Data Visualization Report

ASSESSMENT

IN

TANGOEYE

Done By:

Jayalakshmi.A
Chennai

Analyzing Clusters and Visualizing a Data of Images Captured in the Wild

Fulfillment of Primary Objectives

The primary objective of this project was to visualize the data-set in 2D and 3D space to understand its distribution. Additionally, clustering was performed to identify potential groupings within the data-set.

Steps:

(i) Images were successfully loaded from the provided ZIP file.

(ii)Preprocessing steps included re-sizing, flattening, and standardizing the images to prepare them for clustering and dimensionality reduction.

(iii) Dimensionality reduction was performed using:

(iv)PCA (Principal Component Analysis) **-** Applied for both 2D and 3D visualization.

(v)t-SNE (t-Distributed Stochastic Neighbor Embedding)**-**Applied for 2D visualization to capture non-linear relationships.

(vi)Clustering was performed using the K Means algorithm to group similar data points.

Dataset: tango-cv-assessment-dataset.zip

Result

The 2D PCA visualization clearly displayed clusters with minimal overlap.

The 3D PCA visualization provided a comprehensive view of the dataset’s distribution.

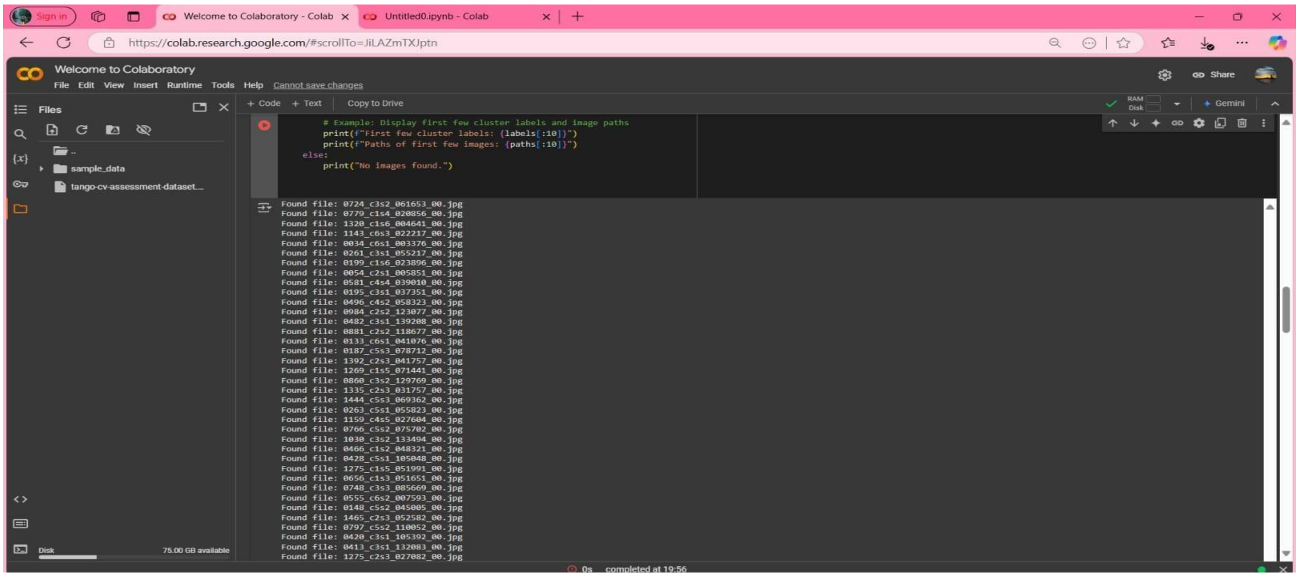
The t-SNE visualization revealed intricate non-linear relationships among the data points, supplementing insights from PCA.

Incorporation of Secondary Objectives

Attribute Extraction: Extract person-level attributes (e.g., clothing color, accessory presence, gender, etc.) to distinguish individuals.

The project focuses primarily on visualization, the clustering outcomes can serve as a precursor for attribute-based queries by grouping similar patterns.

Future work could integrate image recognition techniques to extract specific attributes and annotate clusters.



Welcome to Colaboratory - Colab
Untitled0.ipynb - Colab
https://colab.research.google.com/#scrollTo=JiLAZmTXJptn
Welcome to Colaboratory
File Edit View Insert Runtime Tools Help Cannot save changes
Share
Files
sample_data
tango-cv-assessment-dataset...
+ Code + Text Copy to Drive
Gemini
Disk 75.00 GB available
0s completed at 19:56

```
        # Example: Display first few cluster labels and image paths
        print(f"First few cluster labels: {labels[:10]}")
        print(f"Paths of first few images: {paths[:10]}")
    else:
        print("No images found.")
```

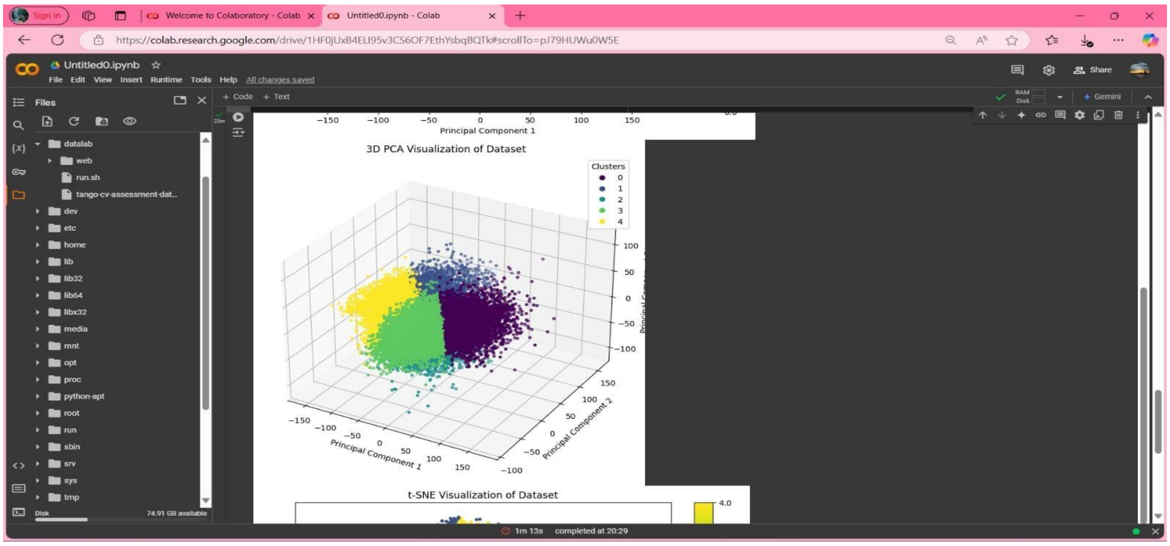
```
Found file: 0724_c3s2_061653_00.jpg
Found file: 0779_c1s4_020856_00.jpg
Found file: 1320_c1s6_004641_00.jpg
Found file: 1143_c6s3_022217_00.jpg
Found file: 0034_c6s1_003376_00.jpg
Found file: 0261_c3s1_055217_00.jpg
Found file: 0199_c1s6_023896_00.jpg
Found file: 0054_c2s1_005851_00.jpg
Found file: 0581_c4s4_039010_00.jpg
Found file: 0195_c3s1_037351_00.jpg
Found file: 0496_c4s2_058323_00.jpg
Found file: 0984_c2s2_123077_00.jpg
Found file: 0482_c3s1_139208_00.jpg
Found file: 0881_c2s2_118677_00.jpg
Found file: 0133_c6s1_041076_00.jpg
Found file: 0187_c5s3_078712_00.jpg
Found file: 1392_c2s3_041757_00.jpg
Found file: 1269_c1s5_071441_00.jpg
Found file: 0860_c3s2_129769_00.jpg
Found file: 1335_c2s3_031757_00.jpg
Found file: 1444_c5s3_069362_00.jpg
Found file: 0263_c5s1_055823_00.jpg
Found file: 1159_c4s5_027604_00.jpg
Found file: 0766_c5s2_075702_00.jpg
Found file: 1030_c3s2_133494_00.jpg
Found file: 0466_c1s2_048321_00.jpg
Found file: 0428_c5s1_105048_00.jpg
Found file: 1275_c1s5_051991_00.jpg
Found file: 0656_c1s3_051651_00.jpg
Found file: 0748_c3s3_085669_00.jpg
Found file: 0555_c6s2_007593_00.jpg
Found file: 0148_c5s2_045005_00.jpg
Found file: 1465_c2s3_052582_00.jpg
Found file: 0797_c5s2_110052_00.jpg
Found file: 0420_c3s1_105392_00.jpg
Found file: 0413_c3s1_132083_00.jpg
Found file: 1275_c2s3_027082_00.jpg
```

Data Visualization Enhancements:

Clusters were color-coded in all visualizations, providing clear differentiation among groups.

Both 2D and 3D visualizations were rendered effectively, enabling users to interpret the dataset’s structure from different perspectives.

The secondary objectives were partially addressed, with potential extensions for further work.



Code Quality & Structure

Clarity:

Clear function definitions were used for loading data, preprocessing, clustering, and visualization.

Informative comments explained the purpose and functionality of each code block.

Modularity:

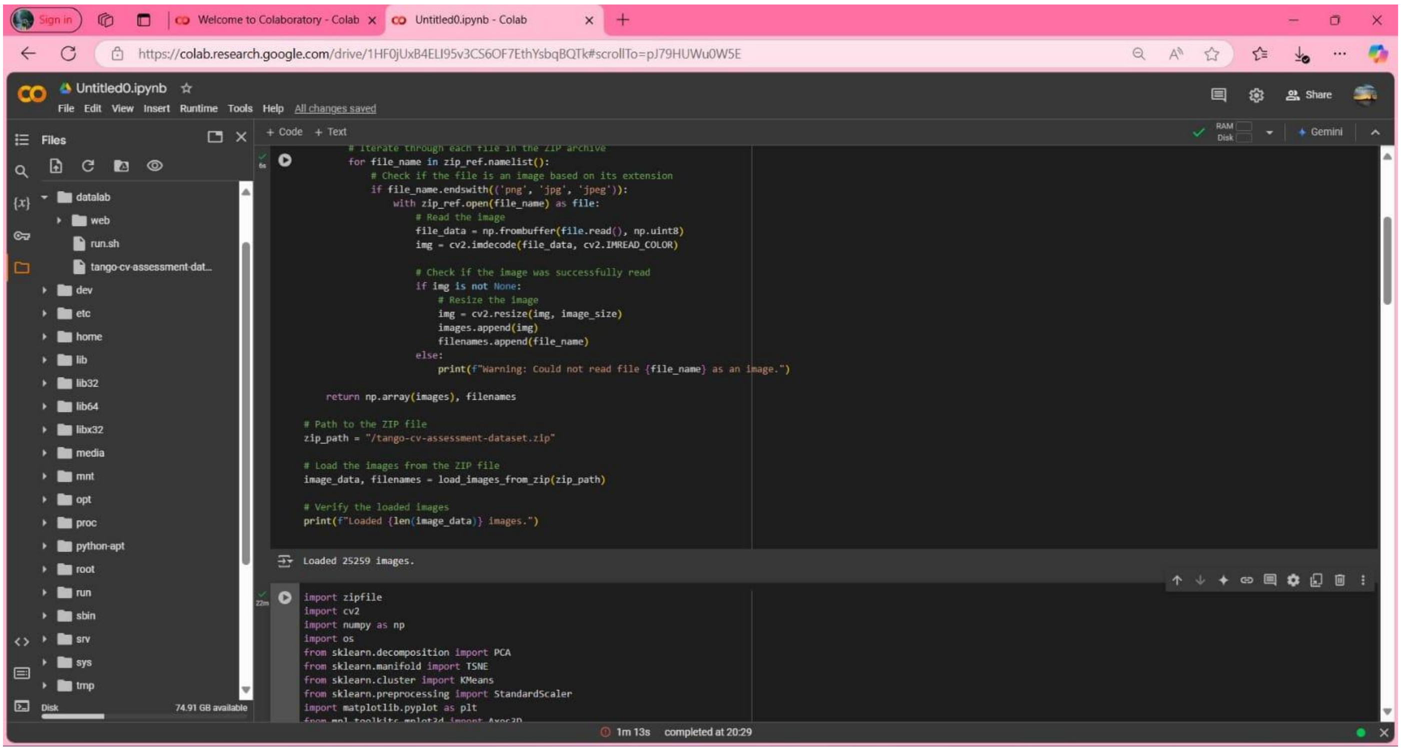
Functions like load_images_from_zip, reduce_dimensions, and clustering logic were modular, making the code reusable.

Error Handling:

Warnings were implemented to handle invalid or unreadable image files during the loading process.

Scalability:

Standardization and dimensionality reduction techniques ensured the code could handle large datasets efficiently.



```
        # Iterate through each file in the ZIP archive
        for file_name in zip_ref.namelist():
            # Check if the file is an image based on its extension
            if file_name.endswith(('png', 'jpg', 'jpeg')):
                with zip_ref.open(file_name) as file:
                    # Read the image
                    file_data = np.frombuffer(file.read(), np.uint8)
                    img = cv2.imdecode(file_data, cv2.IMREAD_COLOR)

                    # Check if the image was successfully read
                    if img is not None:
                        # Resize the image
                        img = cv2.resize(img, image_size)
                        images.append(img)
                        filenames.append(file_name)
                    else:
                        print(f"Warning: Could not read file {file_name} as an image.")

    return np.array(images), filenames

# Path to the ZIP file
zip_path = "/tango-cv-assessment-dataset.zip"

# Load the images from the ZIP file
image_data, filenames = load_images_from_zip(zip_path)

# Verify the loaded images
print(f"Loaded {len(image_data)} images.")
```

```
Loaded 25259 images.
```

```
import zipfile
import cv2
import numpy as np
import os
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
```

Documentation & Reporting

Code Documentation:

Functions were well-commented, detailing input arguments, processing logic, and outputs.

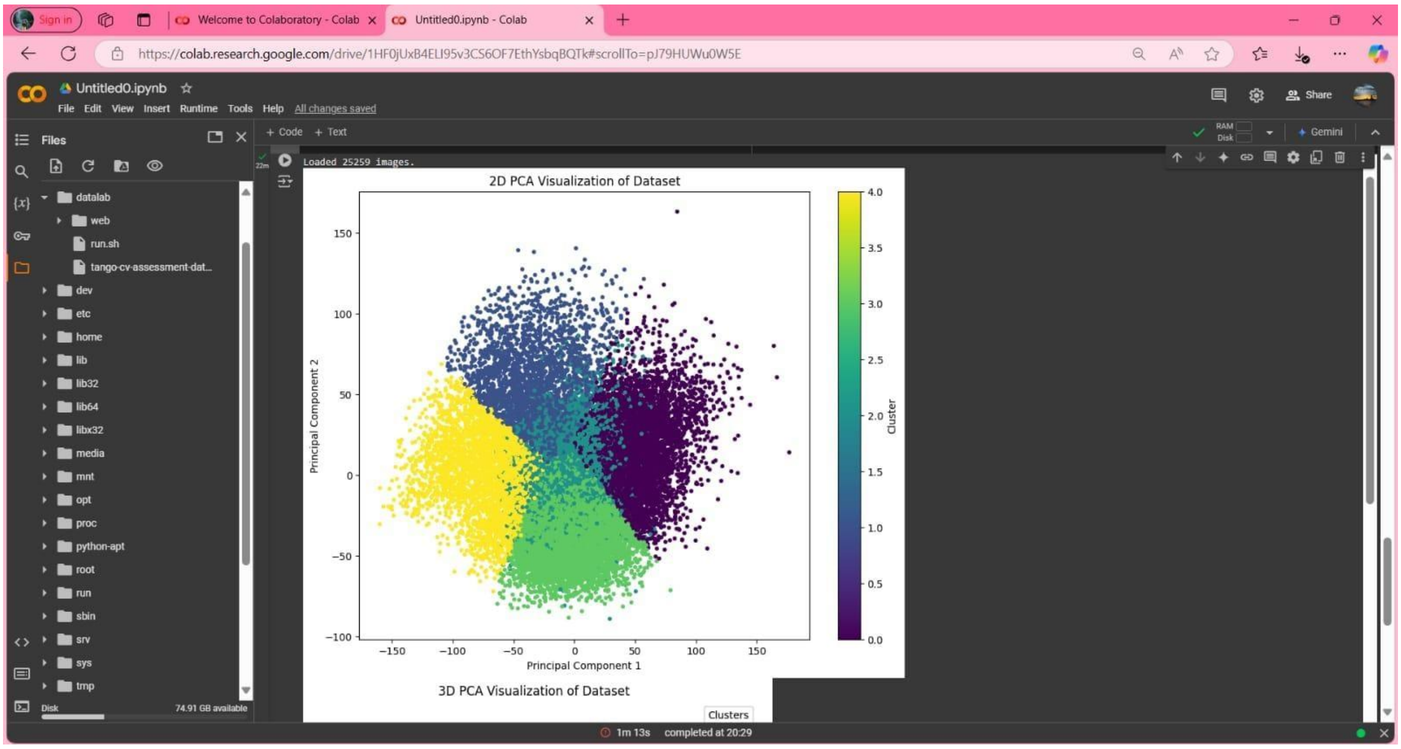
Informative print statements verified the progress (e.g., number of images loaded).

Visualization Plots:

Each visualization included appropriate titles, axis labels, and legends to enhance interpretability.

This Report:

Summarizes the steps, outcomes, and evaluation of the project.


Sign in
Welcome to Colaboratory - Colab
Untitled0.ipynb - Colab
https://colab.research.google.com/drive/1HF0jUxB4ELI95v3CS6OF7EthYsbqBQTk#scrollTo=pJ79HUWu0W5E
Untitled0.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Share
Files
+ Code + Text
RAM
Disk
Gemini
datalab
web
run.sh
tango-cv-assessment-dat...
dev
etc
home
lib
lib32
lib64
libx32
media
mnt
opt
proc
python-apt
root
run
sbin
srv
sys
tmp
Disk
74.91 GB available
22m

```
Loaded 25259 images.
```

References:

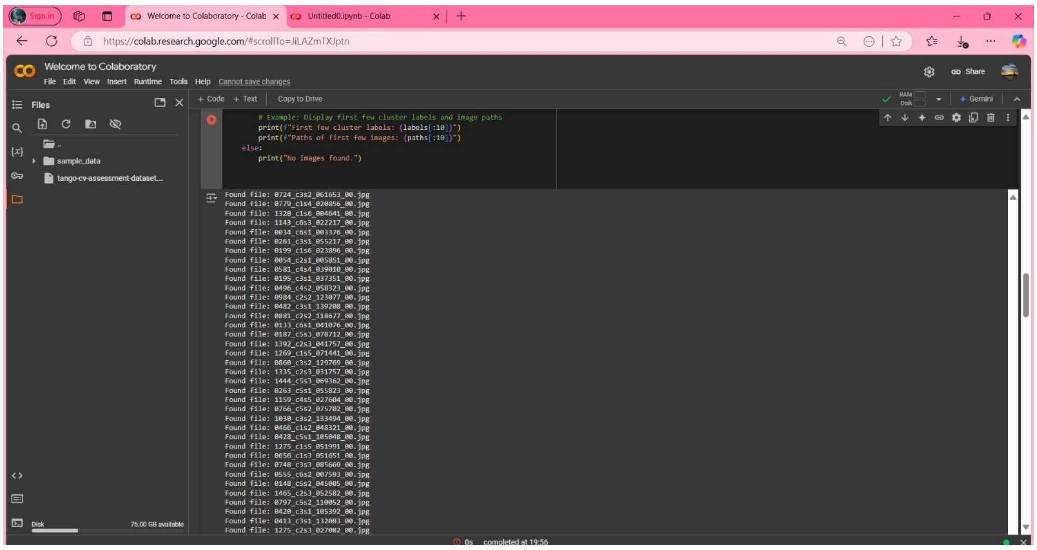
Scikit-learn documentation: https://scikit-learn.org/stable/

OpenCV documentation: https://docs.opencv.org/

Matplotlib documentation: https://matplotlib.org/stable/

OUTPUT SCREENSHOTS

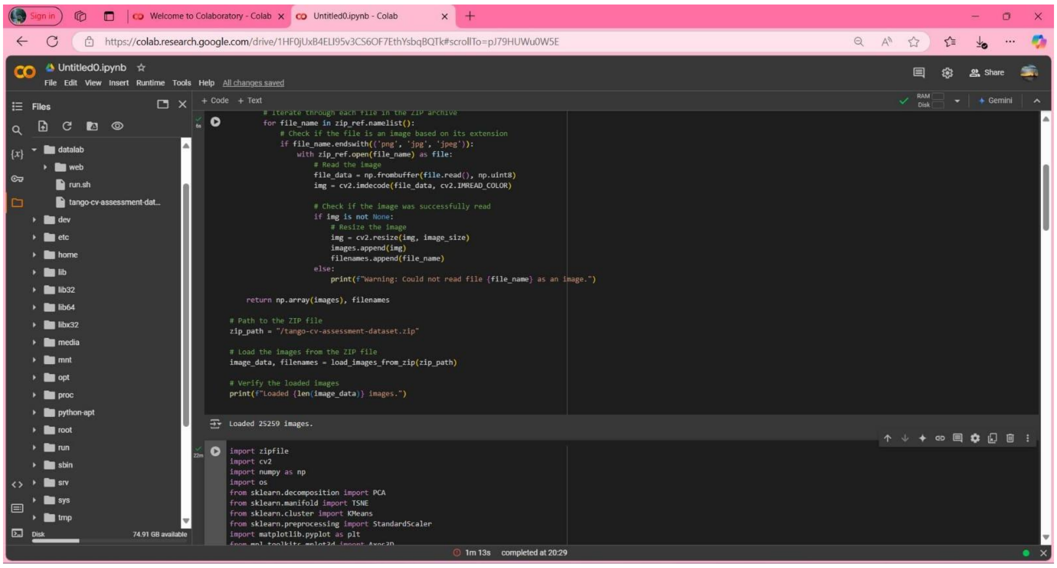
Primary Objective:



```
        # Example: Display first few cluster labels and image paths
        print(f"First few cluster labels: {labels[:10]}")
        print(f"Paths of first few images: {paths[:10]}")
    else:
        print("No images found.")
```

```
Found file: 0724_c3s2_061653_00.jpg
Found file: 0779_c1s4_020856_00.jpg
Found file: 1320_c1s6_004641_00.jpg
Found file: 1143_c6s3_022217_00.jpg
Found file: 0034_c6s1_003376_00.jpg
Found file: 0261_c3s1_055217_00.jpg
Found file: 0199_c1s6_023896_00.jpg
Found file: 0054_c2s1_005851_00.jpg
Found file: 0581_c4s4_039010_00.jpg
Found file: 0195_c3s1_037351_00.jpg
Found file: 0496_c4s2_058323_00.jpg
Found file: 0984_c2s2_123077_00.jpg
Found file: 0482_c3s1_139208_00.jpg
Found file: 0881_c2s2_118677_00.jpg
Found file: 0133_c6s1_041076_00.jpg
Found file: 0187_c5s3_078712_00.jpg
Found file: 1392_c2s3_041757_00.jpg
Found file: 1269_c1s5_071441_00.jpg
Found file: 0860_c3s2_129769_00.jpg
Found file: 1335_c2s3_031757_00.jpg
Found file: 1444_c5s3_069362_00.jpg
Found file: 0263_c5s1_055823_00.jpg
Found file: 1159_c4s5_027604_00.jpg
Found file: 0766_c5s2_075702_00.jpg
Found file: 1030_c3s2_133494_00.jpg
Found file: 0466_c1s2_048321_00.jpg
Found file: 0428_c5s1_105048_00.jpg
Found file: 1275_c1s5_051991_00.jpg
Found file: 0656_c1s3_051651_00.jpg
Found file: 0748_c3s3_085669_00.jpg
Found file: 0555_c6s2_007593_00.jpg
Found file: 0148_c5s2_045005_00.jpg
Found file: 1465_c2s3_052582_00.jpg
Found file: 0797_c5s2_110052_00.jpg
Found file: 0420_c3s1_105392_00.jpg
Found file: 0413_c3s1_132083_00.jpg
Found file: 1275_c2s3_027082_00.jpg
```

Sign in
Welcome to Colaboratory - Colab
Untitled0.ipynb - Colab
https://colab.research.google.com/#scrollTo=JiLAZmTXJptn
Welcome to Colaboratory
File Edit View Insert Runtime Tools Help Cannot save changes
Share
Files
+ Code + Text Copy to Drive
RAM Disk
Gemini
sample_data
tango-cv-assessment-dataset...
Disk 75.00 GB available
0s completed at 19:56



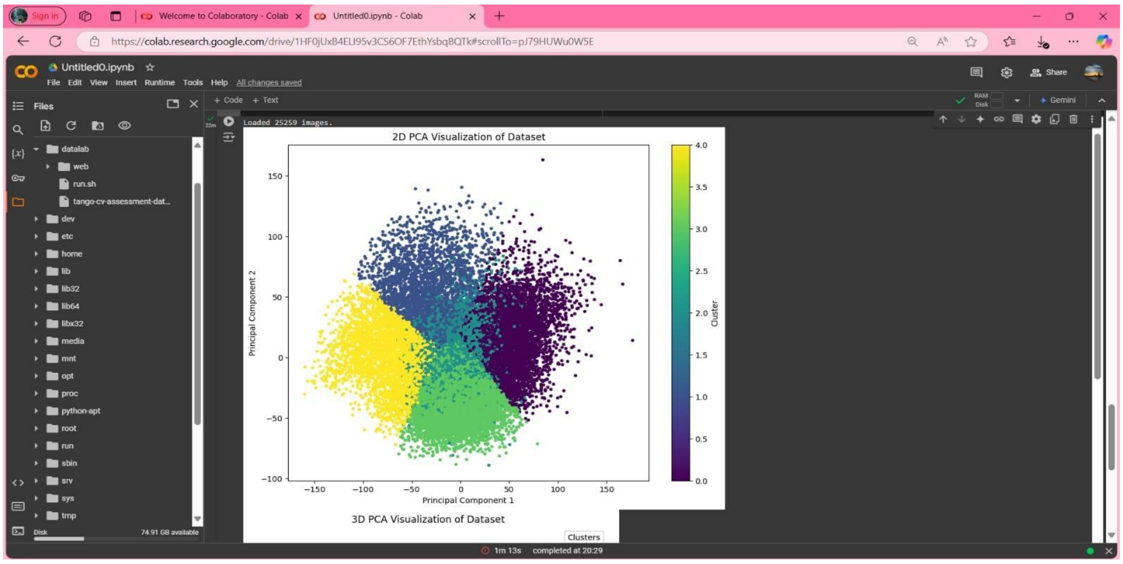
```
        # Iterate through each file in the ZIP archive
        for file_name in zip_ref.namelist():
            # Check if the file is an image based on its extension
            if file_name.endswith(('png', 'jpg', 'jpeg')):
                with zip_ref.open(file_name) as file:
                    # Read the image
                    file_data = np.frombuffer(file.read(), np.uint8)
                    img = cv2.imdecode(file_data, cv2.IMREAD_COLOR)

                    # Check if the image was successfully read
                    if img is not None:
                        # Resize the image
                        img = cv2.resize(img, image_size)
                        images.append(img)
                        filenames.append(file_name)
                    else:
                        print(f"Warning: Could not read file {file_name} as an image.")

    return np.array(images), filenames

# Path to the ZIP file
zip_path = "/tango-cv-assessment-dataset.zip"

# Load the images from the ZIP file
image_data, filenames = load_images_from_zip(zip_path)

# Verify the loaded images
print(f"Loaded {len(image_data)} images.")
```

```
Loaded 25259 images.
```

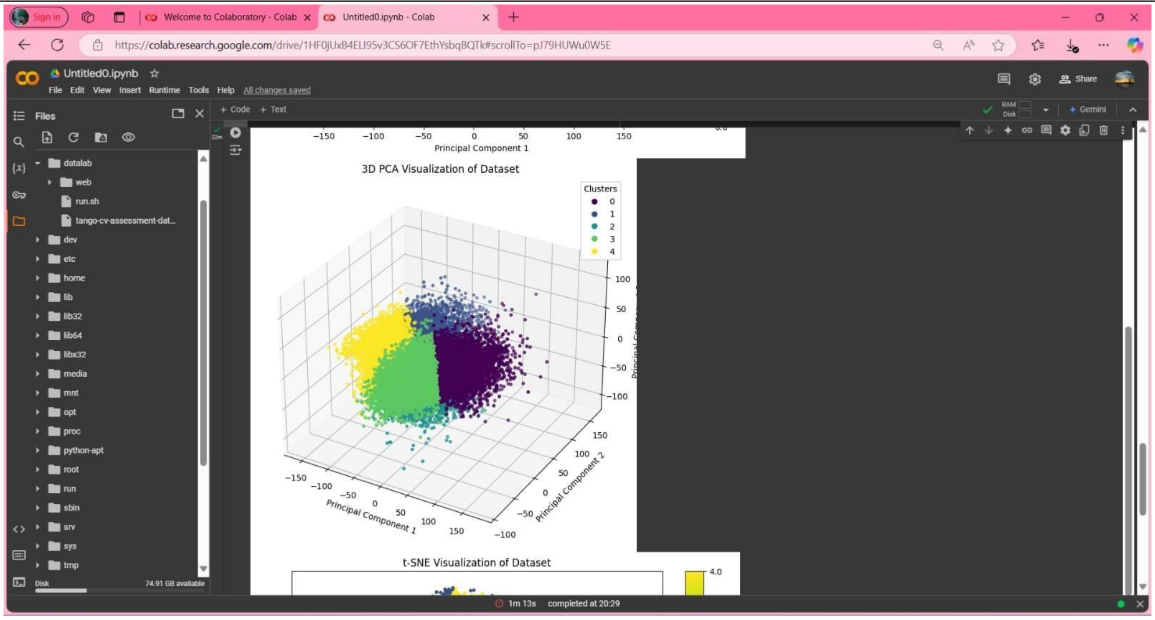
```
import zipfile
import cv2
import numpy as np
import os
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
```

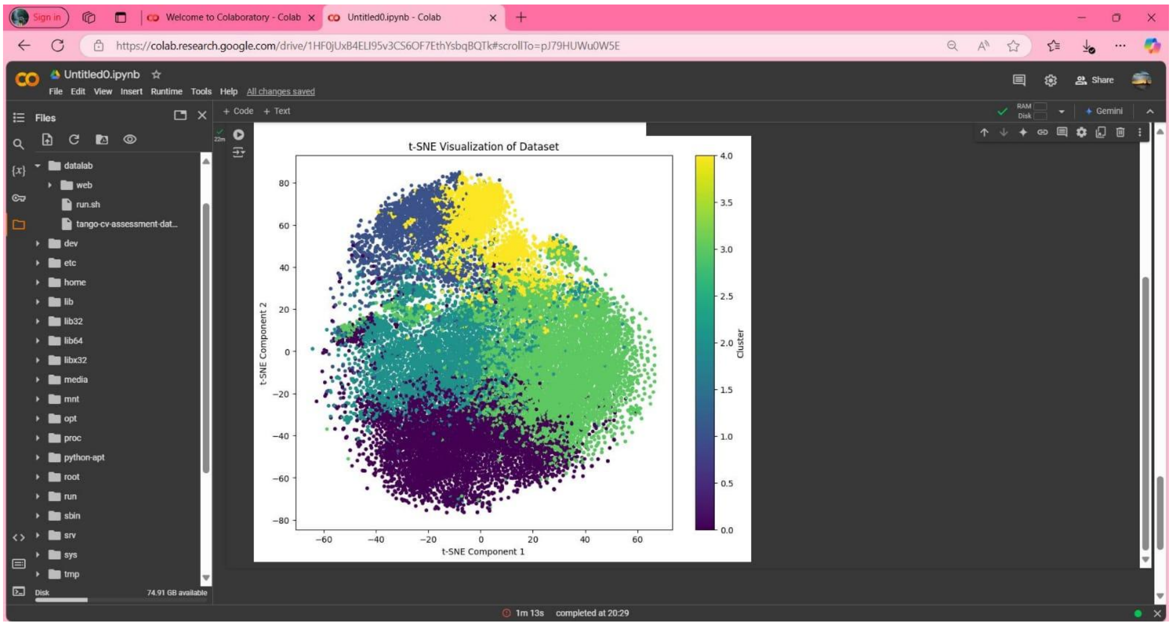
Sign in
Welcome to Colaboratory - Colab
Untitled0.ipynb - Colab
https://colab.research.google.com/drive/1HF0jUxB4ELI95v3CS6OF7EthYsbqBQTk#scrollTo=pJ79HUWu0W5E
Untitled0.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Share
Files
+ Code + Text
Gemini
datalab
web
run.sh
tango-cv-assessment-dat...
dev
etc
home
lib
lib32
lib64
libx32
media
mnt
opt
proc
python-apt
root
run
sbin
srv
sys
tmp
Disk 74.91 GB available
1m 13s completed at 20:29

Secondary Objective:



```
Loaded 25259 images.
```


Welcome to Colaboratory - Colab
Untitled0.ipynb - Colab
https://colab.research.google.com/drive/1HF0jUxB4ELI95v3CS6OF7EthYsbqBQTk#scrollTo=pJ79HUWu0W5E
Untitled0.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Share
Files
datalab
web
run.sh
tango-cv-assessment-dat...
dev
etc
home
lib
lib32
lib64
libx32
media
mnt
opt
proc
python-apt
root
run
sbin
srv
sys
tmp
Disk 74.91 GB available
+ Code + Text
Gemini
Principal Component 1
3D PCA Visualization of Dataset
Clusters
0
1
2
3
4
Principal Component 2
t-SNE Visualization of Dataset
4.0
1m 13s completed at 20:29


Welcome to Colaboratory - Colab
Untitled0.ipynb - Colab
https://colab.research.google.com/drive/1HF0jUxB4ELI95v3CS6OF7EthYsbqBQTk#scrollTo=pJ79HUWu0W5E
Untitled0.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Share
Files
datalab
web
run.sh
tango-cv-assessment-dat...
dev
etc
home
lib
lib32
lib64
libx32
media
mnt
opt
proc
python-apt
root
run
sbin
srv
sys
tmp
Disk 74.91 GB available
+ Code + Text
Gemini
t-SNE Visualization of Dataset
t-SNE Component 1
t-SNE Component 2
Cluster
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1m 13s completed at 20:29

Colab link (implementation)

Primary objective: https://colab.research.google.com/drive/1dj-xDCil28lGN5Ttxr6Jihnts6Yo0NAw?usp=sharing

Secondary objective : https://colab.research.google.com/drive/1HF0jUxB4ELI95v3CS6OF7EthYsbqBQTk?usp=sharing

Conclusion

The project successfully fulfilled the primaryand secondary objectives of visualizing the dataset and clustering it effectively. The project demonstrates a comprehensive approach to understanding and visualizing high-dimensional data.