Data Clustering Project Progress Report

Aumkar Ringe

Dharani Chandra

Meghana Sharath

Viraj Kenekar

Customer Segmentation

Customer segmentation involves analyzing various attributes of customers and organizing them into groups based on their behaviors, demographics, and specific traits. This process enables businesses to employ tailored communication strategies instead of generic approaches, ultimately leading to more effective outcomes for the business.

Demographic segmentation divides customers based on factors like age, gender, income, and location. Psychographic segmentation focuses on lifestyle, personality, and values. Behavioral segmentation looks at purchasing habits, brand loyalty, and interactions with products or services. Geographic segmentation considers where customers are located.

Firmographic segmentation is used in business-to-business (B2B) contexts, looking at attributes of the organization such as industry or company size. Technographic segmentation categorizes customers based on their use of technology, like preferred devices or software.

Once segmented, businesses can tailor their marketing strategies, products, and services to each group's specific needs and preferences. This targeted approach can improve customer satisfaction, loyalty, and ultimately, profitability. By understanding different segments, businesses can allocate resources more efficiently and create more personalized experiences for their customers.

B2C vs B2B customer segmentation

The customer segmentation process may work differently when marketing and selling to consumers (B2C), as opposed to other businesses (B2B), even though both approaches to segmentation involve taking the needs, behaviors, and characteristics of specific customer segments into consideration. Let's take a look at some examples:

- When segmenting customers based on their behaviors, a B2C marketer might look at consumers' browsing activity, their spending habits, or engagement with a brand, while a B2B marketer might look at how contacts within various organizations interact at sales meetings, trade shows, or across email.

- In terms of identity, B2C marketers might segment customers based on demographic details like income, family or relationship status, and age group, while B2B marketers might segment customers based on industry, company size, revenue, and the roles and teams within.

Business use cases:

1. **Targeted Marketing Campaigns**: By segmenting customers based on their preferences, behaviors, or demographics, businesses can create targeted marketing campaigns. For example, a clothing retailer might create different campaigns for segments such as "fashion-forward millennials" or "budget-conscious parents," ensuring that each group receives messages tailored to their interests and needs.

2. **Product Development**: Understanding the specific needs and preferences of different customer segments can inform product development strategies. For instance, a technology company might use segmentation data to identify which features are most important to different customer groups, guiding the development of new products or updates.

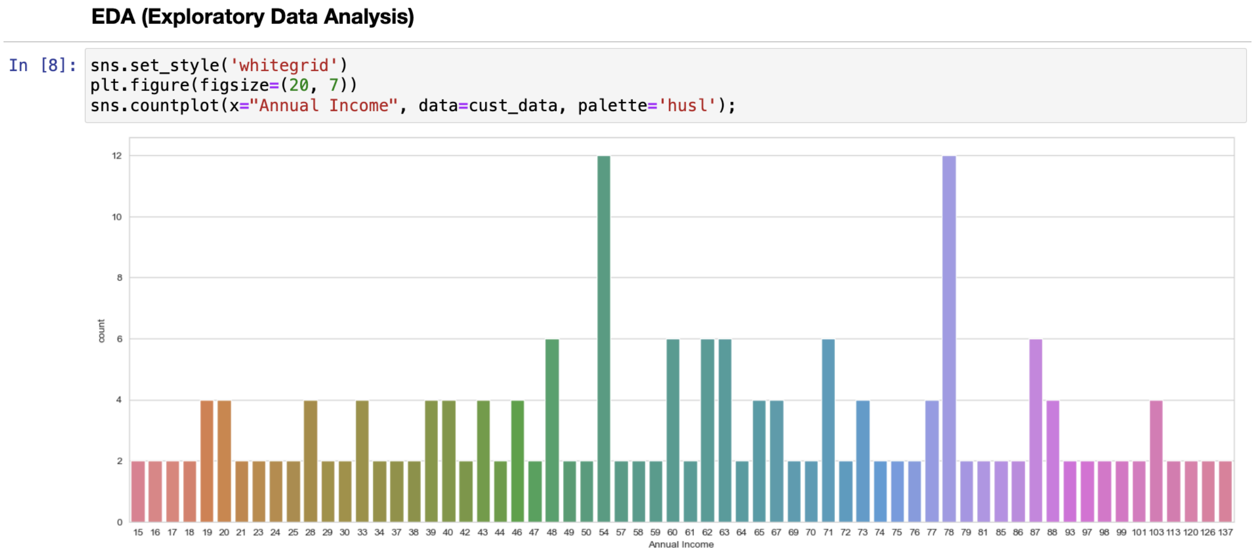
3. **Personalized Customer Service**: Customer segmentation enables businesses to provide personalized customer service experiences. For example, a hotel chain might segment customers based on their previous booking history or loyalty status, allowing them to offer tailored recommendations or perks to each segment.

4. **Pricing Strategies**: Segmentation can also inform pricing strategies, helping businesses optimize pricing for different customer segments. For instance, a software company might offer different pricing tiers based on the needs and budgets of different customer groups.

5. **Customer Retention**: Segmenting customers based on their engagement levels or likelihood to churn can help businesses proactively identify at-risk customers and implement targeted retention efforts. For example, a subscription-based service might offer special promotions or discounts to customers who are showing signs of disengagement.

6. **Market Expansion**: Customer segmentation can identify new market opportunities by uncovering underserved or overlooked customer segments. For example, a food delivery service might identify a segment of health-conscious consumers who are currently underserved in their market and develop a targeted marketing campaign to attract them.

PROGRESS REPORT:



EDA (Exploratory Data Analysis)

```
In [8]: sns.set_style('whitegrid')
        plt.figure(figsize=(20, 7))
        sns.countplot(x="Annual Income", data=cust_data, palette='husl');
```

To visualize the distribution of annual income among customers. This helps identify the most common income ranges and detect any patterns or anomalies in the data.

Analysis of the Output Plot

X-Axis (Annual Income):
Represents the annual income of customers in thousands of dollars.

Each unique value of annual income is displayed as a separate bar.

Y-Axis (Count):
Represents the number of customers for each annual income value.

Bar Heights:
Indicate how many customers have each specific annual income.
For example, the tallest bar (with a height of 12) indicates that there are 12 customers with an annual income of $54,000.

Usefulness for Your Project

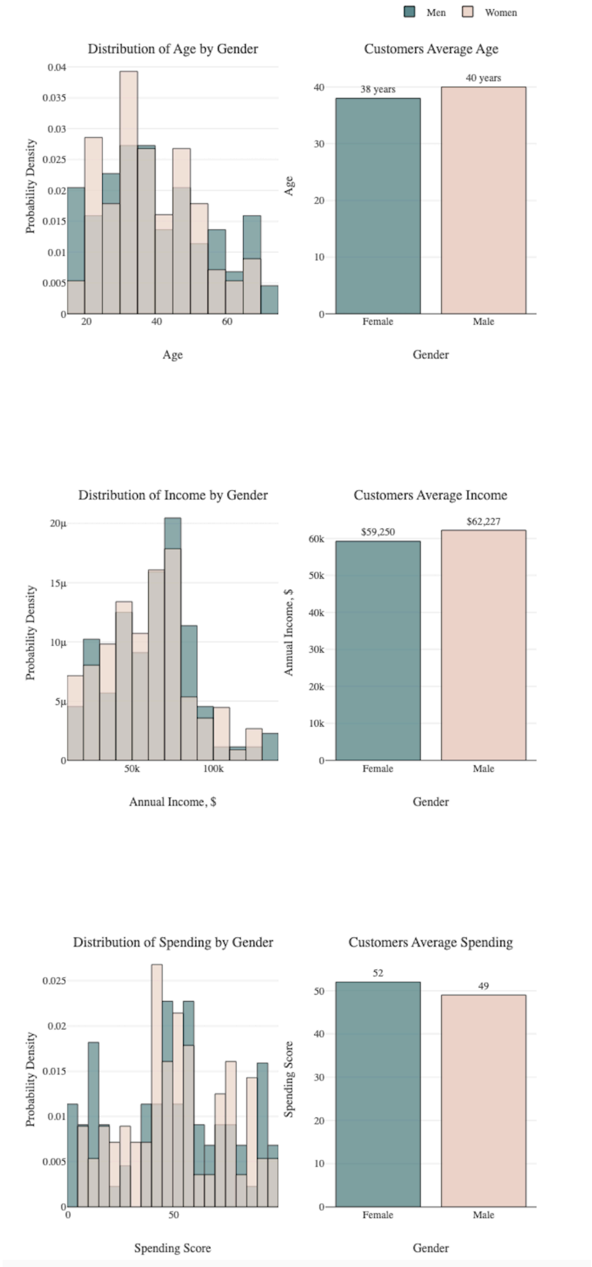
Understanding Customer Income Distribution:

By visualizing the distribution of annual income, you can quickly identify income ranges with higher concentrations of customers. This can help in understanding the financial demographics of the customer base.
Identifying Potential Segments:

Clusters in the income distribution (e.g., many customers at certain income levels) can suggest natural segments within the customer base. These segments can be further analyzed using clustering algorithms.

Data Preparation for Clustering:

Before applying clustering algorithms, understanding the distribution of key features like annual income ensures that you appropriately scale or transform the data if necessary.
Marketing and Sales Strategies:



Age Distribution and Average Age

Left Chart: Distribution of Age by Gender

Type: Histogram

Description: This histogram shows the distribution of ages for male and female customers.

Colors:

Men: Greenish (Teal)

Women: Pinkish (Peach)

Interpretation:

The x-axis represents age ranges.

The y-axis represents the probability density, indicating how common each age range is within the dataset.
Both genders have higher probability densities in the 20-40 age range, but women have a slightly higher density in the early 20s.

Right Chart: Customers Average Age
Type: Bar Chart
Description: This bar chart displays the average age of male and female customers.
Text:
Female: 38 years
Male: 40 years
Interpretation:
On average, male customers are slightly older than female customers.
This quick comparison helps in understanding the general age demographic of your customers.

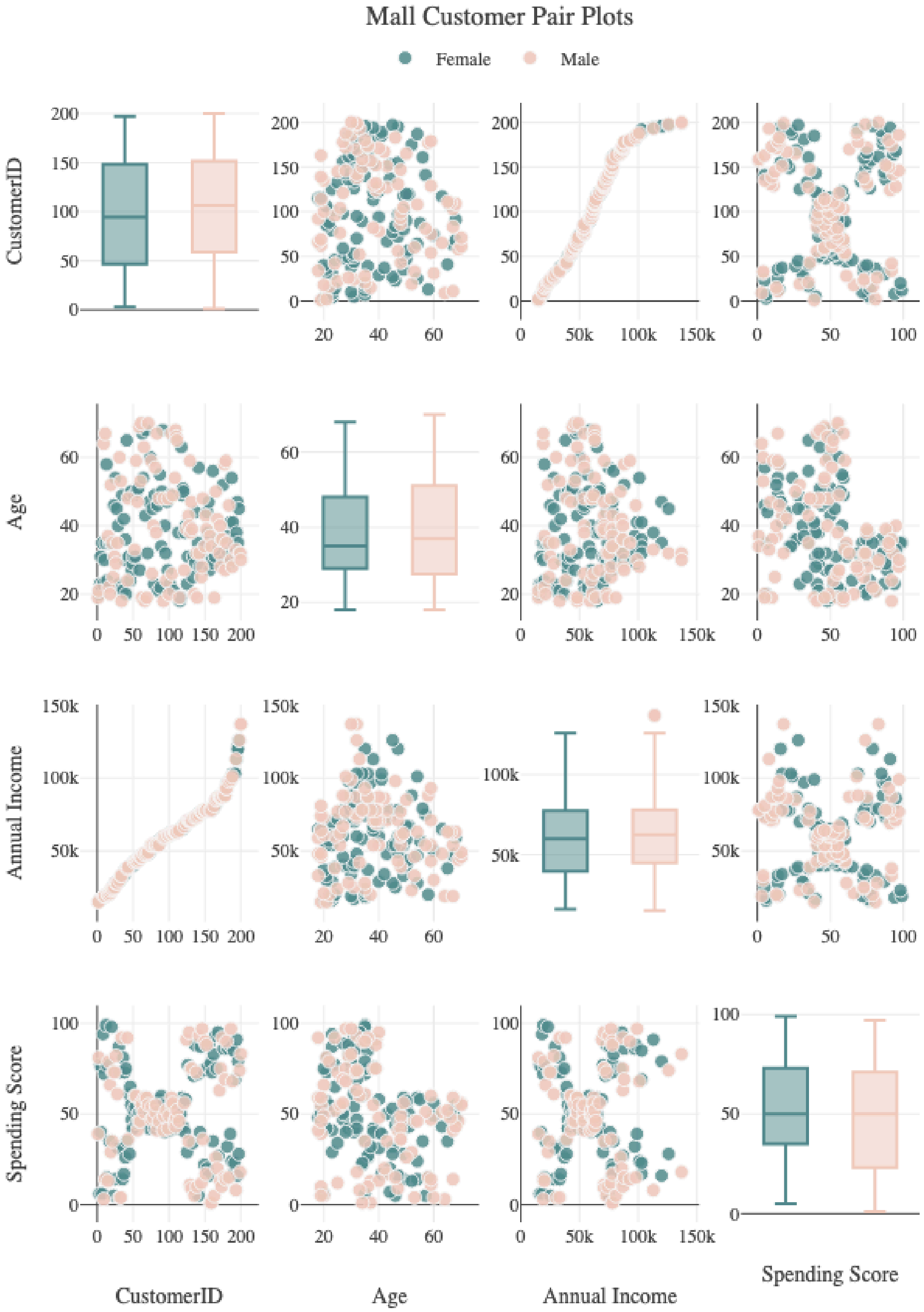
Middle Row: Income Distribution and Average Income
Left Chart: Distribution of Income by Gender
Type: Histogram
Description: This histogram shows the distribution of annual income for male and female customers.
Colors:
Men: Greenish (Teal)
Women: Pinkish (Peach)
Interpretation:
The x-axis represents annual income ranges (in dollars).
The y-axis represents the probability density, showing how common each income range is within the dataset.
Both genders have higher densities around the $50,000 to $75,000 income range, with men having a slightly broader distribution towards higher incomes.

Right Chart: Customers Average Income
Type: Bar Chart
Description: This bar chart shows the average annual income of male and female customers.
Text:
Female: $59,250
Male: $62,227
Interpretation:
On average, male customers have a slightly higher annual income than female customers.
This helps in understanding the financial demographics of your customers.

Bottom Row: Spending Score Distribution and Average Spending Score
Left Chart: Distribution of Spending by Gender
Type: Histogram

Description: This histogram shows the distribution of spending scores for male and female customers.
Colors:
Men: Greenish (Teal)
Women: Pinkish (Peach)
Interpretation:
The x-axis represents spending scores (a score that likely reflects customer spending behavior).
The y-axis represents the probability density, indicating how common each spending score is within the dataset.
Both genders have varied distributions, with noticeable peaks around the middle scores (20-50), suggesting most customers have moderate spending scores.

Right Chart: Customers Average Spending
Type: Bar Chart
Description: This bar chart displays the average spending score of male and female customers.
Text:
Female: 52
Male: 49
Interpretation:
On average, female customers have a slightly higher spending score than male customers.
This quick comparison helps in understanding the spending behavior of your customers.


Mall Customer Pair Plots
● Female
● Male
CustomerID
Age
Annual Income
Spending Score
0
50
100
150
200
20
40
60
0
50k
100k
150k
0
50
100
CustomerID
Age
Annual Income
Spending Score

Components of the Pair Plot
Diagonals (Histograms/Boxplots):

The diagonal subplots show the distribution of individual variables. Here, boxplots are used instead of histograms.
For example, the boxplot for "Age" shows the distribution of ages among male and female customers separately.

Upper and Lower Triangles (Scatter Plots):

The plots above and below the diagonals are scatter plots showing the relationships between pairs of variables.
Each point represents a customer, with colors differentiating males (light pink) and females (green).

Key Insights from the Pair Plot
CustomerID:

The CustomerID is a unique identifier and doesn't provide any specific relationship with other variables.
Boxplots for CustomerID show an even distribution for both genders, which is expected as CustomerID is sequentially assigned.

Age:

The age distribution shows that female customers tend to be slightly younger than male customers.
There is a wide age range (approximately 18 to 70) for both male and female customers.

Annual Income:

Annual Income boxplots reveal that male customers have a slightly higher median annual income compared to female customers.
The scatter plots between Annual Income and other variables show no strong correlation, indicating that Annual Income varies independently.

Spending Score:

The Spending Score distributions are quite varied, with a wider spread for male customers.
The scatter plots between Spending Score and other variables (Age, Annual Income) show clusters but no clear linear relationship, indicating more complex interactions.

Observations from Scatter Plots
Age vs. Annual Income:

There is no strong correlation between Age and Annual Income for both genders, suggesting income levels are independent of age within this customer base.

Age vs. Spending Score:

The scatter plot shows no clear pattern, suggesting Spending Score is not directly related to Age.

Annual Income vs. Spending Score:

There are no obvious patterns or clusters that indicate a direct relationship between Annual Income and Spending Score.

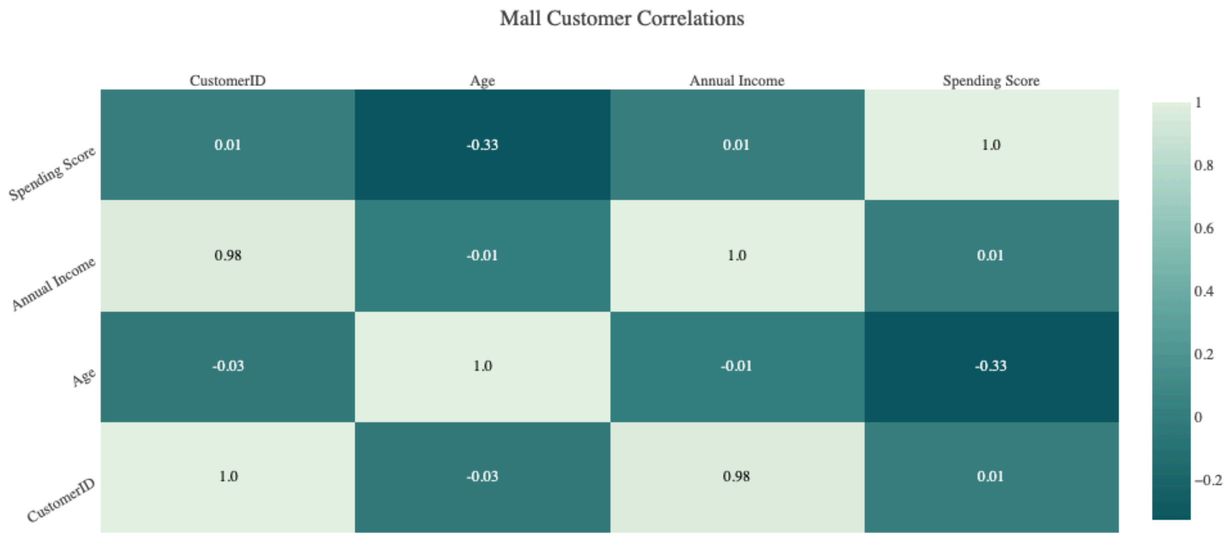
Overall Analysis

The pair plot helps visualize the spread and relationships of different variables in the customer dataset.

Male and female customers exhibit slightly different distributions for Age, Annual Income, and Spending Score.

No single variable appears to strongly dictate another, suggesting that multiple factors might influence spending behavior.

This visual analysis can be the foundation for further statistical analysis or machine learning, such as clustering or segmentation.



This plot is a heatmap showing the correlation matrix for the mall customer data. Correlation values range from -1 to 1, where 1 means a perfect positive correlation, -1 means a perfect negative correlation, and 0 means no correlation. The color intensity indicates the strength of the correlation, with darker shades representing stronger correlations.

Understanding the Heatmap
Diagonal Values:

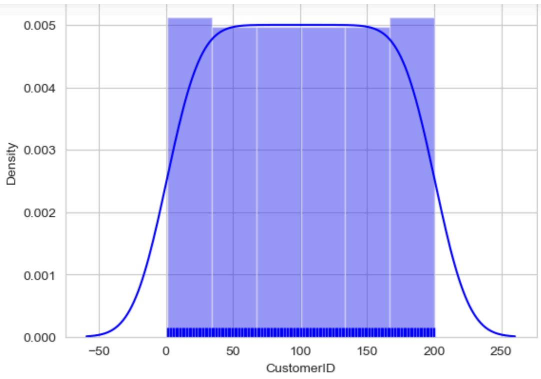
The diagonal values are all 1 because each variable is perfectly correlated with itself.
Correlation Between Different Variables:

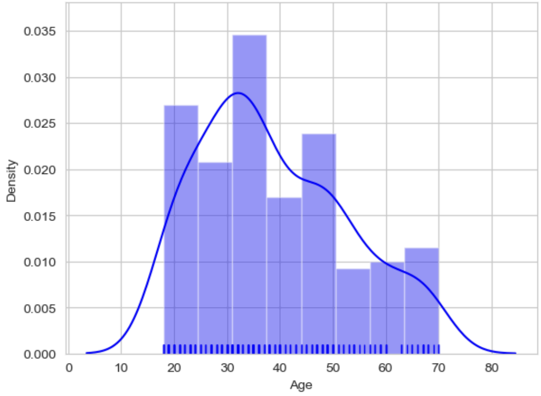
CustomerID and Age: The correlation is -0.03, indicating a very weak negative relationship.
CustomerID and Annual Income: The correlation is 0.98, showing a very strong positive correlation. This suggests that as CustomerID increases (which is sequential), Annual Income also increases. This is likely an artifact of how the data was generated or ordered.
CustomerID and Spending Score: The correlation is 0.01, indicating virtually no relationship.
Age and Annual Income: The correlation is -0.01, indicating no meaningful relationship.
Age and Spending Score: The correlation is -0.33, indicating a moderate negative relationship. This suggests that older customers tend to have lower Spending Scores.
Annual Income and Spending Score: The correlation is 0.01, indicating no meaningful relationship.
Key Insights
Strong Positive Correlation:

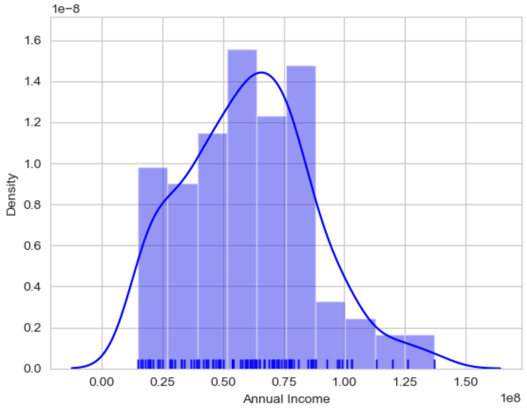
The strongest positive correlation is between CustomerID and Annual Income (0.98). This strong correlation likely results from how the data is structured rather than an inherent relationship between the two variables.
Moderate Negative Correlation:

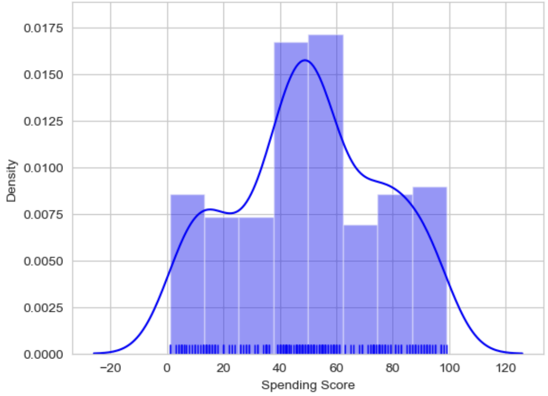
There is a moderate negative correlation between Age and Spending Score (-0.33), suggesting that older customers tend to spend less, as reflected by lower Spending Scores.
Weak or No Correlation:

Most other pairs of variables have weak or no significant correlation. For example, Age and Annual Income, Annual Income and Spending Score, and CustomerID and Spending Score all have correlations close to zero, indicating little to no linear relationship.
Implications
The moderate negative correlation between Age and Spending Score can inform marketing strategies. For instance, targeted campaigns might be designed to increase spending among older customers.
The strong correlation between CustomerID and Annual Income suggests a possible data ordering issue. It is essential to investigate this further to ensure data integrity.


Density
0.005
0.004
0.003
0.002
0.001
0.000
−50
0
50
100
150
200
250
CustomerID


Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
10
20
30
40
50
60
70
80
Age


1e−8
Density
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
Annual Income
1e8


Density
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−20
0
20
40
60
80
100
120
Spending Score

Age Distribution:

KDE peak around 30-35: Most customers are in their early 30s.
Wide spread from 20 to 60: Customers range from young adults to older individuals.
Slight skew towards younger ages: More younger customers than older ones.

Annual Income Distribution:

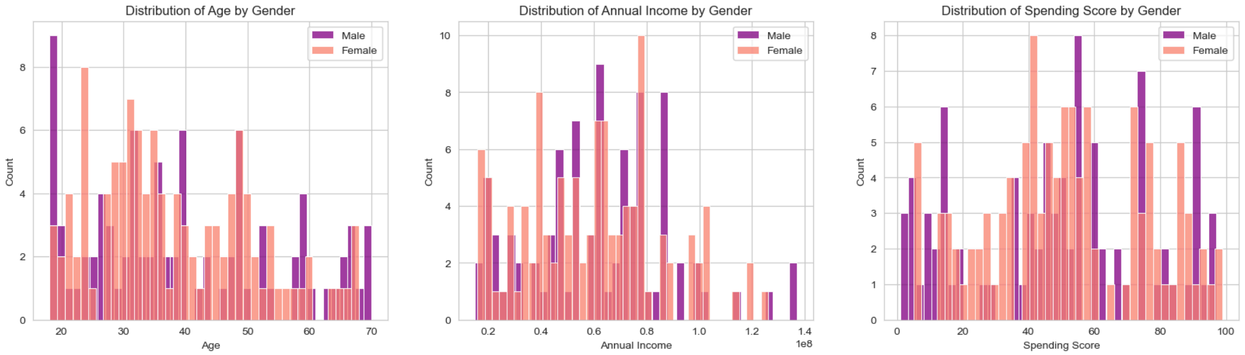
KDE peak around $50,000 to $75,000: Most customers have an annual income in this range.
Long tail to the right: Some customers have very high incomes, but they are less common.
Rug plot shows clusters: Individual incomes might show clustering around certain income levels, indicating income brackets.

Spending Score Distribution:

KDE peak around 40-60: Most customers have moderate spending scores.
Some density around 20-30 and 70-80: Indicates smaller groups of low and high spenders.
Symmetrical shape: Spending scores are relatively evenly distributed without strong skewness.’

Summary
By examining these plots, you can gain insights into the demographic and financial characteristics of your customer base, such as:

The typical age and income ranges of your customers.
Variability in spending behavior.
Presence of distinct customer segments (e.g., high vs. low spenders).



1. Distribution of Age by Gender: This subplot shows the distribution of ages among male and female customers. It helps visualize the age demographics of the customer base and identify any differences or similarities between male and female customers in terms of age.
2. Distribution of Annual Income by Gender: This subplot illustrates the distribution of annual incomes among male and female customers. It provides insights into the income distribution within each gender group and allows for comparison between male and female customers' purchasing power.
3. Distribution of Spending Score by Gender: This subplot displays the distribution of spending scores among male and female customers. It indicates how male and female customers spend across different spending score ranges, highlighting any variations in purchasing behavior between the two groups.

K-Means

K-means clustering is a robust unsupervised machine learning technique used for organizing unlabeled datasets into distinct groups. It aims to partition data into clusters, ensuring that similar data points are grouped together. The algorithm starts by initializing cluster centroids and then repeatedly assigns data points to the closest centroid. After each assignment, it updates the centroids to the mean position of the points in each cluster.

DBSCAN Clustering