GLOBAL UNEMPOYMENT DATASET

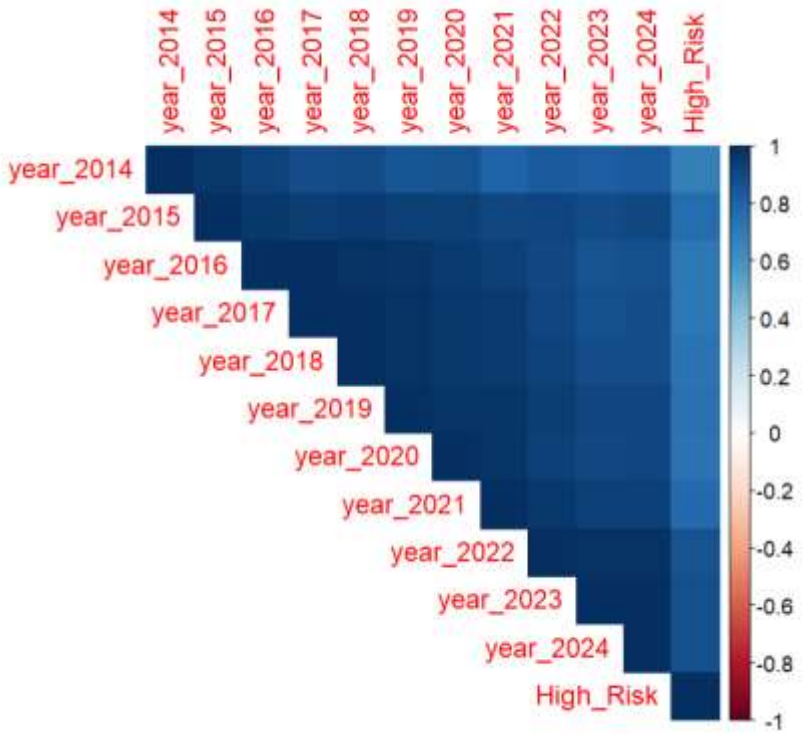
Dataset - https://www.kaggle.com/datasets/sazidthe1/global-unemployment-data/

Data Dictionary -

Country Name: The name of the country where the data is collected.
Sex (Gender): The gender of individuals, categorized as male or female.
Age Group: The grouping of individuals based on their age, such as under 15 years, 15-24 years, and 55+ years.
Age Category: The broader categorization of age groups, such as children (15 years),youth (15-24 years), and adult (55+ years).
year_2014 to year_2024: The population data for each year from 2014 to 2024.

Questions:

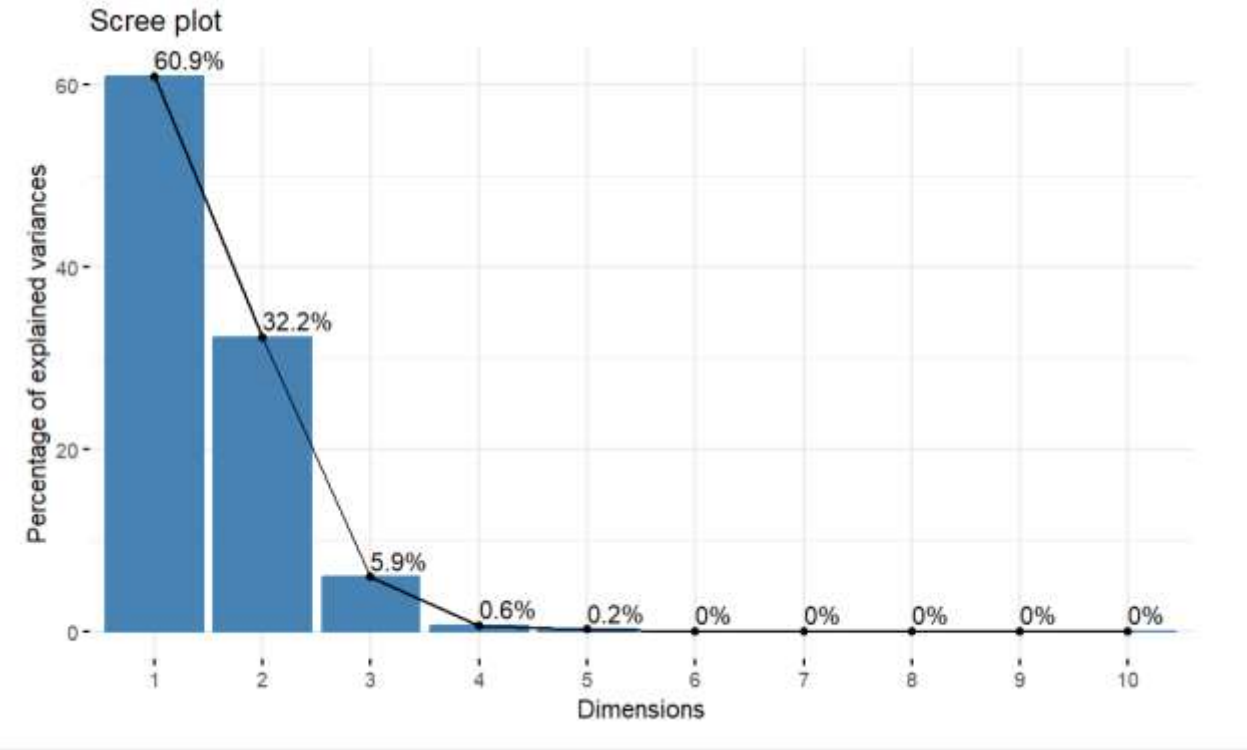
1. Can individuals be classified into different countries based on their demographic characteristics such as sex, age group, and population distribution?
2. Can we predict whether an individual belongs to a high-risk category based on their demographic information and population distribution?



The correlation plot shows correlation, so we start with Principal Component Analysis (PCA) to reduce the number of variables.

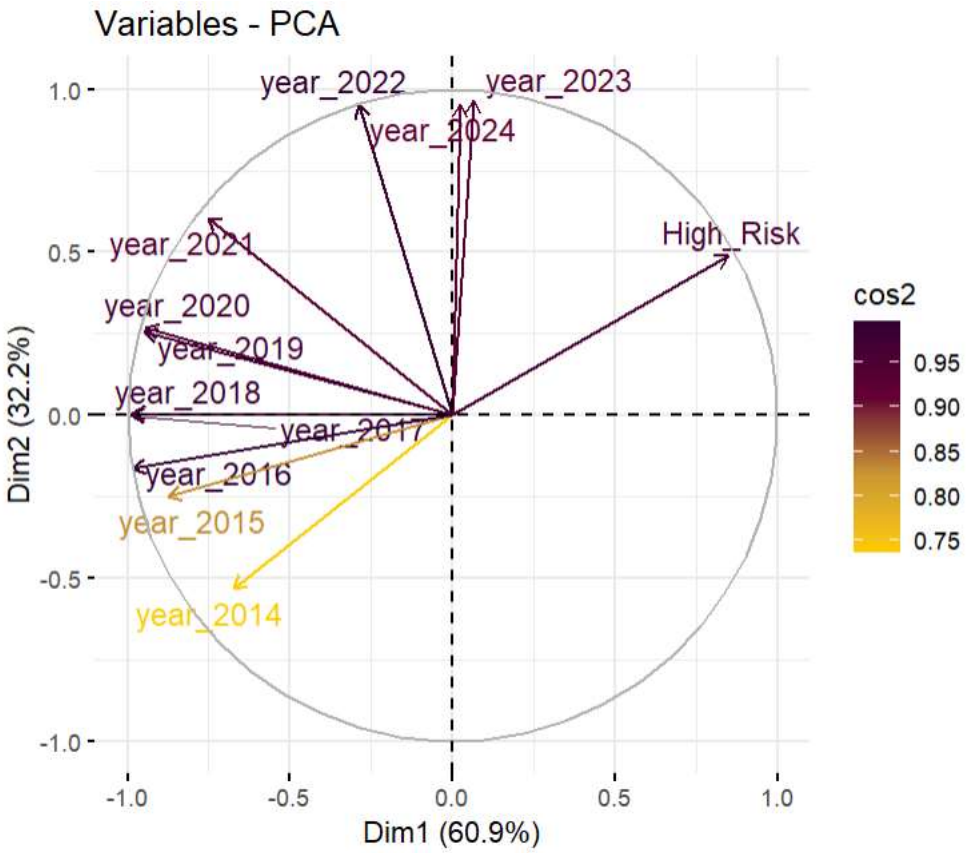
PCA gives the same number of principal components as the number of columns, which, in our analysis, is 12. The principal components that we obtained are as follows:

```
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6     PC7     PC8     PC9
Standard deviation     2.7042 1.9666 0.84397 0.27669 0.1471 0.06923 0.06338 0.01932 0.01156
Proportion of Variance 0.6094 0.3223 0.05936 0.00638 0.0018 0.00040 0.00033 0.00003 0.00001
Cumulative Proportion  0.6094 0.9317 0.99104 0.99742 0.9992 0.99962 0.99996 0.99999 1.00000
                          PC10     PC11      PC12
Standard deviation     0.003934 0.003168 7.763e-16
Proportion of Variance 0.000000 0.000000 0.000e+00
Cumulative Proportion  1.000000 1.000000 1.000e+00
```



The scree plot determines how many Principal Components (PCs) to use for the analysis.

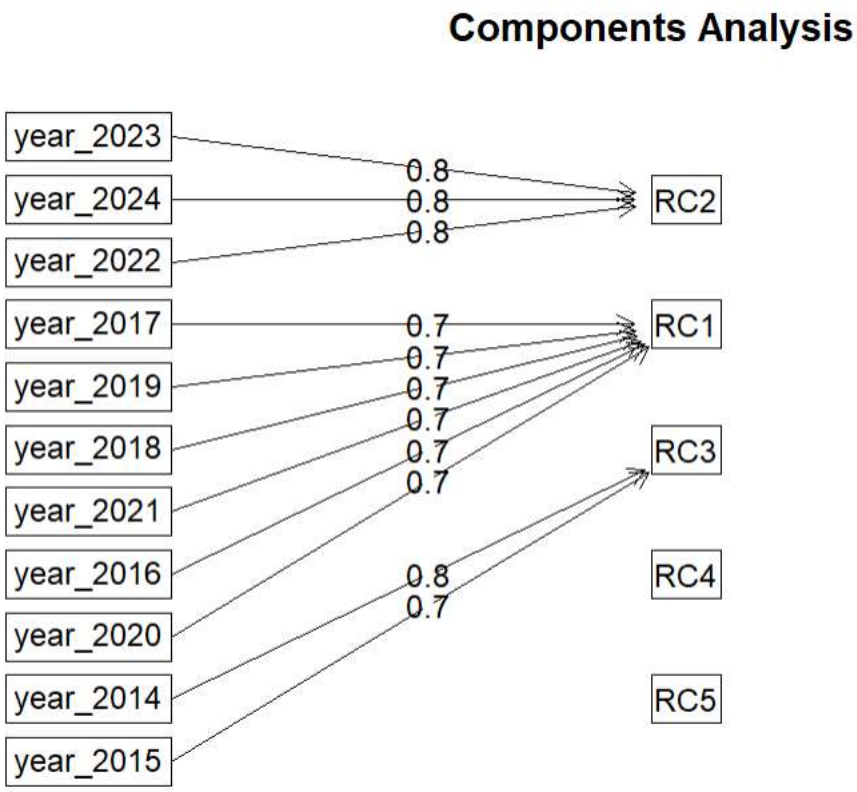
- The significant bend in the plot is used to determine the number of PCs to be used.
- The plot shows us the number of components to be considered is 2. (93.1% of variance)



The distance between points in a biplot reflects the generalized distance between them. The length of the vector reflects the variance of the variable. Correlation of the variables reflected by the angle between them. The smaller the angle, the more significant the correlation.
For example, it shows that year 2015 and year 2016 are all correlated strongly.

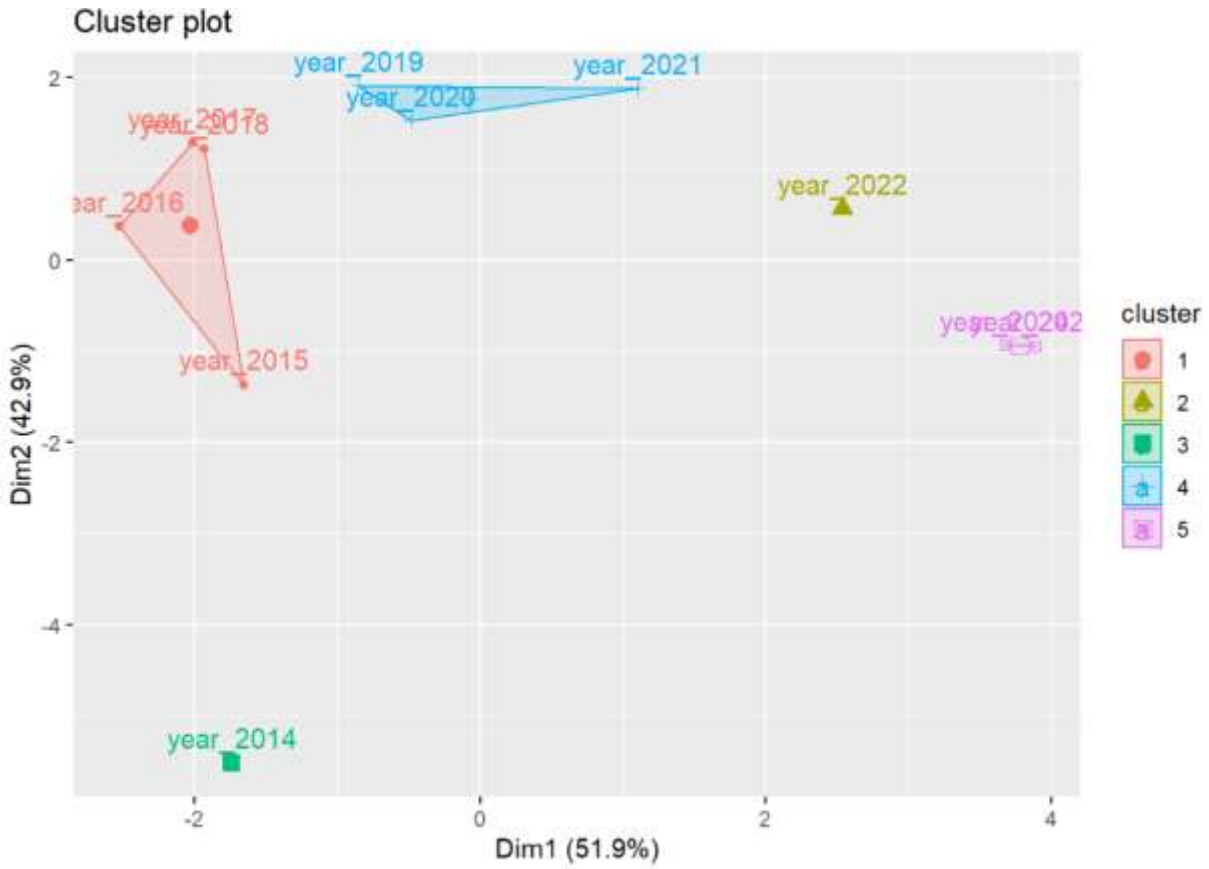
The first principal component (Dim1), which accounts for 60.9% of the variance, is strongly associated with the years ascending from 2014 to 2024. This suggests a trend over time where more recent years have higher values on this component.
The second principal component (Dim2), accounting for 32.2% of the variance, separates earlier years from the latest ones, possibly indicating a change in the pattern of the data over time.

Move on to check Exploratory Factor Analysis (EFA).



RC1 - "Long-term Trend" if it has significant factor loadings from all years, suggesting it captures a pattern persistent across all years.
RC2 - "Recent Change" if it has high loadings for the most recent years, indicating it captures a trend that is more pronounced in recent times.
RC3 - "Mid-term Fluctuation" if it captures variance in the middle years of your dataset, suggesting fluctuations or trends that are not present in the earliest or latest years.

Clustering



Cluster 1 (red, circle): Includes years from 2014 to 2016. This cluster has a downward trajectory on Dim2, suggesting a particular trend or characteristic was diminishing over these years.
Cluster 2 (yellow, triangle): Contains only the year 2022. This might indicate that the population characteristics in 2022 were unique compared to other years.
Cluster 3 (purple, cross): Includes only the year 2023, which, like 2022, may have unique characteristics that separate it from other years.
Cluster 4 (blue, square): Encompasses years from 2018 to 2021, indicating these years share similar population characteristics.
Cluster 5 (green, diamond): Isolated with the year 2014, possibly signifying that 2014's population distribution was significantly different from other years.

We can try to generate a confusion matrix for the above analysis and see the results more clearly.

```
##           Predicted
## Actual     Children Youth Adults
##   Children        6    14      0
##   Youth           7    13      0
##   Adults          7    13      0
```

We see that the clustering has classified the colleges with 31.9% accuracy.

• We can conclude that we cannot classify the unemployment based on the age categories on the data provided.

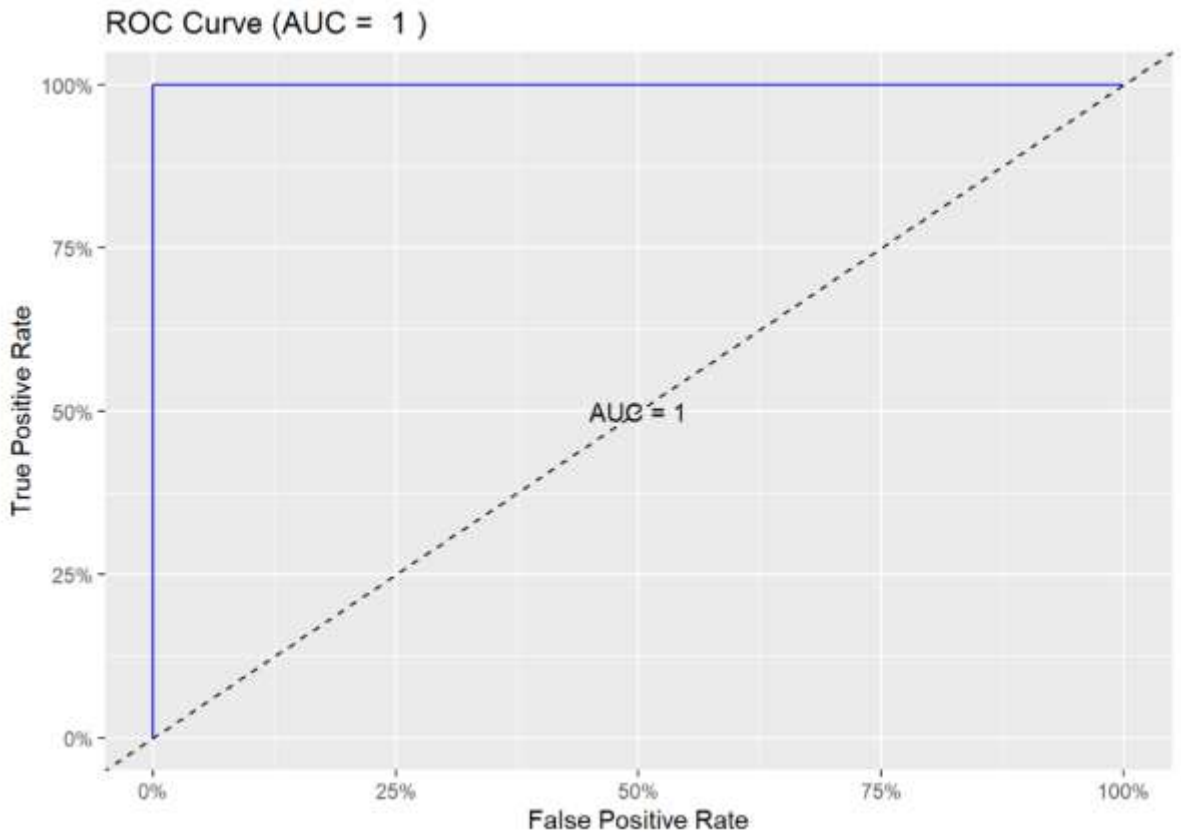
Q2 SOLUTION:

We define training and testing sets to perform logistic regression as the output variable is qualitative and has two factors (High Risk/Low Risk)

```
##
## Call:
## glm(formula = High_Risk ~ . - Country_name - Age_group - Age_categories -
##     year_2014 - year_2015 - year_2016 - year_2017 - year_2018 -
##     year_2019 - year_2020 - year_2021 - year_2022 - year_2023 -
##     High_Risk, family = "binomial", data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2247.5   426960.1  -0.005    0.996
## SexMale       -152.6   349730.5   0.000    1.000
## year_2024      240.7    26632.9   0.009    0.993
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.5193e+01  on 59  degrees of freedom
## Residual deviance: 2.2111e-07  on 57  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 25
```

```
##           actual
## predicted No Yes
##       No  46   0
##       Yes  0  14
```

The regression gives a confusion matrix with an accuracy of 1.



The AUC is obtained to be 1 which is excellent and tells us that our prediction works well.

Based on the given data, we could classify the age_categories
o The accuracy of the classification came out to be 93.1%
o We have used Exploratory Factor Analysis and Clustering for this classification.

We could also predict High Risk and Low Risk based on the variables provided.
o Using logistic regression, we predicted the type of college with 1 accuracy.
o An AUC of 1 for the ROC curve shows good prediction