**Project Report**

**On**

**“Exploratory Data Analysis on WALMART”**

Submitted for fulfilment of requirement for the course of

**Data Science**

Submitted by

**Kush Shah.**

Under the guidance of

**Chintan Patel.**

Affiliated with

**Fly The Nest.**

2025

**“Documentation”**

**Signature with date**

Contents

**Problem Statements & Objectives:**

**Problem Statements -**

1. What is the distribution of purchase amounts?
2. Average Purchase Amount per Product Category
3. Product Popularity Trends (Top Selling Products)
4. Impact of Occupation & Marital Status on Purchases.
5. City-Based Sales Analysis.
6. City-Category Preference for Product Categories.
7. Occupation-Based Spending Patterns.
8. Marital Status Impact on Purchase Behaviour.
9. City Category Impact on Total Sales.
10. City Category Preference for Product Categories.

**Objective -**

**Understanding Challenges:** Address operational inefficiencies and customer engagement issues in underperforming areas.

The objective of this analysis is to **understand customer purchasing behaviour** by exploring trends in demographics, product categories, and sales patterns.

By analysing factors such as **age, gender, occupation, city type, and product categories**, we aim to identify key insights that can help optimize **marketing strategies, inventory management, and customer targeting** to improve overall sales and customer satisfaction.

# Introduction

## 

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize the data, using statistical analysis and data visualization methods.

It is a crucial step in any data analysis process, enabling data analysts to uncover patterns, spot anomalies and gain insights to take actions.

## EDA Pipeline

1. Data Acquisition and Objective:
   1. Obtain Walmart data (Kaggle,CSV, Excel)
   2. Get problem statement from Walmart
   3. Choose tools/environment & programming language
2. Data Loading/Reading:
   1. Load data in MySQL (Remove duplicates)
   2. Load data in Jupyter Notebook (To perform further analysis)
3. Familiarize with Data & Identify Target Variable
   1. Explore data (column names, data types)
   2. Identify target variable based on objective
4. Data Preparation & Transformation
   1. Data Cleaning
   2. Handle missing values (imputation, deletion)
   3. Removal of unwanted data (if present)
   4. Format data types (numerical & categorical)
5. Data Analysis & Visualization
   1. Visualization (Charts: pie, boxplot, histogram, heatmap)

## About the company

* Walmart is a retail store having its chain of operating stores across the world.
* They offer variety of product categories to the customers.
* Walmart wants to leverage its transactions data to improve its business.

## Tools and platforms used in project

### Why Python?

Python is a high level and open-source programming language for mathematical computations and basic visualizations. Also, Python and SQL are most sought-after programming languages in the India today.

### Platforms used

* Jupyter notebook – it is a web-based application for running code and queries in Python.

### Versions of platform

* Jupyter notebook - 7.1.3
* Python version - 3.12.2

# Data Loading/reading

## Load Data in Python

We need to import necessary libraries for performing loading, connecting with Python and doing analysis

## Import Necessary Library

* mysql.connector: Library offers connectivity to MySQL server to query from database
* numpy (np): Provides efficient numerical computation tools
* pandas (pd): Offers data manipulation and analysis structures (DataFrames, Series)
* seaborn (sns): Creates informative statistical data visualizations based on Matplotli
* matplotlib.pyplot (plt): Enables various plot creations for data visualization
* %matplotlib inline (Jupyter Notebook specific): Displays plots within the notebook
* warnings (with warnings.filterwarnings("ignore")): Suppresses warnings

# Familiarize the Data & Identifying the Target Variable

# 

## Explore the provided data (column names, data types)

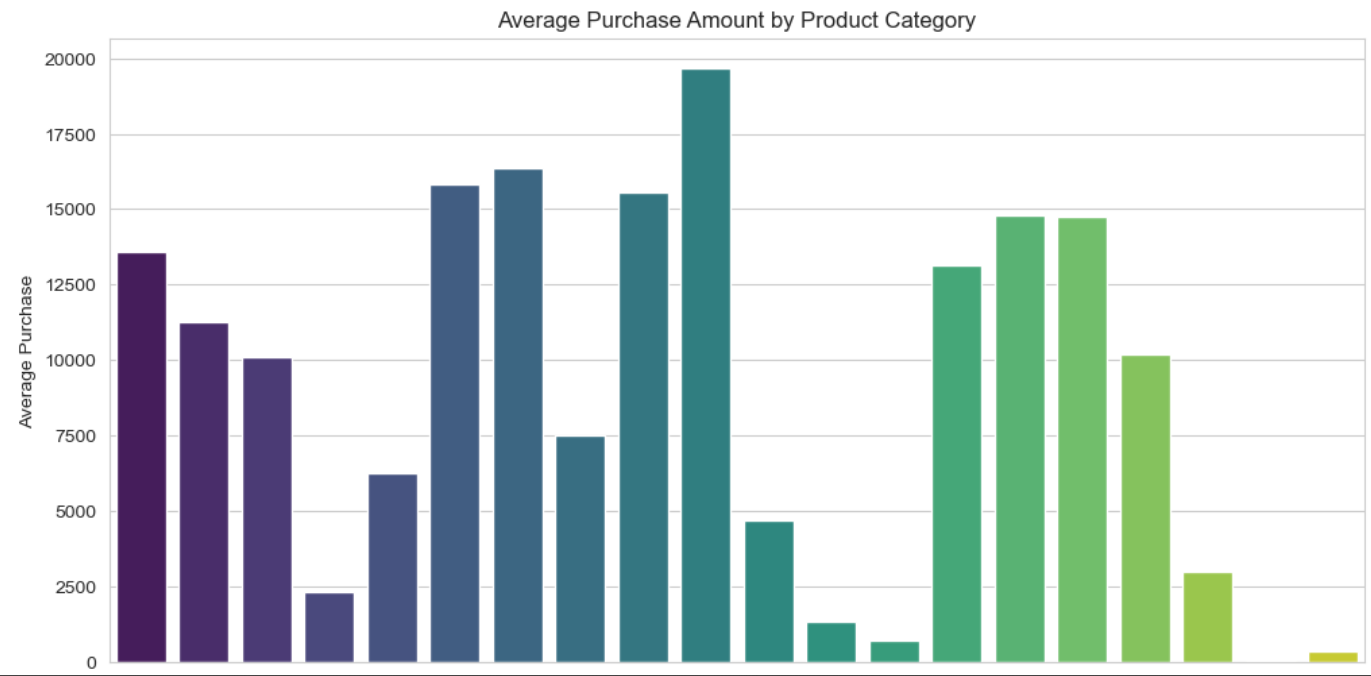
* We need to understand the data before cleaning the data and cross verify if all the required data are provided by Walmart.

### Overview of data

* df.head(); Let's see the data by displaying the first 5 rows
* df.tail(); Let's see the last 5 rows
* df.shape is used to get the dimensions (number of rows and columns) of data
* df.size is used to get the total number of elements in a pandas
* df.info() - used to display concise information about

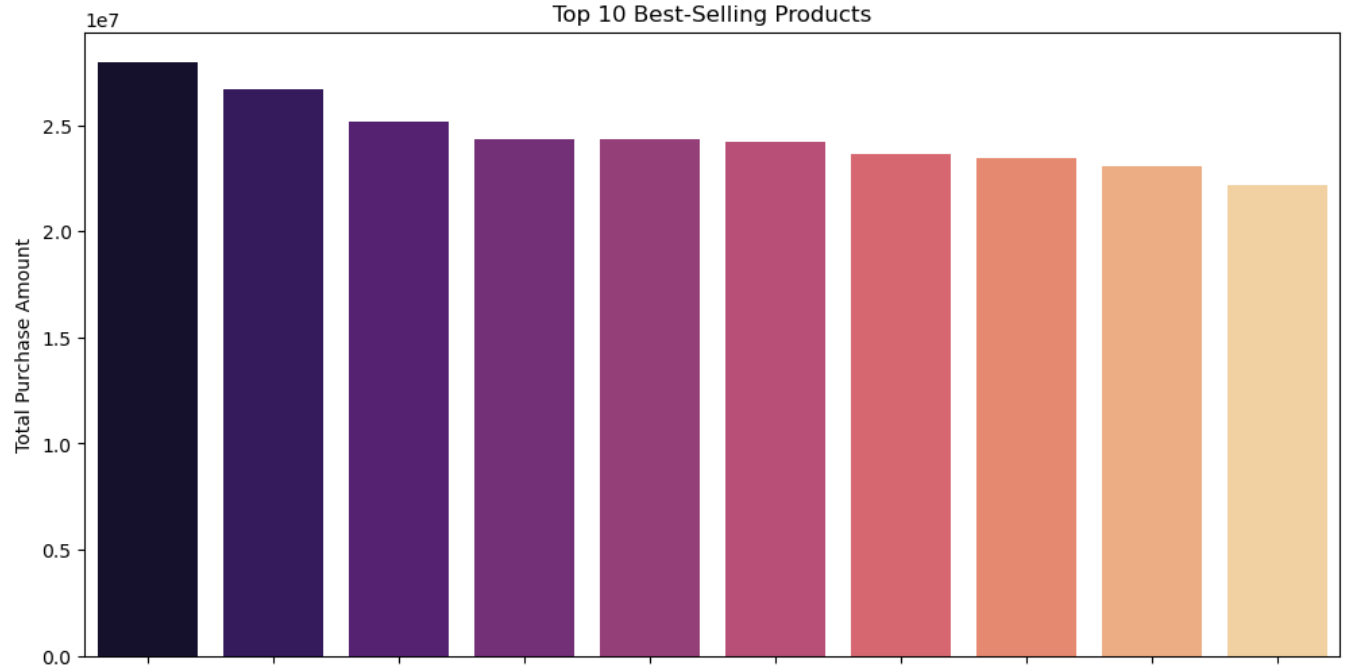
# Graphical Plotting and Interpretations

1. What is the distribution of purchase amounts?



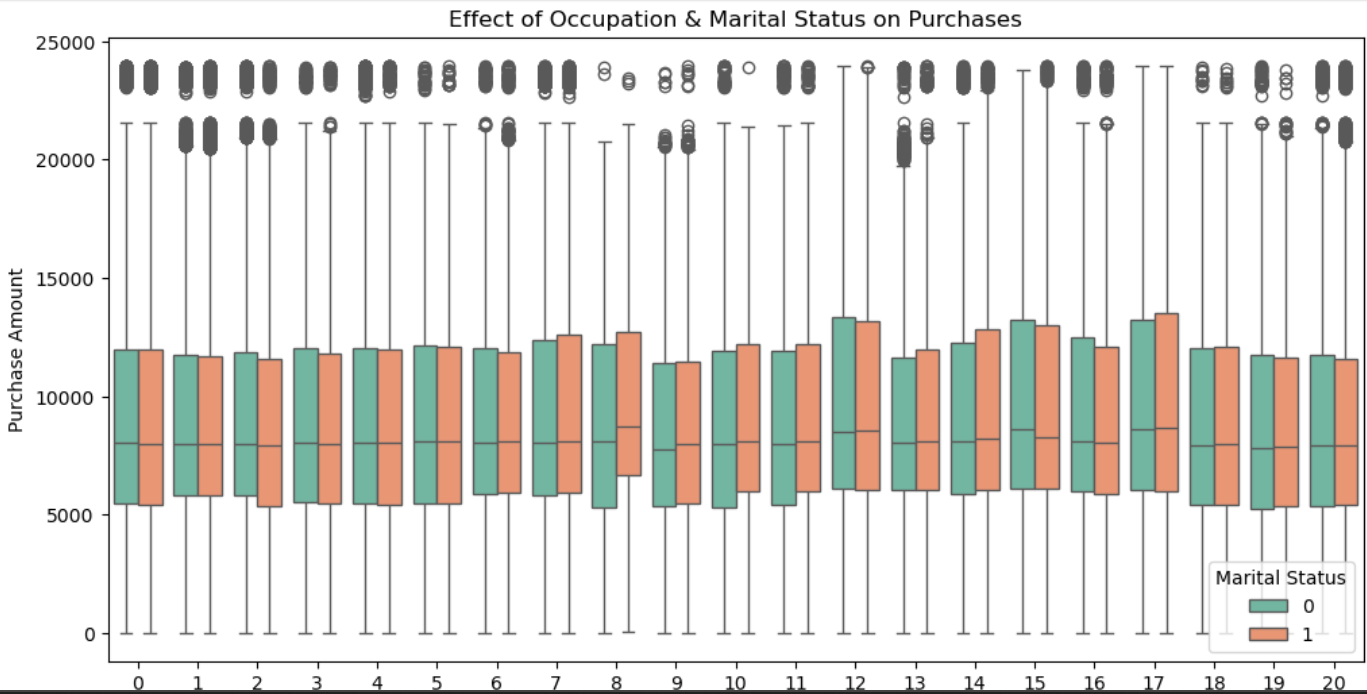
The bar chart shows the **average purchase amount** across different product categories. Categories with **higher average purchases** indicate high-value or premium products, while lower averages suggest more affordable or frequently bought items. This insight can help in **targeting promotions, optimizing pricing, and adjusting inventory** to maximize sales and customer engagement.

1. Product Popularity Trends (Top Selling Products)



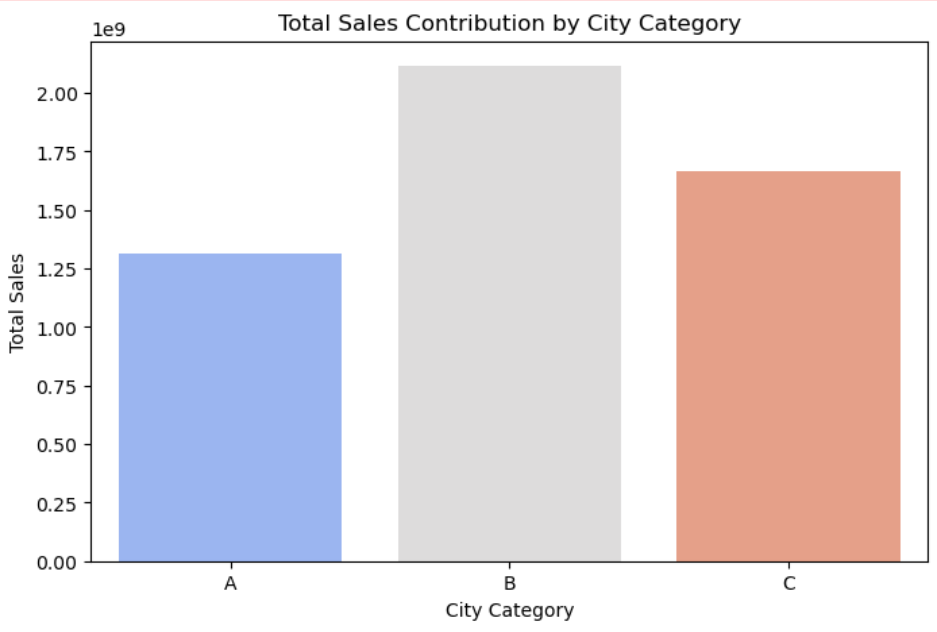
The bar chart highlights the **top 10 best-selling products** based on total purchase amounts. These products drive a significant portion of revenue, indicating **high customer demand**. Ensuring their availability and leveraging targeted promotions can **maximize sales**. Lower-selling products may require **strategic marketing or discounts** to improve performance

1. Impact of Occupation & Marital Status on Purchases



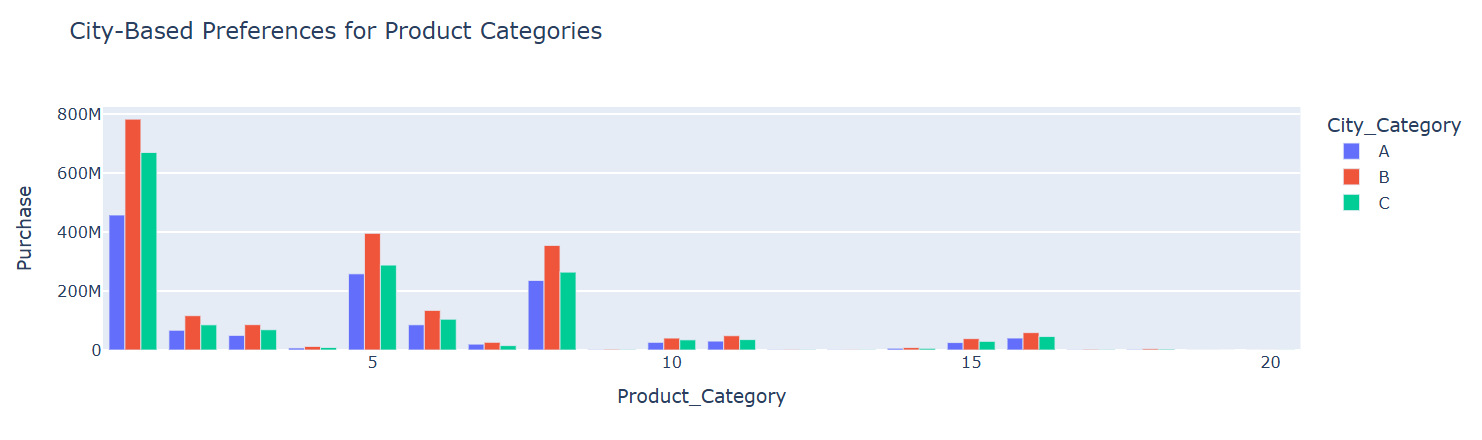
The box plot shows the impact of **occupation and marital status** on purchase amounts. Spending behaviour varies across occupations, with some professions showing **higher median and wider purchase ranges**. Married and single customers may have **different spending patterns**, with one group possibly exhibiting **more consistent or higher purchases**. These insights can help tailor **marketing strategies** based on profession and marital status to drive better engagement and sales.

1. City-Based Sales Analysis



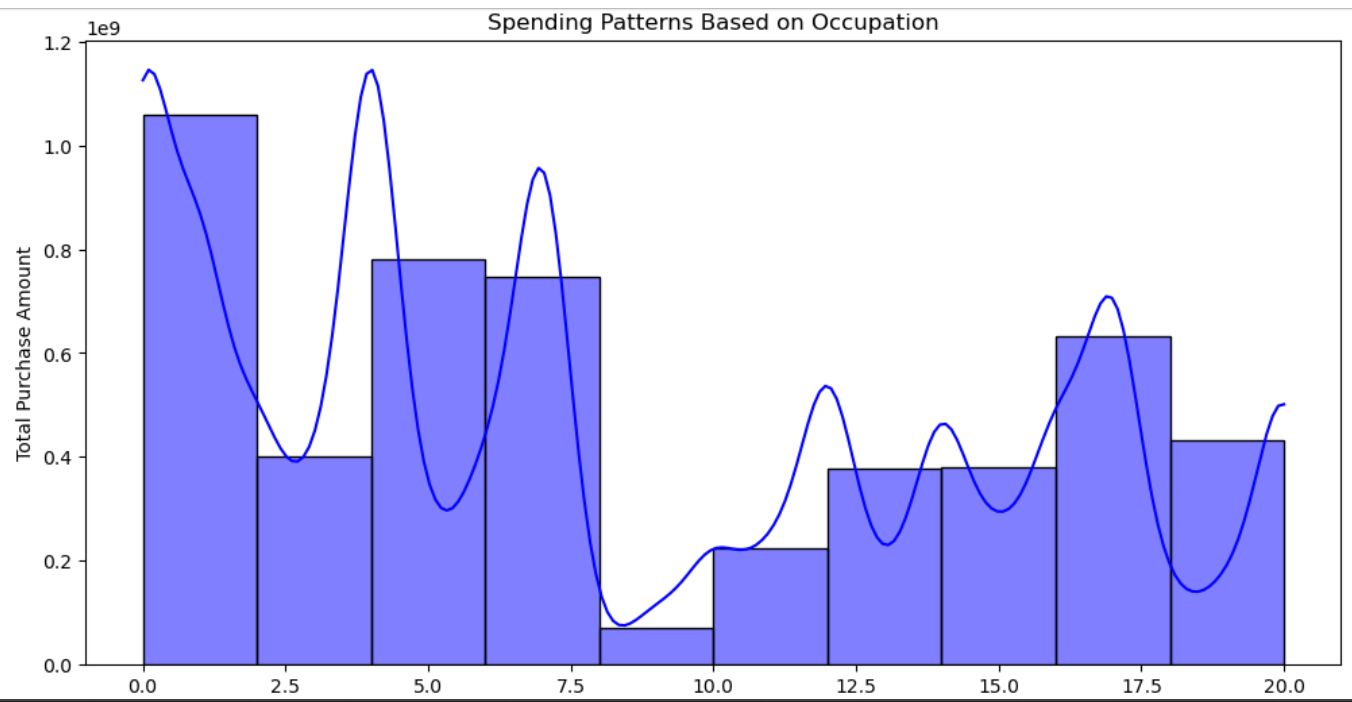
The bar chart illustrates the **total sales contribution** from different city categories. Cities with higher total purchases indicate **greater consumer demand**, possibly due to **higher population density or purchasing power**. Understanding city-wise sales trends helps in **targeted marketing, regional inventory planning, and store expansion strategies** to maximize revenue.

1. City-Category Preference for Product Categories



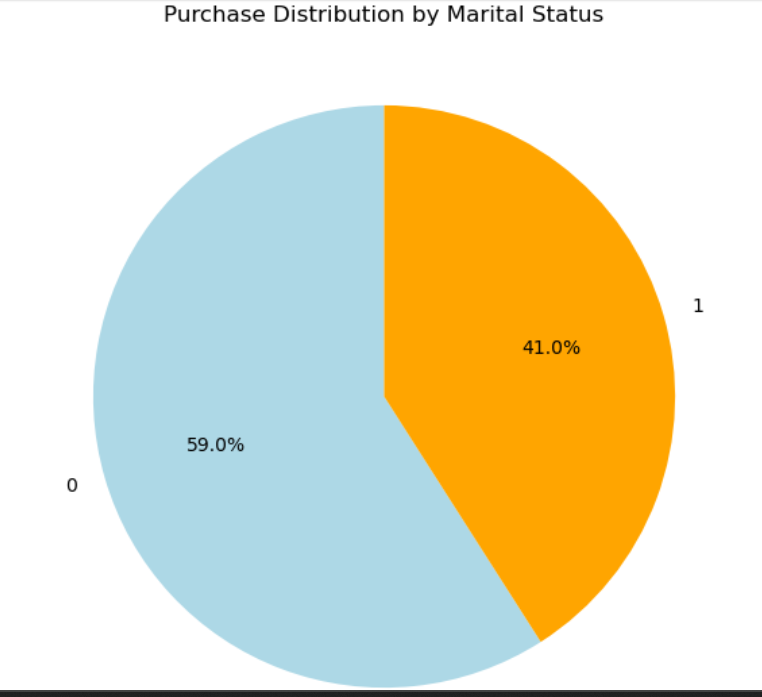
This grouped bar chart shows **city-wise preferences for different product categories** based on total purchase amounts. Some cities may have a **higher demand for specific product categories**, reflecting **regional preferences, income levels, or lifestyle differences**. Retail strategies can be optimized by **stocking popular products in specific cities** and running **localized promotions** to maximize sales.

1. Occupation-Based Spending Patterns



The histogram illustrates **spending patterns across different occupations**, showing which professions contribute the most to total purchases. Peaks in the distribution suggest **higher spending professions**, possibly due to **higher disposable income or purchasing power**. Understanding these trends helps in **tailoring promotions and product recommendations** based on occupational groups.

1. Marital Status Impact on Purchase Behavior



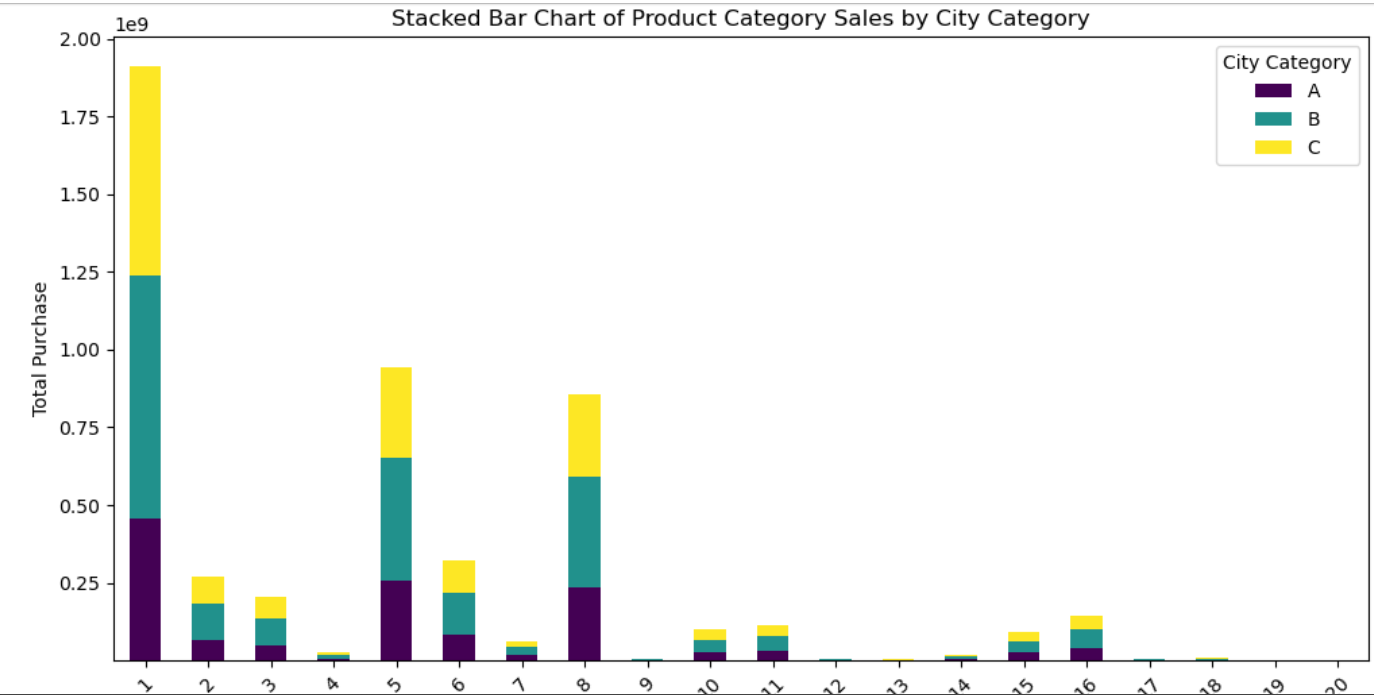
The pie chart shows the **purchase distribution between married and single customers**. A higher percentage for one group indicates **stronger spending behaviour**, which could be influenced by **financial responsibilities, lifestyle choices, or shopping preferences**. These insights help in **targeted marketing campaigns** to better engage each segment and drive sales.

1. City Category Impact on Total Sales



The scatter plot visualizes the **distribution of purchase amounts across different city categories**. A denser concentration of points in a specific city category suggests **higher purchasing activity** in that region. Variations in spending could indicate **differences in economic status, consumer behaviour, or product availability**, helping businesses refine **regional sales strategies and inventory management**.

1. City Category Preference for Product Categories



The stacked bar chart shows **how different city categories contribute to total sales across product categories**. If a particular city dominates a category, it indicates a **strong regional preference for those products**. Balanced distributions suggest **consistent demand across all city types**. These insights help in **regional marketing, inventory allocation, and sales strategy optimization** to boost revenue.