

Industry Standard Documentation

Final Report:

Mall Customer Dataset Analysis and Visualization

- **Executive Summary**

This project focuses on analyzing the Mall Customers dataset to gain insights into customer behavior. The primary objectives were to preprocess the data, perform exploratory data analysis (EDA), segment customers using clustering techniques, and visualize the results. Key outcomes include the identification of distinct customer segments and actionable insights for targeted marketing strategies.

- **Introduction**

Project Background

Understanding customer behavior is crucial for any business aiming to enhance its marketing efforts and customer service. The Mall Customers dataset provides an opportunity to analyze various attributes of customers, such as gender, age, annual income, and spending score.

Objectives

- Preprocess the data to ensure quality and consistency.
- Perform exploratory data analysis to understand the dataset.
- Segment customers using clustering techniques.
- Visualize the results to derive actionable insights.

- **Methodology**

1. Data Preprocessing

- Import Data: Load the Mall Customers dataset.
- Clean Data: Handle missing values, correct data types, and normalize numerical features.

2. Exploratory Data Analysis (EDA)

- Statistical Summary: Generate summary statistics for all features.
- Visualizations: Use histograms, box plots, and scatter plots to understand data distributions and relationships between features.

3. Clustering

- Algorithm Selection: Use K-Means clustering.
- Optimal Clusters: Determine the optimal number of clusters using the elbow method and silhouette score.
- Cluster Interpretation: Analyze the characteristics of each cluster.

4. Visualization

- Static Visualizations: Create 2D and 3D plots using Matplotlib and Seaborn.
- Interactive Dashboards: Develop interactive visualizations using Power BI.

Results

Findings from Data Analysis

- Age Distribution: Customers' ages range from 18 to 70, with a mean age of 38.85.

- Income Distribution: Annual incomes vary between \$15,000 and \$137,000, with a mean income of \$60,560.
- Spending Score: Spending scores range from 1 to 99, with a mean score of 50.2

Customer Segmentation

Using K-Means clustering, customers were segmented into five distinct clusters:

- Cluster 1: Young, high-income individuals with high spending scores.
- Cluster 2: Older, high-income individuals with moderate spending scores.
- Cluster 3: Young, low-income individuals with low spending scores.
- Cluster 4: Middle-aged individuals with moderate incomes and high spending scores.
- Cluster 5: Older, low-income individuals with low spending scores.

Conclusion

- High-Spending Young Customers: Cluster 1 represents young customers with high incomes and spending scores, indicating a potential target for premium products and services.
- Moderate-Spending Older Customers: Cluster 2 includes older customers with high incomes but moderate spending scores, suggesting a focus on loyalty programs and personalized offers.
- Low-Spending Young Customers: Cluster 3 highlights young customers with low incomes and spending scores, suggesting a need for budget-friendly options and promotions.
- High-Spending Middle-Aged Customers: Cluster 4 includes middle-aged customers with moderate incomes but high spending scores, indicating a potential for mid-range products and loyalty programs.

- Low-Spending Older Customers: Cluster 5 represents older customers with low incomes and spending scores, suggesting a focus on essential products and value-for-money promotions.

Recommendations

- Targeted Marketing: Develop marketing strategies tailored to each customer segment.
- Product Recommendations: Personalize product recommendations based on customer segment characteristics.
- Promotional Campaigns: Design promotional campaigns to attract low-spending young customers and retain high-spending individuals.

Appendices

- **Additional Charts**

1. Age Distribution Histogram
2. Income Distribution Box Plot
3. Spending Score Scatter Plot

- **Code Snippets**

```
# Data Import and Preprocessing
import pandas as pd

# Load dataset
data = pd.read_csv('/mnt/data/Mall_Customers.csv')

# Data cleaning
data.dropna(inplace=True)
data['Annual Income (k$)'] = data['Annual Income (k$)'] * 1000

# EDA
import seaborn as sns
import matplotlib.pyplot as plt

# Age distribution
sns.histplot(data['Age'], bins=30)
plt.show()

# Clustering
from sklearn.cluster import KMeans

# Features for clustering
features = data[['Age', 'Annual Income (k$)', 'Spending Score (1-100)']]

# K-Means clustering
kmeans = KMeans(n_clusters=5)
data['Cluster'] = kmeans.fit_predict(features)
|
```