2.4) Perform Principal Component Analysis (on the predictor variables) and extract the Principal Components. Comment on the reason behind choosing the number of Principal Components.

In the given data set, we see that there are lot of variables, i.e., the no. of features are more than 10, hence we are doing PCA to reduce the problem of multicollinearity as seen in question 2.2.

PCA is a statistical technique which reduces the dimensions of the data and help us understand, plot the data with lesser dimension compared to original data. As the name says PCA helps us compute the Principal components in data. Principal components are basically vectors that are linearly uncorrelated and have a variance with in data. From the principal components top n components are picked which have the most variance.

We carry out PCA in a sequential manner as explained in following steps:

**STEP 1 : BEFORE PCA, BUILD CO-VARIANCE MATRIX**

Covariance Matrix

%s [[ 1.00132979e+00 -3.31582115e-02 1.06101321e-01 4.23506899e-01

6.07723445e-01 -5.64225183e-02 -3.11300884e-02 -9.66326289e-03

-9.87300540e-02 5.79406915e-02 1.36890778e-02 -6.03700052e-02

4.05497226e-01 1.46500077e-01]

[-3.31582115e-02 1.00132979e+00 -1.20382861e-01 -3.46051004e-02

-5.83924775e-02 -8.44837698e-02 8.89319009e-01 -1.63266264e-01

2.70507312e-02 -2.34953587e-01 -1.60804404e-01 7.71793469e-02

3.34460049e-01 5.25097894e-02]

[ 1.06101321e-01 -1.20382861e-01 1.00132979e+00 3.18801449e-01

2.67930359e-01 7.90208661e-02 -1.33699059e-01 6.12767546e-01

2.85315019e-01 4.35915402e-01 4.43046609e-01 7.22359587e-02

6.63436673e-02 3.61755369e-01]

[ 4.23506899e-01 -3.46051004e-02 3.18801449e-01 1.00132979e+00

6.52507648e-01 -5.99985405e-02 -3.18782667e-02 1.26391585e-01

6.13711932e-02 9.04253051e-02 9.86077958e-02 -1.27896664e-04

2.50913296e-01 2.31482558e-01]

[ 6.07723445e-01 -5.83924775e-02 2.67930359e-01 6.52507648e-01

1.00132979e+00 -7.08913431e-02 -5.54725311e-02 1.07108799e-01

1.93018022e-02 8.57115891e-02 1.02909091e-01 9.13632427e-03

3.42011117e-01 2.08989211e-01]

[-5.64225183e-02 -8.44837698e-02 7.90208661e-02 -5.99985405e-02

-7.08913431e-02 1.00132979e+00 -9.55138763e-02 1.07988079e-01

-2.36334670e-01 5.34247037e-02 5.04123675e-02 -1.55426002e-01

-9.94983848e-02 1.28311911e-01]

[-3.11300884e-02 8.89319009e-01 -1.33699059e-01 -3.18782667e-02

-5.54725311e-02 -9.55138763e-02 1.00132979e+00 -1.95582291e-01

1.97070726e-02 -2.27759226e-01 -1.35179749e-01 5.31644460e-02

2.72272038e-01 4.05564936e-02]

[-9.66326289e-03 -1.63266264e-01 6.12767546e-01 1.26391585e-01

1.07108799e-01 1.07988079e-01 -1.95582291e-01 1.00132979e+00

3.95189608e-01 3.24906192e-01 3.67187193e-01 5.50903396e-02

-3.63490142e-02 3.77373004e-01]

[-9.87300540e-02 2.70507312e-02 2.85315019e-01 6.13711932e-02

1.93018022e-02 -2.36334670e-01 1.97070726e-02 3.95189608e-01

1.00132979e+00 1.26904382e-01 1.93485974e-01 1.58130174e-01

-1.03443569e-01 7.25986873e-01]

[ 5.79406915e-02 -2.34953587e-01 4.35915402e-01 9.04253051e-02

8.57115891e-02 5.34247037e-02 -2.27759226e-01 3.24906192e-01

1.26904382e-01 1.00132979e+00 5.73833806e-01 1.84266371e-02

-8.22881828e-02 1.60106542e-01]

[ 1.36890778e-02 -1.60804404e-01 4.43046609e-01 9.86077958e-02

1.02909091e-01 5.04123675e-02 -1.35179749e-01 3.67187193e-01

1.93485974e-01 5.73833806e-01 1.00132979e+00 5.86096606e-02

-7.89069628e-02 2.11349596e-01]

[-6.03700052e-02 7.71793469e-02 7.22359587e-02 -1.27896664e-04

9.13632427e-03 -1.55426002e-01 5.31644460e-02 5.50903396e-02

1.58130174e-01 1.84266371e-02 5.86096606e-02 1.00132979e+00

4.44127241e-03 6.16032571e-02]

[ 4.05497226e-01 3.34460049e-01 6.63436673e-02 2.50913296e-01

3.42011117e-01 -9.94983848e-02 2.72272038e-01 -3.63490142e-02

-1.03443569e-01 -8.22881828e-02 -7.89069628e-02 4.44127241e-03

1.00132979e+00 -2.77395342e-02]

[ 1.46500077e-01 5.25097894e-02 3.61755369e-01 2.31482558e-01

2.08989211e-01 1.28311911e-01 4.05564936e-02 3.77373004e-01

7.25986873e-01 1.60106542e-01 2.11349596e-01 6.16032571e-02

-2.77395342e-02 1.00132979e+00]]

**STEP 2 : GET THE EIGEN VALUES AND EIGEN VECTOR**

**EIGEN VECTORS**

[[-0.1686472 0.39843276 -0.30792729 0.05077413 0.04098592 -0.03835352

-0.18317686 0.05250623 -0.16142759 -0.42665959 0.36238807 -0.13407126

0.56424597 -0.01340566]

[ 0.18517543 0.36470814 0.44768491 -0.28649017 -0.02001428 -0.71574255

0.04191019 0.01632308 -0.04891906 0.10816309 -0.00650375 0.06792075

0.1197444 -0.04901206]

[-0.43018553 -0.00260745 0.0785571 -0.20841853 -0.0560183 -0.01224648

-0.05781483 -0.11223917 0.35179612 0.22744303 -0.09946963 -0.71608426

0.09143971 -0.19892952]

[-0.26996195 0.35153623 -0.20018403 0.10455812 0.05315193 -0.0082593

-0.05779293 -0.06109237 -0.10977246 0.58431703 0.49427758 0.16073952

-0.33577236 -0.07876644]

[-0.26740601 0.40103509 -0.26978835 0.11051542 0.01955778 -0.01616753

-0.00348958 -0.05310742 -0.13127128 0.16098962 -0.76064171 0.1719388

0.09777691 0.13542567]

[-0.03056819 -0.13926235 -0.10308487 -0.48352238 0.5607758 -0.01318974

-0.30250752 -0.50125187 -0.21379503 -0.10987928 -0.03385916 0.05313403

-0.1263492 0.01131088]

[ 0.18957364 0.35344005 0.43952227 -0.28601445 -0.02527588 0.68414746

-0.10060173 0.06286632 -0.12694599 0.17009823 -0.01141284 0.02597377

0.18798668 -0.01399267]

[-0.38713204 -0.12120862 0.15726244 -0.09779777 0.08219371 0.0422353

0.09618915 -0.12337821 0.55620203 0.02558035 0.11444548 0.55590766

0.344305 0.13204214]

[-0.27870268 -0.03625154 0.44654483 0.40897585 0.14188969 -0.05691265

-0.66246931 0.19498892 -0.03032182 -0.14284184 -0.06296607 0.03607553

-0.16573607 0.00559485]

[-0.33145193 -0.16721759 -0.03613405 -0.32723755 -0.32926937 0.0083707

-0.00859336 0.21489553 -0.28203532 -0.14676469 -0.0718541 0.27838393

-0.06416815 -0.64123593]

[-0.34043869 -0.14175019 0.05789239 -0.332301 -0.31901438 -0.03558014

0.00903998 0.17083562 -0.30256474 -0.02968831 0.09789142 -0.07523384

-0.10534394 0.70651484]

[-0.04753127 0.0214802 0.20768638 0.26220642 -0.49706982 0.02201835

0.00918648 -0.77052193 -0.16981768 -0.10966079 0.04101181 0.0195832

0.01881239 -0.02754963]

[-0.01484011 0.45848518 -0.04639254 -0.17806551 -0.11438977 0.07937349

0.03463747 -0.01715591 0.41124588 -0.49839555 -0.00298425 0.0240762

-0.56167016 0.03534167]

[-0.33936093 0.069146 0.32839795 0.19588086 0.42334989 0.06623039

0.63781334 0.00382159 -0.27812073 -0.2056163 0.03280928 -0.1062135

-0.09090365 -0.05617807]]

**EIGEN VALUES :**

[3.2443069 2.4192616 1.95369569 1.25214228 1.13938768 0.10637183

0.14362232 0.85996708 0.74053912 0.65891534 0.29050541 0.33216352

0.45173774 0.42600053]

**LET US NOW PRINT THE FIRST EIGEN VECTORS:**

The first eigen vector is:

-0.1686472040447155 \* WorkingHoursWife +

0.39843276446274417 \* WifeAge +

-0.307927291637471 \* EducationWife +

0.050774126356934 \* WifeHourEarnings +

0.04098592317226116 \* WifeWage +

-0.038353521319505125 \* WorkingHoursHusband +

-0.1831768587018411 \* HusbandAge +

0.05250623079205406 \* EducationHusband +

-0.1614275921502438 \* HusbandWage +

-0.4266595911627534 \* EducationWifeMother +

0.3623880688442056 \* EducationWifeFather +

-0.13407125819837415 \* UnemploymentRate +

0.564245973851108 \* WifeExperience +

-0.013405656825398956 \* FamilyIncome

**STEP 3: CALCULATE THE VARIANCE EXPLAINED BY EIGEN VALUES AND THE CUMULATIVE VARIANCE BY THE EIGEN VALUES**

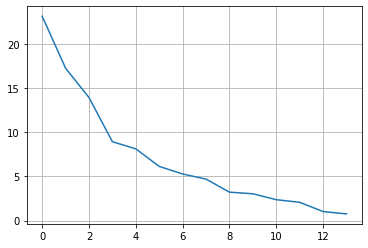
The variance explained by each of eigen values in order is [23.14284563165372, 17.25749118218782, 13.936436695266513, 8.931995747063999, 8.127675346874096, 6.134464460329296, 5.282540513667376, 4.700287742344607, 3.2224130129147035, 3.0388199396878273, 2.369445688748432, 2.072282980718254, 1.024511325400953, 0.7587897331424078]

Cumulative Variance Explained [ 23.14284563 40.40033681 54.33677351 63.26876926 71.3964446

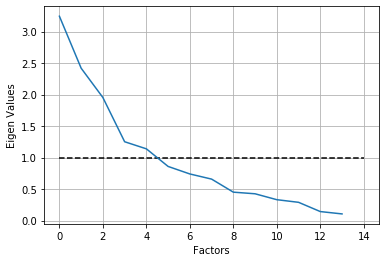
77.53090906 82.81344958 87.51373732 90.73615033 93.77497027

96.14441596 98.21669894 99.24121027 100.

**STEP 4 : NOW PLOT THE VARIANCE EXPLAINED BY EACH EIGEN VALUE WITH THE EIGEN VALUE**

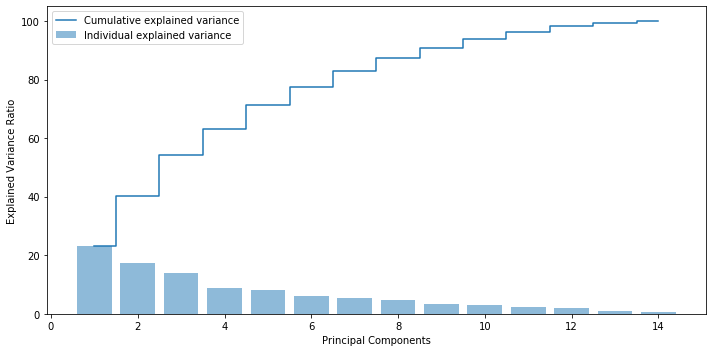


PLOT OF EIGEN VALUES WITH THE NUMBER OF FACTORS OR PRINCIPAL COMPONENTS:

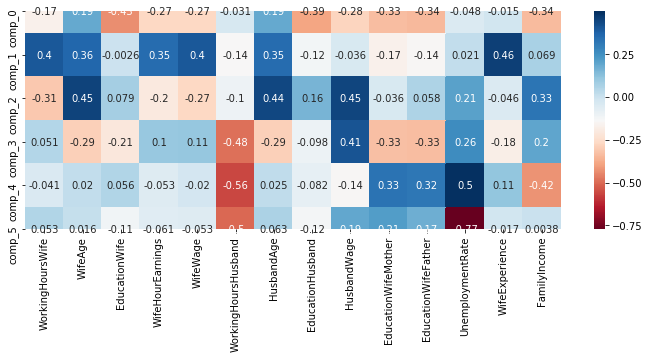


From the above plot, we can see that the number of components that we can probably take is 6.

We also see that if we take 6 components the total amount of variance explained is 77.53090906337545 %



**STEP 5 : PERFOMING PCA WITH SELECTED COMPONENTS AND CHECKING CORRELATION WITH NEW COMPONENTS:**

****

**We see that the correlation between the components has decreased considerably as compared to the correlation observed in question 2.2. We see that the values are falling in a moderate range. Thus, we have successfully solved the problem of multicollinearity.**