies with appropriate caution.

## Data **Availability** and Scope

The data for this study comes from a single dataset, the Amazon Sales Report, which limits the generalizability of the findings. The dataset represents customer transactions within a specific e-commerce platform (Amazon) and may not capture the full diversity of customer behaviors present in other e-commerce contexts. For example, smaller e-commerce platforms or niche retailers may have different customer profiles or purchasing behaviors that are not reflected in the Amazon dataset.

* Impact: The results may be specific to Amazon’s customer base and may not generalize to other platforms. For instance, customer preferences, product choices, and spending behaviors on a niche clothing site might differ from those on a large, multi-category platform like Amazon.
* Mitigation: Future research could incorporate datasets from multiple e-commerce platforms to enhance generalizability.

## Sample Bias

The dataset may have inherent biases based on the geographic distribution of customers, product offerings, or promotional strategies employed by Amazon at the time the data was collected. For instance, certain regions or demographic groups might be over-represented, leading to customer segments that reflect only a subset of Amazon’s overall customer base.

* Impact: Segments identified in the analysis may be skewed toward particular regions or customer demographics, potentially limiting the applicability of marketing strategies to a broader customer base.
* Mitigation: Additional data sources or stratified sampling methods could help balance representation and reduce the impact of bias.

## Temporal Limitations

The dataset used in this project represents a snapshot of customer behavior over a limited time period. Since customer preferences and behaviors evolve over time, this cross-sectional dataset may not capture long-term trends, seasonal variations, or shifts in buying patterns due to external factors such as market conditions or global events (e.g., the COVID-19 pandemic).

* Impact: Insights derived from the data may not account for long-term changes in customer behavior, potentially leading to strategies that are less effective in the future.
* Mitigation: A longitudinal study using time-series data could better capture changes in customer behavior over time, improving the accuracy and relevance of the segmentation.

## Model Assumptions and Limitations

The K-Means clustering algorithm used in this project has its own limitations. One key challenge is determining the optimal number of clusters. The number of clusters must be pre-defined, and while methods like the Elbow Method or Silhouette Analysis can assist, the choice is ultimately somewhat subjective. Additionally, K-Means assumes that clusters are spherical and equally sized, which may not always align with real-world data.

* Impact: Incorrect assumptions or suboptimal cluster numbers could lead to inaccurate segmentation, making it difficult to draw actionable insights from the results.
* Mitigation: Testing different clustering algorithms (e.g., Hierarchical Clustering or DBSCAN) could provide more flexibility and accuracy in identifying natural customer groupings.

## Feature Selection

The effectiveness of the clustering analysis is highly dependent on the selection of features (e.g., Total Sales, Purchase Frequency, Product Category). The choice of these variables impacts the quality of the segmentation results. However, some important customer attributes, such as customer satisfaction or brand loyalty, are not available in the dataset, potentially limiting the insights that can be derived.

* Impact: Focusing only on transactional data may result in customer segments that are incomplete or less actionable because they do not account for softer metrics like loyalty, preferences, or satisfaction.
* Mitigation: Including additional data points such as customer feedback or brand interaction data could lead to more comprehensive and meaningful customer segments.

# Ethical Considerations

In any data-driven project involving customer data, ethical considerations must be carefully examined to ensure the privacy, security, and fairness of the analysis. The following ethical concerns are particularly relevant to this project, which involves the use of customer transactional data from the Amazon Sales Report.

## Data **Privacy** and Anonymization

The dataset used in this project includes detailed customer transactions, which may contain sensitive information such as geographic locations, purchase patterns, and potentially identifying information (e.g., postal codes). Data privacy is a significant concern, especially with regulations like the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in the United States, which protect customer rights regarding data use and access.

* Mitigation:
  + Personal identifiers (e.g., names, addresses, contact information) will be anonymized or removed before analysis to ensure customer privacy.
  + The dataset should contain only non-identifiable transactional data (e.g., total spending, quantity purchased, product category), and customer IDs will be replaced with anonymized identifiers.
  + Compliance with privacy laws such as GDPR and CCPA will be maintained throughout the project by ensuring that the dataset is anonymized and used only for the purposes of this research.

## Informed Consent

In real-world applications, it is crucial to ensure that customers are informed about how their data is being collected and used. Informed consent means that customers should be aware that their transactional data might be analyzed for purposes such as improving marketing strategies or personalizing offers.

* Mitigation:
  + Since the dataset in this project is pre-existing and anonymized, direct consent from customers is not feasible. However, it is essential to ensure that the data was originally collected in a way that aligns with ethical guidelines and regulations.
  + For future applications, businesses should ensure that customers are made aware of how their data will be used, including the potential for analytics and segmentation.

## Data Security

Handling customer data, even in an anonymized form, requires stringent data security measures to protect against unauthorized access or misuse. Data breaches or leaks could lead to reputational damage, legal consequences, or harm to the affected individuals.

* Mitigation:
  + Encryption will be used to protect data during transfer and storage.
  + Access to the dataset will be restricted to authorized personnel, and secure storage environments will be used to prevent unauthorized access.
  + Regular audits and adherence to security protocols will ensure that the data is handled responsibly.

## Bias and Fairness

Machine learning algorithms, including K-Means clustering, can unintentionally reinforce biases present in the data. For example, if certain demographic groups are overrepresented in the dataset, the resulting customer segments may disproportionately reflect their behaviors and preferences, leading to biased marketing strategies or promotions.

* Mitigation:
  + Care will be taken to ensure that the analysis does not perpetuate existing biases in the data, such as geographic or demographic biases.
  + Techniques such as stratified sampling and bias detection algorithms will be employed to ensure that the clusters are fair and representative of the entire customer base.
  + Future studies could include demographic data (with proper anonymization) to assess whether any bias is present in the segmentation process.

## Ethical Use of Findings

The insights derived from customer segmentation must be applied responsibly, ensuring that they benefit both the business and its customers. Ethical concerns arise when businesses use segmentation to exploit vulnerable customer groups or manipulate purchasing behaviors.

* Mitigation:
  + The results of this segmentation should be used to enhance customer experience through personalized, relevant offers rather than manipulative tactics.
  + Any marketing strategies developed from the segments should be transparent and designed to meet customer needs without pressuring them into unwanted purchases.
  + Ethical marketing practices, such as truthful advertising and respect for customer autonomy, will be upheld in any business applications derived from this research.

# Data Analysis-Module 6

The analysis of the Amazon Sales Report dataset aimed to uncover insights into the relationship between order amount, quantity ordered, product categories, and sales channels. Using statistical techniques such as descriptive analysis, correlation, and ANOVA, we tested several hypotheses to evaluate customer behavior and purchasing patterns.

The dataset contained 128,975 entries with 24 variables, including key information such as Order Amount, Quantity Ordered, Product Category, Sales Channel, and Fulfillment Details. The analysis focused on the following research questions and hypotheses:

## Hypothesis Testing

Tested three hypotheses related to customer behavior:

1. **Hypothesis 1: Relationship Between Order Amount and Quantity Ordered**
   * **H₀₁**: There is no significant relationship between order amount and quantity ordered.
   * **Hₐ₁**: There is a significant relationship between order amount and quantity ordered.
2. **Hypothesis 2: Differences in Order Amount Across Product Categories**
   * **H₀₂**: There is no significant difference in the order amount across different product categories.
   * **Hₐ₂**: There is a significant difference in the order amount across different product categories.
3. **Hypothesis 3: Differences in Order Amount Across Sales Channels**
   * **H₀₃**: There is no significant difference in the order amount across different sales channels.
   * **Hₐ₃**: There is a significant difference in the order amount across different sales channels.

## Descriptive Statistics

The first step in the analysis involved calculating the basic descriptive statistics for key variables such as Order Amount, Quantity Ordered, and Sales Channel. This helped provide a broad overview of the data.

## Figure 2

*The screenshot of the summary statistics of sales datasets. (Screenshot A. Yadav)*A screenshot of a computer

Description automatically generated

*Note:* The Figure 2 provide appears to represent sales statistics categorized by sales channels (such as Amazon.in and Non-Amazon) and different product categories (Blouse, Bottom, Dupatta, etc.). For each category, certain metrics such as the number of observations (N Obs), amount, quantity (Qty), mean, standard deviation (Std Dev), minimum, and maximum values are reported. Amazon has a large number of transactions for multiple categories, including kurta (49,826 transactions), Set (50,224), and Top (10,618). The Mean values for both Amount and Quantity vary by category. For instance, Sarees tend to have a higher mean amount (₹799.57) than Blouses (₹520.33). The Maximum values reflect occasional large transactions, such as a maximum sale of ₹5,584 for a Set on Amazon.in.The data for Non-Amazon channels shows fewer observations (e.g., only 4 for Blouses), and the amount data is missing (represented by ".").

**Hypothesis Testing and Results**

## Hypothesis 1: Correlation Between Order Amount and Quantity Ordered

To explore the relationship between Order Amount and Quantity Ordered, a Pearson correlation analysis was conducted.

Null Hypothesis (H₀₁): There is no significant relationship between order amount and quantity ordered.  
Alternative Hypothesis (Hₐ₁): There is a significant relationship between order amount and quantity ordered.

## Figure 3

*The screenshot of the scatter plot result of Amazon sales datasets. (Screenshot A. Yadav)*

# A screenshot of a computer Description automatically generated

Note : This scatter plot represents Amazon sales data, specifically showing the relationship between order amount (on the x-axis) and quantity of items ordered (on the y-axis). The data points reflect individual customer transactions, where each point on the plot represents a specific purchase. The goal is to analyze the relationship between these two variables in the context of customer segmentation and to support or refute the hypothesis that order amount is correlated with quantity ordered.

The scatter plot provides visual evidence suggesting that the relationship between order amount and quantity ordered is weak or non-linear. While there are transactions where higher quantities are associated with higher order amounts, there are also many instances where large amounts are spent on very few items, indicating that other factors, such as the price of individual items, play a significant role in determining the total order amount.

* No clear upward trend: The lack of a clear, upward trend across the plot suggests that increasing the number of items ordered does not consistently result in a proportionate increase in the order amount.
* Outliers and variability: Some transactions with high order amounts correspond to low quantities, which weakens the case for a direct linear relationship between the two variables.

The scatter plot, it appears that there is no strong linear correlation between the amount of money spent and the quantity of items ordered. Instead, the data suggests that factors such as product price significantly influence the order amount, particularly for transactions involving high-priced items. The plot shows significant variability in the relationship, supporting the hypothesis that order amount is not solely dependent on quantity ordered.

To confirm this observation, statistical tests (such as calculating the Pearson correlation coefficient) would be necessary. However, from a visual standpoint, this plot suggests that the null hypothesis (H₀₁)—that there is no significant relationship between order amount and quantity ordered—cannot be rejected based on the current data visualization.

## Hypothesis 2: Differences in Order Amount Across Product Categories

To determine whether Order Amount varied significantly across different Product Categories, an ANOVA (Analysis of Variance) test was conducted.

* Null Hypothesis (H₀₂): There is no significant difference in the order amount across different product categories.
* Alternative Hypothesis (Hₐ₂): There is a significant difference in the order amount across different product categories.

## Figure 4

*The screenshot of the ANOVA result of Amazon sales datasets. (Screenshot A. Yadav)*A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

*Note:* Figure 4 represents the ANOVA test conducted to determine whether the order amount differs significantly across various product categories. The key statistical results are as follows:

F-statistic: 9715.17and p-value: < 0.0001.Since the p-value is much smaller than the standard significance level (0.05), we reject the null hypothesis (H₀₂) and accept the alternative hypothesis (Hₐ₂). This indicates that there are significant differences in the order amount across the different product categories.

* Levene’s Test for Equality of Variances:Levene's test (p < 0.0001) shows that the assumption of equal variances is violated, suggesting that different categories have varying degrees of variability in order amounts. Therefore, the variances in order amounts are not consistent across categories, and this was addressed using Welch’s ANOVA, which is more appropriate when variances are unequal.
* Welch's ANOVA Results:Welch’s ANOVA further confirmed the significant differences across categories:F-value (Welch's ANOVA): 12,482.7 and p-value: < 0.000.This reinforces the conclusion that the order amount varies significantly between product categories.
* Means of Order Amount Across Categories: The average order amounts for each product category are as follows:Blouse: 520.33,Bottom: 358.73,Dupatta: 305.00,Ethnic Dress: 723.90,Saree: 799.57,Set: 833.39,Top: 526.10,Western Dress: 762.79,Kurta: 455.93
* These means show that some categories, like Set, Saree, and Western Dress, have much higher order amounts compared to categories like Bottom and Dupatta.
* Post-Hoc Analysis (Tukey-Kramer Test):The Tukey-Kramer post-hoc test was performed to determine which specific product categories differ from each other. This test revealed significant differences between many pairs of categories. For example:Set has a significantly higher order amount compared to Kurta (p < 0.0001).Saree and Ethnic Dress also have significantly different order amounts compared to Bottom and Blouse.

This indicates that not only are the overall differences across categories significant, but many individual categories differ from each other in terms of average order amount.

Based on the ANOVA and subsequent post-hoc tests, we reject the null hypothesis (H₀₂) and conclude that there are significant differences in the order amount across product categories. This means that the product category has a meaningful impact on the amount spent on orders, with categories like Set, Saree, and Western Dress associated with higher order amounts compared to others like Dupatta and Bottom.

These findings have practical implications for business strategies, as product categories should be considered when setting prices, managing inventory, and devising marketing strategies to maximize order amounts.

## Hypothesis 3: Differences in Order Amount Across Sales Channels

An ANOVA test was also conducted to evaluate whether Order Amount differs across various Sales Channels, such as Amazon.in, Amazon.com, etc.

Null Hypothesis (H₀₃): There is no significant difference in order amount across different sales channels.

Alternative Hypothesis (Hₐ₃): There is a significant difference in order amount across different sales channels.

## Figure 5

*The screenshot of the ANOVA result of Amazon sales datasets. (Screenshot A. Yadav)*

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Note: The Figure 5 represents the Interpretation of the ANOVA Results for Hypothesis 3,there are two levels for the "Sales Channel" variable: Amazon.in and Non-Amazo (appears to be a placeholder for other sales channels or one category).The total number of observations read is 128,975, and 121,180 observations were used in the analysis. R-Square (0.0000): The R-squared value is zero, indicating that 0% of the variability in order amount can be explained by differences between sales channels. In other words, sales channels do not account for any variatiown in the order amounts.Coefficient of Variation (43.36): This measures the dispersion of the data relative to the mean. A high value (43.36%) indicates a high level of variability in order amounts.Root Mean Square Error (281.21): This indicates the standard deviation of residuals (errors) from the predicted values.Amount Mean (648.56): The average order amount across the sales channels is 648.56. ANOVA Table: F Value (.) and Pr > F (.): The results for the "Model" and "Sales Channel" in the ANOVA table are missing (denoted as .), indicating that there is no computed F-value or p-value for this test.

* + This likely happened because the degrees of freedom for the "Model" are zero (DF = 0), meaning there is no variation to be analyzed across the sales channels.

As a result, there is no evidence of a significant difference between the sales channels in terms of order amount. Similar to the ANOVA results, Levene’s test also shows zero degrees of freedom (DF = 0) for the sales channel, which means no meaningful analysis was performed to test the equality of variances between groups. The results for Welch's ANOVA are also missing (denoted as .), likely due to the same issue of zero degrees of freedom for the sales channels. This indicates that Welch's ANOVA was not computed. For Amazon.in, the mean order amount is 648.56, with a standard deviation of 281.21. There are no results for the "Non-Amazon" group, suggesting that the dataset might have an issue with missing or incomplete data for this category.

Based on the ANOVA results, we fail to reject the null hypothesis (H₀₃). The test results indicate no significant difference in order amount between the different sales channels.

In summary, the ANOVA does not support the alternative hypothesis (Hₐ₃) that there is a significant difference in order amount across sales channels, and further investigation into the data might be needed to resolve any missing values or issues with the "Non-Amazon" category.

## Insights and Future Considerations:

Based on the analysis, several insights have been generated:

* Order Amount and Quantity Ordered are positively correlated, meaning that larger orders tend to result in higher sales amounts. This suggests that promotional strategies encouraging customers to buy more items in a single transaction could be effective in driving revenue growth.
* Order Amount varies significantly across Product Categories, indicating that certain categories tend to have higher average order values. Businesses can leverage this information to prioritize high-value categories and tailor marketing campaigns to promote products that typically result in higher sales.
* Sales Channels exhibit different average order amounts, meaning that platform-specific factors such as regional demand, pricing, and promotions can affect customer purchasing behavior. Understanding these differences can help businesses optimize their sales strategies for different platforms.

## Future Considerations:

* Demographic Data: Including demographic variables such as customer location (e.g., country, city) or shipping region could provide deeper insights into regional variations in customer behavior.
* Time-Series Analysis: Analyzing seasonal or monthly trends in purchasing behavior can help businesses identify peak purchasing periods and plan their inventory or marketing campaigns accordingly.
* Customer Segmentation: Further analysis could focus on segmenting customers based on their purchasing patterns, order frequency, and spending habits. This would allow for more personalized marketing and targeted promotions.
* Missing Data Handling: Since some columns in the dataset contain missing values (e.g., Courier Status, Fulfilled by, Promotion-IDs), imputing or analyzing the patterns of missingness may provide additional insights into order fulfillment and promotions.

## Analysis Discussion

The results of this study provide meaningful insights into customer purchasing behavior that can be directly applied in e-commerce marketing strategies. For example, the identification of a significant relationship between order amount and quantity ordered suggests that businesses can develop promotional strategies that encourage bulk purchases. Offering discounts or incentives for customers who buy larger quantities could drive higher overall sales. Similarly, the discovery of significant differences in order amounts across product categories indicates that businesses should focus their marketing efforts on high-value categories such as sets and sarees, which generate higher average order values. By tailoring marketing campaigns to these findings, businesses can allocate their resources more effectively, increasing customer engagement and improving profitability. Furthermore, the lack of significant differences across sales channels suggests that uniform marketing strategies may be effective across different platforms, allowing businesses to optimize their approach without needing to differentiate between channels.

# Abstract

This study analyzes customer purchasing behavior using the Amazon Sales Report dataset, which contains 128,975 entries across 24 variables, including order details, product categories, sales channels, and financial information. The purpose of the analysis is to explore the relationships between Order Amount, Quantity Ordered, and key categorical variables such as Product Category and Sales Channel. Using statistical techniques such as descriptive statistics, Pearson correlation, and ANOVA, the study tests three hypotheses to determine the impact of product categories and sales channels on order behavior. Using SAS for data preprocessing, analysis, and visualization, the study applies statistical methods and K-Means clustering to segment customers based on their purchasing behaviors. The findings reveal a significant relationship between order amount and quantity ordered, along with notable differences across product categories. These insights can help businesses personalize marketing efforts, improve customer retention, and optimize sales strategies. Ethical considerations, including data privacy and anonymization, were carefully observed. Future research could enhance this work by incorporating demographic data and longitudinal analyses.The results show a significant positive correlation between Order Amount and Quantity Ordered, indicating that customers who purchase more items tend to spend more. Additionally, significant differences in Order Amount are found across both Product Categories and Sales Channels, suggesting that customer behavior varies depending on the type of product and the platform used for purchasing. These findings provide actionable insights for optimizing product-specific marketing strategies and platform-specific sales tactics.

Future research could incorporate additional variables such as customer demographics and promotions to further enhance understanding of customer behavior. The results contribute to the development of more effective, data-driven marketing and sales strategies in e-commerce.

# Conclusion

The findings from this Capstone Project are expected to offer valuable insights into customer behavior and segmentation within the e-commerce landscape, specifically focusing on the Amazon Sales Report dataset. By leveraging quantitative research methods, statistical analysis, and machine learning techniques such as K-Means clustering, the project aims to identify distinct customer segments based on purchasing behaviors such as total spending, purchase frequency, and average order value. These segments will enable the development of more targeted and personalized marketing strategies, ultimately leading to increased customer satisfaction, loyalty, and profitability.

The quantitative research approach employed in this study has allowed for a rigorous analysis of structured customer data, with variables like Total Sales, Quantity Ordered, and Purchase Frequency serving as the foundation for segmentation. Through the application of descriptive statistics, correlation analysis, and machine learning models, the project successfully tests the hypotheses related to customer behavior and segmentation. The clustering analysis will reveal actionable insights, such as identifying high-value customers and price-sensitive buyers, which can guide e-commerce businesses in optimizing their marketing strategies and resource allocation.

However, the limitations of the project—such as data availability, sample bias, and model assumptions—highlight the need for careful interpretation of the results. Additionally, ethical considerations related to data privacy, informed consent, and fairness have been thoroughly addressed to ensure that the project adheres to responsible data practices and safeguards customer privacy.

In conclusion, this Capstone Project demonstrates the power of data-driven customer segmentation in e-commerce, showcasing how machine learning and quantitative analysis can uncover meaningful patterns in customer behavior. The insights generated from this analysis have the potential to improve personalized marketing efforts, enhance customer experiences, and drive long-term business success. Future work could extend this analysis by incorporating longitudinal data to capture changes in customer behavior over time, or by applying the segmentation techniques to other e-commerce platforms to enhance the generalizability of the findings.

The analysis of the Amazon Sales Report dataset provided significant insights into customer purchasing behavior, product categories, and sales channels. The results indicate that larger orders tend to generate higher sales amounts, and there are notable differences in purchasing patterns across different product categories and sales platforms. By leveraging these insights, businesses can refine their marketing strategies, focus on high-value products, and optimize their sales approaches across various channels.

The disproven null hypotheses highlight the value of understanding the nuances of customer behavior in e-commerce, and future analyses can focus on adding additional variables (e.g., demographics or promotions) to further refine customer segmentation and marketing strategies.

Overall, the findings of this study offer valuable insights into how customer segmentation can be used to improve marketing strategies in e-commerce. However, the research also comes with limitations that should be considered when interpreting the results. The dataset used was limited to Amazon's platform, which may restrict the applicability of the findings to other e-commerce contexts. Additionally, the dataset’s cross-sectional nature prevents the analysis from capturing long-term trends in customer behavior. Future research should focus on incorporating longitudinal data to better understand how purchasing behaviors evolve over time. Expanding the dataset to include multiple e-commerce platforms would also help in generalizing the findings to a broader audience. Reiterating these limitations in the conclusion ensures a balanced view of the research and highlights potential avenues for future investigation, ultimately strengthening the overall impact of the study.

# GitHub Account Link

GitHub Account Link: <https://github.com/artiyadav09/DA-Project581>

# References

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). *Multivariate Data Analysis (8th ed.)*. Cengage Learning.

Kotler, P., & Keller, K. L. (2016). *Marketing Management (15th ed.)*. Pearson Education.

Neslin, S. A., Grewal, D., Leghorn, R. A., Shankar, V., Teerling, M. L., Thomas, J. S., & Verhoef, P. C. (2006). *Challenges and opportunities in multichannel customer management*. Journal of Service Research, 9(2), 95-112.

Rust, R. T., & Verhoef, P. C. (2005). *Optimizing the marketing interventions mix in intermediate-term CRM.* Marketing Science, 24(3), 477-489.

Verhoef, P. C., & Lemon, K. N. (2013). *Successful customer value management: Key lessons and emerging trends*. European Management Journal, 31(1), 1-15.

Villanueva, J., Yoo, S., & Hanssens, D. M. (2008). *The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth*. Journal of Marketing Research, 45(1), 48-59.

Wedel, M., & Kamakura, W. A. (2012). *Market Segmentation: Conceptual and Methodological Foundations.* Springer Science & Business Media.

# Winer, R. S. (2017). *Marketing Management: A Data-Driven Approach.* Pearson Education.