MDA 620: Capstone Project Report

Analyzing the consumer Shopping Behavior for the online Business start-up

Mentor: Prof. Gurpreet Singh
Project by: Naisarg Bhavsar

1. **Introduction:** The Dataset "Consumer Behavior and Shopping Habits" is taken from Kaggle.

Objectives: Explore shopping behavior, analyze demographics, and build predictive models for the purpose of online business startup.

2. Data Exploration and Visualization:

- Loaded the dataset and checked its structure.

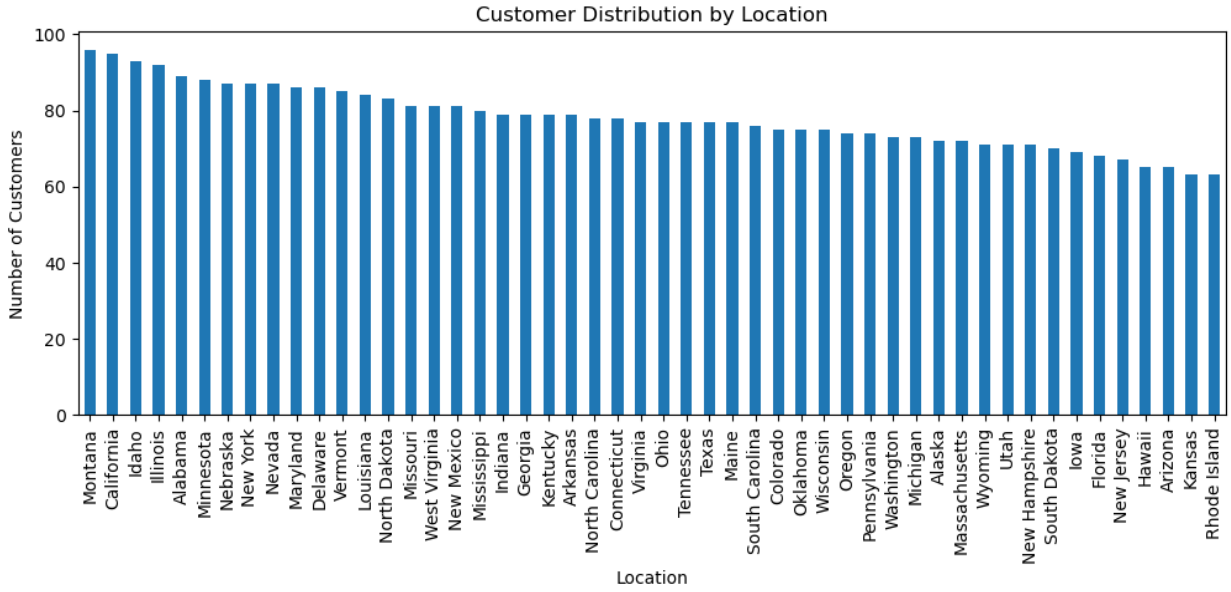
	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes
5	6	46	Male	Sneakers	Footwear	20	Wyoming	M	White	Summer	2.9	Yes	Standard	Yes	Yes
6	7	63	Male	Shirt	Clothing	85	Montana	M	Gray	Fall	3.2	Yes	Free Shipping	Yes	Yes
7	8	27	Male	Shorts	Clothing	34	Louisiana	L	Charcoal	Winter	3.2	Yes	Free Shipping	Yes	Yes
8	9	26	Male	Coat	Outerwear	97	West Virginia	L	Silver	Summer	2.6	Yes	Express	Yes	Yes
9	10	57	Male	Handbag	Accessories	31	Missouri	M	Pink	Spring	4.8	Yes	2-Day Shipping	Yes	Yes

- No duplicates found as the Dataset originally cleaned.
- Examined unique values for each column.

```
Customer ID                [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...
Age                        [55, 19, 50, 21, 45, 46, 63, 27, 26, 57, 53, 3...
Gender                                                    [Male, Female]
Item Purchased             [Blouse, Sweater, Jeans, Sandals, Sneakers, Sh...
Category                        [Clothing, Footwear, Outerwear, Accessories]
Purchase Amount (USD)      [53, 64, 73, 90, 49, 20, 85, 34, 97, 31, 68, 7...
Location                   [Kentucky, Maine, Massachusetts, Rhode Island,...
Size                                                        [L, S, M, XL]
Color                      [Gray, Maroon, Turquoise, White, Charcoal, Sil...
Season                                      [Winter, Spring, Summer, Fall]
Review Rating              [3.1, 3.5, 2.7, 2.9, 3.2, 2.6, 4.8, 4.1, 4.9, ...
Subscription Status                                            [Yes, No]
Shipping Type              [Express, Free Shipping, Next Day Air, Standar...
Discount Applied                                               [Yes, No]
Promo Code Used                                                [Yes, No]
Previous Purchases         [14, 2, 23, 49, 31, 19, 8, 4, 26, 10, 37, 34, ...
Payment Method             [Venmo, Cash, Credit Card, PayPal, Bank Transf...
Frequency of Purchases     [Fortnightly, Weekly, Annually, Quarterly, Bi-...
dtype: object
```

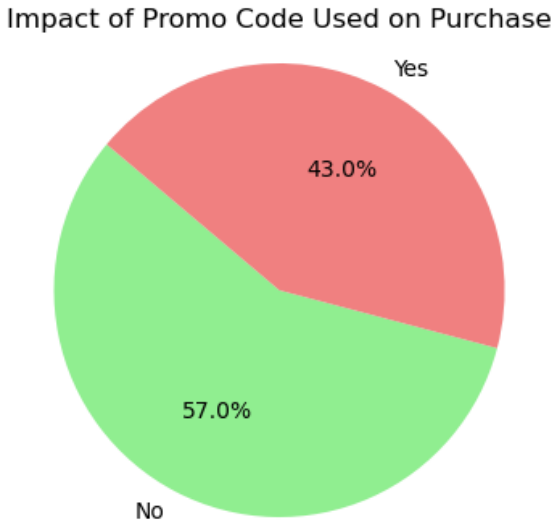
MDA 620: Capstone Project Report

- Visualized customer distribution by location.

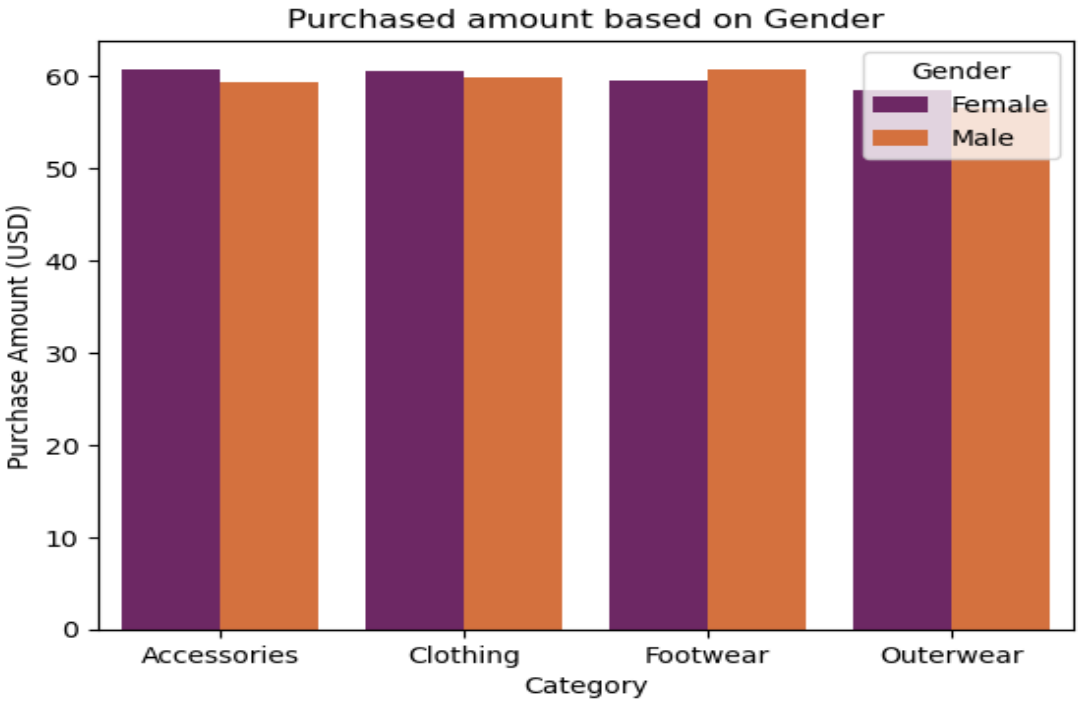


Here Montana is leading all other states with the approximate consumer count of more than 90 and Rhode Island has least count of 85.

- Explored the impact of Promo Code usage with an impact of 43% of all the customer.



- Analyzed purchase amount based on gender and category.
  This shows almost similar values for all categories with one observation that Females are ahead then Males for three categories except only in Footwear.

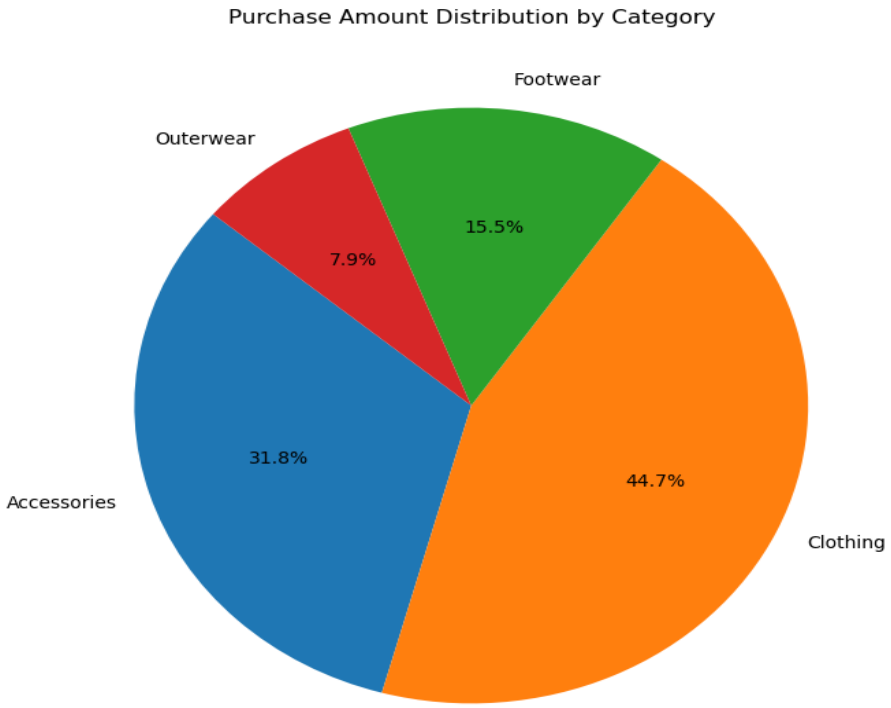


3. Data Manipulation and Visualization:

- Converted 'Purchase Amount' to numeric and explored relationships between 'Purchase Amount,' 'Season,' and 'Category' using Scatter plot.
  where clothing is on top category and outerwear at the last spot.

MDA 620: Capstone Project Report

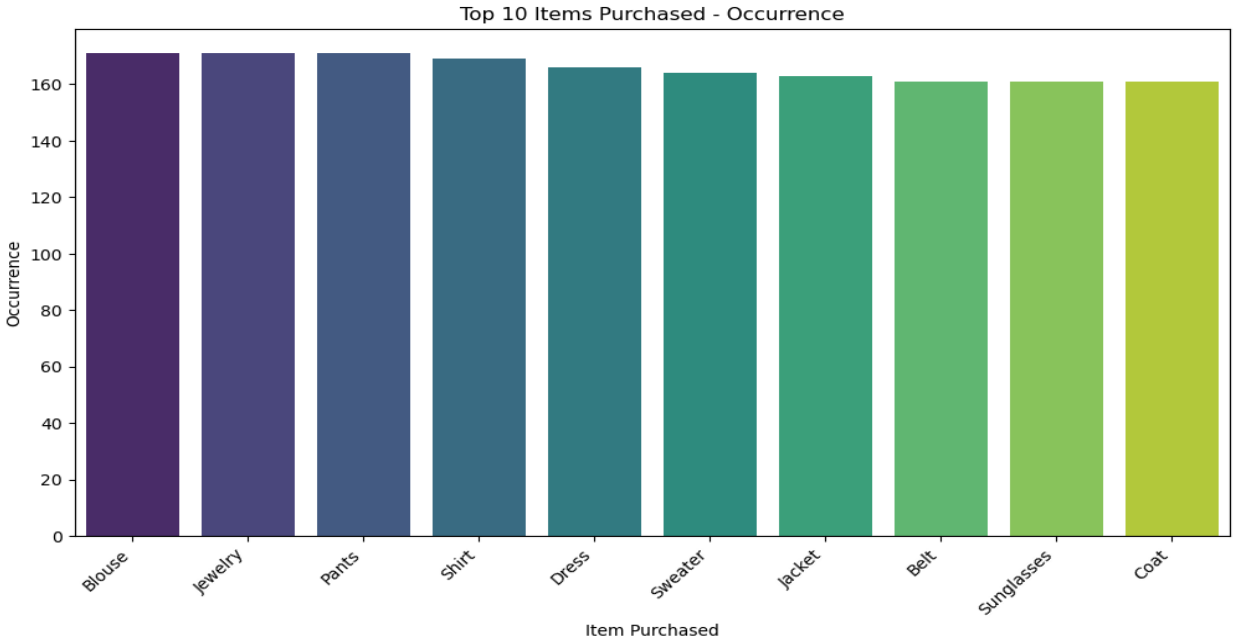
- Grouped data by category and visualized purchase amount distribution.



- Explored the top items purchased and their occurrences where Highest purch ase amount lies near to $100 and clothing again leads the chart here.

MDA 620: Capstone Project Report

- Analyzed customer demographics and purchasing behavior by age group.



- Checked with occurrences for each category for the analysis purposes with t he results as follows.

1. Footwear

	index	Occurrence
0	Sandals	160
1	Shoes	150
2	Sneakers	145
3	Boots	144

3. Outerwear

	index	Occurrence
0	Jacket	163
1	Coat	161

2. Accessories

	index	Occurrence
0	Jewelry	171
1	Sunglasses	161
2	Belt	161
3	Scarf	157
4	Hat	154
5	Handbag	153
6	Backpack	143
7	Gloves	140

4. Clothing

	index	Occurrence
0	Blouse	171
1	Pants	171
2	Shirt	169
3	Dress	166
4	Sweater	164
5	Socks	159
6	Skirt	158
7	Shorts	157
8	Hoodie	151
9	T-shirt	147
10	Jeans	124

4. Building Models:

- Prepared data for modeling by dropping unnecessary columns.
- Trained and evaluated three models:
  Here I tried to check for the models Linear Regression, Random Forest, and Gradient Boosting with the observations shown below for the ease of selecting one of the best and run that.
  But all three models were having almost same values for mean square error hence I selected the Liner regression model which is known for its simplicity and interpretability.

Linear Regression Mean Absolute Error: 20.80
Random Forest Mean Absolute Error: 20.92
Gradient Boosting Mean Absolute Error: 20.72

5. Model Selection: Linear Regression Model

- Developed a Linear Regression model to predict 'Purchase Amount.'
- Achieved a Mean Absolute Error of 20.80.

MDA 620: Capstone Project Report

6. Conclusion:

- Clothing will be the best option for the online business.
- Accessories are not having any impact based on the seasons which is the best category to stock up.
- States and Locations does not really matter for the sales and consumer shopping trend.
- Have to take care of the shipping type and charges based on the purchase amount for the various inexpensive items.
- Discounts are attractive point for the increasing the volume of sales and Consumer.
- Footwear and Outerwear need to be stock up based on requirement.

Libraries used: Pandas, Numpy, Scikit, Matplotlib, seaborn and sklearn
Model used: Liner Regression
Plots used: Scatter, Bar, Pie

GitHub Link: Click here for project files on GitHub