MASTER OF SCIENCE OF INFORMATION SYSTEMS
INTELLIGENT SYSTEMS

CLUSTERING

UCI Wholesale Customers Dataset

K-means vs DBSCAN

STUDENTS (GROUP2)	AINEDEMBE DENIS	2024-M132-23999
	MUSINGUZI BENSON	2024-M132-23947
LECTURER	Dr. Sibitenda Harriet	

Objective

- A comprehensive clustering analysis comparing K-means and DBSCAN algorithms on the UCI Wholesale Customers Dataset.

- This activity performs customer segmentation to identify distinct purchasing patterns and provide actionable business insights.

Dataset

The dataset is about clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on different product categories:

- Fresh, Milk,
- Grocery, Frozen,
- Detergents_Paper,
- Delicassen;

And then

- Channel/Region information.

Loaded the dataset

```
Dataset loaded successfully!

Dataset shape: (440, 8)

First 10 rows:
   Channel  Region  Fresh   Milk  Grocery  Frozen  Detergents_Paper  \
0        2       3  12669   9656     7561     214              2674
1        2       3   7057   9810     9568    1762              3293
2        2       3   6353   8808     7684    2405              3516
3        1       3  13265   1196     4221    6404               507
4        2       3  22615   5410     7198    3915              1777
5        2       3   9413   8259     5126     666              1795
6        2       3  12126   3199     6975     480              3140
7        2       3   7579   4956     9426    1669              3321
8        1       3   5963   3648     6192     425              1716
9        2       3   6006  11093    18881    1159              7425

   Delicassen
0        1338
1        1776
2        7844
3        1788
4        5185
5        1451
6         545
7        2566
8         750
9        2098
```

Showing basic Statistics

```
Basic statistics:
          Channel      Region           Fresh            Milk       Grocery  \
count  440.000000  440.000000      440.000000      440.000000    440.000000
mean     1.322727    2.543182    12000.297727     5796.265909   7951.277273
std      0.468052    0.774272    12647.328865     7380.377175   9503.162829
min      1.000000    1.000000        3.000000       55.000000      3.000000
25%      1.000000    2.000000     3127.750000     1533.000000   2153.000000
50%      1.000000    3.000000     8504.000000     3627.000000   4755.500000
75%      2.000000    3.000000    16933.750000     7190.250000  10655.750000
max      2.000000    3.000000   112151.000000    73498.000000  92780.000000

             Frozen  Detergents_Paper    Delicassen
count    440.000000        440.000000    440.000000
mean    3071.931818       2881.493182   1524.870455
std     4854.673333       4767.854448   2820.105937
min       25.000000          3.000000      3.000000
25%      742.250000        256.750000    408.250000
50%     1526.000000        816.500000    965.500000
75%     3554.250000       3922.000000   1820.250000
max    60869.000000      40827.000000  47943.000000
```

Handle missing data & Removing Duplicates

Handle missing data

```
Missing values per column:
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicassen          0
dtype: int64

Total missing values: 0

No missing values found. No imputation needed.
```

Removing exact duplicates

```
Initial dataset shape: (440, 8)
After removing duplicates: (440, 8)
Duplicates removed: 0

Final dataset dimensions: 440 rows × 8 columns
```

Outlier detection using IQR method

```
Outlier detection using IQR method (values beyond 1.5*IQR):
Fresh: 20 outliers (4.55%)
Milk: 28 outliers (6.36%)
Grocery: 24 outliers (5.45%)
Frozen: 43 outliers (9.77%)
Detergents_Paper: 30 outliers (6.82%)
Delicassen: 27 outliers (6.14%)

Numeric features scaled using RobustScaler (robust to outliers)

Scaled data sample:
      Fresh      Milk   Grocery    Frozen  Detergents_Paper  Delicassen
0  0.301680  1.065712  0.329952 -0.466572          0.506787    0.263810
1 -0.104810  1.092934  0.565993  0.083926          0.675670    0.574008
2 -0.155802  0.915816  0.344418  0.312589          0.736512    4.871459
3  0.344850 -0.429714 -0.062862  1.734708         -0.084442    0.582507
4  1.022092  0.315171  0.287260  0.849573          0.262056    2.988314
```

Encode Channel and Region for interpretation

```
Channel encoding:
Channel
1    298
2    142
Name: count, dtype: int64

Region encoding:
Region
1     77
2     47
3    316
Name: count, dtype: int64

Channel and Region encoded for interpretation only
Note: These categorical features are NOT included in clustering distance calculations
```

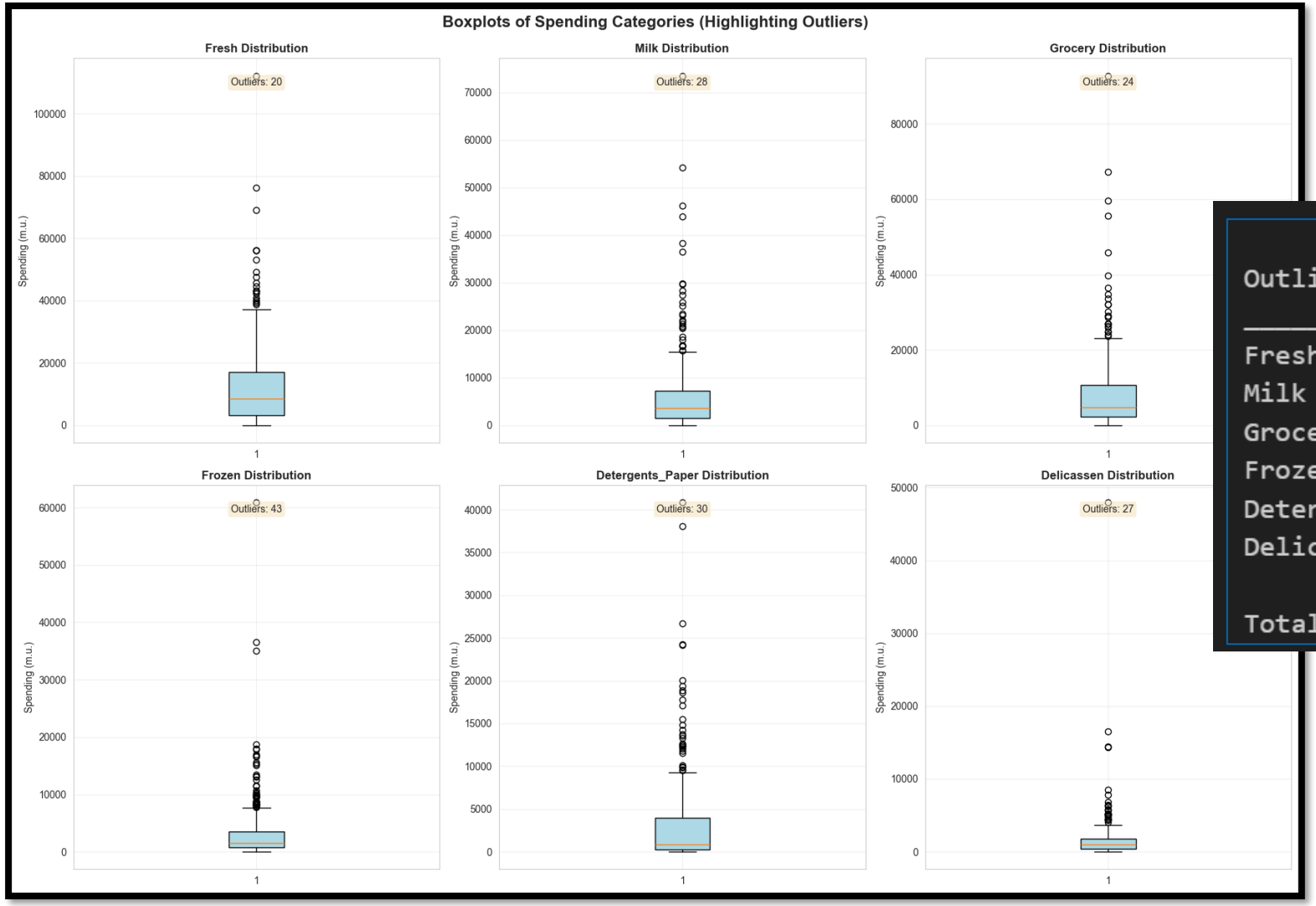
Descriptive statistics for spending categories

Descriptive Statistics for Spending Categories:

	Fresh	Milk	Grocery	Frozen	\
count	440.000000	440.000000	440.000000	440.000000	
mean	12000.297727	5796.265909	7951.277273	3071.931818	
std	12647.328865	7380.377175	9503.162829	4854.673333	
min	3.000000	55.000000	3.000000	25.000000	
25%	3127.750000	1533.000000	2153.000000	742.250000	
50%	8504.000000	3627.000000	4755.500000	1526.000000	
75%	16933.750000	7190.250000	10655.750000	3554.250000	
max	112151.000000	73498.000000	92780.000000	60869.000000	

	Detergents_Paper	Delicassen
count	440.000000	440.000000
mean	2881.493182	1524.870455
std	4767.854448	2820.105937
min	3.000000	3.000000
25%	256.750000	408.250000
50%	816.500000	965.500000
75%	3922.000000	1820.250000
max	40827.000000	47943.000000

Spending categories. Outliers and giving counts




```
Outlier Summary:
_____________________________________________________
Fresh                :  20 outliers ( 4.55%)
Milk                 :  28 outliers ( 6.36%)
Grocery              :  24 outliers ( 5.45%)
Frozen               :  43 outliers ( 9.77%)
Detergents_Paper     :  30 outliers ( 6.82%)
Delicassen           :  27 outliers ( 6.14%)

Total outlier instances across all categories: 172
```

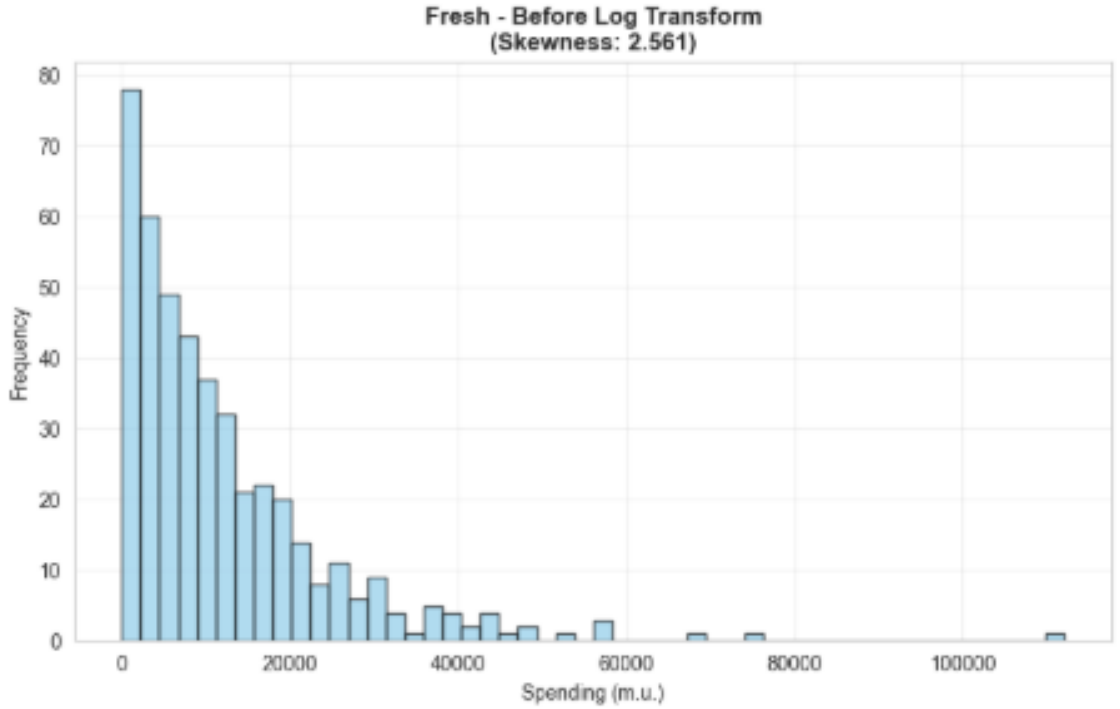
Log-transform highly skewed spending variables

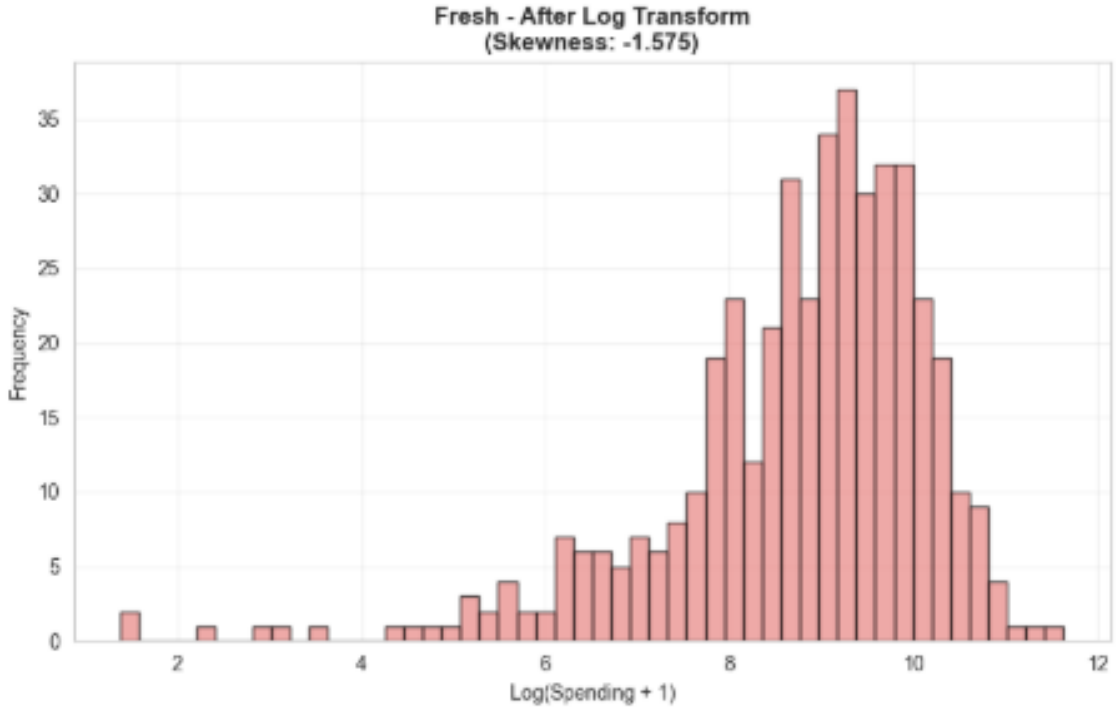
```
Skewness before log transformation:

_____________________________________
Fresh               :   2.561
Milk                :   4.054
Grocery             :   3.587
Frozen              :   5.908
Detergents_Paper    :   3.632
Delicassen          :  11.152

Highly skewed variables (|skewness| > 1): ['Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper', 'Delicassen']

Log-transformed variables: ['Log_Fresh', 'Log_Milk', 'Log_Grocery', 'Log_Frozen', 'Log_Detergents_Paper', 'Log_Delicassen']
```

Identify highly skewed variables (|skewness| > 1)

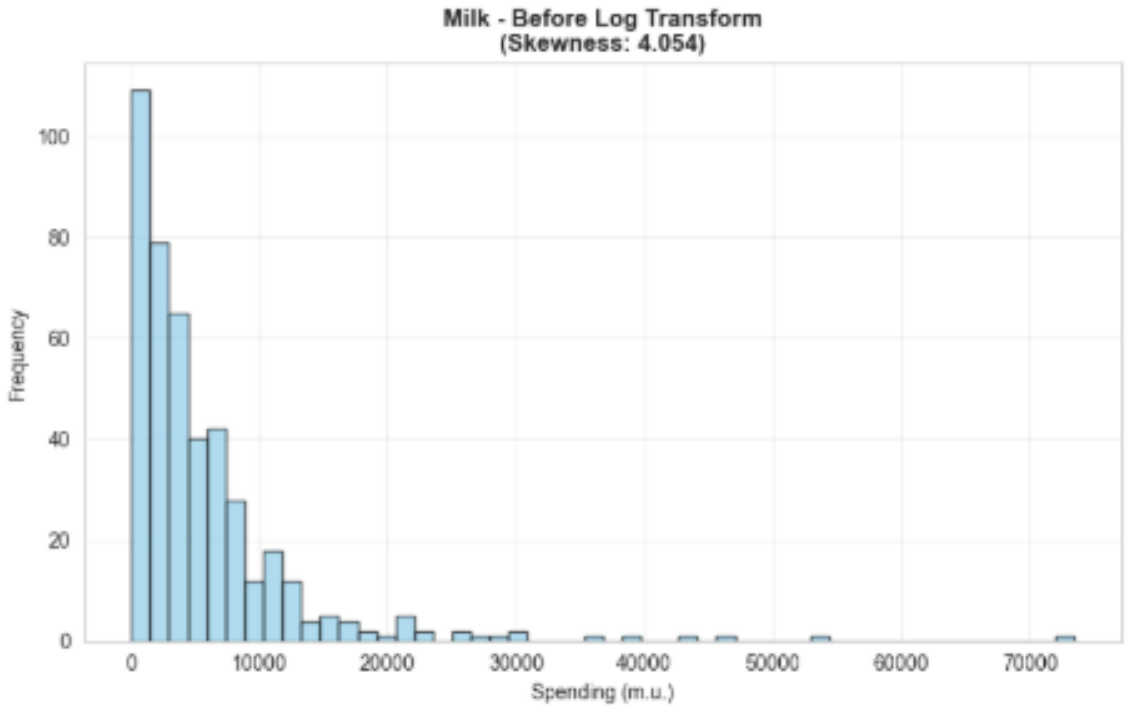
Log transform highly skewed variables

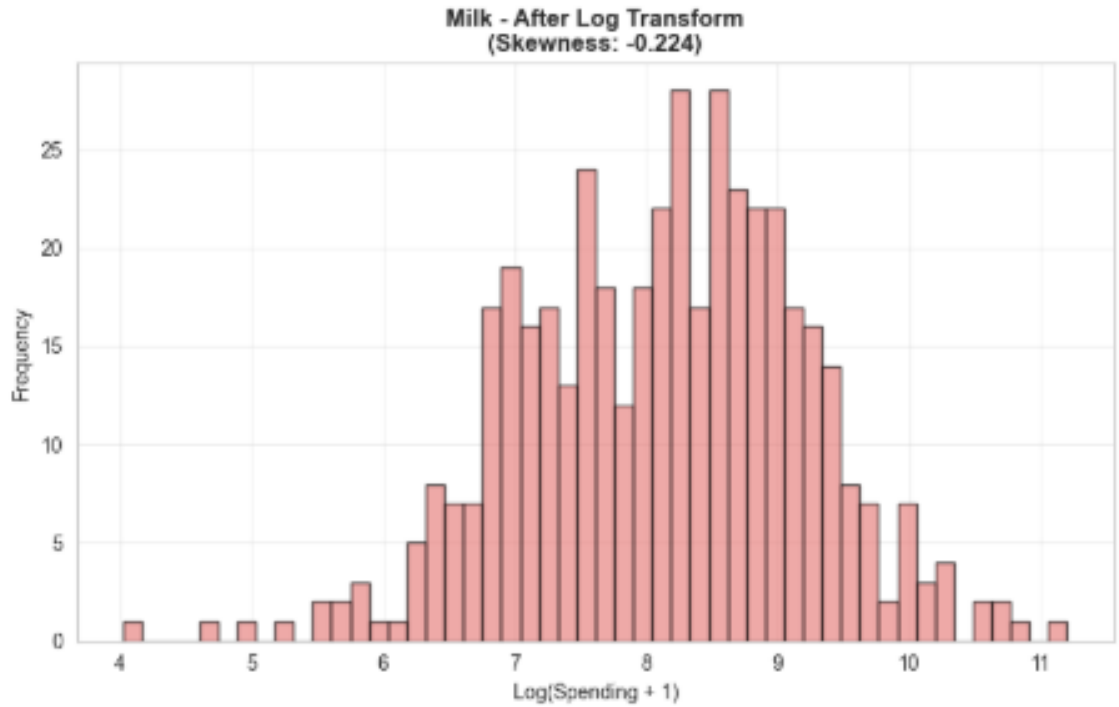
Before/ after histograms for log-transformed variables



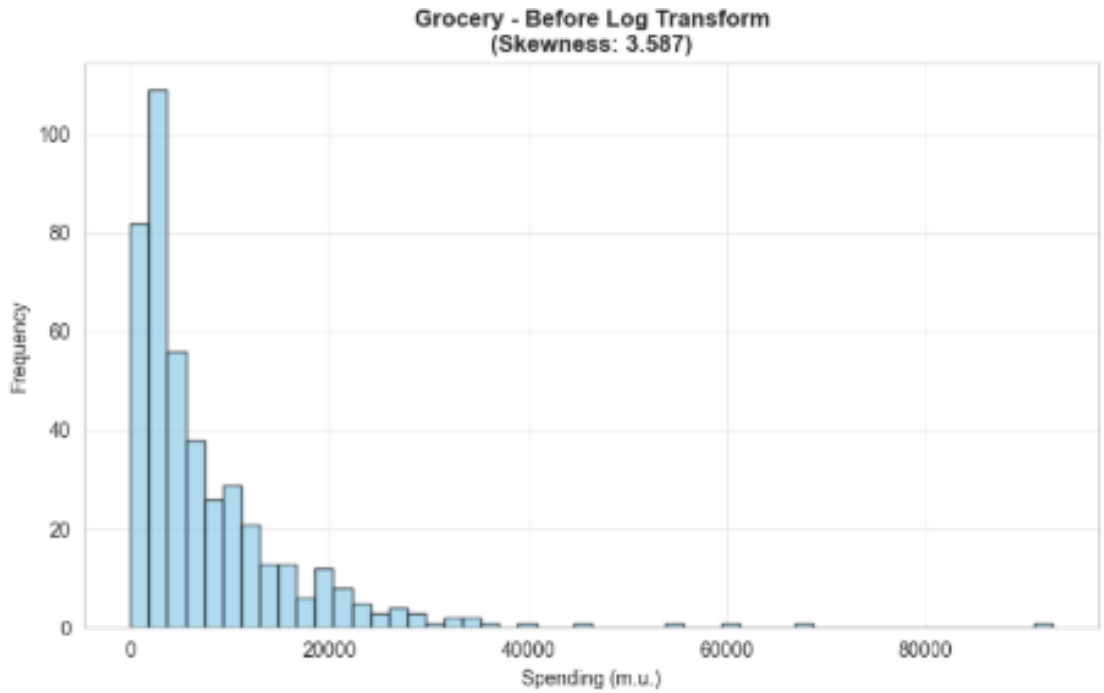


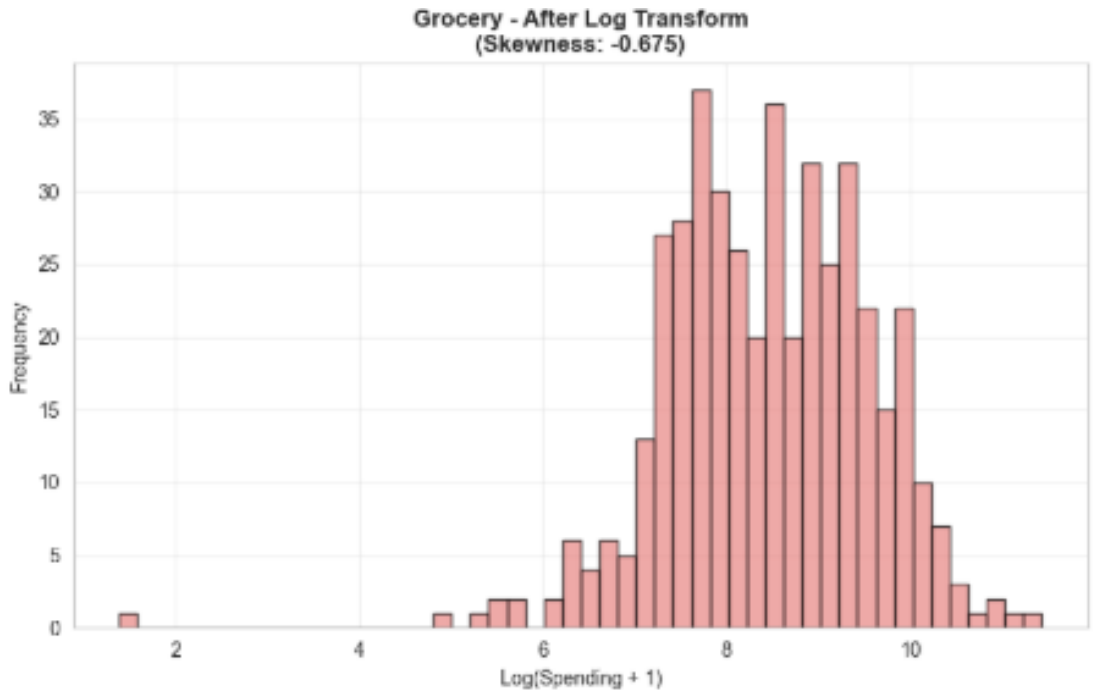
Before/ after histograms for log-transformed variables



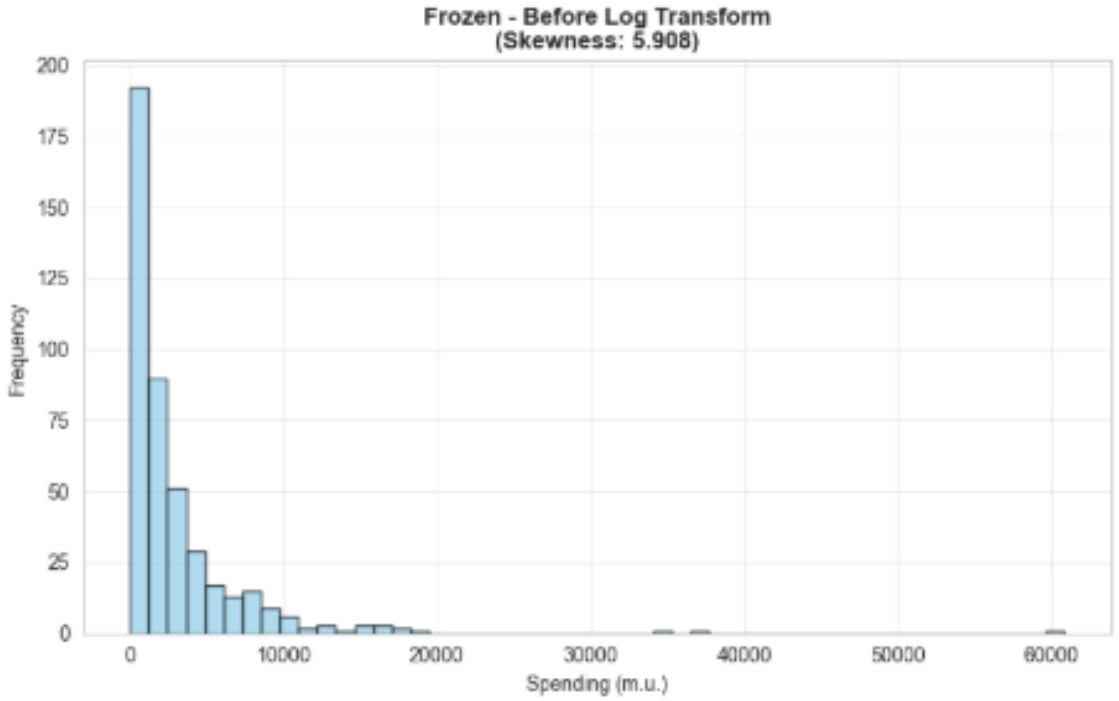


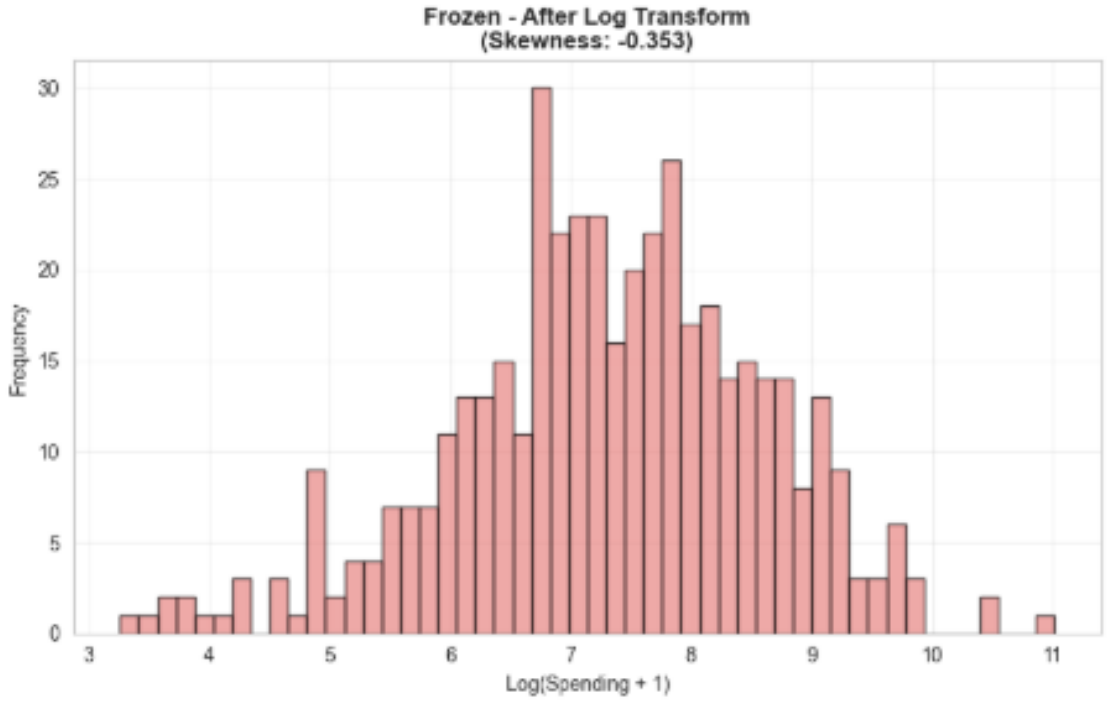
Before/ after histograms for log-transformed variables



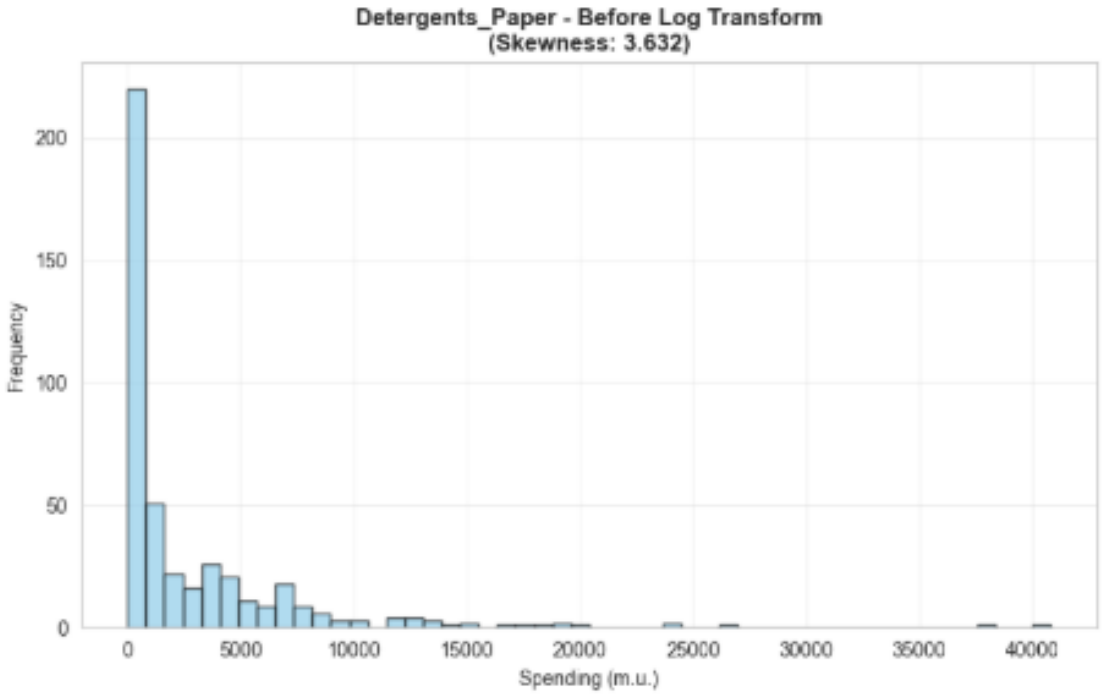


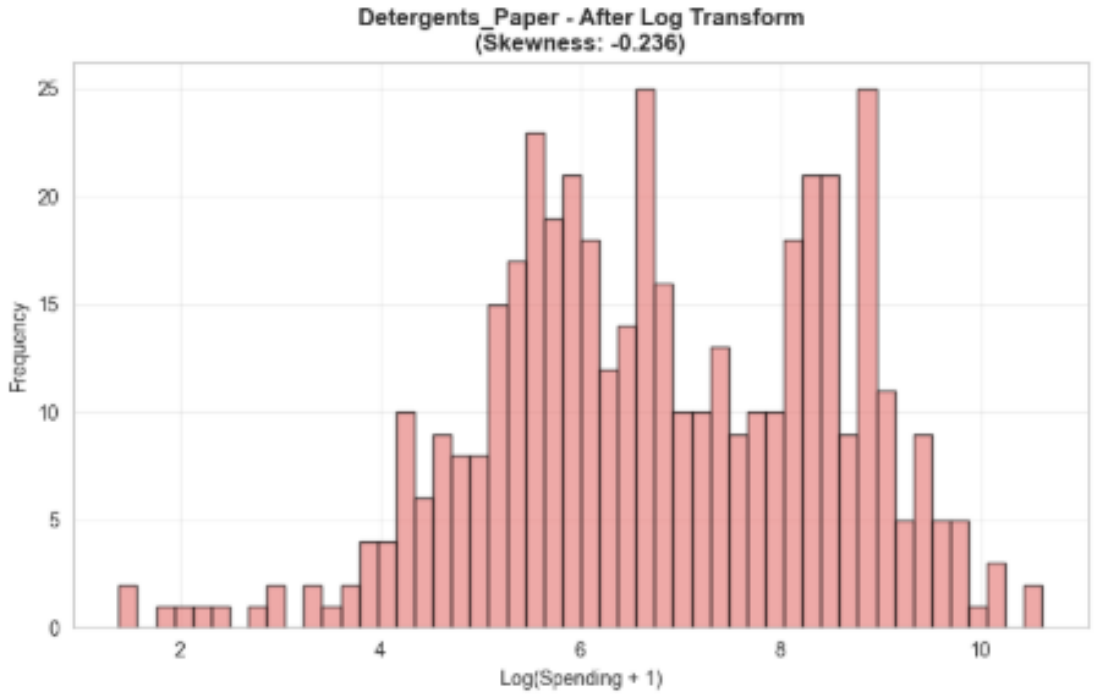
Before/ after histograms for log-transformed variables



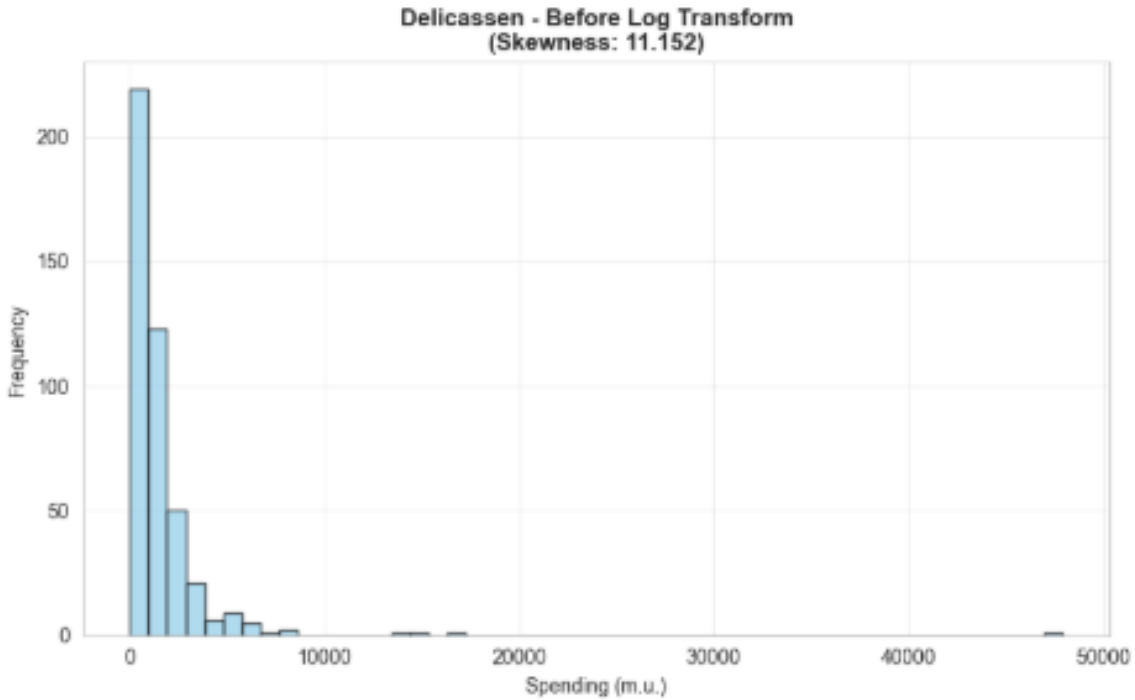


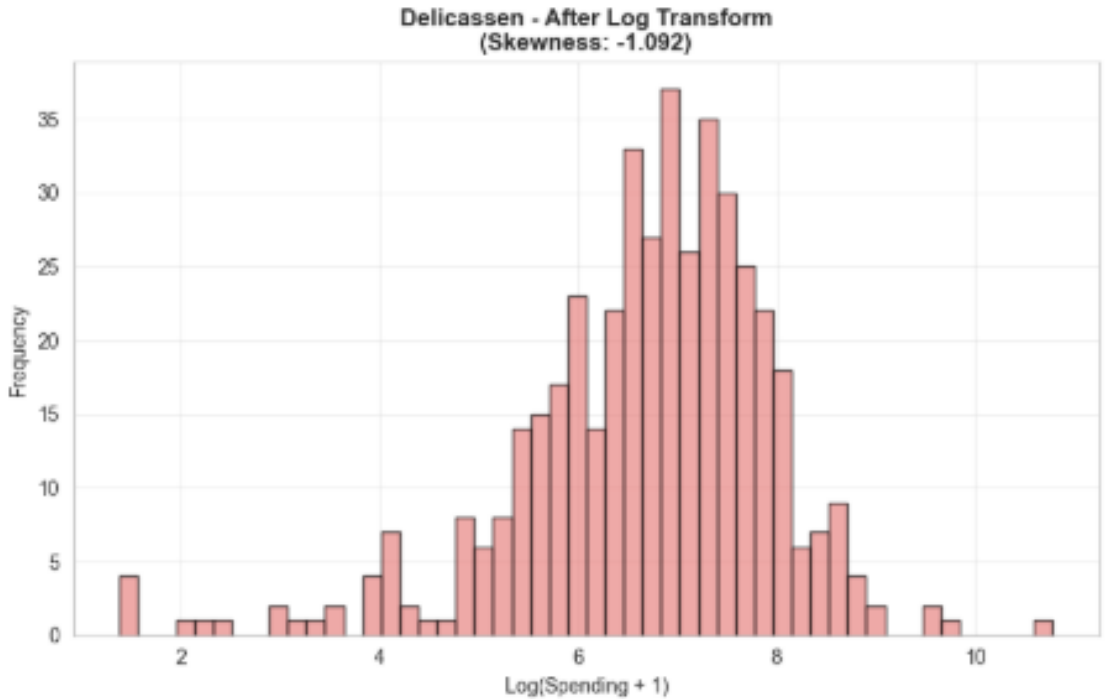
Before/ after histograms for log-transformed variables





Before/ after histograms for log-transformed variables

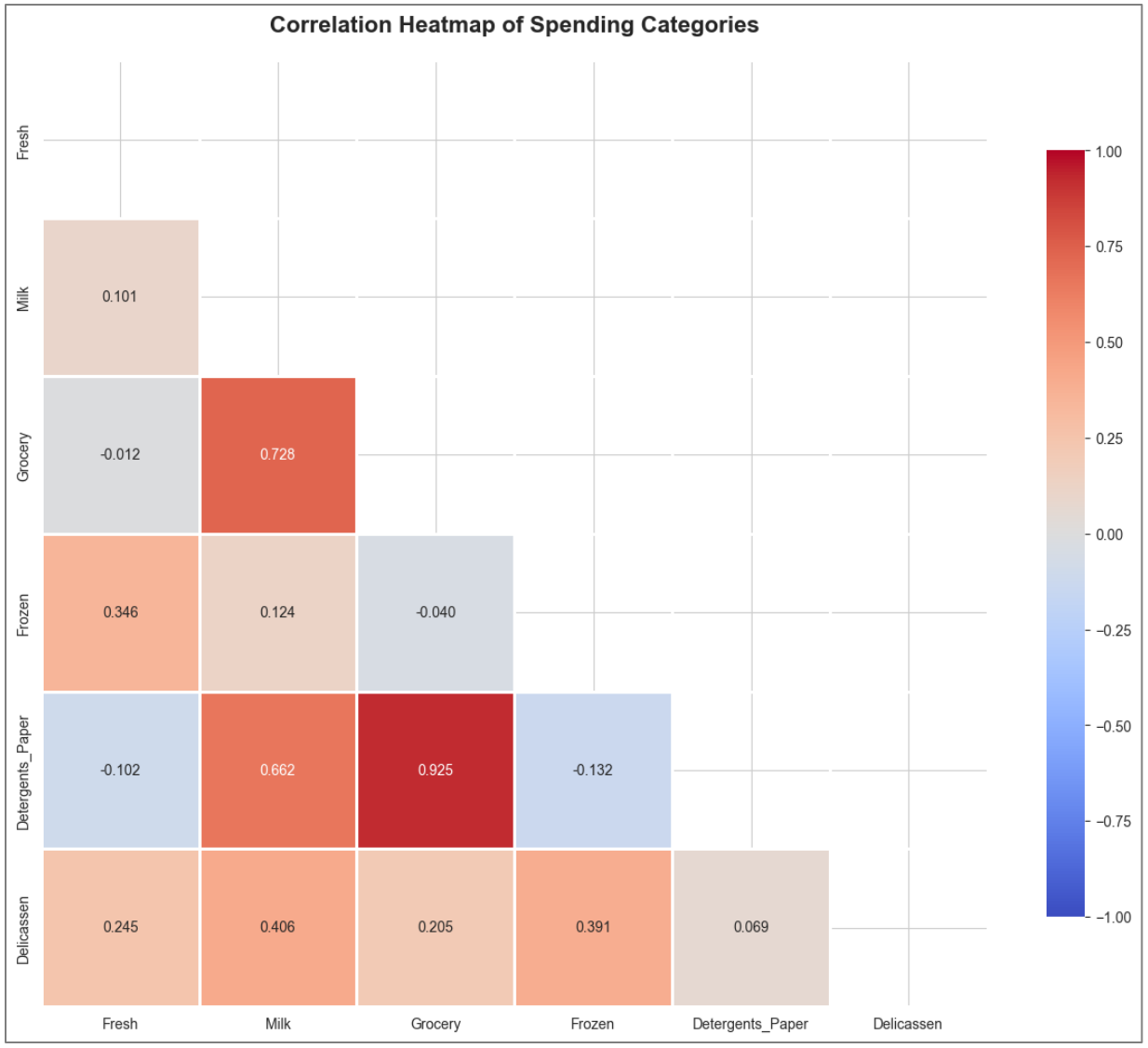




Before/ after histograms for log-transformed variables

```
Skewness after log transformation:

_____________________________________________

Log_Fresh                 :   -1.575
Log_Milk                  :   -0.224
Log_Grocery               :   -0.675
Log_Frozen                :   -0.353
Log_Detergents_Paper      :   -0.236
Log_Delicassen            :   -1.092
```

Correlation heatmap of spending categories




```
Co-purchasing Pattern Analysis:
_____________________________________________________________

Strong correlations (|r| > 0.5) indicate co-purchasing patterns:
  Grocery              ↔ Detergents_Paper    :  0.925
  Milk                 ↔ Grocery             :  0.728
  Milk                 ↔ Detergents_Paper    :  0.662
```

Insights:

- High correlation suggests customers who buy one category also buy the other

- This can inform cross-selling strategies and inventory management

Feature Engineering & Aggregation

```
Feature Engineering Summary:
_____________________________________________________

TotalSpend: sum of all spending categories
    Range: 904.00 - 199891.00 m.u.
    Mean: 33226.14 m.u.

ProportionFresh: Fresh / TotalSpend
    Range: 0.00 - 0.95
    Mean: 0.38

LogTotalSpend: log(1 + TotalSpend)
    Range: 6.81 - 12.21
    Mean: 10.17

GroceryMilkRatio: Grocery / (Milk + 1)
    Range: 0.00 - 21.02
    Mean: 1.82

NonFreshProportion: 1 - ProportionFresh
    Range: 0.05 - 1.00
    Mean: 0.62
```

Feature Engineering & Aggregation

Justification for Derived Features:

- TotalSpend: Captures overall customer value and spending capacity. Essential for segmenting high-value vs low-value customers.

- ProportionFresh: Reveals customer preference for fresh vs processed products. Helps identify customer segments with different product preferences.

- LogTotalSpend: Normalizes the highly skewed total spending distribution, making it more suitable for clustering algorithms that assume normal distributions.

- GroceryMilkRatio: Identifies customers who prefer grocery items over milk products, useful for product mix optimization.

- NonFreshProportion: Complements ProportionFresh by focusing on processed/frozen products, important for supply chain planning

K-means Clustering

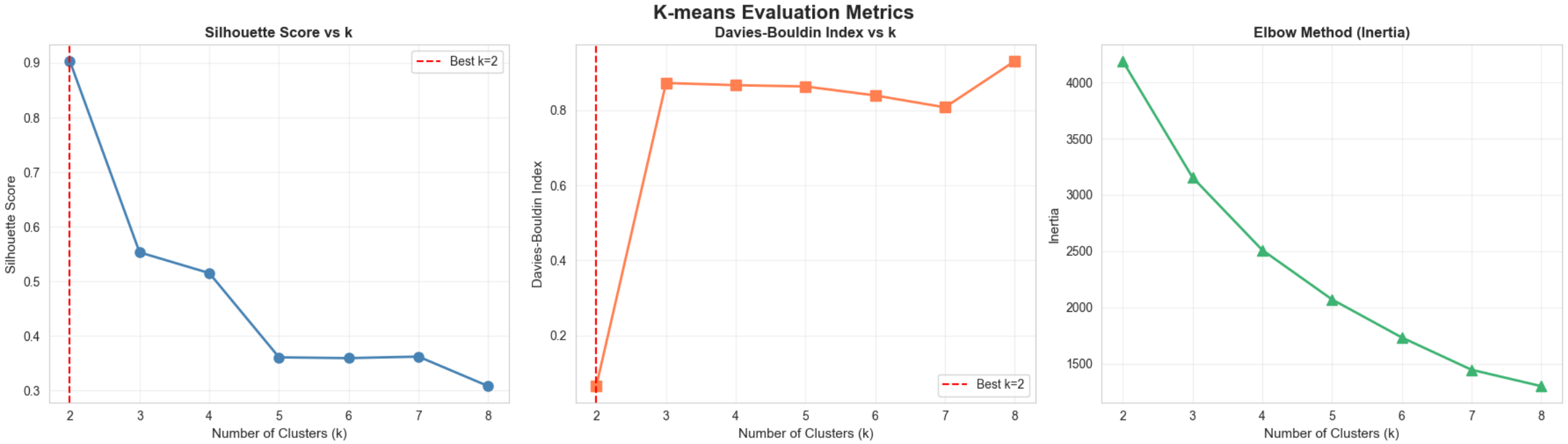
K-means Clustering Results:

k	Silhouette Score	Davies-Bouldin Index	Inertia
2	0.9032	0.0647	4189.88
3	0.5533	0.8728	3156.19
4	0.5153	0.8670	2509.75
5	0.3611	0.8636	2070.29
6	0.3597	0.8396	1731.98
7	0.3624	0.8079	1444.51
8	0.3084	0.9312	1298.50

Note:

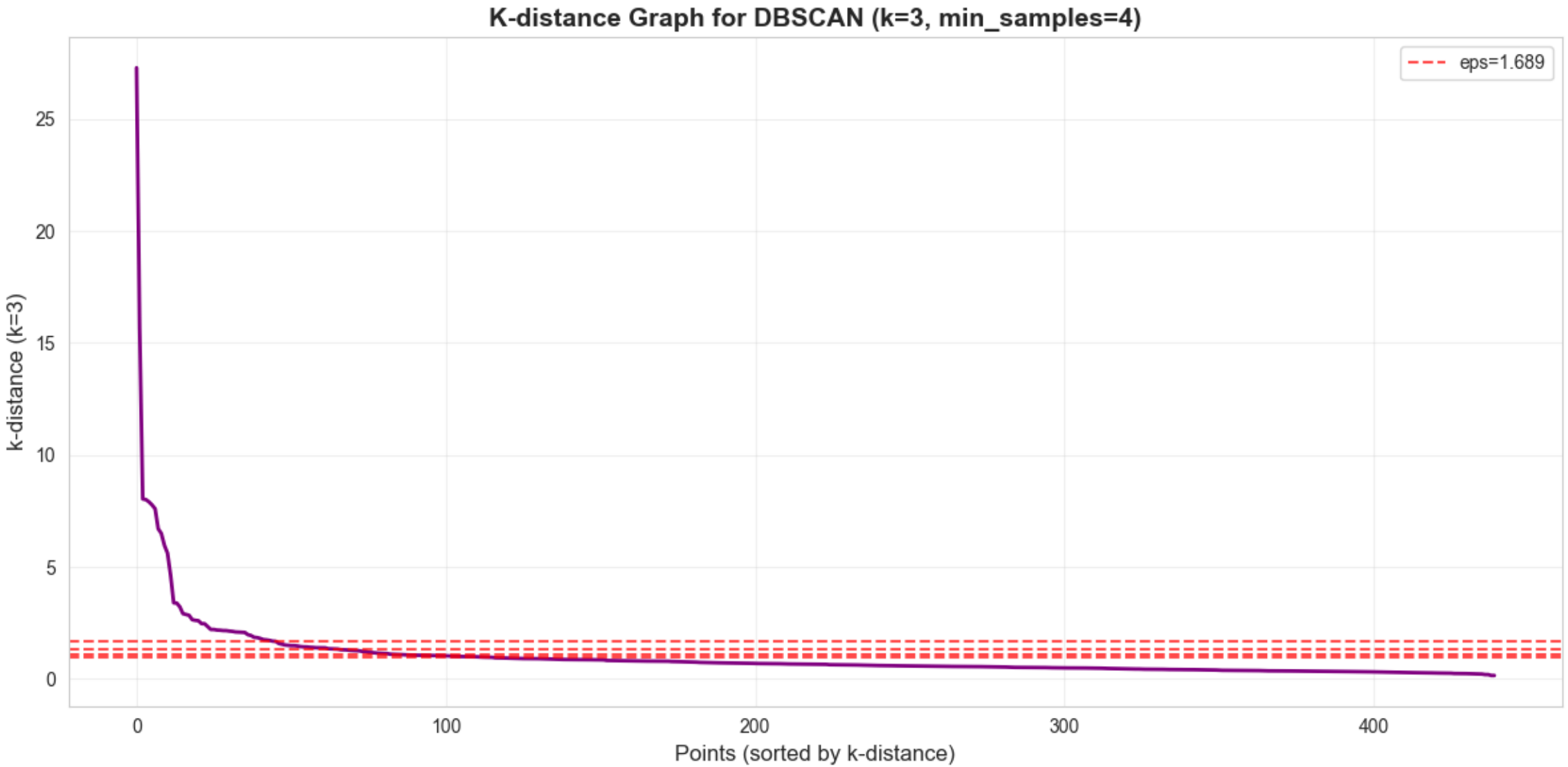
- Higher Silhouette Score = better separation
- Lower Davies-Bouldin Index = better separation
- Lower Inertia = tighter clusters (but may overfit)

K-means evaluation metrics




```
Optimal k based on Silhouette Score: 2 (score: 0.9032)
Optimal k based on Davies-Bouldin Index: 2 (score: 0.0647)

✓ Selected k = 2 (both metrics agree)

K-means clustering complete with k=2
Cluster distribution:
KMeans_Cluster
0    439
1      1
Name: count, dtype: int64
```

DBSCAN Clustering: k-distance plot to choose eps



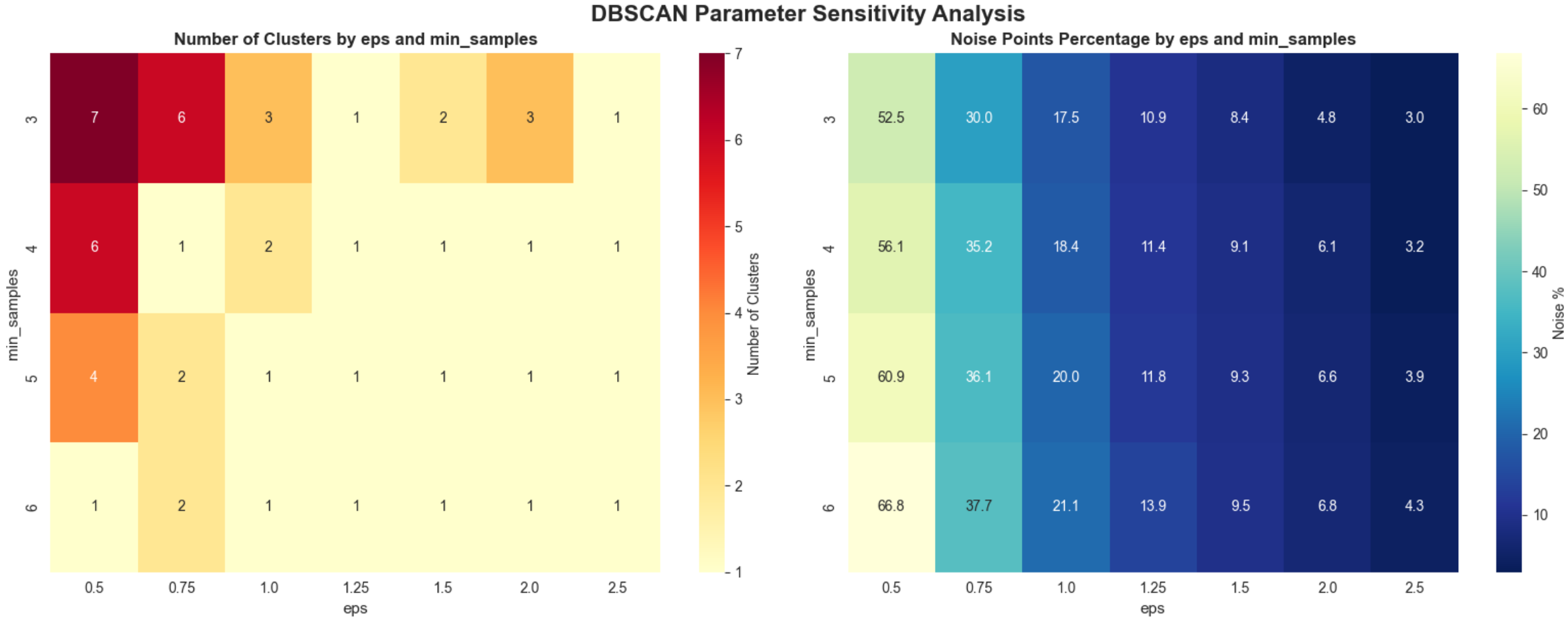
DBSCAN Clustering: k-distance plot to choose eps

```
K-distance statistics (k=3):
  Min: 0.1513
  Max: 27.2875
  Mean: 1.0208
  Median: 0.6512

Suggested eps values (percentiles):
  90th percentile: 1.6893
  85th percentile: 1.3076
  80th percentile: 1.0674
  75th percentile: 0.9877
```

DBSCAN Results for Different Parameters

eps	min_samples	n_clusters	n_noise	% noise
0.50	3	7	231	52.50
0.50	4	6	247	56.14
0.50	5	4	268	60.91
0.50	6	1	294	66.82
0.75	3	6	132	30.00
0.75	4	1	155	35.23
0.75	5	2	159	36.14
0.75	6	2	166	37.73
1.00	3	3	77	17.50
1.00	4	2	81	18.41
1.00	5	1	88	20.00
1.00	6	1	93	21.14
1.25	3	1	48	10.91
1.25	4	1	50	11.36
1.25	5	1	52	11.82
1.25	6	1	61	13.86
1.50	3	2	37	8.41
1.50	4	1	40	9.09
1.50	5	1	41	9.32
1.50	6	1	42	9.55
2.00	3	3	21	4.77
2.00	4	1	27	6.14
2.00	5	1	29	6.59
2.00	6	1	30	6.82
2.50	3	1	13	2.95
2.50	4	1	14	3.18
2.50	5	1	17	3.86
2.50	6	1	19	4.32

DBSCAN parameter sensitivity



DBSCAN parameter sensitivity

```
Selected DBSCAN parameters:
  eps = 1.5
  min_samples = 3
  Expected clusters: 2
  Expected noise: 37 points (8.41%)

DBSCAN clustering complete
Cluster distribution:
DBSCAN_Cluster
-1     37
 0    400
 1      3
Name: count, dtype: int64
```

Second EDA & Statistical Inference

K-means Cluster Centroids (Original Units)

```
K-means Cluster Centroids (Original Units):

                Fresh      Milk   Grocery    Frozen  Detergents_Paper  \
Cluster 0  11943.7   5709.36   7923.44   2995.71            2887.51
Cluster 1  36847.0  43950.00  20170.00  36534.00             239.00

            Delicassen  Size
Cluster 0     1419.13   439
Cluster 1    47943.00     1

Cluster sizes:
KMeans_Cluster
0    439
1      1
Name: count, dtype: int64
```

Second EDA & Statistical Inference

DBSCAN Cluster Medoids (Original Units)

```
DBSCAN Cluster Medoids (Original Units):
_____________________________________________________________________________
             Fresh     Milk  Grocery  Frozen  Detergents_Paper  Delicassen  \
Cluster 0  8257.0   3880.0   6407.0  1646.0            2730.0       344.0
Cluster 1  5181.0  22044.0  21531.0  1740.0            7353.0      4985.0

           Size
Cluster 0   400
Cluster 1     3

Cluster sizes:
DBSCAN_Cluster
-1     37
 0    400
 1      3
Name: count, dtype: int64

Noise points: 37
```

Differences in TotalSpend across Clusters

```
Statistical Test: Differences in TotalSpend across Clusters

1. K-means Clusters:
------------------------------------------------------------
Normality check (Shapiro-Wilk test):
  Cluster 0: p-value = 0.0000 (✗ Not normal)
  Cluster 1: p-value = nan (✗ Not normal)

ANOVA Test:
  F-statistic: 36.2272, p-value: 0.0000
  Result: Significant difference (α=0.05)

Kruskal-Wallis Test (non-parametric):
  H-statistic: 2.9053, p-value: 0.0883
  Result: No significant difference (α=0.05)

2. DBSCAN Clusters:
------------------------------------------------------------
Normality check (Shapiro-Wilk test):
  Cluster -1: p-value = 0.0008 (✗ Not normal)
  Cluster 0: p-value = 0.0000 (✗ Not normal)
  Cluster 1: p-value = 0.6148 (✓ Normal)

ANOVA Test:
  F-statistic: 172.4279, p-value: 0.0000
  Result: Significant difference (α=0.05)

Kruskal-Wallis Test (non-parametric):
  H-statistic: 89.8614, p-value: 0.0000
  Result: Significant difference (α=0.05)
```

THE END

STUDENTS	AINEDEMBE DENIS +256 788-674576 dembedenisjb@gmail.com, ainedembe.denis@stud.umu.ac.ug
	MUSINGUZI BENSON +256 782 942245, musiben@gmail.com, musinguzi.benson@stud.umu.ac.ug
LECTURER	Dr. Sibitenda Harriet +256 777 056581 hsibitenda@umu.ac.ug