2024 Cleveland Cavaliers Sport Business and Analytics Night Hackathon

Andrew Yu

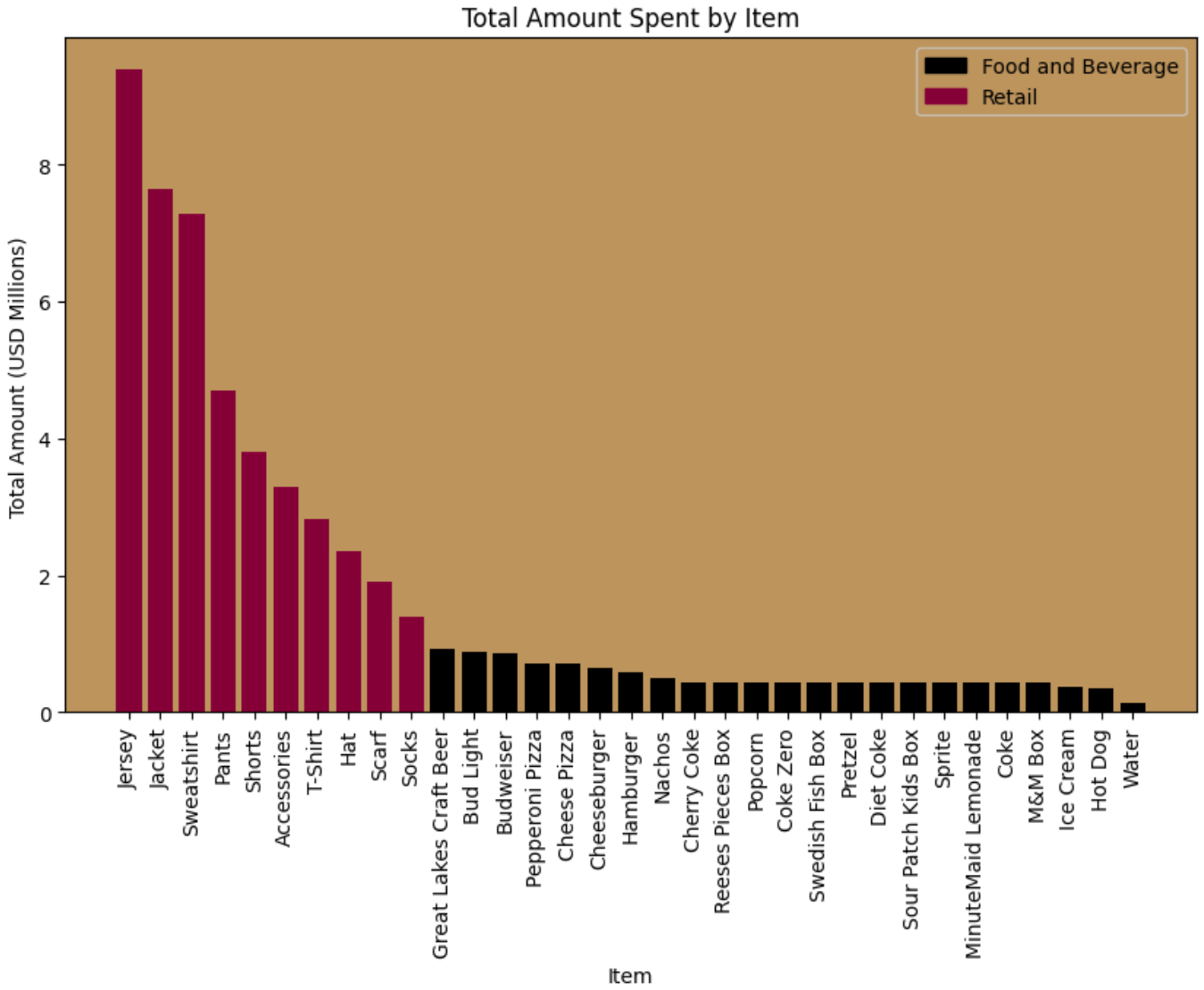
Case Western Reserve University

Problem Statement

- Business intelligence
  - Collect and analyze customer interactions and business operations
  - **Drive growth** and profitability
  - Track trends, monitor performance, and make informed decisions
- Analyze food, beverage, and retail sales for the Cavs
  - **Segment fans** based on purchases
  - Discover behaviors, patterns, and preferences
  - Create opportunities to tailor our offerings

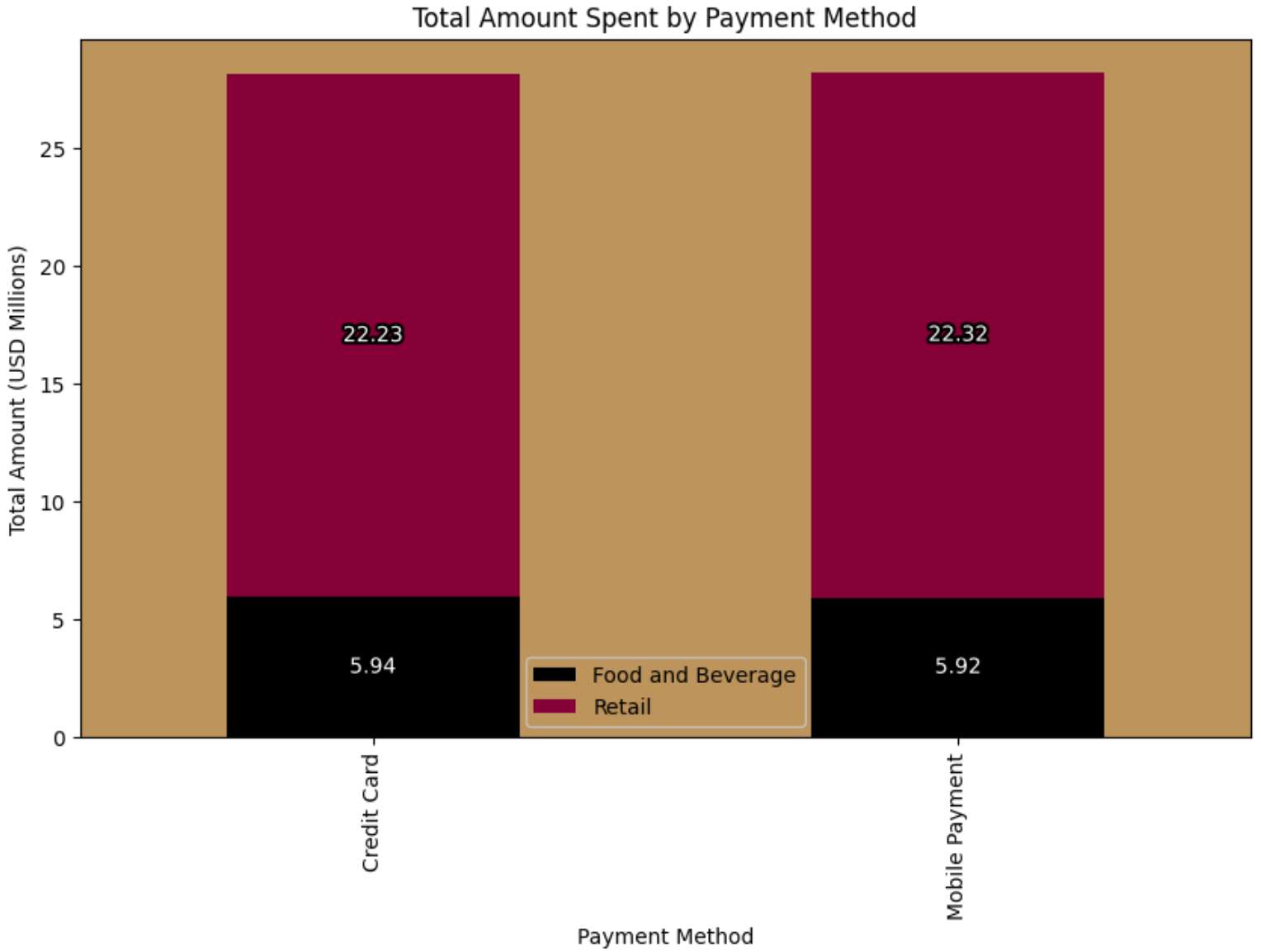
Total Revenue per Item

- Retail overshadows F&B heavily
- In F&B, beer and hot food bring in the most revenue



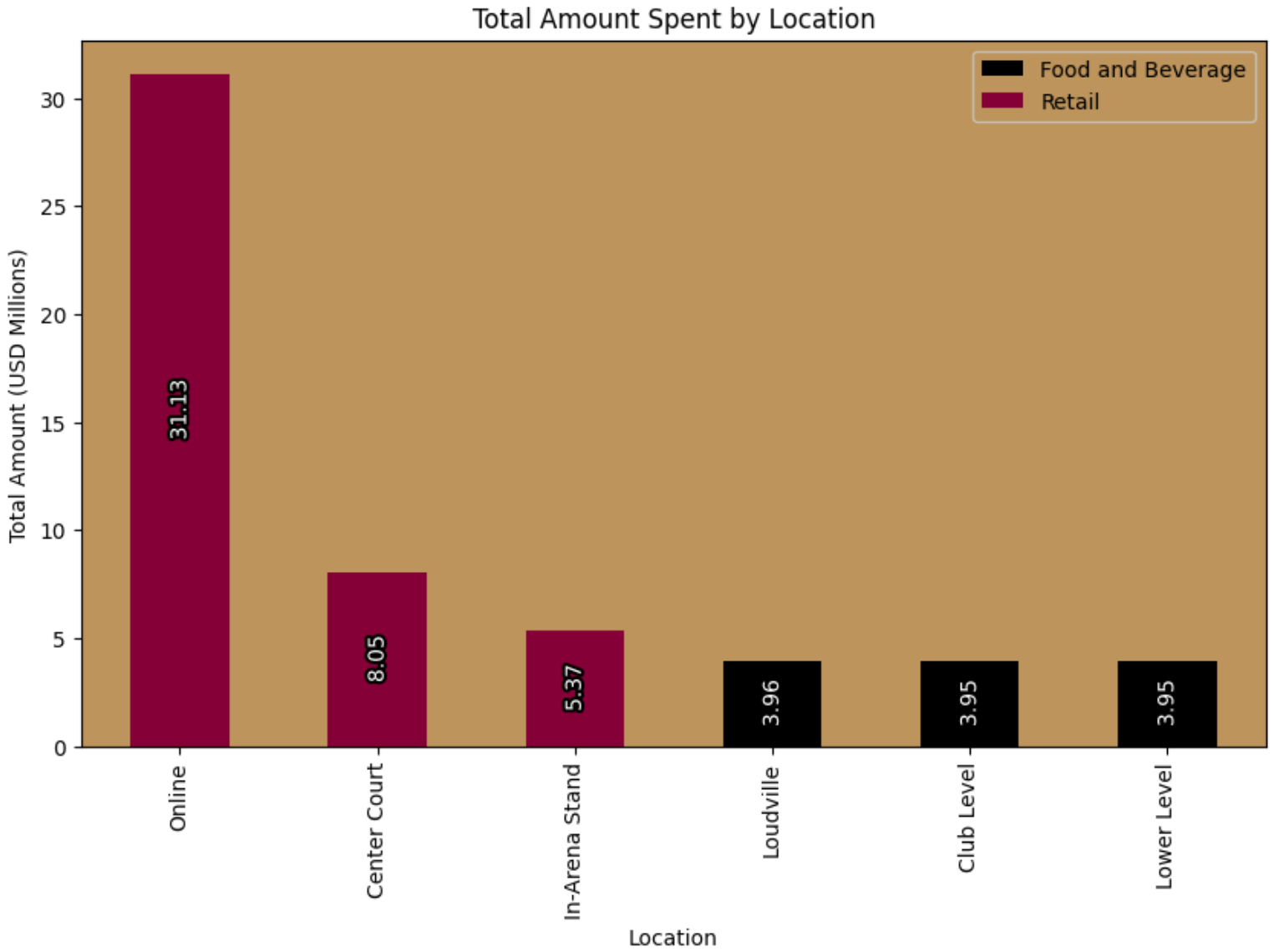
Total Revenue per Payment Method

- Credit card vs. mobile payment
- Distributed evenly across all groupings



Total Revenue per Location

- Online retail dominates
- All location-based revenue is secondary to what they sell (retail vs. F&B)



Feature Derivation: 110 Total Features

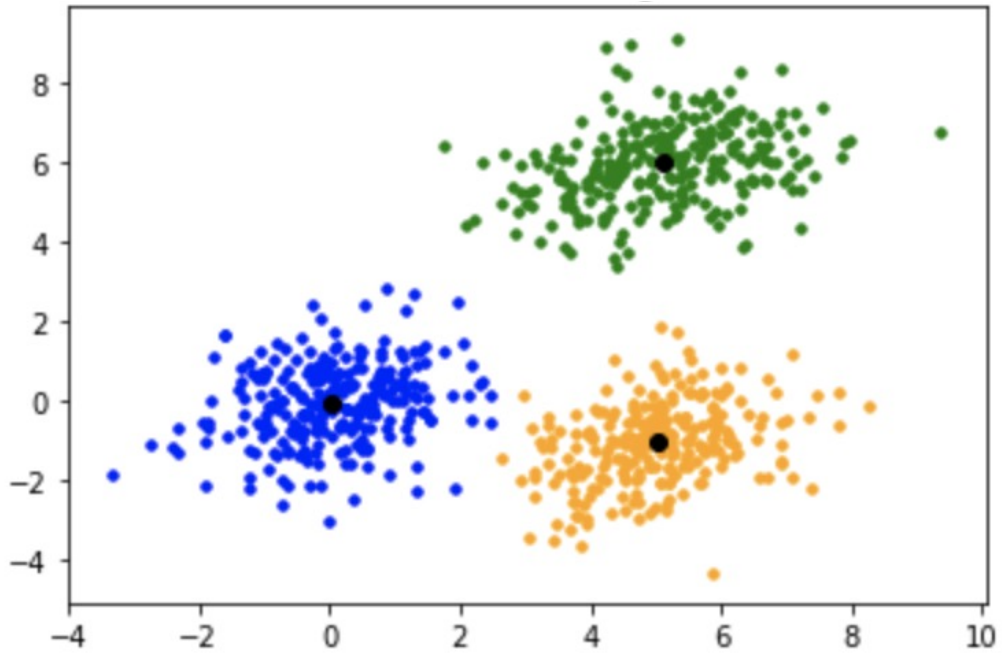
- Time features:
  - Days since season start (1)
  - Seconds since game start (1)
- External data:
  - Win / loss (1)
  - East or West conference (1)
- Continuous values:
  - Total amount (1)
  - Quantity (1)
- Categorical Values:
  - Item (33)
  - Category (2)
  - Payment method (2)
  - Location (6)
- Derived features:
  - Subcategories:
    - Alcoholic / non-alcoholic (2)
    - Hot food / snack (2)
    - Clothing / accessories (2)

- Aggregation: Sum and mean (x2 for all)
- Scaling: Adjust values to standard normal distribution

Unsupervised K-Means Clustering

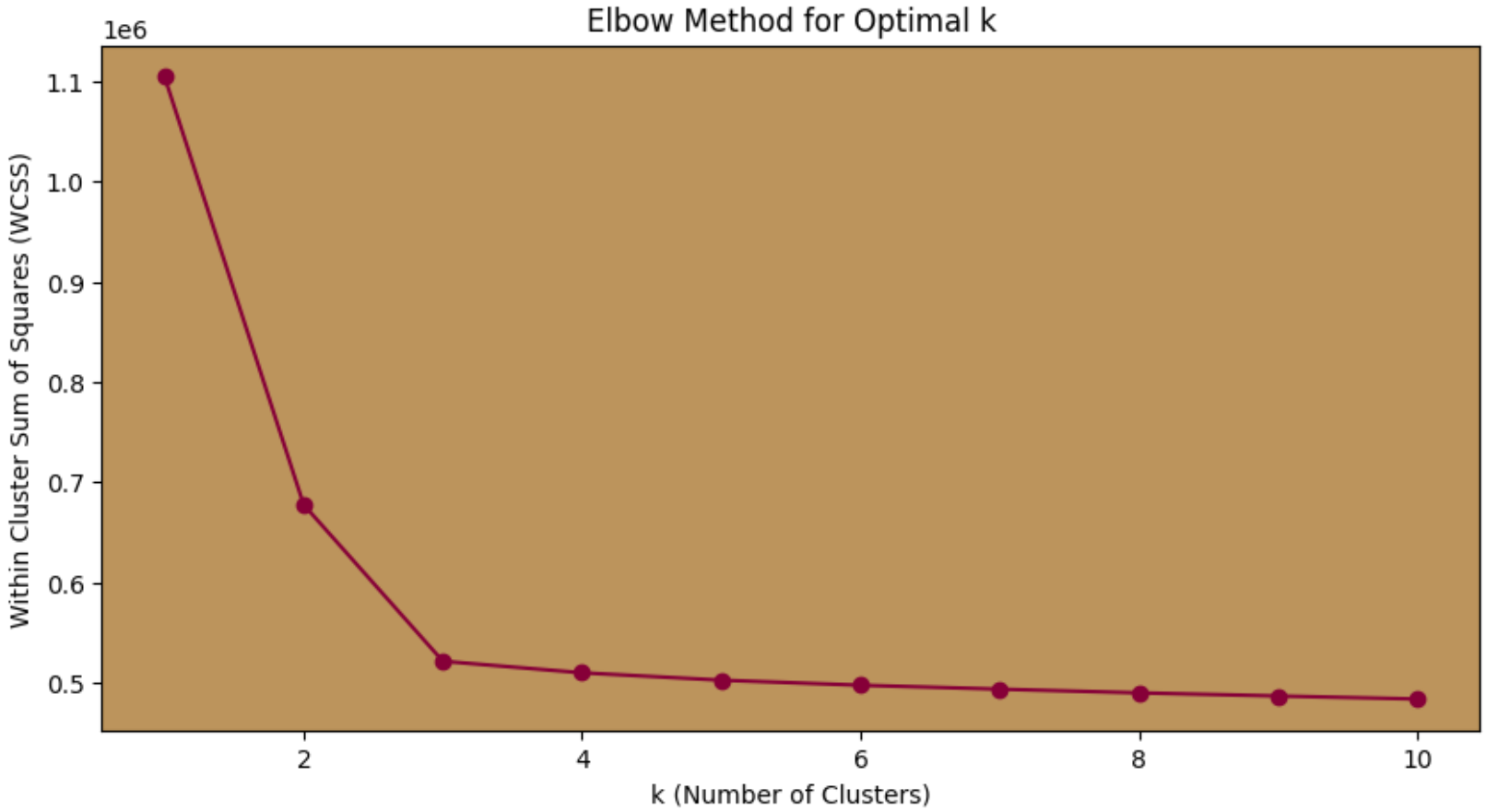
- Clustering samples into groups of similar features without any labeled data

- Steps:
  1. (Find number of clusters)
  2. Assign each sample to nearest centroid
  3. Update centroids' locations
  4. Repeat 2 and 3 until converged



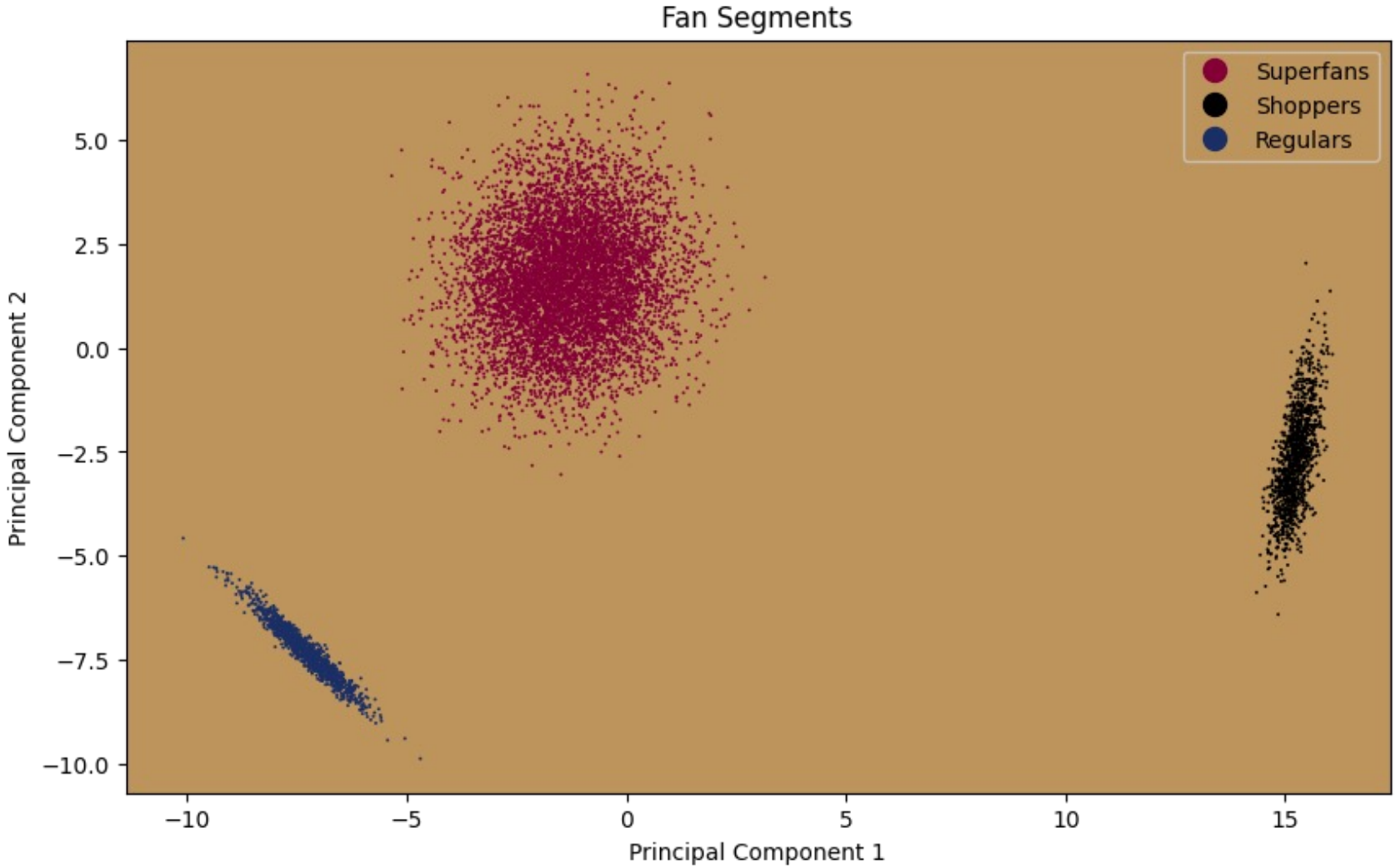
Choosing k (Number of Clusters)

- Run k-means for each possible number of clusters
- Compute the cluster spread from the centroids (WCSS)
- Determine elbow (point of diminishing returns)
- In this case, k=3



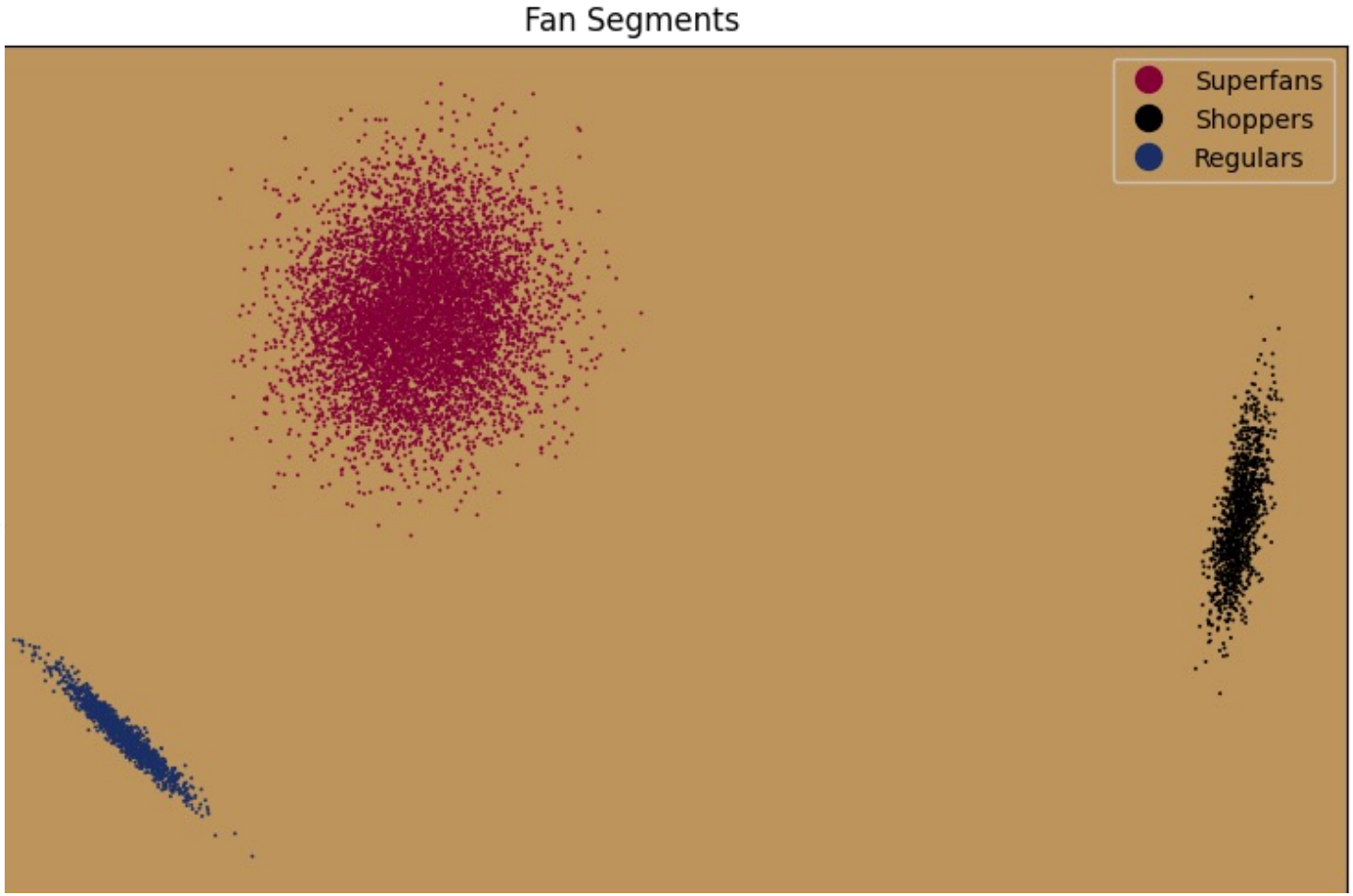
Principal component analysis (PCA)

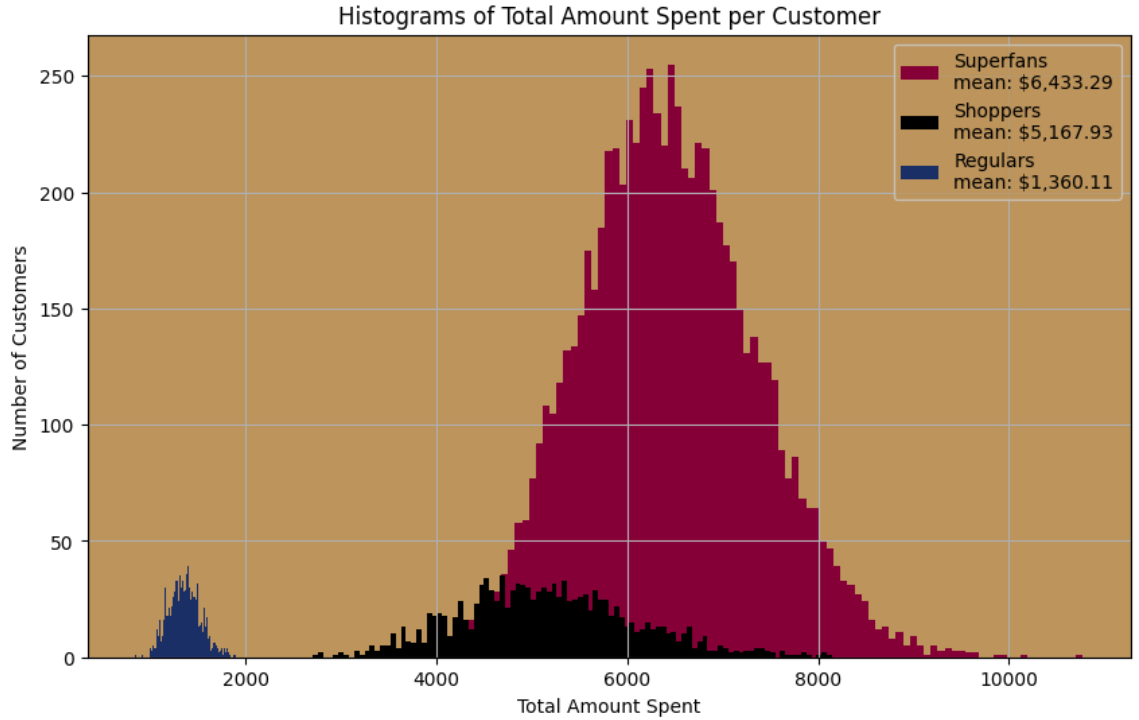
- Derive linear combinations of original features
- Trade accuracy and interpretability for simplicity
- Visualize complex data in 2D

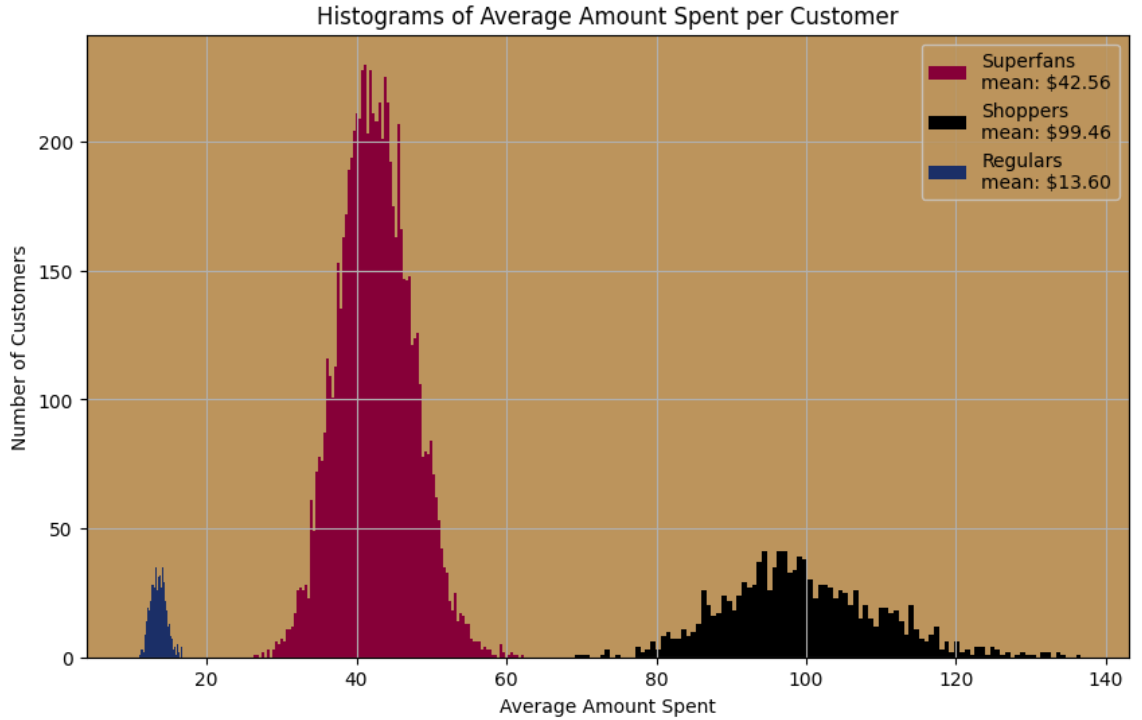


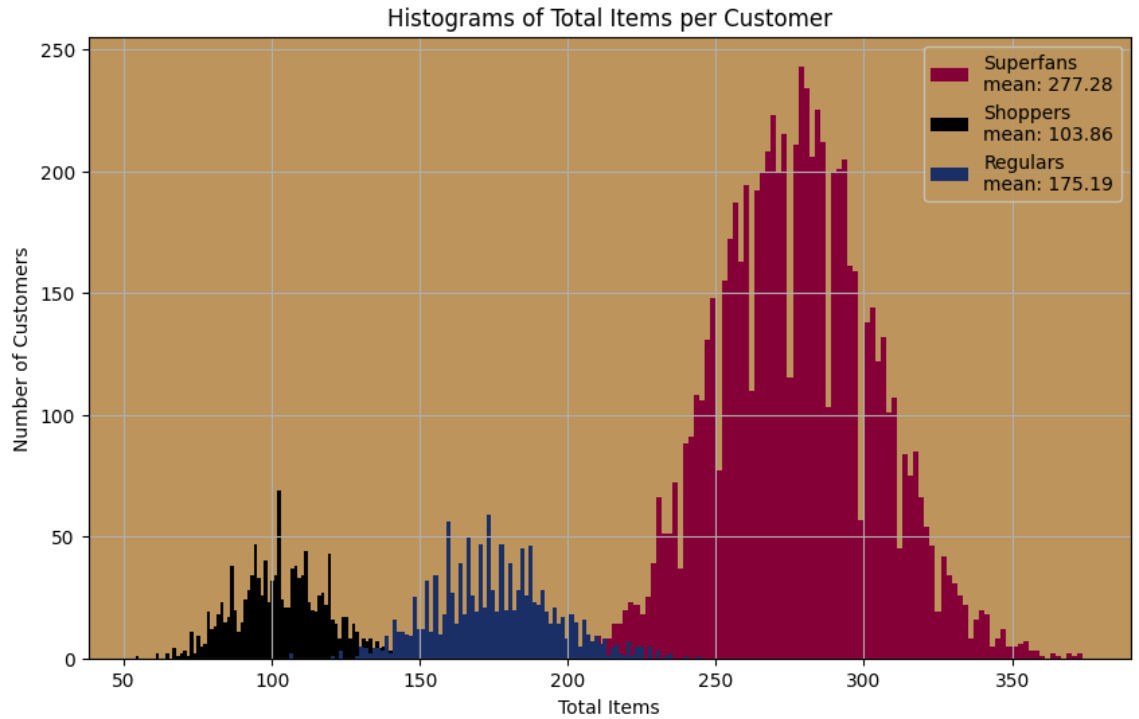
Results: Fan Segments

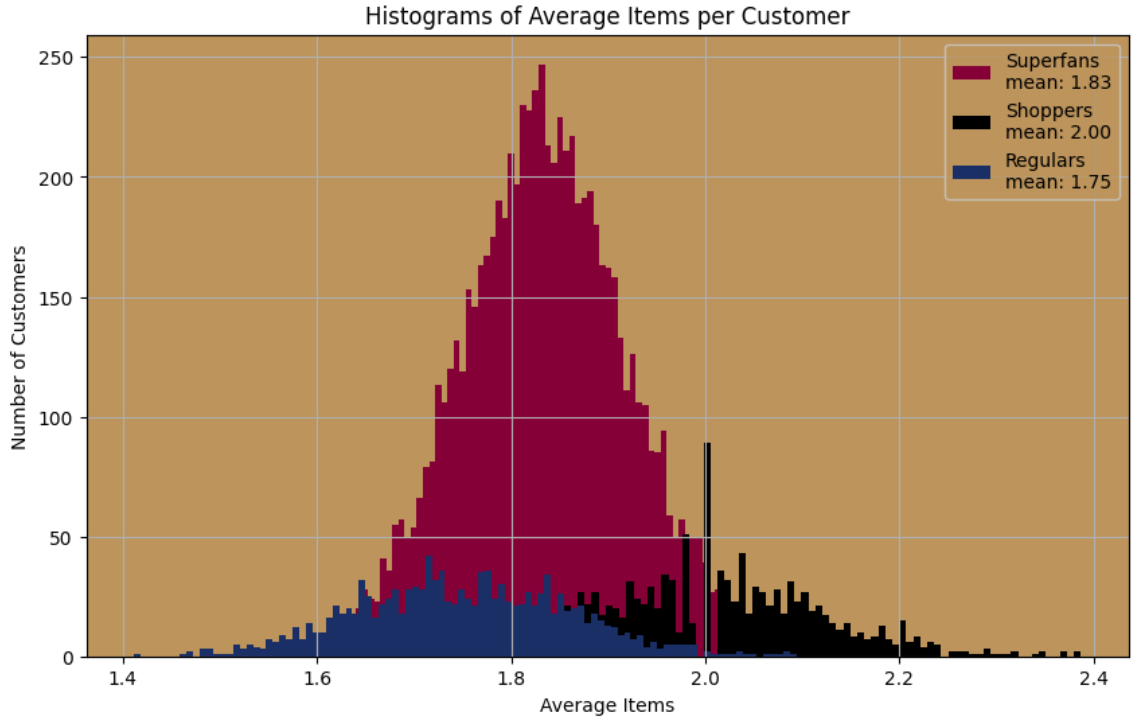
- 7,500/10,000: Superfans
  - ~$6,430 per customer
  - High attendance (~38/39 games)
- 1,250/10,000: Shoppers
  - ~$5,160 per customer
  - No food & beverage; mostly online
  - Low attendance (~13/39 games)
- 1,250/10,000: Regulars
  - ~$1,360 per customer
  - No retail
  - High attendance (~36/39 games)




Histograms of Total Amount Spent per Customer
Superfans
mean: $6,433.29
Shoppers
mean: $5,167.93
Regulars
mean: $1,360.11
Number of Customers
Total Amount Spent
0
50
100
150
200
250
2000
4000
6000
8000
10000


Histograms of Average Amount Spent per Customer
Superfans
mean: $42.56
Shoppers
mean: $99.46
Regulars
mean: $13.60
Number of Customers
Average Amount Spent
0
50
100
150
200
20
40
60
80
100
120
140


Histograms of Total Items per Customer
Superfans
mean: 277.28
Shoppers
mean: 103.86
Regulars
mean: 175.19
Number of Customers
Total Items
0
50
100
150
200
250
50
100
150
200
250
300
350

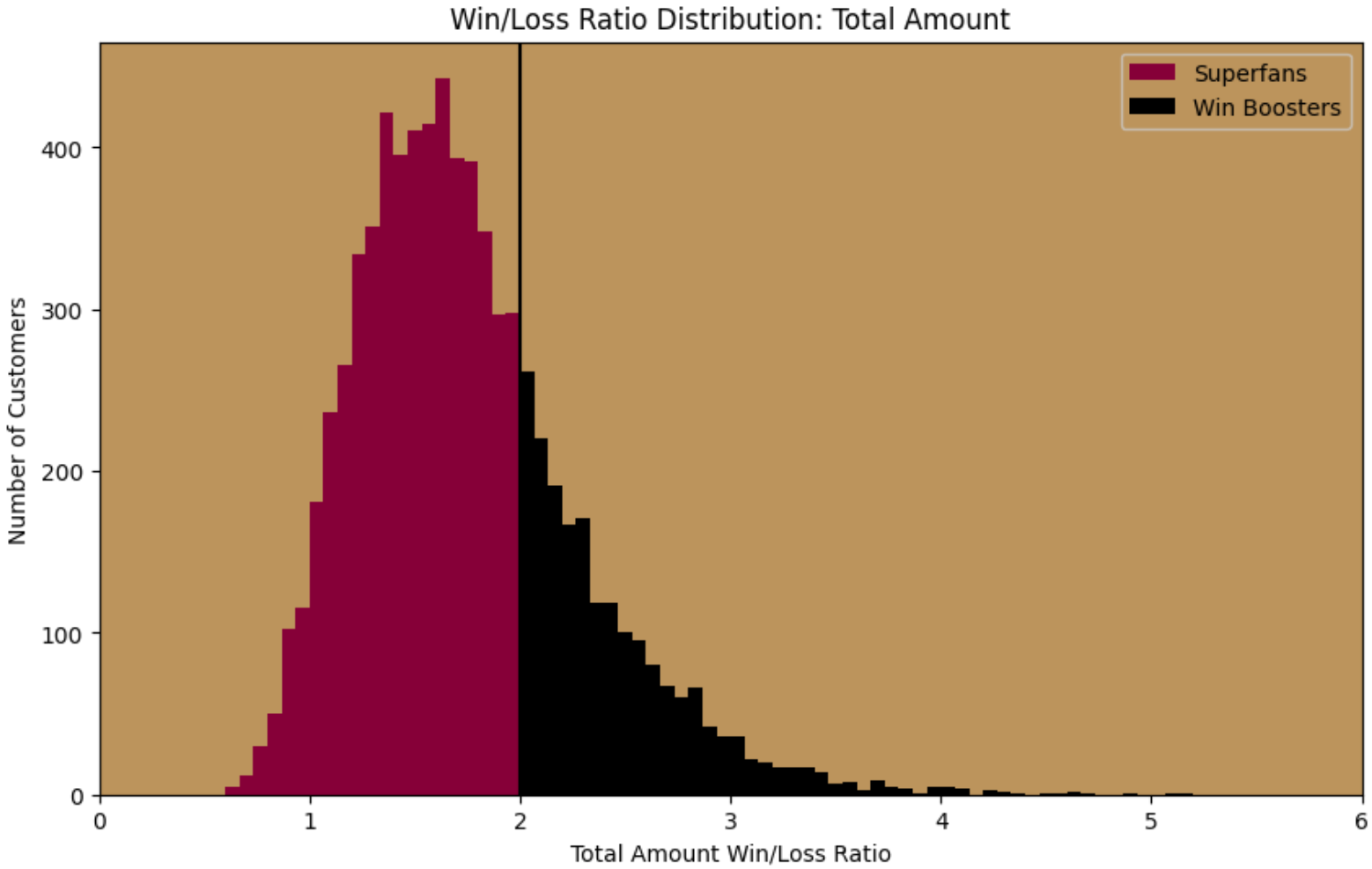

Histograms of Average Items per Customer
Superfans
mean: 1.83
Shoppers
mean: 2.00
Regulars
mean: 1.75
Number of Customers
Average Items
0
50
100
150
200
250
1.4
1.6
1.8
2.0
2.2
2.4

Fan Segment Analysis: Opportunities

- Retail brings in most of the revenue
- Regulars attend many games, but never buy retail
  - Incentivize them with retail discounts for attendance
- Regulars and Superfans buy lots of F&B
  - Incentivize them with retail discounts for F&B purchases
- Shoppers don’t attend many games
  - Incentivize them with ticket discounts for retail purchases

Win Boosters

- Subgroup of Superfans
  - 2,004/7,500
- 2x or more spent when the Cavs win



Other Methods Tested

- Derived Data
  - Day of the week
  - Opponent
  - Box score data
- Correlation matrix

- Clustering methods
  - Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
  - Hierarchical (agglometerative) clustering
- Choosing k
  - Silhouette scores
- Dimensionality reduction
  - t-distributed Stochastic Neighbor Embedding (t-SNE)

References & Software

- Box score data: NBA.com/stats
- Programming language: Python
- Data loading: pandas
- Data processing: pandas + scikit-learn
- Analysis: scikit-learn
- Visualization: matplotlib

Thank You