LOVELY PROFESSIONAL UNIVERSITY

Cipher SCHOOLS

SUMMER INTERSHIP TRAINING

- **Project Title:** Customer Behaviour Analytics for Retail Stores
- **Project Manager:** Yashwanth Booram
- **Start Date:** 13 July 2024
- **End Date:** 16 July 2024
- **Objectives:** To divide consumers into various categories according on how they make purchases.
- **Scope:** Data cleaning, EDA, RFM(Recency, Frequency, Monetary) Analysis, Customer Segmentation using K-Means Clustering, Visualization using Matplotlib and Power BI.
- **Deliverables:** Conclusions, insights, and suggestions

1. Introduction

Project Objective:

The primary objective of this project is to segment retail customers based on their purchasing behavior. By understanding these segments, the retail store can develop targeted marketing strategies to enhance customer engagement and increase sales.

Scope:

The project involves several key steps:

- Data cleaning to ensure data quality.

- Exploratory Data Analysis (EDA) to understand the data distribution and relationships.

- RFM (Recency, Frequency, Monetary) analysis to evaluate customer value.

- K-means clustering to identify distinct customer groups.

- Visualization of the results using Matplotlib and Power BI.

2. Data Overview

Dataset Description:

The dataset used for this project is the Mall_Customers dataset, which includes the following features:

- **`CustomerID`**: Unique identifier for each customer.

- **`Gender`:** Gender of the customer.

- **`Age`:** Age of the customer.

- **`Annual Income (k$)`:** Annual income of the customer in thousand dollars.

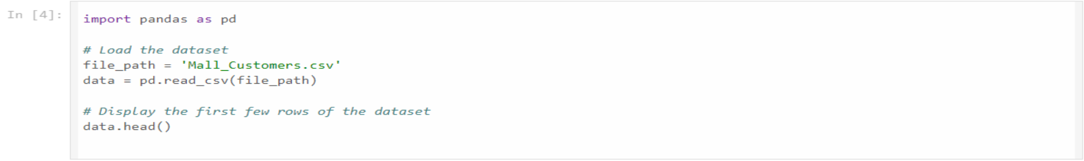
- **`Spending Score (1-100)`:** A score assigned by the mall based on customer behavior and spending nature.

Data Cleaning:

Missing Values: Checked for missing values and found none, ensuring a complete dataset.

Duplicates: Identified and removed duplicate records to maintain data integrity.

Data Types: Ensured all columns have appropriate data types for analysis.



```
import pandas as pd

# Load the dataset
file_path = 'Mall_Customers.csv'
data = pd.read_csv(file_path)

# Display the first few rows of the dataset
data.head()
```

3. Exploratory Data Analysis (EDA)

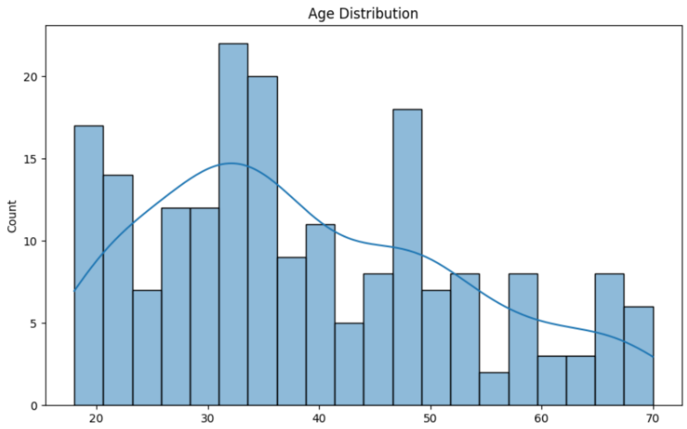
Summary Statistics:

The dataset contains 200 records with no missing values. Below are key summary statistics for numerical features:

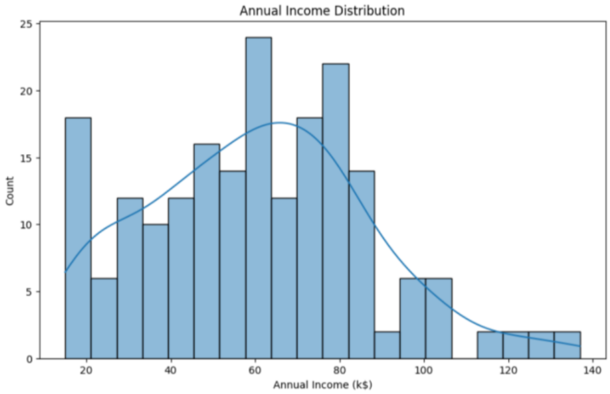
Feature	Mean	Median	Std Dev	Min	Max
Age	38.85	36	13.97	18	70
Annual Income (k$)	60.56	61.5	26.26	15	137
Spending Score (1-100)	50.20	50	25.82	1	99

Visualizations:

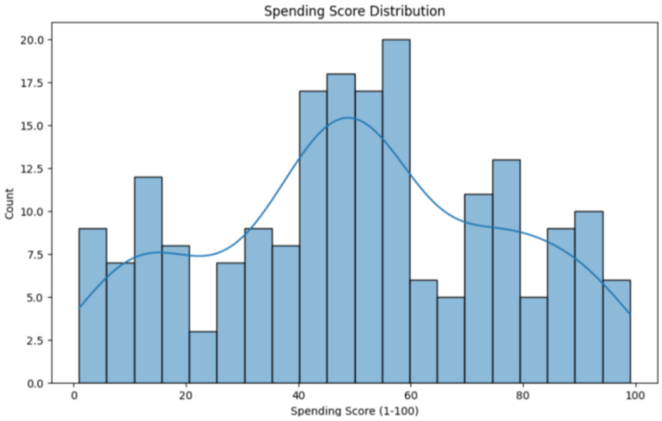
- Age Distribution:



- Annual Income Distribution:



- Spending Score Distribution:



4. RFM Analysis

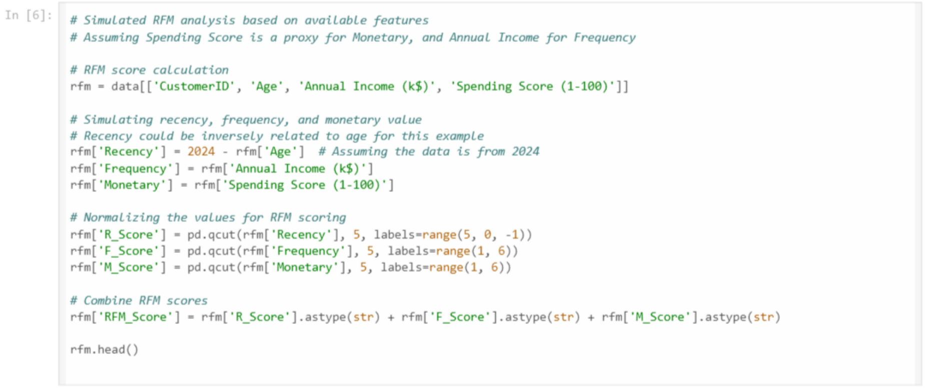
RFM Model:

The RFM model is used to evaluate customers based on three metrics:

- **Recency (R):** How recently a customer made a purchase.
- **Frequency (F):** How often a customer makes a purchase.
- **Monetary (M):** How much money a customer spends.

RFM Scoring:

For this dataset, we simulate RFM analysis using the available features:



```
In [6]:
# Simulated RFM analysis based on available features
# Assuming Spending Score is a proxy for Monetary, and Annual Income for Frequency

# RFM score calculation
rfm = data[['CustomerID', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)']]

# Simulating recency, frequency, and monetary value
# Recency could be inversely related to age for this example
rfm['Recency'] = 2024 - rfm['Age']  # Assuming the data is from 2024
rfm['Frequency'] = rfm['Annual Income (k$)']
rfm['Monetary'] = rfm['Spending Score (1-100)']

# Normalizing the values for RFM scoring
rfm['R_Score'] = pd.qcut(rfm['Recency'], 5, labels=range(5, 0, -1))
rfm['F_Score'] = pd.qcut(rfm['Frequency'], 5, labels=range(1, 6))
rfm['M_Score'] = pd.qcut(rfm['Monetary'], 5, labels=range(1, 6))

# Combine RFM scores
rfm['RFM_Score'] = rfm['R_Score'].astype(str) + rfm['F_Score'].astype(str) + rfm['M_Score'].astype(str)

rfm.head()
```

RFM Segments:

The RFM analysis segments customers into different groups based on their RFM scores, such as:

- **Champions:** High recency, high frequency, and high monetary value.
- **Loyal Customers:** High frequency and monetary value, not necessarily recent.
- **At Risk:** High monetary value but not recent.
- **Hibernating:** Low recency, low frequency, and low monetary value.

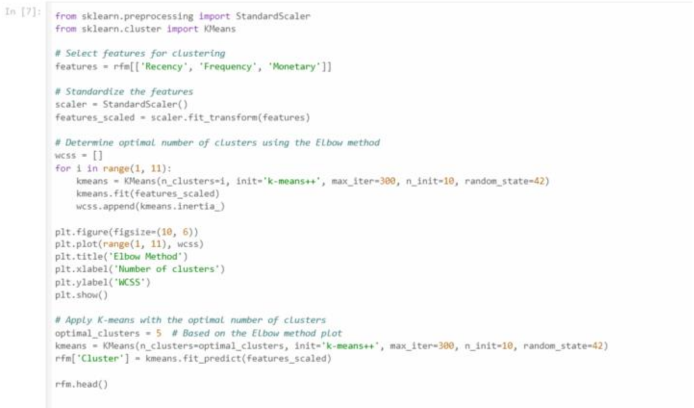
5. K-means Clustering

Feature Selection:

Selected features for clustering: **`Recency`, `Frequency`, and `Monetary`.**

Normalization:

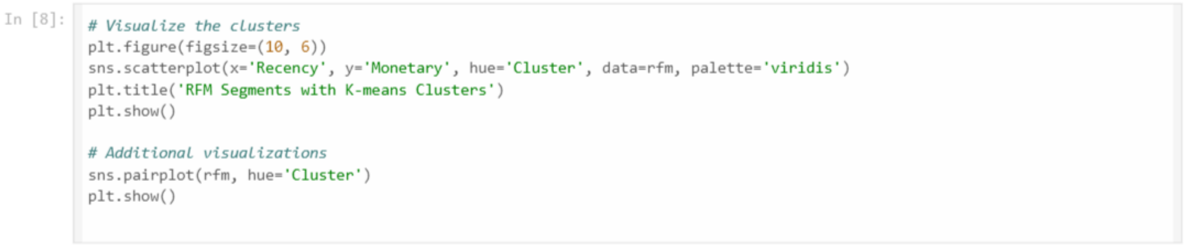
Standardized the features to ensure they are on a similar scale.



```
In [7]:
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

# Select features for clustering
features = rfm[['Recency', 'Frequency', 'Monetary']]

# Standardize the features
scaler = StandardScaler()
features_scaled = scaler.fit_transform(features)

# Determine optimal number of clusters using the Elbow method
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=42)
    kmeans.fit(features_scaled)
    wcss.append(kmeans.inertia_)

plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), wcss)
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()

# Apply K-means with the optimal number of clusters
optimal_clusters = 5  # Based on the Elbow method plot
kmeans = KMeans(n_clusters=optimal_clusters, init='k-means++', max_iter=300, n_init=10, random_state=42)
rfm['Cluster'] = kmeans.fit_predict(features_scaled)

rfm.head()
```

Optimal Clusters:

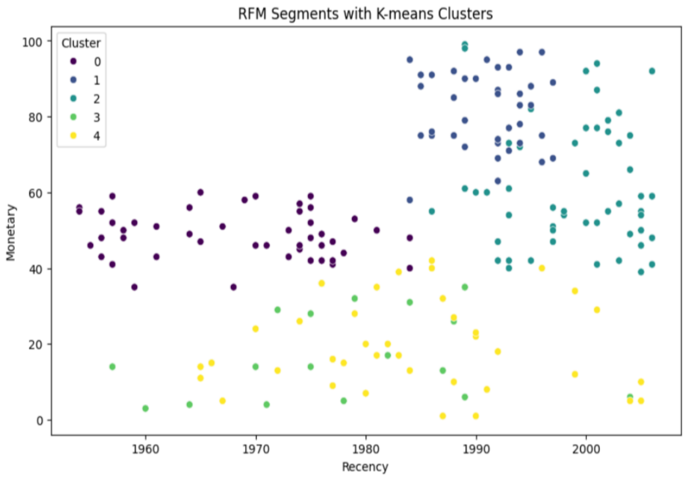
Determined the optimal number of clusters using the Elbow method:



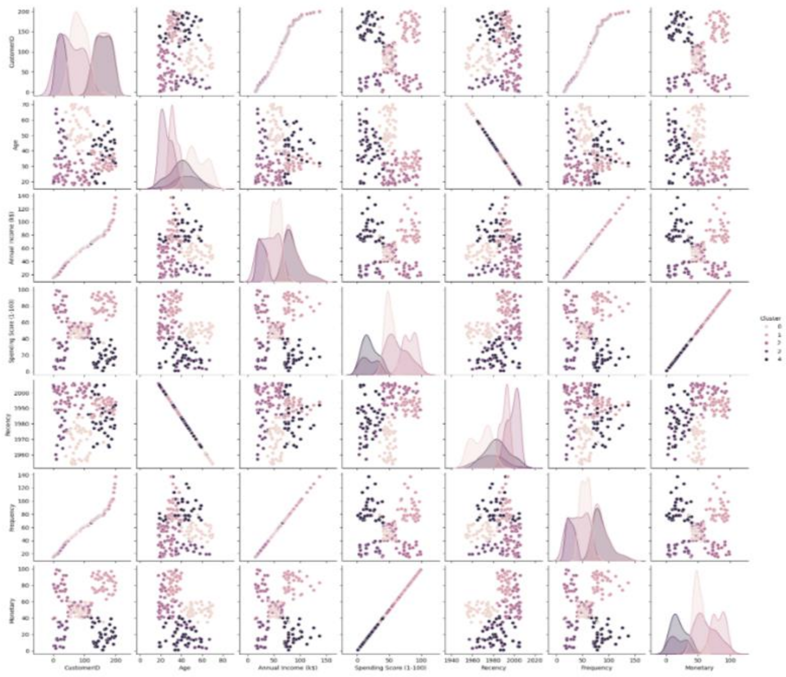
```
In [8]:  # Visualize the clusters
         plt.figure(figsize=(10, 6))
         sns.scatterplot(x='Recency', y='Monetary', hue='Cluster', data=rfm, palette='viridis')
         plt.title('RFM Segments with K-means Clusters')
         plt.show()

         # Additional visualizations
         sns.pairplot(rfm, hue='Cluster')
         plt.show()
```

Clustering Results:

Applied K-means clustering with the optimal number of clusters and visualized the results:



6. Visualization



Power BI Dashboards:

- Imported the RFM data with cluster assignments into Power BI.

- Created interactive dashboards to explore customer segments.

- Used various visualizations like bar charts, scatter plots, and heat maps to present the RFM segments and clusters.

7. Insights and Recommendations

Customer Segments:

Summarized the key characteristics of each customer segment identified through RFM analysis and K-means clustering:

- **Champions:** Recent, frequent, and high spenders.
- **Loyal Customers:** Frequent and high spenders.
- **At Risk:** High spenders but haven't purchased recently.
- **Hibernating:** Infrequent and low spenders.

Marketing Strategies:

Provided recommendations for targeted marketing strategies based on the characteristics of each segment:

- **Champions:** Exclusive offers and loyalty programs.
- **Loyal Customers:** Personalized recommendations and rewards.
- **At Risk:** Re-engagement campaigns and special discounts.
- **Hibernating:** General promotions and awareness campaigns.

Business Impact:

Discussed the potential impact of the findings on the retail store's business strategy:

- Improved customer retention and satisfaction.

- Increased sales through targeted marketing.

- Efficient allocation of marketing resources.

8. Conclusion

Summary:

Recapped the main findings and the overall process:

- Performed data cleaning, EDA, RFM analysis, and K-means clustering.
- Identified distinct customer segments and provided actionable insights.

Future Work:

Suggested areas for further research or additional analyses:

- Incorporate additional features such as online behaviour data.
- Explore other clustering techniques like hierarchical clustering.
- Continuously monitor and update customer segments