CUSTOMER SHOPPING TRENDS

Presented by:
Keerthana , Sai Krishna , Likitha

INTRODUCTION

ABOUT THE DATASET

The dataset contains customer information with unique IDs, demographics (age, gender), purchase details (item, category, amount, location, size, color, season), review ratings, subscription status, shipping preferences, discount usage, promo codes, previous purchase counts, preferred payment methods, and purchase frequency. This comprehensive dataset enables the exploration of diverse customer shopping trends and behaviors.

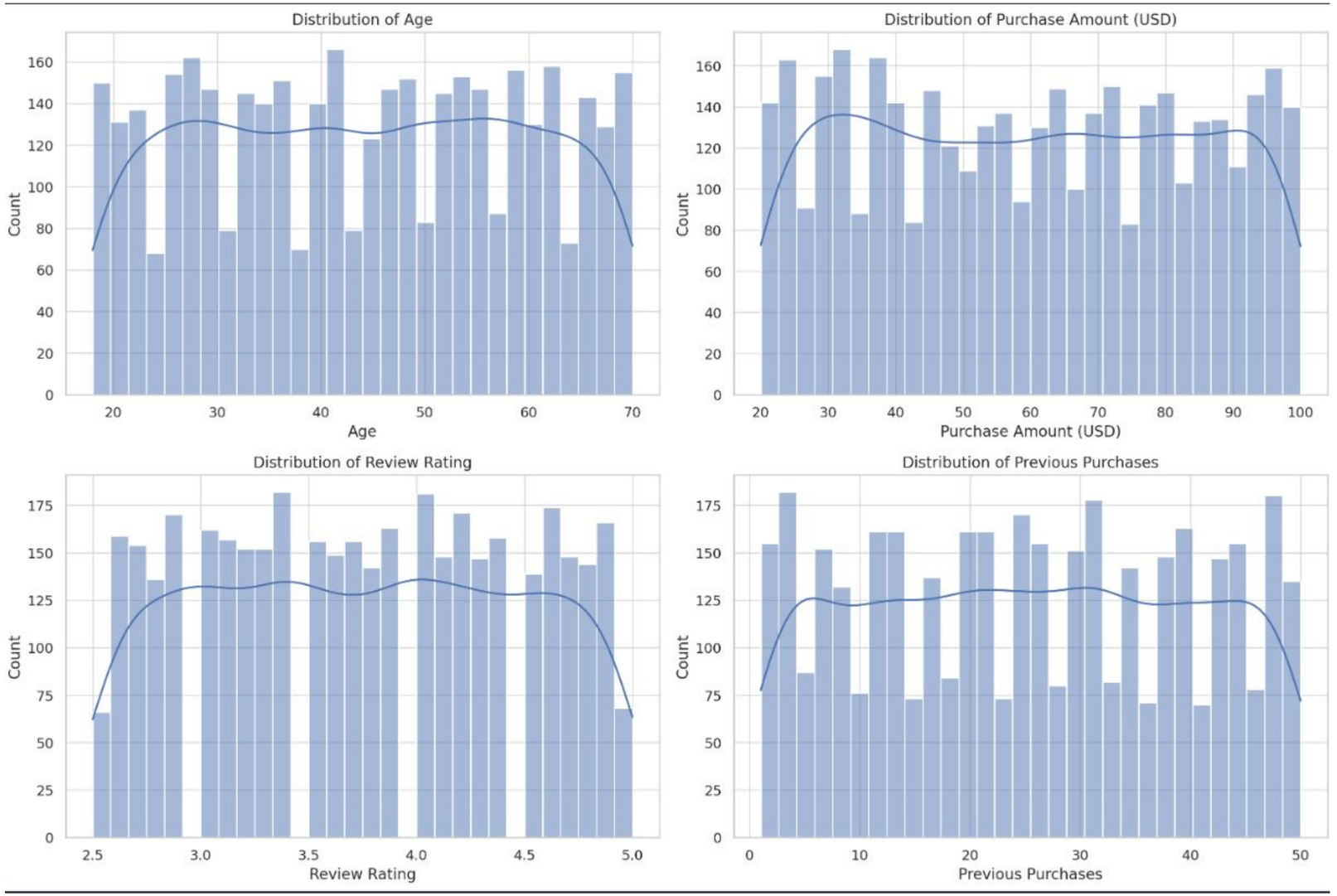
COMPRISING 4100 ROWS AND 18 COLUMNS,

DATASET

	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	Previ Purcha
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	

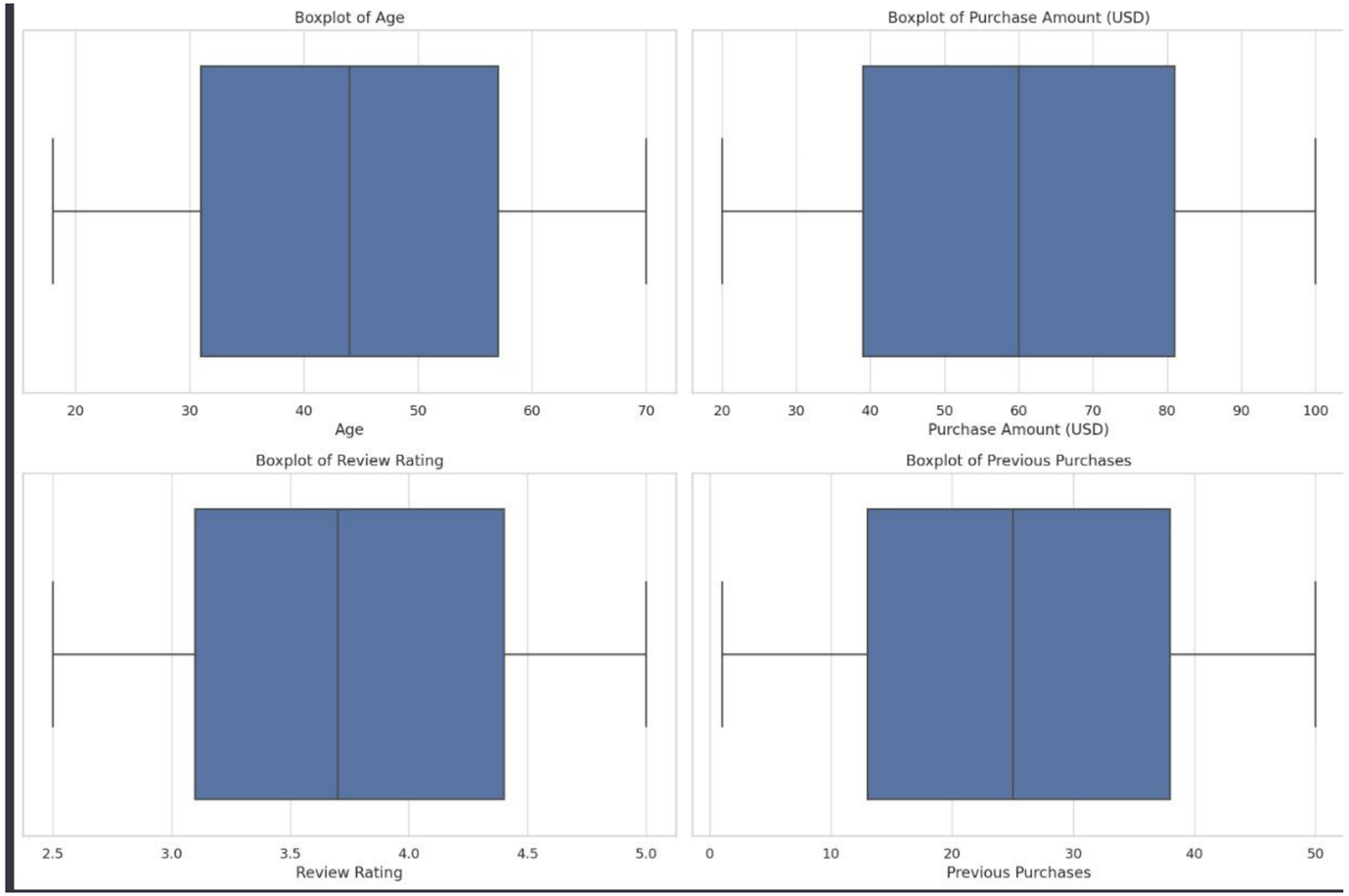
EXPLORATORY DATA ANALYSIS

HISTOGRAMS

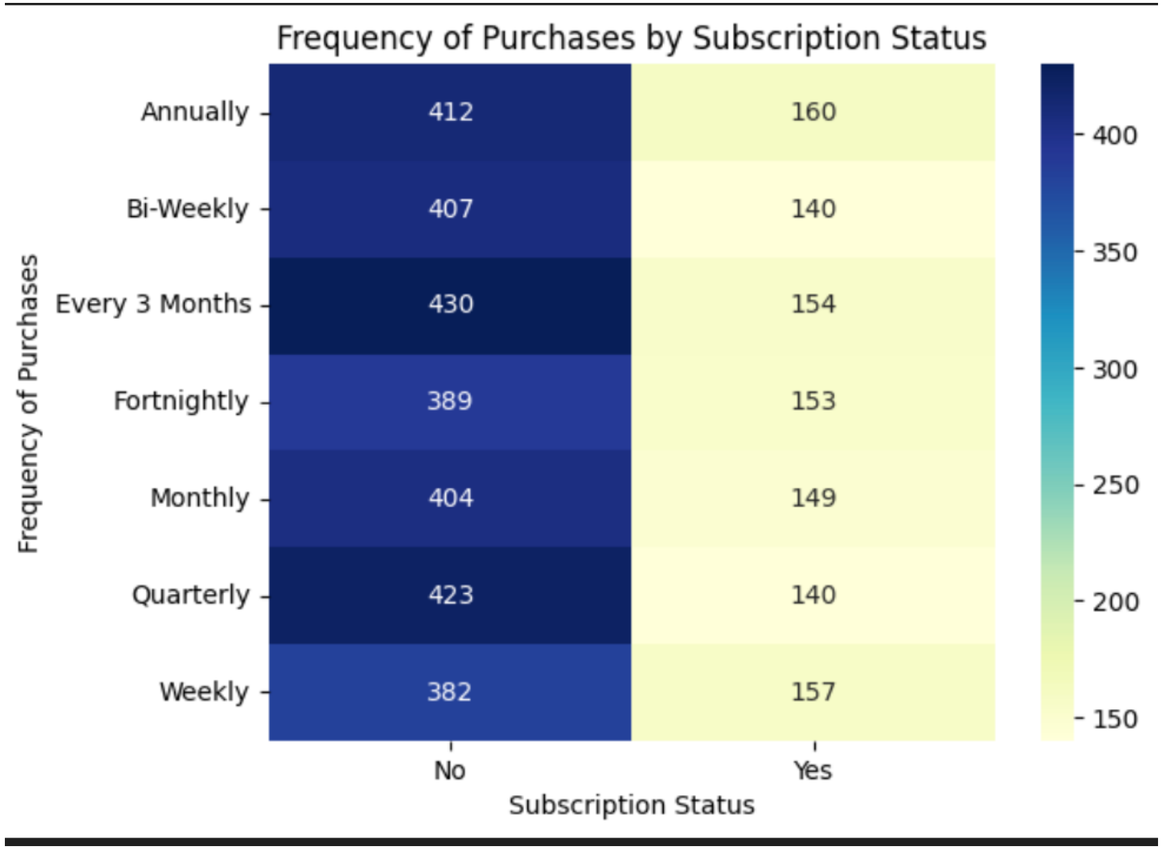

Distribution of Age
Count
Age
Distribution of Purchase Amount (USD)
Count
Purchase Amount (USD)
Distribution of Review Rating
Count
Review Rating
Distribution of Previous Purchases
Count
Previous Purchases

BOX PLOTS

These boxplots indicating that there are no significant outliers in these numerical columns.

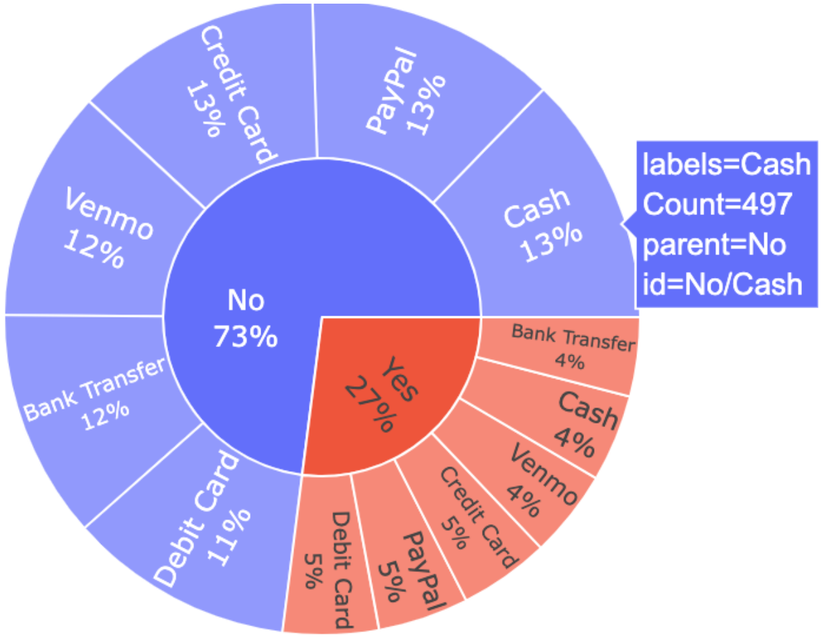


Heatmap of Frequency VS Subscription Status

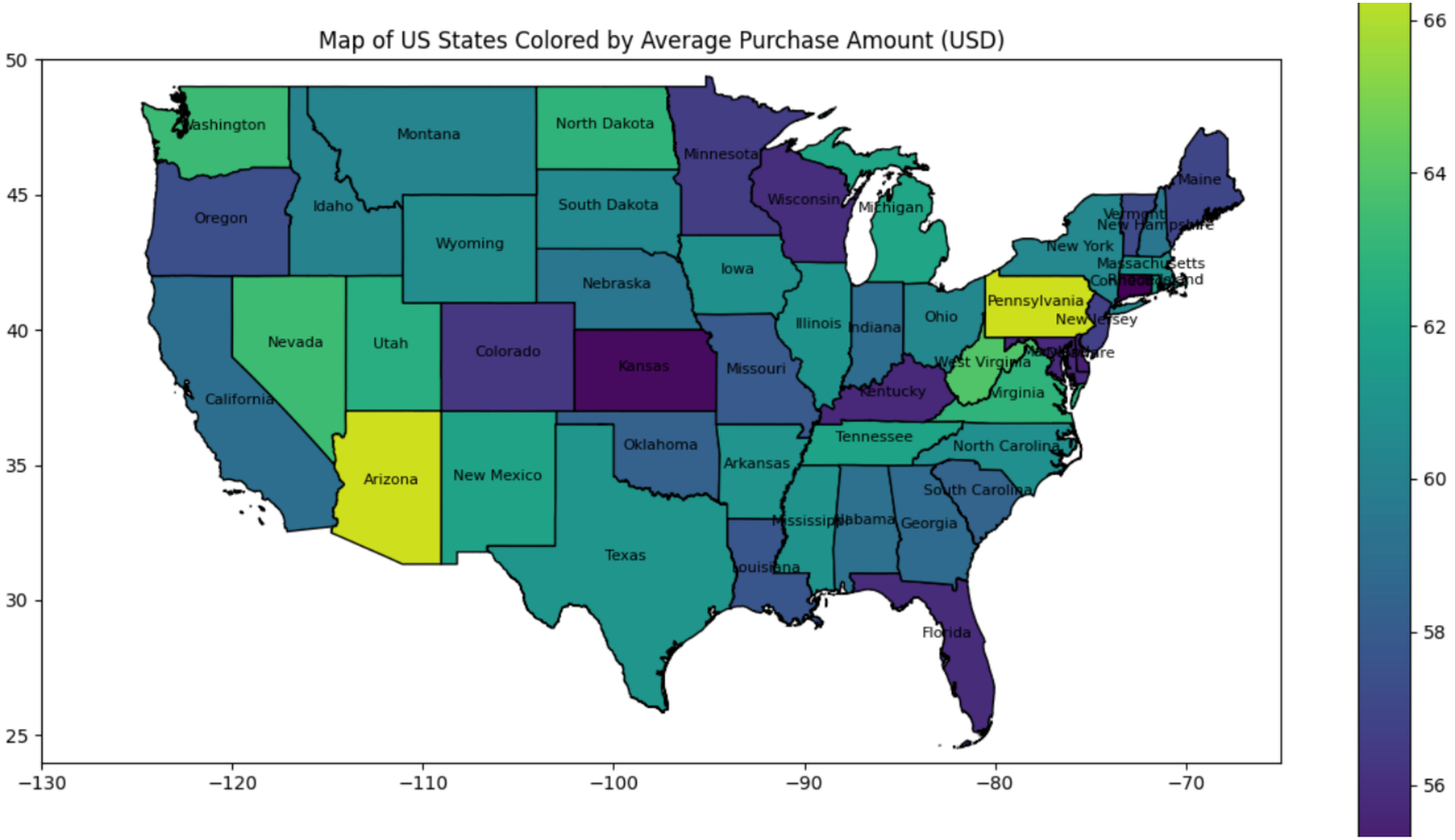


Product Category Sales by Season

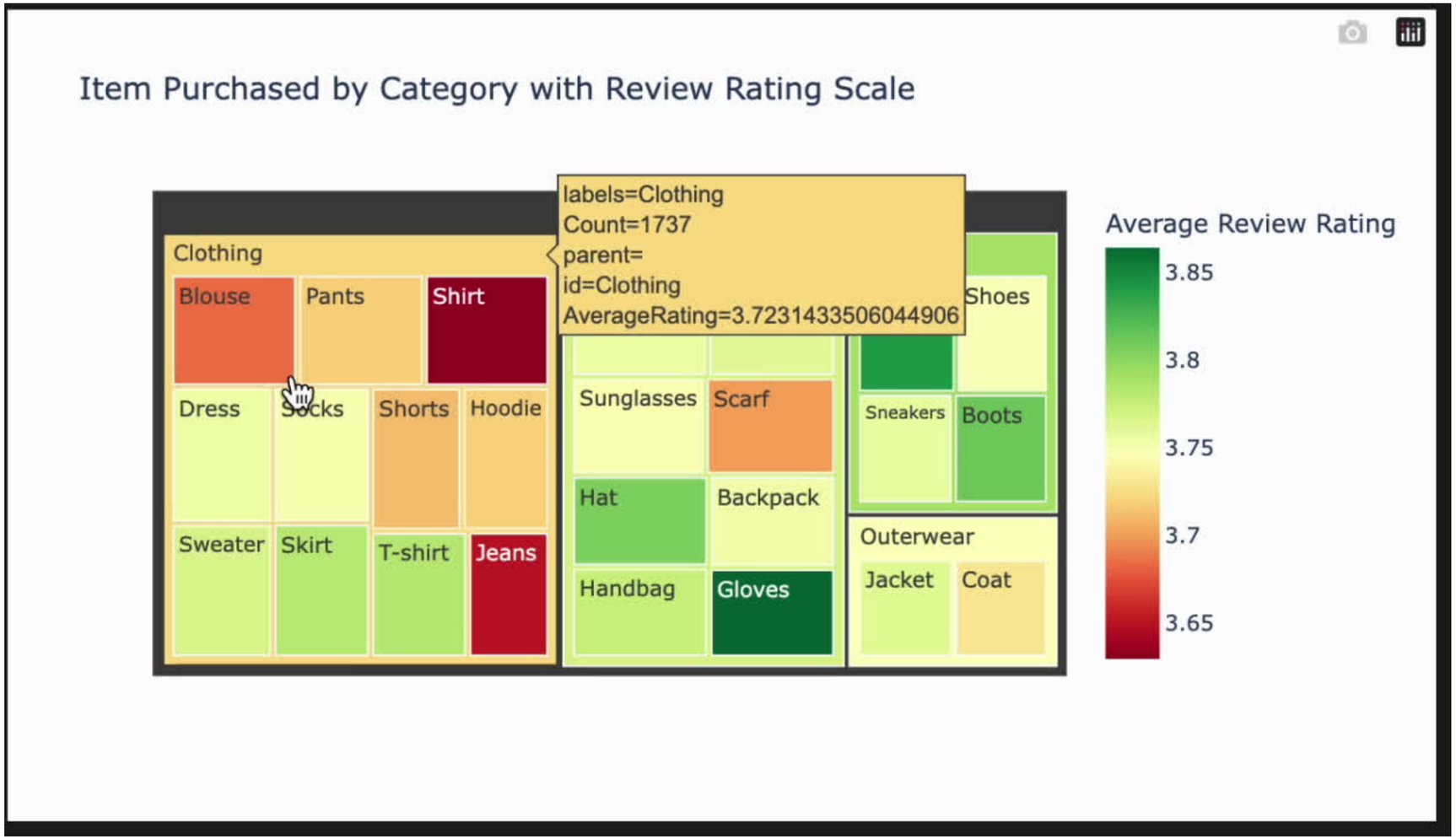
Subscription status vs Payment Method



Maps Of Location Colored By Average Purchase Amount

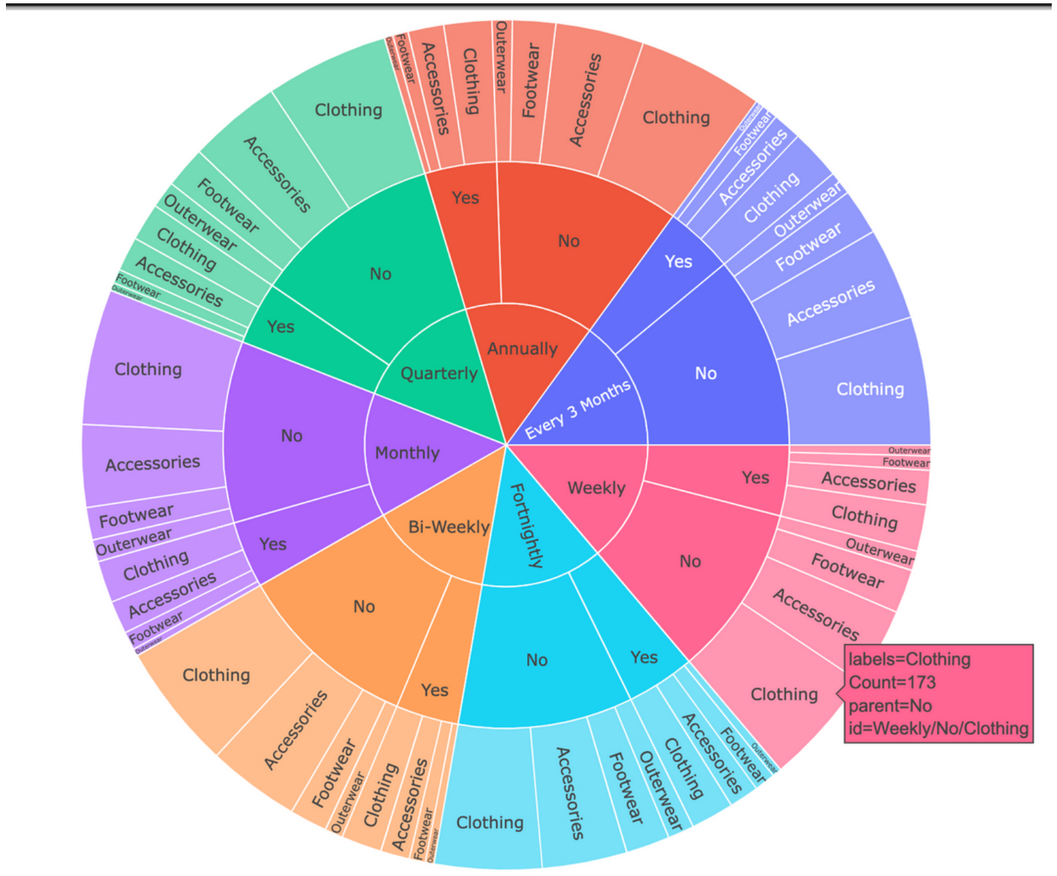


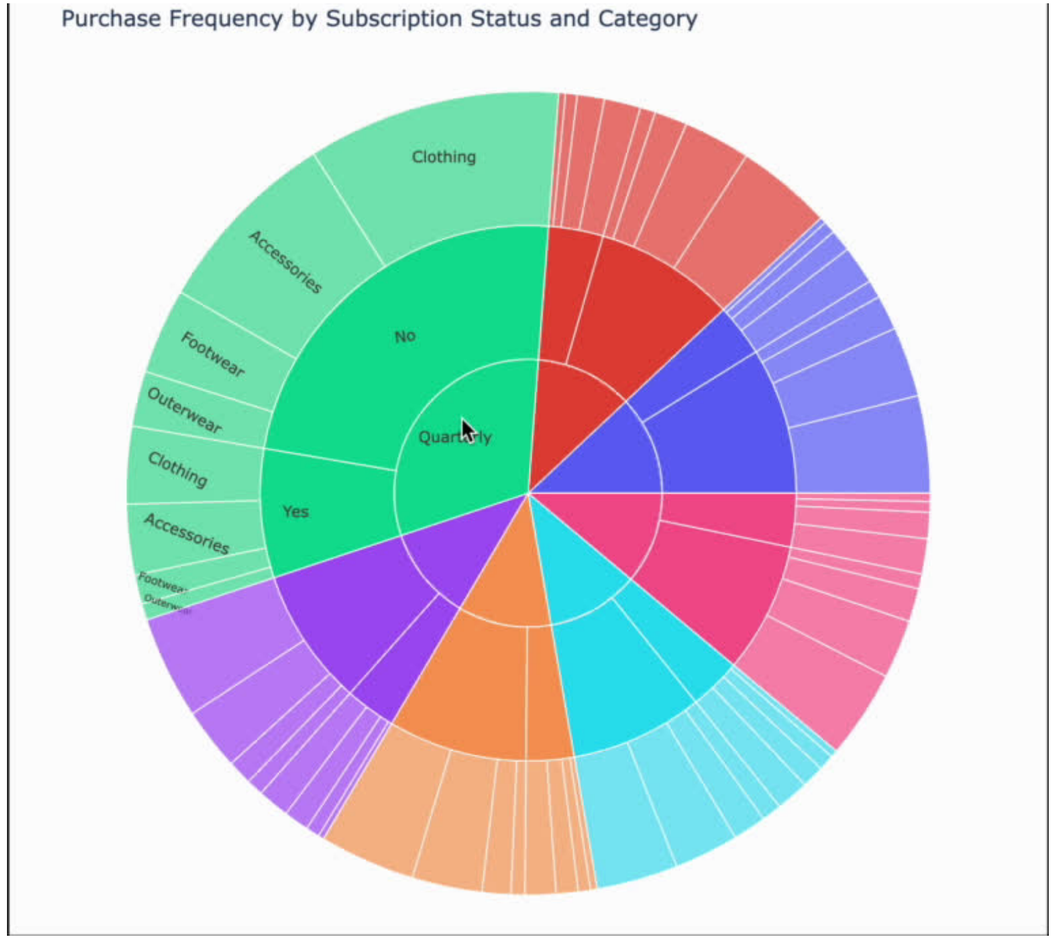
Tree Map


Item Purchased by Category with Review Rating Scale
labels=Clothing
Count=1737
parent=
id=Clothing
AverageRating=3.7231433506044906
Clothing
Blouse
Pants
Shirt
Dress
Socks
Shorts
Hoodie
Sweater
Skirt
T-shirt
Jeans
Sunglasses
Scarf
Hat
Backpack
Handbag
Gloves
Shoes
Sneakers
Boots
Outerwear
Jacket
Coat
Average Review Rating
3.85
3.8
3.75
3.7
3.65

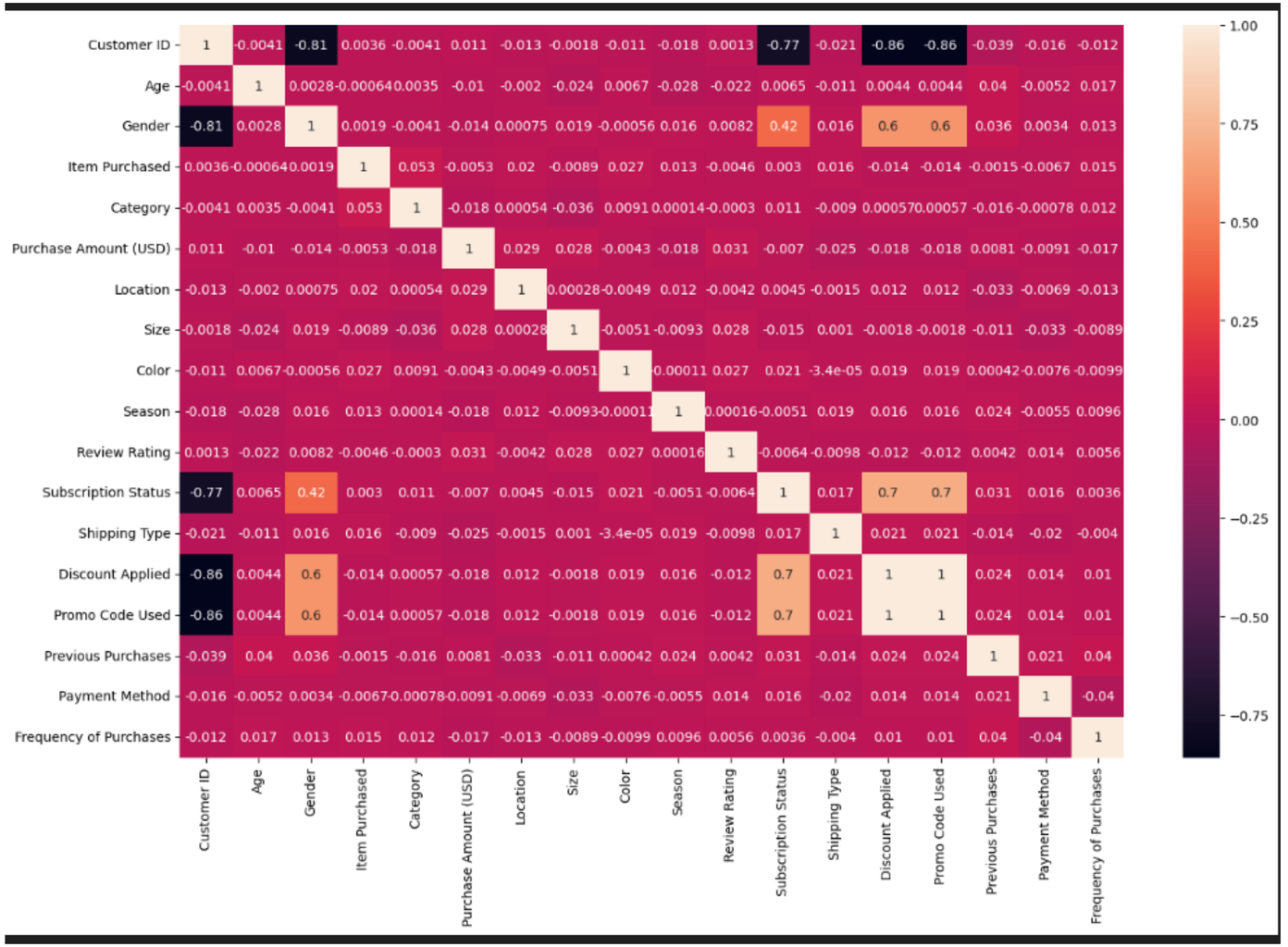
MULTIVARIANT ANALYSIS

Frequency Of Purchases VS. By Subscription VS Catrgory





Corelation between Purchases VS Subscription Status



MODELS

SMOTE ANALYSIS

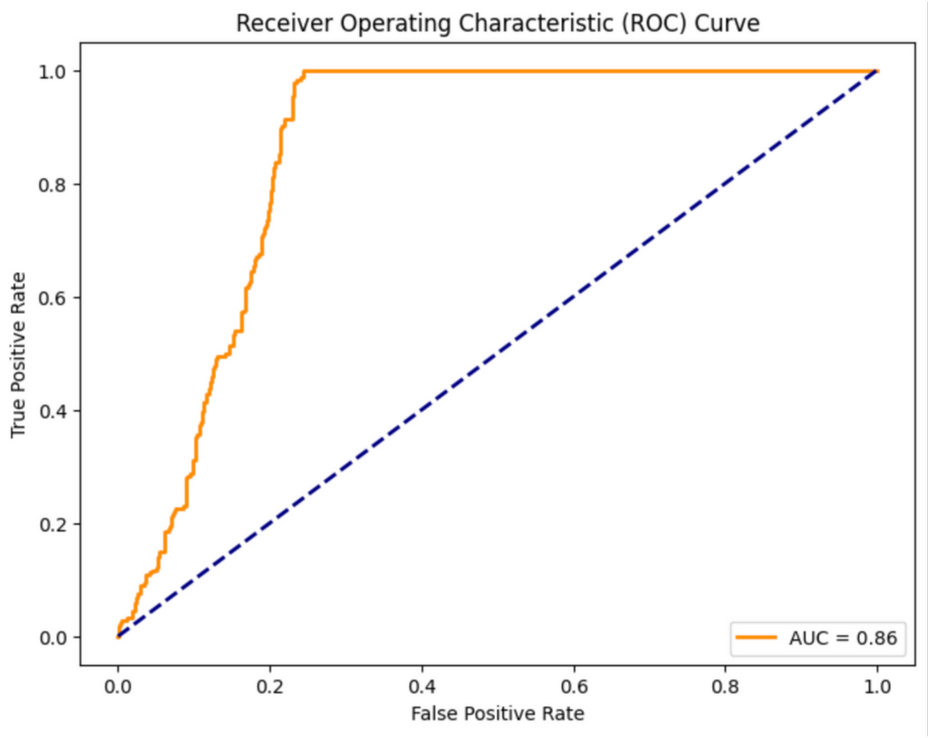
- SMOTE is a specific type of oversampling that generates synthetic examples for the minority class rather than duplicating existing instances.

- Using SMOTE with logistic regression can be particularly beneficial when dealing with imbalanced datasets, as it helps to alleviate issues related to biased models and poor generalization on the minority class.

- Variance Inflation Factor (VIF) and Multicollinearity:VIF quantifies the extent of multicollinearity in regression models by measuring how much the variance of an estimated regression coefficient increases if predictors are correlated.

Logistic Regression

ROC CURVE



Results

```
Accuracy: 0.8256410256410256
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.77      0.86       558
           1       0.62      0.98      0.76       222

    accuracy                           0.83       780
   macro avg       0.81      0.87      0.81       780
weighted avg       0.88      0.83      0.83       780

Confusion Matrix:
[[427 131]
 [  5 217]]
AUC: 0.86
```

Random Forest

RESULTS

```
Training Accuracy of the Random Forest model: 1.00
Testing Accuracy of the Random Forest model: 0.98
Accuracy of the Random Forest model: 0.98

Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       558
           1       1.00      0.94      0.97       222

    accuracy                           0.98       780
   macro avg       0.99      0.97      0.98       780
weighted avg       0.98      0.98      0.98       780
```

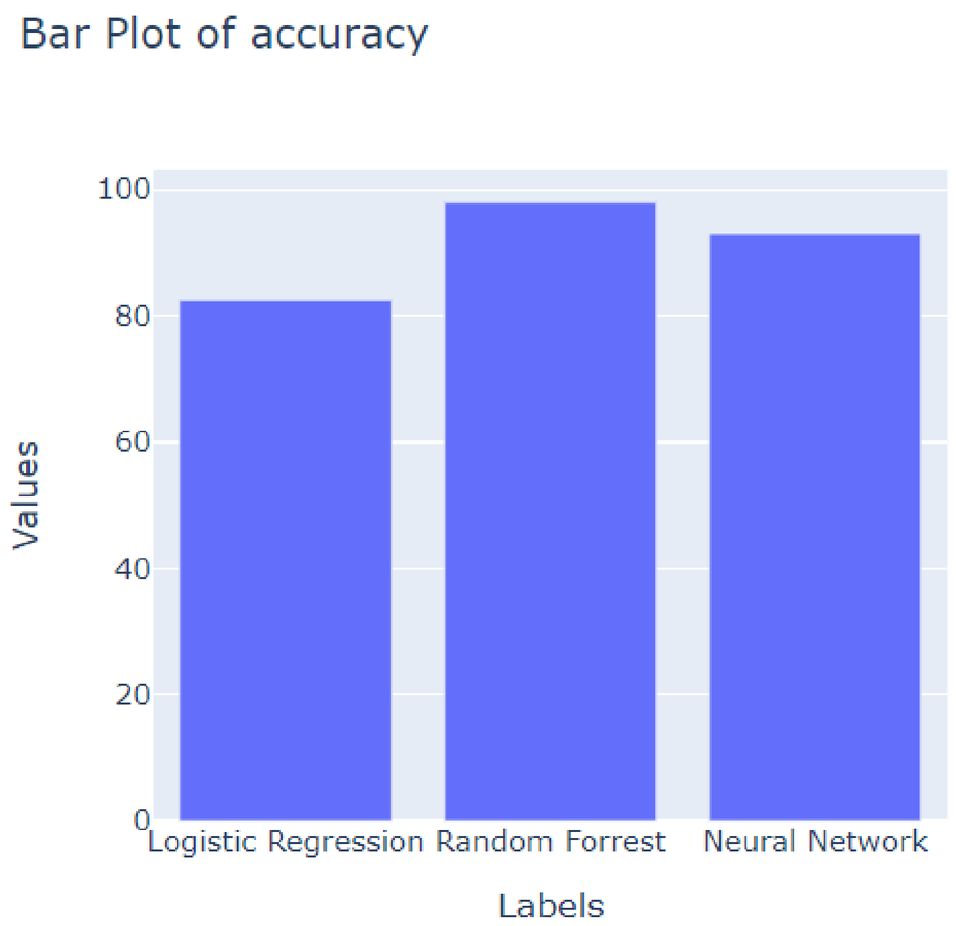
CROSS VALIDATION

```
✓ # # Cross Validation ···

Cross-Validation Accuracy Scores for 5 folds: [0.50512821 0.94230769 1.         1.         0.9474359 ]
Average CV Accuracy Score: 0.88
```

Neural Network

```
✓ # # Neural Network : MLPClassifier ···

···  Accuracy of the Neural Network: 0.93

     Classification Report:
                   precision    recall  f1-score   support

               No       0.96      0.94      0.95       834
              Yes       0.87      0.90      0.88       336

         accuracy                           0.93      1170
        macro avg       0.91      0.92      0.92      1170
     weighted avg       0.93      0.93      0.93      1170
```

RESULTS


Bar Plot of accuracy
100
80
60
40
20
0
Values
Logistic Regression
Random Forrest
Neural Network
Labels

Thank You