**Student name: GOUTHAM SELVAKUMAR**

**DSC 324/424 Assignment 3**  
**Due**: Tuesday, April 26th at 11:59pm CST

**1)** **Principal Component Analysis**

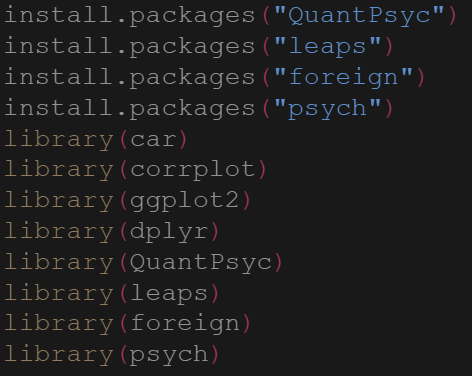
**The data given in the file ‘Employment.txt’ is the percentage employed in different industries in Europe countries during 1979. Techniques such as Principal Component Analysis (PCA) can be used to examine which countries have similar employment patterns. There are 26 countries in the file and 10 variables as follows:**

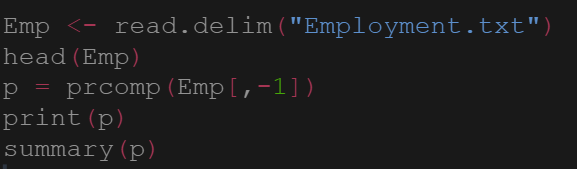
**Variable Names:**

**1. Country: Name of country  
2. Agr: Percentage employed in agriculture  
3. Min: Percentage employed in mining  
4. Man: Percentage employed in manufacturing  
5. PS: Percentage employed in power supply industries  
6. Con: Percentage employed in construction  
7. SI: Percentage employed in service industries  
8. Fin: Percentage employed in finance  
9. SPS: Percentage employed in social and personal services  
10. TC: Percentage employed in transport and communications.**

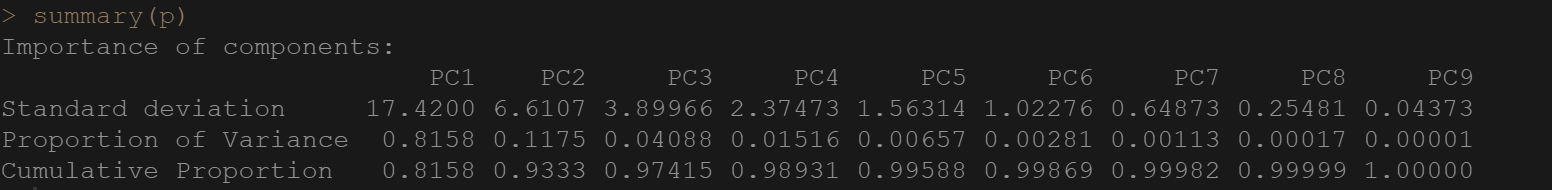
**Perform a principal component analysis using the covariance matrix:**

1. **How many principal components are required to explain 90% of the total variation for this data?**



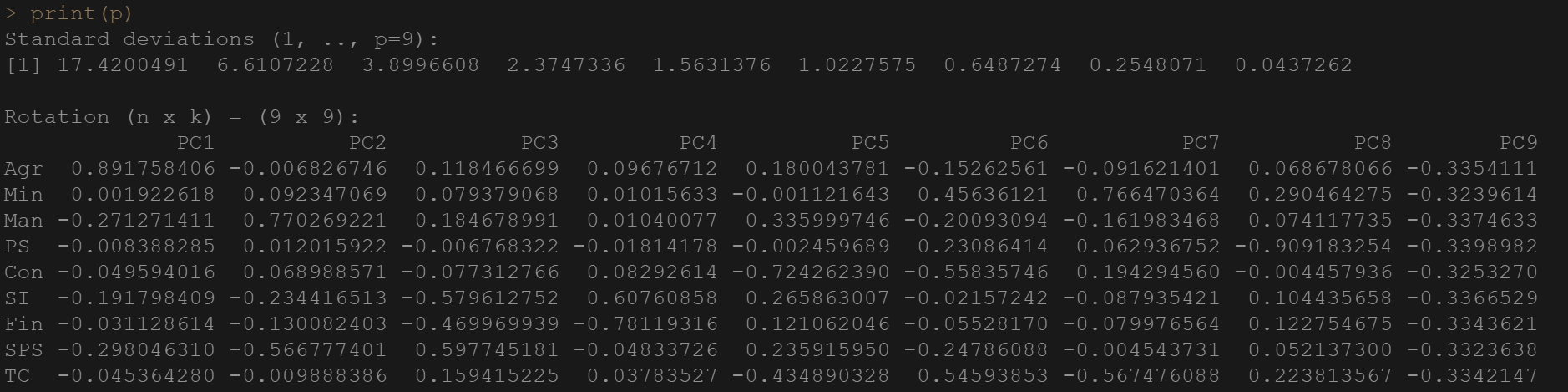


The summary of PCA shows that two principal components are required to explain about 93% of the variance for the data.



1. **For each of component from part a, give the formula for each component and a brief interpretation. Think about what kind of label you might give to each component.**

Two components are required to explain 90% of variance.



PC1 = (0.89\*Agr) + (0.0019\*Min) – (0.271\*Man) – (0.0083\*PS) – (0.049\*Con) – (0.191\*SI) – (0.031\*Fin) – (0.298\*SPS) – (0.045\*TC)

In PC1 agriculture has the highest value and the mining has almost zero value. Except for agriculture and mining all other industries have negative values. So, PC1 may indicate the people employed in agricultural industry.

PC2 = - (0.006\*Agr) + (0.092\*Min) + (0.770\*Man) + (0.012\*PS) + (0.068\*Con) – (0.234\*SI) – (0.130\*Fin) – (0.566\*SPS) – (-0.0453\*TC)

Manufacturing has highest value in PC2 and Mining, Manufacturing, Power Supply and construction has positive values. This indicates that PC2 is related to Hardware industries. Without rotation it is difficult to interpret the components.

PC3 = (0.118\*Agr) + (0.079\*Min) + (0.184\*Man) – (0.006\*PS) – (0.077\*Con) – (0.579\*SI) – (0.469\*Fin) + (0.597\*SPS) + (0.159\*TC)

PC4 = (0.096\*Agr) + (0.010\*Min) + (0.010\*Man) – (0.018\*PS) + (0.082\*Con) + (0.607\*SI) – (0.781\*Fin) – (0.048\*SPS) + (0.037\*TC)

PC5 = (0.180\*Agr) – (0.001\*Min) + (0.335\*Man) – (0.002\*PS) – (0.724\*Con) + (0.265\*SI) + (0.121\*Fin) + (0.235\*SPS) – (0.434\*TC)

PC6 = - (0.152\*Agr) + (0.456\*Min) – (0.200\*Man) + (0.230\*PS) – (0.558\*Con) – (0.021\*SI) – (0.55\*Fin) – (0.247\*SPS) + (0.545\*TC)

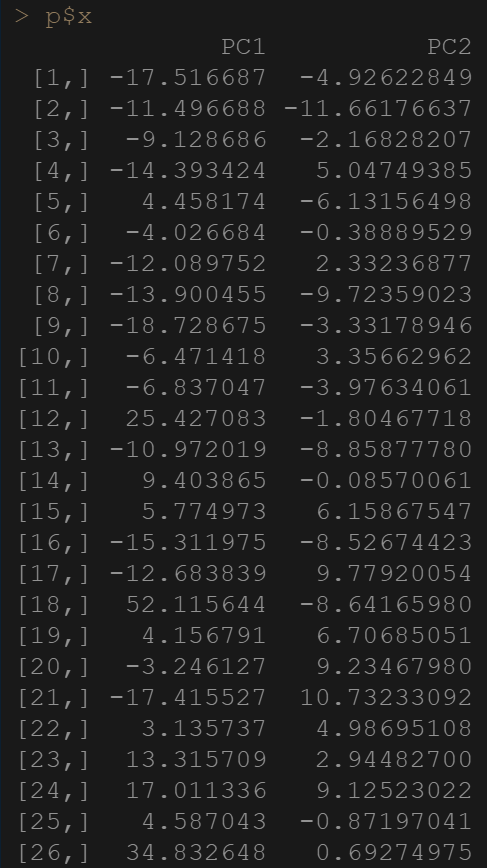
PC7 = - (0.091\*Agr) + (0.766\*Min) – (0.161\*Man) + (0.062\*PS) + (0.194\*Con) – (0.087\*SI) – (0.079\*Fin) – (0.004\*SPS) – (0.567\*TC)

PC8 = (0.068\*Agr) + (0.290\*Min) + (0.074\*Man) – (0.909\*PS) – (0.004\*Con) + (0.104\*SI) + (0.122\*Fin) + (0.052\*SPS) + (0.22\*TC)

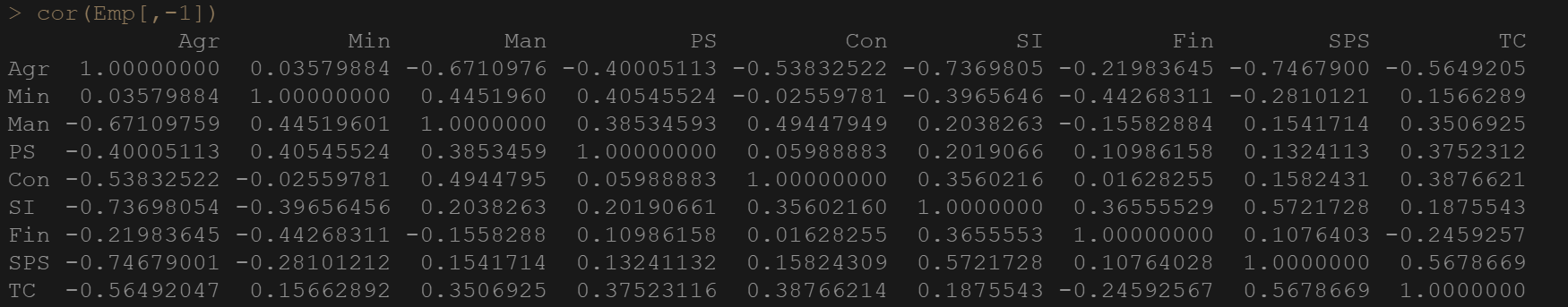
PC9 = - (0.335\*Agr) – (0.323\*Min) – (-0.337\*Man) – (0.339\*PS) – (0.325\*Con) – (0.336\*SI) – (0.334\*Fin) – (0.332\*SPS) – (0.334\*TC)

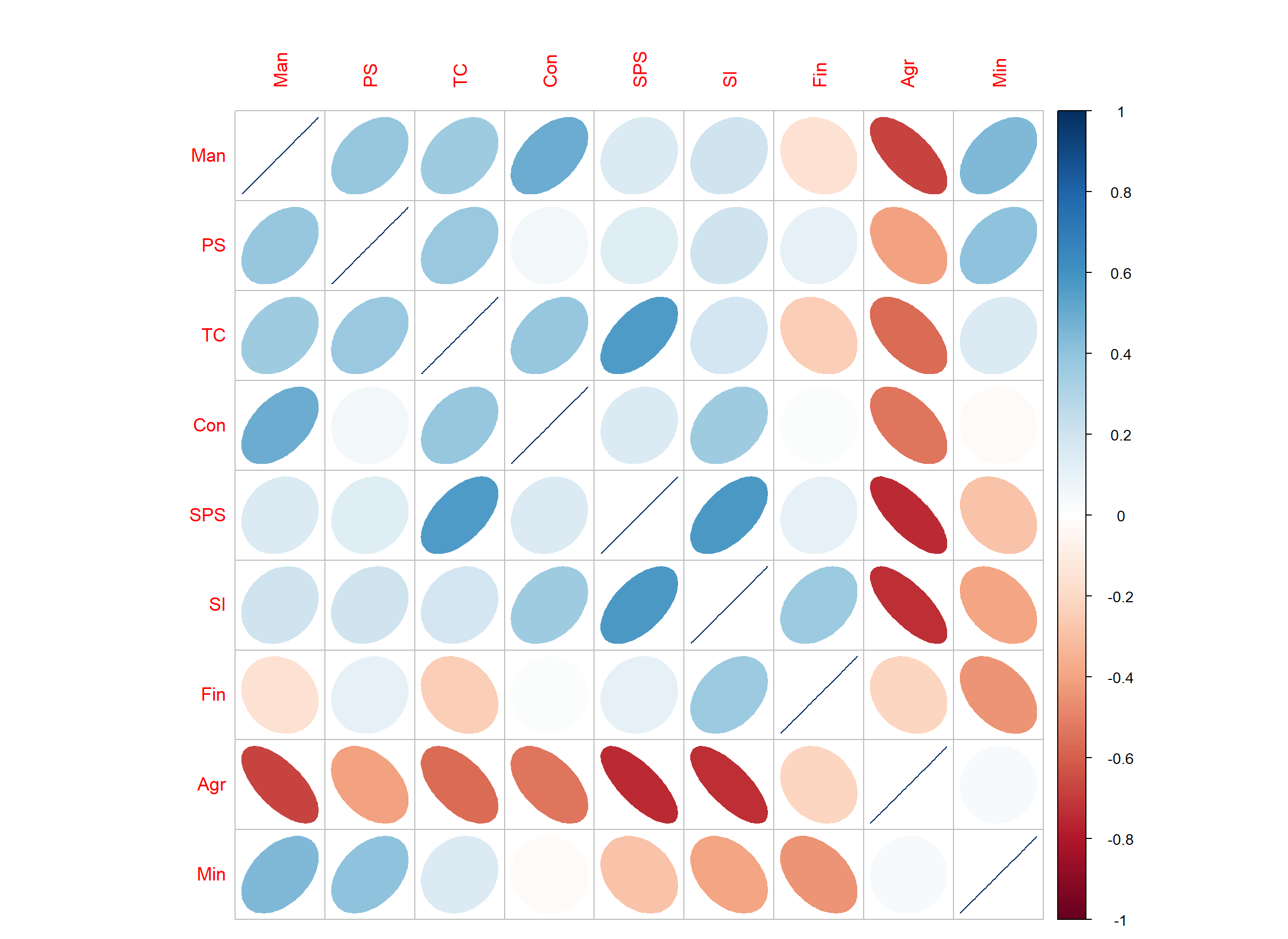
1. **What countries have the highest and lowest values for each principal component (only include the number of components specified in part a). For each of those countries, give their principal component scores (again only for the number of components specified in part a).**

Computed the principle components of each country using p$x. From the below output we can say that “Turkey” has the highest value (52.11) and “United Kingdom” has the lowest value (-18.72) for PC1. Furthermore, for PC2 “E.Germany” has the highest (10.73) and “Denmark” has the lowest value (-11.66). The principal component values for all the countries are shown in the below output.

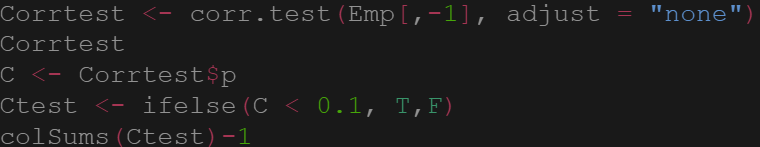


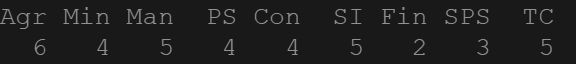
**d. Analyze the entries in the correlation matrix for fields that are highly correlated or completely uncorrelated with the other fields. If there are highly uncorrelated fields, try removing them from the model. Does this help your interpretation from part b)?**

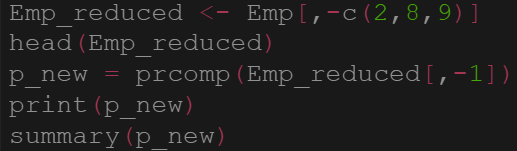


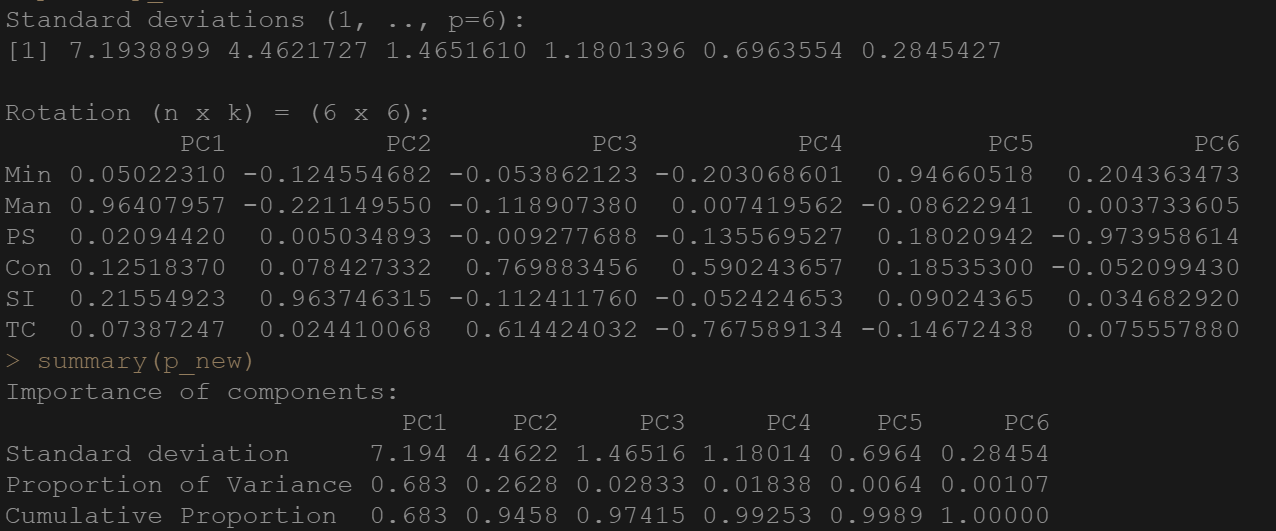


From the above correlation plot we can see that “Agr” is highly correlated with other fields. For better inference I used the corrtest and printed the column sums with 90% confidence interval. The R-Code and the outputs are shown below.









After excluding variables, we observe that two principal components are required to explain 94% of variance. Man has the highest value for PC1 so, PC1 might indicate the persons working in Manufacturing Sector. The ability to interpret the components has been improved after removing the variables.

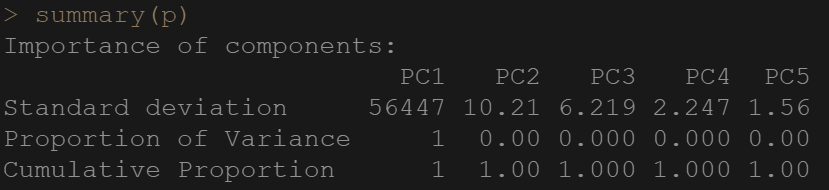
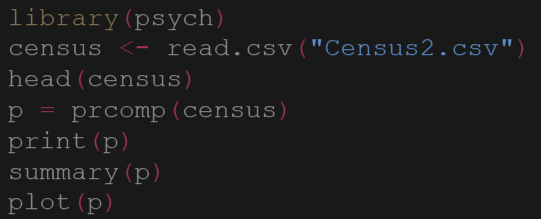
**2)** **Principal Component Analysis**

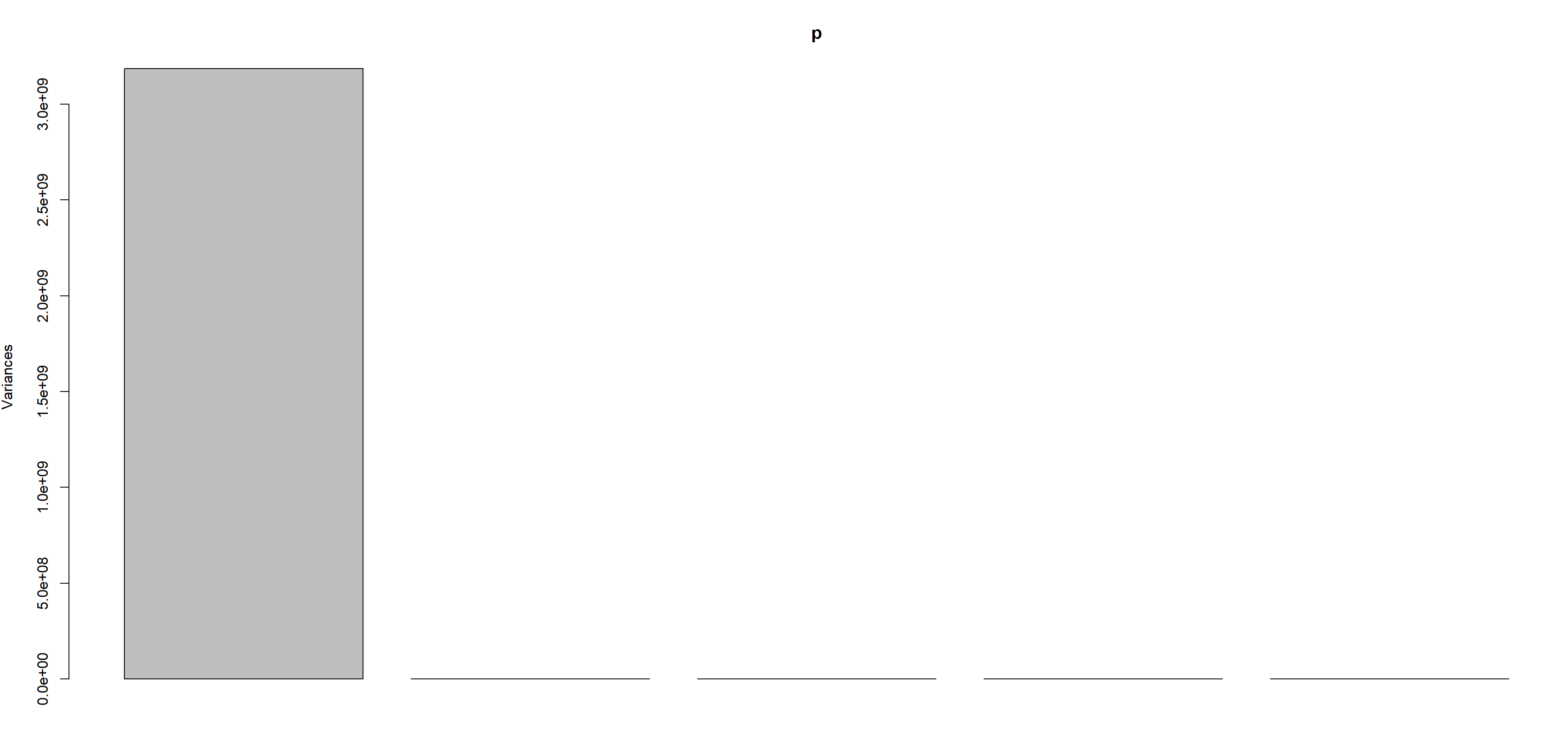
**Begin with the “census2.csv” datafile, which contains census data on various tracts in a district. The fields in the data are:**

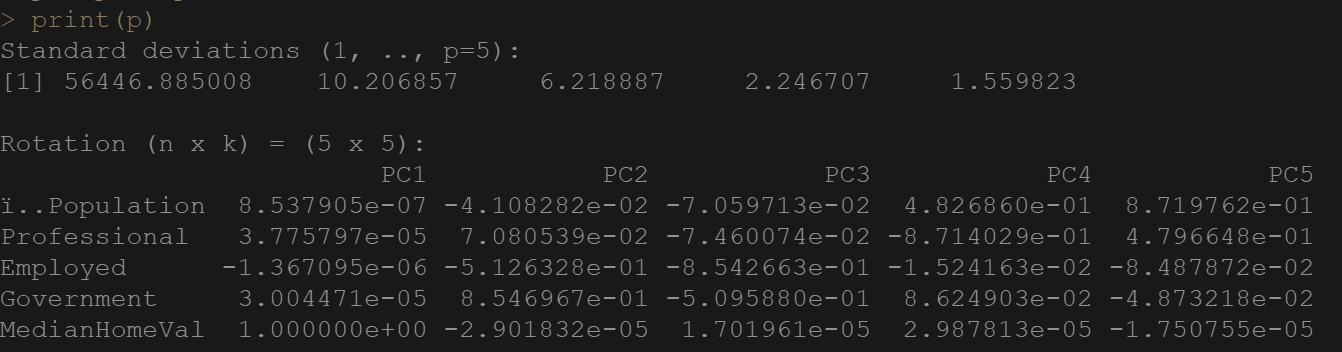
* **Total Population (thousands)**
* **Professional degree (percent)**
* **Employed age over 16 (percent)**
* **Government employed (percent)**
* **Median home value (dollars)**

1. **Conduct a principal component analysis using the covariance matrix. How much of the variance is accounted for in the first component? Why is this happening?**

Performed the principle component analysis using the covariance matrix. The R-code and the outputs are shown below,



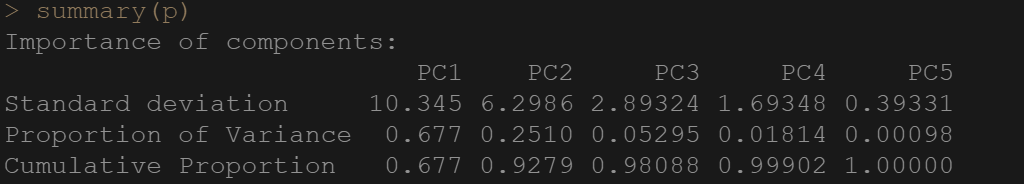


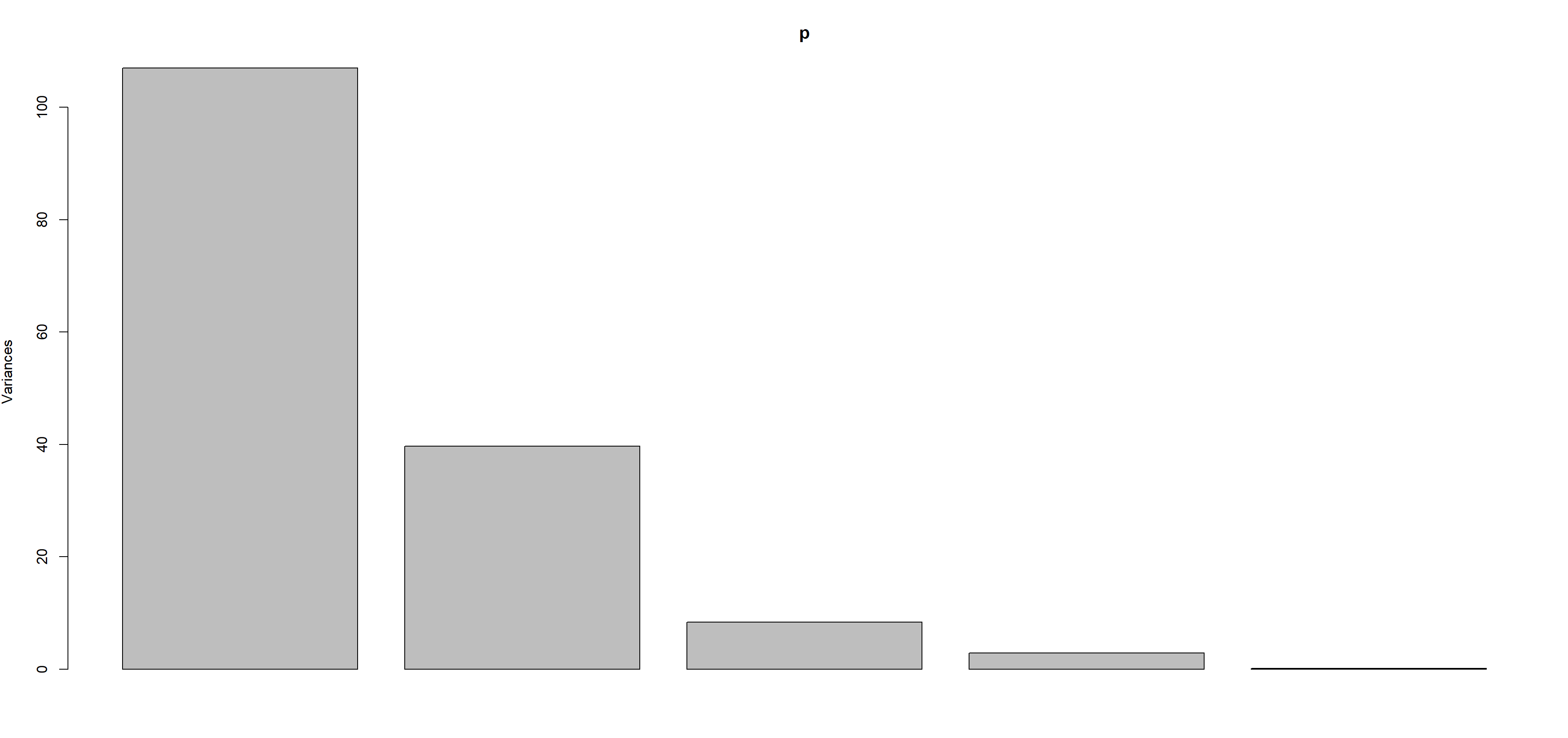


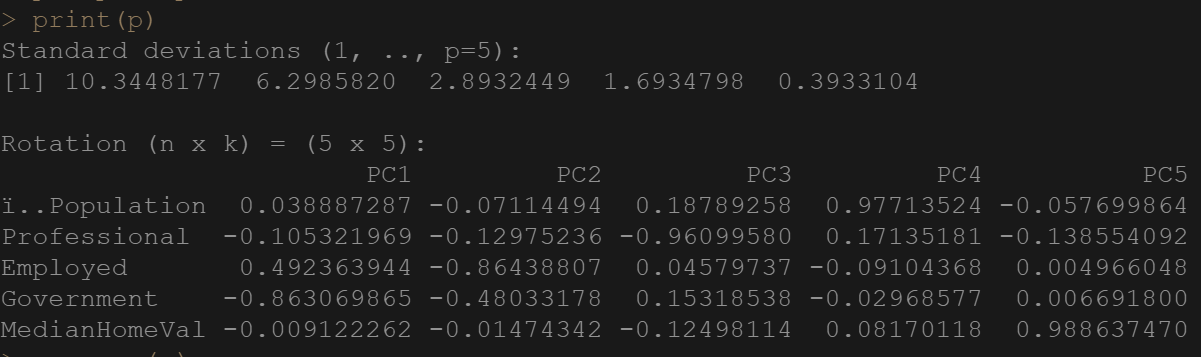
The summary statistics and screeplot indicate that the 100% variance is accounted by first component. The scores in PC1 indicate that median home value has the highest score of 1.0. This is because PCA will give more emphasis to variables that have high variance and, in the data, median home value is in 100 thousand. Since it has more variance PCA gave more weight to median home value so, the first component accounted for 100% variance.

1. **Try dividing the MedianHomeValue field by 100,000 so that the median home value in the dataset is measured in $100,000’s rather than in dollars. How does this change the analysis?**

After dividing the MedianHomeValue by 100,000 the data is scaled and other components will also contribute for variance so that, PCA seeks to maximize the variances by each component. Two components are required to explain 92% of variance in the data.



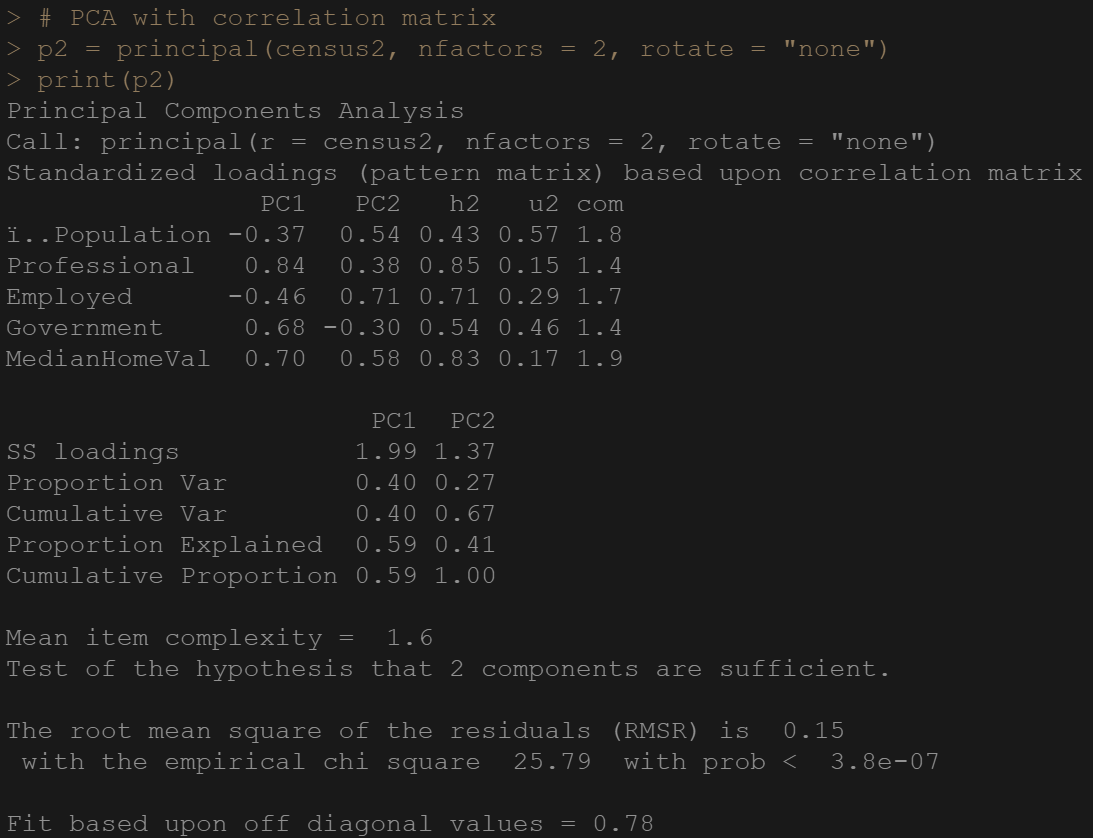




Employed has highest value for PC1 and government has lowest value. PC1 accounts for people who are employed above 16 years of age and not sure in what department they were employed in. PC2 accounts for all variables. It seems to be difficult to interpret the variables.

1. **Compute the PCA with the correlation matrix instead. How does this change the result and how does your answer compare with your answer in b)?**

Computed the PCA with the correlation matrix using the principal function. The R-code and the output is shown below. Now the interpretability of first component is quite opposite. In answer (b) professional degree, Government employed, and Median home value have negative value. This indicates that the PC1 represents the people over 16 years of age who are employed and do not have a professional degree. Furthermore, in this answer the PC1 represents the people who have a professional degree and have a government job with good median home value.



1. **Discuss what using the correlation matrix does and why it may or may not be appropriate in this case.**

In this data we have three different units of measurement (Population count, Value, and Percentage). In this case it is important to find both the strength and direction of relationship so, I think the correlation matrix for PCA is better. In our analysis correlation matrix for PCA helped in reducing the dimensionality and thereby improved the ability to interpret the variables.

**3. Common Factor Analysis**

**For this problem, you will analyze partial from intelligence tests given to children in the 'wiscsem.csv' dataset. Each child was given 11 tests on which they were rated. These were:**

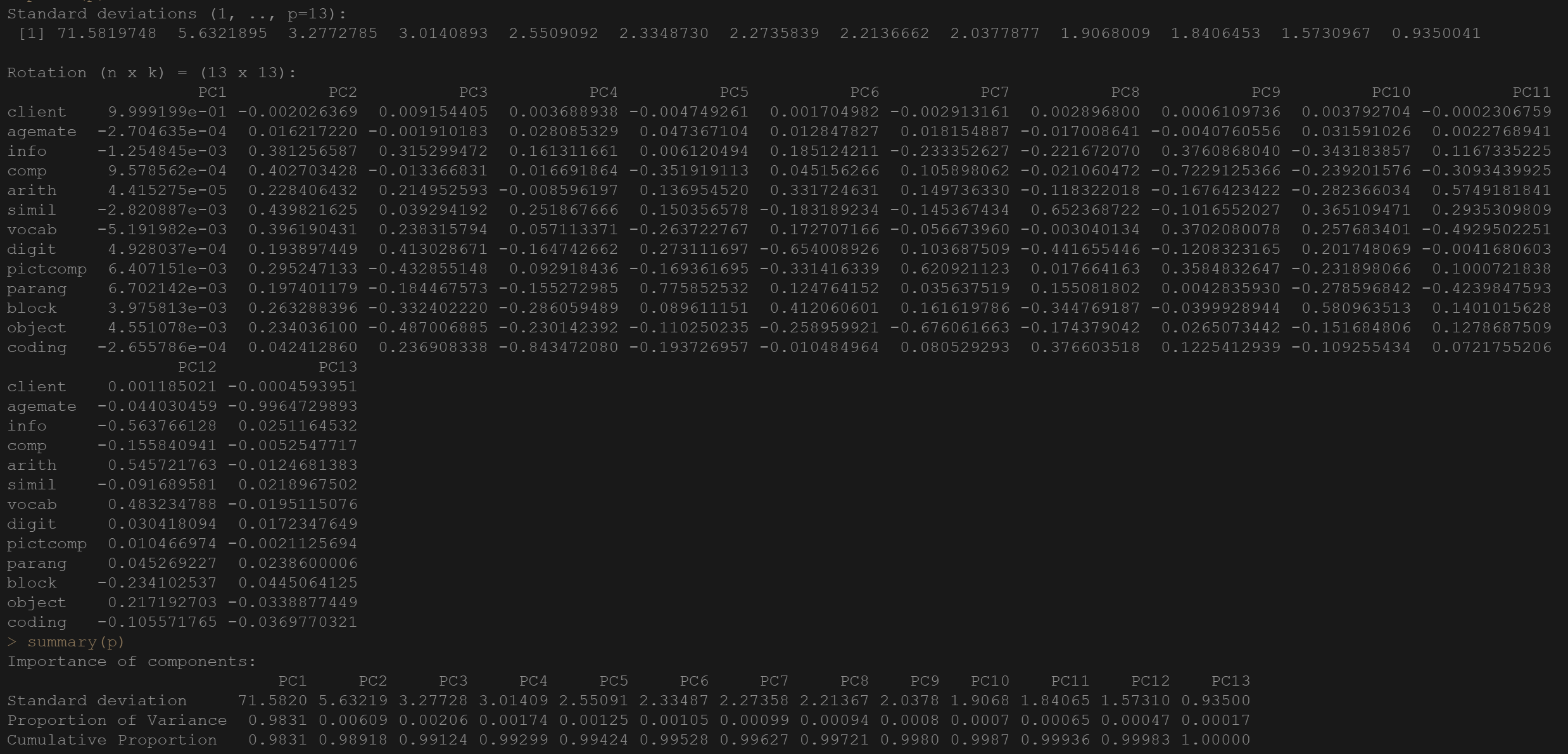
**Info = 'Information'  
comp = 'Comprehension'  
arith = 'Arithmetic'  
simil = 'Similarities'**  
**vocab = 'Vocabulary'  
digit = 'Digit Span'  
pictcomp = 'Picture Completion'  
parang = 'Paragraph Arrangement'  
block = 'Block Design'  
object = 'Object Assembly'  
coding = 'Coding';**

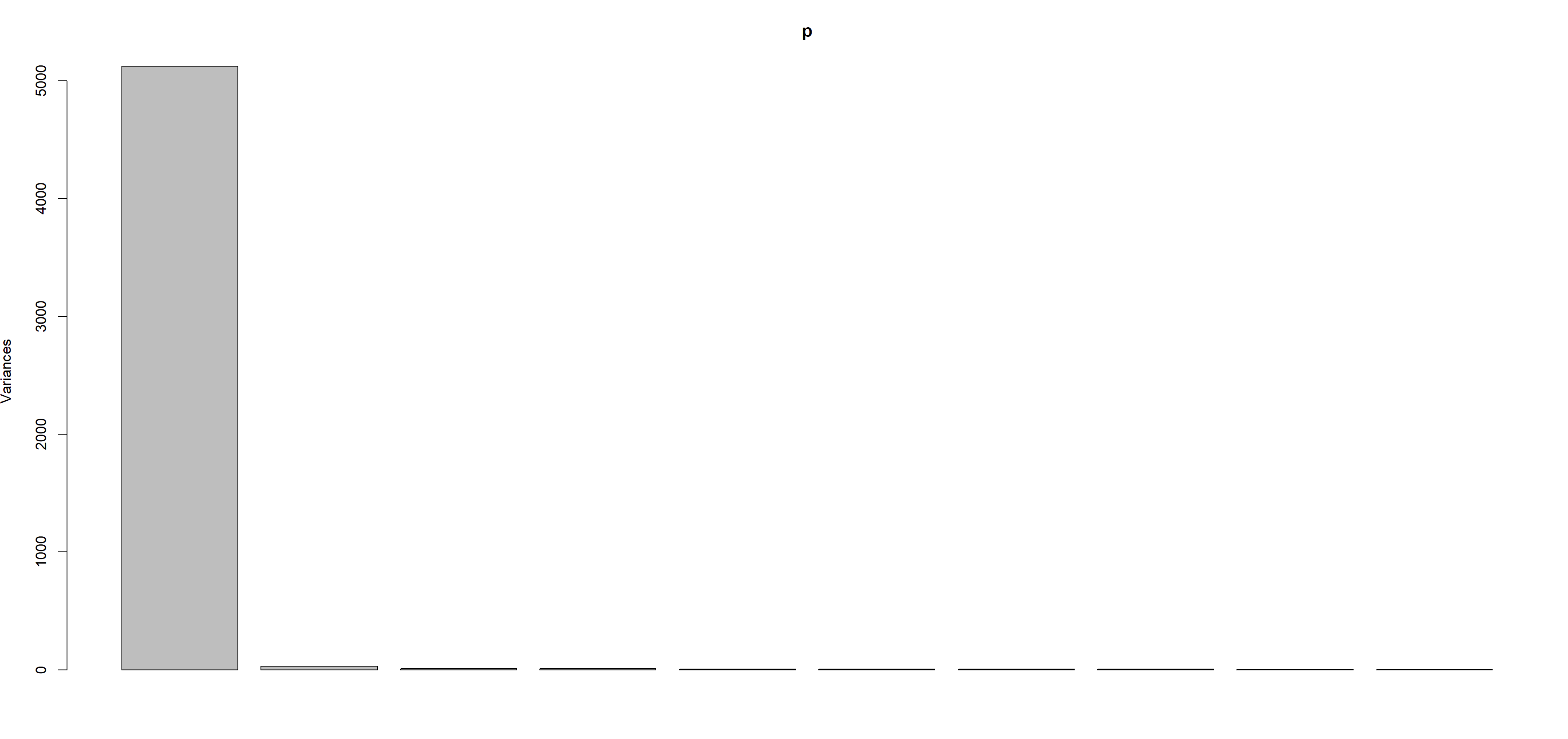
1. **Should the data be scaled or not for running PCA? Explain why/why not in detail.**

Client column indicates that the test number is different from rest of the variables. Since an ID scaling is not appropriate, it is better to exclude this column while performing PCA. Except “Client” and “Agemate” all other observations in the data are the intelligence test results given to children they are in same scale so, I think scaling is not required (If we exclude “Client” and “Agemate”) for running PCA. But if we run principal factor analysis it will automatically scale the data.

1. **Run an initial corrplot and an initial PCA. Use the corrplot and a scree plot to determine the appropriate number of factors to extract.**

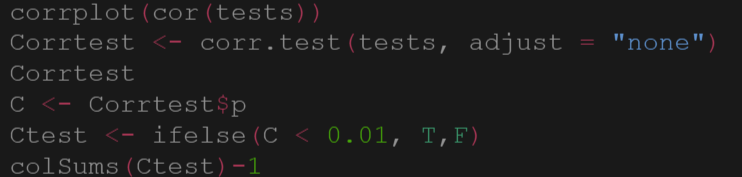
Computed the initial PCA and ran the correlation plot. The R-code and the outputs are shown below.

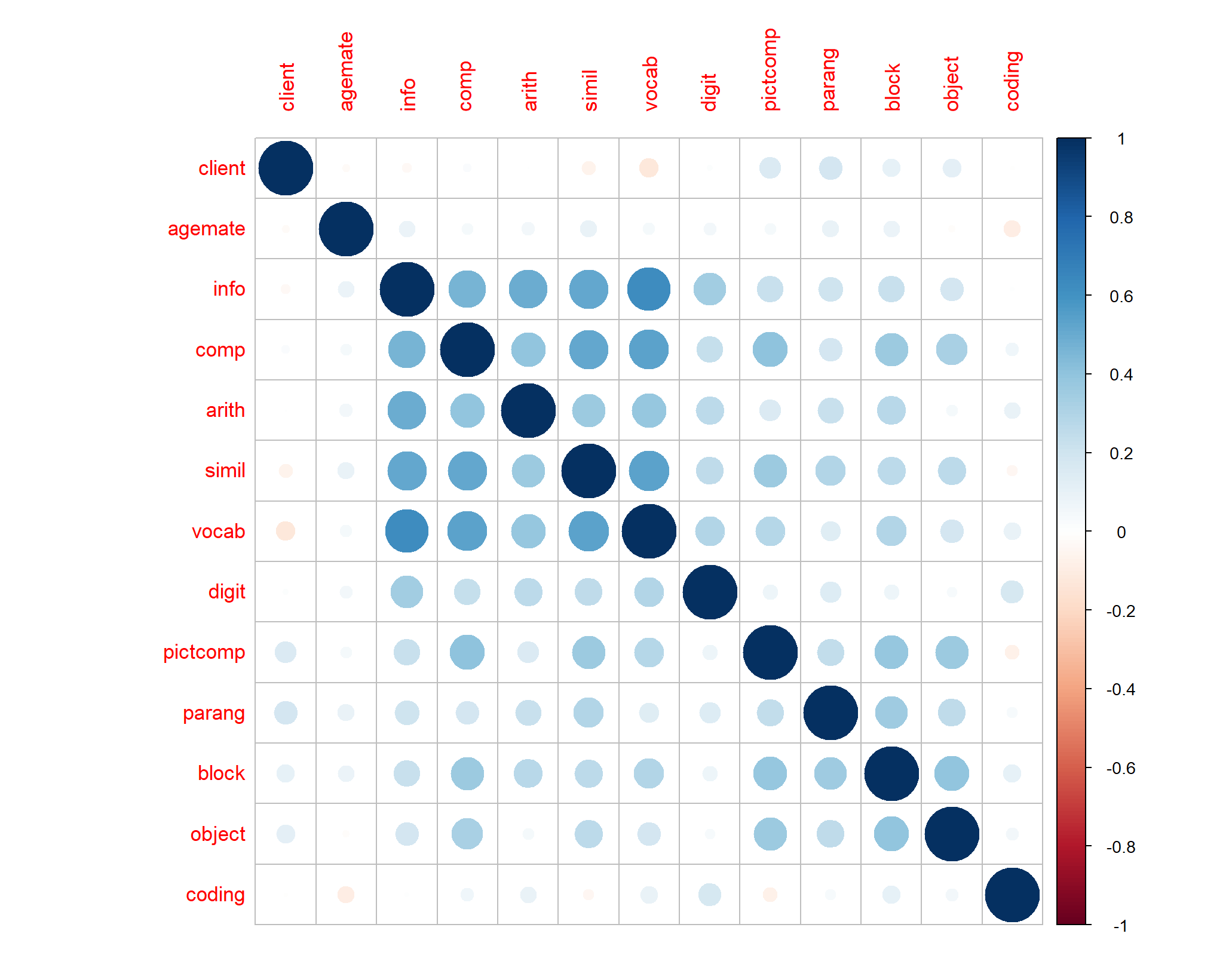


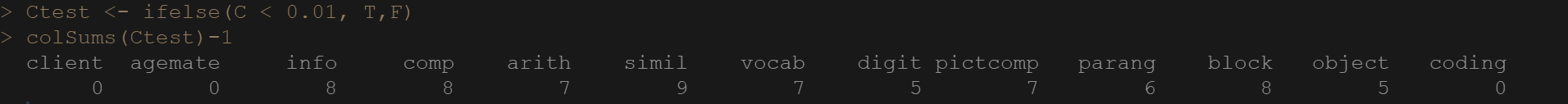


From the above tables we can infer that the first component captured 98% of variance. This is because there is large variance in “Client”. The screeplot indicates the same. Next, I run corrplot and calculated the column sums from correlation test matrix to find the appropriate number of factors.

From the below corrplot we can observe that there are two groups and “Client”, “Agemate” and “coding” are uncorrelated with others. I run the corrtest and calculated the column sums to get the exact factors.







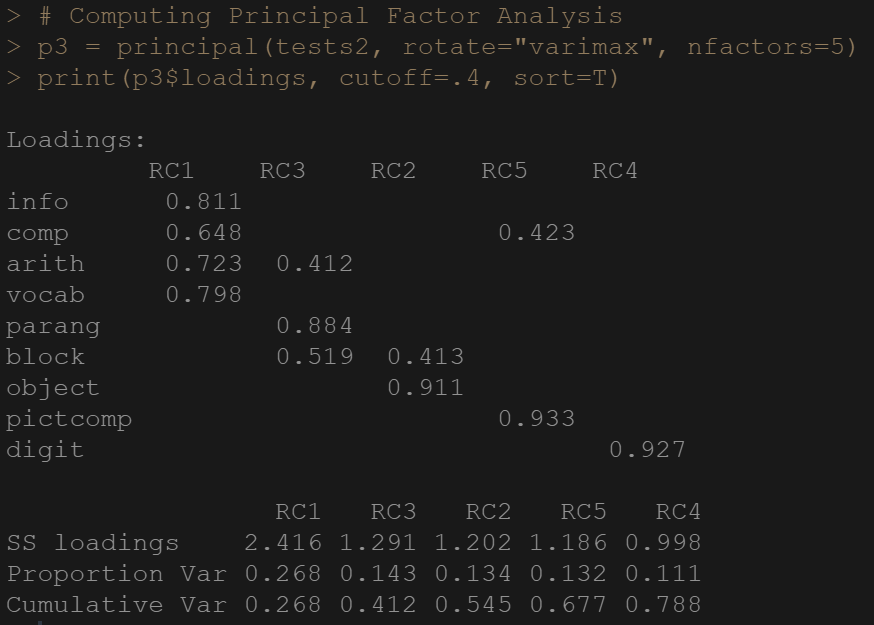
From the column sums it is evident that the “Client”, “Agemate” and “Coding” are uncorrelated and “Simil” is correlated with most of the variables. So, we can exclude these variables and perform the analysis. Now there will be 9 optimal number of factors.

1. **Are there any variables that will likely be single-variable factors? Explain.**

The highest PC1 are Client with 9.99, Comp 9.57, PictComp 6.40, and Parang 6.70. The highest PC2 are Comp is 0.40. The Lowest PC1 is info with -1.25, Agemate with -1.25 and Coding of -2.66. The Lowest for PC2 Coding is 0.04, Agemate 0.01.

1. **Run a Principal Factor Analysis with VARIMAX rotation and report the loadings with a cutoff of 0.4. How clean are the variable separations? Give a name to each factor.**

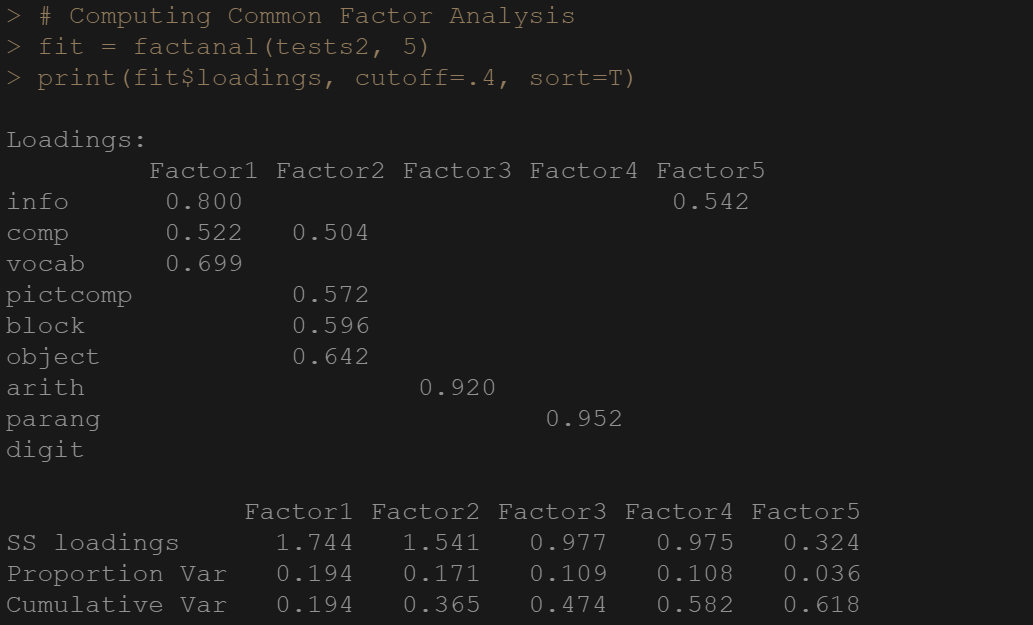
Five components are required to get 79% of variance so, I computed PFA with 5 factors, with VARIMAX and then reported the loading with a cut-off 0.4.

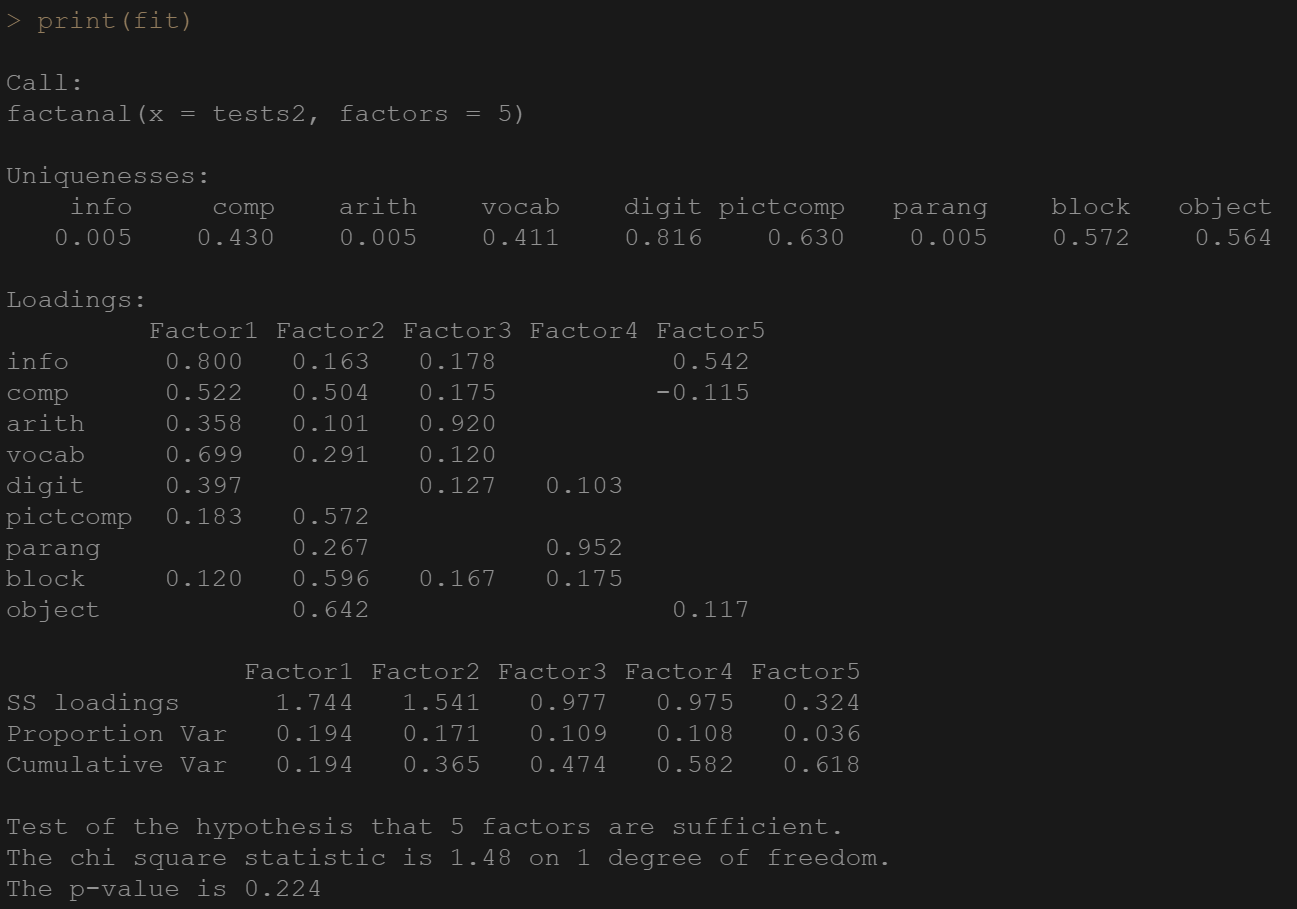


After performing PFA the components are easy to interpret when compared with previous one. RC1 represents the students who performed well in Comprehension and Arithmetic. RC2 represents Design and Assemble. RC3 represents Comprehension and Assemble. RC4 represents Digit span and RC5 represents the Comprehension and Puzzle.

1. **Run a Common Factor Analysis (exploratory) and compare the loadings to those of the principal factor analysis. Note any significant differences and explain how they affect the factors practically.**

Computed common factor analysis with five factors and displayed the loadings with cut-off 0.4. The R-Code and outputs are shown below,





When we compare the loadings of principal and common factor analysis, the ability to interpret the components is improved for CFA. The difference is in factor 1 Arith has low loading factor1 which indicates that the student are good at vocabulary. Comp, pictcomp are added in factor2. Arithmetic, Parang and Info has the highest loadings in Factors 3, 4 and 5 respectively. In PCA the cumulative variance is not linear. In CFA, the cumulative variance is linear and 61% of variance is captured by five factors.