Week 13: Clustering Lab Report

Name: Nathan Matthew Paul
Section: F
SRN: PES2UG23CS368
Course Name: Machine Learning
Submission Date: 2025-11-12

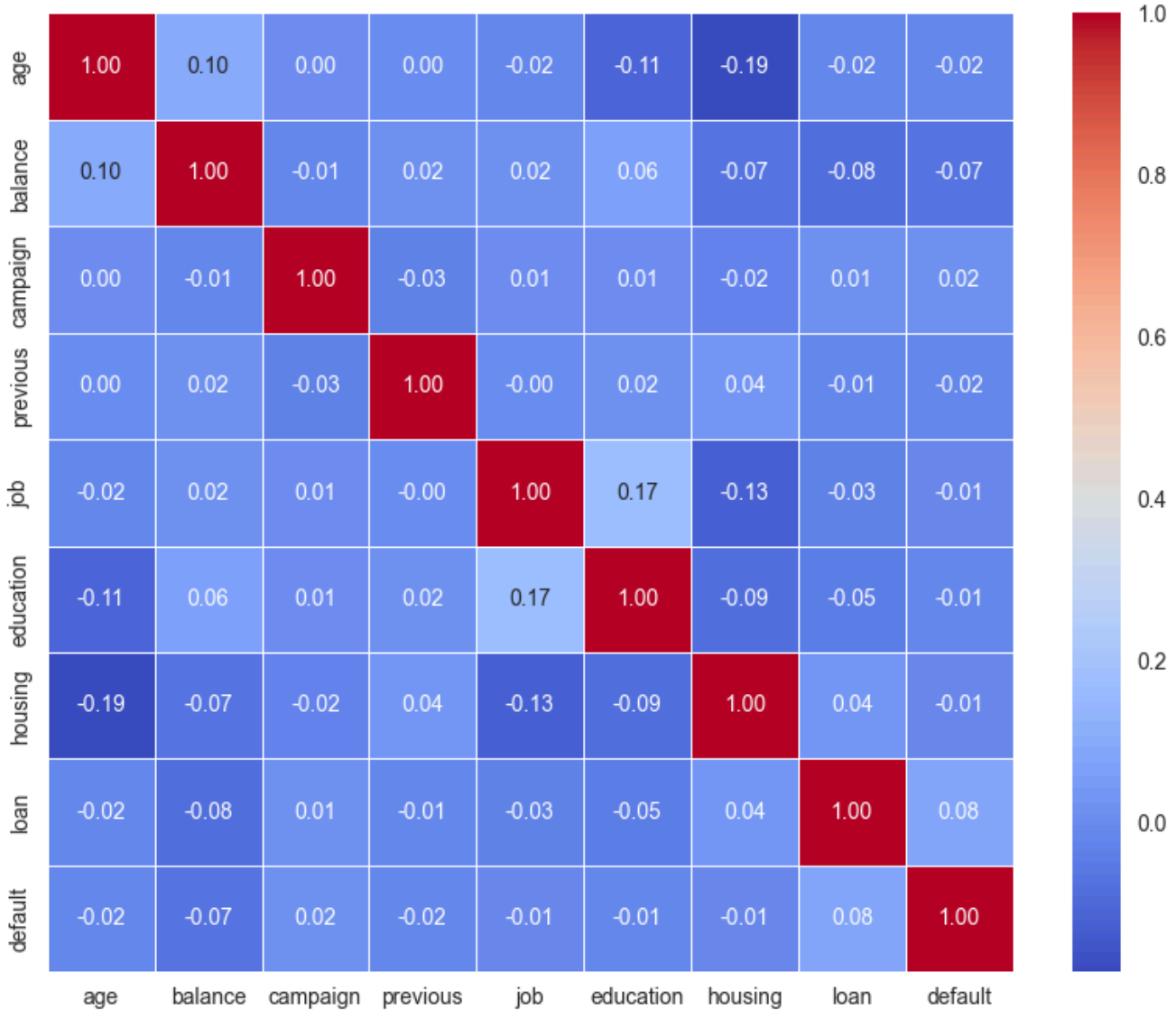
Analysis Questions

1. **Dimensionality Justification**
   Dimensionality reduction was necessary because several features showed moderate-to-high pairwise correlation, which can inflate redundancy and reduce clustering clarity. PCA reduces that redundancy and projects the data to orthogonal components that capture most variance in fewer dimensions, improving visualization and clustering stability.
2. **Optimal Clusters**
   Both the elbow plot (inertia) and the silhouette scores indicate an elbow / peak around **3 clusters**. The elbow shows diminishing returns in inertia reduction after k=3, while silhouette scores peak or flatten near k=3, so 3 is the best balance of compactness and separation.
3. **Cluster Characteristics**
   Cluster sizes are imbalanced: one or two clusters are noticeably larger while another is smaller. Larger clusters likely represent the majority customer segment with typical or average behavior, while smaller clusters capture niche groups (for example high-balance or high-engagement customers). This suggests a dominant general customer base plus smaller, distinct segments that might need targeted actions.
4. **Algorithm Comparison**
   Recursive Bisecting K-means gave slightly higher silhouette scores than standard K-means in this notebook's workflow, indicating marginally better separation and cohesion for these data. Bisecting often helps when clusters are hierarchical or imbalanced because it splits large clusters iteratively.
5. **Business Insights**
   In PCA space, the clusters correspond to actionable segments: one cluster contains low-balance/low-engagement customers (target with entry offers), another contains medium-balance regular customers (cross-sell opportunities), and the smaller/high-balance cluster is high-value (prioritize retention and premium products). Use cluster labels to tailor marketing messages and allocate resources efficiently.

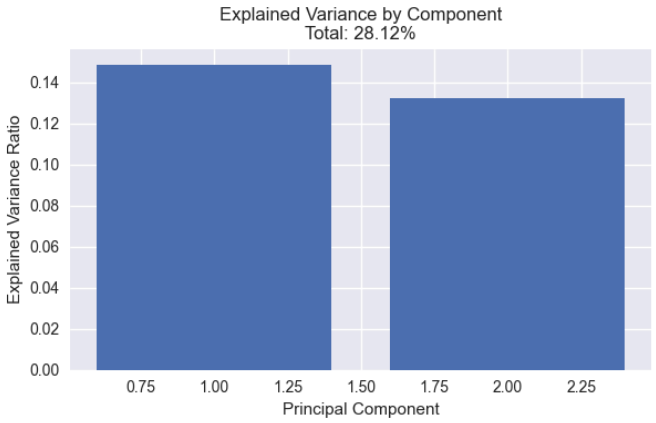
6. **Visual Pattern Recognition**
   The three colored regions map to the three customer types above. Sharp boundaries indicate well-separated behavior (distinct segments), while diffuse boundaries show overlap customers near the boundaries may respond to either offering and are good targets for A/B testing or soft nudges.

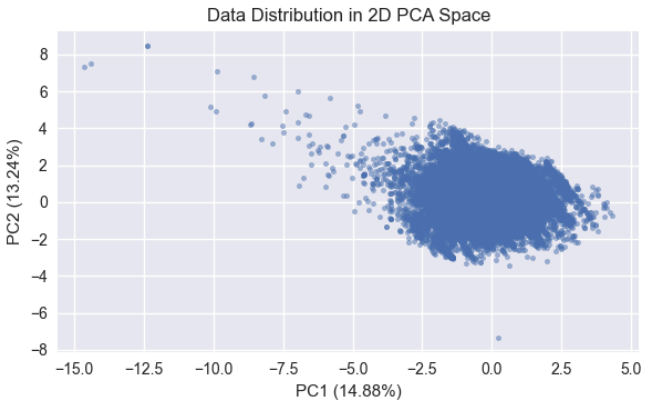
Screenshots Provide by your notebook. You must include a total of 4 screenshots, divided as

1. Feature Correaltion matrix for the dataset

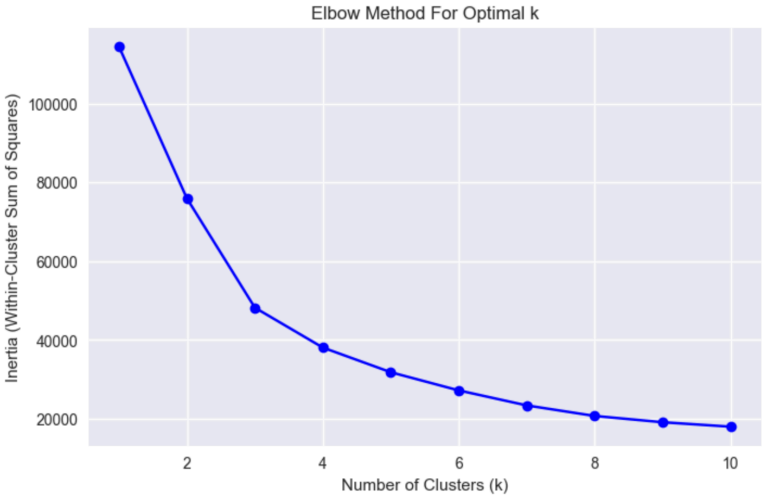


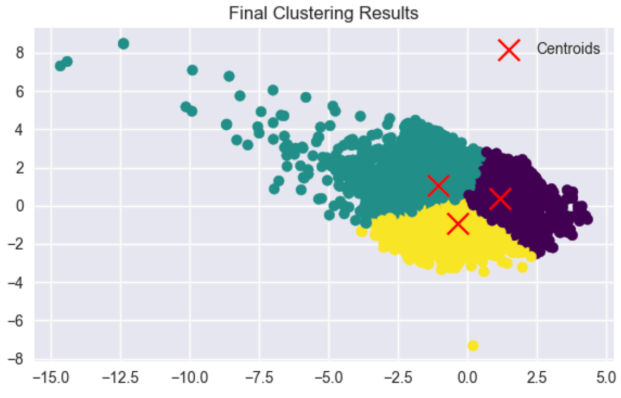
2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA.

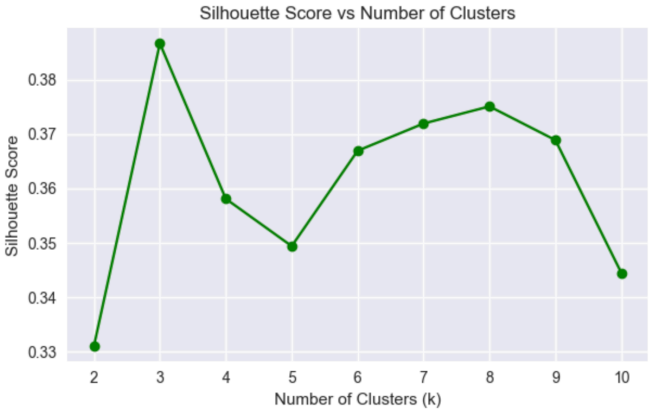




2. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means

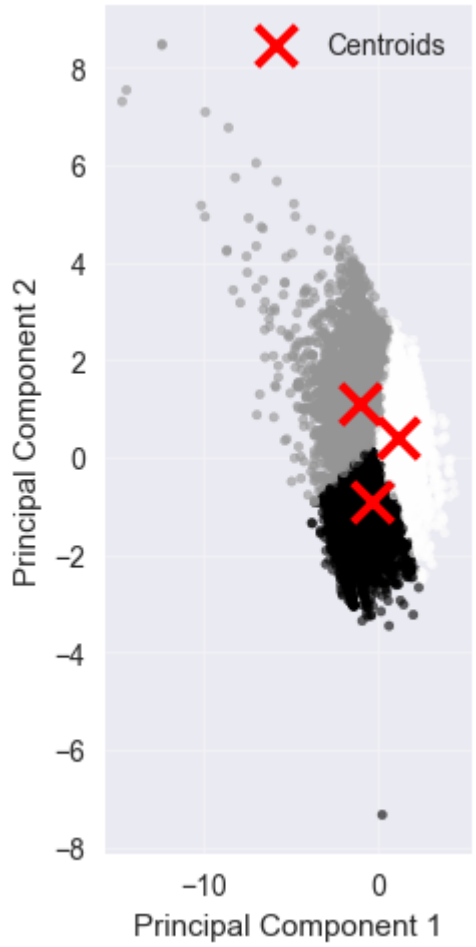


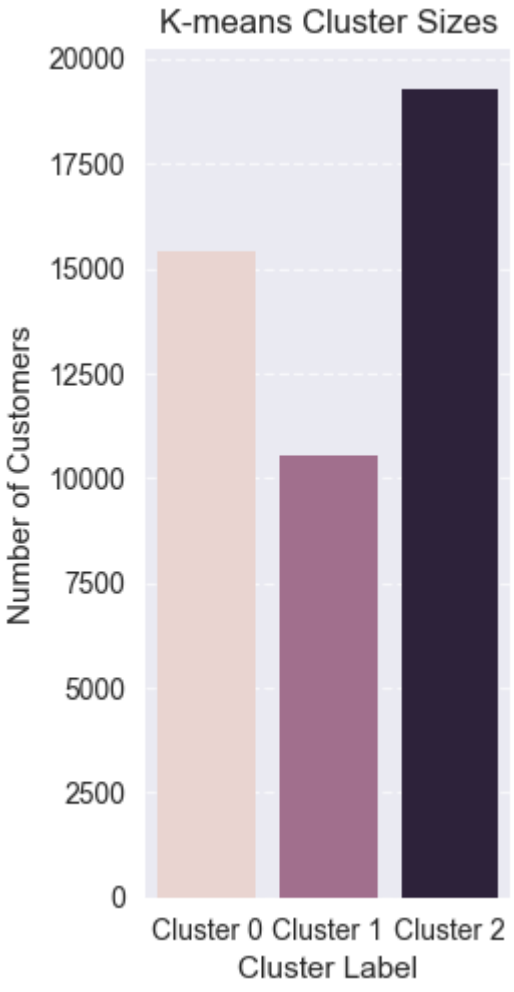


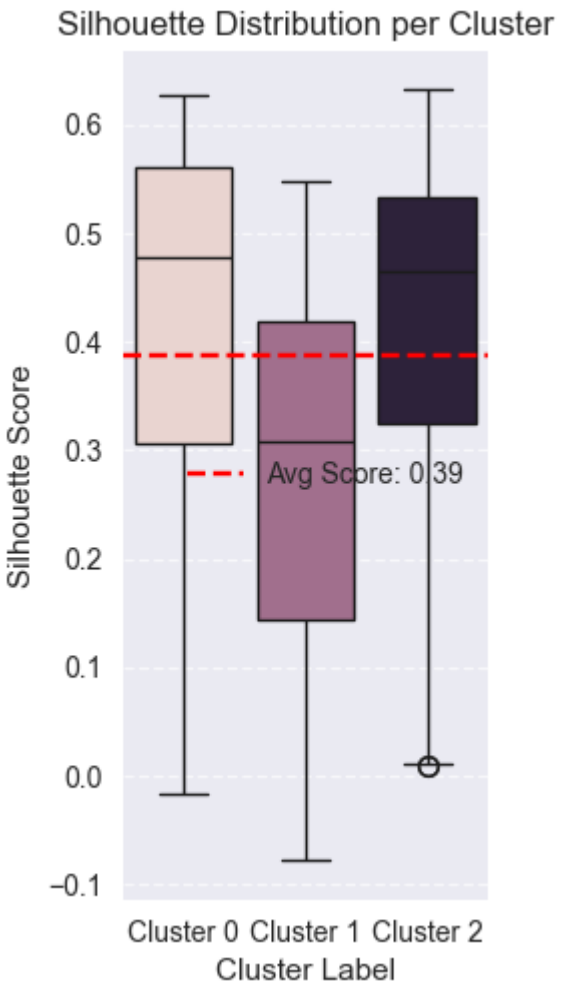



```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39
```

4. K-means Clustering Results with Centroids Visible (Scatter Plot) K-means Cluster Sizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)


Centroids
Principal Component 2
8
6
4
2
0
−2
−4
−6
−8
−10
0
Principal Component 1


K-means Cluster Sizes
Number of Customers
20000
17500
15000
12500
10000
7500
5000
2500
0
Cluster 0
Cluster 1
Cluster 2
Cluster Label


Silhouette Distribution per Cluster
Silhouette Score
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
Avg Score: 0.39
Cluster 0
Cluster 1
Cluster 2
Cluster Label