資料分析方法-HW5

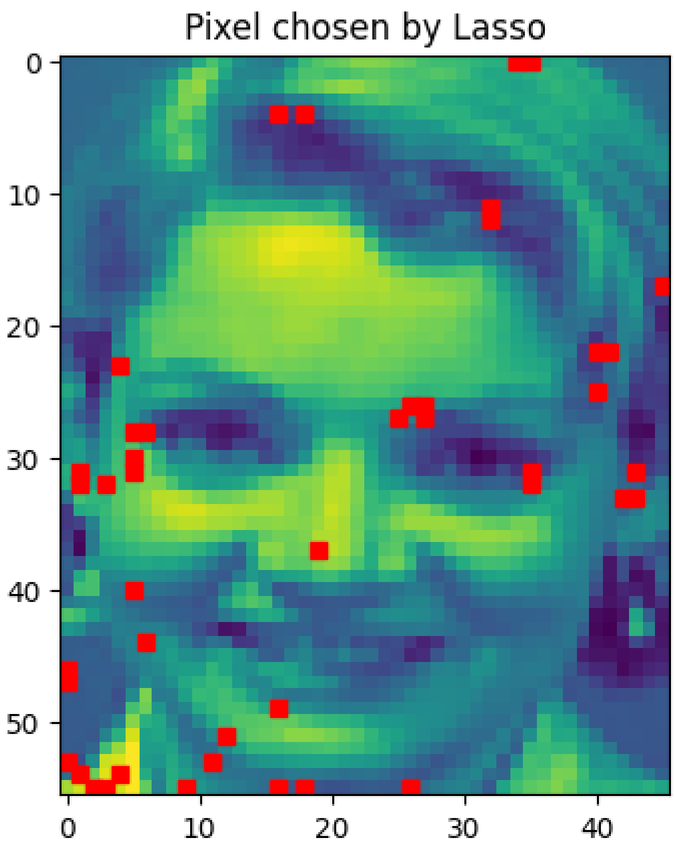
經研一 江彥亨 R11323040

1.a

```
MSE of stepwise regression is 0.012474149409729338

MSE of Ridge regression is 3.0289602143414327e-23

MSE of LASSO regression is 0.03982085228233607
```

Based on results, Ridge regression has the best prediction.

1.b



2.a

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Value Added   R-squared:                       0.817
Model:                            OLS   Adj. R-squared:                  0.786
Method:                 Least Squares   F-statistic:                     26.78
Date:                Sun, 02 Apr 2023   Prob (F-statistic):           3.76e-05
Time:                        17:45:02   Log-Likelihood:                 19.196
No. Observations:                  15   AIC:                            -32.39
Df Residuals:                      12   BIC:                            -30.27
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -9.6745      2.907     -3.328      0.006     -16.009      -3.341
Capital        0.5201      0.506      1.028      0.324      -0.583       1.623
Labor          0.8345      0.422      1.975      0.072      -0.086       1.755
==============================================================================
Omnibus:                        6.533   Durbin-Watson:                   1.924
Prob(Omnibus):                  0.038   Jarque-Bera (JB):                3.365
Skew:                           1.010   Prob(JB):                        0.186
Kurtosis:                       4.142   Cond. No.                     2.98e+03
==============================================================================
```

Based on results with a significance level of 0.05, both two coefficients are insignificant.

2.b

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            Value Added   No. Observations:                   15
Model:                            GLM   Df Residuals:                       13
Model Family:                Gaussian   Df Model:                            1
Link Function:               identity   Scale:                       0.0064940
Method:                          IRLS   Log-Likelihood:                 17.566
Date:                Sun, 02 Apr 2023   Deviance:                     0.084422
Time:                        17:45:09   Pearson chi2:                   0.0844
No. Iterations:                     1   Pseudo R-squ. (CS):             0.9478
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.7126      0.021   -225.980      0.000      -4.753      -4.672
Capital        0.0235      0.444      0.053      0.958      -0.846       0.893
Labor          0.9765      0.444      2.201      0.028       0.107       1.846
==============================================================================

Model has been estimated subject to linear equality constraints.
```

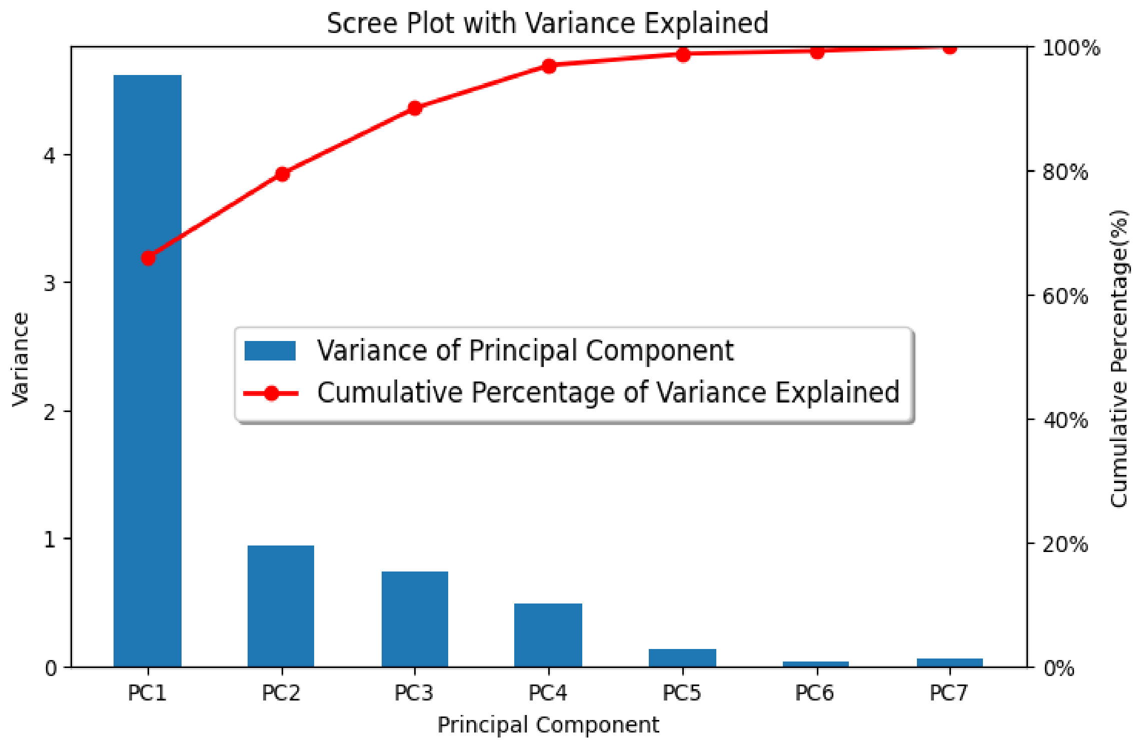
Based on the results with a significance level of 0.05 and the CRTS constraint, it can be concluded that the coefficient of capital share is insignificant and small, whereas the coefficient of labor share is significant and large. These findings suggest that the value added of the U.S. between the years 72 to 86 was largely driven by labor share.

3.a

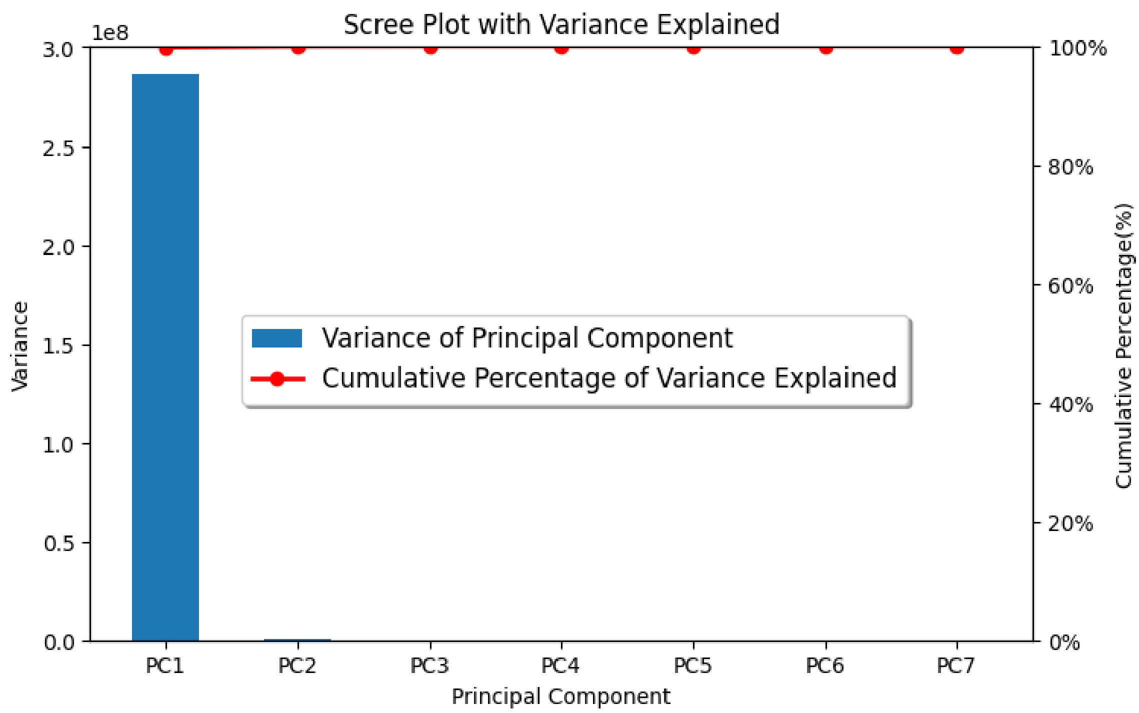
```python
def PCA_func(X=None, isCorrMX=False):
    n = len(X) # sample size
    p = len(X.columns) # number of variables
    Z = np.dot((np.eye(n)-np.ones((n,n))/int(n)), X) # center X
    S = np.dot(Z.T, Z)
    if isCorrMX == True:
        R = np.corrcoef(Z,rowvar=False)
        eig_vals, eig_vecs = np.linalg.eig(R)
    else:
        eig_vals, eig_vecs = np.linalg.eig(S)

    P = eig_vecs # loading matrix
    eigenvalues = list(np.real(eig_vals))
    eigenvector = eig_vecs
    T = np.dot(X,P) # score matrix

    #Scree plot
    fig, ax1 = plt.subplots(figsize=(12,8))

    PC = pd.Series({f"PC{i+1}":eigenvalues[i] for i in range(p)}) # Variance of each pc
    ax1.bar(PC.keys(), PC.values, width=0.5, align='center',
            label='Variance of Principal Component')
    ax1.set_title('Scree Plot with Variance Explained')
    ax1.set_xlabel('Principal Component')
    ax1.set_ylabel('Variance')
    if len(PC.keys()) > 10:
       ax1.xaxis.set_major_locator(ticker.MultipleLocator(500))

    ax2 = ax1.twinx()
    variance_ratio = np.cumsum(PC) / np.sum(PC)
    ax2.plot(PC.keys(), variance_ratio, 'o-', linewidth=2, c='r',
             label='Cumulative Percentage of Variance Explained')
    ax2.set_ylabel('Cumulative Percentage(%)')
    ax2.set_ylim(0, 1)
    ax2.yaxis.set_major_formatter(ticker.PercentFormatter(1.0))

    fig.legend(loc='center', fontsize=12, shadow=True)

    return P, eigenvalues, eigenvector, T
```

3.b

Correlation matrix:



Covariance matrix:



As shown in the two images above, using the covariance matrix produces different results from using the correlation matrix. Scale transformation can have an impact on the outcome of spectral decomposition, leading to differences in eigenvalues and eigenvectors. This means that the results of PCA can vary depending on the scaling method used, and therefore, PCA is scale-variant.

4.a

```
2 principal components are needed to explain 50% of total variance

4 principal components are needed to explain 60% of total variance

7 principal components are needed to explain 70% of total variance

17 principal components are needed to explain 80% of total variance

50 principal components are needed to explain 90% of total variance
```

4.b