**CONSUMER E-COMMERCE ANALYSIS**

Prepared in the partial fulfilment of the Summer Internship Program on Data Analysis

AT



*Under the guidance of*

**Mrs. Narmada Mani, APSSDC**

**Mrs. K Meenakshi, APSSDC**

*Submitted by*

**Teja Sai Siddharth Gandu, AP21110011421**

**Srinadh Gonugunta, AP21110011437**

**Lakshmipathi Pathagunta, AP21110011457**

**SRM University, AP**

(June-Aug, 2023)

**CONTENTS**

**Acknowledgement** ………………………………………………………….. 3

**Abstract** …………………………………………………………... 4

**Introduction** …………………………………………………………... 5

**Overview** …………………………………………………………... 6

* Data Collection …………………………………………………..6
* Data Pre-processing …………………………………………....6
* Queries Resolution ………………………………………………7
* Data Visualization ………………………………………………..8
* Behaviour Analysis ……………………………………………..11

**Requirements** …………………………………………………………... 12

**Conclusion** ………………………………………………………….... 13

**References** …………………………………………………………... 14

**ACKNOWLEDGEMENT**

We would like to extend our sincere appreciation to Meenakshi Ma’am and Narmada Ma’am, for their invaluable guidance and consistent support during our summer internship program, which was made possible through the efforts of APSSDC. Your expertise not only improved our technical skills but also broadened our industry perspective.

We also want to express our gratitude to APSSDC that actively promotes skill development and empowers students like us to gain real-world experience. The resources, opportunities, and exposure provided by APSSDC were invaluable in shaping my understanding of the industry and preparing me for future challenges.

Once again, thank you, Narmada Ma’am, Meenakshi Ma’am, and APSSDC, for your significant contribution to our professional growth. We look forward to carrying forward the knowledge and insights gained during this internship as we embark on the next phase of my academic and professional journey. Your guidance will always be a beacon of inspiration for us.

Thank you.

Sincerely,

GTS Siddharth, AP21110011421

G Srinadh, AP21110011437

P Lakshmipathi, AP21110011457

**ABSTRACT**

Consumer E-commerce analysis is a crucial endeavour in understanding the landscape of modern business. This abstract highlights the key aspects and objectives of such analysis. It involves examining the behaviour, preferences, and decision-making processes of consumers in the context of commerce.

In this project we have mainly focused data cleaning and data visualization. Having clean data will ultimately increase overall productivity and allow for the highest quality information in decision-making. Data cleansing is further needed to remove the duplicate, and any inaccurate data from databases.

Data visualization provides a quick and effective way to communicate information in a universal manner using visual information. Starting from a clean dataset allows you to focus on creating an effective visualization rather than trying to diagnose and fix issues while creating visualizations.

**INTRODUCTION**

Each customer's information, including purchases made, number of purchases, category of products, frequency of purchases, order hike from the previous year, and other data, is kept on file by the e-commerce company. Analysing the purchase behaviour of customers is one of the most important metrics for a brand. Data analysis provides insights into how a customer is satisfactory with the service and areas to be improved to increase number of customers.

Customer analysis is also helpful to compare sales of previous years and present years to know a particular sale of product is increased or decreased. Knowing customer behaviour can greatly enhance decision-making processes and can further help reduce churn to improve profitability. In this project we have taken data of customers of e-commerce Company and analyse the data.

Churn Analysis plays a crucial role in revealing the patterns that indicate the common motivators for customers to part ways with your company. These could be anything from poor satisfaction score to greater distance from the company’s warehouse to consumer’s home. It is also instrumental in demonstrating the exact way customers engage with your product throughout their lifecycle. You can use this to maximize what your customers already love, and improve upon everything they don’t.

**OVERVIEW**

1. **DATA COLLECTION –**

* This step involves collecting dataset from a reliable website (like Kaggle ), as it directly affects the quality of your analysis.
* The dataset used is **E-Commerce.csv**. It has multiple columns such as CustomerID, Tenure, CityTier, PreferredPaymentMode, Gender, Satisfaction Score, Category and many more.
* Import this csv file to Jupyter Notebook using **read\_csv()** method in pandas.

1. **DATA PREPROCESSING** **–**

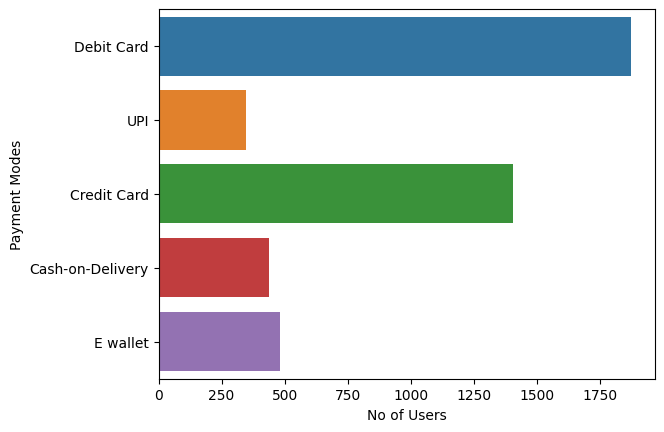
* This step involves several tasks aimed at cleaning, transforming, and organizing the data to improve its quality, relevance, and suitability for the specific analysis.
* Renaming few columns using **rename()** method such as Tenure to Years, PreferredPaymentMode to PaymentMode , etc, to increase readability.
* Replacing and renaming the duplicate values in the dataset using **replace()** method, such as COD and Cash on Delivery to Cash-on-Delivery, etc. Deleting columns that are relatively insignificant using **drop()** method such as HourSpentOnApp column , etc.
* **Isnull()** method is used to display all the null values in a column of the dataset. Null values are needed to handle as null values can lead to inaccurate results, predictions, negatively impact the analysis, and can cause inconsistence in the analysis of the dataset.
* SimpleImputer is a scikit- learn class which is helpful in handling the missing data in the dataset. It replaces the NaN/Null values, with a specified placeholder. **fit\_transform()** method is used to calculate the statistical value(mean, median ,mode ,or any constant) that will be used to impute(fill in ) in the missing values( mean in our case) and replace the missing values with this value in the given column.
* Later using **dropna()** method , drop the rows that contains null value and keeping inplace as True.

1. **QUERIES RESOLUTION-**

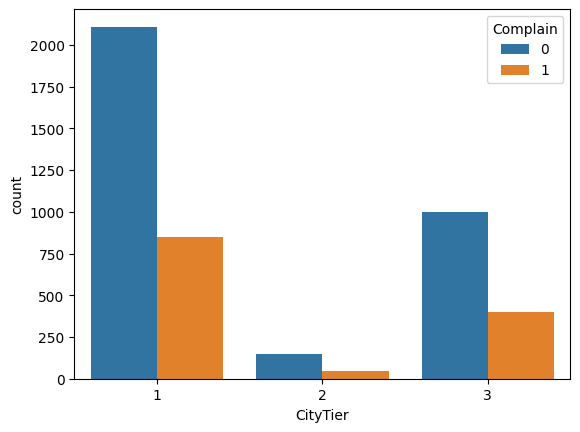
* Query resolution is critical for data analysis because it ensures that the inquiries or questions on the data are addressed accurately and efficiently. This allows focusing on deriving valuable insights from the data.
* **groupby()** method is used to group data based on one or more categorical variables, allowing one to apply aggregate functions to each group independently. For example, groupby() can be used for Category to find out all details based on each category.
* **sort\_values()** method is used to sort the rows in a dataset, which allows one to better understand data’s distribution, find patterns, and present data in more organized manner. For example, sort\_values() can be used to find out payment mode that provides more cash back in descending /ascending order.
* **value\_counts()** method helps to count the occurrences of each unique value and provide summary of the data’s frequencies. For example, value\_counts() can be used to find the number of male or female and can also be used to find relationship status of the customer, such as married, single or divorced.
* **Few Queries that are resolved are-**
  + Which category saw the highest growth in orders from last year?
  + Does marital status affect the satisfaction score for the product?
  + Which mode of payment gives highest cashback amount?
  + City-Tier wise distribution of the number of orders received.

1. **DATA VISUALIZATION –**

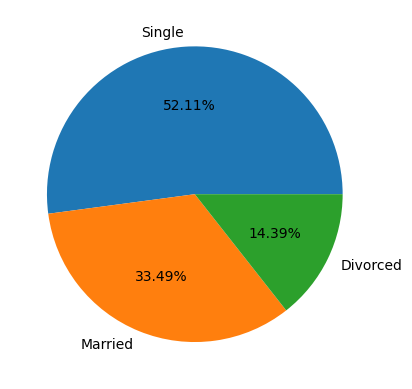
* It involves creating graphical representations of data to visually present patterns, trends, relationships, and insights, that might be difficult to understand from the raw data alone.
* Matplotlib and Seaborn are extensive used to plot graphs.
* **Bar graphs** are represents frequency or count of a particular category as rectangular bars. For example- plot showing number of users on the basis of the mode of payment.



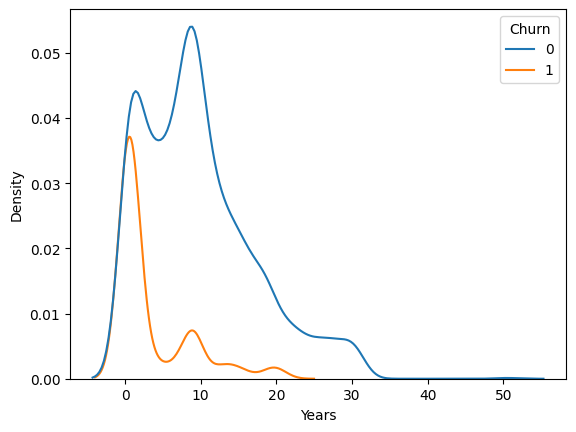
* **Count plot** is type of bar plot that displays the count of observations in each category of a categorical variable. For example- Plot showing number of complaints received based on City Tier.



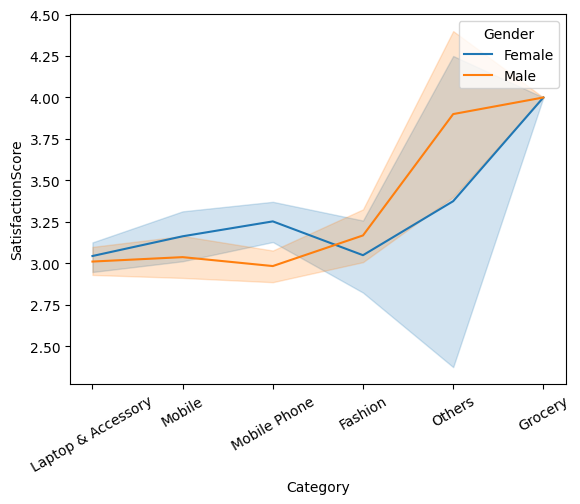
* **Pie chart** is a circular statistical graph used to display the proportions of different categories or components within a whole. Highly useful when one want to convey relative proportions. For example – Marital status of the consumers



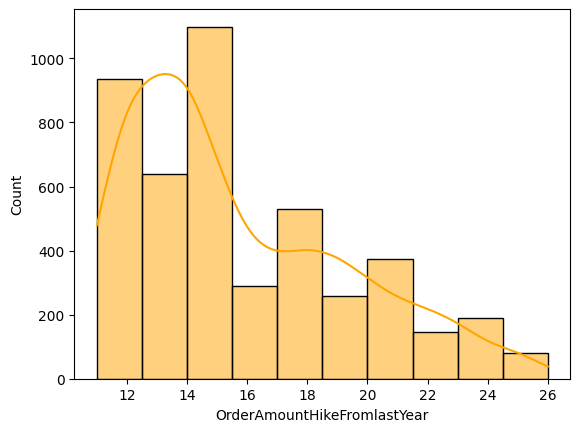
* **Kernel Density Estimation** (KDE) plot is used for estimating the probability density function (PDF) of a continuous random variable. This allows one to examine how likely different values are within the observed data. For example- KDE Plot that shows relationship between number of years and the churn.



* **Line Plot-** Used to display the relationship between two continuous variables over a continuous interval. This allows one to visualize how a variable changes over time, identifying trends, and patterns. For example – Plot showing relationship between satisfaction score and category on the basis of gender.



* **Histogram** shows how data is spread across different ranges or interval (bins), helping to understand central tendencies, spread and shape of the data. For example- Plot shows the amount of order hike from the last year. (In the following example, kde is set as True, hence, kde plot is also present alongside).



1. **BEHAVIOUR ANALYSIS-**

* Behaviour analysis provides critical insights that can drive decision-making, improve user experience, optimize marketing strategies, and enhance overall business performance.
* **Important remarks that can be made by behaviour analysis-**
  + Laptop and Accessory has seen the most growth, while grocery has seen the least growth since the last year.
  + Male are more likely to churn as compared to female
  + Debit card provides highest cash back amount to the users.
  + There are more mobile users than desktop users.
  + People from Tier-2 Cities have the least number of orders.
  + If more number of coupons is used, then more number of orders is purchased.
  + Most of the consumers stay within 20 km from the warehouse.
  + If people are associated with the E-Commerce Company for more than 25 years, then they are highly unlikely to churn.

**REQUIREMENTS**

**Operating System Used** –

* Windows OS

**Software requirements** -

* **Data Analytic Tools**: Pandas and Scikit-learn - These popular libraries provide essential tools for data processing and analysis.
* **Data Visualisation Tools**: Matplotlib and Seaborn - Libraries in Python for creating data visualizations.
* **IDE** – Jupyter Notebook

**Hardware requirements**-

* Multicore processor with atleast 2.5 GHz clock speed (recommended 4 core or more)
* Minimum 8 GB RAM for handling large dataset and memory-intensive operations.
* Adequate storage space for storing dataset and analysis results.

**CONCLUSION**

## In conclusion, by understanding the factors driving consumer decisions, businesses can tailor their offerings, optimize marketing strategies, and engage stronger customer relationships. This data can further enable the companies to gain deep insights into consumer behaviour, preferences and trends. Data analysis is important in research because it makes studying data a lot simpler and more accurate, and in return, enhances decision-making capabilities of the companies. Analysing these technological shifts is essential for businesses seeking to adapt and remain competitive in the digital age.

## REFERENCES

## Dataset - <https://www.kaggle.com/datasets>

## Documentations –

## <https://matplotlib.org/>

## <https://scikit-learn.org/stable/modules/>

## <https://pandas.pydata.org/docs/user_guide/>

## 3. AP Skill Development Corporation (2023) – SRM Data Analysis Summer Internship – <https://github.com/AP-Skill-Development-Corporation/SRM-Data-analysis-summer-internship-2023/>

## 