Assignment DAI-101 Aadit - 23114001

Introduction

This is a Customer Transaction Dataset. This dataset consists of multiple numerical features, including **Age, Annual Income, Spending Score, Purchase Frequency, and Transaction Amount**. The goal of this analysis is to uncover relationships, trends, and potential outliers in the dataset to aid in data-driven decision-making. Various statistical and visualization techniques have been used to achieve this objective.

Steps for Data Cleaning Process:

1. **For Handling Missing Values:**
    - I checked for missing values in all columns and replaced them with mean for numerical data and with mode for categorical data.
    - I found no missing values in Customer_ID.
2. **Duplicate Removal:**
    - I identified 200 duplicate rows and removed duplicates, reducing dataset from 10,200 to 10,000 entries.
3. **Inconsistent Data Handling:**
    - I defined inconsistent data as those whose (Transaction Amount)* (Purchase Frequency)> (Annual Income).
    - I found and removed 10 inconsistent rows.
    - The new dataset shape: **9,990 rows, 6 columns**.
4. **Outlier Detection and Removal:**
    - I used **IQR method** to remove extreme outliers, resulting in **9,777 rows**.
    - I also applied **Z-Score method** to further refine, yielding a final dataset of **9,912 rows**.
5. **Feature Cleaning(Can be done for any other Categorical data):**
    - Customer_ID: In this analysis I treated this as categorical.

- I removed special characters using **REGEX** and standardized case using .upper().
- Stripped unnecessary spaces.
- Formatting in a specific format can be done if required like date, time, etc.

Exploratory Data Analysis (EDA):

1. **Outlier Analysis:**
   - Boxplots and histograms confirmed the presence of outliers.
   - I addressed using IQR and Z-score methods.
2. **Correlation Analysis:**
   - Annual_Income and Spending_Score showed a weak positive correlation.
   - Purchase_Frequency and Transaction_Amount had a moderate correlation.
   - Age and Spending_Score exhibited no significant correlation.
3. **Categorical Data Analysis:**
   - Customer_ID was cleaned and formatted(although they were quite regular).
   - Distribution of unique values examined.

Final Dataset:

- Shape: **(9,912, 6)**
- Cleaned, formatted, and outlier-adjusted dataset ready for modeling.

Methods for Exploratory Data Analysis

1) Correlation Matrix

- A correlation matrix was computed to determine the relationships between numerical features.
- This helps in identifying potential dependencies and redundant variables.

2) Pairwise Scatter Plots & Pairplot

- This gives visualized distributions and relationships between numerical variables.
- Allowed for the detection of patterns and potential linear/nonlinear associations.

3) Box Plots with Z-Score Outlier Detection

- Box plots are used to identify potential outliers based on statistical thresholds.
- Outliers may indicate anomalies or special cases in the dataset that require further investigation.

4) Line and Bar Graphs for Trend Analysis

- This is used to study the variation of **Annual Income, Spending Score, and Transaction Amount** with **Purchase Frequency**.
- Helps in determining trends and potential predictive patterns.

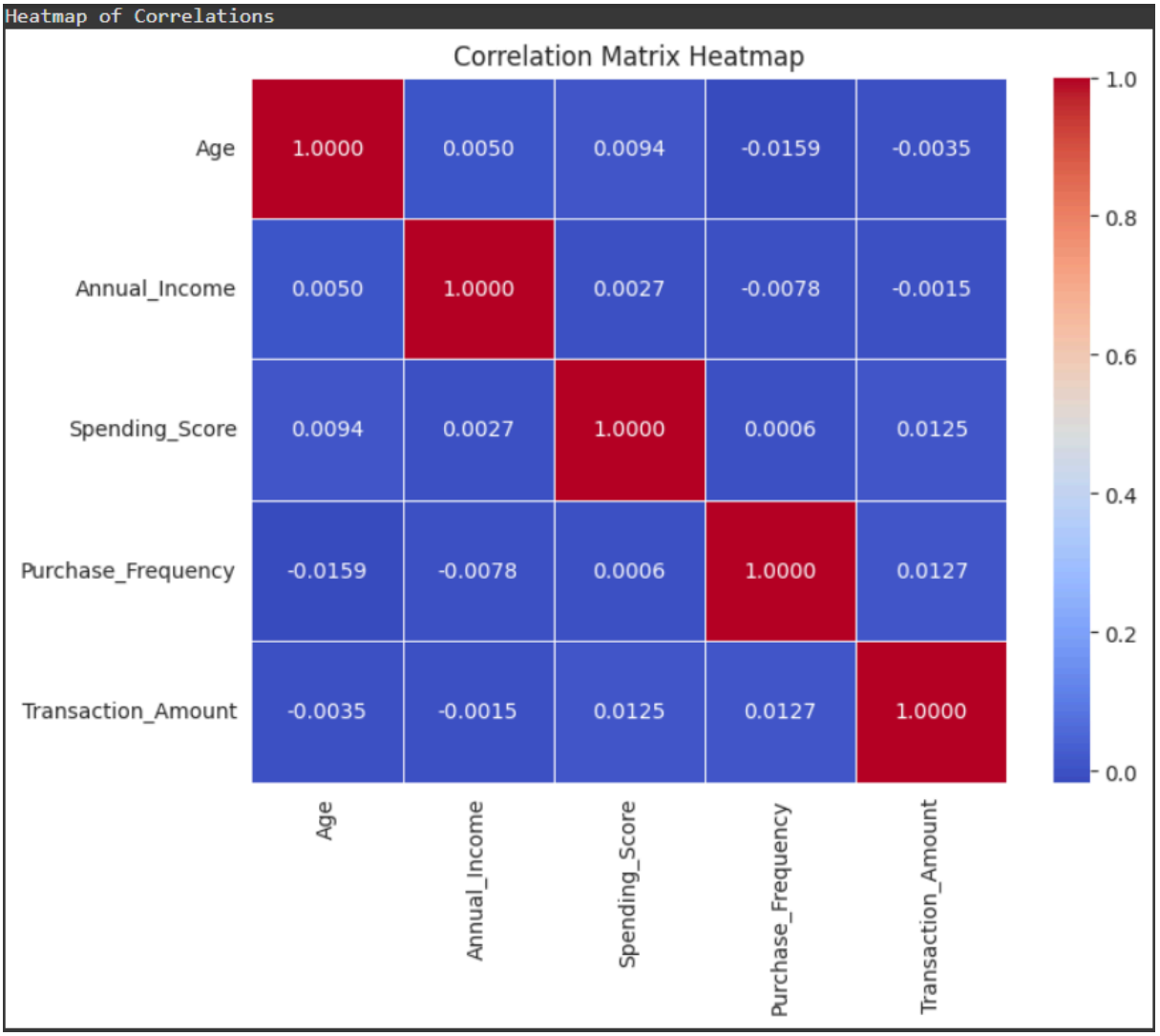
5) Scatter Plots for Cluster Identification

- Plotted **Age vs Spending Score**, **Age vs Transaction Amount**, etc., to check for natural groupings.
- This aids in segmenting the customers based on their purchasing behavior.

Observations and Inferences from EDA Analysis

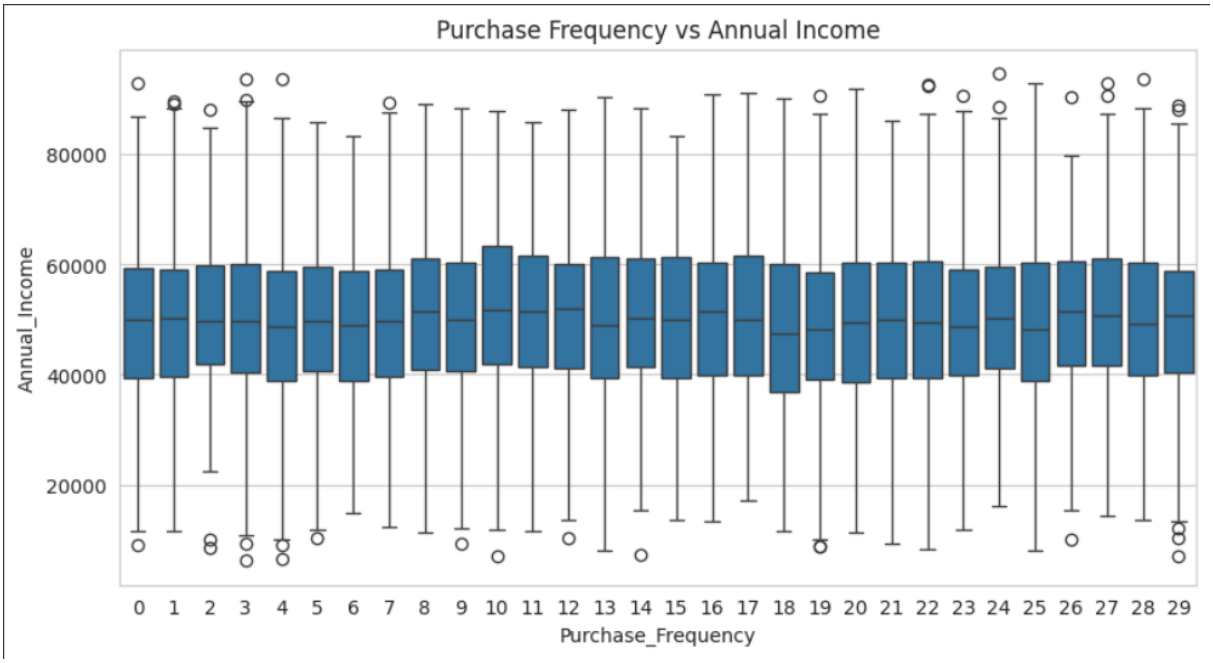
1) Correlation Matrix

- Low correlation among variables, indicating minimal linear dependency.
- **Annual Income and Purchase Frequency (-0.0077)** show no strong relationship, suggesting income does not significantly influence purchase frequency.
- **Spending Score and Transaction Amount (0.0125)** have a minor positive correlation, indicating customers with higher spending scores tend to have slightly higher transaction amounts.



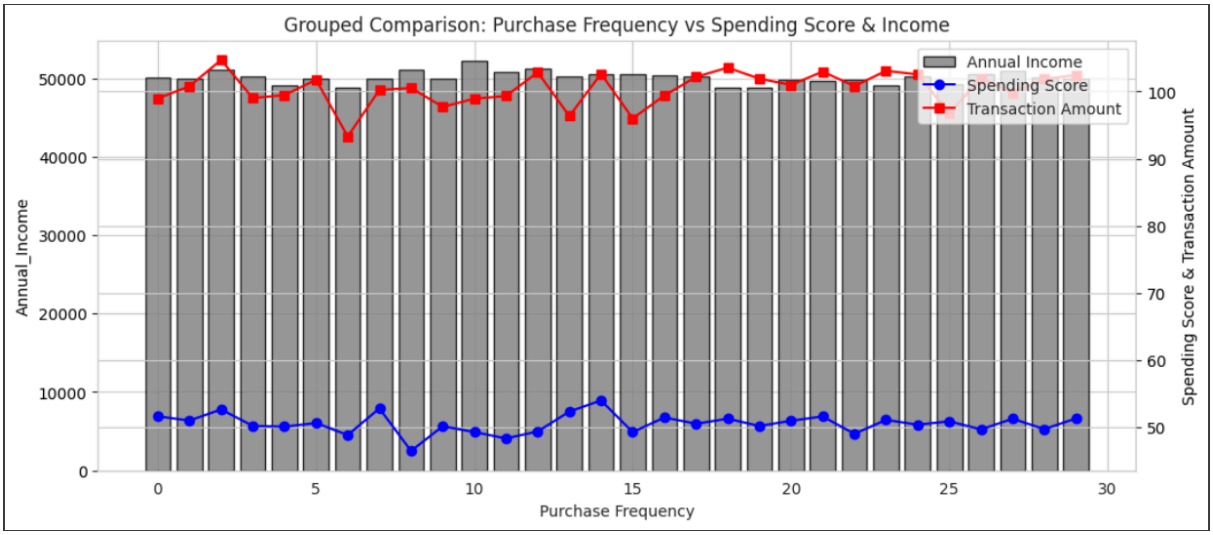
2) Purchase Frequency vs Annual Income (Screenshot 1)

- **No clear upward or downward trend.** Income remains relatively stable across purchase frequencies.
- Income fluctuates significantly with purchase frequency.
- Peak income occurs at around **Purchase Frequency = 10**, suggesting high-spending users in this range.



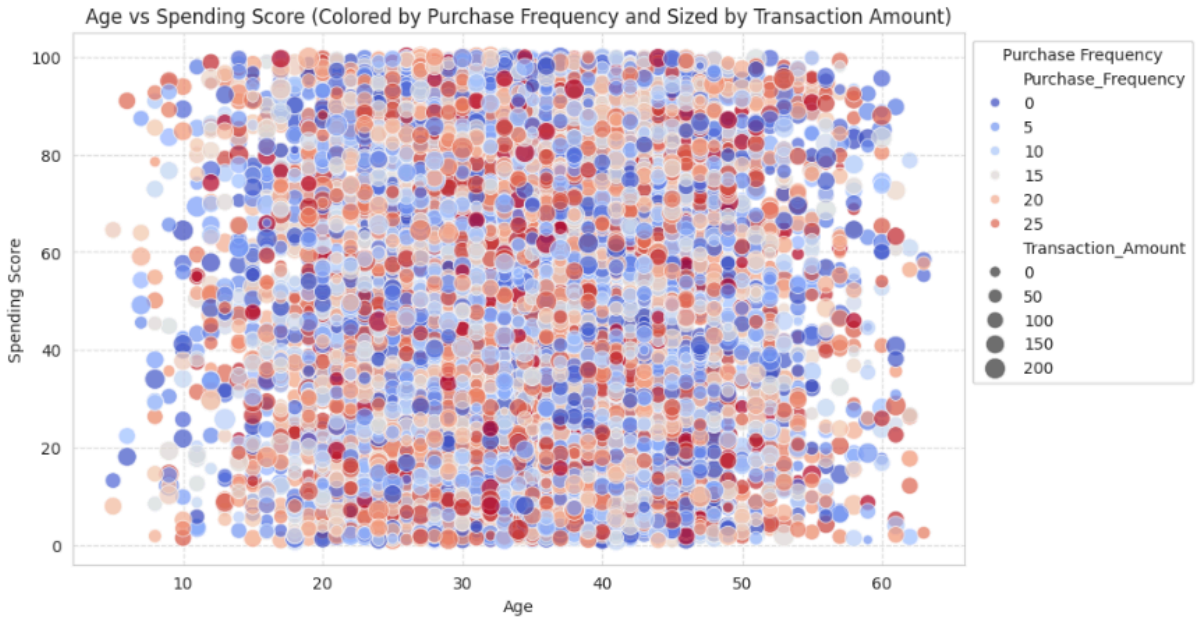
3) Grouped Comparison: Purchase Frequency vs Spending Score & Transaction Amount

- **Annual income remains stable across purchase frequencies describes people have saving habits and spend thrifting habits almost equally and there can be a linear distribution.**
- Spending Score (blue line) remains relatively constant, indicating spending patterns are independent of purchase frequency.
- Transaction amounts (red line) show minor variations but generally align with increasing purchase frequency.



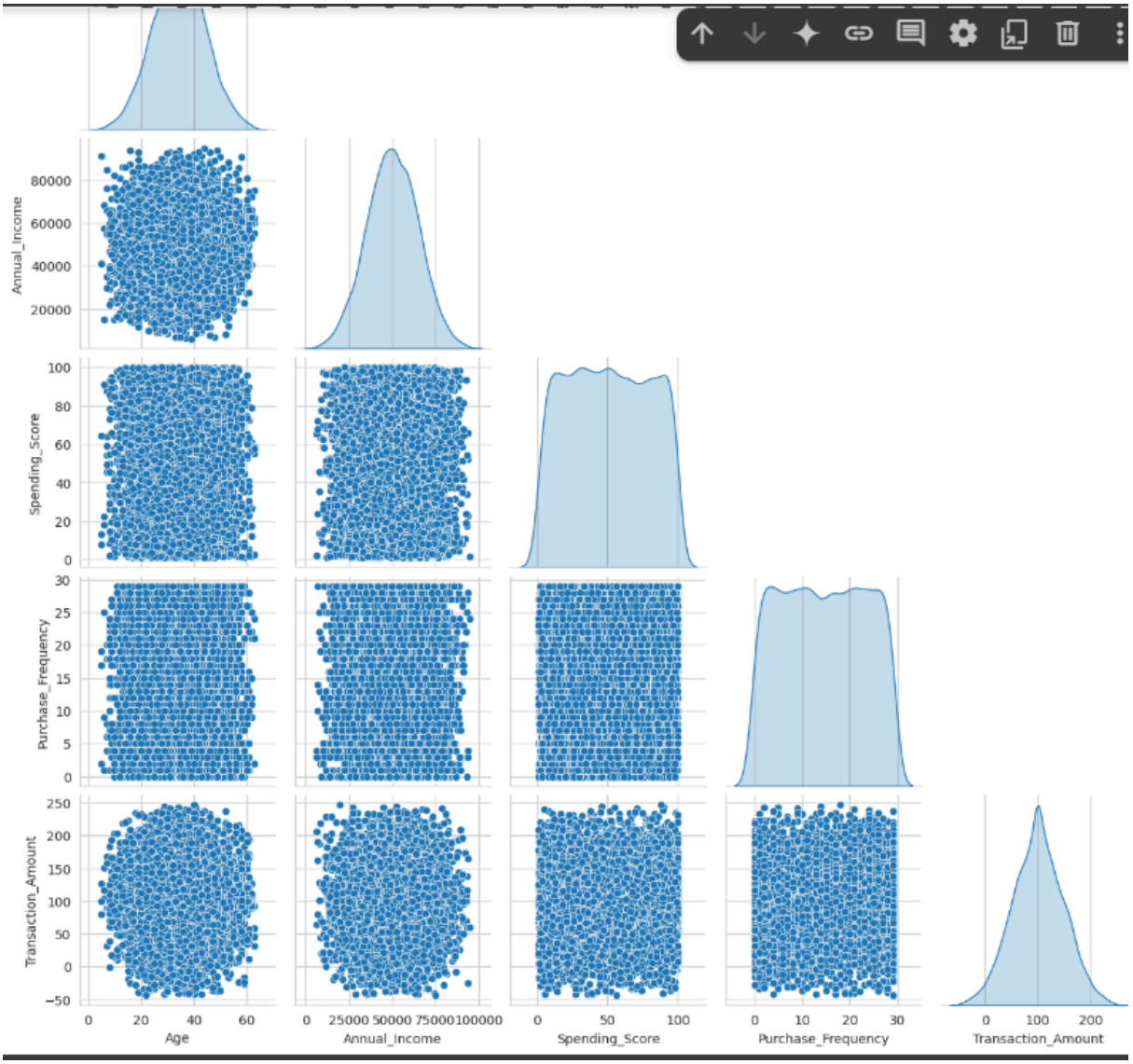
4) Age vs Spending Score

- There is no clear relationship between age and spending score.
- Higher spending scores are distributed across all age groups, suggesting spending behavior is independent of age.
- **Purchase Frequency (Color) and Transaction Amount (Size) show dispersed behavior,** indicating no strong clustering.



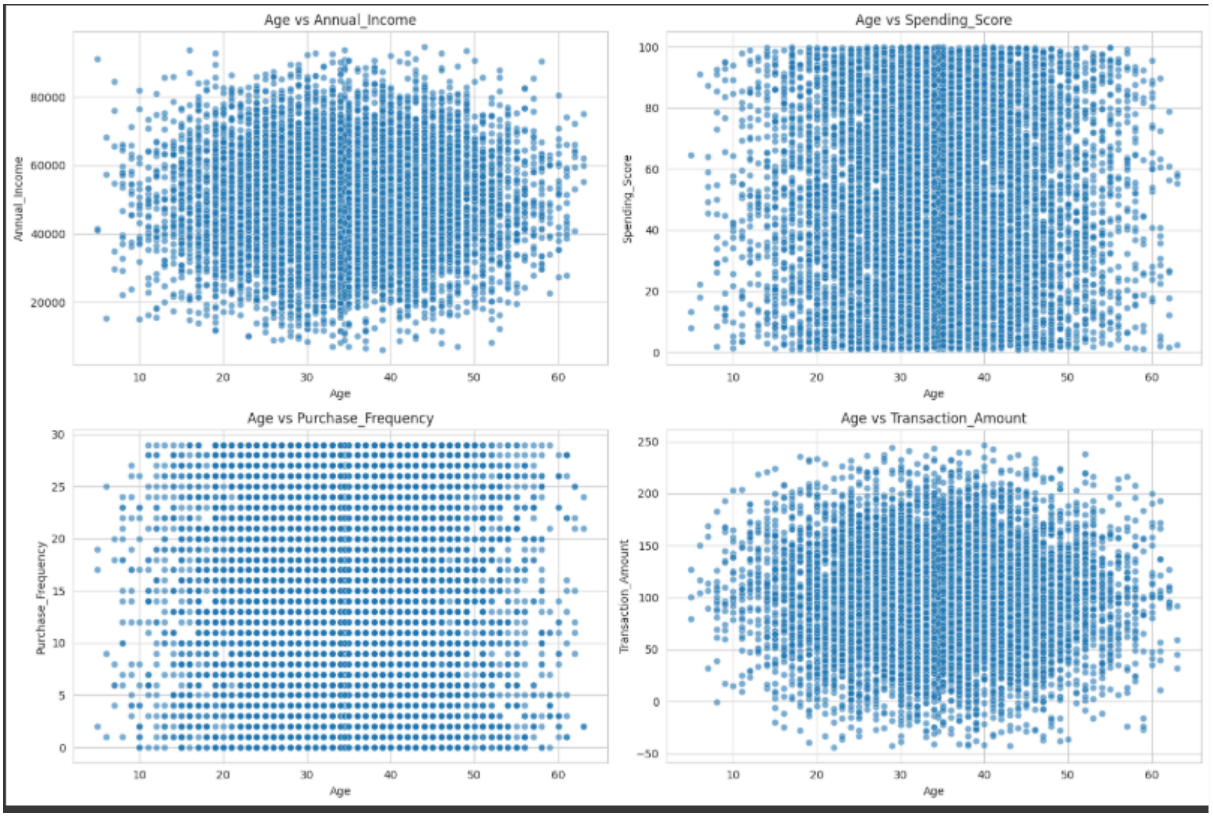
5) Pairplot Analysis

- Features show near-uniform distributions without strong patterns.
- Some variables like **Annual Income and Transaction Amount** exhibit minor skewness.



6) Age vs Various Metrics

- **Age does not significantly influence Spending Score**, as all age groups exhibit a diverse range of spending habits.
- **Young customers (10-30) show higher variability in spending**, with more extreme values in the higher range.
- **Older customers (40+) have a more uniform spending score.**



Conclusion

- **Annual Income** follows a normal distribution but with a slight right skew.
- **Purchase Frequency and Spending Score** appear to be uniformly distributed.
- **Spending Score and Purchase Frequency are independent variables, suggesting that frequent customers do not necessarily spend more.**
- **Young customers exhibit more varied spending habits**, which may require targeted marketing strategies.
- **High-income customers have unpredictable spending behaviors**, indicating luxury vs. necessity-based spending should be separately analyzed.
- **Extreme transaction amounts indicate potential high-value customers,** suggesting further segmentation could be useful.