[6]:
```
import glob
import pandas as pd

# Get list of matching files
file_list = glob.glob("/Users/admin/Documents/AMS/clean_datasets/DMC2_S*.csv")

# Read and combine them
df = pd.concat([pd.read_csv(file) for file in file_list], ignore_index=True)

# View combined shape and head
print(df.shape)
df.head()
```

```
(305929, 56)
```

[6]:

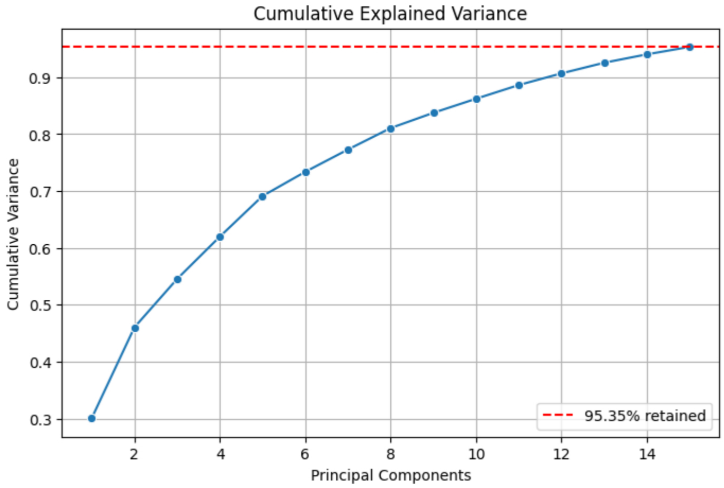
	LOAD|1	LOAD|2	LOAD|3	LOAD|6	ENC_POS|1	ENC_POS|2	ENC_POS|3	ENC_POS|6	CTRL_DIFF2|1	CTRL_DIFF2|2	...	ENC1_POS|3	ENC1_POS|6	ENC
0	0.341797	6.475830	0.421143	0.0	22.200148	284.270207	483.336283	181.730994	-0.00001	-0.000010	...	483.319422	181.730994	22
1	0.341797	6.481934	0.415039	0.0	22.200139	284.270197	483.336283	181.730994	0.00000	0.000000	...	483.319422	181.730994	22
2	0.341797	6.481934	0.415039	0.0	22.200139	284.270197	483.336283	181.730994	0.00000	0.000000	...	483.319418	181.730994	22
3	0.341797	6.481934	0.421143	0.0	22.200139	284.270197	483.336273	181.730994	0.00000	0.000000	...	483.319418	181.730994	22
4	0.341797	6.481934	0.421143	0.0	22.200139	284.270227	483.336273	181.730994	0.00000	-0.000029	...	483.319420	181.730994	22

5 rows × 56 columns

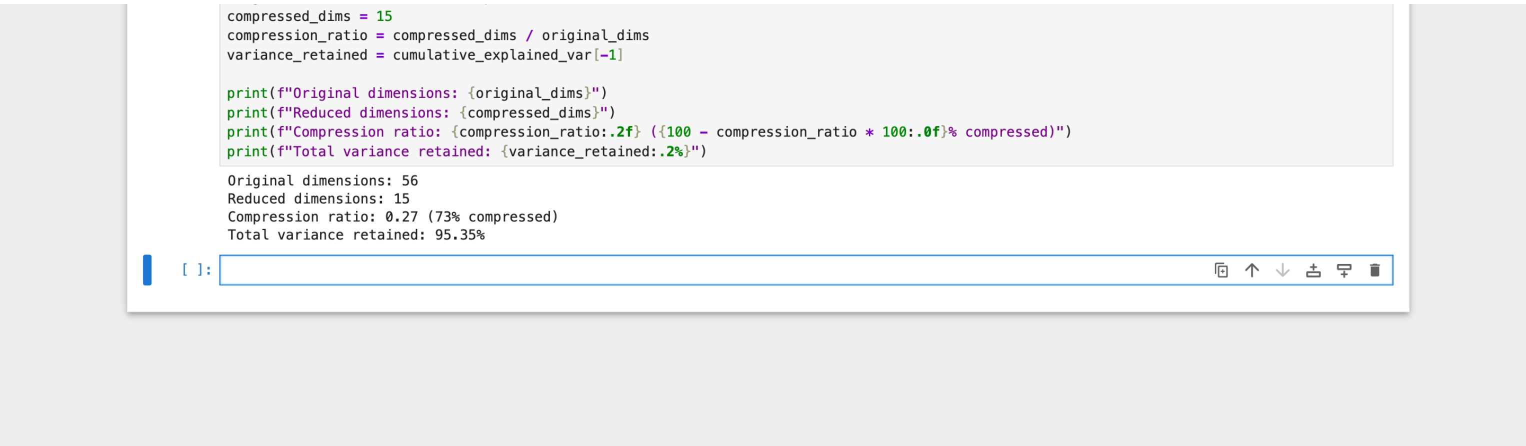
[7]:
```
dfs = []
for file in file_list:
    df_single = pd.read_csv(file)
    print(f"Read file: {file}, Shape: {df_single.shape}")
    dfs.append(df_single)

# Combine into one
combined_df = pd.concat(dfs, ignore_index=True)
print(f"Combined shape: {combined_df.shape}")
```

```
Read file: /Users/admin/Documents/AMS/clean_datasets/DMC2_S_CP2.csv, Shape: (42016, 56)
Read file: /Users/admin/Documents/AMS/clean_datasets/DMC2_S_CP1.csv, Shape: (263913, 56)
Combined shape: (305929, 56)
```

[8]:
```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Select only numeric columns and drop NaNs
df_numeric = combined_df.select_dtypes(include=['float64', 'int64']).dropna()

# Standardize
scaler = StandardScaler()
scaled_data = scaler.fit_transform(df_numeric)

# PCA with 15 components
pca = PCA(n_components=15)
pca_result = pca.fit_transform(scaled_data)

# Cumulative variance plot
cumulative_explained_var = np.cumsum(pca.explained_variance_ratio_)
plt.figure(figsize=(8, 5))
sns.lineplot(x=range(1, 16), y=cumulative_explained_var, marker='o')
plt.axhline(cumulative_explained_var[-1], color='red', linestyle='--', label=f'{cumulative_explained_var[-1]:.2%} retained')
plt.title("Cumulative Explained Variance")
plt.xlabel("Principal Components")
plt.ylabel("Cumulative Variance")
plt.legend()
plt.grid(True)
plt.show()
```



[9]:
```
original_dims = scaled_data.shape[1]
```



```
compressed_dims = 15
compression_ratio = compressed_dims / original_dims
variance_retained = cumulative_explained_var[-1]

print(f"Original dimensions: {original_dims}")
print(f"Reduced dimensions: {compressed_dims}")
print(f"Compression ratio: {compression_ratio:.2f} ({100 - compression_ratio * 100:.0f}% compressed)")
print(f"Total variance retained: {variance_retained:.2%}")
```

```
Original dimensions: 56
Reduced dimensions: 15
Compression ratio: 0.27 (73% compressed)
Total variance retained: 95.35%
```