**CSCI 6223 – Practical Data Science**

**Fall 2023**

1. **Project Title:** Customer Shopping Trends
2. **Report:**

**Final draft report on the customer shopping trends.**

1. **Project Objectives:**

In this project we used Data Analysis and Machine Learning concepts in different areas of the project in order to get the desired results. They are listed below:

1. **Classification:**

Classification is used when you want to categorize data points into predefined classes or categories. In this project we used classification to predict whether a customer will make a purchase or not (binary classification) or categorize customers into different segments based on their shopping behaviour (multi-class classification).

Example: Predicting whether a customer will buy a product (yes/no) based on features like age, gender, browsing history, etc.

1. **Regression:**

Regression is used when you want to predict a continuous numerical value. In this project we can use regression to predict the amount of money a customer is likely to spend on a shopping trip or forecast future sales figures.

Example: Predicting the total purchase amount of a customer based on various features such as time spent shopping, number of items in the cart, etc.

1. **Clustering:**

Clustering is used when you want to group similar data points together based on their inherent patterns or similarities without predefined categories. In this project we used clustering to identify different customer segments based on their shopping behaviour.

Example: Grouping customers into clusters based on their shopping habits and preferences without specifying the number of clusters in advance.

In this project, we mainly focused on clustering to identify customer segments.

**4.0 Project Description:**

The Customer Shopping Preferences Dataset offers valuable insights into consumer behaviour and purchasing patterns. Understanding customer preferences and trends is critical for businesses to tailor their products, marketing strategies, and overall customer experience. This dataset captures a wide range of customer attributes including age, gender, purchase history, preferred payment methods, frequency of purchases, and more. Analysing this data can help businesses make informed decisions, optimize product offerings, and enhance customer satisfaction. The dataset stands as a valuable resource for businesses aiming to align their strategies with customer needs and preferences.

**4.1 Datasets:**

Find the dataset which is used in this project through this link:

[**https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset/data**](https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset/data)

This dataset is a synthetic creation to simulate a realistic customer shopping experience. Its purpose is to provide a platform for beginners to learn from a dataset that mirrors real-world customer shopping behaviour. The aim is to foster learning and experimentation in a simulated environment, encouraging a deeper understanding of data analysis and interpretation in the context of consumer preferences and retail scenarios.

****

**4.2 Data mining Tasks:**

Clustering is used when you want to group similar data points together based on their inherent patterns or similarities without predefined categories. In this project we used clustering to identify different customer segments based on their shopping behaviour.

Classification is used when you want to categorize data points into predefined classes or categories. In this project we use classification to predict whether a customer will make a purchase or not.

Regression is used when you want to predict a continuous numerical value. In this project we can use regression to predict the amount of money a customer is likely to spend on a shopping trip or forecast future sales figures.

**5.0 Data Preprocessing:**

Missing Data - There was no missing data in the dataset since the customer shopping trends is continuous process and on go.

Outliers – The data which is less considered and has no common points is taken out as outliers to show the data as processed and clean.

**6.0 Exploratory Data Analysis:**

The purpose of this analysis is to examine the customer purchase behaviour and also the accuracy of the models used in the customer-shopping-trends dataset and then the result will predict the age of the buyers who are purchasing frequently. Analysing this data can help businesses make informed decisions, optimize product offerings, and enhance customer satisfaction.

**Employment\_Type:**

This will compare both male and female then give the count of total number of males and females.

A blue and pink rectangles

Description automatically generated

**Age:**

Here, the age distribution is plotted below along with the density or count of the people based on age by a curve.

A graph of a graph

Description automatically generated with medium confidence

**Categories:**

The data is categorised into different types like Clothing, Footwear etc., and then obtaining the count.

A graph of different colored squares

Description automatically generated

**Item\_Purchased:**

Segregating the items and plotting the number of times it is purchased by people.

A graph of colored lines

Description automatically generated

**Location:** Counting the number of purchases by people from different locations.

A graph of different colored rectangular shapes

Description automatically generated with medium confidence

**Size:**

Segregating all the items and putting them in their respective sizes which are small, medium, large, extra large and finding the count of which size has the highest occurrences of purchase.

A graph of different colored squares

Description automatically generated

**Colour:**

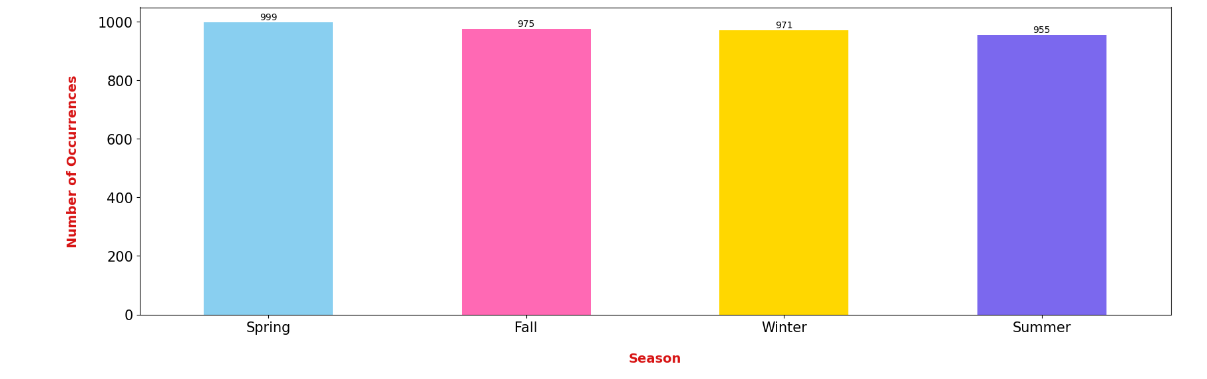
Differentiating each colour and separating the items according to that and based on the colour plotting the highest purchased colour item.

A graph with different colored lines

Description automatically generated with medium confidence

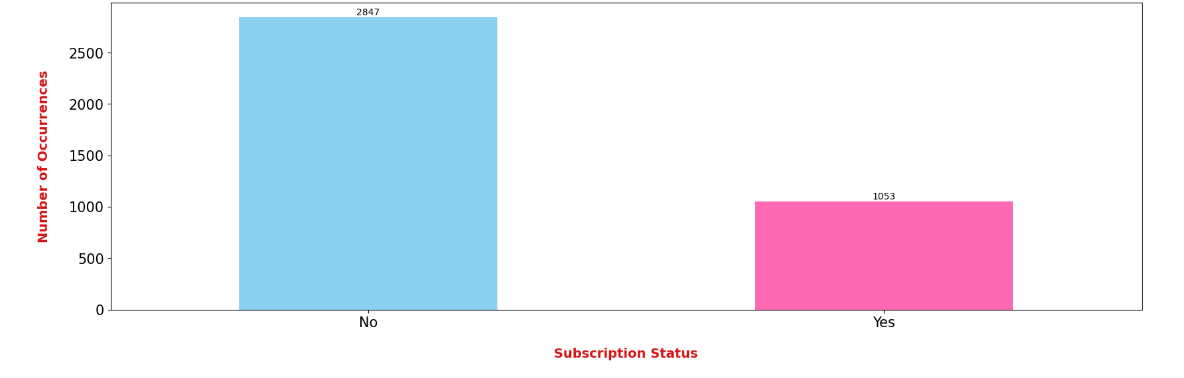
**Season:**

Plotted the season wise data in-order to know the number of purchases in each season.



**Subscription\_Status:**

Checking the count of customers who have subscription and who doesn’t to keep track and then noticed that the customers who does not have subscription are in higher rate of purchasing.



**Payment\_Method:**

Plotted the type of payment from the users and counted the number of people for each payment type.

A graph of different colored rectangles

Description automatically generated

**Shipping\_Type:**

Plotted the mode of receiving the item purchased and counted the no. of people preferred each shipping type.

A chart of different colored rectangular shapes

Description automatically generated

**Frequency\_Of\_Purchases:**

Analysing the time period between the purchases by the customers and counted all the possibilities.

A graph of different colored rectangular shapes

Description automatically generated with medium confidence

**Final prediction plot of the age and their purchasing frequency:**

A graph of blue bars

Description automatically generated

**7.0 Implementation Details:**

In this project we used the data analysis and machine learning concepts mainly clustering to get the similar data points as a group based on their similarities.

* In this we mainly used Python and its libraries like pandas, numpy, sklearn, seaborn and matplotlib.
* matplotlib is used in the code in order to plot the output as clusters.
* This project will mainly focus on two parts- Observing the customers shopping behaviours through code and finally by using clustering in the code to get the result as a plot with different clusters.
* The code was written by using the above-mentioned data mining algorithms, data mining techniques and also with the python platform.

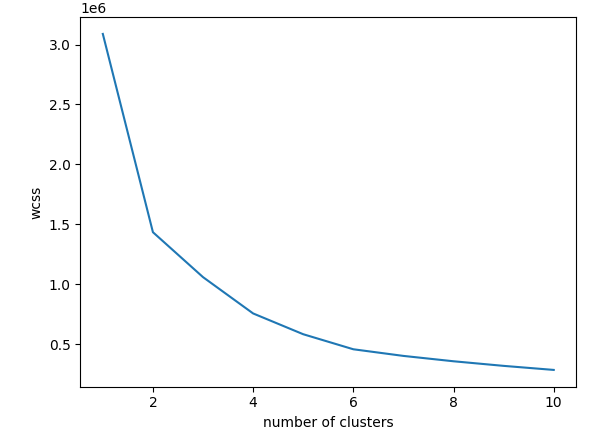
**7.1 Model Selection:**

We are using clustering model to predict the number of possible clusters that can be used for this prediction.

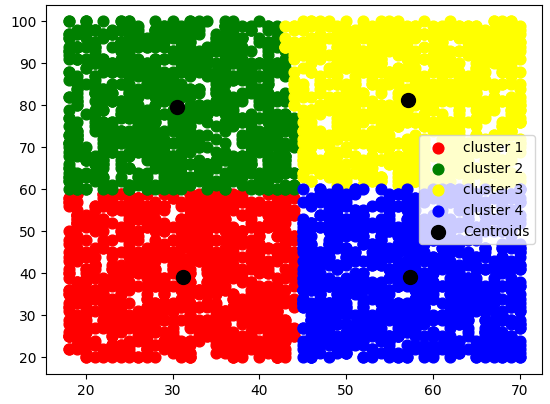
Here we are using,

**K-means clustering:** This is an [unsupervised learning](https://deepai.org/machine-learning-glossary-and-terms/unsupervised-learning) technique to [classify](https://deepai.org/machine-learning-glossary-and-terms/classifier) labelled or unlabelled data by grouping them by features, rather than pre-defined categories. The variable K represents the number of groups or categories created. The goal is to split the data into K different clusters and report the location of the centre of mass for each cluster. Each centroid of a cluster is a collection of feature values which define the resulting groups.

Dataset is loaded and WCSS (Within-Cluster Sum of Square) is used to decide the number of clusters. WCSS is the sum of the squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow.



To plot the final number of clusters we used scatter plot with optimal no. of centroids.



**8.0 Conclusion:**

By comparing different models, we concluded that K-means clustering model is the best as this is giving the prediction from the analysis and customer satisfaction. Also, we used classification and regression which helped to separate and classify the data according to the need for better understanding of customer purchase trends and this will benefit in improving the sales.

**9.0** **References:**

1. Journals, Research papers: Smith, J. A., & Johnson, B. C. (2020). "The Impact of Social Media Marketing on Consumer Purchase Behaviour." Journal of Consumer Research, 45(2), 256-274.
2. Market Research Reports: Nielsen. (2022). "Global Consumer Shopping Trends Report 2022." [Include the specific report title and publication date].
3. Books: Underhill, P. (2019). "Why We Buy: The Science of Shopping." Simon & Schuster.
4. Government Publications: U.S. Census Bureau. (2022). "Monthly Retail Trade Report." [Include the specific report title and publication date].
5. E-commerce and Retail Reports: Amazon. (2021). "Amazon Customer Behaviour Insights 2021." [Include the specific report title and publication date].
6. Consumer Surveys and Reports: National Retail Federation (NRF). (2022). "Consumer View: The Impact of Retail's Evolution on Shopping Behaviour." [Include the specific report title and publication date].
7. Trade Journals and Magazines: Retail Dive. (2023). "Trends Shaping the Future of In-Store Shopping." [Include the article title and publication date].
8. Blogs and Online Articles: Shopify Blog. (2023). "E-commerce Trends to Watch in 2023." [Include the article title and publication date].

-----------------------------------------------------END------------------------------------------------------------------