
# MARKETING DATASET PROJECT

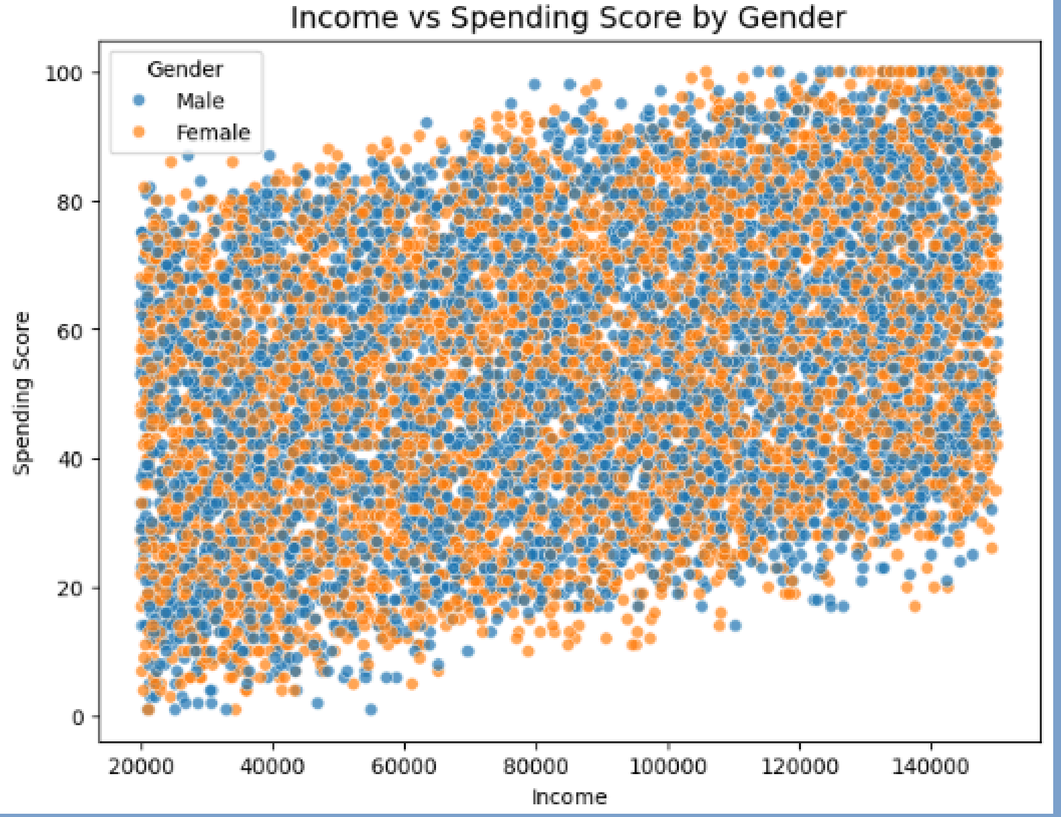
## ENT 2017: AI and Data Analytics

Group 8: LaldinPuia Hmar, Isha, Kushal Shahi, Shruti, Adarsh Tiwari

# Introduction

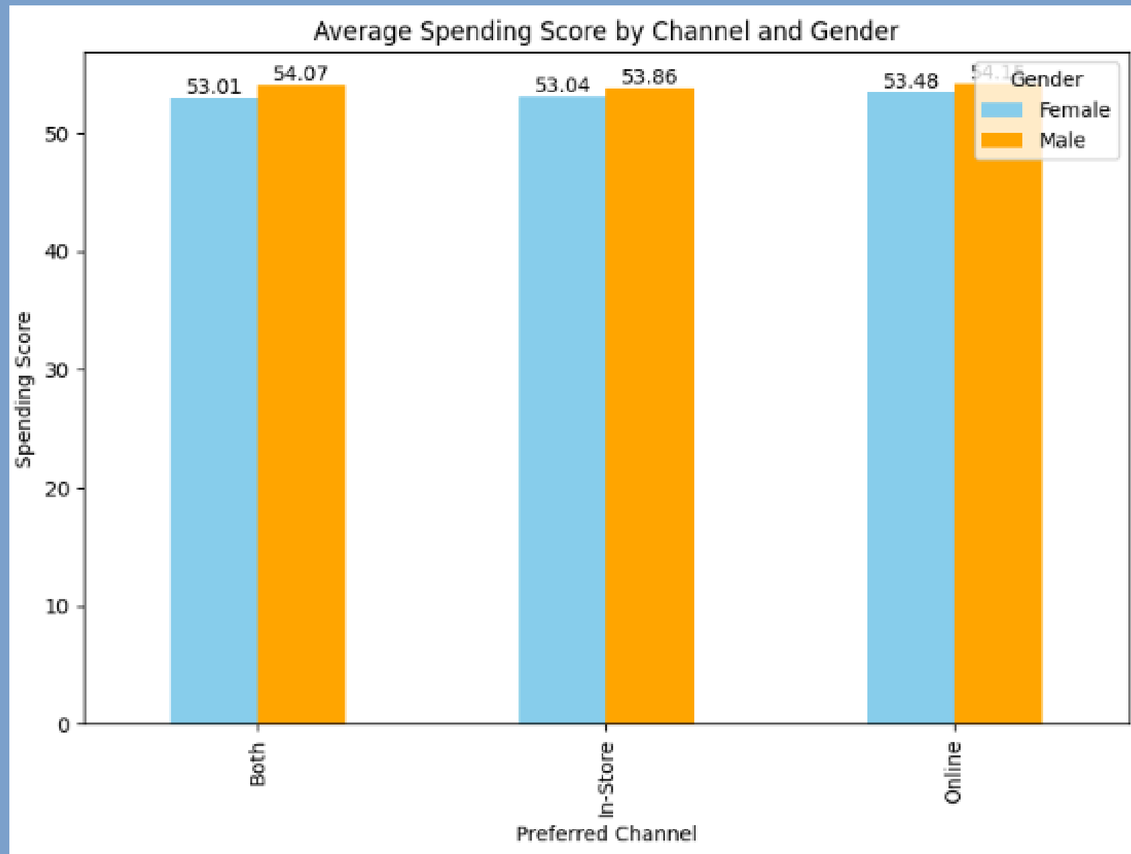
**Purpose: This project analyzes a marketing dataset with customer-level data using various AI and data analytics techniques. The goal is to identify patterns, predict behaviors, and provide actionable insights for improved marketing strategies.**

# DATA OVERVIEW



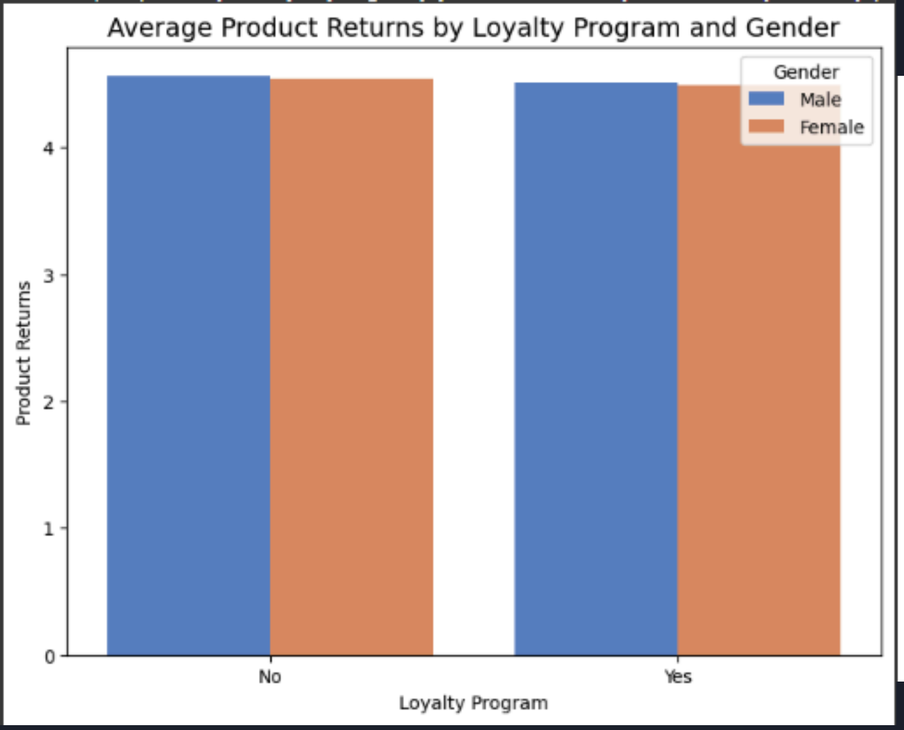

# Dataset:

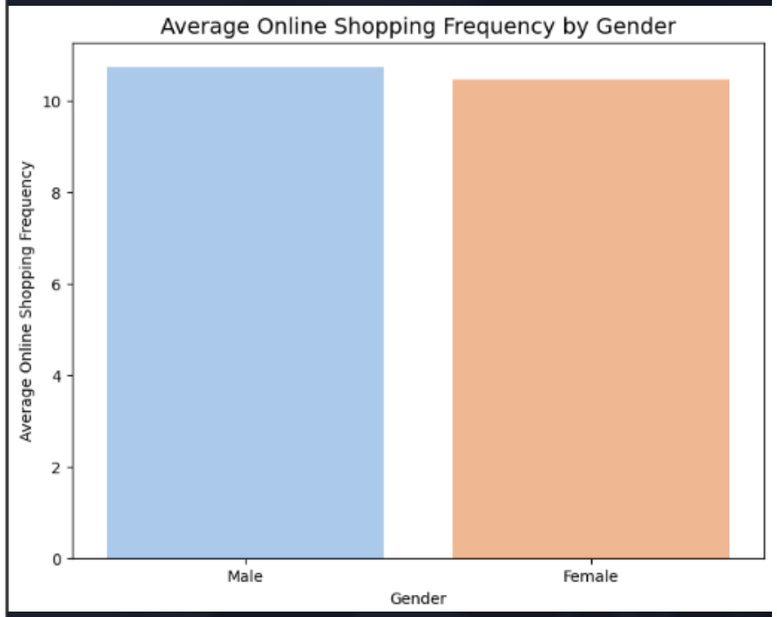
- Information on 10,000 customers
- Includes their Age (18–70 years), Gender, Income, SpendingScore (1–100), PreferredChannel of shopping (Online, In-Store, Both), etc.

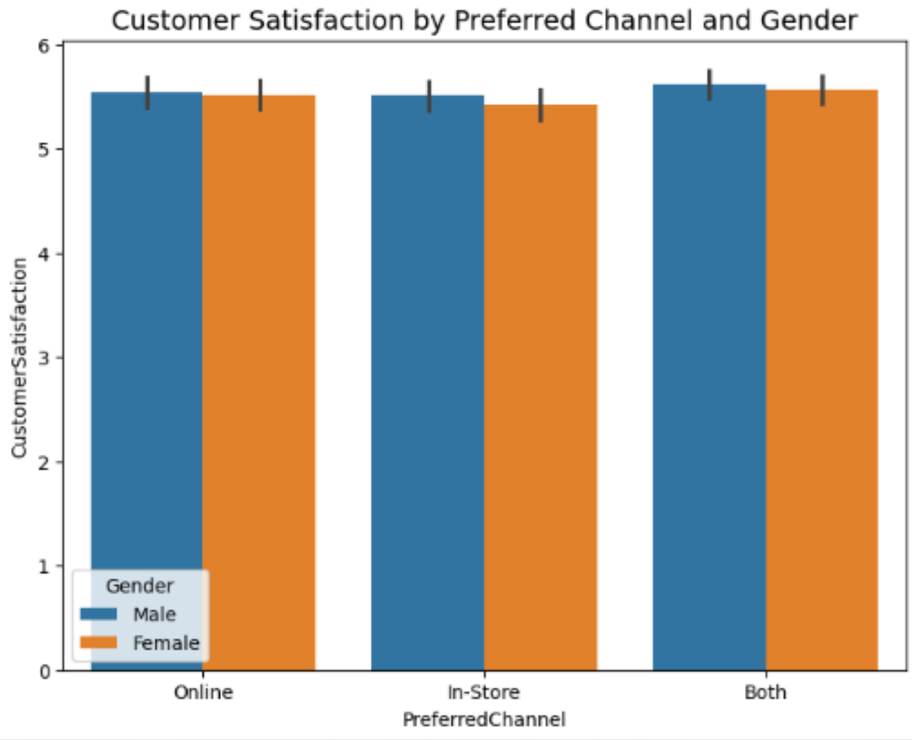



# Preliminary Observations:

- There’s a clear positive correlation b/w income and spending score
- Spending Score vs Gender: Males tend to spend slightly more than females



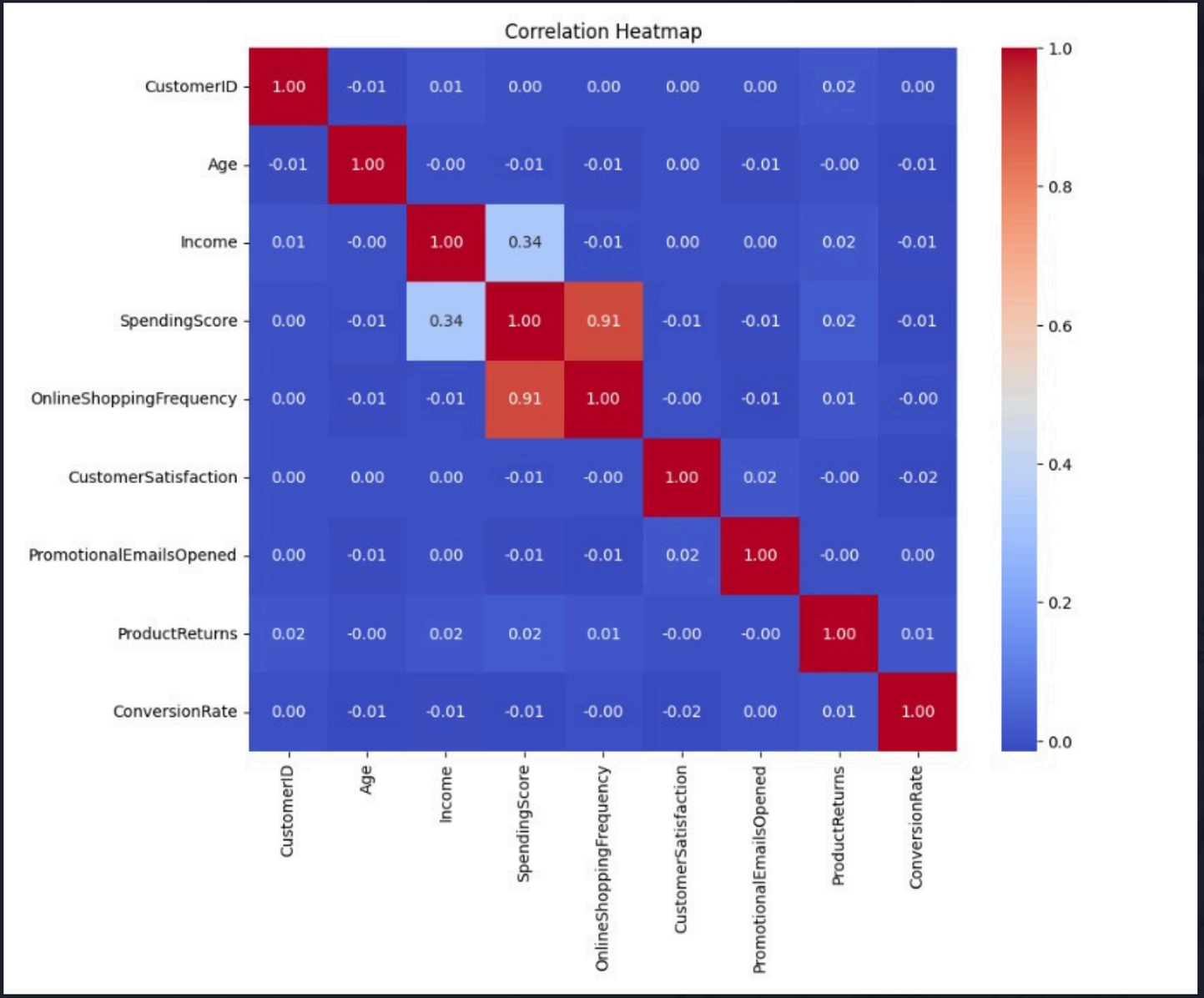




# Preliminary Observations on the dataset:

- **Online shopping freq vs Gender**
- **Customer Satisfaction vs Gender**
- **Product Returns vs Gender**

**The above 3 graphs show males tend to shop more through every channel, have greater satisfaction with products and average product returns are higher than females.**

# CORRELATION HEATMAP




- We observe high correlation between customer income and spending score
- We observe a very high correlation between spending score and Online shopping frequency of customers
- Very low to negligible correlation between all the other features

# Data Modelling

- Data Preprocessing and Feature engineering
- KMeans Clustering
- ANOVA
- Regression Analysis
- KNN
- Naive Bayes
- CART
- Logistic Regression

# Data Preprocessing

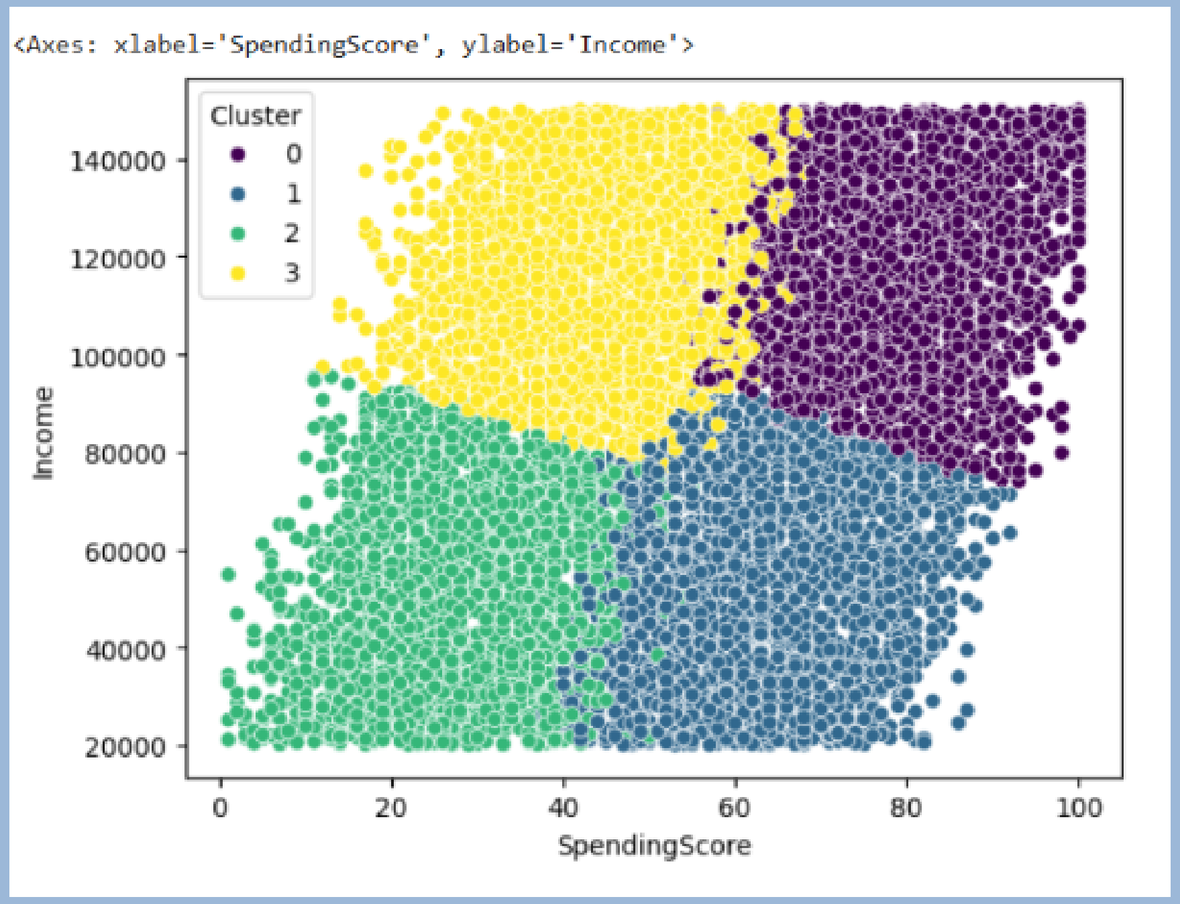
- Checked for distribution of the data using histogram-- no skewness found
- We checked for null values -- No null values in the data
- There are no duplicate values in the data
- Checked for outliers using IQR and Box plots
- Checked for balance and imbalance in the target variable-- its balanced!

# Feature Engineering

- We used LabelEncoder and OneHotencoder to transform categorical variables into binary values
- Used minmax scaler to normalize the features in the train and test data

# KMeans Clustering

## Cluster wise customer segment analysis – SpendingScore & Income

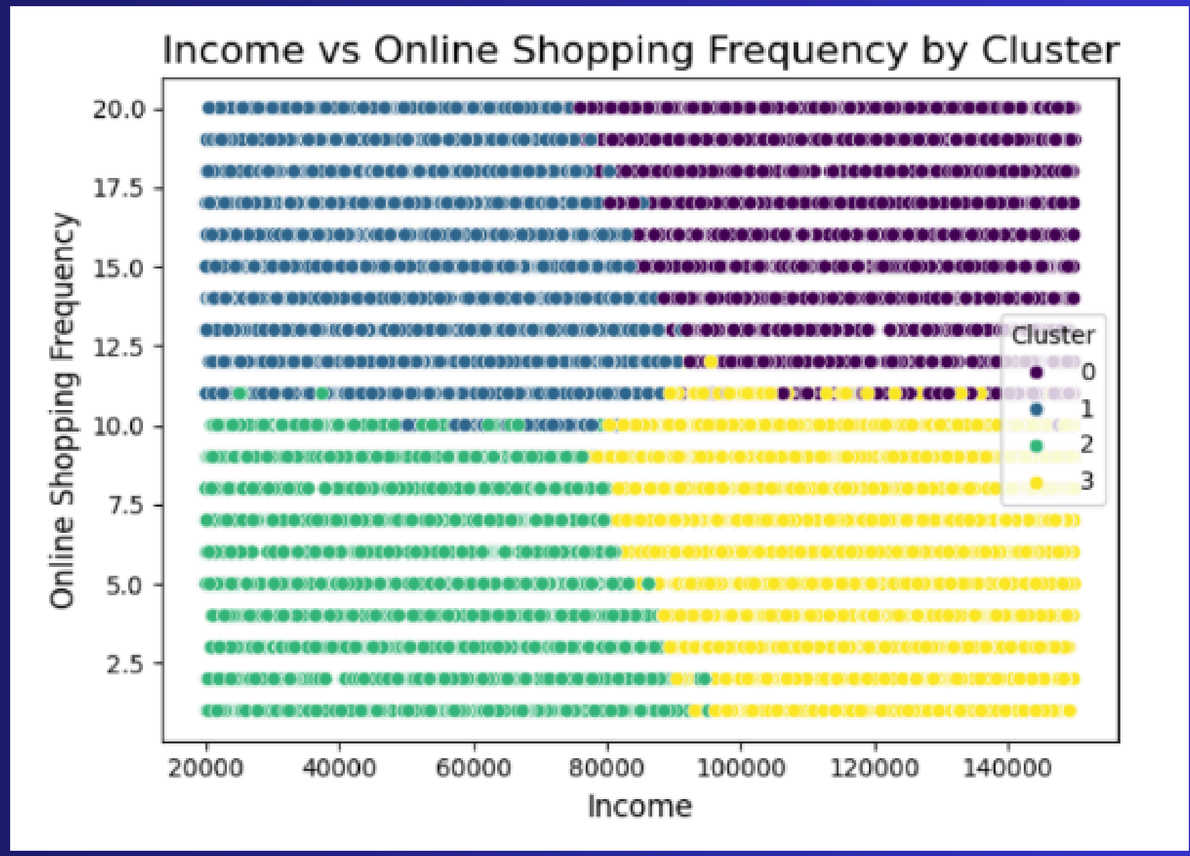



- **Cluster 0: High-income & high-spending score: Target luxury or high-priced items for this group**
- **Cluster 1: Low-income & high-spending score: Focus on affordable, value-oriented marketing**
- **Cluster 2: Low-income & low-spending score: Costly to target due to minimal purchasing behavior and potential to purchase**
- **Cluster 3: High-income & low-spending score: Opportunity to influence spending behavior for increased sales**

# KMeans Clustering

## Cluster wise customer segment Analysis – Income & Online Shopping frequency

- Cluster 0: High-income & high online shopping frequency: Target with premium online shopping campaigns.
- Cluster 1: Low-income & high online shopping frequency: promote value-oriented online purchase
- Cluster 2: Low-income & low online shopping frequency: Limited potential for targeted marketing efforts.
- Cluster 3: High-income & low online shopping frequency: Focus on increasing their online purchases through premium products



# KMeans Clustering


## Cluster wise customer segment Analysis – Online Shopping Frequency & Spending Score

- Cluster 0: High online shopping frequency & High spending score
- Cluster 1: High online shopping frequency, moderate spending score
- Cluster 2: Low online shopping frequency & lowest spending score
- Cluster 3: Low shopping frequency & moderate spending score

## Cluster wise customer segmentation & Marketing Recommendations

- Cluster 0: High-spending score, high-income, frequent online shoppers → Luxury & high-price goods
- Cluster 1: High (moderate)-spending score, low-income, High online shopping → Affordable items, frequent promotions.
- Cluster 2: Low-spending score, low-income, low online shopping → Avoid marketing focus.
- Cluster 3: High-income, low-spending score, low online shopping → Campaigns to shape spending behavior, especially online

# KMeans Clustering

## Cluster overview (Average values and Max frequencies)

| | Cluster | Age | Income | SpendingScore | OnlineShoppingFrequency | CustomerSatisfaction | PromotionalEmailsOpened | ProductReturns | ConversionRate | PreferredChannel | Gender | LoyaltyProgram |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 43.353807 | 117485.172486 | 79.278502 | 16.004207 | 5.491796 | 24.469499 | 4.587295 | 0.152780 | In-Store | Male | No |
| 1 | 1 | 43.478959 | 52506.434966 | 63.049732 | 15.079572 | 5.545907 | 24.263963 | 4.545524 | 0.155913 | In-Store | Female | Yes |
| 2 | 2 | 43.416807 | 51518.632179 | 28.067568 | 5.288851 | 5.539274 | 24.689611 | 4.404983 | 0.154892 | In-Store | Female | No |
| 3 | 3 | 43.876183 | 116669.348353 | 44.018932 | 6.045816 | 5.535025 | 24.942067 | 4.558501 | 0.157439 | Both | Male | No |

| Model | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|---|
| Naive Bayes | 47% | 50% | 51% | 47% |
| KNN | 50% | 50% | 46% | 53% |
| CART | 50% | 53% | 47% | 50% |
| Logistic Regression | 51% | 51% | 49% | 51% |

**Cluster wise highest accuracies:**

**Cluster 0: Logistic Regression has highest accuracy**

**Cluster 1: CART delivers the highest accuracy.**

**Cluster 2: Naive Bayes is most effective.**

**Cluster 3: KNN provides the best classification accuracy.**

# ANOVA Analysis

**Determine if average spending scores differ across shopping channels (Online, In-Store, Both) for the dataset**

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| PreferredChannel | 2.442923e+02 | 2.0 | 0.25624 | 0.773961 |
| Residual | 4.765431e+06 | 9997.0 | NaN | NaN |

**Findings**

- **At the 5% significance level, the average spending scores are equal across all shopping channels.**
- **Shopping channel preference does not significantly influence customer spending behavior.**

## Marketing Implications & Strategies

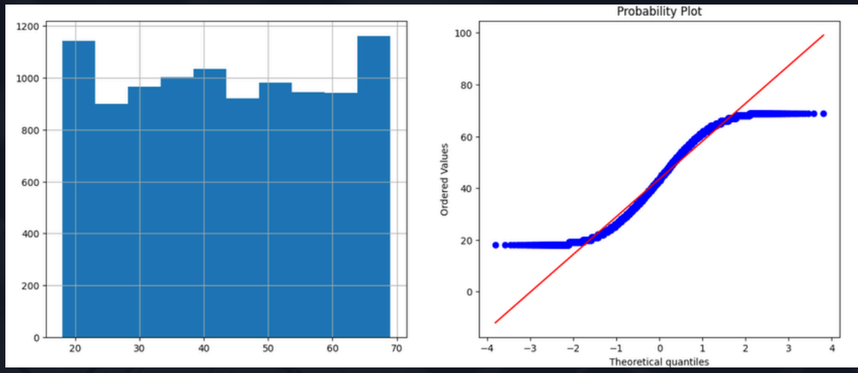
1. **Uniform Spending Behavior Across Channels: Focus on broad, cross-channel marketing campaigns instead of channel-specific promotions.**
2. **Cross-Channel Opportunities: Promote omnichannel experiences. Ex: Shop online, pick up in-store. Provide consistent promotions across all channels.**
3. **Customer-Centric Segmentation: Shift focus to customer-centric promotions such as income, spending score, or product preferences.**
4. **Channel Investment Decisions: Evaluate operational costs of each channel & Reallocate resources to improve efficiency or customer experience uniformly.**
5. **Experimentation Potential: Experiment with promotions targeting specific customer segments in each channel. Measure the impact of discounts, loyalty rewards, or new offers.**

# Regression Analysis

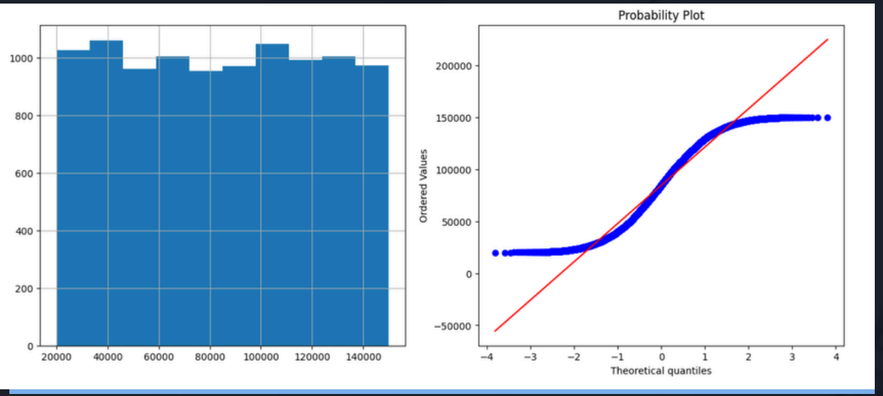
## Assumptions & Methodology

- **Linearity: Pairwise scatter plots indicate mostly linear relationships (except Age vs. Spending Score)**
- **Normality: QQ plots confirm approximate normality for features and residuals (5% significance level)**
- **Homogeneity of Variances: Levene's Test confirms equality of variances among features**
- **Mean Equality: Two-tailed t-tests show no statistical differences between mean values across features**

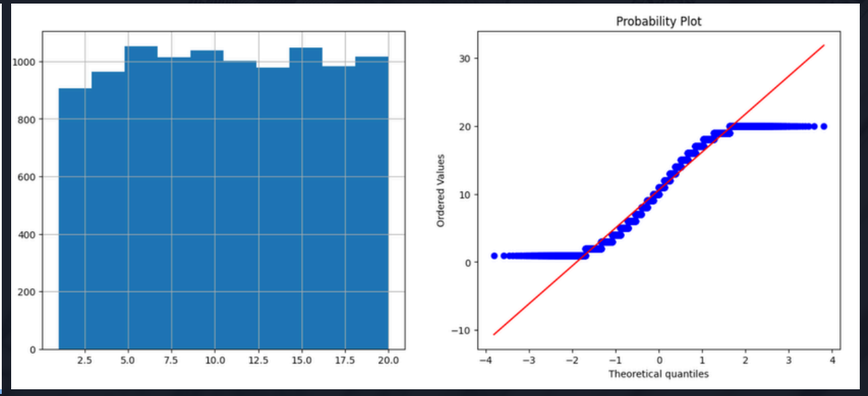
AGE



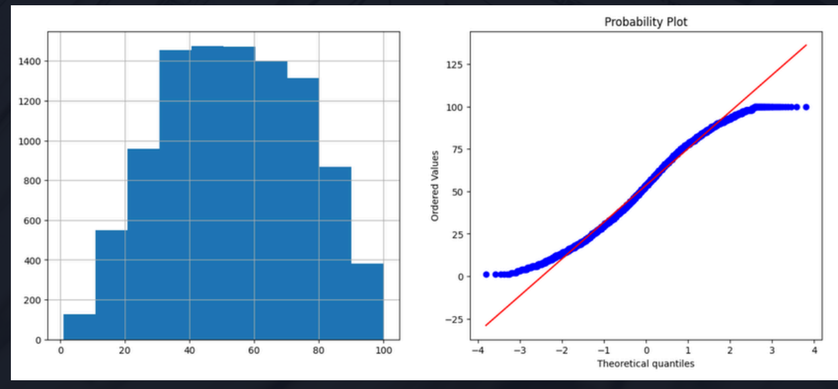

Income




Online Shopping Frequency




Spending Score



# Regression Analysis

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0621 | 0.002 | 39.600 | 0.000 | 0.059 | 0.065 |
| Age | 0.0016 | 0.002 | 0.919 | 0.358 | -0.002 | 0.005 |
| Income | 0.2609 | 0.002 | 150.948 | 0.000 | 0.257 | 0.264 |
| OnlineShoppingFrequency | 0.6706 | 0.002 | 402.318 | 0.000 | 0.667 | 0.674 |

**Coefficient Interpretation**

- **Income (0.2609): Every 1-unit increase in income raises Spending Score by 0.2609 points (statistically significant).**
- **Online Shopping Frequency (0.6706): Every 1-unit increase in shopping frequency raises Spending Score by 0.6706 points (statistically significant).**
- **Age (0.0016): A 1-year increase in age raises Spending Score by 0.0016 points (not statistically significant).**

# Regression Analysis

## Model Performance

- **Mean Squared Error (MSE) is 0.25 : This indicates that, on average, the squared difference between the actual Spending Scores and the predicted values is minimal, signifying high predictive accuracy.**
- **R-squared ($R^2$): The model explains 94.8% of the variance in Spending Score using the predictors (Age, Income, and Online Shopping Frequency). This high $R^2$ value shows that the model fits the data well, leaving only 5.2% of the variance unexplained, potentially due to factors not included in the model.**

## Insights & Implications

- **Income & Online Shopping Frequency: Significant predictors of Spending Score. Focus marketing efforts on high-income and frequent online shoppers for better returns.**
- **Age Factor: Age has minimal influence; prioritize other predictors for segmentation.**
- **High Model Accuracy: Confirms strong predictive capabilities for Spending Score using selected variables.**

# K-Nearest Neighbors (KNN) Analysis

**Classify customer loyalty program membership (Yes/No) using demographic and behavioral variables.**

**Optimization Process**

**K Value Selection:**

- **Explored k values in the range of 1 to 10.**
- **Optimal k=8, with a threshold of 0.5, provided the maximum model accuracy.**

```
Accuracy: 0.52
Precision: 0.51
Recall: 0.39
F1 Score: 0.44

Classification Report:

              precision    recall  f1-score   support

         0.0       0.52      0.64      0.57      1015
         1.0       0.51      0.39      0.44       985

    accuracy                           0.52      2000
   macro avg       0.52      0.51      0.51      2000
weighted avg       0.52      0.52      0.51      2000
```

## Overall Model Performance

| Metric | Value | Interpretation |
|---|---|---|
| **Accuracy** | 0.52 | Correct **52%** of the time, slightly better than random guessing, indicating low predictive power. |
| **Precision** | 0.51 | Only **51%** of Class 1 predictions were correct; indicates many false positives. |
| **Recall** | 0.39 | Identifies **39%** of true Class 1 samples; misses many positives (high false negative rate). |
| **F1 Score** | 0.44 | Low F1 score reflects difficulty balancing precision and recall. |

# Naive Bayes Analysis

**Classify loyalty program membership (Yes/No) using predictors such as Age, Income, Spending Score, and Online Shopping Frequency.**

```
Accuracy: 0.51
Precision: 0.00
Recall: 0.00
F1 Score: 0.00

Classification Report:

              precision    recall  f1-score   support

         0.0       0.51      1.00      0.67      1015
         1.0       0.00      0.00      0.00       985

    accuracy                           0.51      2000
   macro avg       0.25      0.50      0.34      2000
weighted avg       0.26      0.51      0.34      2000
```

1. **Overall Accuracy:**
   - **The model is correct in 51% of predictions.**
   - **However, accuracy alone is not meaningful as the model appears heavily biased toward Class 0.0.**
2. **Class 1.0 (Loyalty Program Members):**
   - **Precision: 0.00 – None of the predicted Class 1.0 instances are correct.**
   - **Recall: 0.00 – The model fails to identify any true instances of Class 1.0.**
   - **F1-Score: 0.00 – Indicates a complete failure to balance precision and recall for Class 1.0.**

# Naive Bayes Analysis

3. Class 0.0 (Non-Members):
   - Precision: 0.51 – About half of the predictions for Class 0.0 are correct.
   - Recall: 1.00 – Captures all actual Class 0.0 instances (perfect recall).
   - F1-Score: 0.67 – Higher than Class 1.0 due to perfect recall, despite limited precision.
4. Class Imbalance Handling:
   - For class 0.0:
     - Precision (0.51): About half of the predictions for class 0.0 are correct.
     - Recall (1.00): The model predicts every instance as 0.0, so it captures all actual 0.0 labels perfectly.
     - F1-Score (0.67): Higher than class 1.0 because the model completely neglects class 1.0.

# Logistic Regression Analysis

**Predict loyalty program membership (Yes/No) using variables such as Age, Income, Spending Score, and Online Shopping Frequency.**

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

         0.0       0.49      0.62      0.55      1015
         1.0       0.46      0.34      0.39       985

    accuracy                           0.48      2000
   macro avg       0.48      0.48      0.47      2000
weighted avg       0.48      0.48      0.47      2000
```

## Overall Metrics

| Metric | Value | Interpretation |
|---|---|---|
| **Accuracy** | 0.48 | The model is correct **48%** of the time, which is only slightly better than random guessing (~50%). |
| **Macro Average** | Precision: 0.48, Recall: 0.48, F1-Score: 0.47 | Indicates poor balance between the two classes. |
| **Weighted Average** | Precision: 0.48, Recall: 0.48, F1-Score: 0.47 | Weighted results are slightly skewed toward Class 0.0 due to its higher representation (1015 vs. 985). |

# Logistic Regression Analysis

## Feature-Specific Results

| Feature | Coefficient | Odds Ratio | Interpretation |
|---|---|---|---|
| **Age** | 0.0838 | 1.0874 | A one-unit increase in age raises the odds by **8.74%**, making it the strongest predictor. |
| **Income** | 0.0281 | 1.0284 | A one-unit increase in income raises the odds by **2.84%**. |
| **Spending Score** | 0.0102 | 1.0102 | Spending Score has a minimal effect, increasing odds by only **1.02%**. |
| **Online Shopping Frequency** | -0.0253 | 0.9750 | A one-unit increase reduces the odds by **2.50%**, indicating a negative effect. |

# Logistic Regression Analysis

## Insights from the Visualization

- **Online Shopping Frequency: Negatively affects the likelihood of loyalty program membership (Odds Ratio<1).**
- **Age, Income, and Spending Score: Positively affect membership, with Age having the most significant impact.**

## General Observations

1. **Age as the Strongest Predictor: A one-unit increase in age raises the odds by 8.74%, indicating its substantial influence on program membership.**
2. **Minimal Impact of Spending Score and Income: While both have a positive effect, their contributions to the odds are small.**
3. **Negative Effect of Online Shopping Frequency: This surprising result suggests that customers with high online shopping frequency are slightly less likely to join the loyalty program.**

# Final Recommendations

1. **Customer-Centric Marketing:**
- **Cluster 0: Target high-income, high-spending customers with premium, luxury products.**
- **Cluster 1: Use frequent promotions for affordable items to attract low-income, high online shoppers.**
- **Cluster 3: Encourage spending through personalized campaigns for high-income, low-spending customers.**

**2. Optimize Omnichannel Experience:**
- **Promote "shop online, pick up in-store" campaigns.**
- **Ensure consistent promotions across all channels to drive engagement.**

**3. Focus on High-Impact Predictors:**
- **Prioritize marketing to high-income and frequent online shoppers for maximum ROI.**
- **Deprioritize age and low-spending customers for resource efficiency.**

**4. Experiment & Adapt:**
- **Test loyalty rewards and targeted promotions to improve retention.**
- **Address barriers to loyalty program adoption for frequent online shoppers.**

# LIMITATIONS

- **Low Model Accuracy: The model does not perform well in predicting Loyalty Program membership.**
- **Hyperparameter Tuning: Despite using GridSearchCV and experimenting with thresholds and parameters, results remain suboptimal.**
- **Data Quality Concerns: Even after extensive preprocessing, the model's performance is limited, possibly due to issues with data quality.**
- **Balanced Target Variable: Despite the balanced data our model has poor performance suggesting possibility of issue with the data**
- **Future Suggestion: Acquire more detailed and high-quality data to improve model performance.**