# 

# Capstone Project Proposal: Customer Segmentation in E-commerce

# Critical Thinking- Module 8– Week 8

# Arti Govardhan Yadav

# Colorado State University Global

# MIS 581-Capstone - Business Intelligence and Data Analytics

# Dr. Justin Bateh

# Capstone Project Proposal: Customer Segmentation in E-commerce

## Introduction

The capstone project aims to analyze customer purchasing behavior on an e-commerce platform using the Amazon Sales Report dataset. This dataset includes various order-related details such as order amounts, product categories, sales channels, and fulfillment information, comprising over 128,000 entries and 24 variables. Understanding customer behavior is critical for e-commerce businesses like Amazon, as it enables the development of data-driven strategies for targeted marketing, personalized customer experiences, and revenue optimization.

In this project, the primary focus is to segment customers based on their purchasing patterns, identify the key factors influencing sales, and determine how different variables such as product categories and sales channels impact order values. The insights gained from this analysis will help the business design tailored marketing campaigns, improve product recommendations, and enhance overall customer satisfaction.

To achieve these objectives, various data analysis techniques are employed using SAS, a powerful statistical software suite. The analysis includes calculating descriptive statistics, performing correlation and ANOVA tests, and applying clustering algorithms to create customer segments. This approach not only uncovers significant relationships between order amount and quantity ordered but also highlights differences in purchasing behavior across product categories and sales channels.

This document provides the programming code used for data analysis, screenshots of the analysis outcomes, and a link to the GitHub repository where the final project is hosted. By combining these elements, the project offers a comprehensive view of the data-driven strategies that can enhance customer engagement and drive business growth in the e-commerce sector.

## Programming Code and Analysis Outcomes

Below are the steps and programming code for analyzing the dataset using **SAS**. This includes statistical analysis, visualizations, and the relevant code to run in **SAS**.

## 1.1 Correlation Analysis Between Order Amount and Quantity Ordered

The goal is to check the correlation between Amount and Ordered and visualize it with a scatter plot.

## Figure 1

## *The screenshot of the* bar chart - Order Status Distribution of sales datasets*. (Screenshot A. Yadav)*A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated

Explanation :This scatter plot allows you to see if higher quantities ordered correlate with higher order amounts.The purpose of this analysis was to determine if there is a linear relationship between the Order Amount (total amount spent) and the Quantity Ordered (number of items purchased) by customers. The Pearson Correlation method was used to evaluate this relationship. Below is a detailed summary and interpretation of the results:

Before calculating the correlation, the descriptive statistics for both variables were analyzed:The Order Amount variable had an average value of 648.56, with a standard deviation of 281.21, indicating moderate variability in the order values. The maximum value observed was 5,584, suggesting that while most orders are moderate in value, there are some high-value transactions.

The Quantity Ordered variable showed an average of 0.904, with a standard deviation of 0.313. The minimum value was 0, and the maximum was 15, indicating that while most orders consist of less than one item (likely due to fractional or partial measures), some customers place orders with multiple items.

2. Pearson Correlation Coefficient :The correlation analysis provided the following key results:

Correlation Coefficient (r): 0.0669,p-value: <.0001,Number of Observations: 121,180

This correlation coefficient (r = 0.0669) indicates a very weak positive correlation between Order Amount and Quantity Ordered. This means that there is a slight tendency for higher order amounts to be associated with higher quantities, but the relationship is not strong. Despite the weak correlation, the p-value of <.0001 indicates that the relationship is statistically significant, suggesting that it is unlikely to have occurred by chance.

3. Interpretation of Results :The weak correlation implies that while there is a relationship between the quantity ordered and the total amount spent, it is not strong enough to make reliable predictions. Other factors, such as the type of product, price range, or customer-specific behaviors, might play a more significant role in determining the total order amount.

Given the statistical significance of the relationship (p < 0.001), it is important to acknowledge that the dataset is large enough to detect even small correlations. This suggests that, although the effect is weak, it is consistent throughout the dataset.

A scatter plot was used to visualize the relationship between Order Amount and Quantity Ordered. The scatter plot displayed individual orders as points on the graph, with the x-axis representing the Quantity Ordered and the y-axis representing the Order Amount. Due to the weak correlation, the scatter plot showed a dispersed pattern with no clear upward trend, further supporting the finding that the relationship is minimal.

The findings indicate that increasing the quantity of items ordered may not significantly impact the total order amount. To increase order values, the business might need to focus on promoting higher-value products or upselling more expensive items rather than focusing solely on increasing the number of items per transaction.

The correlation analysis demonstrates that while there is a statistically significant, albeit weak, positive relationship between Order Amount and Quantity Ordered, it is insufficient for making substantial predictions. Future analyses may explore other variables, such as product category or customer demographics, to identify stronger predictors of order value and refine marketing strategies accordingly.

## 1.2 ANOVA Test for Order Amount Across Product Categories

The purpose is to determine if there are significant differences in order amounts across different product categories.

## Figure 2

## *The screenshot of the* Code and ouput *(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

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Explanation: The ANOVA test is used to determine if there are statistically significant differences in the average order amounts across different product categories such as Blouse, Bottom, Dupatta, Ethnic Dress, Saree, Set, Top, Western Dress, and Kurta.
* Results Summary:
  + Model Degrees of Freedom (DF): 8 (corresponding to the number of product categories minus one).
  + Sum of Squares: Indicates the variance between the groups (categories).
  + F Value: 9715.17, which is quite large. This high F-value suggests that there are significant differences between the categories.
  + p-value: <.0001. This very low p-value indicates that the differences in means are statistically significant at the 0.05 level. Therefore, we reject the null hypothesis, confirming that at least one product category's mean order amount is significantly different from the others.
* R-Square: 0.390771 indicates that about 39.08% of the variability in order amounts is explained by the product categories.
* Coefficient of Variation (Coeff Var): 33.84440% shows the ratio of the standard deviation to the mean, giving a sense of the data's dispersion.
* Root MSE: 219.5018, which represents the standard deviation of the residuals or errors.
* Amount Mean: 648.5615 is the overall mean of order amounts across all observations.

Tukey's HSD Test for Post-Hoc Analysis : Tukey's HSD test further examines which specific pairs of product categories have significant differences in their mean order amounts.

* Alpha Level: 0.05 (standard level for significance).
* Error Degrees of Freedom: 121171.
* Error Mean Square: 48181.03 (an indicator of within-group variance).

The output includes several comparisons between product categories. Here’s how to interpret the comparisons:

* Comparisons Significant at the 0.05 Level:
  + Significant comparisons are indicated by "\*\*\*". For example:
    - Set vs. Western Dress: The difference in means (70.595) is significant, as indicated by the \*\*\* symbol, showing that the average order amount for 'Set' is significantly higher than for 'Western Dress.'
    - Set vs. Ethnic Dress: A significant difference of 109.490 indicates that the 'Set' category has a higher mean order amount than the 'Ethnic Dress' category.
* Non-Significant Comparisons:
  + Some comparisons, like Saree vs. Set (-33.813), do not have a significant difference, as the confidence limits (-88.590, 20.964) include zero, indicating no statistical significance.

Interpretation of Significant Differences: The ANOVA and Tukey's HSD tests show that there are substantial differences in the order amounts across product categories:

* Categories like Bottom and Dupatta tend to have significantly higher order amounts compared to others such as Top or Blouse.
* Categories like Kurta and Western Dress show lower average order amounts when compared to some other categories like Dupatta and Set.

The output confirms that the Order Amount varies significantly across different product categories. This insight can be used to prioritize marketing efforts or inventory management for high-revenue categories. The results suggest that understanding which categories generate higher sales can help the organization optimize product offerings and develop targeted marketing strategies.

## 1.3 K-Means Clustering for Customer Segmentation

Segment customers into groups based on Order Amount and Quantity Ordered.

## Figure 3

## *The screenshot of the* Code and output *(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

Explanation: The output from the K-Means Clustering analysis for customer segmentation based on Order Amount and Quantity Ordered provides insights into how customers can be grouped based on their purchasing behavior. Below is a detailed explanation of the results:

K-Means Clustering Summary : The clustering procedure was executed using three clusters based on Order Amount and Quantity Ordered. Here’s a breakdown of the key details:

1. Initial Seeds

The clustering algorithm started with three initial seed points:

* Cluster 1: Initialized with an Amount of 5584 and a Quantity of 8.
* Cluster 2: Initialized with an Amount of 2796 and a Quantity of 4.
* Cluster 3: Initialized with an Amount of 0 and a Quantity of 1.

These seeds represent the starting points for each cluster before the algorithm iteratively assigns data points and recalculates cluster centroids.

2. Cluster Summary

The table provides information on each cluster's frequency, standard deviation, and the distance between clusters:

* Cluster 1:
  + Frequency: 6 customers
  + RMS Standard Deviation: 533.2
  + Maximum distance from seed to observation: 869.2
* Cluster 2:
  + Frequency: 12,096 customers
  + RMS Standard Deviation: 114.1
  + Maximum distance from seed to observation: 1522.6
* Cluster 3:
  + Frequency: 116,873 customers
  + RMS Standard Deviation: 145.6
  + Maximum distance from seed to observation: 634.1

Clusters 2 and 3 have significantly higher frequencies compared to Cluster 1, suggesting that most customers fall into these two groups.

Statistics for Variables (Amount and Quantity):-

R-Square Values:

* + The R-Square for Amount is 0.484, indicating that approximately 48.4% of the variability in order amount is explained by the clustering.
  + The R-Square for Quantity is much lower, at 0.036, showing that quantity variability is less well-explained by the clustering model.

This suggests that Order Amount is a stronger determinant for customer segmentation than Quantity Ordered.

* Pseudo F Statistic: The Pseudo F value is 60,579.13, which is quite high and indicates the model has effectively formed well-separated clusters.
* Cubic Clustering Criterion (CCC): The CCC value is -448.336, which may indicate that while the clusters are statistically valid, the clustering structure might not be entirely optimal. The warning indicates that the R-squared values could be invalid for correlated variables.

Cluster Means

The average Order Amount and Quantity Ordered for each cluster:

* Cluster 1:
  + Average Order Amount: 5104.91
  + Average Quantity Ordered: 8.33
* Cluster 2:
  + Average Order Amount: 1235.67
  + Average Quantity Ordered: 1.00
* Cluster 3:
  + Average Order Amount: 584.06
  + Average Quantity Ordered: 0.89

Interpretation:

* Cluster 1 likely represents high-value frequent buyers who tend to place large orders and purchase multiple items.
* Cluster 2 appears to include moderate-value occasional buyers who make smaller orders but may still shop regularly.
* Cluster 3 corresponds to low-value buyers or those who make infrequent and small purchases.

Cluster Standard Deviations

* Cluster 1 has higher standard deviations in both Order Amount and Quantity Ordered compared to Clusters 2 and 3, indicating greater variability in purchasing behavior within this group.
* Clusters 2 and 3 have relatively lower standard deviations, suggesting more consistency in customer behavior in these segments.

Conclusion

The clustering analysis effectively segments customers into three distinct groups:

* Cluster 1: High-value, frequent buyers who spend significantly more and purchase larger quantities.
* Cluster 2: Moderate-value, less frequent buyers with smaller, but consistent order patterns.
* Cluster 3: Low-value buyers with infrequent and small purchases.

These clusters provide valuable insights that can inform targeted marketing strategies:

* Cluster 1 can be targeted with loyalty programs and exclusive offers to maintain their high purchasing levels.
* Cluster 2 could benefit from upselling and cross-selling promotions to increase their average order amount.
* Cluster 3 might respond well to discount-driven campaigns aimed at encouraging larger purchases or more frequent shopping.

This segmentation is crucial for optimizing marketing efforts, improving customer retention, and boosting overall revenue.

# 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.