In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.ensemble import IsolationForest
from sklearn.covariance import EllipticEnvelope
from scipy import stats
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.tools.tools import add_constant
```

In [ ]:
```
def comprehensive_analysis(df):
    # 1. Basic statistics and missing values
    print("Basic Statistics:")
    print(df.describe())
    print("\nMissing Values:")
    print(df.isnull().sum())

    # 2. Correlation analysis
    corr_matrix = df.corr()
    plt.figure(figsize=(20, 16))
    sns.heatmap(corr_matrix, cmap='coolwarm', annot=False)
    plt.title('Correlation Heatmap')
    plt.tight_layout()
    plt.show()

    # 3. Identify highly correlated features
    high_corr = np.where(np.abs(corr_matrix) > 0.8)
    high_corr_pairs = [(corr_matrix.index[x], corr_matrix.columns[y]) for x, y in zip(*high_corr) if x != y and x < y
    print("\nHighly correlated feature pairs:")
    for pair in high_corr_pairs:
        print(f"{pair[0]} - {pair[1]}: {corr_matrix.loc[pair[0], pair[1]]:.2f}")

    # 4. Variance Inflation Factor (VIF) for multicollinearity
    X = add_constant(df)
    vif_data = pd.DataFrame()
    vif_data["feature"] = X.columns
    vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
    print("\nVariance Inflation Factors:")
    print(vif_data.sort_values('VIF', ascending=False))

    # 5. PCA for dimensionality assessment
    scaler = StandardScaler()
    scaled_data = scaler.fit_transform(df)
    pca = PCA()
    pca_result = pca.fit_transform(scaled_data)

    cumulative_variance_ratio = np.cumsum(pca.explained_variance_ratio_)
    plt.figure(figsize=(10, 6))
    plt.plot(cumulative_variance_ratio)
    plt.xlabel('Number of Components')
    plt.ylabel('Cumulative Explained Variance Ratio')
    plt.title('PCA Cumulative Explained Variance')
    plt.tight_layout()
    plt.show()

    # 6. Scree plot
    plt.figure(figsize=(10, 6))
    plt.plot(pca.explained_variance_ratio_)
    plt.xlabel('Principal Component')
    plt.ylabel('Explained Variance Ratio')
    plt.title('Scree Plot')
    plt.tight_layout()
    plt.show()

    # 7. Outlier detection using multiple methods
    # Isolation Forest
    iso_forest = IsolationForest(contamination=0.1, random_state=42)
    outliers_iso = iso_forest.fit_predict(scaled_data)

    # Elliptic Envelope
    ee = EllipticEnvelope(contamination=0.1, random_state=42)
```

```
    outliers_ee = ee.fit_predict(scaled_data)

    # Z-score
    z_scores = np.abs(stats.zscore(df))
    outliers_z = np.where(z_scores > 3)

    print(f"\nNumber of potential outliers (Isolation Forest): {sum(outliers_iso == -1)}")
    print(f"Number of potential outliers (Elliptic Envelope): {sum(outliers_ee == -1)}")
    print(f"Number of potential outliers (Z-score > 3): {len(outliers_z[0])}")

    # 8. Skewness and Kurtosis analysis
    skewness = df.skew()
    kurtosis = df.kurtosis()
    print("\nSkewness of features:")
    print(skewness)
    print("\nKurtosis of features:")
    print(kurtosis)

    # 9. Feature-to-feature relationships (sample for efficiency)
    sample_df = df.sample(min(1000, len(df)))
    sns.pairplot(sample_df, diag_kind='kde', plot_kws={'alpha': 0.2})
    plt.tight_layout()
    plt.show()

    # 10. Noise-to-Signal Ratio (approximation using PCA)
    total_variance = np.sum(pca.explained_variance_)
    noise_variance = np.sum(pca.explained_variance_[int(0.95 * len(pca.explained_variance_)):])
    nsr = noise_variance / (total_variance - noise_variance)
    print(f"\nApproximate Noise-to-Signal Ratio: {nsr:.4f}")
```

To interpret the results:

Multicollinearity:

Look for high correlations in the heatmap and the list of highly correlated pairs. Check VIF values. VIF > 5 suggests moderate multicollinearity, while VIF > 10 indicates severe multicollinearity. In the PCA results, if a small number of components explain most of the variance, it suggests high multicollinearity.

Noisiness:

Check the number of outliers detected by different methods. A high number of outliers might indicate noisy data. Look for features with high skewness or kurtosis, which might indicate noise or the need for transformation. In the PCA results, if many components are needed to explain most of the variance, it might indicate noisy data. A high Noise-to-Signal Ratio suggests noisier data.

Overall data quality:

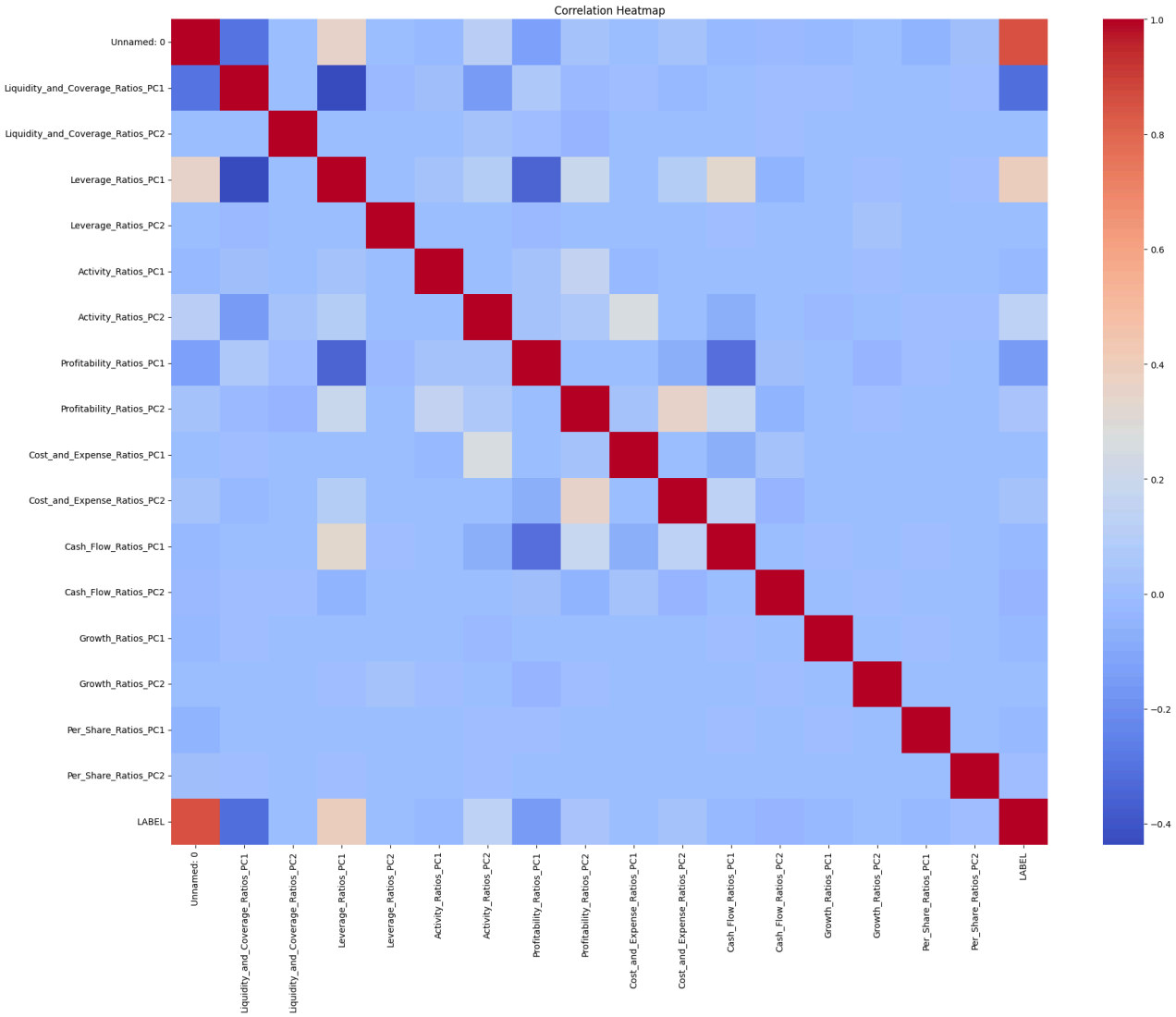
Examine the pairplots for unexpected patterns or inconsistencies. Check if the cumulative explained variance in PCA reaches a high level (e.g., 95%) with a reasonable number of components.

This comprehensive analysis will give you a thorough understanding of the multicollinearity and noisiness in your high-dimensional financial dataset. Based on the results, you can make informed decisions about feature selection, dimensionality reduction, or data transformation techniques to address any issues identified. C

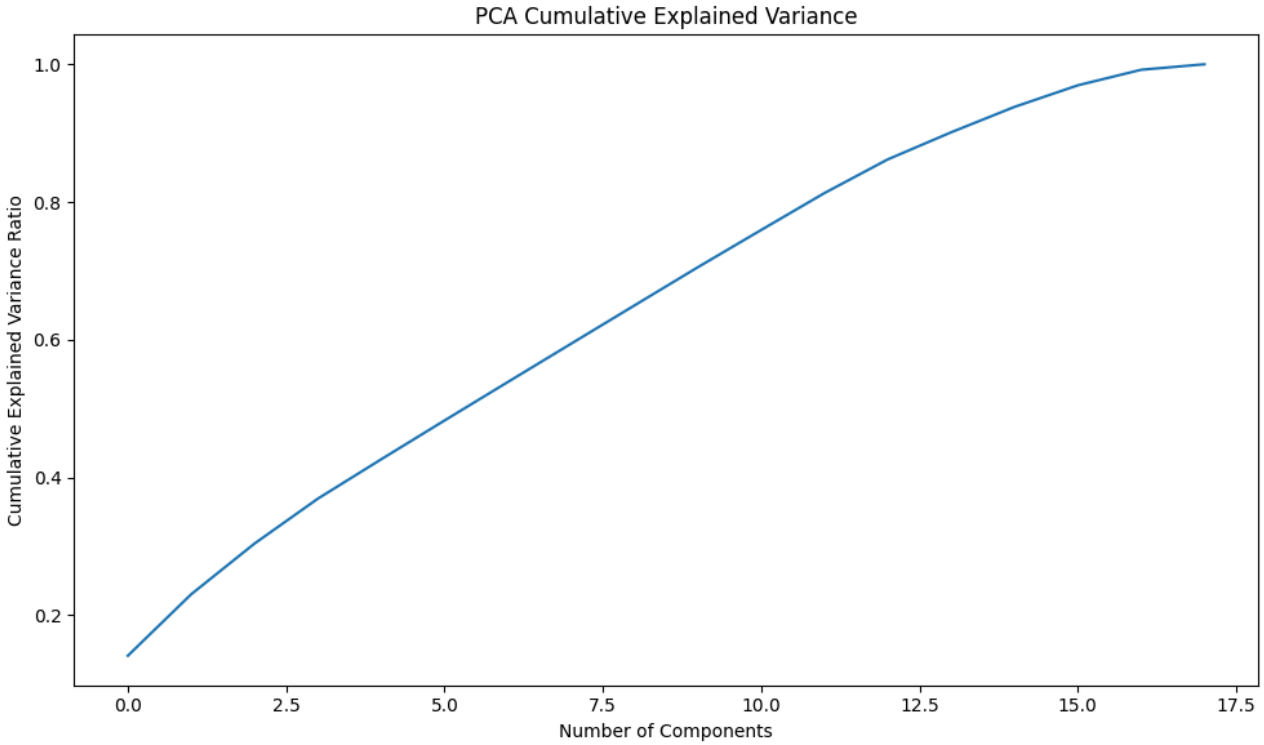
In [ ]:
```
df_mice = pd.read_excel("C:\\Users\\dev\\Desktop\\Msc thesis Prior RS\\ML training\\df_mice_labeled_after_PCA.xlsx")
df_AE = pd.read_excel("C:\\Users\\dev\\Desktop\\Msc thesis Prior RS\\ML training\\df_autoencoder_labeled_after_PCA.x


# Run the comprehensive analysis
print("Analysis of Df_mice:")
comprehensive_analysis(df_mice)
print(" ")
print(" ")
print(" ")

print("Analysis of Df_AE:")
comprehensive_analysis(df_AE)
```

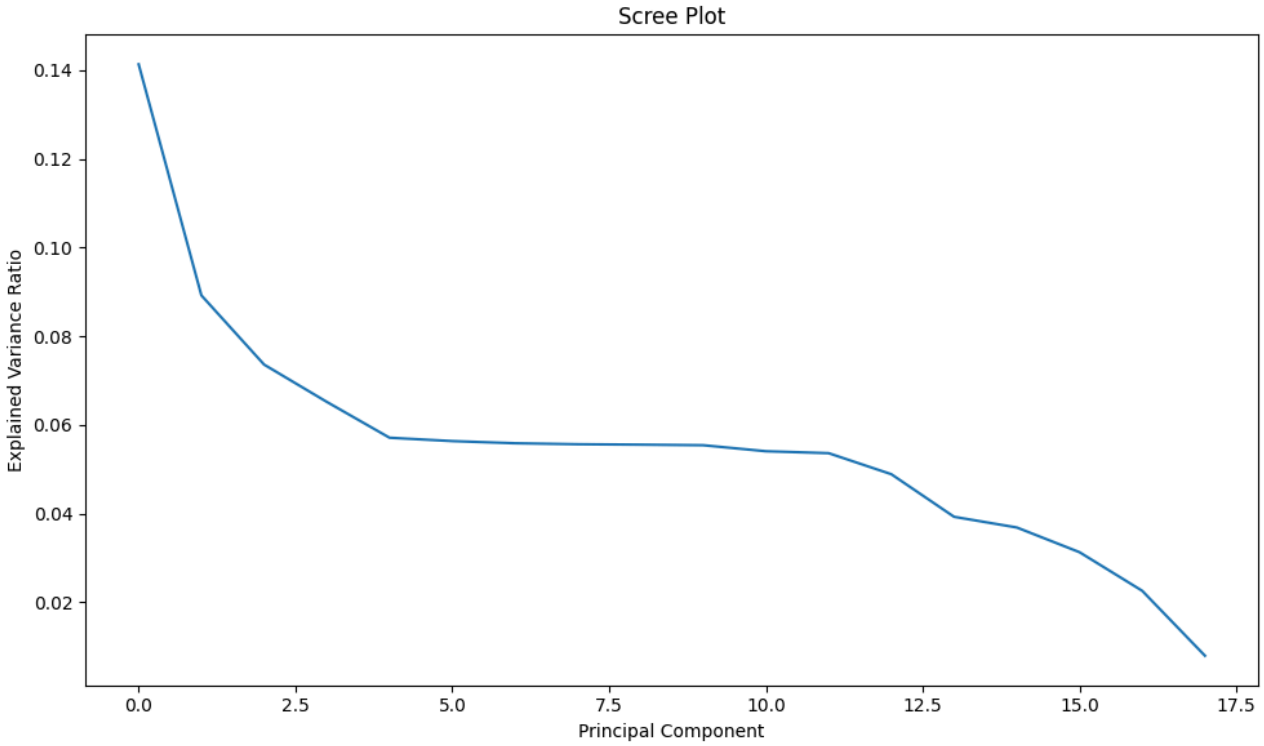
```
Analysis of Df_mice:
Basic Statistics:
         Unnamed: 0  Liquidity_and_Coverage_Ratios_PC1  \
count  20125.000000                       2.012500e+04
mean   10062.000000                      -4.519228e-17
std     5809.731419                       1.787012e+00
min        0.000000                      -7.841332e+00
25%     5031.000000                      -6.208395e-01
50%    10062.000000                      -3.530034e-01
75%    15093.000000                       1.324803e-01
max    20124.000000                       1.193724e+02

       Liquidity_and_Coverage_Ratios_PC2  Leverage_Ratios_PC1  \
count                       2.012500e+04         2.012500e+04
mean                       -1.059194e-17        -1.129807e-17
std                         1.000021e+00         1.140464e+00
min                        -7.799666e+00        -2.633598e+01
25%                        -1.485416e-02        -3.732928e-01
50%                        -8.511489e-03        -1.956420e-01
75%                        -4.440269e-03         1.697937e-01
max                         1.360365e+02         3.844270e+01

       Leverage_Ratios_PC2  Activity_Ratios_PC1  Activity_Ratios_PC2  \
count         2.012500e+04         2.012500e+04         2.012500e+04
mean         -7.061294e-19         2.824518e-18         3.389421e-17
std           1.069642e+00         1.412573e+00         1.023095e+00
min          -4.001687e+01        -6.190249e+00        -1.685172e+00
25%          -1.634621e-02        -5.442500e-02        -5.514452e-01
50%          -6.461196e-03        -3.149699e-02        -7.421455e-02
75%           8.430890e-03        -5.143283e-03         4.685639e-01
max           1.027550e+02         1.583244e+02         7.944953e+01

       Profitability_Ratios_PC1  Profitability_Ratios_PC2  \
count              2.012500e+04              2.012500e+04
mean              -1.129807e-17             -9.885812e-18
std                1.105665e+00              1.013237e+00
min               -5.804650e+01             -8.416007e+00
25%               -2.888133e-02             -1.179799e-01
50%                1.304414e-01             -9.300224e-02
75%                2.208723e-01             -4.626775e-02
max                6.034696e+01              1.059168e+02

       Cost_and_Expense_Ratios_PC1  Cost_and_Expense_Ratios_PC2  \
count                 2.012500e+04                 2.012500e+04
mean                  6.002100e-18                -1.412259e-18
std                   1.417167e+00                 1.065329e+00
min                  -3.318473e-02                -1.817444e+01
25%                  -3.036870e-02                -9.907991e-02
50%                  -2.979116e-02                -9.456594e-02
75%                  -2.736684e-02                -8.030290e-02
max                   1.982758e+02                 5.828151e+01

       Cash_Flow_Ratios_PC1  Cash_Flow_Ratios_PC2  Growth_Ratios_PC1  \
count          20125.000000          2.012500e+04       2.012500e+04
mean               0.000000          2.824518e-18       2.824518e-18
std                1.015163          1.002641e+00       1.329023e+00
min              -16.347361         -1.260259e+01      -6.667743e+00
25%               -0.251294         -3.974870e-02      -5.550133e-02
50%               -0.185254         -2.674195e-02      -3.524864e-02
75%               -0.012509         -8.834717e-03      -2.903311e-03
max               39.479475          8.305408e+01       1.235511e+02

       Growth_Ratios_PC2  Per_Share_Ratios_PC1  Per_Share_Ratios_PC2  \
count       2.012500e+04          2.012500e+04          2.012500e+04
mean        6.355165e-18          1.412259e-18          8.826618e-19
std         1.309731e+00          1.601925e+00          6.291807e-01
min        -2.703377e+01         -3.469094e+00         -3.621296e+01
25%        -1.586664e-02         -4.715991e-02          6.044305e-03
50%        -8.521549e-03         -4.715826e-02          6.046776e-03
75%        -2.524055e-03         -4.714832e-02          6.048398e-03
max         1.737075e+02          9.931935e+01          3.935410e+01

              LABEL
count  20125.000000
mean       0.500025
```

```
std        0.500012
min        0.000000
25%        0.000000
50%        1.000000
75%        1.000000
max        1.000000

Missing Values:
Unnamed: 0                          0
Liquidity_and_Coverage_Ratios_PC1   0
Liquidity_and_Coverage_Ratios_PC2   0
Leverage_Ratios_PC1                 0
Leverage_Ratios_PC2                 0
Activity_Ratios_PC1                 0
Activity_Ratios_PC2                 0
Profitability_Ratios_PC1            0
Profitability_Ratios_PC2            0
Cost_and_Expense_Ratios_PC1         0
Cost_and_Expense_Ratios_PC2         0
Cash_Flow_Ratios_PC1                0
Cash_Flow_Ratios_PC2                0
Growth_Ratios_PC1                   0
Growth_Ratios_PC2                   0
Per_Share_Ratios_PC1                0
Per_Share_Ratios_PC2                0
LABEL                               0
dtype: int64
```

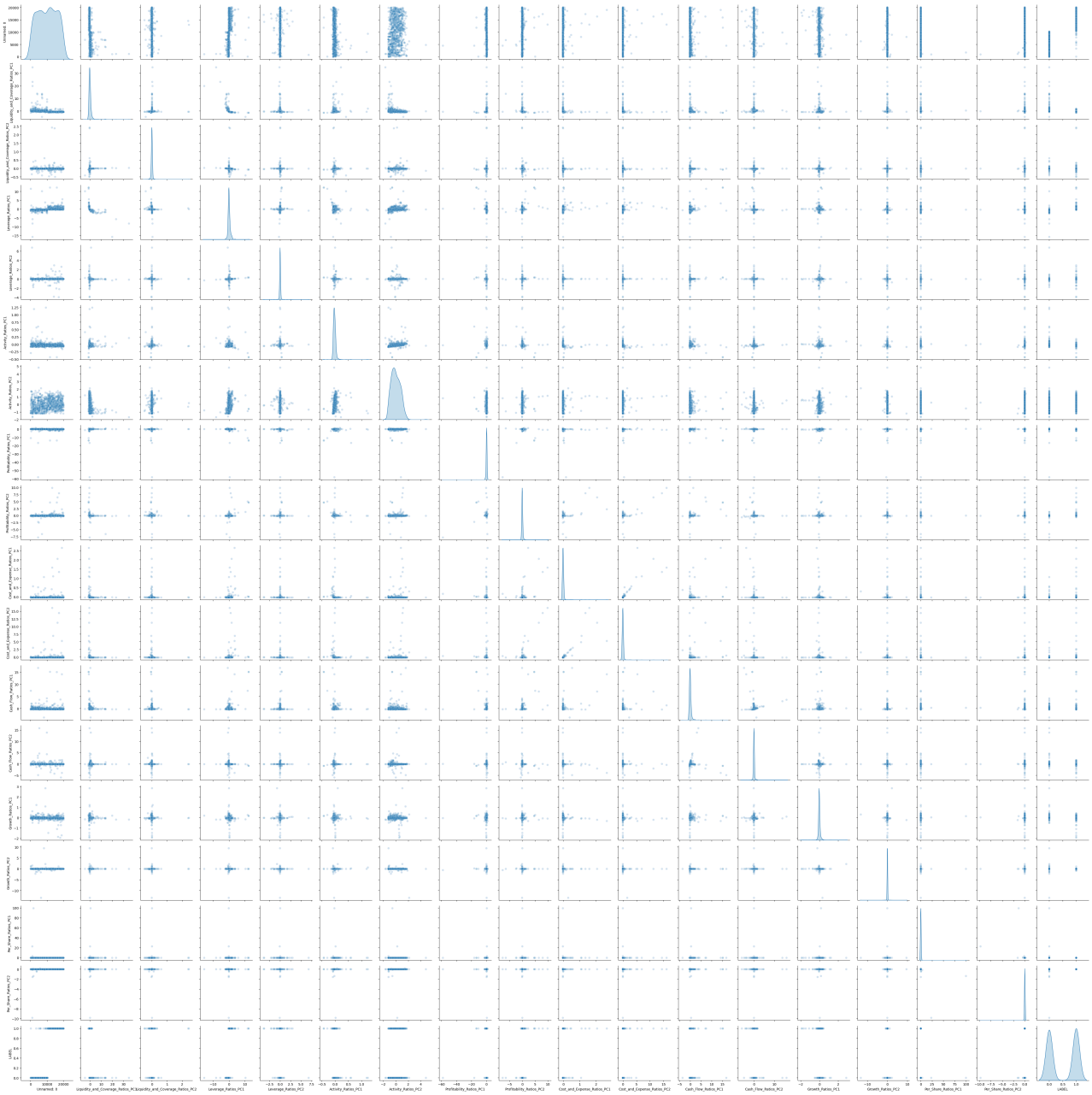


```
Highly correlated feature pairs:
Unnamed: 0 - LABEL: 0.85

Variance Inflation Factors:
                             feature       VIF
0                              const  5.201196
18                             LABEL  3.911864
1                        Unnamed: 0  3.709760
4                Leverage_Ratios_PC1  1.754381
2   Liquidity_and_Coverage_Ratios_PC1  1.327540
12               Cash_Flow_Ratios_PC1  1.280179
9            Profitability_Ratios_PC2  1.246659
8            Profitability_Ratios_PC1  1.235338
11       Cost_and_Expense_Ratios_PC2  1.154909
7                Activity_Ratios_PC2  1.127904
10       Cost_and_Expense_Ratios_PC1  1.087557
6                Activity_Ratios_PC1  1.036868
13               Cash_Flow_Ratios_PC2  1.007364
3   Liquidity_and_Coverage_Ratios_PC2  1.003643
15                 Growth_Ratios_PC2  1.002928
16               Per_Share_Ratios_PC1  1.002398
5                Leverage_Ratios_PC2  1.001355
14                 Growth_Ratios_PC1  1.000891
17               Per_Share_Ratios_PC2  1.000439
```



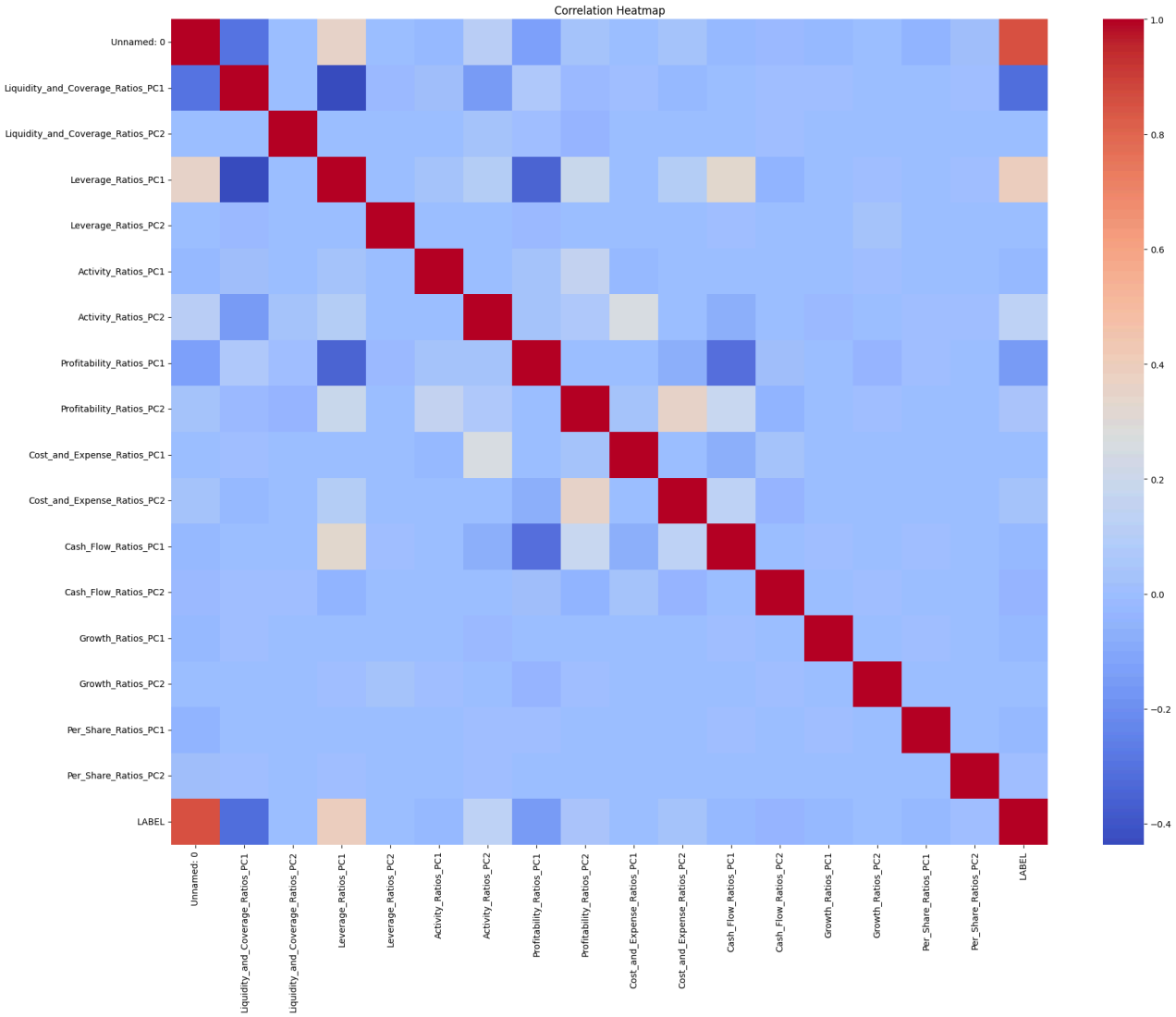



```
Number of potential outliers (Isolation Forest): 2013
Number of potential outliers (Elliptic Envelope): 2013
Number of potential outliers (Z-score > 3): 1461

Skewness of features:
Unnamed: 0                              0.000000
Liquidity_and_Coverage_Ratios_PC1      20.965352
Liquidity_and_Coverage_Ratios_PC2     126.707336
Leverage_Ratios_PC1                     6.765075
Leverage_Ratios_PC2                    45.339471
Activity_Ratios_PC1                    89.479799
Activity_Ratios_PC2                    32.946907
Profitability_Ratios_PC1                0.184323
Profitability_Ratios_PC2               62.990896
Cost_and_Expense_Ratios_PC1           136.238073
Cost_and_Expense_Ratios_PC2            23.286295
Cash_Flow_Ratios_PC1                   13.745306
Cash_Flow_Ratios_PC2                   33.851644
Growth_Ratios_PC1                      73.846641
Growth_Ratios_PC2                     115.774302
Per_Share_Ratios_PC1                   45.792300
Per_Share_Ratios_PC2                   -9.983996
LABEL                                  -0.000099
dtype: float64

Kurtosis of features:
Unnamed: 0                             -1.200000
Liquidity_and_Coverage_Ratios_PC1    1063.301370
Liquidity_and_Coverage_Ratios_PC2   17066.452609
Leverage_Ratios_PC1                   186.124284
Leverage_Ratios_PC2                  4661.370905
Activity_Ratios_PC1                  8996.830039
Activity_Ratios_PC2                  2238.236080
Profitability_Ratios_PC1             1033.375682
Profitability_Ratios_PC2             6073.802352
Cost_and_Expense_Ratios_PC1         19044.190222
Cost_and_Expense_Ratios_PC2           813.913966
Cash_Flow_Ratios_PC1                  318.179706
Cash_Flow_Ratios_PC2                 2480.313546
Growth_Ratios_PC1                    5959.381758
Growth_Ratios_PC2                   15401.110570
Per_Share_Ratios_PC1                 2297.285649
Per_Share_Ratios_PC2                 2447.330325
LABEL                                  -2.000199
dtype: float64
```

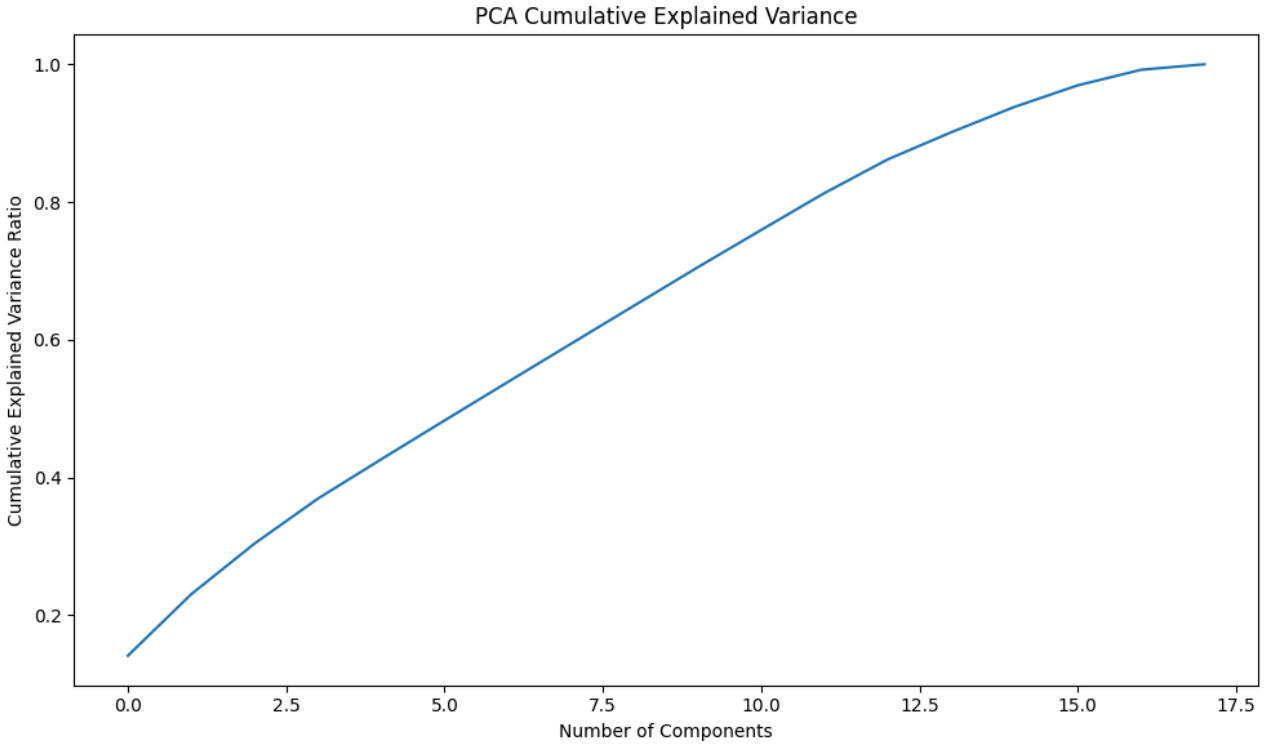


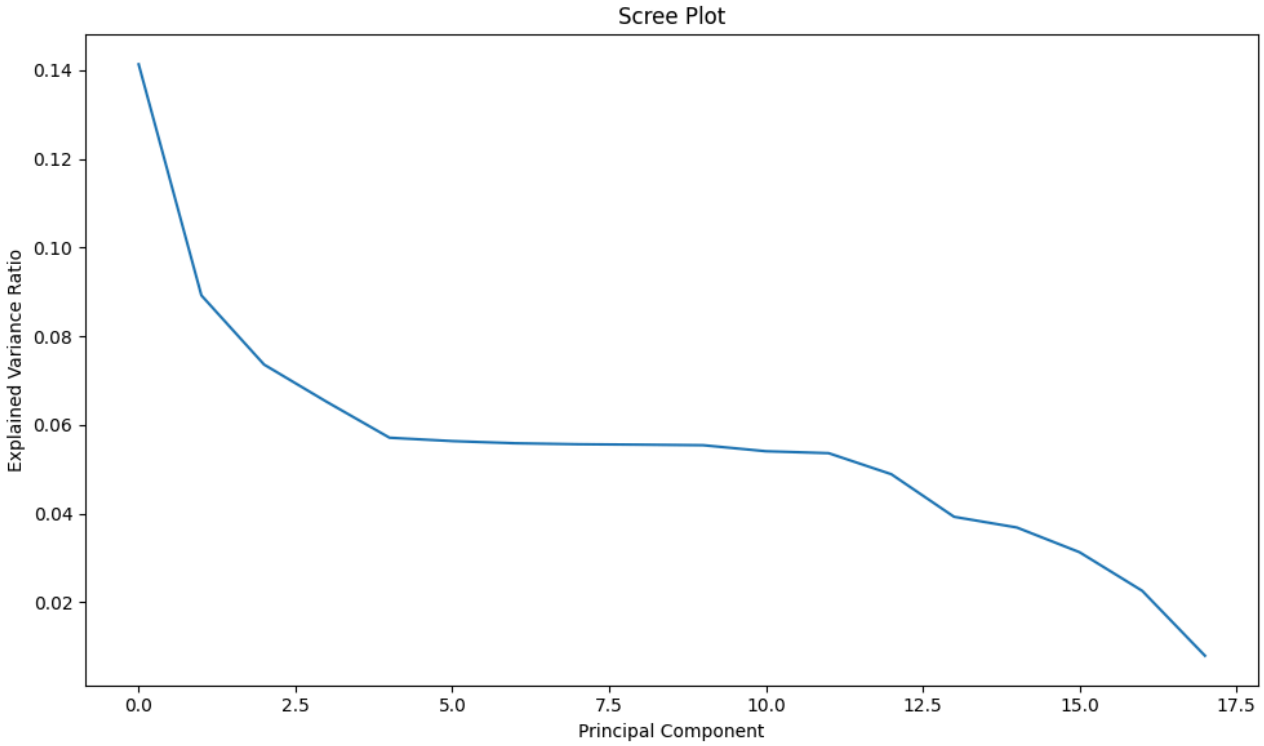
```
Approximate Noise-to-Signal Ratio: 0.0080



Analysis of Df_AE:
Basic Statistics:
         Unnamed: 0  Liquidity_and_Coverage_Ratios_PC1  \
count  20125.000000                       2.012500e+04
mean   10062.000000                      -4.519228e-17
std     5809.731419                       1.787012e+00
min        0.000000                      -7.841332e+00
25%     5031.000000                      -6.208395e-01
50%    10062.000000                      -3.530034e-01
75%    15093.000000                       1.324803e-01
max    20124.000000                       1.193724e+02

       Liquidity_and_Coverage_Ratios_PC2  Leverage_Ratios_PC1  \
count                       2.012500e+04         2.012500e+04
mean                       -1.059194e-17        -1.129807e-17
std                         1.000021e+00         1.140464e+00
min                        -7.799666e+00        -2.633598e+01
25%                        -1.485416e-02        -3.732928e-01
50%                        -8.511489e-03        -1.956420e-01
75%                        -4.440269e-03         1.697937e-01
max                         1.360365e+02         3.844270e+01

       Leverage_Ratios_PC2  Activity_Ratios_PC1  Activity_Ratios_PC2  \
count         2.012500e+04         2.012500e+04         2.012500e+04
mean         -7.061294e-19         2.824518e-18         3.389421e-17
std           1.069642e+00         1.412573e+00         1.023095e+00
min          -4.001687e+01        -6.190249e+00        -1.685172e+00
25%          -1.634621e-02        -5.442500e-02        -5.514452e-01
50%          -6.461196e-03        -3.149699e-02        -7.421455e-02
75%           8.430890e-03        -5.143283e-03         4.685639e-01
max           1.027550e+02         1.583244e+02         7.944953e+01

       Profitability_Ratios_PC1  Profitability_Ratios_PC2  \
count              2.012500e+04              2.012500e+04
mean              -1.129807e-17             -9.885812e-18
std                1.105665e+00              1.013237e+00
min               -5.804650e+01             -8.416007e+00
25%               -2.888133e-02             -1.179799e-01
50%                1.304414e-01             -9.300224e-02
75%                2.208723e-01             -4.626775e-02
max                6.034696e+01              1.059168e+02

       Cost_and_Expense_Ratios_PC1  Cost_and_Expense_Ratios_PC2  \
count                 2.012500e+04                 2.012500e+04
mean                  6.002100e-18                -1.412259e-18
std                   1.417167e+00                 1.065329e+00
min                  -3.318473e-02                -1.817444e+01
25%                  -3.036870e-02                -9.907991e-02
50%                  -2.979116e-02                -9.456594e-02
75%                  -2.736684e-02                -8.030290e-02
max                   1.982758e+02                 5.828151e+01

       Cash_Flow_Ratios_PC1  Cash_Flow_Ratios_PC2  Growth_Ratios_PC1  \
count          20125.000000          2.012500e+04       2.012500e+04
mean               0.000000          2.824518e-18       2.824518e-18
std                1.015163          1.002641e+00       1.329023e+00
min              -16.347361         -1.260259e+01      -6.667743e+00
25%               -0.251294         -3.974870e-02      -5.550133e-02
50%               -0.185254         -2.674195e-02      -3.524864e-02
75%               -0.012509         -8.834717e-03      -2.903311e-03
max               39.479475          8.305408e+01       1.235511e+02

       Growth_Ratios_PC2  Per_Share_Ratios_PC1  Per_Share_Ratios_PC2  \
count       2.012500e+04          2.012500e+04          2.012500e+04
mean        6.355165e-18          1.412259e-18          8.826618e-19
std         1.309731e+00          1.601925e+00          6.291807e-01
min        -2.703377e+01         -3.469094e+00         -3.621296e+01
25%        -1.586664e-02         -4.715991e-02          6.044305e-03
50%        -8.521549e-03         -4.715826e-02          6.046776e-03
75%        -2.524055e-03         -4.714832e-02          6.048398e-03
max         1.737075e+02          9.931935e+01          3.935410e+01
```

```
              LABEL
count  20125.000000
mean       0.500025
std        0.500012
min        0.000000
25%        0.000000
50%        1.000000
75%        1.000000
max        1.000000

Missing Values:
Unnamed: 0                          0
Liquidity_and_Coverage_Ratios_PC1   0
Liquidity_and_Coverage_Ratios_PC2   0
Leverage_Ratios_PC1                 0
Leverage_Ratios_PC2                 0
Activity_Ratios_PC1                 0
Activity_Ratios_PC2                 0
Profitability_Ratios_PC1            0
Profitability_Ratios_PC2            0
Cost_and_Expense_Ratios_PC1         0
Cost_and_Expense_Ratios_PC2         0
Cash_Flow_Ratios_PC1                0
Cash_Flow_Ratios_PC2                0
Growth_Ratios_PC1                   0
Growth_Ratios_PC2                   0
Per_Share_Ratios_PC1                0
Per_Share_Ratios_PC2                0
LABEL                               0
dtype: int64
```

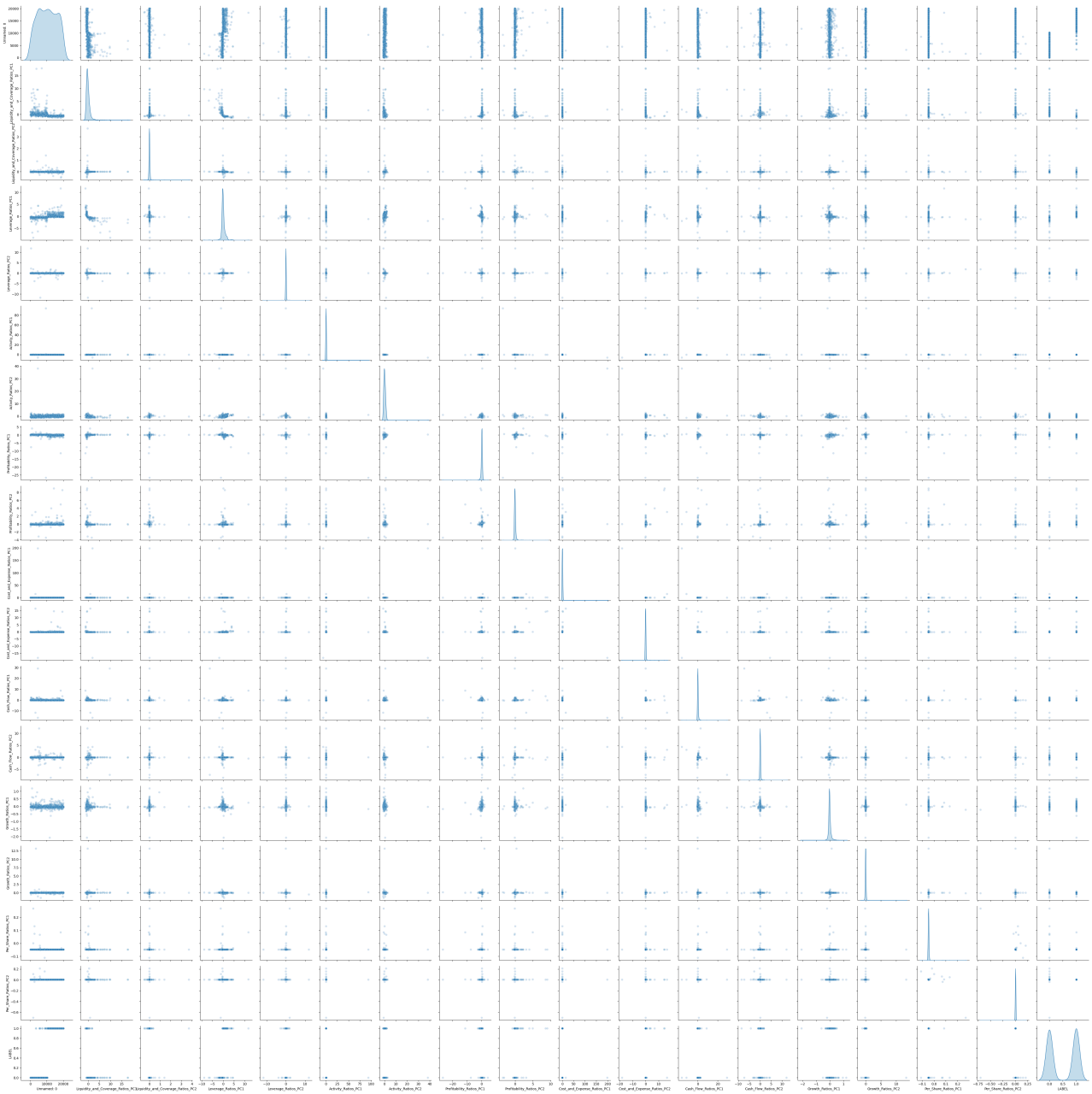


```
Highly correlated feature pairs:
Unnamed: 0 - LABEL: 0.85

Variance Inflation Factors:
                             feature       VIF
0                              const  5.201196
18                             LABEL  3.911864
1                        Unnamed: 0  3.709760
4                Leverage_Ratios_PC1  1.754381
2   Liquidity_and_Coverage_Ratios_PC1  1.327540
12               Cash_Flow_Ratios_PC1  1.280179
9            Profitability_Ratios_PC2  1.246659
8            Profitability_Ratios_PC1  1.235338
11       Cost_and_Expense_Ratios_PC2  1.154909
7                Activity_Ratios_PC2  1.127904
10       Cost_and_Expense_Ratios_PC1  1.087557
6                Activity_Ratios_PC1  1.036868
13               Cash_Flow_Ratios_PC2  1.007364
3   Liquidity_and_Coverage_Ratios_PC2  1.003643
15                 Growth_Ratios_PC2  1.002928
16               Per_Share_Ratios_PC1  1.002398
5                Leverage_Ratios_PC2  1.001355
14                 Growth_Ratios_PC1  1.000891
17               Per_Share_Ratios_PC2  1.000439
```






```
Number of potential outliers (Isolation Forest): 2013
Number of potential outliers (Elliptic Envelope): 2013
Number of potential outliers (Z-score > 3): 1461

Skewness of features:
Unnamed: 0                              0.000000
Liquidity_and_Coverage_Ratios_PC1      20.965352
Liquidity_and_Coverage_Ratios_PC2     126.707336
Leverage_Ratios_PC1                     6.765075
Leverage_Ratios_PC2                    45.339471
Activity_Ratios_PC1                    89.479799
Activity_Ratios_PC2                    32.946907
Profitability_Ratios_PC1                0.184323
Profitability_Ratios_PC2               62.990896
Cost_and_Expense_Ratios_PC1           136.238073
Cost_and_Expense_Ratios_PC2            23.286295
Cash_Flow_Ratios_PC1                   13.745306
Cash_Flow_Ratios_PC2                   33.851644
Growth_Ratios_PC1                      73.846641
Growth_Ratios_PC2                     115.774302
Per_Share_Ratios_PC1                   45.792300
Per_Share_Ratios_PC2                   -9.983996
LABEL                                  -0.000099
dtype: float64

Kurtosis of features:
Unnamed: 0                             -1.200000
Liquidity_and_Coverage_Ratios_PC1    1063.301370
Liquidity_and_Coverage_Ratios_PC2   17066.452609
Leverage_Ratios_PC1                   186.124284
Leverage_Ratios_PC2                  4661.370905
Activity_Ratios_PC1                  8996.830039
Activity_Ratios_PC2                  2238.236080
Profitability_Ratios_PC1             1033.375682
Profitability_Ratios_PC2             6073.802352
Cost_and_Expense_Ratios_PC1         19044.190222
Cost_and_Expense_Ratios_PC2           813.913966
Cash_Flow_Ratios_PC1                  318.179706
Cash_Flow_Ratios_PC2                 2480.313546
Growth_Ratios_PC1                    5959.381758
Growth_Ratios_PC2                   15401.110570
Per_Share_Ratios_PC1                 2297.285649
Per_Share_Ratios_PC2                 2447.330325
LABEL                                  -2.000199
dtype: float64
```



Approximate Noise-to-Signal Ratio: 0.0080