

Amit Shendge, Disha Kacha, Saniya Shinde, and Rachel Culbreath

Project 1 proposal:

At first glance we think that the dataset could be accurately represented by utilizing bar graphs, scatterplots, and histograms to answer the proposed questions. The dataset includes order details, customer information, product details, shipping information, and regions. In our project a few key areas we will focus on and break down further are sales performance, customer information, regional comparisons, and the supply chain. Our data set came from the source Kaggle and it has 9800 observations. The smart questions we came up with as a team are:

- How does the average time between order date and shipment date vary across different product categories?
- Can we identify any seasonal purchasing patterns among customers within specific product categories?
- What is the distribution of sales across different customer segments (e.g., consumer, corporate, home office)?
- Which product category generates the highest sales revenue, and which one contributes the least?

- How many unique product names are there within each product category?
- Are there any correlations between the chosen ship mode and the time it takes to ship orders, based on order date and shipped date?
- Which shipping mode is most commonly used for specific product categories and subcategories?
- What are the total sales figures for each region, and how do they compare?
- Which states and cities have the highest sales, and can we identify any patterns in sales based on segments within these locations?
- Is there any correlation between postal codes and the categories, subcategories, and sales amounts of products purchased?
- Who is the customer with the highest number of purchases, and what are their characteristics?
- How does the time it takes to ship orders (order date to shipped date) vary across different regions?

The link for the github: <https://github.com/DishaKacha7/>

Intro_Data_Science_Project_1.git