Homework 4 : 6D DATA HUNT

Algorithm

1. **Initialize Spark Session**:
    - Create a Spark session to process the data using Apache Spark.
2. **Load Data**:
    - Read the `.dat` file into an RDD.
    - Convert the RDD into a Spark DataFrame with 6 columns (`x1, x2, x3, x4, x5, x6`).
3. **Preprocess Data**:
    - Assemble the 6 columns into a single "features" column using `VectorAssembler`.
4. **Determine Optimal Clusters (k) Using Silhouette Score**:
    - For each `k` in the range 2 to 10:
        1. Perform K-Means clustering with `k` clusters.
        2. Calculate the silhouette score to evaluate the clustering.
    - Choose the `k` that gives the highest silhouette score.
5. **Perform K-Means Clustering**:
    - Fit the K-Means model with the optimal `k`.
    - Assign each data point to a cluster and store the cluster label.
6. **Analyze Each Cluster**:
    - For each cluster `i`:
        1. Count the number of points in the cluster.
        2. Calculate the bounding box (min, max) values in each dimension.
        3. Print the cluster center, count, size in each dimension, and bounding box.
7. **Apply PCA to Reduce Dimensions**:
    - For each cluster `i`, apply PCA to reduce the data from 6 dimensions to 3 dimensions for visualization.
    - Calculate the explained variance for each of the 6 principal components.
8. **Generate Visualizations**:
    - **Static 3D Plot**:
        1. Plot the cluster points in 3D using PCA results (PCA1, PCA2, PCA3).
        2. Display the explained variance ratios for the first 3 principal components.
        3. Save the plot as a PNG image.
    - **Interactive 3D Plot**:
        1. Create an interactive 3D scatter plot of the cluster points.
        2. Display the explained variance ratios.
        3. Save the interactive plot as an HTML file.
9. **Save Results**:
    - Save the static plots as PNG files.
    - Save the interactive plots as HTML files for each cluster.
10. **Stop Spark Session**:
    - Terminate the Spark session to free up resources.

Homework 4 : 6D DATA HUNT

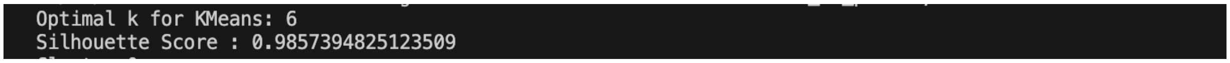
Code:

Is attached to the mail as 6D_Datahunt.py

Output and Inference:

1. Optimal k for KMeans:

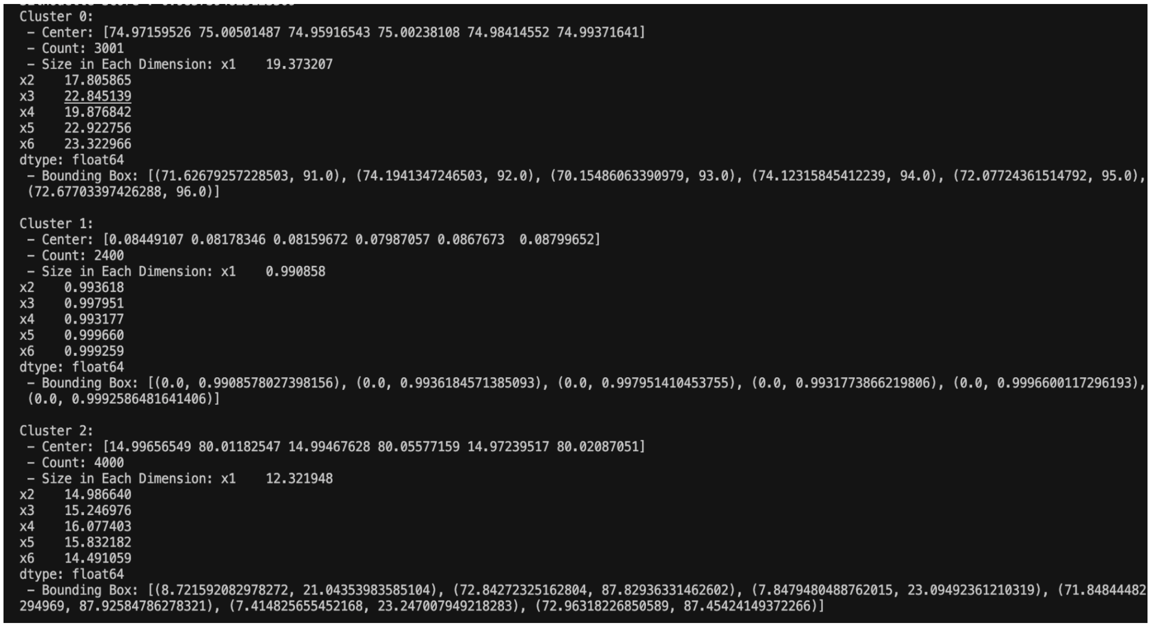
Output:



```
Optimal k for KMeans: 6
Silhouette Score : 0.9857394825123509
```

- This output indicates the best number of clusters (6) chosen based on the highest silhouette score(0.98), which measures how well-separated the clusters are. A higher silhouette score generally means the clusters are more distinct from each other.

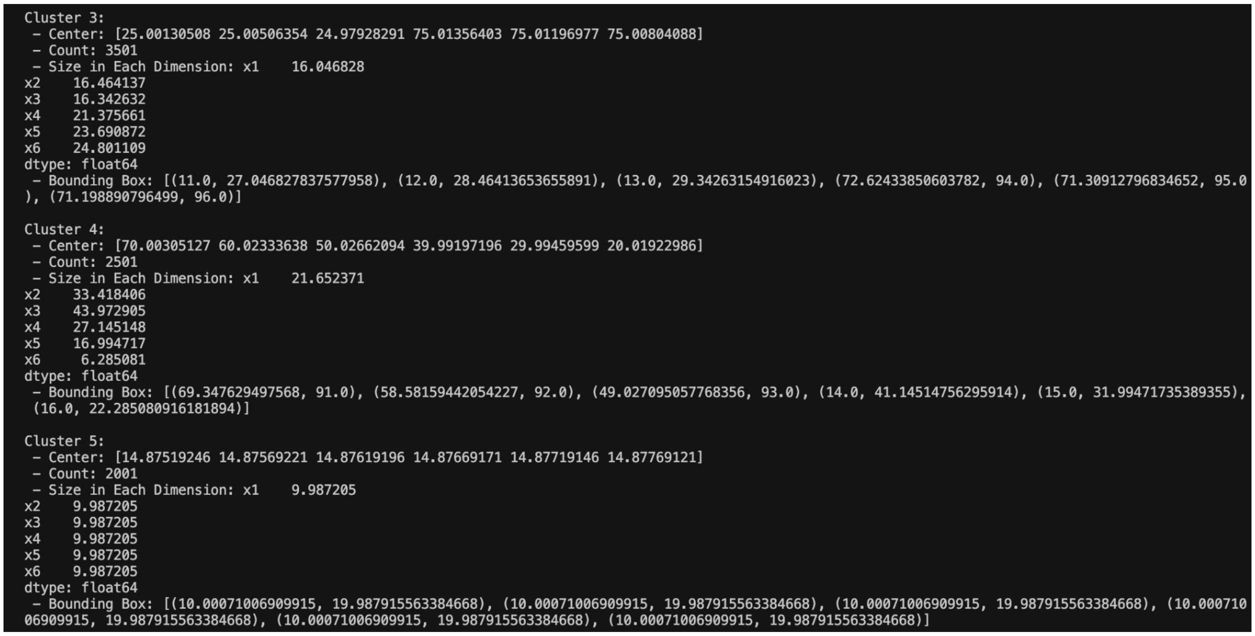
2. Cluster Shape and Information:

- **Output for each cluster** (e.g., Cluster 0, Cluster 1, etc.):
- **Center**: This shows the coordinates of the center of the cluster in the 6-dimensional space.
- **Count**: The number of data points that belong to the current cluster.
- **Size in Each Dimension**: The range of values (difference between maximum and minimum) in each of the 6 dimensions for the points in the cluster.
- **Bounding Box**: This represents the boundaries of the cluster in each of the 6 dimensions, indicating the minimum and maximum values along each axis (i.e., the 6D "shape" of the cluster).
- 

```
Cluster 0:
 - Center: [74.97159526 75.00501487 74.95916543 75.00238108 74.98414552 74.99371641]
 - Count: 3001
 - Size in Each Dimension: x1    19.373207
x2    17.805865
x3    22.845139
x4    19.876842
x5    22.922756
x6    23.322966
dtype: float64
 - Bounding Box: [(71.62679257228503, 91.0), (74.1941347246503, 92.0), (70.15486063390979, 93.0), (74.12315845412239, 94.0), (72.07724361514792, 95.0),
 (72.67703397426288, 96.0)]

Cluster 1:
 - Center: [0.08449107 0.08178346 0.08159672 0.07987057 0.0867673  0.08799652]
 - Count: 2400
 - Size in Each Dimension: x1    0.990858
x2    0.993618
x3    0.997951
x4    0.993177
x5    0.999660
x6    0.999259
dtype: float64
 - Bounding Box: [(0.0, 0.9908578027398156), (0.0, 0.9936184571385093), (0.0, 0.997951410453755), (0.0, 0.9931773866219806), (0.0, 0.9996600117296193),
 (0.0, 0.9992586481641406)]

Cluster 2:
 - Center: [14.99656549 80.01182547 14.99467628 80.05577159 14.97239517 80.02087051]
 - Count: 4000
 - Size in Each Dimension: x1    12.321948
x2    14.986640
x3    15.246976
x4    16.077403
x5    15.832182
x6    14.491059
dtype: float64
 - Bounding Box: [(8.721592082978272, 21.04353983585104), (72.84272325162804, 87.82936331462602), (7.8479480488762015, 23.09492361210319), (71.84844482
294969, 87.92584786278321), (7.414825655452168, 23.247007949218283), (72.96318226850589, 87.45424149372266)]
```



```
Cluster 3:
 - Center: [25.00130508 25.00506354 24.97928291 75.01356403 75.01196977 75.00804088]
 - Count: 3501
 - Size in Each Dimension: x1    16.046828
x2    16.464137
x3    16.342632
x4    21.375661
x5    23.690872
x6    24.801109
dtype: float64
 - Bounding Box: [(11.0, 27.046827837577958), (12.0, 28.46413653655891), (13.0, 29.34263154916023), (72.62433850603782, 94.0), (71.30912796834652, 95.0
), (71.198890796499, 96.0)]

Cluster 4:
 - Center: [70.00305127 60.02333638 50.02662094 39.99197196 29.99459599 20.01922986]
 - Count: 2501
 - Size in Each Dimension: x1    21.652371
x2    33.418406
x3    43.972905
x4    27.145148
x5    16.994717
x6     6.285081
dtype: float64
 - Bounding Box: [(69.347629497568, 91.0), (58.58159442054227, 92.0), (49.027095057768356, 93.0), (14.0, 41.14514756295914), (15.0, 31.99471735389355),
 (16.0, 22.285080916181894)]

Cluster 5:
 - Center: [14.87519246 14.87569221 14.87619196 14.87669171 14.87719146 14.87769121]
 - Count: 2001
 - Size in Each Dimension: x1    9.987205
x2    9.987205
x3    9.987205
x4    9.987205
x5    9.987205
x6    9.987205
dtype: float64
 - Bounding Box: [(10.00071006909915, 19.987915563384668), (10.00071006909915, 19.987915563384668), (10.00071006909915, 19.987915563384668), (10.000710
06909915, 19.987915563384668), (10.00071006909915, 19.987915563384668), (10.00071006909915, 19.987915563384668)]
```

3. **Explained Variance for PCA**:

- This shows how much of the total variance in the dataset is explained by each of the 6 principal components (PCA). Each percentage represents how much of the data's spread (variation) is captured by that specific principal component. The higher the percentage for a component, the more important it is for understanding the data's structure.
- If the value is close to zero in a particular component it means that object is not in that dimension

4. Static 3D PCA Plot:

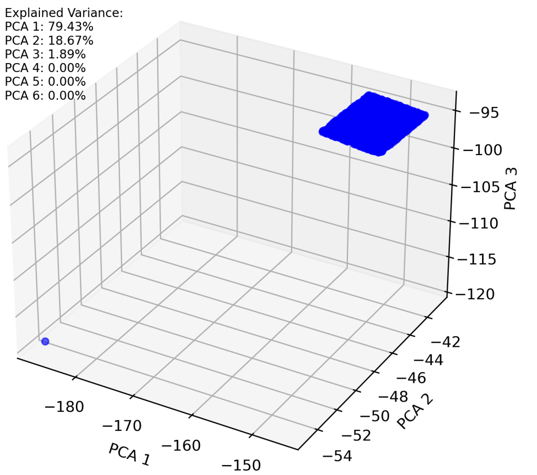
- **Output**: A 3D scatter plot (saved as a PNG file).
- The plot visualizes the data in the reduced 3D space after applying PCA. It shows the distribution of the points for each cluster using the first three principal components (`PCA1`, `PCA2`, `PCA3`).
- The cluster points are displayed in a 3D space with labeled axes (PCA1, PCA2, PCA3), and the plot includes a text annotation with the explained variance ratios for the first three principal components.

5. Interactive 3D Plot:

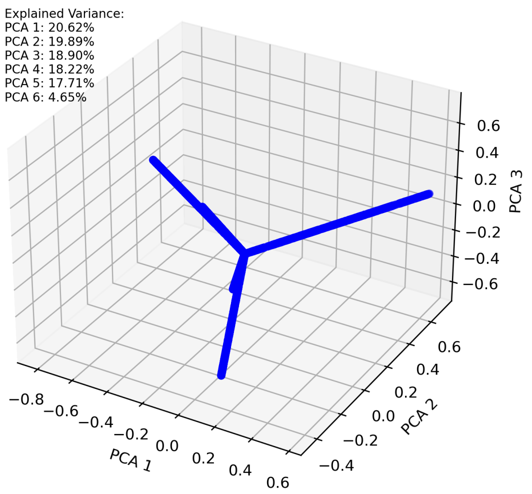
- **Output**: A dynamic 3D scatter plot (saved as an HTML file).
- This is an interactive version of the static 3D plot generated using Plotly. Users can rotate, zoom, and pan the plot to explore the cluster's distribution in the 3D space.

Homework 4 : 6D DATA HUNT

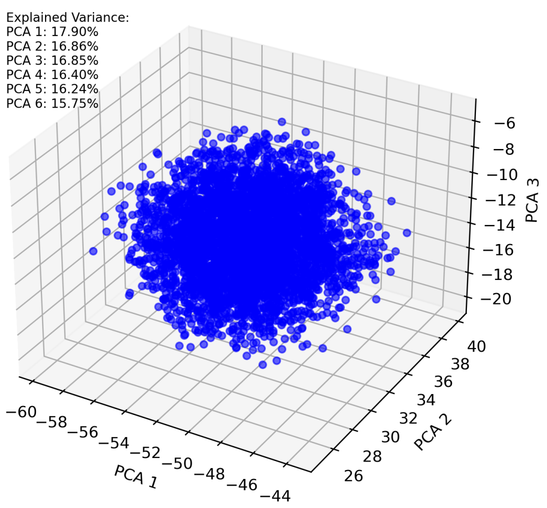
Cluster 0


Explained Variance:
PCA 1: 79.43%
PCA 2: 18.67%
PCA 3: 1.89%
PCA 4: 0.00%
PCA 5: 0.00%
PCA 6: 0.00%

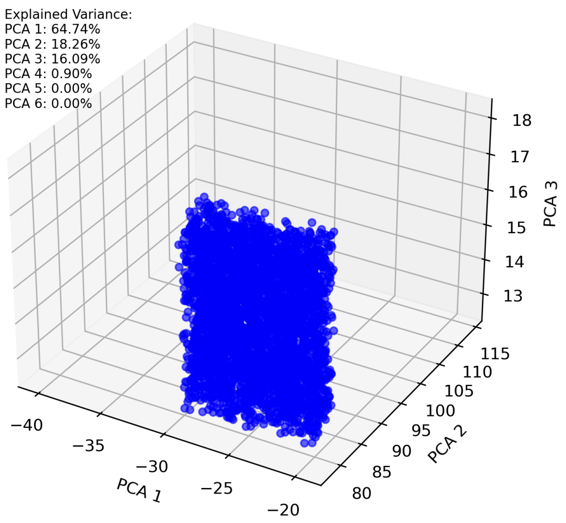
Cluster 1


Explained Variance:
PCA 1: 20.62%
PCA 2: 19.89%
PCA 3: 18.90%
PCA 4: 18.22%
PCA 5: 17.71%
PCA 6: 4.65%

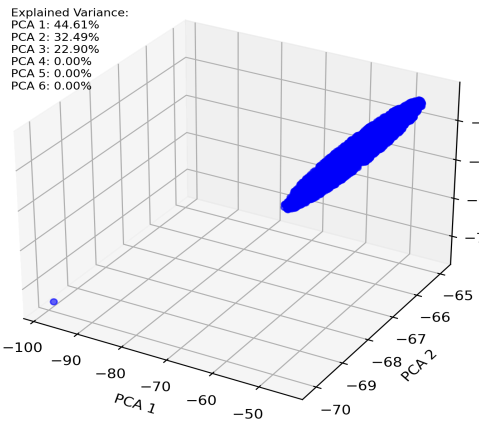
Cluster 2


Explained Variance:
PCA 1: 17.90%
PCA 2: 16.86%
PCA 3: 16.85%
PCA 4: 16.40%
PCA 5: 16.24%
PCA 6: 15.75%

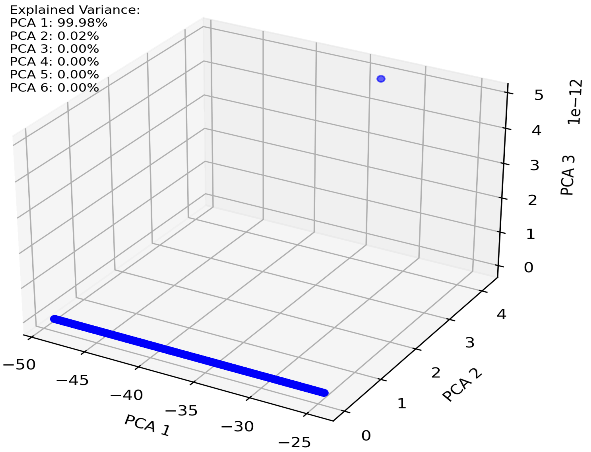
Cluster 3


Explained Variance:
PCA 1: 64.74%
PCA 2: 18.26%
PCA 3: 16.09%
PCA 4: 0.90%
PCA 5: 0.00%
PCA 6: 0.00%

Cluster 4


Explained Variance:
PCA 1: 44.61%
PCA 2: 32.49%
PCA 3: 22.90%
PCA 4: 0.00%
PCA 5: 0.00%
PCA 6: 0.00%

Cluster 5


Explained Variance:
PCA 1: 99.98%
PCA 2: 0.02%
PCA 3: 0.00%
PCA 4: 0.00%
PCA 5: 0.00%
PCA 6: 0.00%

Homework 4 : 6D DATA HUNT

Cluster	Object	Location	Size	Dimensions	Orientation	Count
0	Rectangle	[74.97159526 75.00501487 74.95916543 75.00238108 74.98414552 74.99371641]	**21-Length**(x1 19.373207,x2 17.805865,x3 22.845139 x4 19.876842,x5 22.922756,x6 23.322966)	2D	Parallel to all dimensions	3001
1	6 line(intersecting at origin)	[0.08449107 0.08178346 0.08159672 0.07987057 0.0867673 0.08799652]	**1-Length**(x1 0.990858,x2 0.993618,x3 0.997951 x4 0.993177,x5 0.999660,x6 0.999259)	6D	Oriented in all dimension	2400
2	6D - Ellipsoid	[14.99656549 80.01182547 14.99467628 80.05577159 14.97239517 80.02087051]	**15-Diameter**(x1 12.321948,x2 14.986640,x3 15.246976 x4 16.077403,x5 15.832182,x6 14.491059)	6D	Oriented in All dimension	4000
3	3D Cuboid	[25.00130508 25.00506354 24.97928291 75.01356403 75.01196977 75.00804088]	**LBH-10x10x20**(x1 16.046828,x2 16.464137,x3 16.342632 x4 21.375661,x5 23.690872,x6 24.801109)	3D	Parallel to 3 dimension and elongated along PC3	3501
4	2D Disc	[70.00305127 60.02333638 50.02662094 39.99197196 29.99459599 20.01922986]	**44-Diameter(**x1 21.652371,x2 33.418406,x3 43.972905 x4 27.145148,x5 16.994717,x6 6.285081)	2D(object is oriented in 3D)	Parallel to 3 dimension but is angled along pc1,2,3	2501
5	Line	[14.87519246 14.87569221 14.87619196 14.87669171 14.87719146 14.87769121]	**10-Length**(x1 9.987205,x2 9.987205,x3 9.987205 x4 9.987205,x5 9.987205,x6 9.987205)	1D	1 Dimension	2001

The necessary outputs are attained.