Stage 1 Algorithm Explanation:

In the initial phase of customer segmentation analysis, we employed a comprehensive methodology combining the **RFM model** with **K-means** clustering. First, within the RFM framework, we established three core customer value dimensions: Recency measures customer activity, Frequency assesses customer loyalty, and Monetary reflects customer contribution value. Through quantitative calculations across these dimensions, we generate multidimensional feature vectors for each customer.

Subsequently, we applied the unsupervised K-means machine learning algorithm to cluster these RFM features. The elbow rule determined the optimal number of clusters as **k=4**, striking a balance between model complexity and interpretability. The clustering results naturally segmented customers into four distinct groups with markedly different RFM characteristics: high-value customers, regular customers, at-risk customers, and new/low-activity customers.

This combined RFM+K-means approach enables us to transcend simplistic demographic categorization, achieving precise customer segmentation based on actual purchasing behavior. This provides a data-driven foundation for subsequent personalized marketing strategy formulation.

Algorithm Inputs and Outputs

Inputs:

1. customers_2.csv - Customer basic information data
2. products_2.csv - Product catalog and pricing information
3. sales_2.csv - Sales transaction records

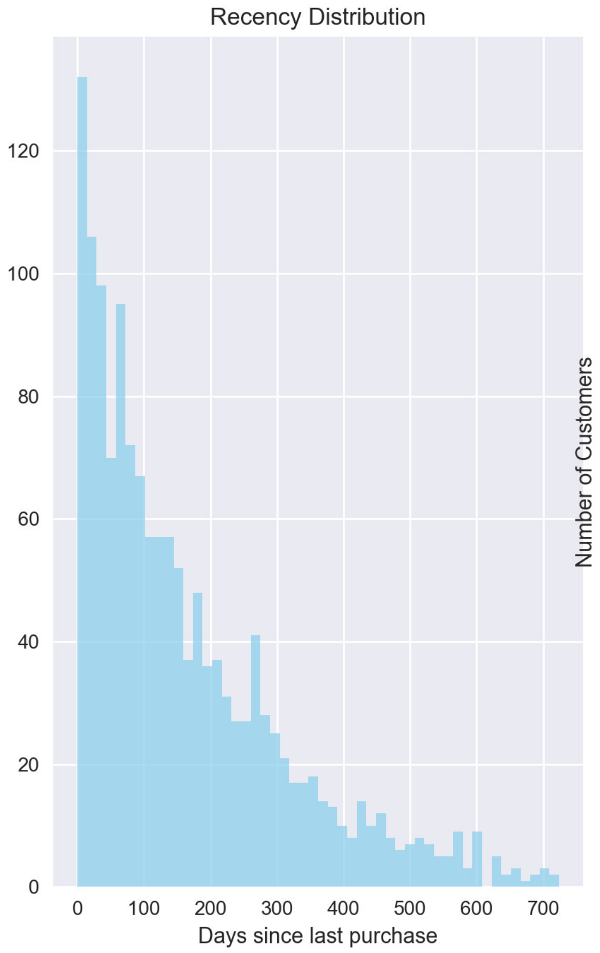
Outputs:

1. Customer Segmentation Results (s1_customer_segmentation_results.csv)
2. RFM Analysis Visualizations (direct display)
3. Customer Segment Distribution (direct display)
4. Business Recommendations Report (direct display)

Detailed Explanation:

1. Rency Distribution

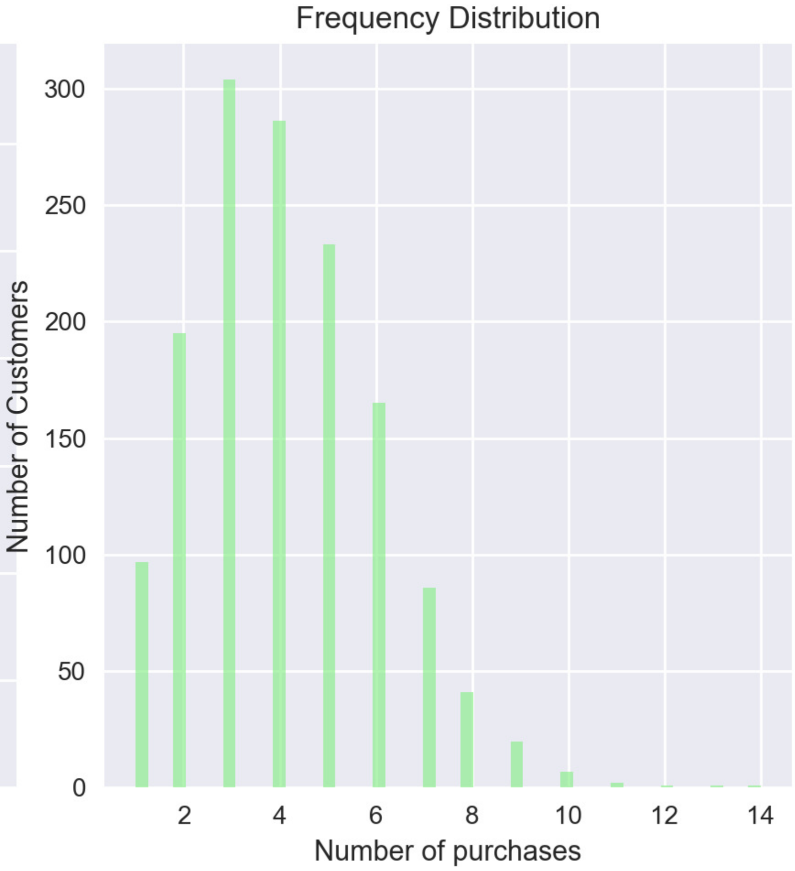
- What it does:
  - This histogram visualizes how recently customers have made their last purchase.
  - Recency is calculated as the number of days since a customer's most recent purchase.
  - It helps identify customer engagement levels. A lower recency values means a more active customer while vice versa, higher recency means a more churning customer.
- Key data analysed:
  - **Count**: 1,439 customers.
  - **Mean Recency**: 165.95 days.
  - **Standard Deviation**: 151.22 days.
  - **Minimum**: 0 days
  - **25th Percentile**: 48 days.
  - **Median (50th Percentile)**: 123 days.
  - **75th Percentile**: 246.5 days.
  - **Maximum**: 724 days.
  - Binned Distribution (10 bins for overview):
    - [-0.725, 72.4]: 501 customers
    - [72.4, 144.8]: 310 customers
    - [144.8, 217.2]: 210 customers
    - [217.2, 289.6]: 154 customers
    - [289.6, 362.0]: 100 customers
    - [362.0, 434.4]: 57 customers
    - [434.4, 506.8]: 43 customers
    - [506.8, 579.2]: 34 customers
    - [579.2, 651.6]: 19 customers
    - [651.6, 724.0]: 11 customers



- Trend we observed:
  - Distribution is right-skewed with a peak in the lowest bin (containing 501 customers total)
  - Followed by a decline as recency increases
  - Strong recent engagement from a large number of customers

2. Frequency Distribution Graph

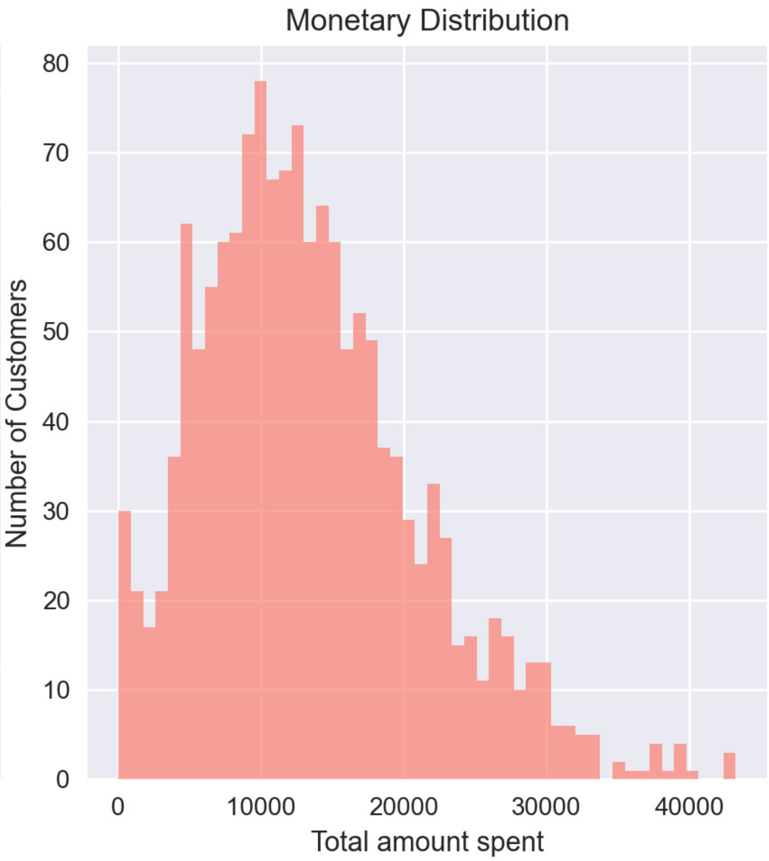
- What it does:
    - Depicting how often customers make purchases
    - Frequency is the number of unique invoices per customer
    - Highlights customer loyalty and their behaviour. With higher frequencies, it points to loyal and repeat buyers. While vice versa, with low frequencies, it points to infrequent buyers.
- Key data analyzed:
    - **Count**: 1,439 customers.
    - **Mean Frequency**: 4.13 purchases.
    - **Standard Deviation**: 1.93 purchases.
    - **Minimum**: 1 purchase.
    - **25th Percentile**: 3 purchases.
    - **50th Percentile**: 4 purchases.
    - **75th Percentile**: 5 purchases.
    - **Maximum**: 14 purchases.
    - Exact Value Counts:
        - 1 purchase: 97 customers
        - 2 purchases: 195 customers
        - 3 purchases: 304 customers
        - 4 purchases: 286 customers
        - 5 purchases: 233 customers
        - 6 purchases: 165 customers
        - 7 purchases: 86 customers
        - 8 purchases: 41 customers
        - 9 purchases: 20 customers
        - 10 purchases: 7 customers
        - 11 purchases: 2 customers
        - 12 purchases: 1 customer
        - 13 purchases: 1 customer
        - 14 purchases: 1 customer



- Trend we observed:
    - Distribution is right-skewed with a peak around 3-5 purchases
    - Shows most customers are moderately frequent buyers, every time buy a few stuffs only.
    - Pointing to potential for programs to boost repeat purchases from customers.

3. ***Monetary Distribution Graph***

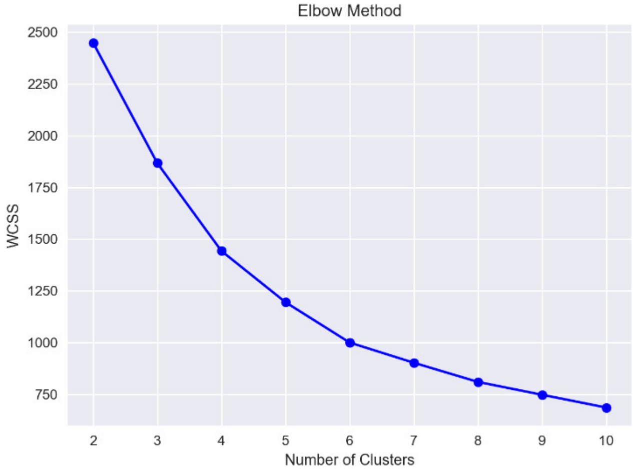
- What it does:
  - Illustrating the total amount spent by customers
  - Monetary value is the sum of prices from all products purchased
  - Assesses customer value: with higher monetary value, it identifies high-spenders. While vice versa, lower monetary value identifies budget-conscious or low-engagement customers.
- Key data analysed:
  - **Count**: 1,439 customers.
  - **Mean Monetary**: 13,469.37
  - **Standard Deviation**: 7,621.49.
  - **Minimum**: 55.15.
  - **25th Percentile**: 7,973.92.
  - **Median (50th Percentile)**: 12,451.86.
  - **75th Percentile**: 17,732.47.
  - **Maximum**: 43,257.97.
  - Binned Distribution (10 bins):
    - (11.946, 4,375.432]: 125
    - (4,375.432, 8,695.714]: 286
    - (8,695.714, 13,015.996]: 358
    - (13,015.996, 17,336.278]: 284
    - (17,336.278, 21,656.56]: 175
    - (21,656.56, 25,976.842]: 102
    - (25,976.842, 30,297.124]: 70
    - (30,297.124, 34,617.406]: 22
    - (34,617.406, 38,937.688]: 9
    - (38,937.688, 43,257.97]: 8



- Trend we observed:
  - Distribution is right-skewed with peaks in mid-lower bins
  - Highlights that revenue is driven by a broad base of moderate spenders,
  - With a small group of high-value customers contributing disproportionately
  - Suggesting targeted upselling to increase overall spend.

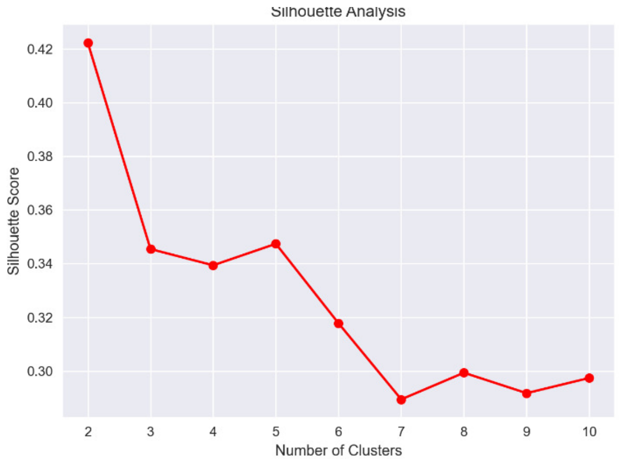
4. ***Elbow Method Graph***

- What it does:
  - determine the optimal number of clusters in K-Means clustering
  - The "elbow" point, where the rate of WCSS decrease slows significantly, indicates the ideal k, balancing underfitting and overfitting.
- Key data analysed:
  - The elbow point is at k=4, where the drop in WCSS begins to flatten (decrease from 1750 at k=4 to 1500 at k=5 is less steep than prior drops).
- Trend we observed:
  - suggesting diminishing returns beyond 4 clusters; this supports using k=4 for meaningful segmentation without excessive complexity.



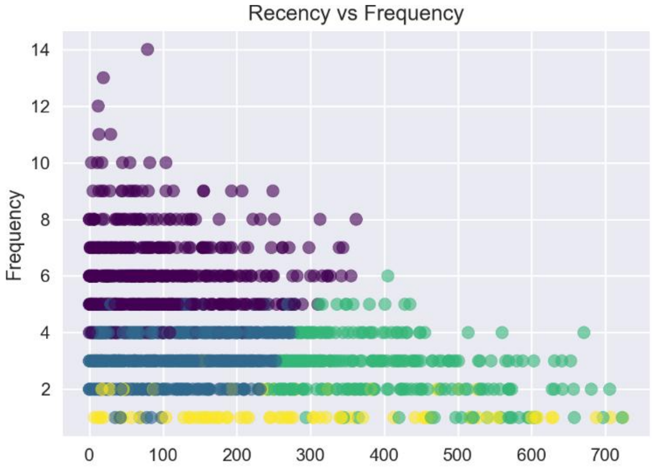
5. ***Silhouette Analysis Graph***

- What it does:
  - evaluating cluster quality
  - displays the average Silhouette Score
  - Higher scores indicate better-defined clusters
- Key data analysis:
  - Highest score at k=2 (0.42), but k=4 (0.34) is selected as a balance with elbow method.
- Trend we observe:
  - k=4 offers a practical compromise for interpretability without significant quality loss.



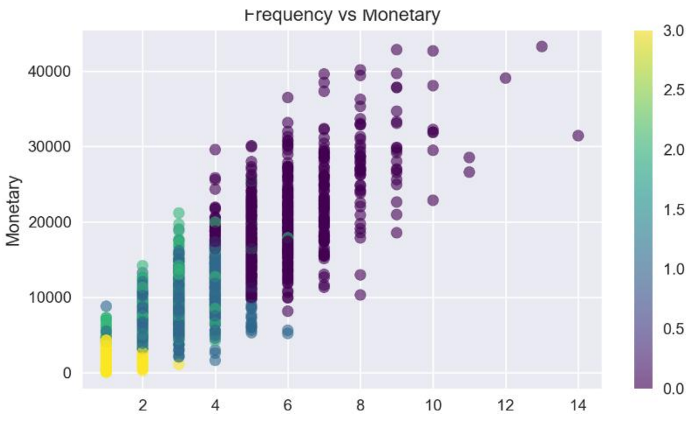
6. ***Recency VS Frequency Scatter Plot***

- What it does:
  - visualizes customer distribution with Recency versus Frequency
- Key data analysis:
  - Correlation: Negative overall
  - (lower Recency higher Frequency)
- Trend we observe:
  - Negative correlation shows recent buyers purchase more often
  - with outliers in high Frequency despite moderate Recency, suggesting loyalty persists in some cases.



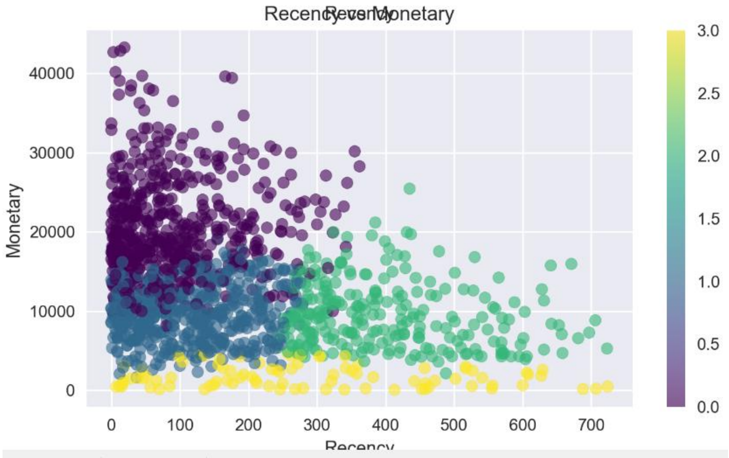
7. ***Frequency VS Monetary Scatter Plot***

- What it does:
    - shows Frequency VS Monetary
    - reveals how purchase count drives spending
- Key data analysis
    - Correlation: Positive
    - (higher Frequency higher Monetary).
- Trend we observe:
    - indicating repeat purchases significantly boost spending, but diminishing returns beyond Frequency 10



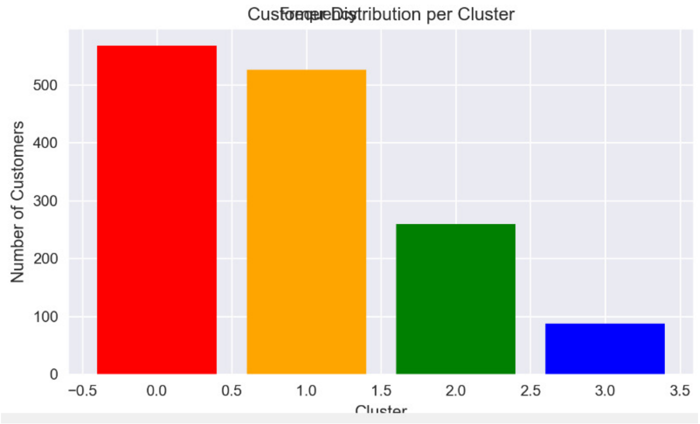
8. Recency VS Monetary Scatter Plot

- What it does:
    - Shows recency VS monetary
    - illustrates how recent activity relates to spending
- key data analysis:
    - Correlation: Negative
    - (lower Recency higher Monetary).
- Trend we observe:
    - highlighting potential revenue loss from churn in the upper right quadrant

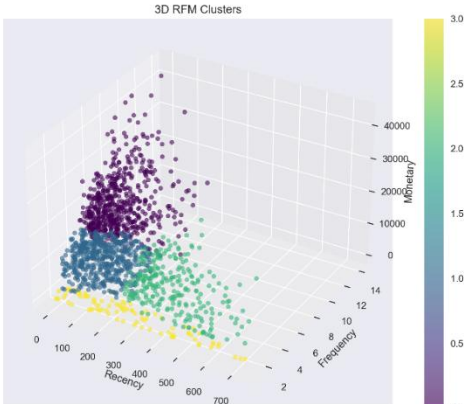


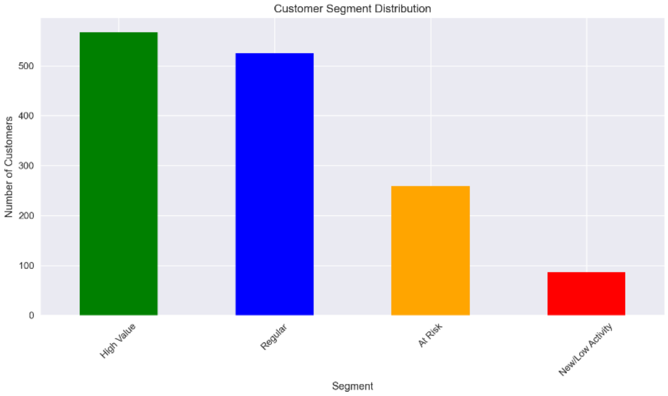
9. Customer Size Distribution per Cluster Bar Chart

- What it does:
    - Displays the number of customers per cluster
    - Quantifies segment sizes post-clustering
- Key data analysis:
    - Customer Counts: High Value: 567, Regular: 525, At Risk: 259, New/Low Activity: 88
- Trend we observe:
    - uneven distribution emphasizes focus on retaining large valuable segments while nurturing smaller ones for growth.



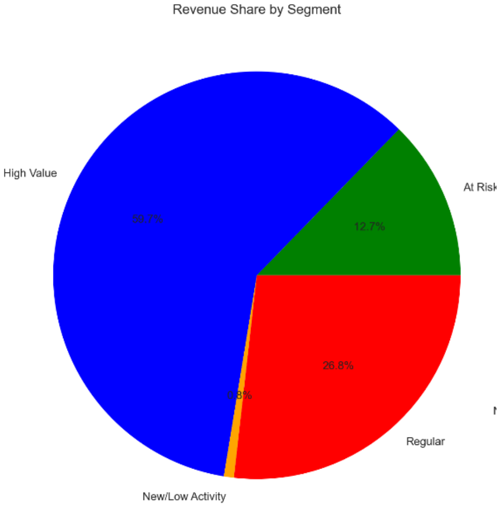
10. 3D RFM Clusters Scatter Plot





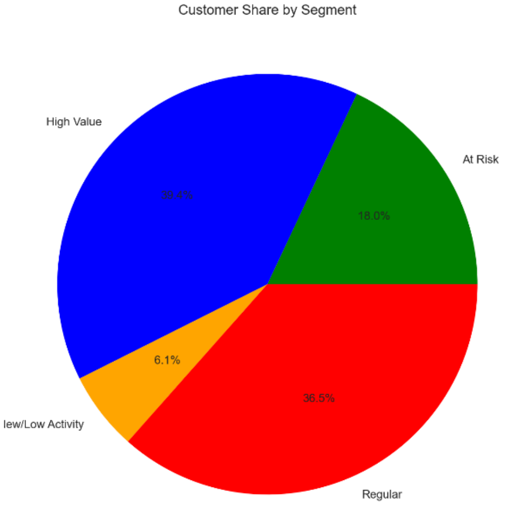
- What it does:
  - visualizes customers in RFM space: Recency, Frequency, Monetary
  - provides a multi-dimensional view of cluster separation
- Key data analysis:
  - Cluster centres (approx.): High Value (low Recency ~50, high Frequency ~6.5, high Monetary ~17,333),
  - Regular (mid ~120/4/8,160), At Risk (high ~300/5/19,756), New/Low (~200/2/1,762).
- Trend we observed:
  - Clusters separate along axes: purple (high value) in low Recency/high Frequency/high Monetary front; green/cyan mid; yellow (low) in high Recency/low Frequency/low Monetary back

11. Revenue Share by Segment Pie Chart



- What it does:
  - Shows percentage of total revenue contributed by each segment
- key data analysis:
  - Revenue Shares: High Value: 50.7%, At Risk: 26.4%, Regular: 22.1%, New/Low: 0.8%
- Trend we observe:
  - Dominance by High Value

12. Customer Share by Segment Pie Chart



- What it does:
  - depicts the percentage of total customers in each segment
  - shows demographic distribution for targeted marketing.
- Key data analysis:
  - Customer Shares: High Value: 39.4%, Regular: 36.5%, At Risk: 18.0%, New/Low: 6.1%
- Trend we observe:
  - Near-equal large shares for High Value and Regular