Dataset overview

Generated on: 2025-11-17 12:55:29 IST

Task type: classification

Target column: Gender

Rows: 200

Columns: 5

Numeric columns: 3

Categorical columns: 0

Overall missing: 0.00%

Model Performance (test split):

Accuracy: 0.5500

ROC AUC: 0.6705

Top Influencing Features:

Spending Score (1-100): 0.3430

Age: 0.3360

Annual Income (k$): 0.3209

ID-like columns removed: 1

Top columns by missing values

CustomerID: 0 (0.0%)

Gender: 0 (0.0%)

Age: 0 (0.0%)

Annual Income (k$): 0 (0.0%)

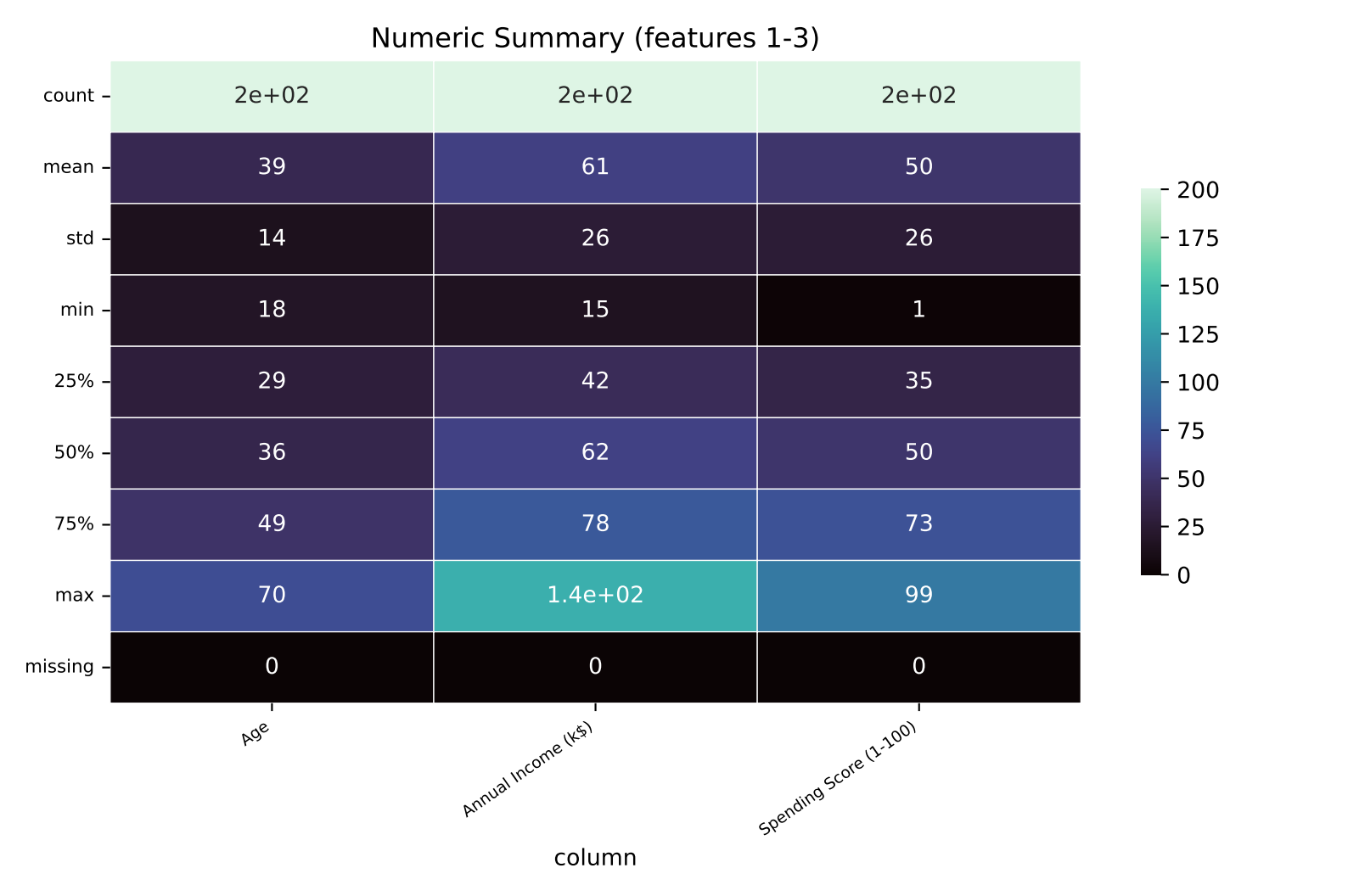
Spending Score (1-100): 0 (0.0%)

Top 5 Columns by Missing Values

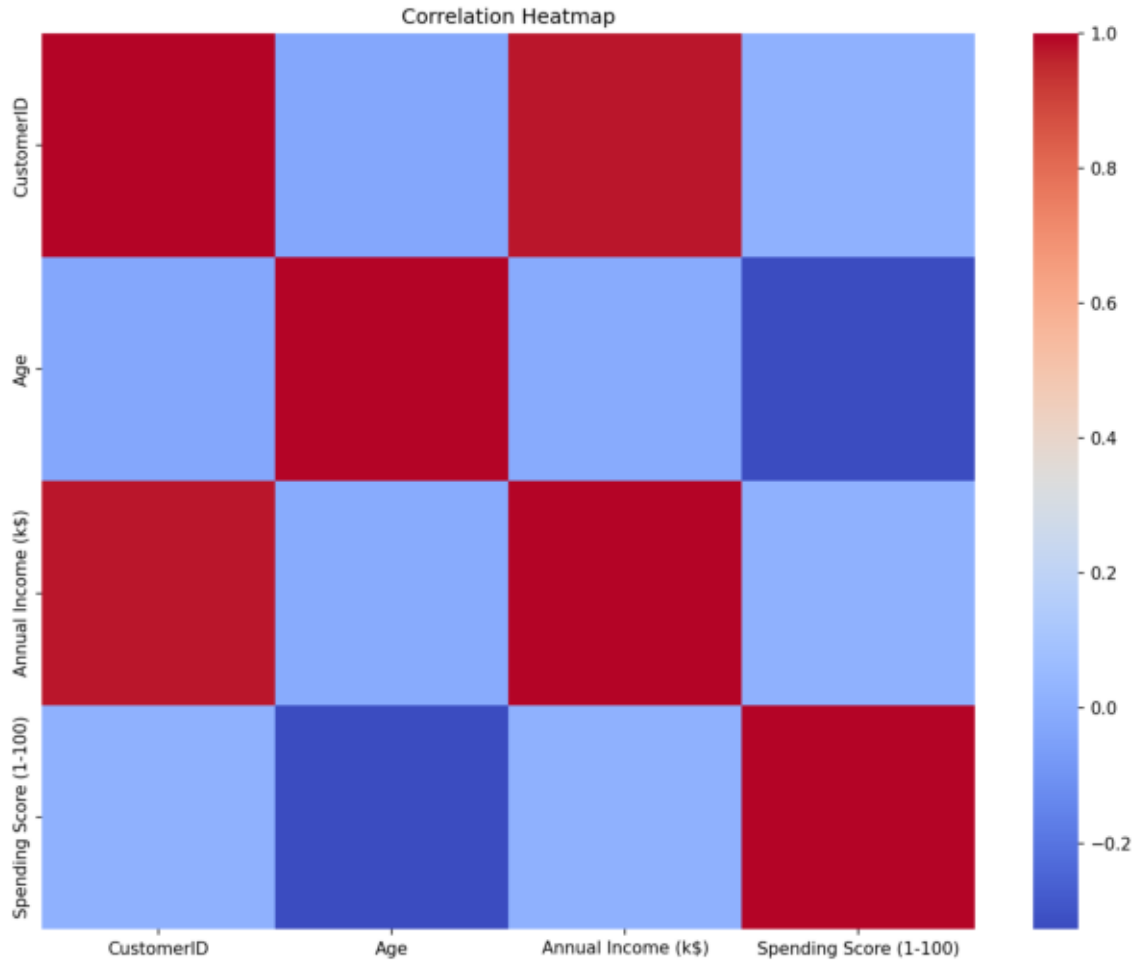
column	dtype	unique	missing_count	missing_percent
CustomerID	int64	200	0	0.0
Gender	object	2	0	0.0
Age	int64	51	0	0.0
Annual Income (k$)	int64	64	0	0.0
Spending Score (1-100)	int64	84	0	0.0

Numeric Columns

Age | Annual Income (k$) | Spending Score (1-100)


Numeric Summary (features 1-3)
count
mean
std
min
25%
50%
75%
max
missing
2e+02
2e+02
2e+02
39
61
50
14
26
26
18
15
1
29
42
35
36
62
50
49
78
73
70
1.4e+02
99
0
0
0
Age
Annual Income (k$)
Spending Score (1-100)
column
200
175
150
125
100
75
50
25
0

Correlation Heatmap



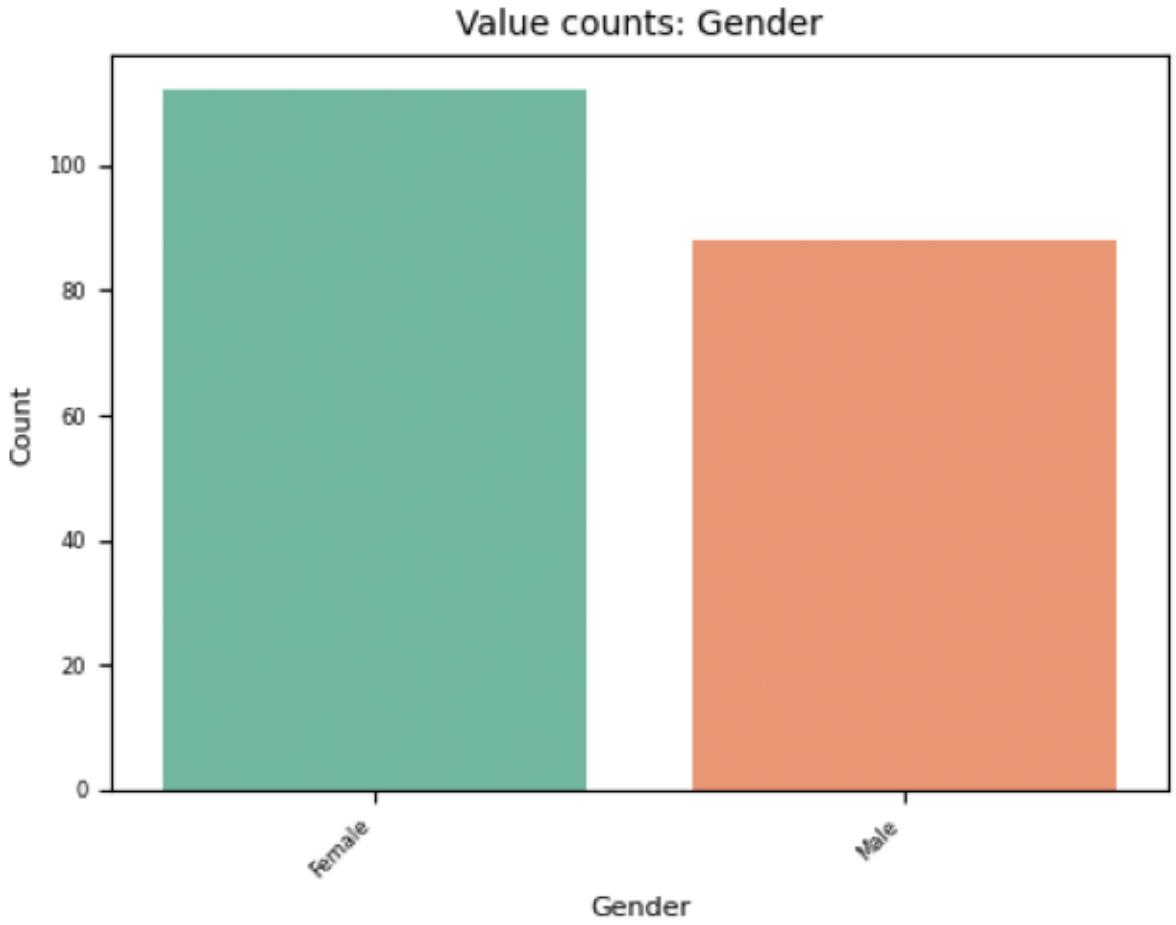
Categorical Columns

Gender

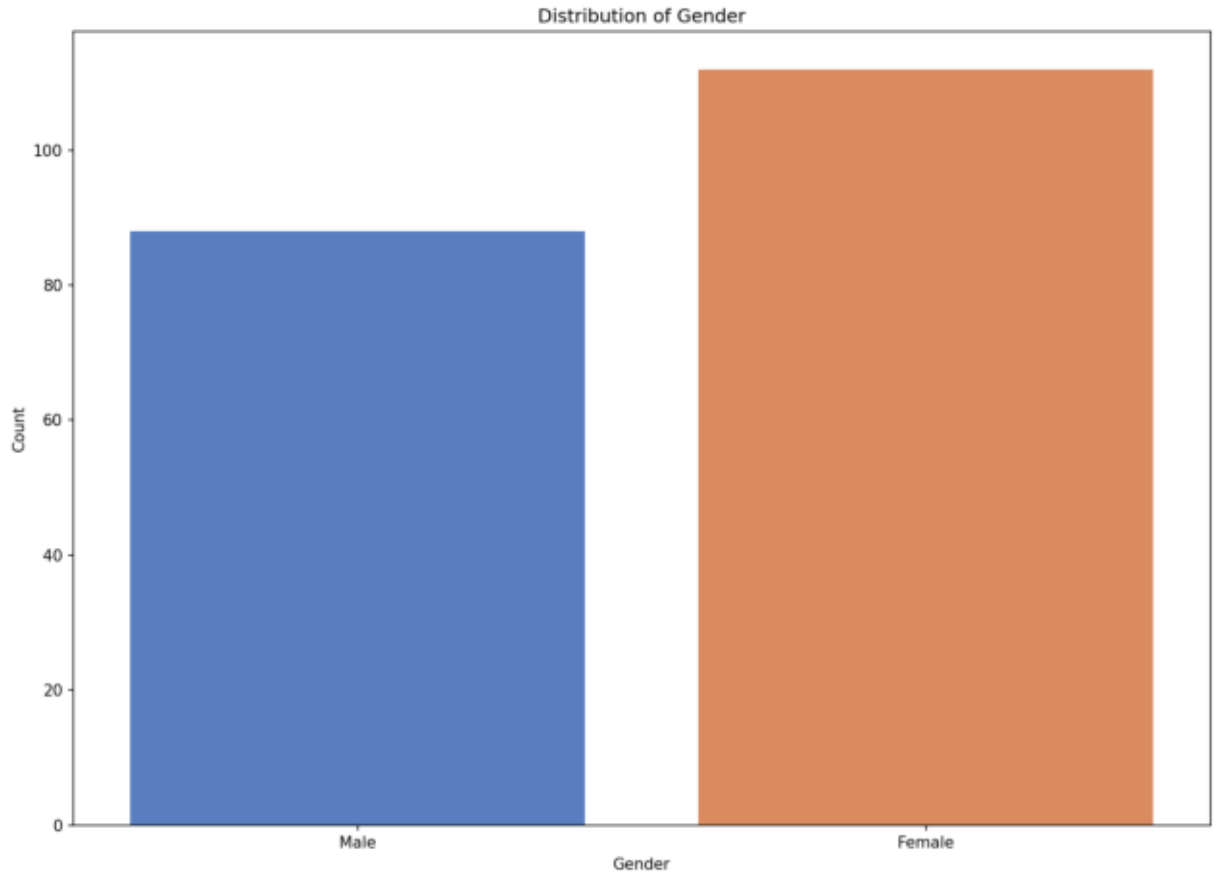
Categorical Columns Summary

column	count	unique	top	freq
Gender	200	2	Female	112

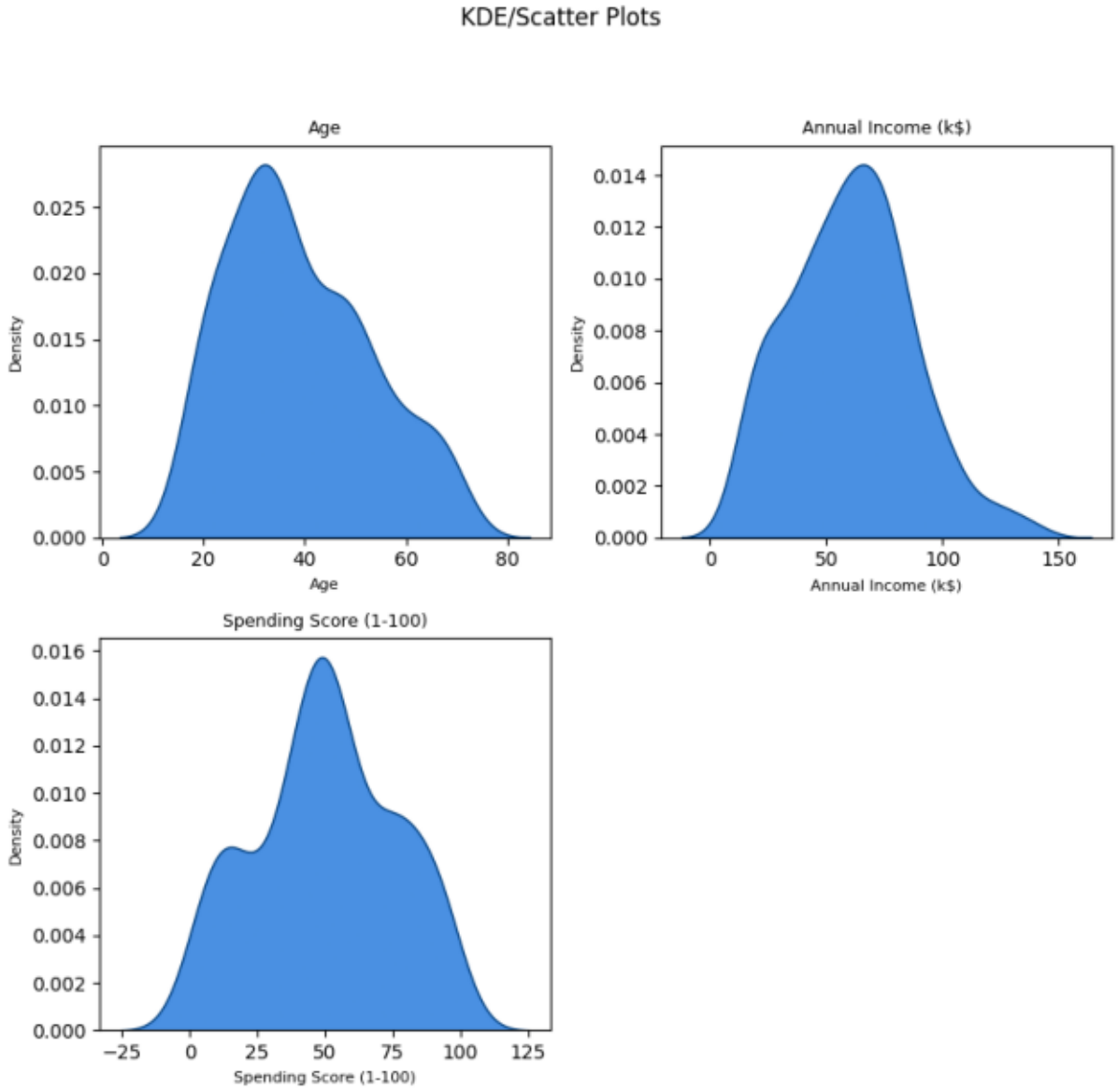
Value counts: Gender



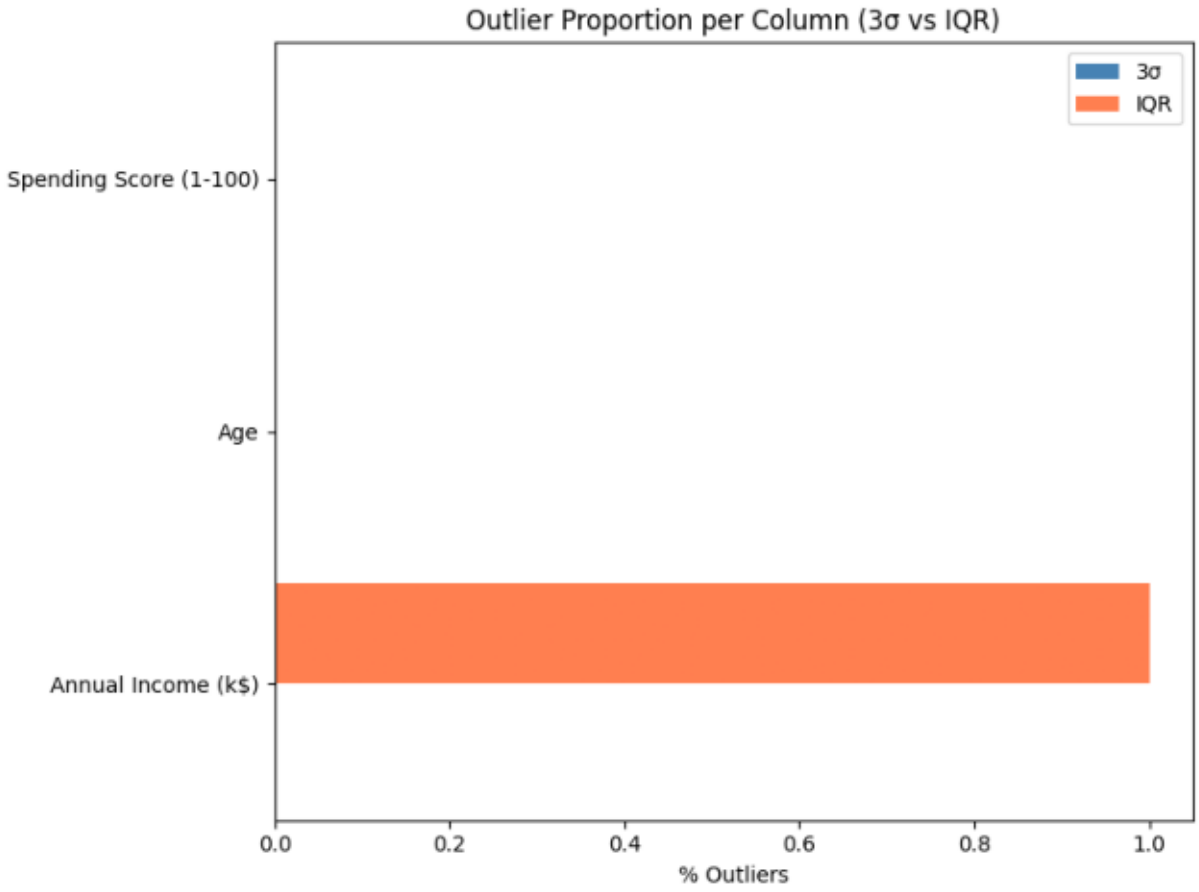
Distribution of Gender



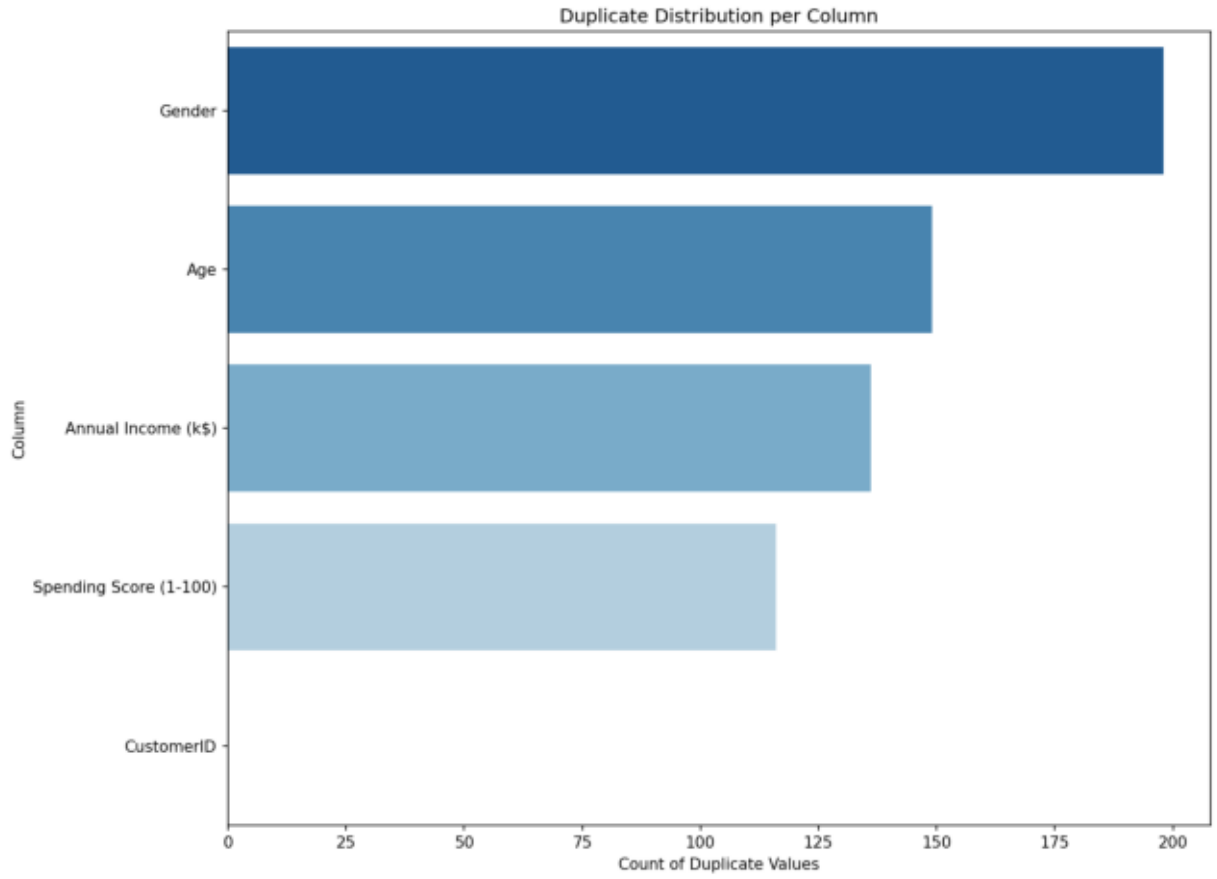
KDE Plots – Page 1



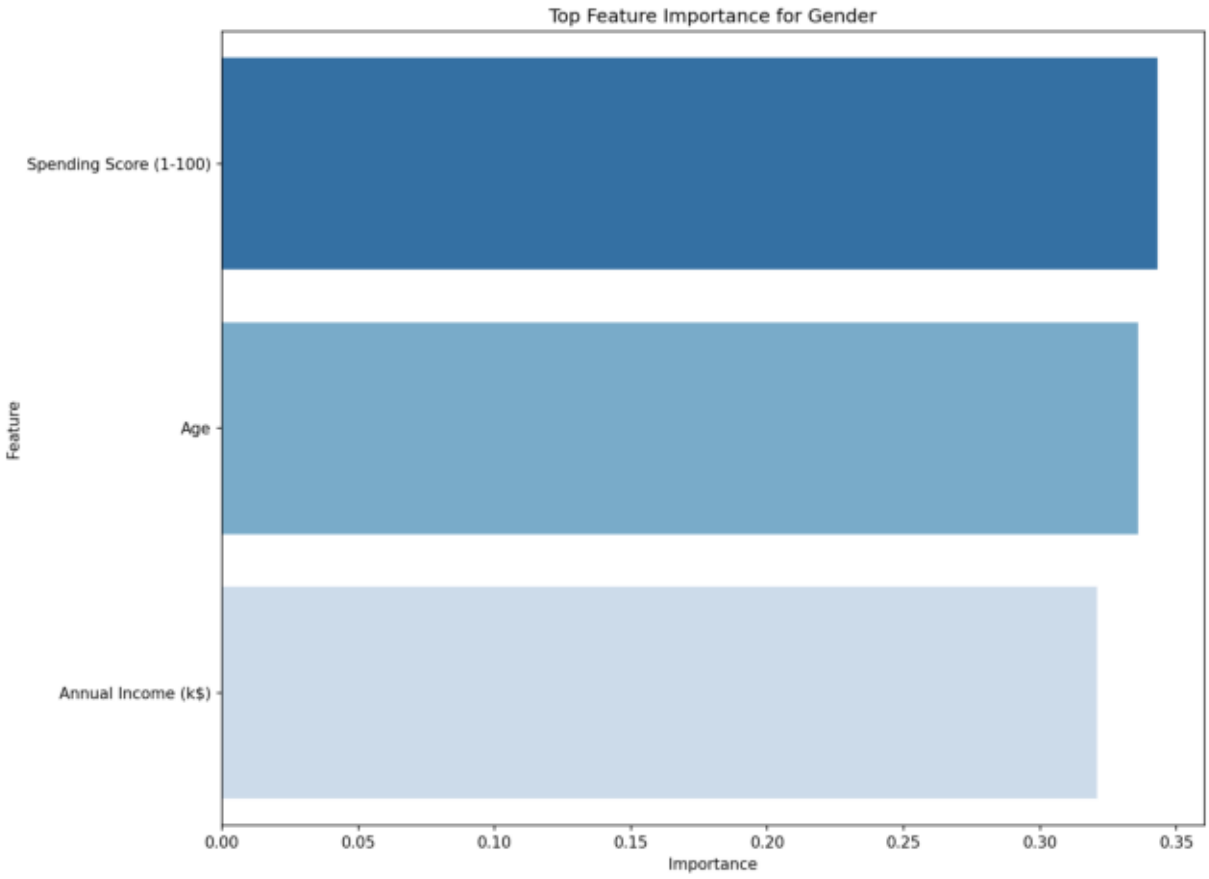
Outlier Proportion per Column



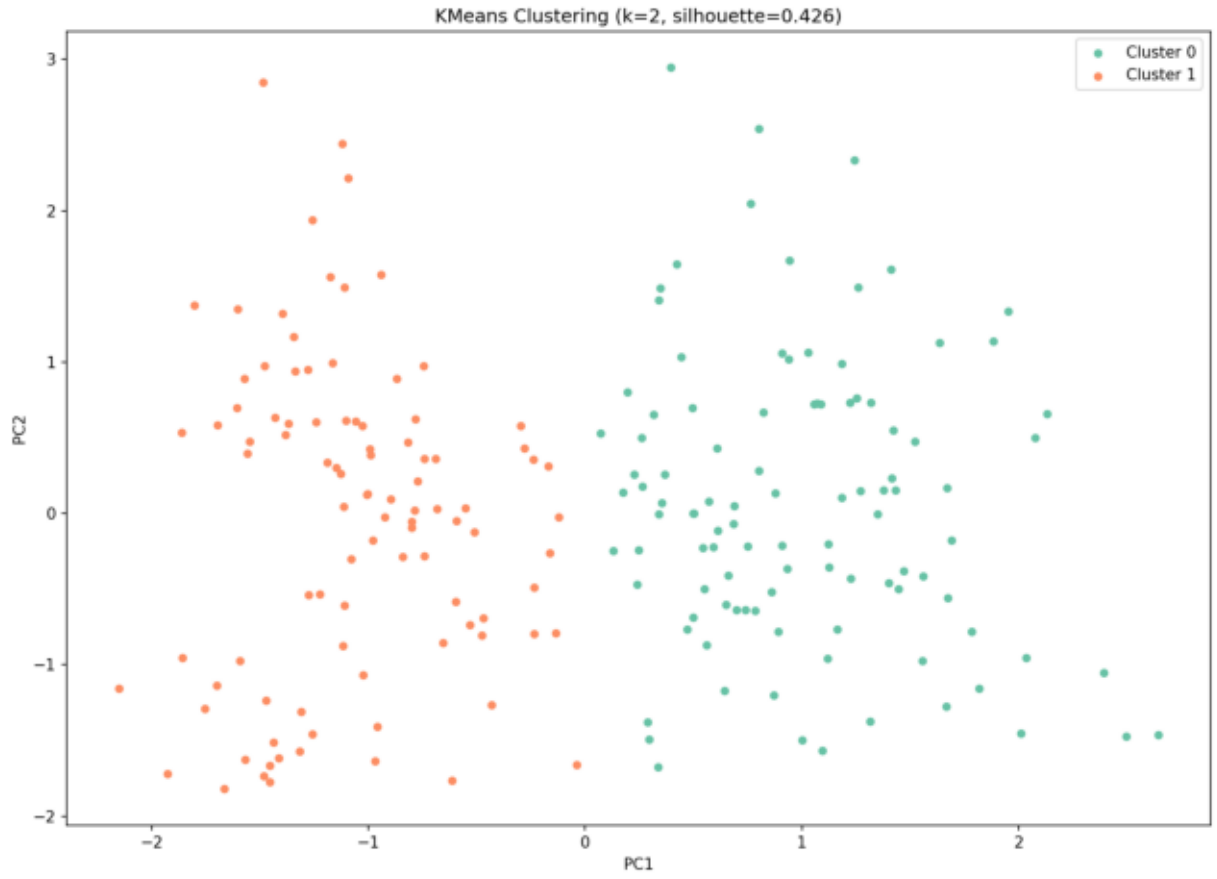
Duplicate Distribution per Column



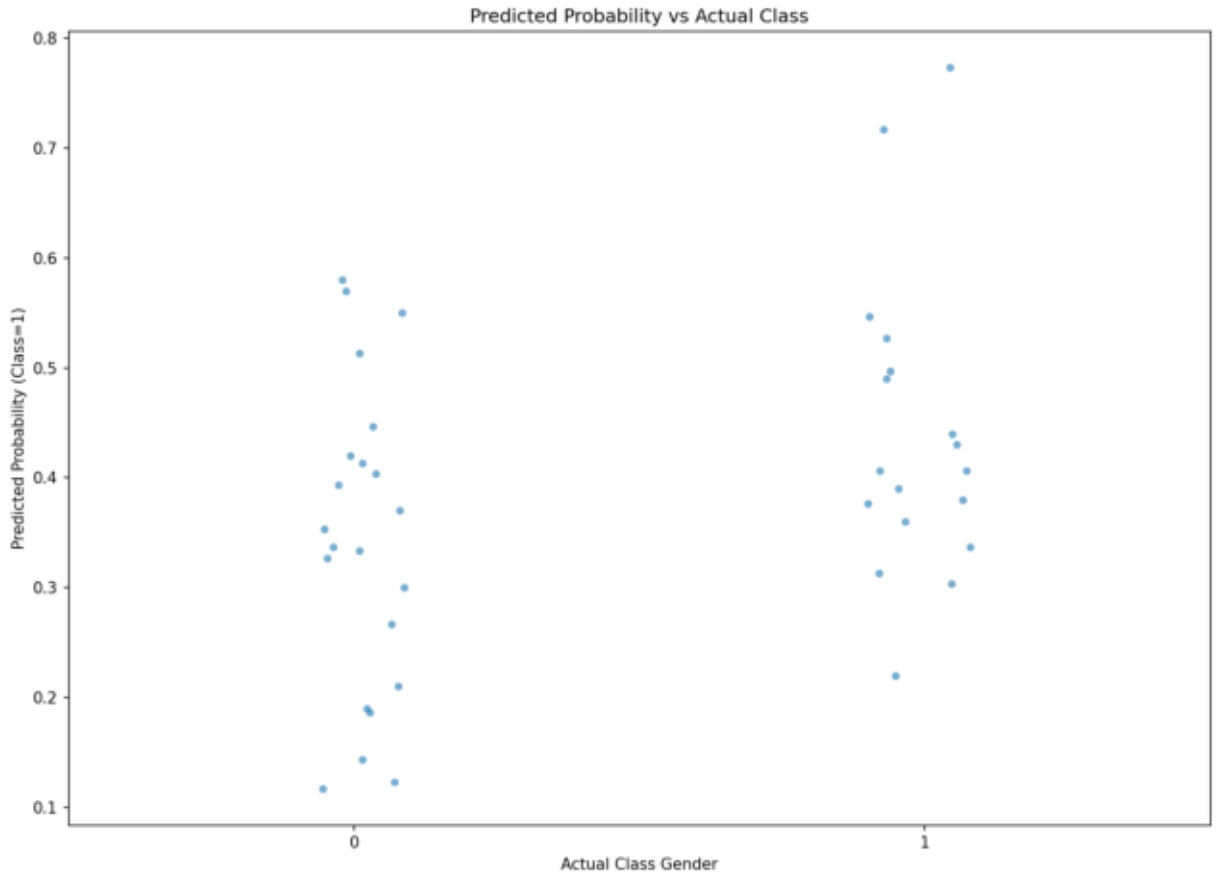
Feature Importance



PCA Clusters (Silhouette = 0.426)



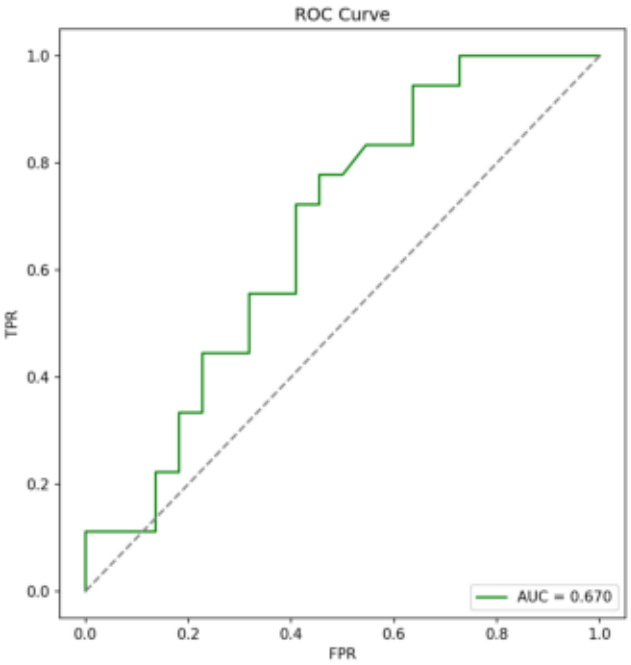
Predicted Probability vs Actual



Model Summary & ROC/Residuals

Model Summary

metric	value
train_accuracy	1.0
test_accuracy	0.55
roc_auc	0.6704545454545454
#id_like_cols_not_used	1.0



LLM Analysis

AI Agent Interpretation

The dataset contains 200 observations across 5 columns, with 3 numeric and 0 categorical features

Approximately 0.00% of the data is missing.

The columns with the highest missing rates are: CustomerID, Gender, Age, Annual Income (k$), Spe

This analysis addresses a classification task using 'Gender' as the target variable.

Outlier analysis (3σ vs. IQR) indicates that some features contain high percentages of extreme obse

Duplicate value analysis shows varying degrees of redundancy across columns; features with many

Correlation assessment reveals strong relationships among several measurement pairs (e.g. radius

PCA reduction followed by K-Means clustering suggests two distinct groups in the data (silhouette s

The classifier achieves a test accuracy of 0.5500.

The area under the ROC curve (AUC) is 0.6705, indicating excellent separability between classes.

The model identifies the most influential features as: Spending Score (1-100) (0.343), Age (0.336),

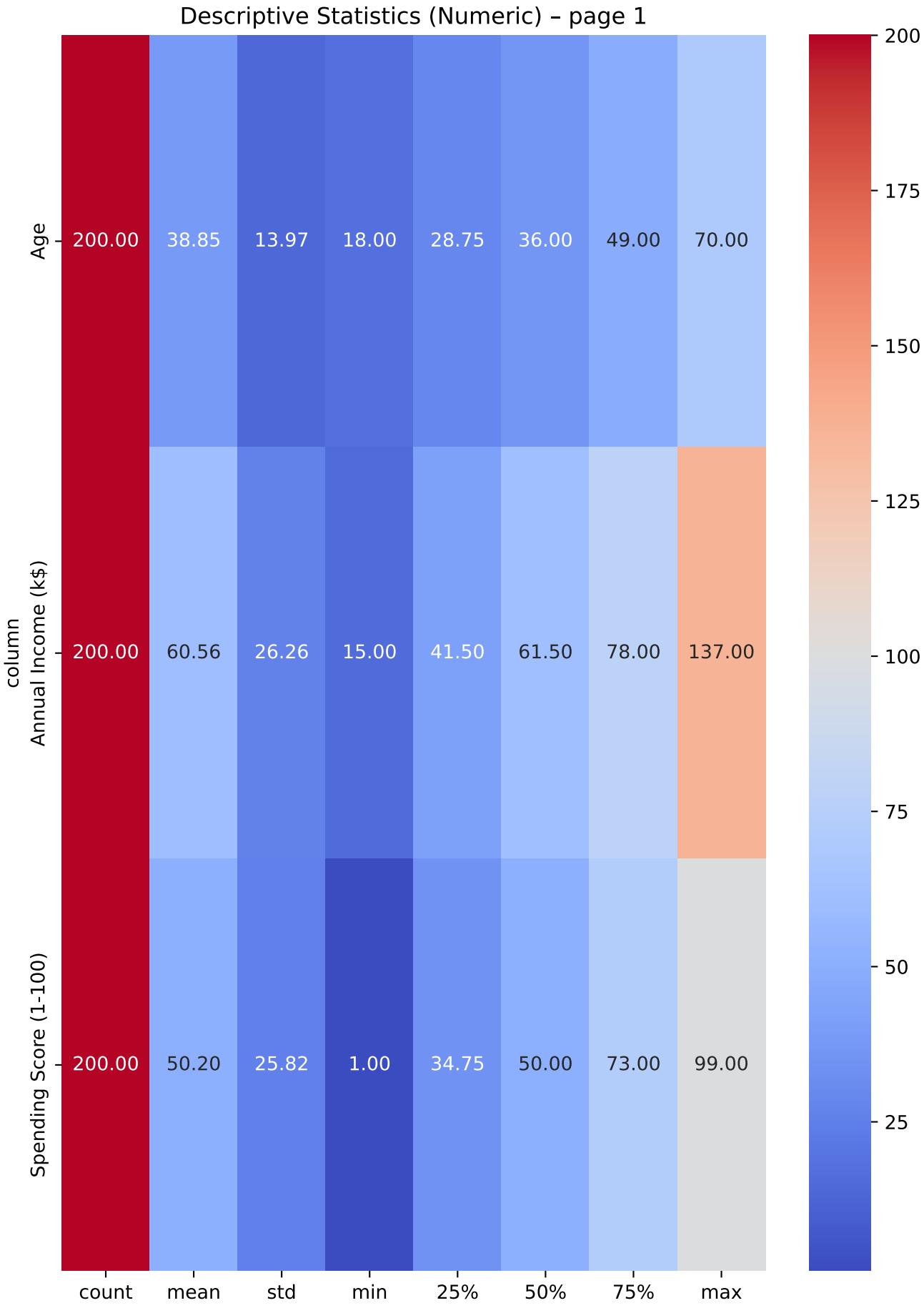
Based on these findings, efforts should prioritise imputing missing values, managing outliers, and e

Categorical Columns

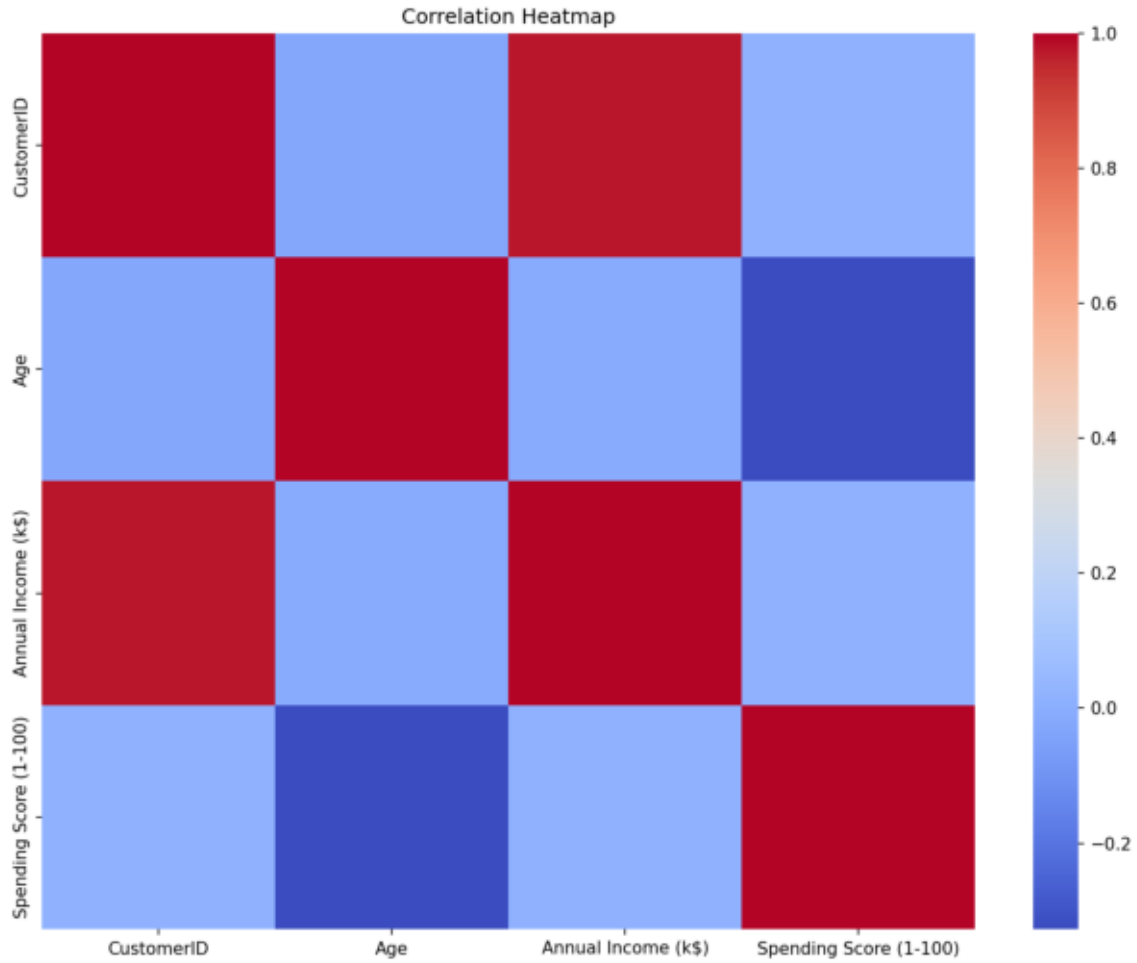
```
Gender
```

Categorical Summary

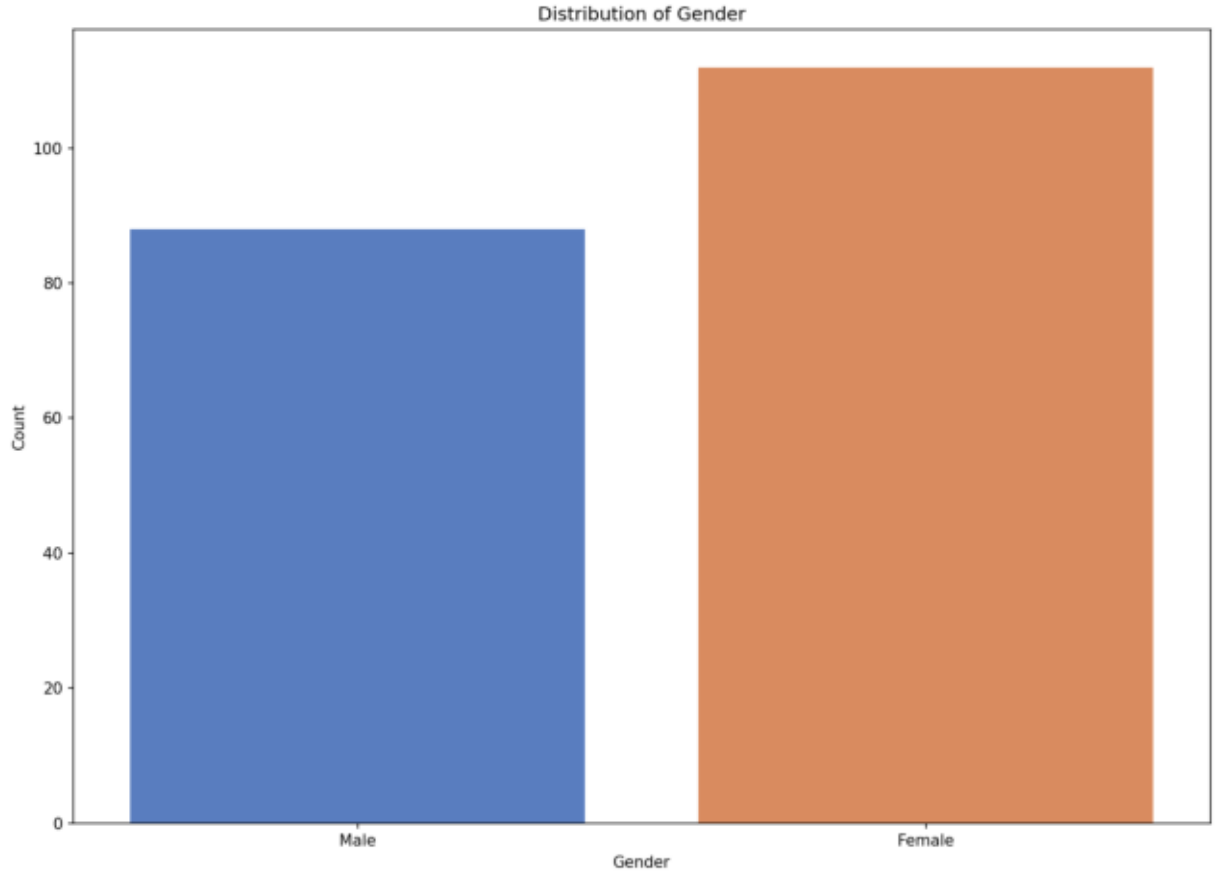
column	count	unique	top	freq
Gender	200	2	Female	112


Descriptive Statistics (Numeric) – page 1
column
Age
Annual Income (k$)
Spending Score (1-100)
200.00
38.85
13.97
18.00
28.75
36.00
49.00
70.00
200.00
60.56
26.26
15.00
41.50
61.50
78.00
137.00
200.00
50.20
25.82
1.00
34.75
50.00
73.00
99.00
count
mean
std
min
25%
50%
75%
max
200
175
150
125
100
75
50
25

Correlation Heatmap



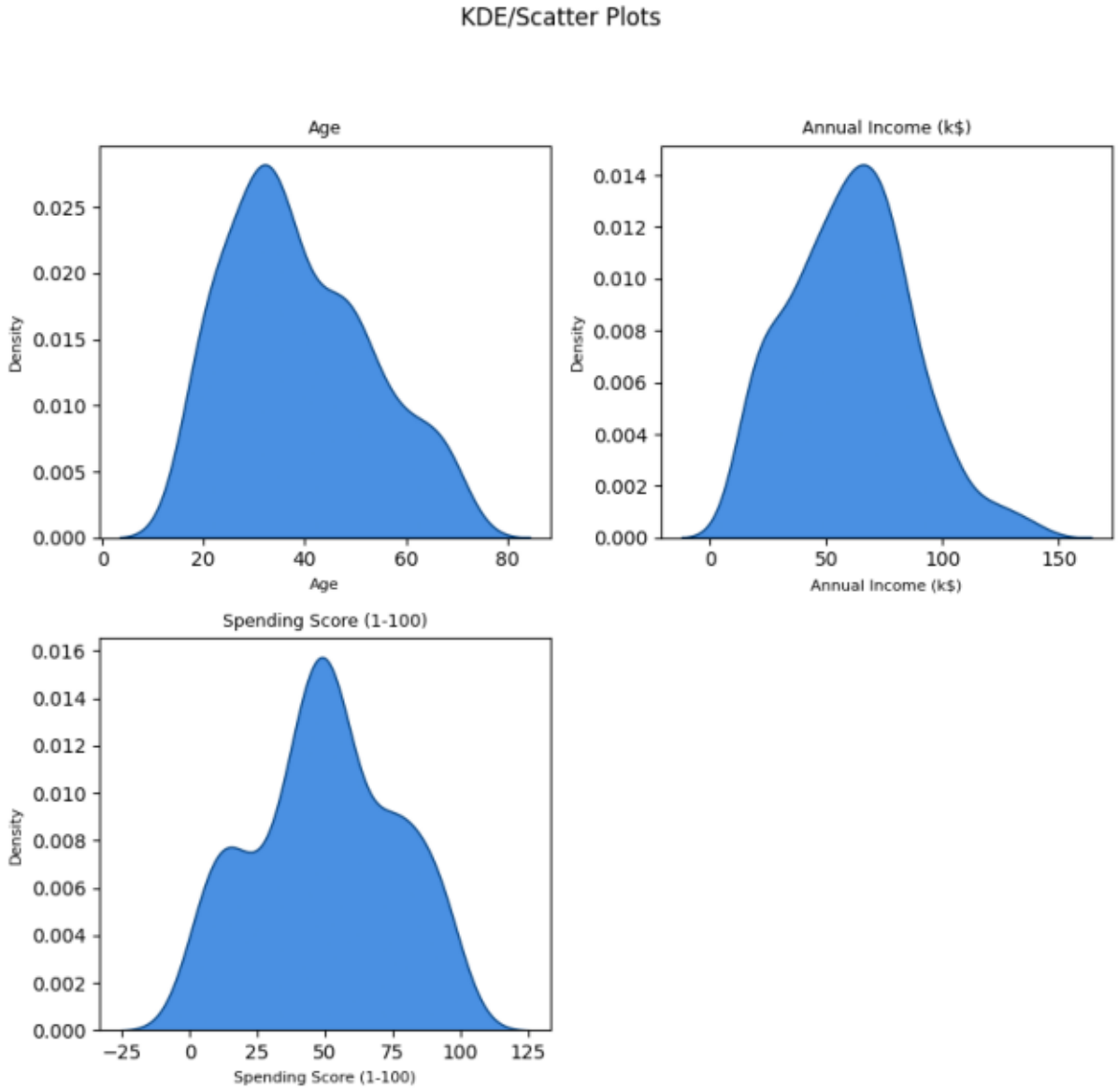
Target Distribution



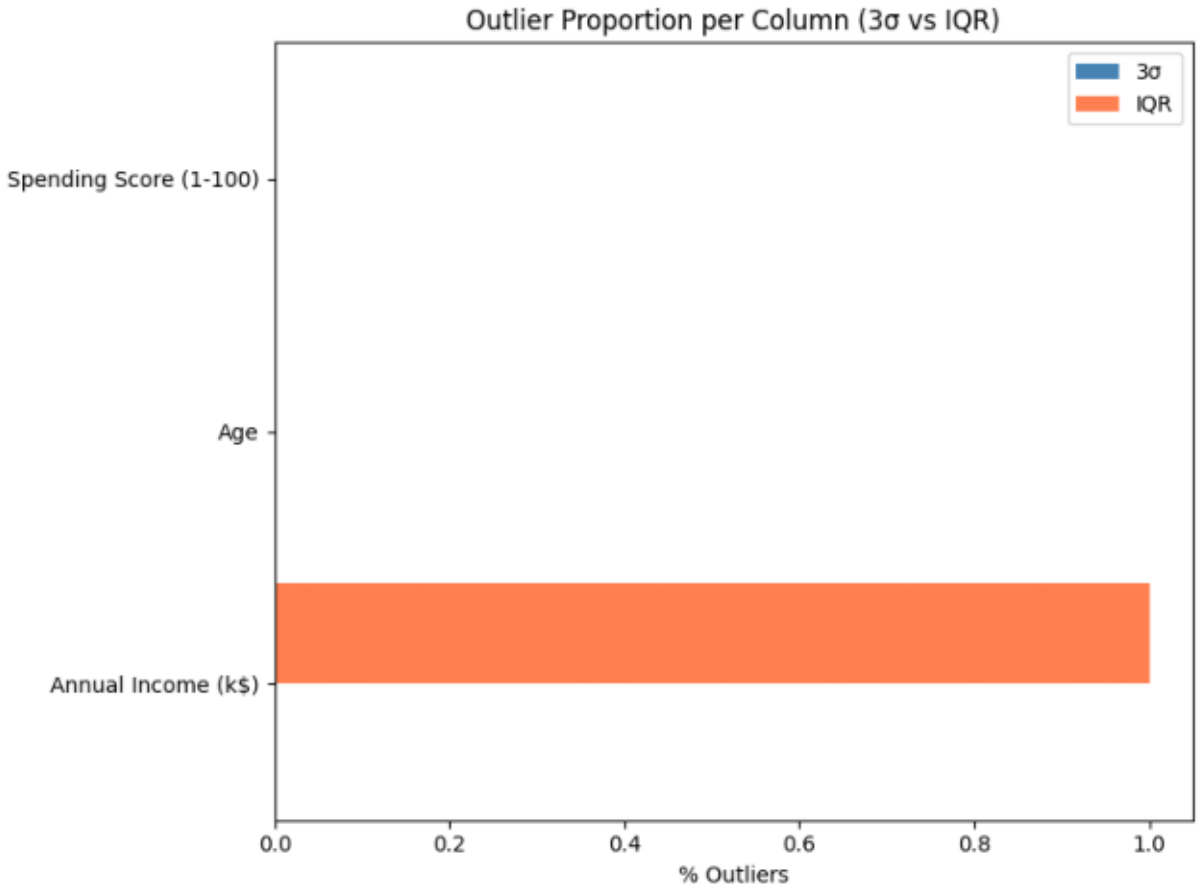
Target Summary

stat	value
count	200
unique	2
top	Female
freq	112

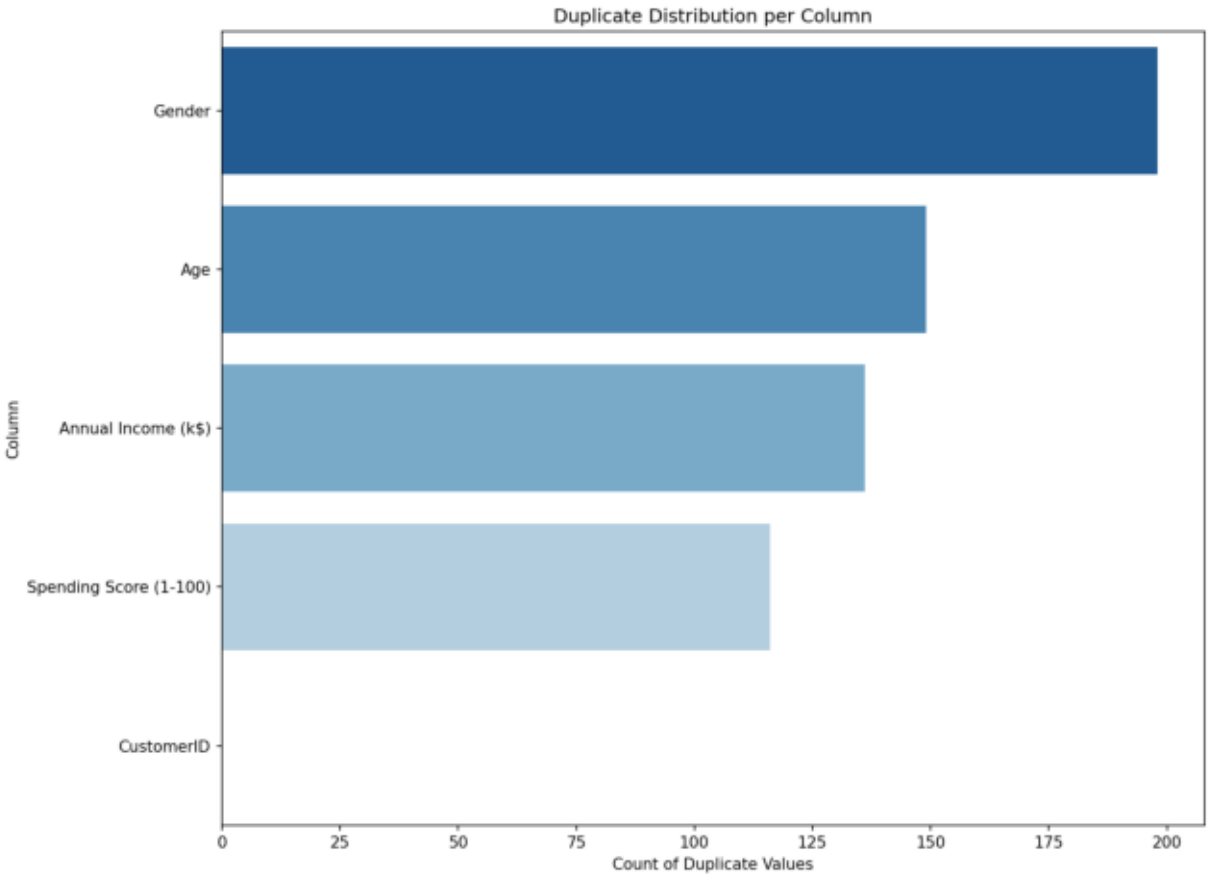
Feature vs Target (page 1)



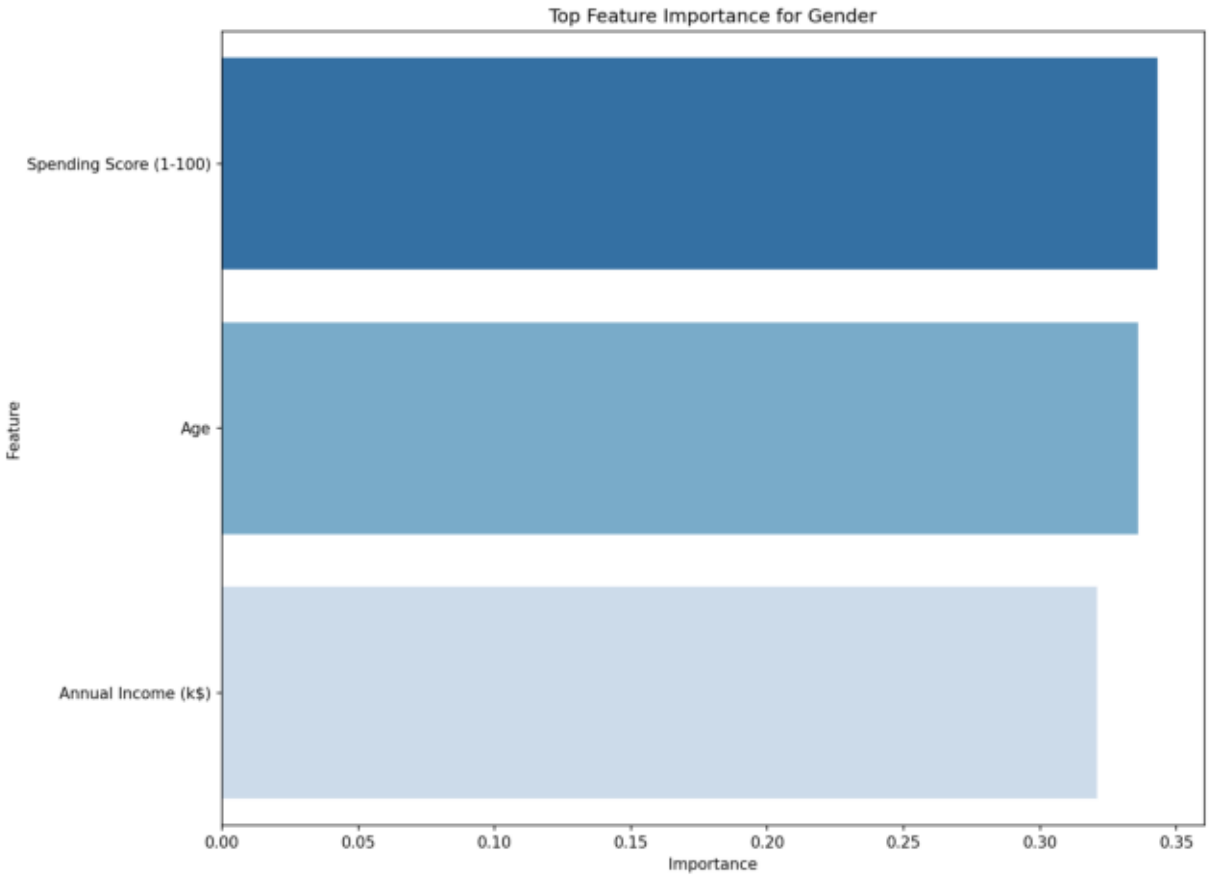
Outlier Proportion



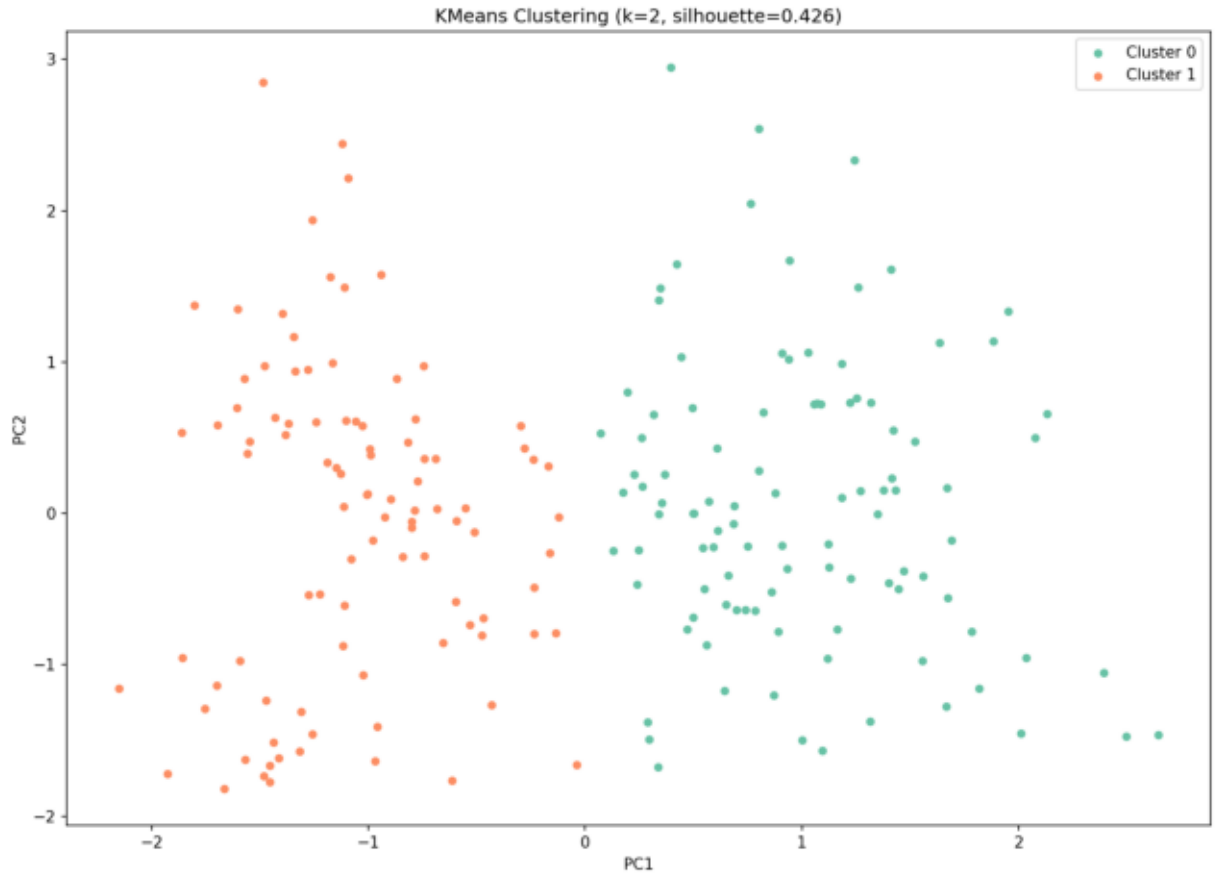
Duplicate Distribution



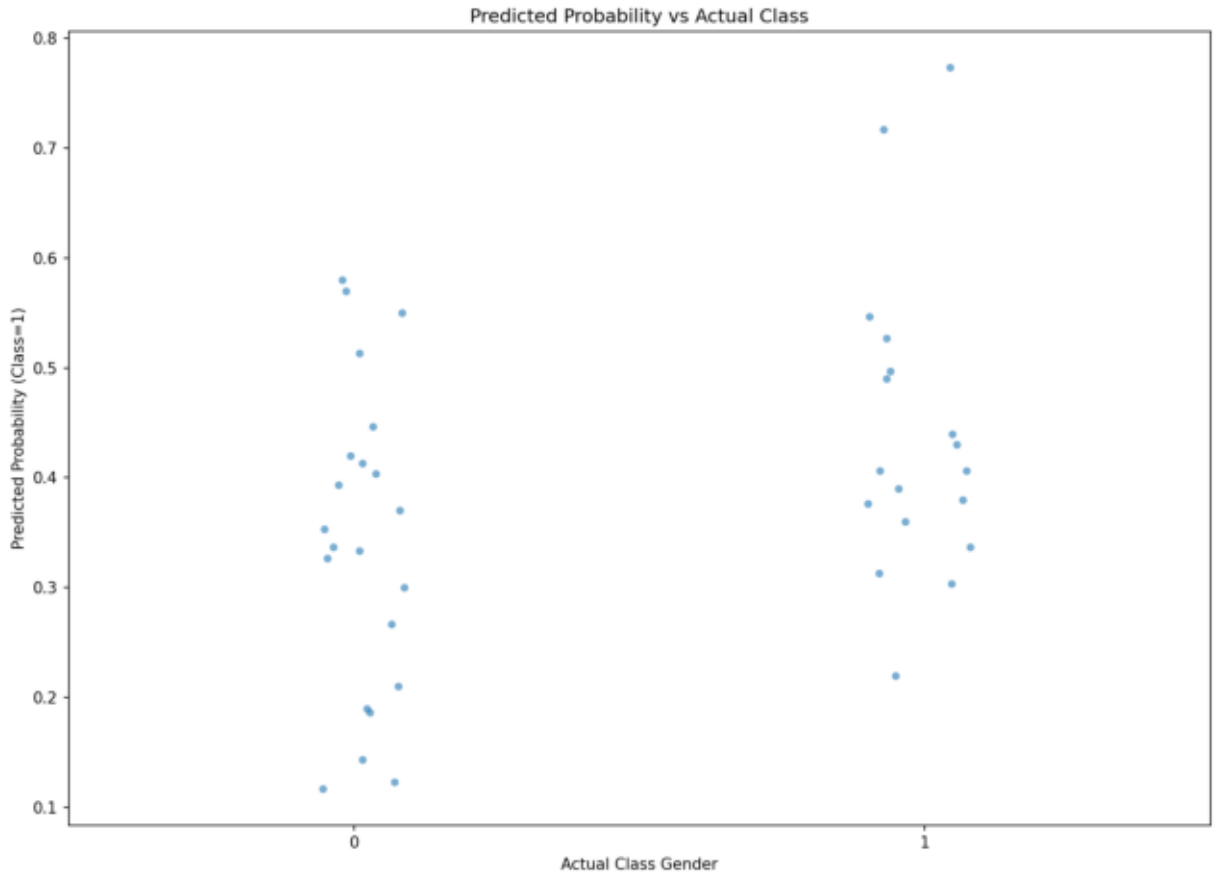
Feature Importance



PCA Clusters (silhouette=0.426)



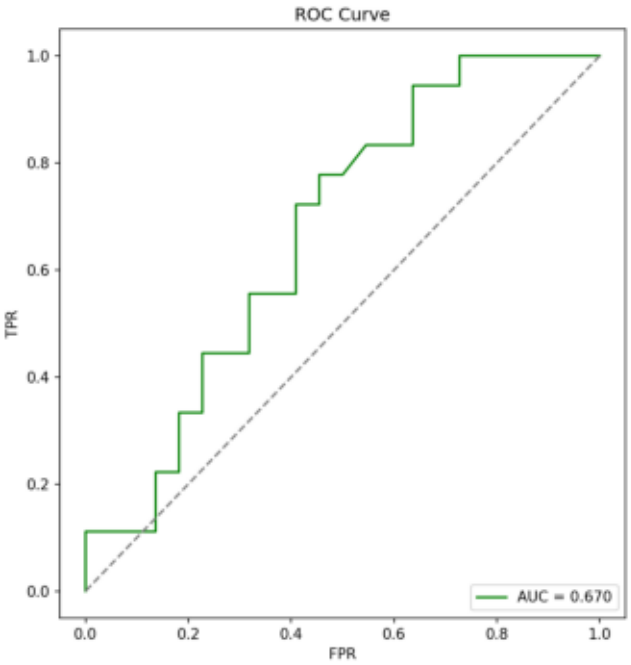
Predicted Probability vs Actual



Model Summary & Performance

Model Summary

metric	value
train_accuracy	1.0
test_accuracy	0.55
roc_auc	0.6704545454545454
#id_like_cols_not_used	1.0



LLM Interpretation

The dataset contains 200 observations across 5 columns, with 3 numeric and 0 categorical features

Approximately 0.00% of the data is missing.

The columns with the highest missing rates are: CustomerID, Gender, Age, Annual Income (k$), Spe

This analysis addresses a classification task using 'Gender' as the target variable.

Outlier analysis (3σ vs. IQR) indicates that some features contain high percentages of extreme obse

Duplicate value analysis shows varying degrees of redundancy across columns; features with many

Correlation assessment reveals strong relationships among several measurement pairs (e.g. radius

PCA reduction followed by K-Means clustering suggests two distinct groups in the data (silhouette s

The classifier achieves a test accuracy of 0.5500.

The area under the ROC curve (AUC) is 0.6705, indicating excellent separability between classes.

The model identifies the most influential features as: Spending Score (1-100) (0.343), Age (0.336),

Based on these findings, efforts should prioritise imputing missing values, managing outliers, and e