# Exploring the Link Between Household and Climate Change Factors and Educational Achievement in Rural Pakistan: A PCA and Logistic Regression Study

Ali Abid

2024-12-03

## Contents

# Introduction and Situational Context

Despite being the fifth most populous country in the world, Pakistan is only able to spend as much as 1.7 percent of its GDP on education (Abbasi, 2023). The tumultuous domestic political climate, in conjunction with the global geopolitical landscape, has undeniably resulted in a series of macroeconomic crises within the nation. These crises have unfortunately led to widespread inflation, a rise in poverty, and a profound literacy crisis that poses significant challenges for future generations. A sizable children population (around 26 million) remains out of school (Haider, 2024) which can have devastating impact on the opportunities of the country to grow out of these crises.

# Purpose

The purpose of this research project is to understand the the factors that influence students' general knowledge by utilizing principal component analysis (PCA) to reduce the dataset into key components.

The dataset used in this analysis is the ASER 2023 rural dataset which is a household survey conducted in Pakistan to assess the learning levels of children in the country. The dataset contains information on the academic outcomes including whether a student exhibited general knowledge while being surveyed, the said students' household characteristics, and the impact of climate change as reported by their households. The analysis will focus on the factors that contribute to the general knowledge of students in Pakistan and will use logistic regression to predict the outcome in a probabilistic form based on these predictors.

# Data Description

**Predictor Variables**:

- **HouseholdCounter**: Number of people in the household
- **EarningMembers**:Number of earning members in the household
- **TravelTime**:Time taken to travel to school
- **Car**: Number of cars in the household
- **MotorCycle**: Number of motorcycles in the household
- **ClimateChange**: Whether the household has been impacted by climate change and to which severity on a scale of 1-4
- **MigrantIDP**: Whether the household was categorized as *migrant* due to flood impact (0 - Not a migrant, 1 - Migrant due to flood impact)
- **FloodImpacted**: Whether the household has been impacted by floods and to which severity on a scale of 1-3
    - 1.Yes, significantly 2. Yes, moderately 3. No, not affected

- **EarningImpacted**: Whether the household has been impacted by loss of earnings due to climate change and to which severity
  - 1.Less than 10% 2. btw 11%-25% 3. btw 26%-50% 4. More than 50% 5. No affect
- **PsychologicalImpacted**: Whether the household has been impacted by psychological distress due to climate change and to which severity on a scale of 1-4
  - 1. Substantially 2. Somewhat affected 3. Affected only a bit 4. Not at all
- **SchoolingAffected**: Whether the household reports that the student's schooling has been impacted due to climate change and to which severity on a scale of 1-4
  - 1. Extremely affected 2. Moderately affected 3. Somewhat affected 4. Not at all

**Outcome Variables**:

- GeneralKnowledge: Whether the student exhibited general knowledge or not while tested by the ASER data collection team.
  - 1. Yes 0. No

# Research Question

Can student's household characteristics, and climate-change factors affecting the student's household be combined into principal components to identify underlying factors, and how well do these factors predict a student's general knowledge level?

# Analysis

```r
rm(list=ls(all=TRUE))
```

## Required Libraries

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tibble' was built under R version 4.3.3

## Warning: package 'tidyr' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
## Warning: package 'stringr' was built under R version 4.3.3
```

```
## Warning: package 'forcats' was built under R version 4.3.3
```

```
## Warning: package 'lubridate' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(readr)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.3.3
```

```
##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.3.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

## Load the data

The two distinct datasets are loaded into R. The first dataset contains the child-level data, while the second dataset contains the household-level data. The two datasets are merged on the *HouseholdId* column by first renaming the variable to *HHID* which is unique to both the datasets to create a single dataset for analysis.

```r
# Load the data
aser_child <- read_csv("ITAASER2023Child.csv")
```

```
## Rows: 214014 Columns: 47
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): RNAME, DNAME, C06, C08, C09, C10, C14, ICH02, BasicVaccines, Aller...
## dbl (36): Id, PrvCode, DstCode, VCODES, C03, C04, C05, C11, C12, C07, C13, C...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
aser_household <- read_csv("ITAASER2023Household.csv") %>%
  rename(HHID = HouseholdId) # Renaming the HouseholdId column to HHID for merging with the Child Datas
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 89602 Columns: 58
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): MotherLanguage, Religion
## dbl (54): HouseholdId, VillageMapSurveyId, HouseholdCounter, IsFamilyHead, G...
## lgl  (2): TotalFuctionalToilets, TravelMode
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

## Data Wrangling

The two datasets are merged on the HHID column to create a single dataset *aser_child_household_data* for analysis. The variables of interest as discussed in the section of predictor variables and outcome variable are selected and renamed for clarity.

```r
# Merging the two datasets
aser_child_household_data <- aser_child %>%
  left_join(aser_household, by = "HHID")

aser_child_household_data <- aser_child_household_data %>%
  select(
    # Household Characteristics
    HouseholdCounter, EarningMembers, Car, MotorCycle,
    # Time to school
    TravelTime,

    # Climate Change Impact
    ClimateChange, FloodImpacted, EarningImpacted, PsychologicalImpacted, SchoolingAffected, MigrantIDP

    # Child Characteristics - GeneralKnowledge
    C27
  ) %>%
  rename(

    # Child Characteristics - GeneralKnowledge - whether the child exhibits GK or not
    GeneralKnolwedge = C27

  )
```

## Data Cleaning

The dataset is cleaned by removing rows with missing values and changing column types to numeric for further analysis.

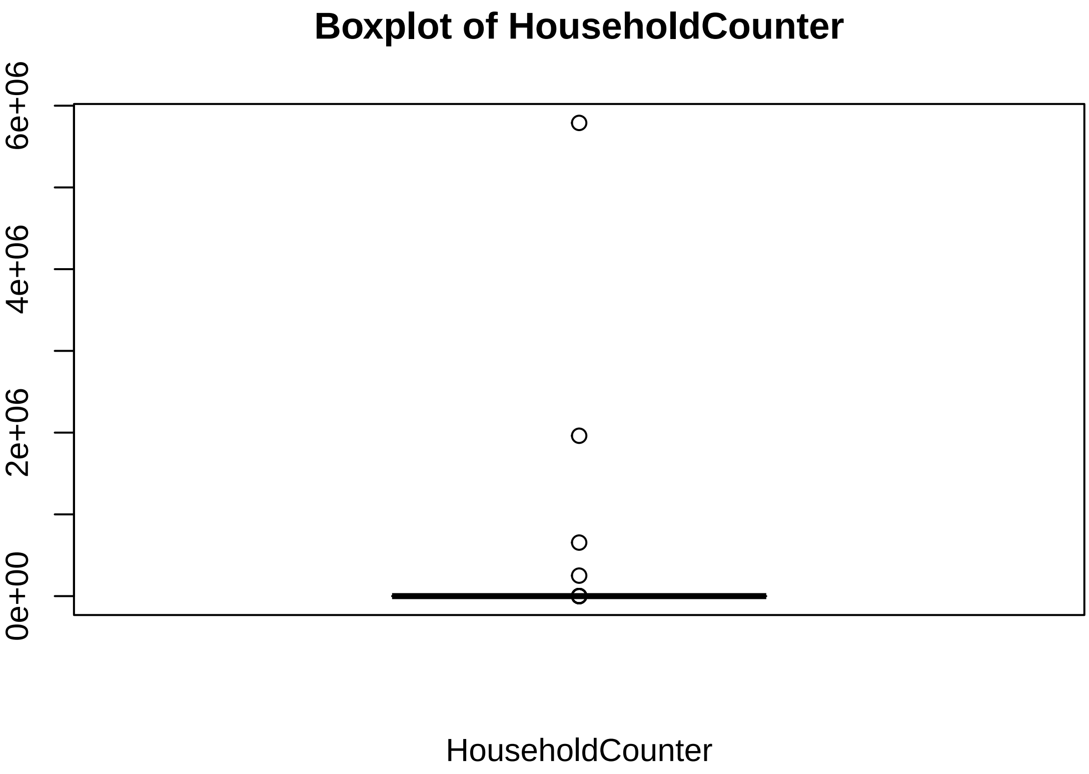
```
# Removing the rows with missing values
aser_child_household_tib <- aser_child_household_data %>%
  na.omit() %>% #get rid of rows with NAs
  mutate_at(c(1:12),as.numeric) #change all columns to numeric
  #mutate(GeneralKnolwedge = as.factor(GeneralKnolwedge))  #change GK Score to factor
  #mutate_at(c(1:12), ~(scale(.) %>% as.vector))
  #scale all variables so mean is zero and values are standardized to SD from zero
  #as.vector ensures columns are vectors

library(psych)

psych::describe(aser_child_household_tib) #gives you a lot of descriptives quickly
```

```
##                       vars    n    mean        sd median trimmed  mad min
## HouseholdCounter         1 2378 3651.27 126147.96     11   10.50 7.41   0
## EarningMembers           2 2378   21.78    535.60      1    1.03 1.48   0
## Car                      3 2378    0.79      0.92      1    0.71 0.00   0
## MotorCycle               4 2378    0.89      0.57      1    0.86 0.00   0
## TravelTime               5 2378    1.52      0.70      1    1.40 0.00   1
## ClimateChange            6 2378    2.00      0.81      2    2.00 1.48   1
## FloodImpacted            7 2378    2.56      0.73      3    2.70 0.00   1
## EarningImpacted          8 2378    4.06      1.40      5    4.32 0.00   1
## PsychologicalImpacted    9 2378    3.30      1.07      4    3.50 0.00   1
## SchoolingAffected       10 2378    3.41      1.00      4    3.62 0.00   1
## MigrantIDP              11 2378    0.04      0.19      0    0.00 0.00   0
## GeneralKnolwedge        12 2378    0.49      0.50      0    0.49 0.00   0
##                           max   range  skew kurtosis      se
## HouseholdCounter      5789321 5789321 42.20  1882.46 2586.87
## EarningMembers          15000   15000 27.63   768.76   10.98
## Car                         9       9  4.91    37.32    0.02
## MotorCycle                  5       5  0.49     2.91    0.01
## TravelTime                   3        2  0.98     -0.35     0.01
## ClimateChange                3        2  0.00     -1.49     0.02
## FloodImpacted                3        2 -1.31      0.14     0.01
## EarningImpacted              5        4 -1.15     -0.17     0.03
## PsychologicalImpacted        4        3 -1.13     -0.32     0.02
## SchoolingAffected            4        3 -1.44      0.61     0.02
## MigrantIDP                   1        1  4.72     20.32     0.00
## GeneralKnolwedge             1        1  0.02     -2.00     0.01
```

```
# There seems to be an outlier value in HouseholdCounter, which needs investigation.
```

The boxplot is used to identify and remove an outlier value in the *HouseholdCounter* variable.

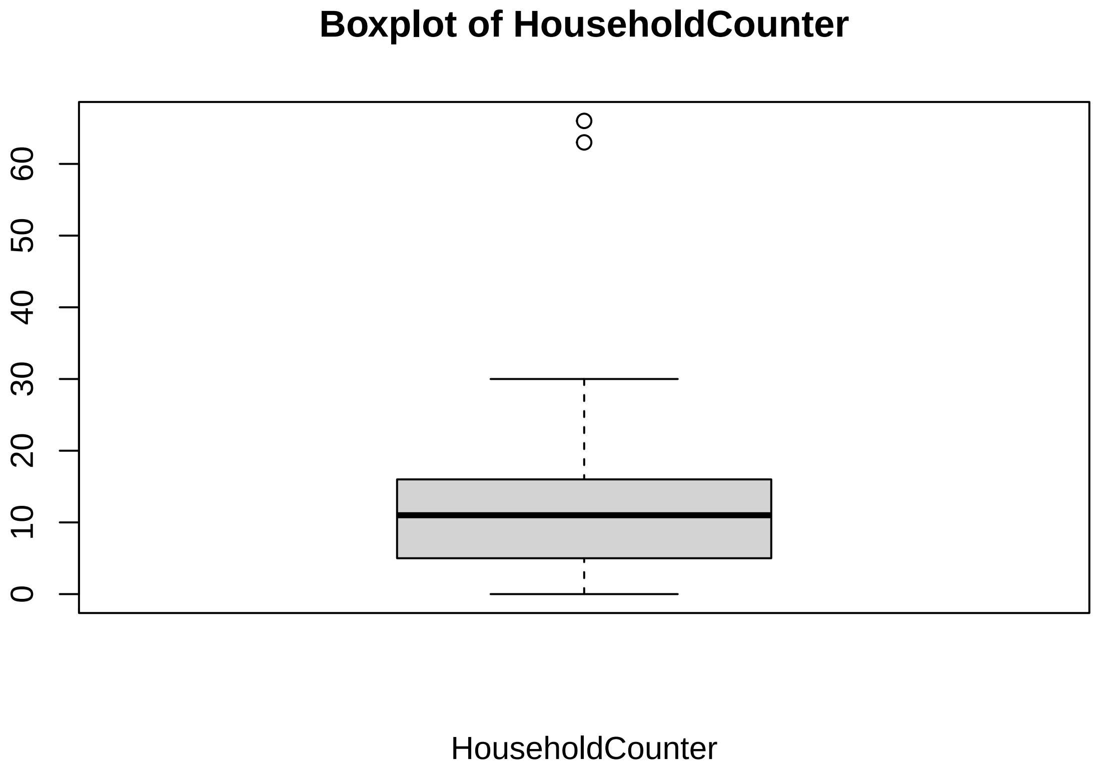
```
boxplot(aser_child_household_tib$HouseholdCounter,
        main = "Boxplot of HouseholdCounter",
        xlab = "HouseholdCounter")
```



```
# Let's just try to find out which is the outlier value

max(aser_child_household_tib$HouseholdCounter)
```

```
## [1] 5789321
```

```
# There is an outlier value in HouseholdCounter, which is 5789321. We will remove this row.

aser_child_household_tib <- aser_child_household_tib %>%
  filter(HouseholdCounter < 100)

# Checking the boxplot again to see if the outlier has been removed

boxplot(aser_child_household_tib$HouseholdCounter,
        main = "Boxplot of HouseholdCounter",
        xlab = "HouseholdCounter")
```



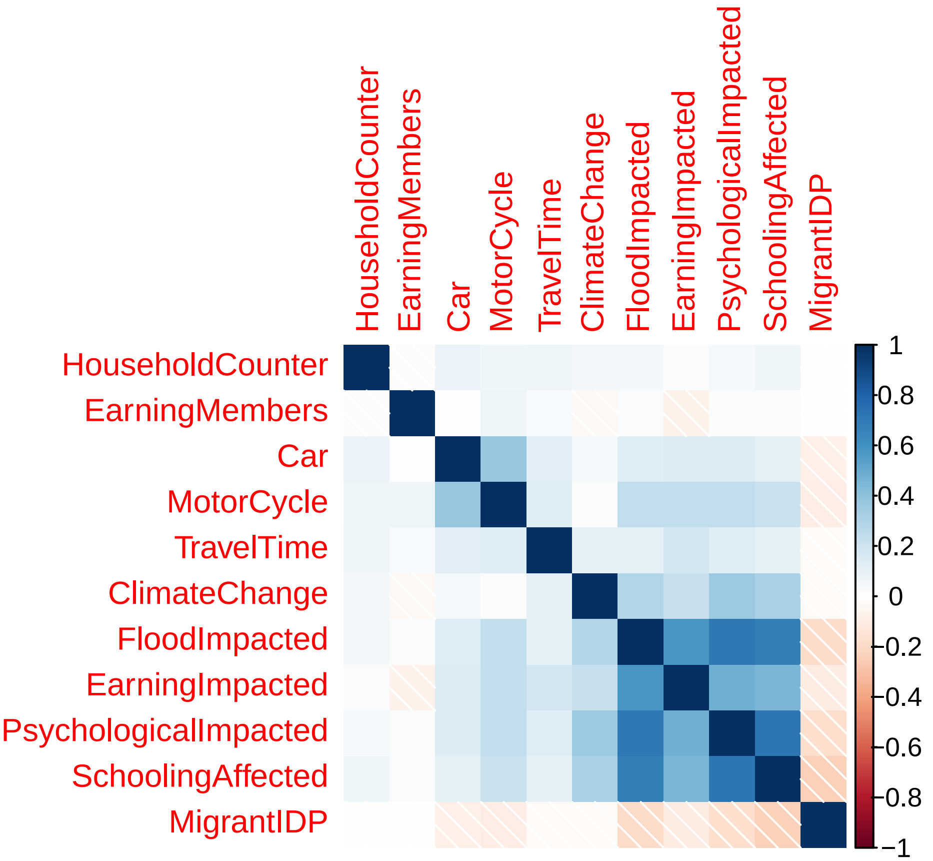
## STEP 1: Correlations for Strong Multicollinearity

The correlation matrix is used to identify variables with strong multicollinearity (r>0.9) to ensure that the components accurately represent the variance in the data without redundancy.

```
# PCA requires the removal of variables with strong multicollinearity (r>0.9) to ensure that the compon

aser_child_household_tib_excluded <- aser_child_household_tib[, -12] # Excluding the GeneralKnowledge v

corr_aser_pca <- cor(aser_child_household_tib_excluded)

corr_aser_pca
```

```
##                       HouseholdCounter EarningMembers          Car  MotorCycle
## HouseholdCounter           1.000000000   -0.017261914  0.085308876  0.06665192
## EarningMembers            -0.017261914    1.000000000  0.009557159  0.07543932
## Car                        0.085308876    0.009557159  1.000000000  0.38640268
## MotorCycle                 0.066651920    0.075439321  0.386402682  1.00000000
## TravelTime                 0.076883764    0.031267210  0.120566064  0.13990614
## ClimateChange              0.058505703   -0.039697296  0.042269546  0.01899966
## FloodImpacted              0.059498928    0.023669087  0.135088357  0.24800475
## EarningImpacted            0.027323823   -0.074970812  0.143746094  0.24003726
## PsychologicalImpacted      0.045746481    0.025524851  0.149841450  0.24096941
## SchoolingAffected          0.060334189    0.023126603  0.114129583  0.21400680
## MigrantIDP                -0.008076353   -0.007850098 -0.085987829 -0.09129039
##                         TravelTime ClimateChange FloodImpacted EarningImpacted
## HouseholdCounter       0.07688376    0.05850570    0.05949893      0.02732382
## EarningMembers         0.03126721   -0.03969730    0.02366909     -0.07497081
## Car                    0.12056606    0.04226955    0.13508836      0.14374609
## MotorCycle             0.13990614    0.01899966    0.24800475      0.24003726
## TravelTime             1.00000000    0.10666199    0.10251889      0.18390152
## ClimateChange          0.10666199    1.00000000    0.29932533      0.22076860
## FloodImpacted          0.10251889    0.29932533    1.00000000      0.58428918
## EarningImpacted        0.18390152    0.22076860    0.58428918      1.00000000
## PsychologicalImpacted  0.13357679    0.36542153    0.71669734      0.48547527
## SchoolingAffected      0.10252093    0.32677557    0.68446208      0.45134086
## MigrantIDP            -0.02415864   -0.02916086   -0.18589842     -0.10559782
##                       PsychologicalImpacted SchoolingAffected   MigrantIDP
## HouseholdCounter                 0.04574648        0.06033419 -0.008076353
## EarningMembers                   0.02552485        0.02312660 -0.007850098
## Car                              0.14984145        0.11412958 -0.085987829
## MotorCycle                       0.24096941        0.21400680 -0.091290388
## TravelTime                       0.13357679        0.10252093 -0.024158636
## ClimateChange                    0.36542153        0.32677557 -0.029160859
## FloodImpacted                    0.71669734        0.68446208 -0.185898417
## EarningImpacted                  0.48547527        0.45134086 -0.105597815
## PsychologicalImpacted            1.00000000        0.72618058 -0.175009170
## SchoolingAffected                0.72618058        1.00000000 -0.232119203
## MigrantIDP                      -0.17500917       -0.23211920  1.000000000
```

```
corrplot::corrplot(corr_aser_pca, method = "shade") #corrplot for visualizing the correlation matrix
```



```
# There are no variables with strong multicollinearity (r>0.9) in the dataset.The climate change factor
```

## STEP 2: Scale all the variables

The variables are scaled to ensure that the mean is zero and the values are standardized to a standard deviation of 1. This is done to ensure that all variables are on the same scale for the PCA analysis.

```
scaled_aser_child_household_tib <- aser_child_household_tib_excluded %>%
  mutate(across(where(is.numeric), ~ scale(.) %>% as.vector)) #scale all variables

glimpse(scaled_aser_child_household_tib)
```

```
## Rows: 2,367
## Columns: 11
## $ HouseholdCounter      <dbl> -0.89874246, -0.41033883, -0.41033883, -1.387146~
## $ EarningMembers        <dbl> -0.03889119, -0.03889119, -0.03889119, -0.038891~
## $ Car                   <dbl> 0.2251277, 0.2251277, 0.2251277, 0.2251277, 0.22~
## $ MotorCycle            <dbl> 0.1840658, 0.1840658, 0.1840658, 0.1840658, 0.18~
## $ TravelTime            <dbl> -0.7425363, 2.1176484, 2.1176484, -0.7425363, 2.~
## $ ClimateChange         <dbl> 1.2300809648, -1.2290420451, -1.2290420451, -1.2~
## $ FloodImpacted         <dbl> 0.6049237, 0.6049237, 0.6049237, 0.6049237, 0.60~
## $ EarningImpacted       <dbl> 0.67523329, 0.67523329, 0.67523329, 0.67523329, ~
## $ PsychologicalImpacted <dbl> 0.6569238, 0.6569238, 0.6569238, 0.6569238, 0.65~
## $ SchoolingAffected     <dbl> 0.5939522, -1.4073075, -1.4073075, -1.4073075, -~
## $ MigrantIDP            <dbl> -0.2033164, -0.2033164, -0.2033164, -0.2033164, ~
```

```
psych::describe(scaled_aser_child_household_tib) #make sure all means are 0, and sd is 1. This is gives
```

```
##                       vars    n mean sd median trimmed  mad   min   max range
## HouseholdCounter         1 2367    0  1   0.08   -0.01 1.21 -1.71  9.03 10.74
## EarningMembers           2 2367    0  1  -0.04   -0.04 0.00 -0.04 27.90 27.94
## Car                      3 2367    0  1   0.23   -0.09 0.00 -0.86  8.93  9.79
## MotorCycle               4 2367    0  1   0.18   -0.06 0.00 -1.57  7.18  8.75
## TravelTime               5 2367    0  1  -0.74   -0.17 0.00 -0.74  2.12  2.86
## ClimateChange            6 2367    0  1   0.00    0.00 1.82 -1.23  1.23  2.46
## FloodImpacted            7 2367    0  1   0.60    0.19 0.00 -2.14  0.60  2.75
## EarningImpacted          8 2367    0  1   0.68    0.19 0.00 -2.19  0.68  2.86
## PsychologicalImpacted    9 2367    0  1   0.66    0.18 0.00 -2.14  0.66  2.80
## SchoolingAffected       10 2367    0  1   0.59    0.22 0.00 -2.41  0.59  3.00
## MigrantIDP              11 2367    0  1  -0.20   -0.20 0.00 -0.20  4.92  5.12
##                        skew kurtosis   se
## HouseholdCounter       0.59     3.67 0.02
## EarningMembers        27.57   765.18 0.02
## Car                    4.91    37.27 0.02
## MotorCycle             0.50     2.94 0.02
## TravelTime             0.98    -0.35 0.02
## ClimateChange          0.00    -1.49 0.02
## FloodImpacted         -1.31     0.13 0.02
## EarningImpacted       -1.16    -0.17 0.02
## PsychologicalImpacted -1.12    -0.32 0.02
## SchoolingAffected     -1.44     0.60 0.02
## MigrantIDP             4.71    20.20 0.02
```

## STEP 3: Visualizing PCA

The PCA is visualized to understand the underlying data.

```
# This is just to understand the underlying data. The correlation between a variable and a principal co

library(factoextra) #extract and visualize the output of multivariate data analyses, including 'PCA'
library(FactoMineR) #multivariate exploratory data analysis
```

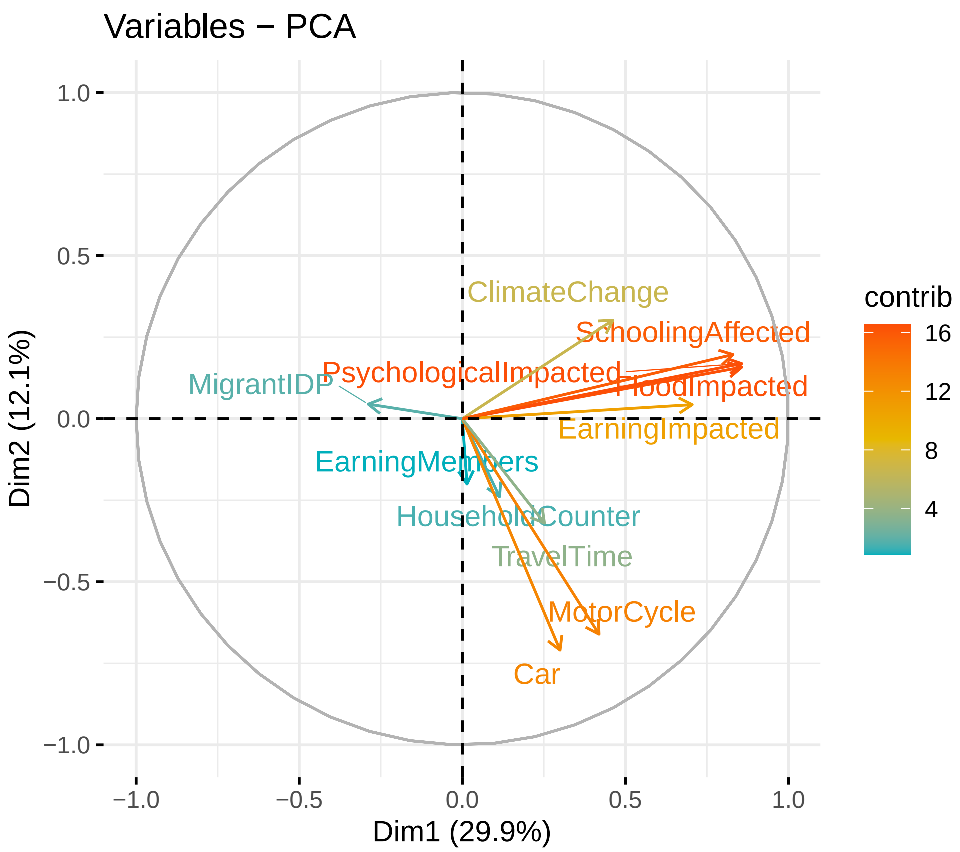
```
## Warning: package 'FactoMineR' was built under R version 4.3.3
```

```
aser_pca_eigen <- PCA(scaled_aser_child_household_tib, scale.unit = TRUE, graph = FALSE)
aser_pca_eigen$eig #eigenvalues
```

```
##         eigenvalue percentage of variance cumulative percentage of variance
## comp 1   3.2915171              29.922883                          29.92288
## comp 2   1.3269389              12.063081                          41.98596
## comp 3   1.0680099               9.709181                          51.69515
## comp 4   0.9984998               9.077271                          60.77242
## comp 5   0.9541435               8.674032                          69.44645
## comp 6   0.8960133               8.145575                          77.59202
## comp 7   0.7949933               7.227212                          84.81924
## comp 8   0.5810486               5.282260                          90.10150
## comp 9   0.5455105               4.959187                          95.06068
## comp 10  0.2869054              2.608231                          97.66891
## comp 11  0.2564195              2.331087                         100.00000
```

Kaiser rule suggests that we should keep components with eigenvalues greater than 1. Hence, it shows that we can keep **components 1, 2 and 3.**

```
#line below runs a simple PCA with a component for each variable.

viz_pca <- prcomp(scaled_aser_child_household_tib, center = TRUE,scale. = TRUE)

#Graph of variables. Positive correlated variables point to the same side of the plot. Negative correla

fviz_pca_var(viz_pca,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE #Avoid overlapping text if possible
             )
```



Through visualization, it seems that the following are loading together: **Dimension 1** (29.9%) and **Dimension 2** (12.1%)

1. **Climate Change Factors:** PsychologicalImpacted, SchoolingAffected, EarningImpacted, FloodImpacted, ClimateChange
2. **Household Factors:** MigrantIDP, HouseholdCounter, EarningMembers, Car, MotorCycle, TravelTime

## STEP 4: Bartlett's test

Bartlett's test checks if the data is suitable for PCA by determining whether the variables are related enough (correlated) to find meaningful patterns.

```
psych::cortest.bartlett(scaled_aser_child_household_tib, 2367) #there are 2367 observations
```

```
## R was not square, finding R from data

## $chisq
## [1] 6154.402
##
## $p.value
## [1] 0
##
## $df
## [1] 55
```

The p-value is very close to zero (it shows 0 in the result), so we reject the null hypothesis that the correlation matrix is an identity matrix, hence, PCA is justified because it will capture meaningful variance from the data.

## STEP 5: KMO Test

The Kaiser-Meyer-Olkin (KMO) test tells you whether the data is good enough for PCA or factor analysis by checking how well the variables are grouped together. A KMO value of 0.5 or higher is considered suitable for PCA.

```
psych::KMO(scaled_aser_child_household_tib)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: psych::KMO(r = scaled_aser_child_household_tib)
## Overall MSA =  0.81
## MSA for each item = 
##      HouseholdCounter        EarningMembers                   Car 
##                  0.69                  0.36                  0.65 
##            MotorCycle            TravelTime         ClimateChange 
##                  0.72                  0.75                  0.88 
##         FloodImpacted       EarningImpacted PsychologicalImpacted 
##                  0.81                  0.84                  0.81 
##     SchoolingAffected            MigrantIDP 
##                  0.83                  0.85 
```

```
#all data above .50 and overall MSA is strong (0.81) except for EarningMembers which is 0.36

# For reference: KMO > 0.8: Great for PCA! Variables are strongly related. KMO 0.7-0.8: Good, you can p

# We are going to remove EarningMembers because it has a KMO of 0.36, which is below the suitable thres

scaled_aser_child_household_tib <- scaled_aser_child_household_tib %>%
  select(-EarningMembers) #remove EarningMembers

KMO(scaled_aser_child_household_tib) #re-run KMO to make sure it is above .50, it is, hence, we can mov
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = scaled_aser_child_household_tib)
## Overall MSA =  0.81
## MSA for each item =
##      HouseholdCounter                   Car            MotorCycle
##                  0.71                  0.65                  0.73
##            TravelTime         ClimateChange         FloodImpacted
##                  0.76                  0.88                  0.81
##       EarningImpacted PsychologicalImpacted     SchoolingAffected
##                  0.86                  0.81                  0.83
##            MigrantIDP
##                  0.85
```

Through the KMO test, we found that the variable *EarningMembers* is not suitable for PCA as it has a KMO value of 0.36, which is below the suitable threshold of 0.5. Hence, we removed it from the dataset.

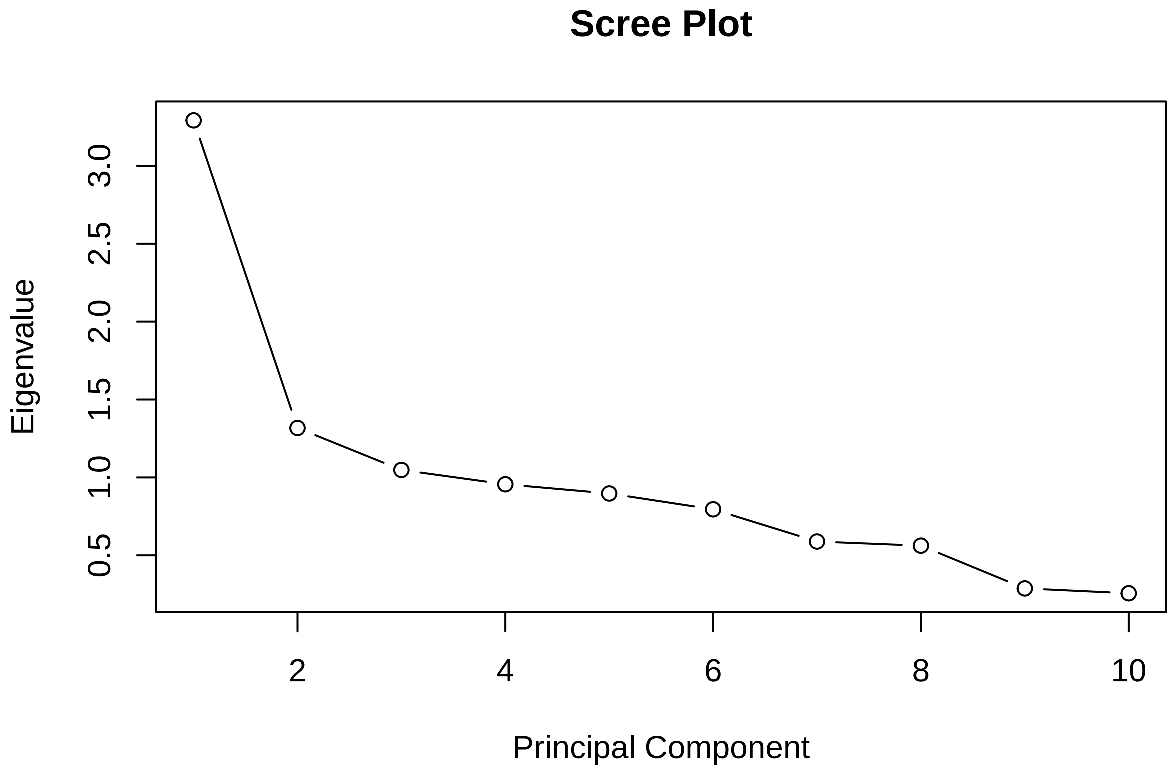
## STEP 6: Baseline PCA to select number of components

The baseline PCA is run to determine the number of components to keep in the final PCA analysis.

```r
#This is our initial PCA to see how many components we should keep.
pca_base <- principal(scaled_aser_child_household_tib, rotate = "none") #baseline PCA

pca_base
```

```
## Principal Components Analysis
## Call: principal(r = scaled_aser_child_household_tib, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                         PC1    h2   u2 com
## HouseholdCounter       0.11 0.013 0.99   1
## Car                    0.30 0.090 0.91   1
## MotorCycle             0.42 0.175 0.82   1
## TravelTime             0.25 0.063 0.94   1
## ClimateChange          0.46 0.214 0.79   1
## FloodImpacted          0.86 0.734 0.27   1
## EarningImpacted        0.71 0.498 0.50   1
## PsychologicalImpacted  0.86 0.734 0.27   1
## SchoolingAffected      0.83 0.688 0.31   1
## MigrantIDP            -0.29 0.083 0.92   1
##
##                 PC1
## SS loadings    3.29
## Proportion Var 0.33
##
## Mean item complexity =  1
## Test of the hypothesis that 1 component is sufficient.
##
## The root mean square of the residuals (RMSR) is  0.09
##  with the empirical chi square  1621.32  with prob <  9.999987e-319
##
## Fit based upon off diagonal values = 0.9
```

```
plot(pca_base$values, type = "b", xlab = "Principal Component", ylab = "Eigenvalue", main = "Scree Plot
```



```
#the plot shows the variance explained by 2-3 linear components. We will keep 3 components for now.
```

The elbow in the scree plot suggests that we should keep **3 components** for the final PCA analysis. This is because the variance explained by the components starts to level off after the third component. In other words, the elbow is bending sharply after the third component, indicating that the first three components capture most of the variance in the data.

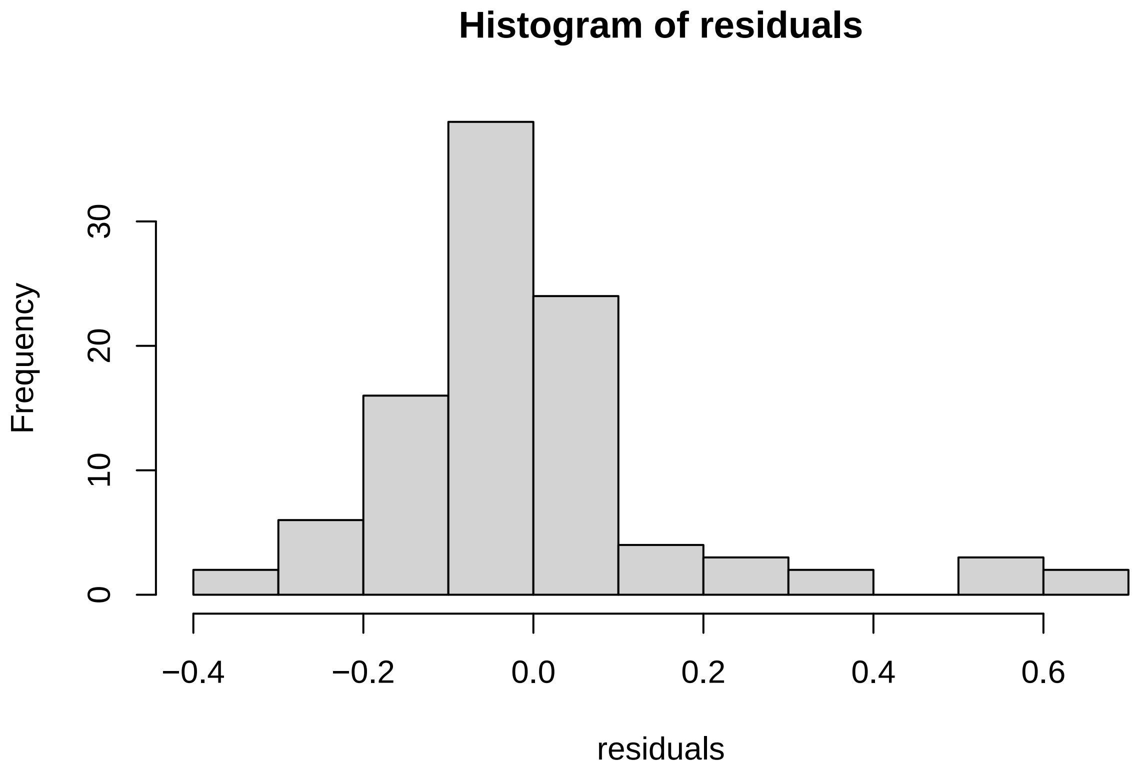
## STEP 7: Check that residuals are normally distributed

The residuals are checked to ensure that they are normally distributed.

```
pca_resid <- principal(scaled_aser_child_household_tib, nfactors = 3 , rotate = "none")
pca_resid #results.
```

```
## Principal Components Analysis
## Call: principal(r = scaled_aser_child_household_tib, nfactors = 3,
##     rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                          PC1   PC2   PC3   h2   u2 com
## HouseholdCounter        0.11  0.26  0.56 0.39 0.61 1.5
## Car                     0.30  0.73 -0.11 0.63 0.37 1.4
## MotorCycle             0.42  0.66 -0.17 0.64 0.36 1.9
## TravelTime             0.25  0.33  0.49 0.40 0.60 2.3
## ClimateChange          0.46 -0.29  0.40 0.45 0.55 2.7
## FloodImpacted          0.86 -0.16 -0.06 0.76 0.24 1.1
## EarningImpacted        0.71 -0.02  0.03 0.50 0.50 1.0
## PsychologicalImpacted  0.86 -0.18 -0.02 0.77 0.23 1.1
## SchoolingAffected      0.83 -0.21 -0.07 0.74 0.26 1.1
## MigrantIDP            -0.29 -0.04  0.54 0.37 0.63 1.5
##
##                        PC1  PC2  PC3
## SS loadings           3.29 1.32 1.05
## Proportion Var        0.33 0.13 0.10
## Cumulative Var        0.33 0.46 0.57
## Proportion Explained  0.58 0.23 0.19
## Cumulative Proportion 0.58 0.81 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.11
##  with the empirical chi square  2484.85  with prob <  0
##
## Fit based upon off diagonal values = 0.85
```

```r
#require correlation matrix for final data
corMatrix<-cor(scaled_aser_child_household_tib)

#next,create an object from the correlation matrix and the pca loading. Call it residuals. It will cont
residuals<-factor.residuals(corMatrix, pca_resid$loadings)

#call a histogram to check residuals
hist(residuals)
```



The residuals are somewhat normally distributed but exhibit positive skewness.

### STEP 8: Informed PCA with specific number of components

The final PCA is run with the specific number of components (3) to identify the underlying factors in the data.

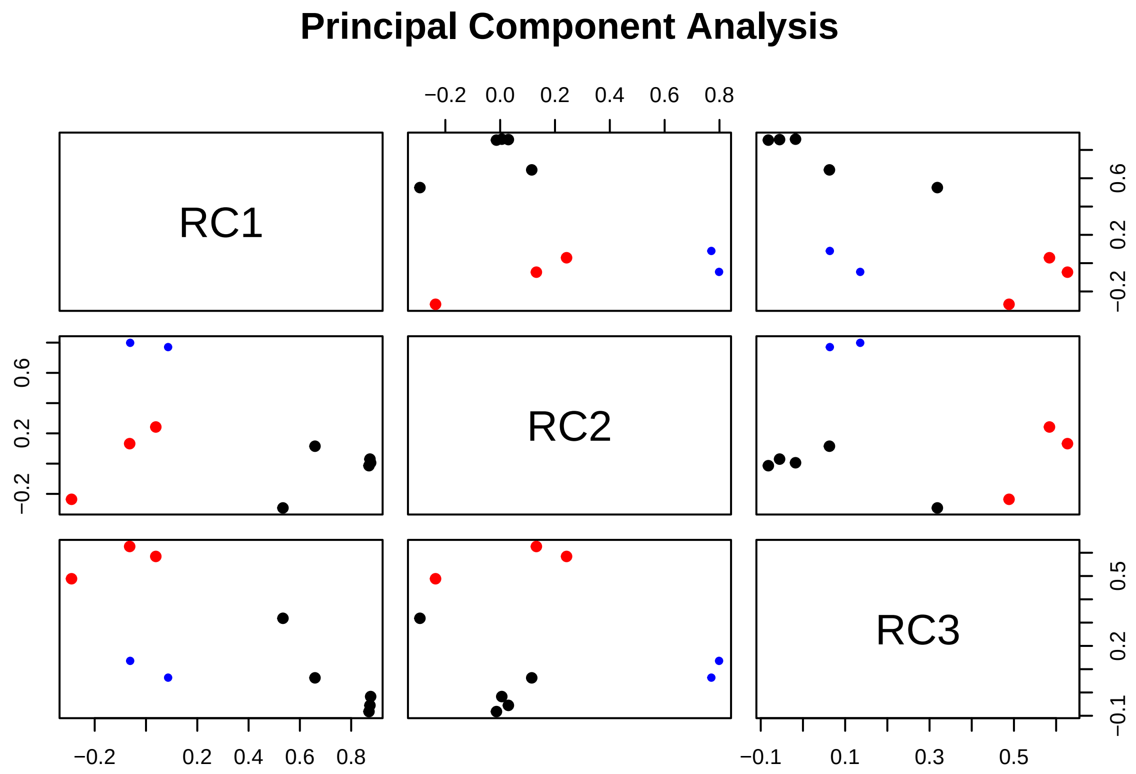
```
# rotation. Since factors should be related that's our assumption, use oblique technique (promax).
pca_final <- principal(scaled_aser_child_household_tib, nfactors = 3, rotate = "promax")
pca_final #results.
```

```
## Principal Components Analysis
## Call: principal(r = scaled_aser_child_household_tib, nfactors = 3,
##     rotate = "promax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                          RC1   RC2   RC3   h2   u2 com
## HouseholdCounter       -0.06  0.13  0.63 0.39 0.61 1.1
## Car                    -0.06  0.80  0.14 0.63 0.37 1.1
## MotorCycle              0.09  0.77  0.06 0.64 0.36 1.0
## TravelTime              0.04  0.24  0.58 0.40 0.60 1.3
## ClimateChange           0.53 -0.29  0.32 0.45 0.55 2.3
## FloodImpacted           0.87  0.03 -0.06 0.76 0.24 1.0
## EarningImpacted         0.66  0.12  0.06 0.50 0.50 1.1
## PsychologicalImpacted  0.88  0.01 -0.02 0.77 0.23 1.0
## SchoolingAffected       0.87 -0.01 -0.08 0.74 0.26 1.0
## MigrantIDP            -0.29 -0.24  0.49 0.37 0.63 2.1
##
##                         RC1  RC2  RC3
## SS loadings           3.10 1.46 1.09
## Proportion Var        0.31 0.15 0.11
## Cumulative Var        0.31 0.46 0.57
## Proportion Explained  0.55 0.26 0.19
## Cumulative Proportion 0.55 0.81 1.00
##
##  With component correlations of
##      RC1   RC2   RC3
## RC1 1.00  0.23  0.15
## RC2 0.23  1.00 -0.03
## RC3 0.15 -0.03  1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.11
##  with the empirical chi square  2484.85  with prob <  0
##
## Fit based upon off diagonal values = 0.85
```

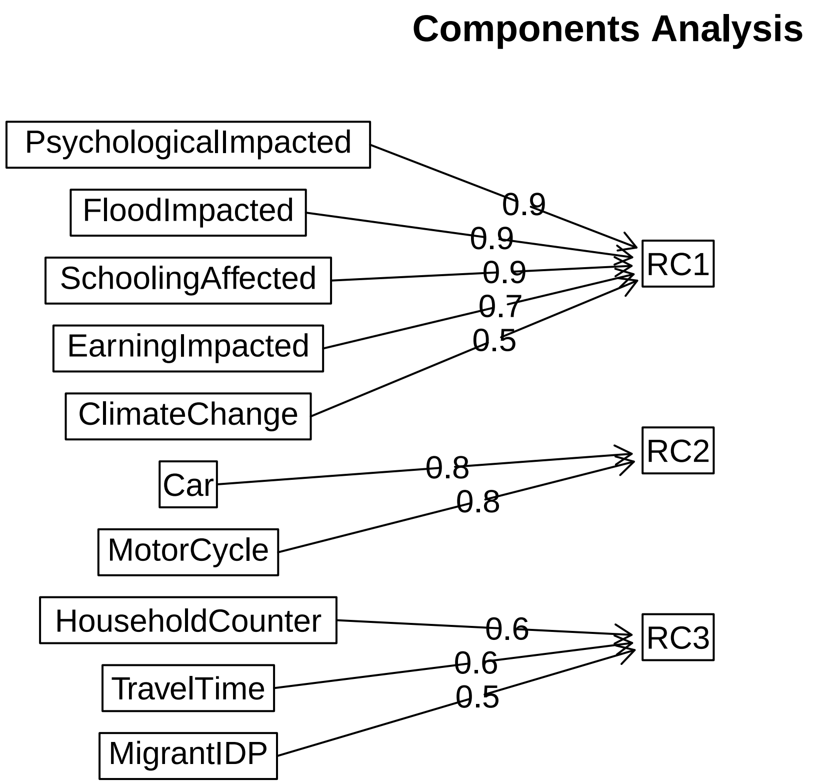
```
#RMSR is 0.11 and fit measure is 0.85. This is a good fit.

print.psych(pca_final, cut = 0.3, sort = TRUE) #print the results
```

```
## Principal Components Analysis
## Call: principal(r = scaled_aser_child_household_tib, nfactors = 3,
##     rotate = "promax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                       item   RC1   RC2   RC3    h2   u2 com
## PsychologicalImpacted    8  0.88              0.77 0.23 1.0
## FloodImpacted            6  0.87              0.76 0.24 1.0
## SchoolingAffected        9  0.87              0.74 0.26 1.0
## EarningImpacted          7  0.66              0.50 0.50 1.1
## ClimateChange            5  0.53        0.32 0.45 0.55 2.3
## Car                      2        0.80        0.63 0.37 1.1
## MotorCycle               3        0.77        0.64 0.36 1.0
## HouseholdCounter         1              0.63 0.39 0.61 1.1
## TravelTime               4              0.58 0.40 0.60 1.3
## MigrantIDP              10              0.49 0.37 0.63 2.1
##
##                         RC1  RC2  RC3
## SS loadings           3.10 1.46 1.09
## Proportion Var        0.31 0.15 0.11
## Cumulative Var        0.31 0.46 0.57
## Proportion Explained  0.55 0.26 0.19
## Cumulative Proportion 0.55 0.81 1.00
##
##  With component correlations of
##      RC1   RC2   RC3
## RC1 1.00  0.23  0.15
## RC2 0.23  1.00 -0.03
## RC3 0.15 -0.03  1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.11
##  with the empirical chi square  2484.85  with prob <  0
##
## Fit based upon off diagonal values = 0.85
```

```
plot(pca_final)
```



```
#component 1 is black
#component 2 is red
#component 3 is blue
#component 4 is grey in case of four components
fa.diagram(pca_final)
```



### Components

As the diagram shows, the components identified are as follows:

- Component 1: Climate Change Factors (PsychologicalImpacted, SchoolingAffected, EarningImpacted, FloodImpacted, ClimateChange)
- Component 2: Household Vehicles (Car, MotorCycle)
- Component 3: Displacement Factors (MigrantIDP, HouseholdCounter, TravelTime)

## STEP 9: Collect factor scores

The factor scores are collected for each text on each factor. These scores provide a way to understand how strongly each observation is associated with the patterns captured by the factors. The factor scores are then combined with the original dataset for further analysis. Similarly, we rename the columns for clarity.

```r
pca_final_scores <- as.data.frame(pca_final$scores) #scores for each text on each factor.
head(pca_final_scores)
```

```
##           RC1         RC2        RC3
## 1  0.89038732 -0.08160222 -0.5605457
## 2 -0.04729858  0.83997927  0.4401560
## 3 -0.04729858  0.83997927  0.4401560
## 4 -0.11551260  0.52664957 -1.5699130
## 5 -0.04813936  0.84346818  0.5314087
## 6  0.66493379  0.21989954 -0.9794028
```

```
#rename columns
pca_final_scores <- pca_final_scores %>%
  rename(climate_change_factors = RC1, household_vehicles = RC2, people_time = RC3)

#combine this dataframe with earlier dataframe

glimpse(aser_child_household_tib)
```

```
## Rows: 2,367
## Columns: 12
## $ HouseholdCounter      <dbl> 5, 8, 8, 2, 9, 5, 2, 1, 6, 6, 6, 18, 18, 18, 3, ~
## $ EarningMembers        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, ~
## $ Car                   <dbl> 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, ~
## $ MotorCycle            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, ~
## $ TravelTime            <dbl> 1, 3, 3, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ ClimateChange         <dbl> 3, 1, 1, 1, 1, 2, 2, 1, 3, 3, 3, 3, 3, 3, 3, 2, ~
## $ FloodImpacted         <dbl> 3, 3, 3, 3, 3, 3, 2, 3, 3, 3, 3, 3, 3, 3, 3, 2, ~
## $ EarningImpacted       <dbl> 5, 5, 5, 5, 5, 5, 4, 5, 5, 5, 5, 2, 2, 2, 5, 3, ~
## $ PsychologicalImpacted <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 2, ~
## $ SchoolingAffected     <dbl> 4, 2, 2, 2, 2, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 2, ~
## $ MigrantIDP            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ GeneralKnolwedge      <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, ~
```

```
final_aser_child_household_tib <- cbind(aser_child_household_tib, pca_final_scores)

glimpse(final_aser_child_household_tib)
```

```
## Rows: 2,367
## Columns: 15
## $ HouseholdCounter       <dbl> 5, 8, 8, 2, 9, 5, 2, 1, 6, 6, 6, 18, 18, 18, 3,~
## $ EarningMembers         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,~
## $ Car                    <dbl> 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0,~
## $ MotorCycle             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,~
## $ TravelTime             <dbl> 1, 3, 3, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ ClimateChange          <dbl> 3, 1, 1, 1, 1, 2, 2, 1, 3, 3, 3, 3, 3, 3, 3, 2,~
## $ FloodImpacted          <dbl> 3, 3, 3, 3, 3, 3, 2, 3, 3, 3, 3, 3, 3, 3, 3, 2,~
## $ EarningImpacted        <dbl> 5, 5, 5, 5, 5, 5, 4, 5, 5, 5, 5, 2, 2, 2, 5, 3,~
## $ PsychologicalImpacted  <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 2,~
## $ SchoolingAffected      <dbl> 4, 2, 2, 2, 2, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 2,~
## $ MigrantIDP             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,~
## $ GeneralKnolwedge       <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0,~
## $ climate_change_factors <dbl> 0.89038732, -0.04729858, -0.04729858, -0.115512~
## $ household_vehicles     <dbl> -0.08160222, 0.83997927, 0.83997927, 0.52664957~
## $ people_time            <dbl> -0.5605457, 0.4401560, 0.4401560, -1.5699130, 0~
```

## Step 10: Balancing the classes

The classes are balanced to ensure that the model is not biased towards the majority class. The classes are balanced by randomly sampling the data to ensure that the number of observations in each class is the same.

```
# First, checking, whether classes are balanced at the outcome level

table(final_aser_child_household_tib$GeneralKnolwedge) #shows the distribution of the dependent variabl
```

```
##
##    0    1
## 1192 1175
```

```
#this gives us sample sizes.
#IMPORTANTLY, tells us the baseline class
#Some_GK = 1. No_GK = 0. In our data, we have 1192 observations of No_GK and 1175 observations of Some_

#balance the factors
set.seed(123)
final_aser_child_household_tib_balanced <- final_aser_child_household_tib %>%
  group_by(GeneralKnolwedge) %>%
  sample_n(1175) %>%
  ungroup()

table(final_aser_child_household_tib_balanced$GeneralKnolwedge) #shows the distribution of the dependen
```

```
##
##    0    1
## 1175 1175
```

```
glimpse(final_aser_child_household_tib_balanced)
```

```
## Rows: 2,350
## Columns: 15
## $ HouseholdCounter       <dbl> 3, 17, 7, 10, 15, 9, 13, 19, 12, 10, 11, 12, 18~
## $ EarningMembers         <dbl> 2, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1,~
## $ Car                    <dbl> 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1,~
## $ MotorCycle             <dbl> 1, 0, 0, 1, 1, 2, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1,~
## $ TravelTime             <dbl> 1, 1, 2, 1, 2, 1, 1, 3, 1, 1, 2, 1, 2, 1, 3, 3,~
## $ ClimateChange          <dbl> 1, 1, 2, 1, 1, 3, 1, 3, 1, 3, 1, 3, 3, 3, 2, 1,~
## $ FloodImpacted          <dbl> 1, 1, 2, 2, 3, 3, 1, 3, 1, 3, 3, 3, 3, 2, 3, 1,~
## $ EarningImpacted        <dbl> 2, 2, 4, 2, 5, 5, 3, 5, 1, 5, 5, 5, 1, 3, 5, 3,~
## $ PsychologicalImpacted  <dbl> 4, 1, 2, 3, 4, 4, 1, 4, 4, 3, 3, 4, 4, 2, 3, 1,~
## $ SchoolingAffected      <dbl> 4, 1, 4, 3, 4, 4, 1, 4, 4, 4, 4, 4, 4, 3, 4, 2,~
## $ MigrantIDP             <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ GeneralKnolwedge       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ climate_change_factors <dbl> -0.75997649, -2.43690218, -0.37748475, -1.35181~
## $ household_vehicles     <dbl> -0.27308760, -1.12240523, -1.21478167, -0.73614~
## $ people_time            <dbl> -1.66326583, -0.20348888, -0.06742178, 1.413309~
```

```
final_aser_child_household_tib_balanced <- final_aser_child_household_tib_balanced %>%
  mutate(GeneralKnolwedge = as.factor(GeneralKnolwedge)) %>% #change GK Score to factor
  select(GeneralKnolwedge, climate_change_factors, household_vehicles, people_time) #remove other varia

levels(final_aser_child_household_tib_balanced$GeneralKnolwedge) #check levels of the factors of GK
```

```
## [1] "0" "1"
```

```r
table(final_aser_child_household_tib_balanced$GeneralKnolwedge) #check the distribution of the factors
```

```
##
##    0    1
## 1175 1175
```

```r
glimpse(final_aser_child_household_tib_balanced) #check the final dataset
```

```
## Rows: 2,350
## Columns: 4
## $ GeneralKnolwedge       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ climate_change_factors <dbl> -0.75997649, -2.43690218, -0.37748475, -1.35181~
## $ household_vehicles     <dbl> -0.27308760, -1.12240523, -1.21478167, -0.73614~
## $ people_time            <dbl> -1.66326583, -0.20348888, -0.06742178, 1.413309~
```

### Step 11: Correlation Matrix for the final dataset

The correlation matrix is used to check for multicollinearity between the components. The Variance Inflation Factor (VIF) is used to check for multicollinearity between the components. A VIF value greater than 5 indicates multicollinearity.

```r
final_cor_aser <- final_aser_child_household_tib_balanced[,2:4]  # Exclude outcome variable, GeneralKno
corr_aser_final <- cor(final_cor_aser)
corr_aser_final #all r<0.7, no munlticollinearity, ready for next step
```

```
##                        climate_change_factors household_vehicles people_time
## climate_change_factors              1.0000000         0.22966803  0.15293579
## household_vehicles                  0.2296680         1.00000000 -0.02449072
## people_time                         0.1529358        -0.02449072  1.00000000
```

```r
car::vif(lm(climate_change_factors ~ household_vehicles + people_time, data = final_aser_child_househol
```

```
## household_vehicles        people_time
##             1.0006             1.0006
```

```r
#overall low correlations between the components but no multicollinearity issue
```

The correlations between the factors are low, with no multicollinearity issues r<0.7. The VIF value is less than 5, indicating that there is no multicollinearity between the components.

### Step 12: Cross-validated logistic regression

```r
library(caret)
set.seed(123)

# Set up 10-fold cross-validation
train.control <- trainControl(method = "cv", number = 10, verbose = FALSE)
```

```
# method = cross validation, number = ten times (10 fold cross-validation)

# Logistic regression with stepwise selection
lr_cv10 <- train(GeneralKnolwedge ~ .,
                 data = final_aser_child_household_tib_balanced,
                 method = "glmStepAIC",
                 direction = "backward",
                 trControl = train.control,
                 family = "binomial",
                 verbose = FALSE
)
```

```
# Cross-validated model results
lr_cv10 # kappa is 0.16 and accuracy is 0.58
```

```
## Generalized Linear Model with Stepwise Feature Selection
##
## 2350 samples
##    3 predictor
##    2 classes: ’0’, ’1’
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2115, 2114, 2115, 2116, 2115, 2115, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.5723065  0.1446917
```

```
#print(lr_cv10$finalModel) # Final model)
summary(lr_cv10$finalModel) # Coefficients and model summary
```

```
##
## Call:
## NULL
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.003161   0.042346   0.075    0.941
## household_vehicles 0.498194   0.047941  10.392   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3257.8  on 2349  degrees of freedom
## Residual deviance: 3134.2  on 2348  degrees of freedom
## AIC: 3138.2
##
## Number of Fisher Scoring iterations: 4
```

The stepwise logistic regression model removed the two components, climate change factors, and the displacement factors, leaving only the household vehicles component as a significant predictor of general knowledge.

### Step 13: Confusion Matrix

```
#get predicted values
predicted <- unname(lr_cv10$finalModel$fitted.values) #change from a named number vector
#print(predicted)

#add predicted values to tibble

final_aser_child_household_tib_balanced$predicted.probabilities <- predicted


final_aser_child_household_tib_balanced <- final_aser_child_household_tib_balanced %>%
  mutate(actual = ifelse(GeneralKnolwedge == "1", 1, 0)) %>%
  #assign 0 to .50 and less and 1 to anything else
  mutate(predicted = ifelse(predicted.probabilities > 0.5, 1, 0)) #criteria is arbitary, but >0.5 means


#both need to be factors
final_aser_child_household_tib_balanced$predicted <- as.factor(final_aser_child_household_tib_balanced$p
final_aser_child_household_tib_balanced$actual <- as.factor(final_aser_child_household_tib_balanced$actu

glimpse(final_aser_child_household_tib_balanced)
```

```
## Rows: 2,350
## Columns: 7
## $ GeneralKnolwedge        <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ climate_change_factors  <dbl> -0.75997649, -2.43690218, -0.37748475, -1.3518~
## $ household_vehicles      <dbl> -0.27308760, -1.12240523, -1.21478167, -0.7361~
## $ people_time             <dbl> -1.66326583, -0.20348888, -0.06742178, 1.41330~
## $ predicted.probabilities <dbl> 0.4668263, 0.3644700, 0.3538779, 0.4100921, 0.~
## $ actual                  <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ predicted               <fct> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1~
```

```
table(final_aser_child_household_tib_balanced$actual) #what are final numbers
```

```
##
##    0    1
## 1175 1175
```

```
# create confusion matrix using CARET
confusionMatrix(final_aser_child_household_tib_balanced$actual, final_aser_child_household_tib_balanced$
                mode = "everything", #what you want to report in stats
                positive="1") #positive here is some general knowledge
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 642 533
##          1 452 723
##
```

```
##                Accuracy : 0.5809
##                  95% CI : (0.5606, 0.6009)
##     No Information Rate : 0.5345
##     P-Value [Acc > NIR] : 3.409e-06
##
##                   Kappa : 0.1617
##
##  Mcnemar's Test P-Value : 0.0108
##
##             Sensitivity : 0.5756
##             Specificity : 0.5868
##          Pos Pred Value : 0.6153
##          Neg Pred Value : 0.5464
##               Precision : 0.6153
##                  Recall : 0.5756
##                      F1 : 0.5948
##              Prevalence : 0.5345
##          Detection Rate : 0.3077
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.5812
##
##        'Positive' Class : 1
##
```

The low Kappa value (0.16) suggests that there is only slight agreement between predicted and actual values, adjusted for chance. Similarly, the low accuracy (0.58) suggests that the model is not highly accurate in predicting whether a student exhibits general knowledge or not.

## Step 14: Final Model Interpretation

```
final_aser_model <- lr_cv10$finalModel
summary(final_aser_model) # household_vehicles is the only significant variable
```

```
##
## Call:
## NULL
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.003161   0.042346   0.075    0.941
## household_vehicles 0.498194   0.047941  10.392   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3257.8  on 2349  degrees of freedom
## Residual deviance: 3134.2  on 2348  degrees of freedom
## AIC: 3138.2
##
## Number of Fisher Scoring iterations: 4
```

household_vehicles is the only significant variable. We are going to compute probabilities to understand the effect of household vehicles on the likelihood of a student exhibiting general knowledge.

```
exp(final_aser_model$coefficients)
```

```
##        (Intercept) household_vehicles 
##           1.003166           1.645747
```

```
#create function for computing probabilities
probabilities <- function(coef) {
  odds <- exp(coef)
  prob <- odds / (1 + odds)
  return(prob)
}

#compute probabilities
probabilities(final_aser_model$coefficients)
```

```
##        (Intercept) household_vehicles 
##          0.5007902          0.6220349
```

**Interpretation:**

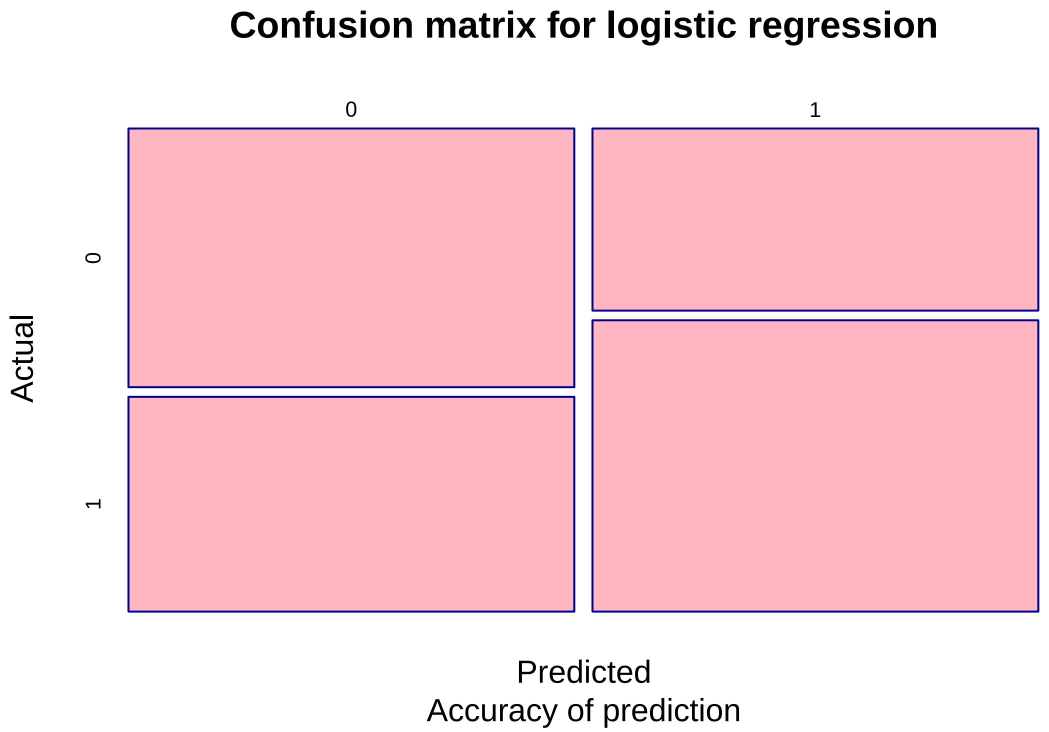
The probability of having general knowledge for students from households with no vehicles is approximately 50%. For each additional vehicle in a household, the probability of student possessing general knowledge increases to about 62%. This suggests that students whose households have one or more vehicles are associated with a higher likelihood of exhibiting general knowledge compared to those students who come from households without any vehicles.

## Step 15: Mosaic Plot

```
#put the actual and predicted values into a table
mosaic_table <- table(final_aser_child_household_tib_balanced$actual, final_aser_child_household_tib_bal
mosaic_table #check on that table
```

```
##      
##         0   1
##   0 642 533
##   1 452 723
```

```
#simple mosaic plot
mosaicplot(mosaic_table,
           main = "Confusion matrix for logistic regression",
           sub = "Accuracy of prediction",
           xlab = "Predicted",
           ylab = "Actual",
           color = "lightpink",
           border = "darkblue")
```



## Conclusion

The PCA analysis identified three components in the data: climate change factors, household vehicles, and displacement factors. The logistic regression model showed that only the household vehicles component was a significant predictor of general knowledge. The model had a low accuracy of 0.58 and a Kappa value of 0.16, indicating slight agreement between predicted and actual values. The probability of a student exhibiting general knowledge increased from 50% to 62% for each additional vehicle in the household. The mosaic plot visualizes the confusion matrix for the logistic regression model.

Since the household vehicles was a significant factor in determining general knowledge, it means that there is a potential association between access to vehicles and educational outcomes reflecting the impact of socioeconomic factors on a student's educational outcome.