Cluster Analysis on Credit Card Customer Data

Introducing the Dataset

```
Veri seti boyutu: (660, 6)

İlk 5 satır:
   Customer Key  Avg_Credit_Limit  Total_Credit_Cards  Total_visits_bank
0         87073            100000                   2                  1
1         38414             50000                   3                  0
2         17341             50000                   7                  1
3         40496             30000                   5                  1
4         47437            100000                   6                  0

   Total_visits_online  Total_calls_made
0                    1                 0
1                   10                 9
2                    3                 4
3                    1                 4
4                   12                 3
```

```
Eksik değerler:
Customer Key            0
Avg_Credit_Limit        0
Total_Credit_Cards      0
Total_visits_bank       0
Total_visits_online     0
Total_calls_made        0
dtype: int64
```

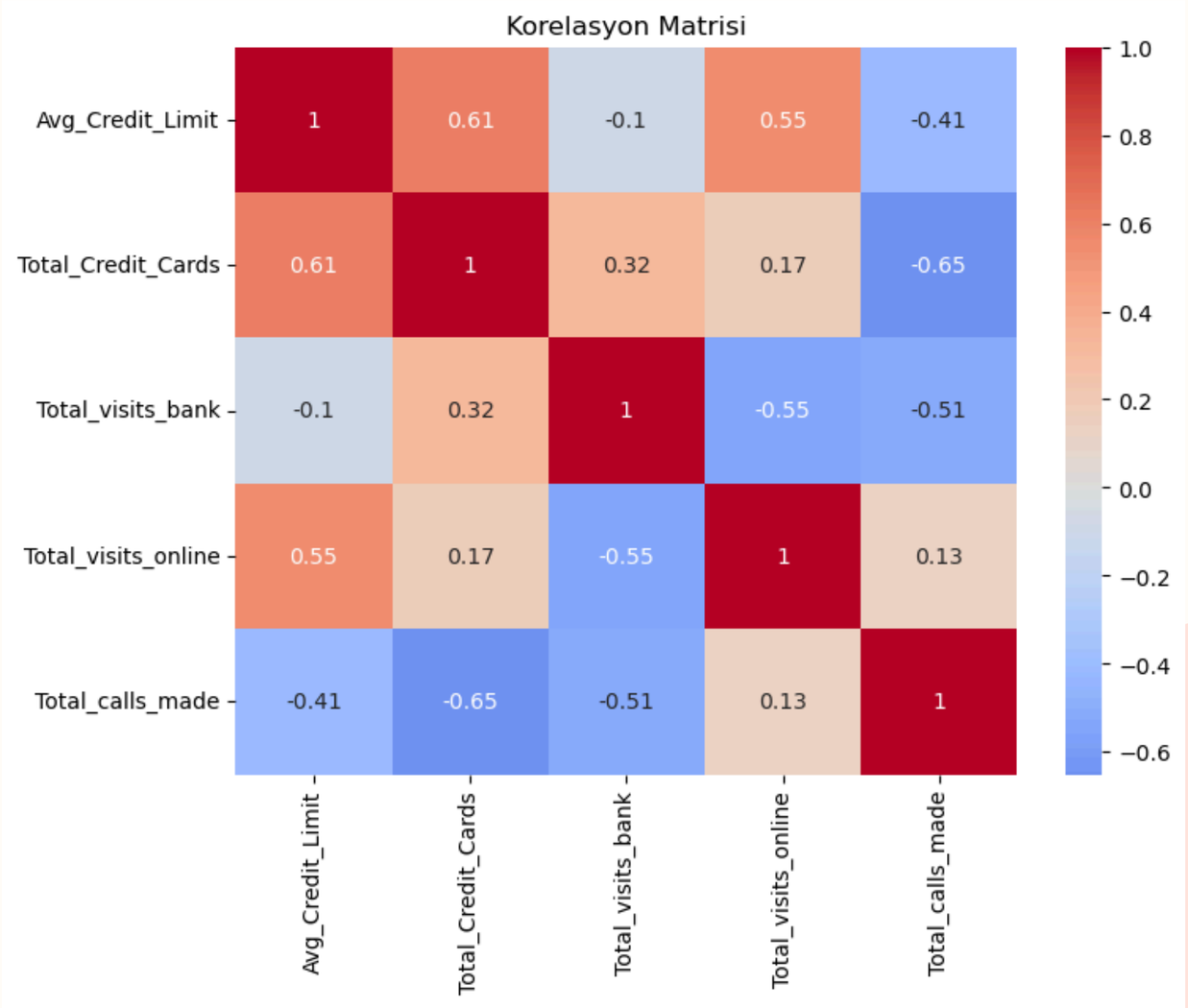
DESCRİPTİVE STATİSTİCS

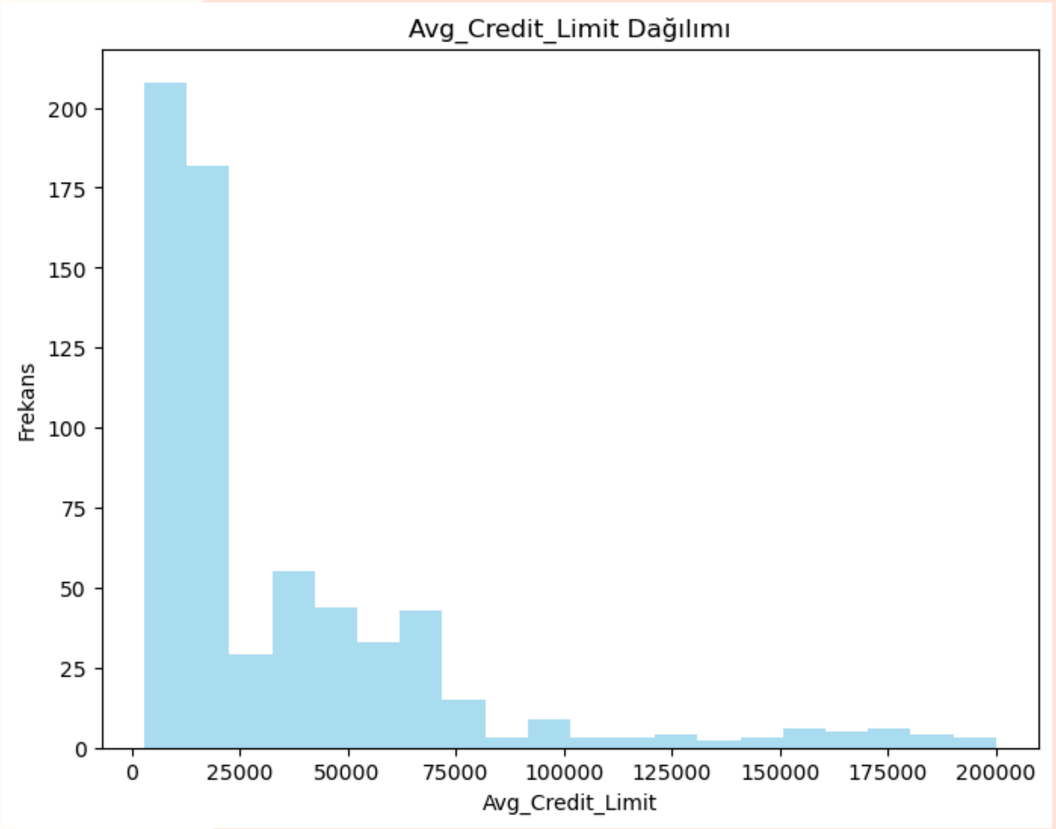
Tanımlayıcı İstatistikler:

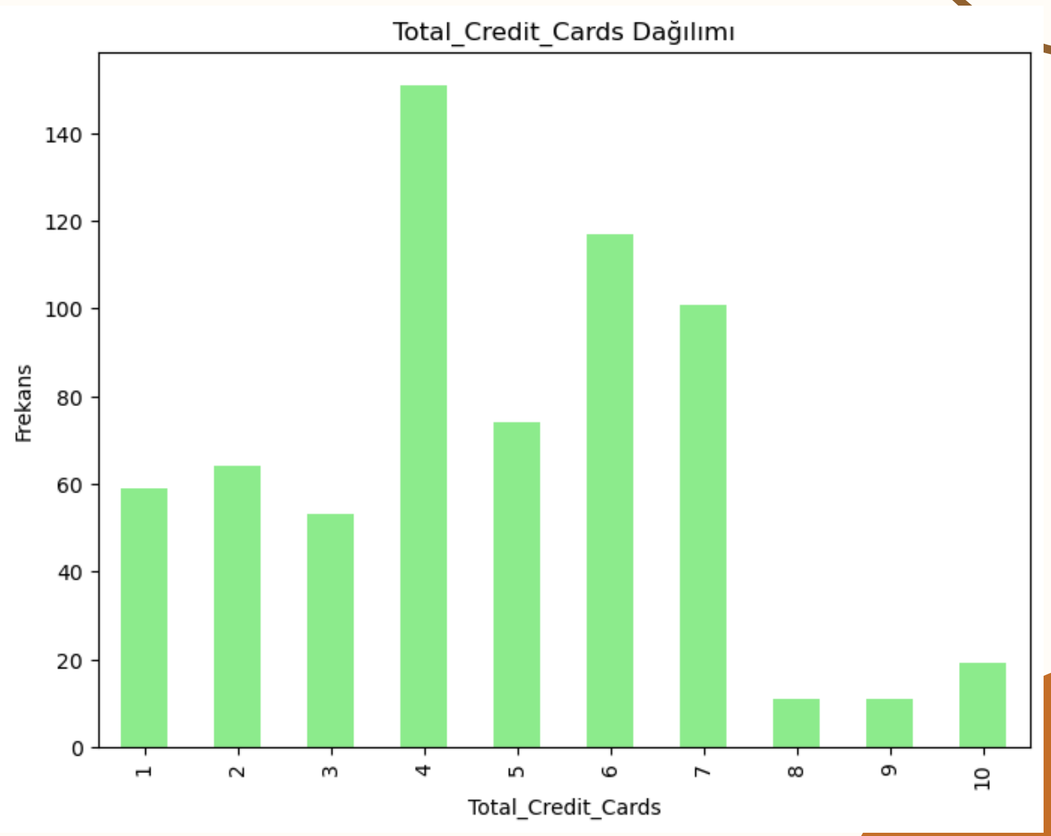
	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	\
count	660.000000	660.000000	660.000000	
mean	34574.242424	4.706061	2.403030	
std	37625.487804	2.167835	1.631813	
min	3000.000000	1.000000	0.000000	
25%	10000.000000	3.000000	1.000000	
50%	18000.000000	5.000000	2.000000	
75%	48000.000000	6.000000	4.000000	
max	200000.000000	10.000000	5.000000	

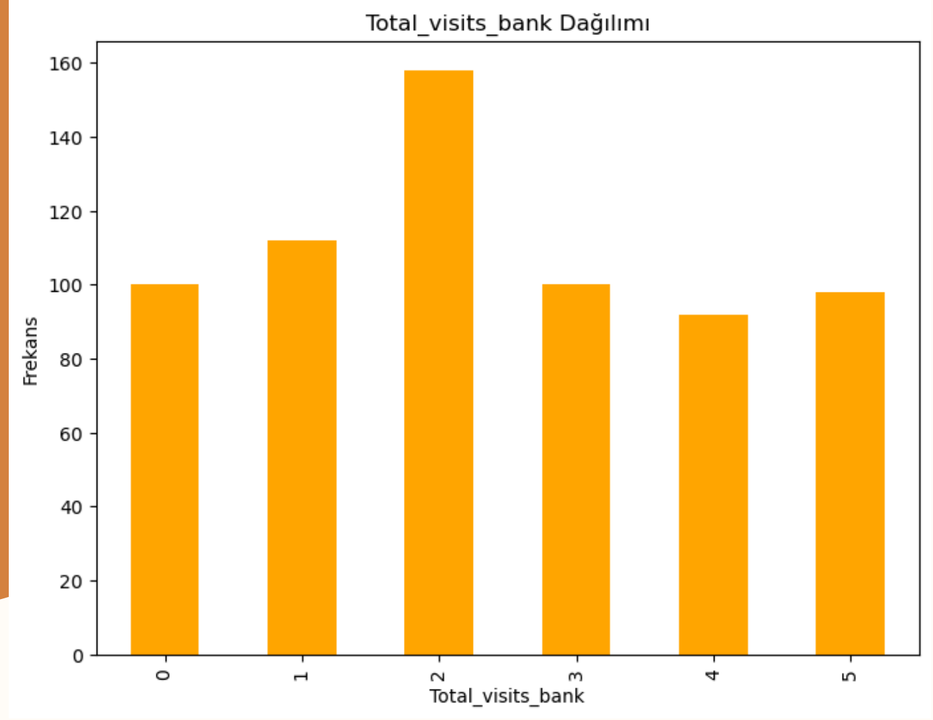
	Total_visits_online	Total_calls_made
count	660.000000	660.000000
mean	2.606061	3.583333
std	2.935724	2.865317
min	0.000000	0.000000
25%	1.000000	1.000000
50%	2.000000	3.000000
75%	4.000000	5.000000
max	15.000000	10.000000

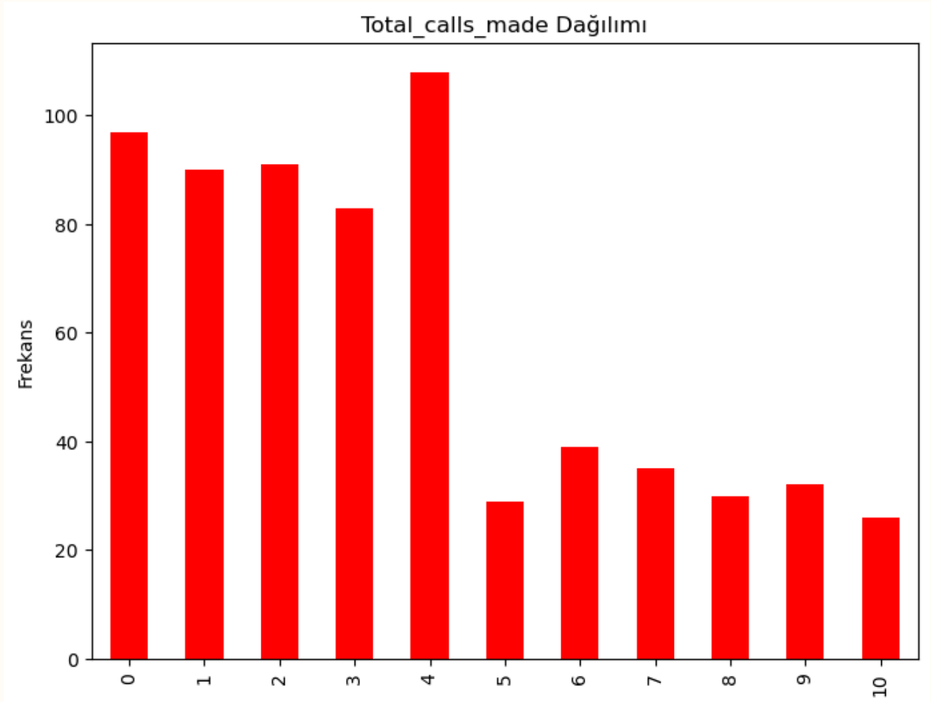
Correlation Matrix

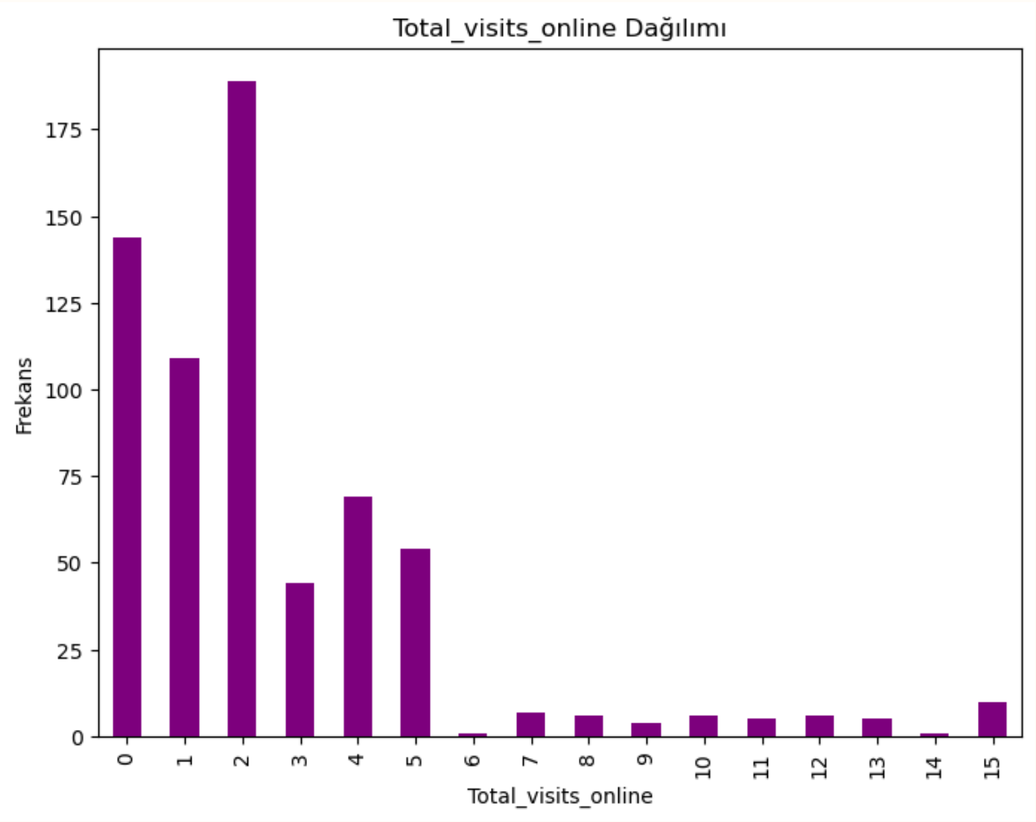


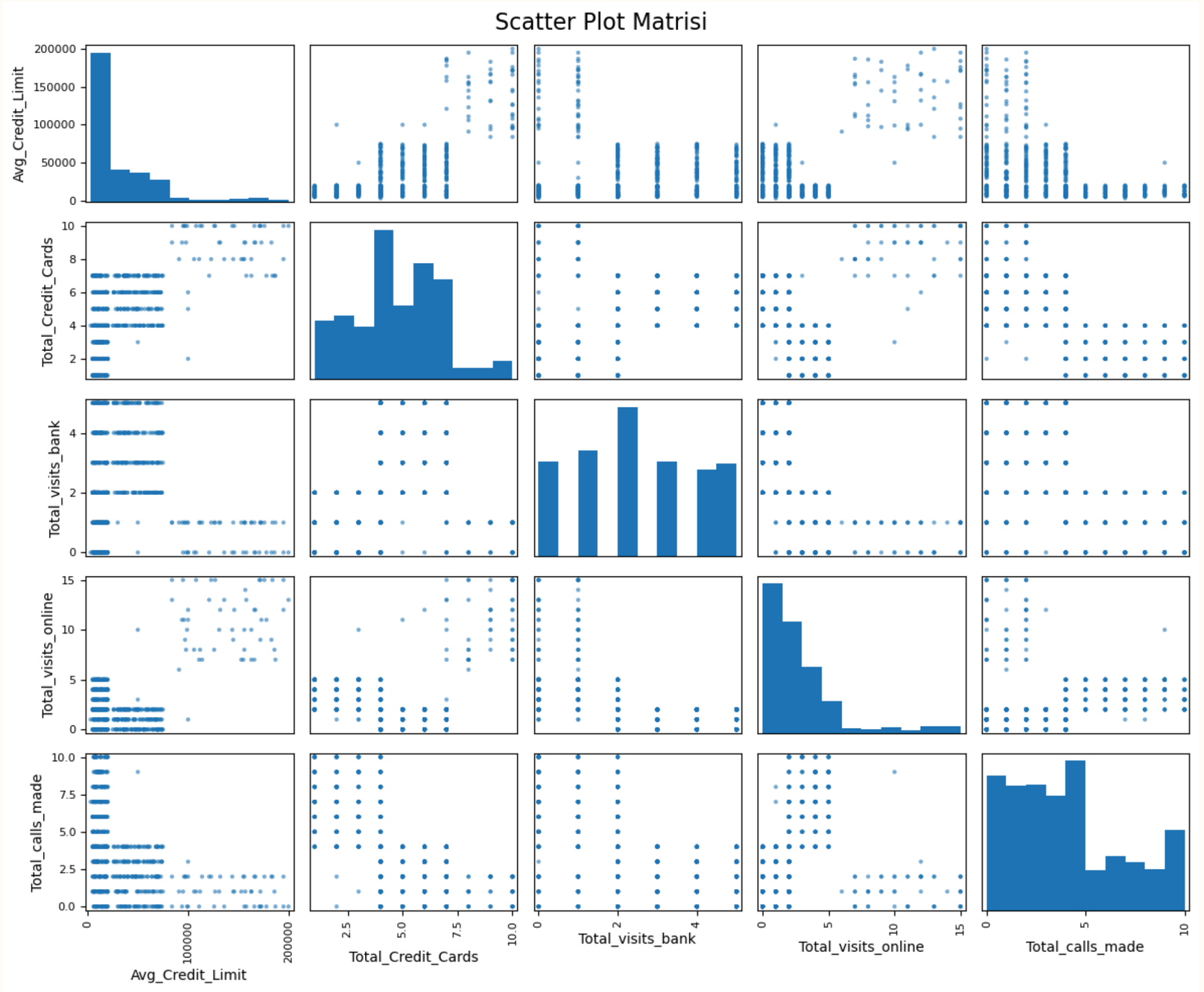

Avg_Credit_Limit Dağılımı
Frekans
Avg_Credit_Limit
0
25
50
75
100
125
150
175
200
0
25000
50000
75000
100000
125000
150000
175000
200000


Total_Credit_Cards Dağılımı
Frekans
Total_Credit_Cards
0
20
40
60
80
100
120
140
1
2
3
4
5
6
7
8
9
10

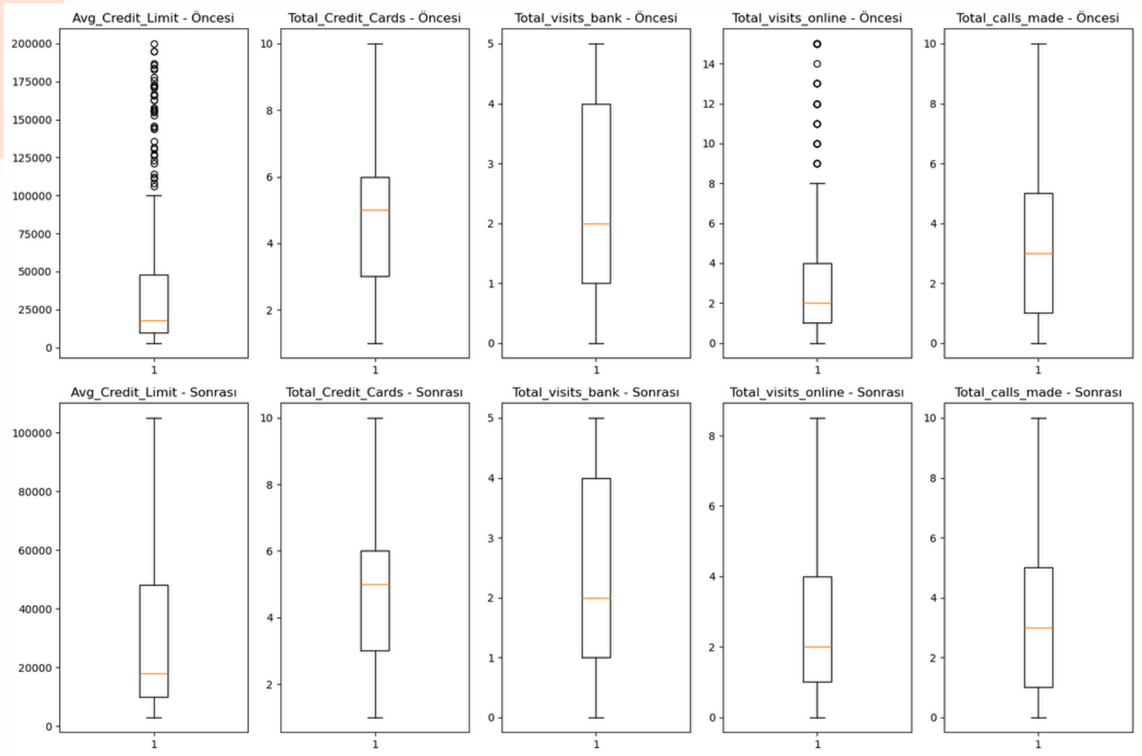

Total_visits_bank Dağılımı
Frekans
Total_visits_bank
0
20
40
60
80
100
120
140
160
0
1
2
3
4
5

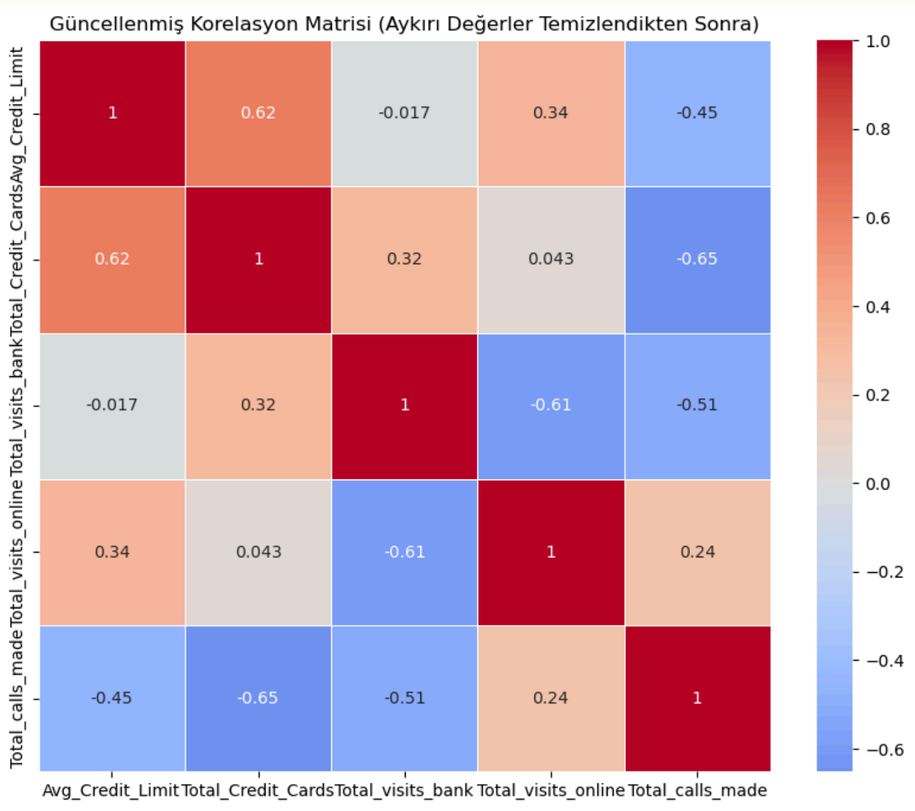

Total_calls_made Dağılımı
Frekans
0
20
40
60
80
100
0
1
2
3
4
5
6
7
8
9
10


Total_visits_online Dağılımı
Frekans
Total_visits_online
0
25
50
75
100
125
150
175
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

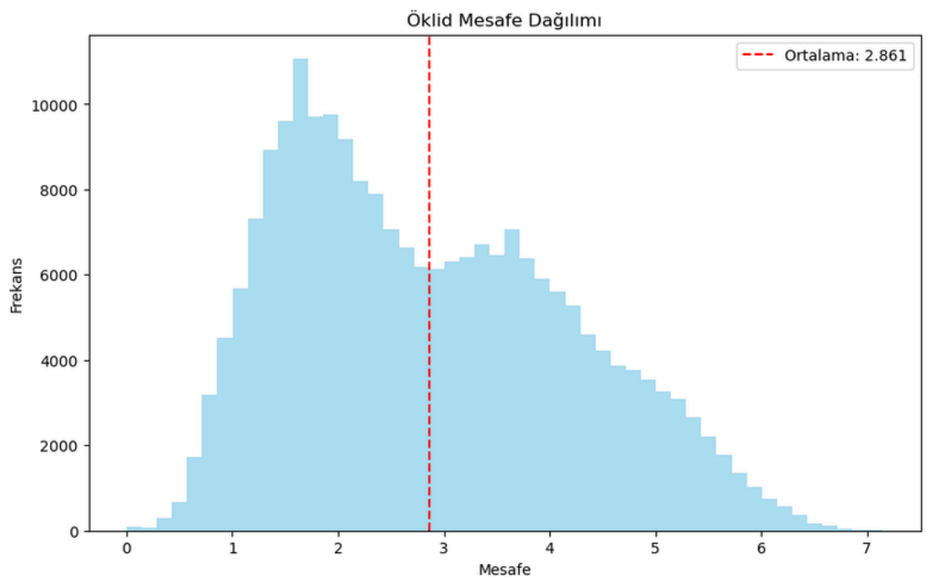

Scatter Plot Matrisi
Avg_Credit_Limit
Total_Credit_Cards
Total_visits_bank
Total_visits_online
Total_calls_made

Boxplots - Correlation Matrix



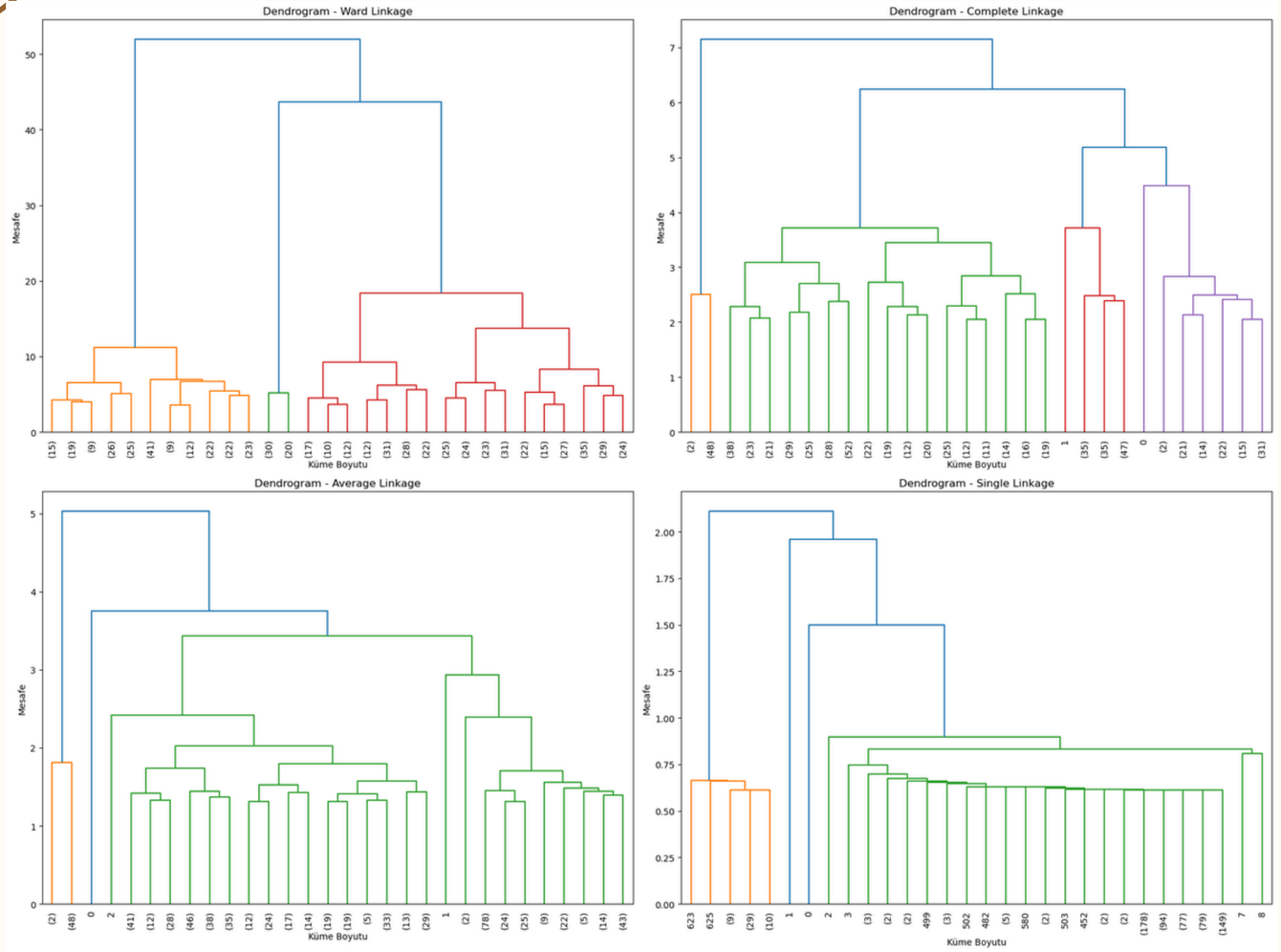


Euclidean Distance Distribution -Euclidean Distance Matrix

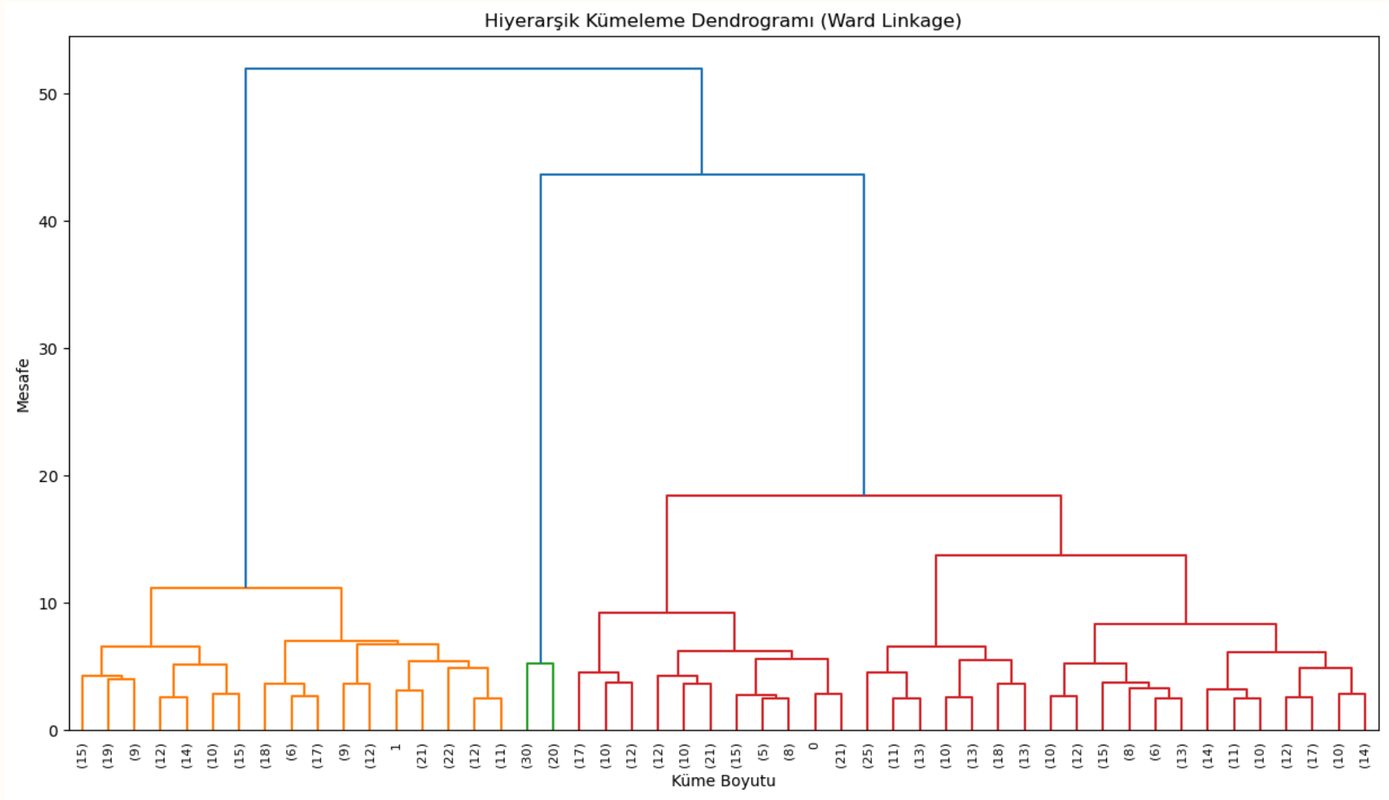


The first chart shows frequency of distance values, with an average of 2.861 and data peaking around 3.
The second chart represents distances between observations with colors, where dark indicates low and light indicates high distances, with the diagonal showing zero distance

Ward, Complete, Average, Single dendogram



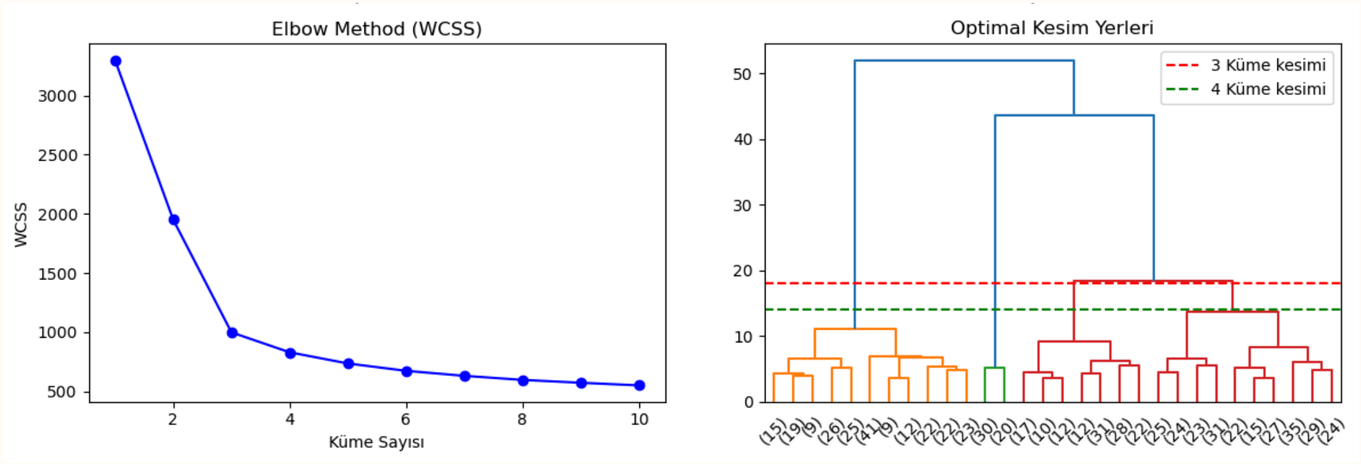
While the Ward and Complete methods tend to cluster at higher distance values, the Average and Single methods tend to cluster at lower distances.




```
2 küme için dağılım:
  Küme 1: 223 gözlem
  Küme 2: 437 gözlem

3 küme için dağılım:
  Küme 1: 223 gözlem
  Küme 2: 50 gözlem
  Küme 3: 387 gözlem

4 küme için dağılım:
  Küme 1: 223 gözlem
  Küme 2: 50 gözlem
  Küme 3: 132 gözlem
  Küme 4: 255 gözlem

5 küme için dağılım:
  Küme 1: 223 gözlem
  Küme 2: 50 gözlem
  Küme 3: 132 gözlem
  Küme 4: 103 gözlem
  Küme 5: 152 gözlem

6 küme için dağılım:
  Küme 1: 94 gözlem
  Küme 2: 129 gözlem
  Küme 3: 50 gözlem
  Küme 4: 132 gözlem
  Küme 5: 103 gözlem
  Küme 6: 152 gözlem
```

The chart displays a hierarchical clustering dendrogram using the Ward Linkage method, with distance values reaching up to 50, though clusters mostly merge at lower distances (0-10 range).

The text lists the number of observations in each cluster for different cluster counts (2-6); for example, with 2 clusters, the 1st cluster has 223 and the 2nd has 437 observations.



According to the "Optimal Cutting Points" chart, the 3-cluster choice is marked as an optimal cut point at approximately 20 distance levels, suggesting that dividing the data into 3 groups is appropriate. The "Elbow Method (WCSS)" chart also supports this, as 3 clusters appear where WCSS stabilizes after a noticeable drop.

3 Küme Analizi:
Silhouette Skoru: 0.517
Küme 1: 223 gözlem (33.8%)
Küme 2: 50 gözlem (7.6%)
Küme 3: 387 gözlem (58.6%)

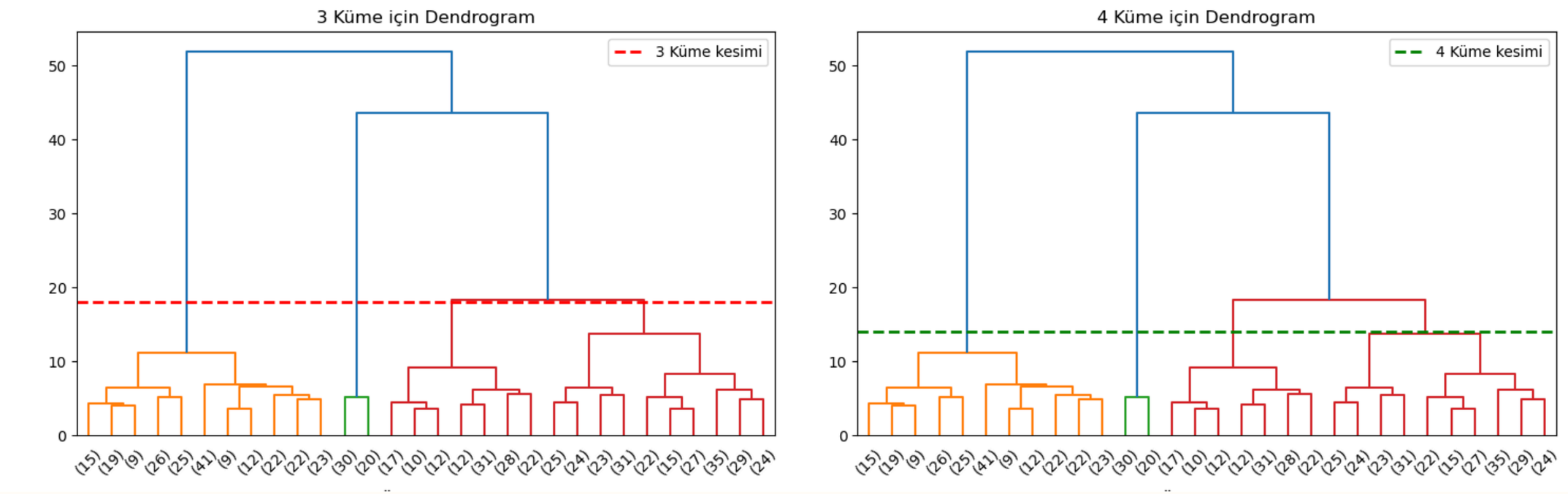
4 Küme Analizi:
Silhouette Skoru: 0.372
Küme 1: 223 gözlem (33.8%)
Küme 2: 50 gözlem (7.6%)
Küme 3: 132 gözlem (20.0%)
Küme 4: 255 gözlem (38.6%)

3 Küme - Özellik Ortalamaları:

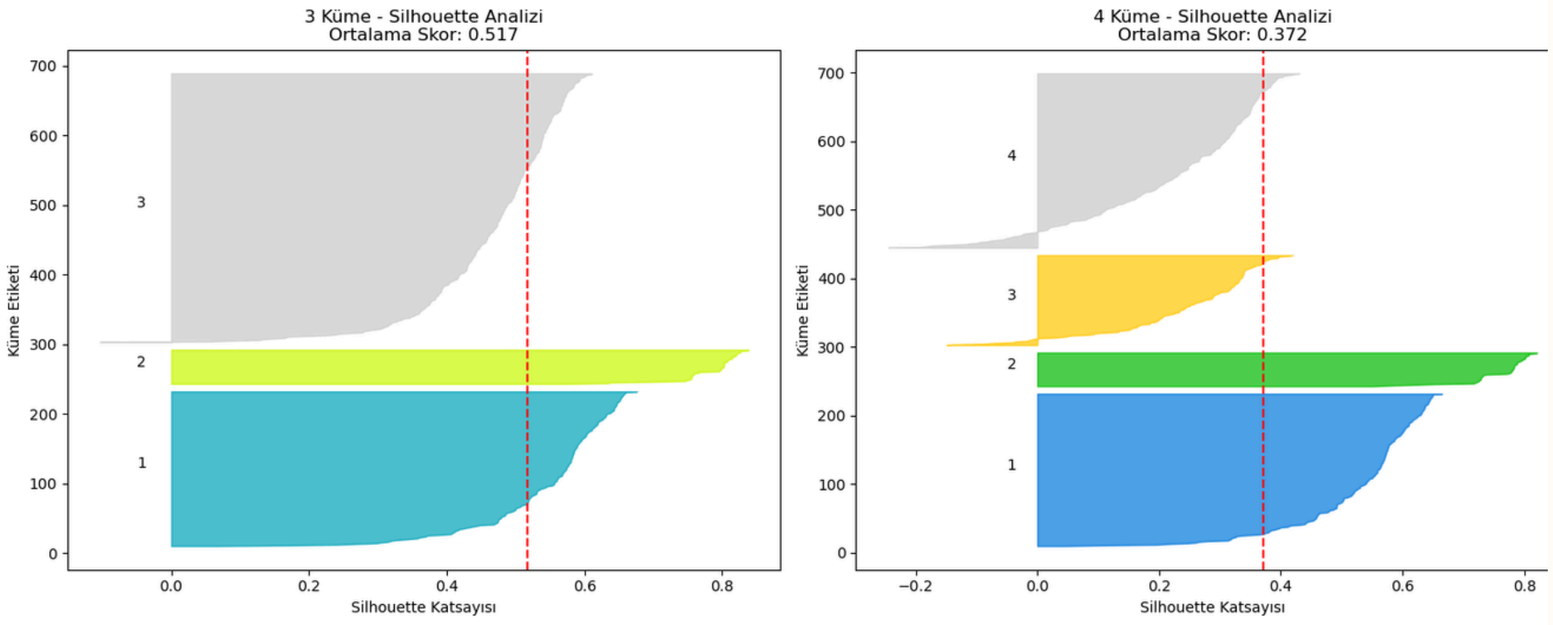
Cluster	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
1	12197.309417	2.403587	0.928251	3.553812	6.883408
2	102660.000000	8.740000	0.600000	8.180000	1.080000
3	33713.178295	5.511628	3.485788	0.984496	2.005168

4 Küme - Özellik Ortalamaları:

Cluster	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
1	12197.309417	2.403587	0.928251	3.553812	6.883408
2	102660.000000	8.740000	0.600000	8.180000	1.080000
3	58916.666667	5.583333	3.113636	0.848485	1.909091
4	20666.666667	5.474510	3.678431	1.054902	2.054902


3 Küme için Dendrogram
3 Küme kesimi
50
40
30
20
10
0
(15) (19) (9) (26) (25) (41) (9) (12) (22) (22) (23) (30) (20) (17) (10) (12) (12) (31) (28) (22) (25) (24) (23) (31) (22) (15) (27) (35) (29) (24)
4 Küme için Dendrogram
4 Küme kesimi
50
40
30
20
10
0
(15) (19) (9) (26) (25) (41) (9) (12) (22) (22) (23) (30) (20) (17) (10) (12) (12) (31) (28) (22) (25) (24) (23) (31) (22) (15) (27) (35) (29) (24)

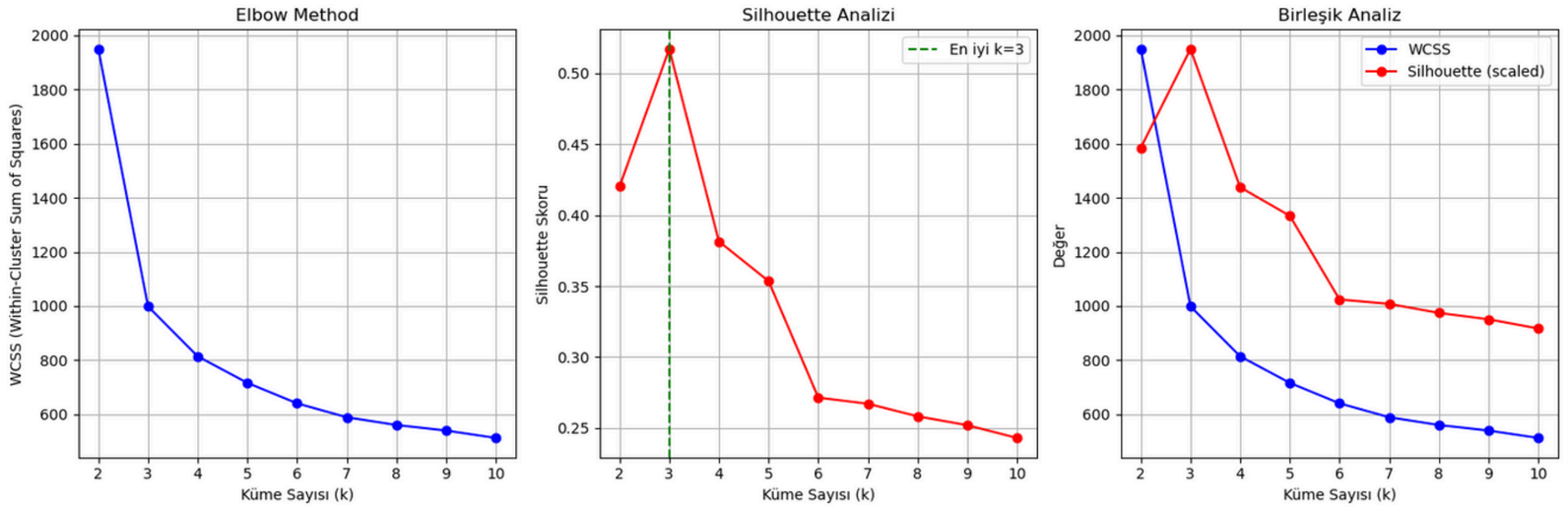
Silhouette Analysis



- "K-Means Clusters (PCA)" chart shows 3 clusters (red, blue, green) with centroids (X); PCA explains 81.7% of variance, indicating good separation.
- "Silhouette Analysis" chart, with an average score of 0.517, confirms 3 clusters are suitable, especially for green and blue clusters.
- "Cluster Size Distribution" pie chart shows clusters distributed as 58.5% (red), 33.9% (green), and 7.6% (blue).
- Conclusion: 3 clusters well represent the data.

- "3 Clusters - Silhouette": Average 0.517, cluster 2 (yellow) well-separated.
- "4 Clusters - Silhouette": Average 0.372, cluster 4 (gray) negative, 3 better.
- Conclusion: 3 clusters best.

K-Means



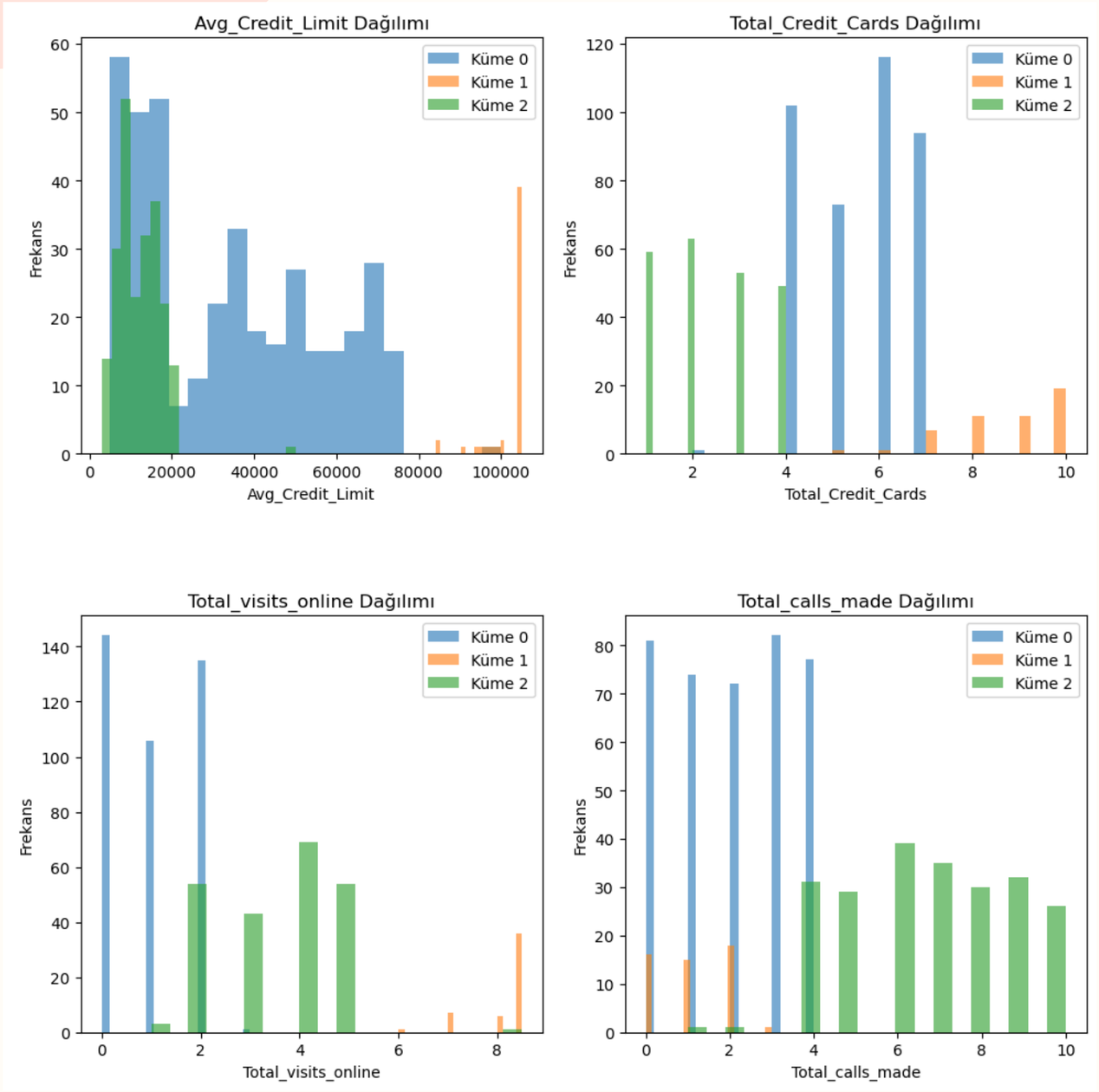

```
Optimal k değeri analizi:
k=2: WCSS=1949.90, Silhouette=0.420
k=3: WCSS=998.47, Silhouette=0.517
k=4: WCSS=813.87, Silhouette=0.381
k=5: WCSS=715.78, Silhouette=0.354
k=6: WCSS=640.18, Silhouette=0.272
k=7: WCSS=588.73, Silhouette=0.267
k=8: WCSS=560.07, Silhouette=0.258
k=9: WCSS=539.55, Silhouette=0.252
k=10: WCSS=512.20, Silhouette=0.243

En yüksek Silhouette skoruna sahip k: 3
```

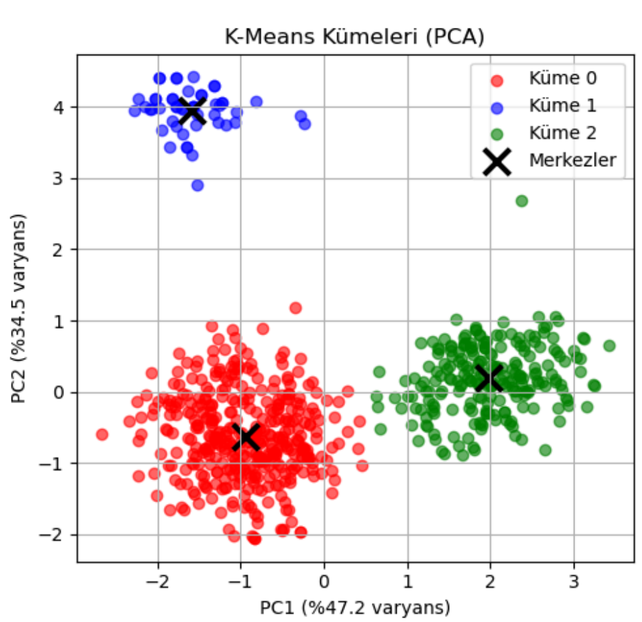
```
K-Means k=3 Sonuçları:
Silhouette Skoru: 0.517
WCSS: 998.47
Küme 0: 386 gözlem (58.5%)
Küme 1: 50 gözlem (7.6%)
Küme 2: 224 gözlem (33.9%)

K-Means k=4 Sonuçları:
Silhouette Skoru: 0.381
WCSS: 813.87
Küme 0: 223 gözlem (33.8%)
Küme 1: 50 gözlem (7.6%)
Küme 2: 223 gözlem (33.8%)
Küme 3: 164 gözlem (24.8%)
```

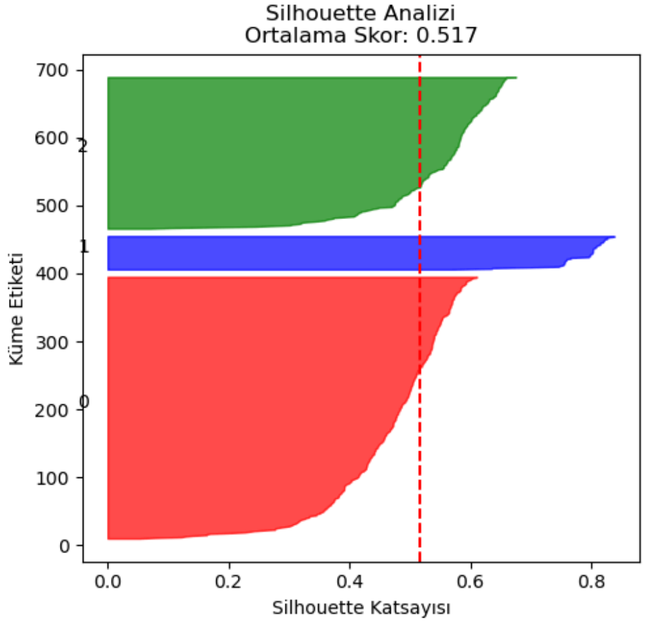
K=3

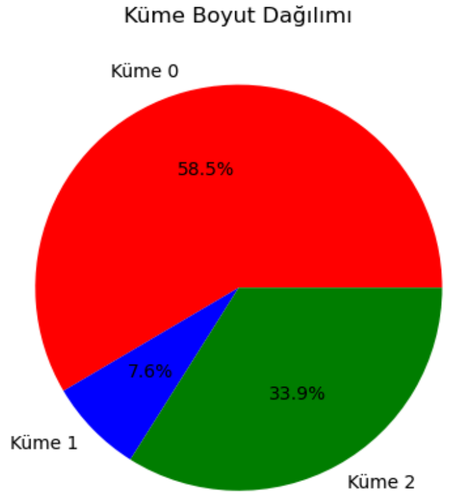
```
Küme Merkezleri (Normalize edilmiş):
   Avg_Credit_Limit  Total_Credit_Cards  Total_visits_bank  \
0          0.074275            0.373690           0.666395
1          2.492325            1.862226          -1.105763
2         -0.684315           -1.059623          -0.901518

   Total_visits_online  Total_calls_made
0            -0.627808         -0.553005
1             2.563922         -0.874330
2             0.509544          1.148109

Küme Merkezleri (Orijinal ölçek):
   Avg_Credit_Limit  Total_Credit_Cards  Total_visits_bank  \
0      33782.383420            5.515544           3.489637
1     102660.000000            8.740000           0.600000
2      12174.107143            2.410714           0.933036

   Total_visits_online  Total_calls_made
0             0.981865          2.000000
1             8.180000          1.080000
2             3.546875          6.870536
```

```
Küme Profilleri:
        Avg_Credit_Limit                    Total_Credit_Cards           \
                    mean           std count               mean       std
Cluster
0           33782.383420  22169.460696   386           5.515544  1.140069
1          102660.000000   5161.118373    50           8.740000  1.274715
2           12174.107143   5249.048524   224           2.410714  1.100783

              Total_visits_bank                 Total_visits_online           \
        count              mean       std count               mean       std
Cluster
0         386          3.489637  1.135563   386           0.981865  0.857167
1          50          0.600000  0.494872    50           8.180000  0.612456
2         224          0.933036  0.803567   224           3.546875  1.187109

              Total_calls_made
        count             mean       std count
Cluster
0         386         2.000000  1.430648   386
1          50         1.080000  0.876915    50
2         224         6.870536  1.990161   224
```


Avg_Credit_Limit Dağılımı
Küme 0
Küme 1
Küme 2
Frekans
Avg_Credit_Limit
Total_Credit_Cards Dağılımı
Küme 0
Küme 1
Küme 2
Frekans
Total_Credit_Cards
Total_visits_bank Dağılımı
Küme 0
Küme 1
Küme 2
Frekans
Total_visits_bank
Total_visits_online Dağılımı
Küme 0
Küme 1
Küme 2
Frekans
Total_visits_online
Total_calls_made Dağılımı
Küme 0
Küme 1
Küme 2
Frekans
Total_calls_made







- "K-Means Clusters (PCA)" chart shows 3 clusters (red, blue, green) with centroids (X); PCA explains 81.7% of variance, indicating good separation.
- "Silhouette Analysis" chart, with an average score of 0.517, confirms 3 clusters are suitable, especially for green and blue clusters.
- "Cluster Size Distribution" pie chart shows clusters distributed as 58.5% (red), 33.9% (green), and 7.6% (blue).
- Conclusion: 3 clusters well represent the data.

```
K-Means (k=3) Sonuç Özeti:
Silhouette Skoru: 0.517
WCSS: 998.47
PCA ile açıklanan toplam varyans: %81.7
```

K=4

K=4 Küme Merkezleri (Normalize edilmiş):

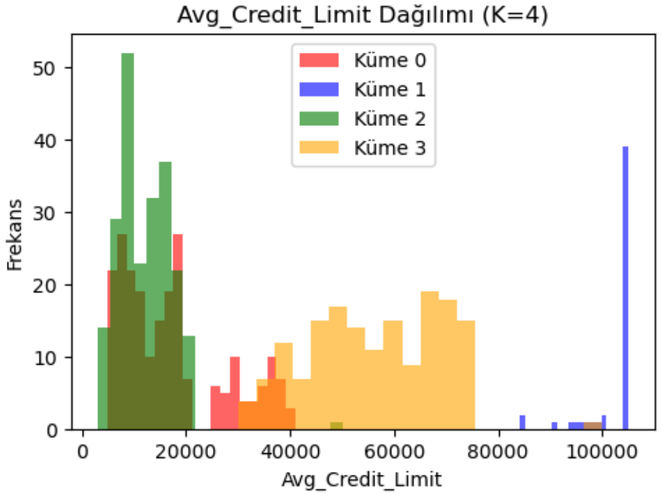
	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	-0.507146	0.364434	0.802588	-0.613775	-0.576604
1	2.492325	1.862226	-1.105763	2.563922	-0.874330
2	-0.683500	-1.062913	-0.904453	0.512620	1.152605
3	0.850853	0.381902	0.477615	-0.643953	-0.517020

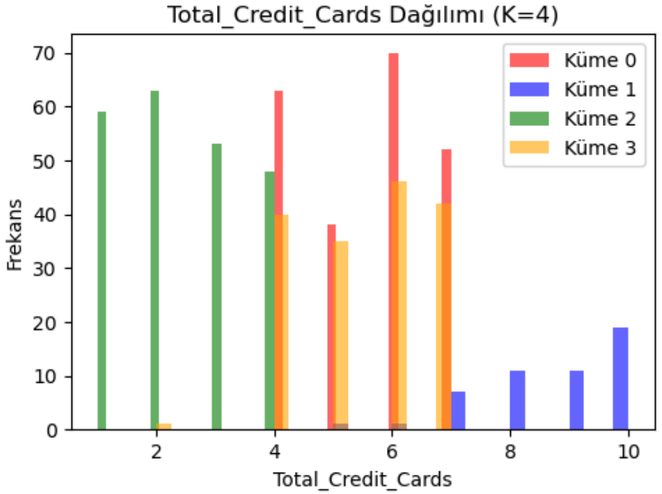
K=4 Küme Merkezleri (Orijinal ölçek):

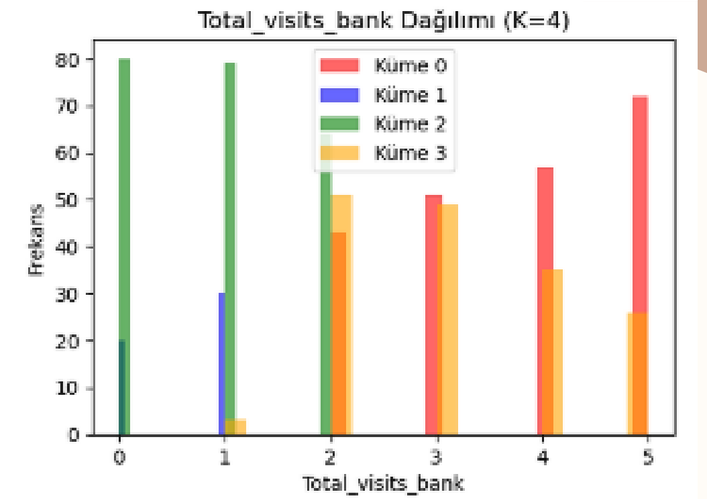
	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	17220.720721	5.495495	3.711712	1.013514	1.932432
1	102660.000000	8.740000	0.600000	8.180000	1.080000
2	12197.309417	2.403587	0.928251	3.553812	6.883408
3	55903.030303	5.533333	3.181818	0.945455	2.103030

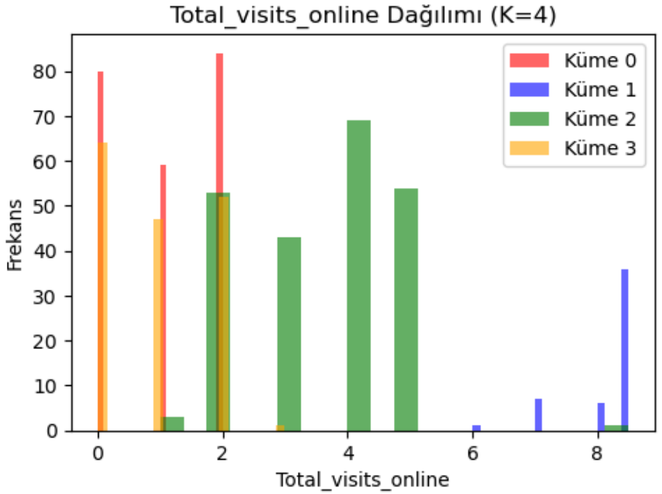
K=4 Küme Profilleri:

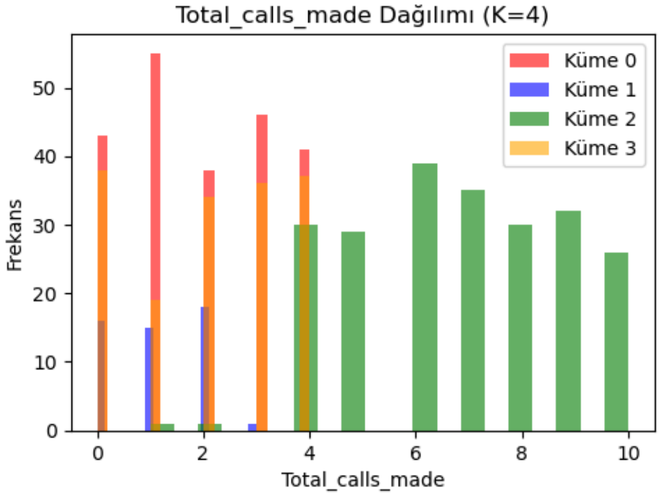
	Avg_Credit_Limit			Total_Credit_Cards		Total_visits_bank			Total_visits_online		Total_calls_made		
	mean	std	count	mean	std	count	mean	std	mean	std	count	mean	std
Cluster													
0	17295.964126	9996.951420	223	5.497758	1.134529	223	3.708520	1.115117	1.017937	0.859311	223	1.941704	1.401804
1	102660.000000	5161.118373	50	8.740000	1.274715	50	0.600000	0.494872	8.180000	0.612456	50	1.080000	0.876915
2	12197.309417	5249.332169	223	2.403587	1.098068	223	0.928251	0.802171	3.553812	1.185221	223	6.883408	1.985271
3	56036.585366	12610.196597	164	5.530488	1.153409	164	3.182927	1.097996	0.939024	0.855836	164	2.091463	1.472921

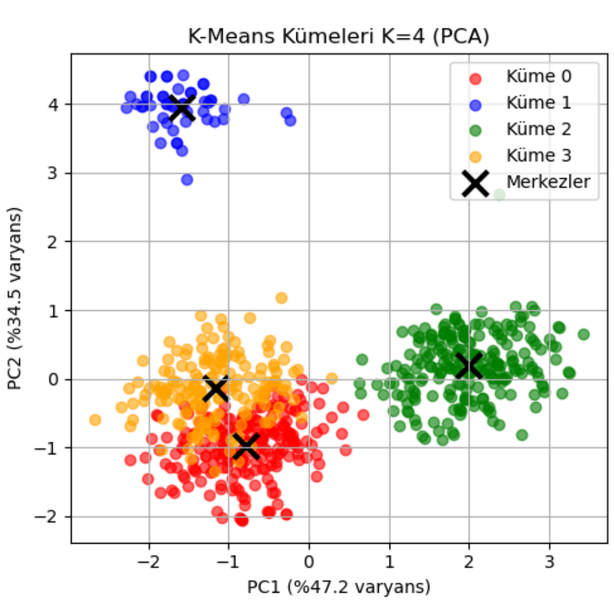

Avg_Credit_Limit Dağılımı (K=4)
Küme 0
Küme 1
Küme 2
Küme 3
Frekans
Avg_Credit_Limit

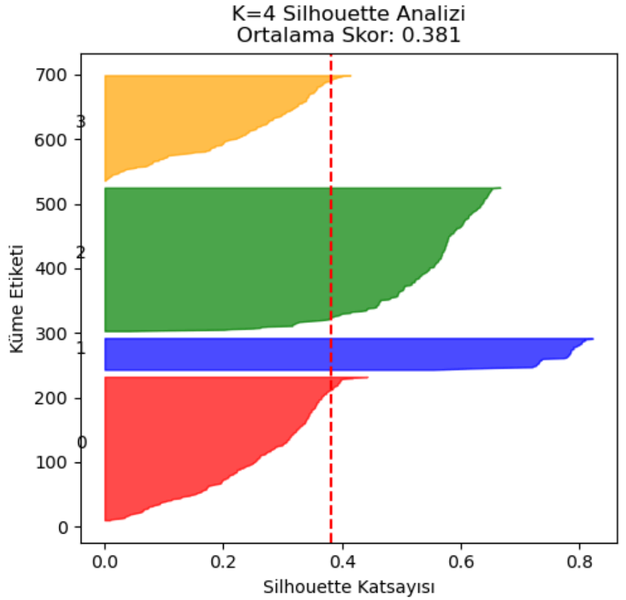

Total_Credit_Cards Dağılımı (K=4)
Küme 0
Küme 1
Küme 2
Küme 3
Frekans
Total_Credit_Cards

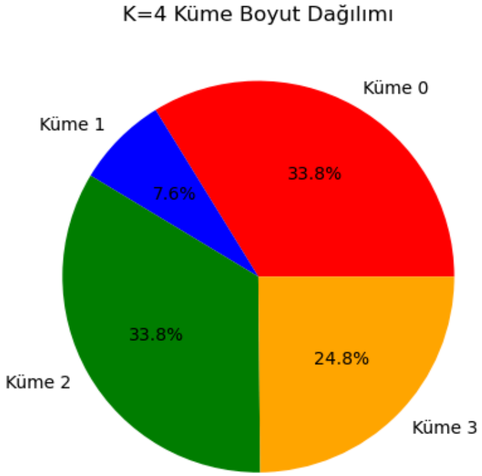

Total_visits_bank Dağılımı (K=4)
Küme 0
Küme 1
Küme 2
Küme 3
Frekans
Total_visits_bank


Total_visits_online Dağılımı (K=4)
Küme 0
Küme 1
Küme 2
Küme 3
Frekans
Total_visits_online


Total_calls_made Dağılımı (K=4)
Küme 0
Küme 1
Küme 2
Küme 3
Frekans
Total_calls_made



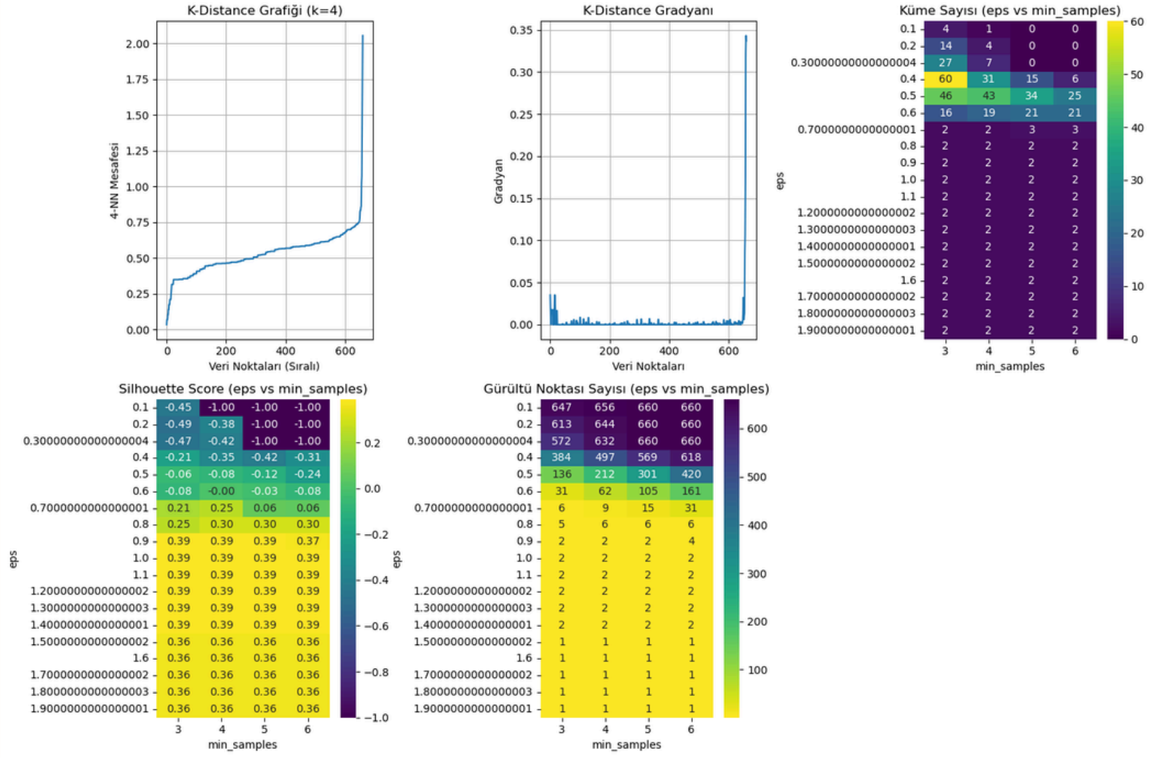




- "K-Means Clusters K=4 (PCA)" chart shows 4 clusters (red, blue, green, yellow) with centroids (X); PCA explains 47.2% variance, clusters appear separated.
- "K=4 Silhouette Analysis" chart, with an average score of 0.381, indicates moderate separation for 4 clusters, with green cluster (2) showing good separation.
- "K=4 Cluster Size Distribution" pie chart shows clusters distributed as 33.8% (red), 33.8% (green), 7.8% (blue), and 24.8% (yellow).
- Conclusion: 4 clusters support separation, but the score is lower than for 3 clusters.

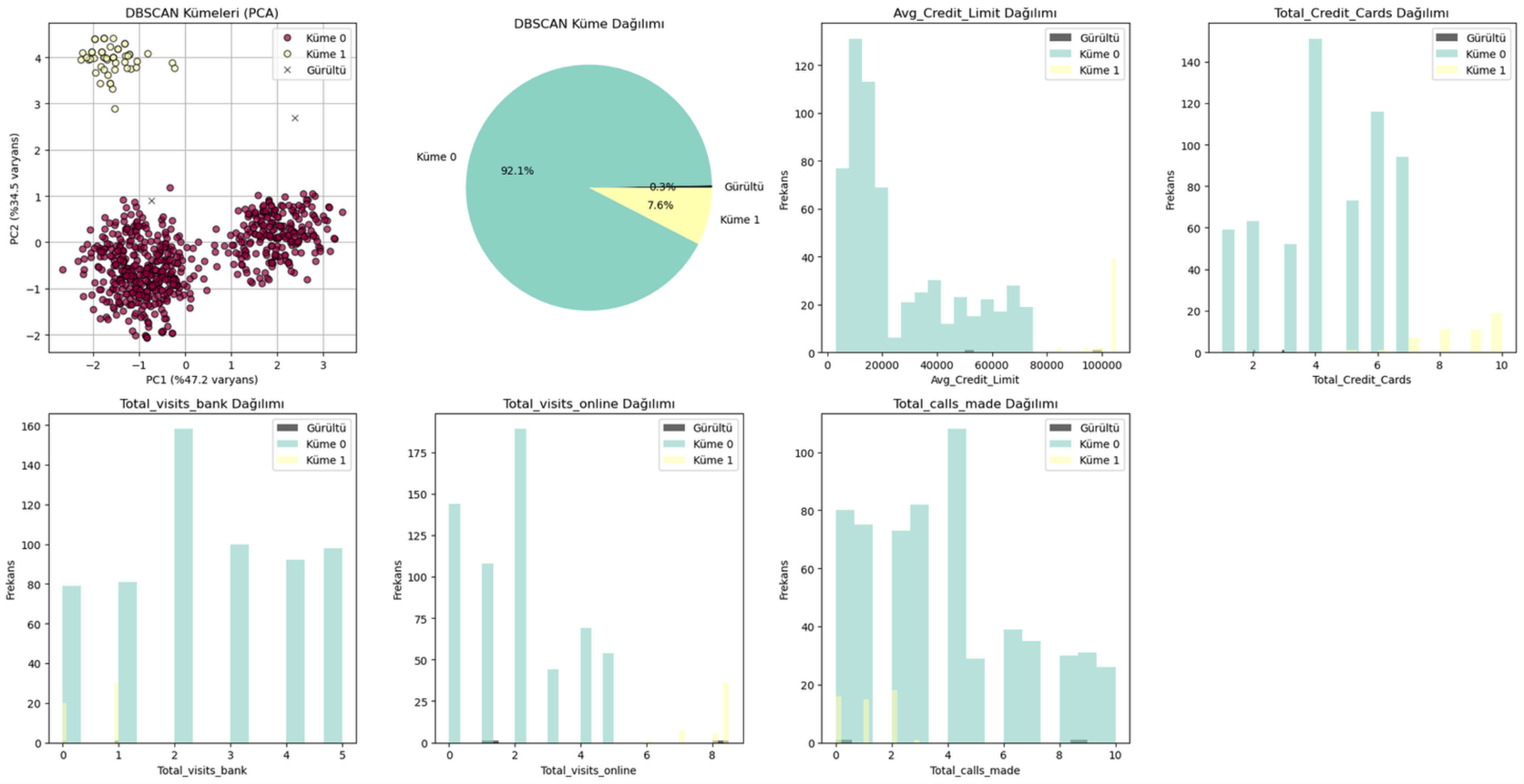
```
K-Means (k=4) Sonuç Özeti:
Silhouette Skoru: 0.381
WCSS: 813.87

K=3 vs K=4 Karşılaştırması:
K=3 Silhouette: 0.517, WCSS: 998.47
K=4 Silhouette: 0.381, WCSS: 813.87

Küme boyut dağılımları:
K=3: [386  50 224]
K=4: [223  50 223 164]
```

DBSCAN




```
Optimal DBSCAN Sonuçları:
eps: 0.9
min_samples: 3
Küme sayısı: 2
Gürültü noktası sayısı: 2
Gürültü oranı: 0.3%
Silhouette Score: 0.519

Küme dağılımı:
Gürültü: 2 nokta
Küme 0: 608 nokta
Küme 1: 50 nokta
```

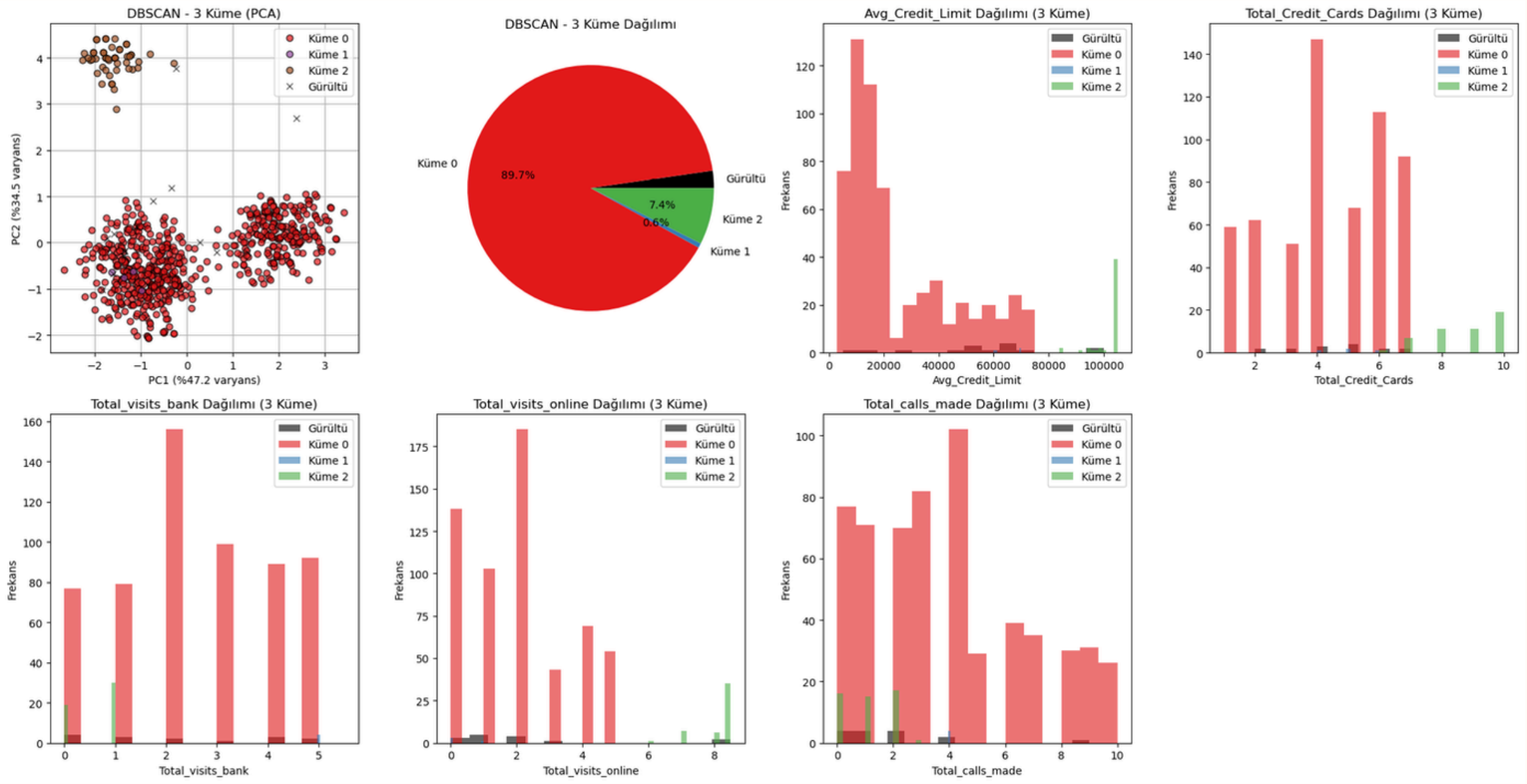
- K-Distance Graph (K=4) shows a knee point around 1.5, indicating a suitable eps value for DBSCAN.
- K-Distance Gradient" chart reflects data density, with a sharp increase between 400-600, marking a critical clustering region.
- Silhouette Score (eps vs min_samples) table gives the highest score of 0.45 with eps=0.3 and min_samples=4-5, indicating good separation.
- Noise Points (eps vs min_samples) table shows the lowest noise count (3-4) with eps=0.3 and min_samples=4-5, supporting these settings as optimal.
- Conclusion: eps=0.3 and min_samples=4-5 provide the best clustering performance.



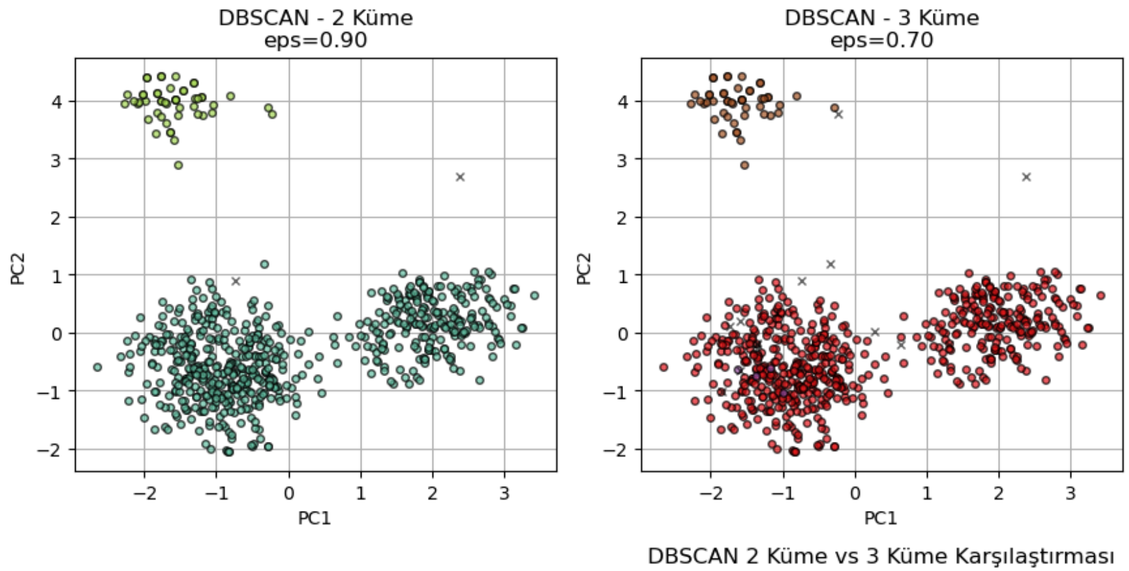

```
DBSCAN Algoritması Özeti:
- Toplam 2 küme bulundu
- 2 gürültü noktası tespit edildi
- Gürültü oranı: %0.3
- Silhouette Score: 0.519
```

- DBSCAN Clusters (PCA) chart shows 2 clusters (red, yellow) and noise (X) points; separation is clear, noise ratio is low.
- DBSCAN Cluster Distribution pie chart indicates clusters as 92.1% (cluster0), 7.6% (cluster1), and 0.3% noise.
- Avg_Credit/Limit Distribution bar chart shows cluster0 has a higher average credit limit, while cluster1 is lower.
- Total_Credit_Cards Distribution bar chart reveals cluster0 dominates in credit card numbers.
- "Total_visits_bank", "Total_visits_online", and "Total_calls_made" bar charts show cluster0 has higher frequencies across all categories.
- Conclusion: Cluster0 is the dominant group, cluster1 is smaller, and noise impact is minimal.

What happens if 3 clusters are selected?




```
3 küme için seçilen parametreler:
eps: 0.7000000000000001
min_samples: 5

3 Küme DBSCAN Sonuçları:
Gerçek küme sayısı: 3
Gürültü noktası sayısı: 15
Gürültü oranı: 2.3%
Silhouette Score: 0.108

Küme dağılımı:
Gürültü: 15 nokta
Küme 0: 592 nokta
Küme 1: 4 nokta
Küme 2: 49 nokta
```



Parametre	2 Küme DBSCAN	3 Küme DBSCAN
eps	0.900	0.700
min_samples	3	5
Küme Sayısı	2	3
Gürültü Sayısı	2	15
Gürültü Oranı	0.3%	2.3%
Silhouette Score	0.519	0.108

CONCLUSION

In this study, three different clustering algorithms (Hierarchical, K-Means, and DBSCAN) were compared. Based on the analyses and evaluation metrics, the K-Means algorithm provided the most successful results. Particularly for k=3, the average silhouette score reached its highest value of 0.517, and the clusters were clearly separated in the PCA plane. The Elbow method also supports this finding.

Hierarchical clustering produced visually meaningful results, but at k=3 and k=4 cuts, some clusters overlapped. The silhouette score was also lower compared to K-Means.

The DBSCAN algorithm successfully identified some noise points in the dataset but showed weaker performance with a low silhouette score and a smaller number of clusters.
Overall, K-Means was determined to be the most suitable clustering method for this dataset.