A Cluster Analysis of Rural Pakistani Households: Socioeconomic, Climate Change, and Academic Achievement Factors

Ali Abid

2024-12-06

```
knitr::opts_chunk$set(echo = TRUE)
```

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 primary objective of this research project is to classify rural households across 151 surveyed districts of Pakistan into distinct clusters based on socioeconomic, climate change, and academic achievement factors. This clustering approach is instrumental in uncovering patterns and relationships within the data that may not be immediately apparent. Hence, it will help us first, discarding the apparent binary of rich versus poor households, and second, understanding the nuanced differences between households that may have been impacted by climate change and those that have not, and how it influences their decision-making for schooling.

By employing this technique, policymakers can gain nuanced insights into how rural households perceive climate change information, how climate change impacts influence educational enrollment decisions, and how internet availability correlates with socioeconomic status. The resulting clusters will provide a detailed understanding of various household types and their characteristics, informing policy decisions and interventions aimed at enhancing educational outcomes and addressing the challenges faced by rural households in Pakistan.

Data

The two datasets (*child*, and *household*) utilized in this project are from the ASER 2023 rural dataset, a household survey conducted in Pakistan to assess the learning levels of children and the current socioeconomic status of the household that they belong to. The survey encompasses a stratified population of 89,551 rural households, across 4,381 villages in 151 rural districts in Pakistan.

Data Description

- **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
- **IsInternetAvailable**: Binary variable for understanding whether the household has internet facility or not.
    - 1. Yes 0. No
- **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
- **Institution Type**: The type of institution that the student is enrolled in.
    - 1. Government
    - 2. Private
    - 3. Madrassah
    - 4. NFE (Non-Formal Education) / Other
- **LocalLangReadingLevel**: The local language reading level of the student tested
    - 1. Beginner/Nothing
    - 2. Letters
    - 3. Words
    - 4. Sentences
    - 5. Story
- **ArithmeticLevel**: The arithmetic level of the student tested
    - 1. Beginner/Nothing
    - 2. Recognition of 1-9
    - 3. Recognition of 10-99
    - 4. Recognition of 100-200
    - 5. Subtraction 2-digit
    - 6. Subtraction 4-digit
    - 7. Division
- **EnglishReadingLevel**: The English reading level of the student tested
    - 1. Beginner/Nothing
    - 2. Capital Letters
    - 3. Small letters
    - 4. Words

- 5. Sentences

Research Questions

- What distinct clusters of rural households can be identified in Pakistan based on socioeconomic, climate change, and academic achievement factors?
- How do these clusters differ in terms of household access to Internet and the types of institutions that the students are enrolled in?

Analysis

Required Libraries

Loading the required packages

```
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 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become e
rrors
```

```
library(readr)
```

Load the data

Loading the two datasets into RStudio for further analysis.

```
# 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.
```

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

```
## 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

Merging the two datasets and selecting the required columns for further analysis. The variables are renamed for additional clarity.

```
# 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(
    #Identifier
    Id,

    # Household Characteristics
    HouseholdCounter, EarningMembers, Car, MotorCycle,

    # Time to school
    TravelTime,

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

    # Socioeconomic Factor
    IsInternetAvailable,

    # School Characteristics
    C11,

    # Child Characteristics
    C15, C19, C20
  ) %>%
  rename(

    # Renaming Variables for Readability
    InstitutionType = C11, LocalLangReadingLevel = C15, ArithmeticLevel = C19, EnglishReadingLevel
= C20

  )
```

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(1:16),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) # Loading the psych package
```

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

```
##
## Attaching package: 'psych'
```

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

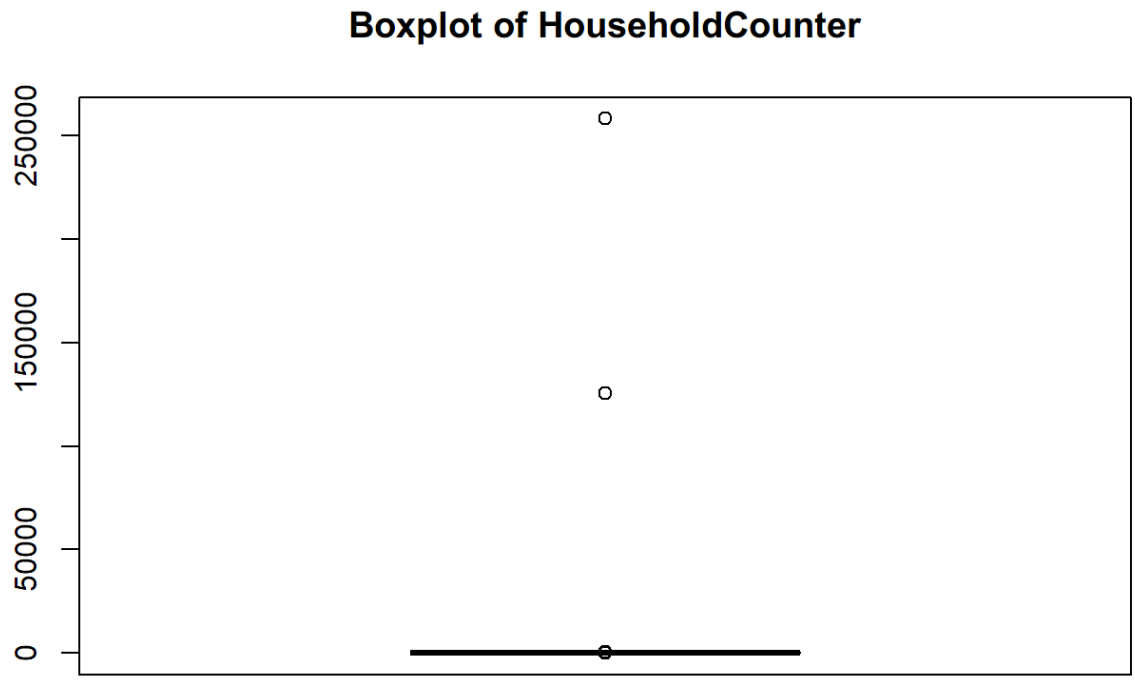
```
psych::describe(aser_child_household_tib) #gives you a lot of descriptives quickly
```

```
##                       vars    n       mean       sd  median    trimmed      mad
## Id                       1 2474 1565984.63 24563.87 1562075 1564782.76 32230.24
## HouseholdCounter         2 2474     166.22  5768.51      11      10.56     7.41
## EarningMembers           3 2474      20.90   525.12       1       0.96     1.48
## Car                      4 2474       0.73     0.82       1       0.67     0.00
## MotorCycle               5 2474       0.89     0.56       1       0.87     0.00
## TravelTime               6 2474       1.49     0.68       1       1.36     0.00
## ClimateChange            7 2474       1.98     0.81       2       1.97     1.48
## FloodImpacted            8 2474       2.53     0.75       3       2.67     0.00
## EarningImpacted          9 2474       4.04     1.37       5       4.29     0.00
## PsychologicalImpacted   10 2474       3.29     1.08       4       3.49     0.00
## SchoolingAffected       11 2474       3.39     1.01       4       3.60     0.00
## IsInternetAvailable     12 2474       0.34     0.47       0       0.30     0.00
## InstitutionType         13 2474       1.42     0.60       1       1.35     0.00
## LocalLangReadingLevel   14 2474       3.42     1.37       4       3.53     1.48
## ArithmeticLevel         15 2474       4.66     2.02       5       4.79     2.97
## EnglishReadingLevel     16 2474       3.61     1.38       4       3.76     1.48
##                           min     max  range  skew kurtosis     se
## Id                    1532168 1617421  85253  0.37    -1.16 493.85
## HouseholdCounter            0  258096 258096 40.28  1702.68 115.97
## EarningMembers              0   15000  15000 28.19   800.02  10.56
## Car                         0       9      9  4.72    40.83   0.02
## MotorCycle                  0       5      5  0.50     3.47   0.01
## TravelTime                  1       3      2  1.06    -0.16   0.01
## ClimateChange               1       3      2  0.04    -1.46   0.02
## FloodImpacted               1       3      2 -1.23    -0.13   0.02
## EarningImpacted             1       5      4 -1.13    -0.17   0.03
## PsychologicalImpacted       1       4      3 -1.12    -0.34   0.02
## SchoolingAffected           1       4      3 -1.36     0.35   0.02
## IsInternetAvailable         0       1      1  0.68    -1.53   0.01
## InstitutionType             1       4      3  1.56     3.42   0.01
## LocalLangReadingLevel       1       5      4 -0.28    -1.23   0.03
## ArithmeticLevel             1       7      6 -0.25    -1.28   0.04
## EnglishReadingLevel         1       5      4 -0.55    -1.03   0.03
```

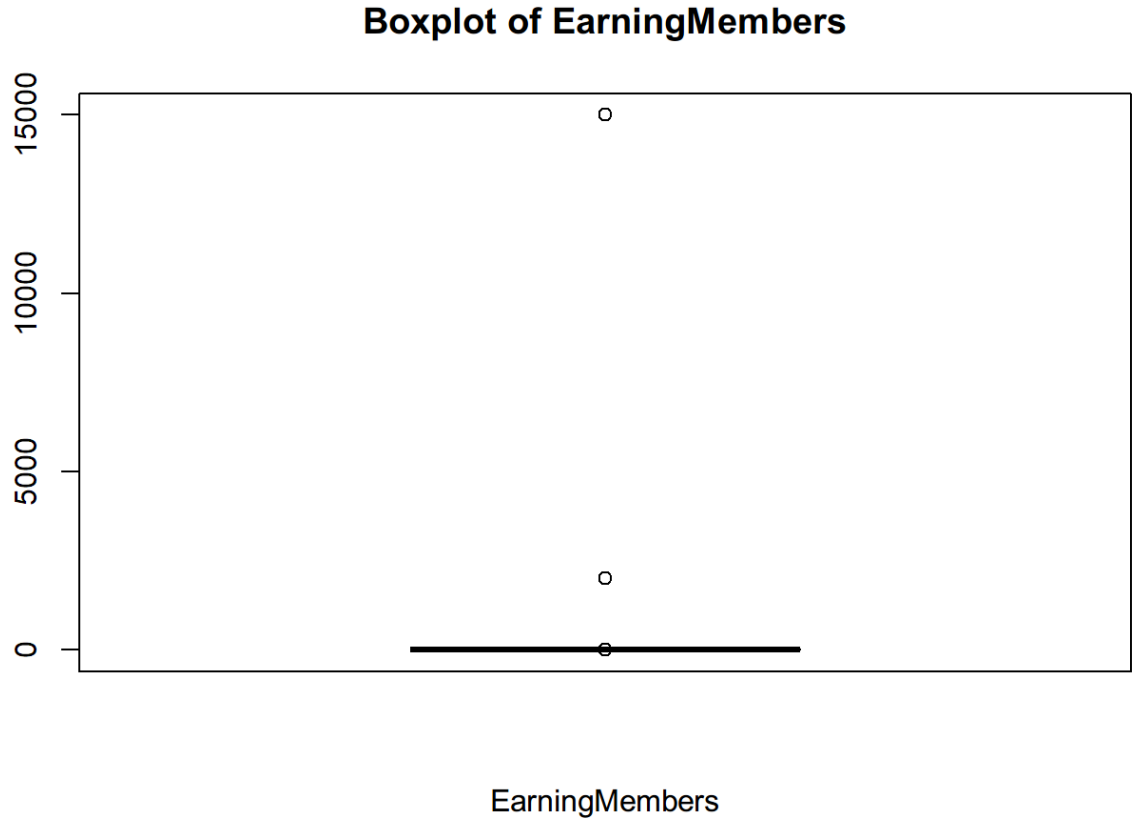
```
# There seems to be an outlier values in HouseholdCounter and EarningMembers, which needs investig
ation.
```

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

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




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



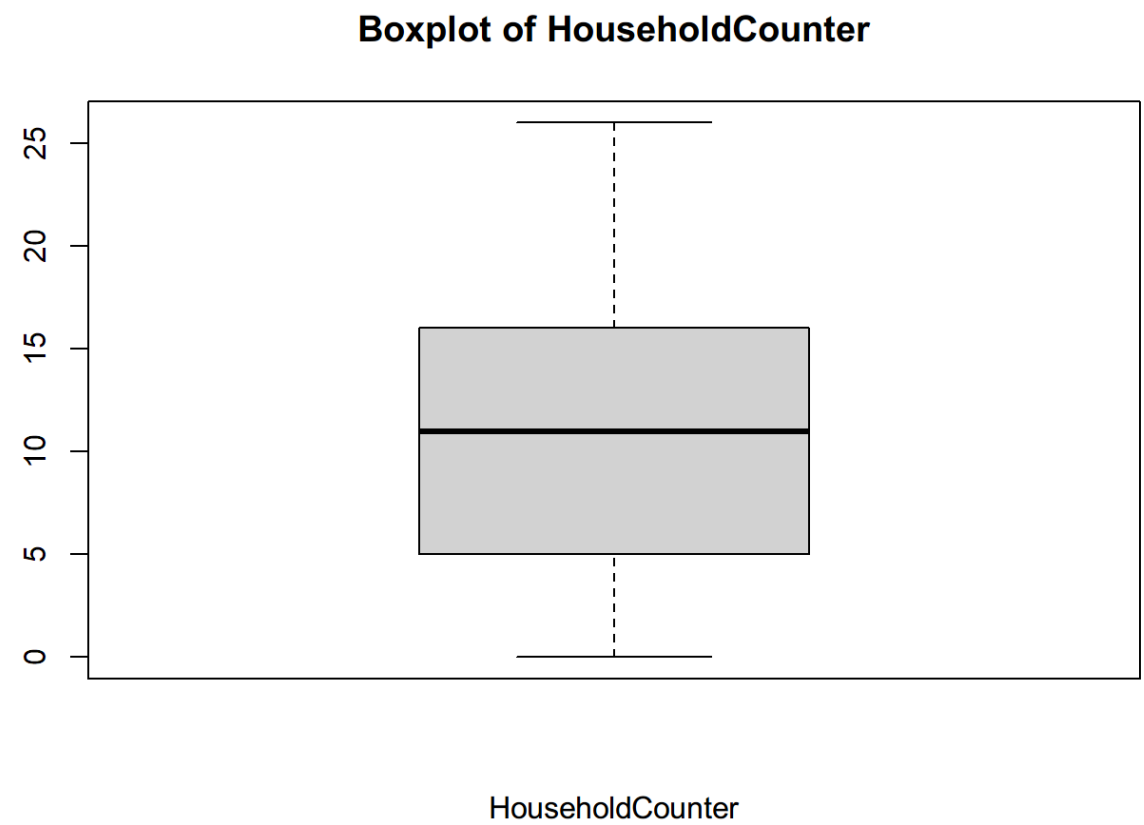
```
# Let's just try to find out which is the outlier value

max(aser_child_household_tib$HouseholdCounter)
```

```
## [1] 258096
```

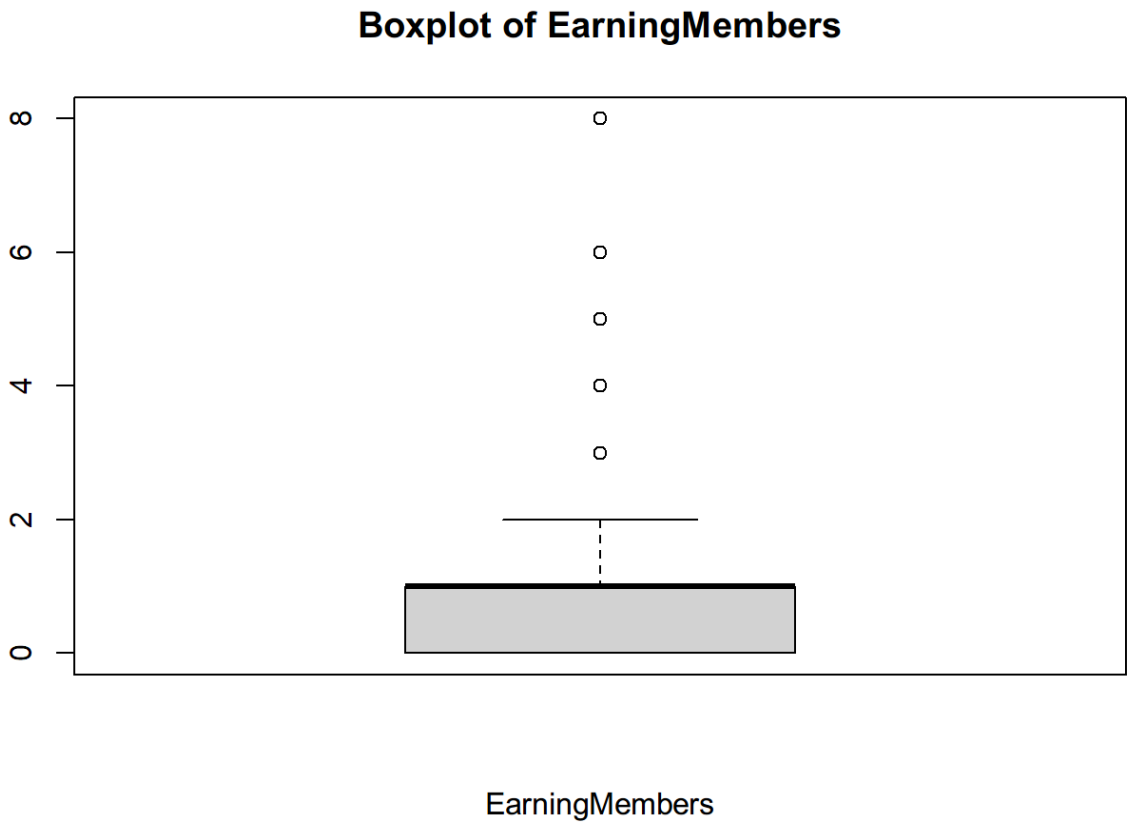
```
max(aser_child_household_tib$EarningMembers)
```

```
## [1] 15000
```

```
# Fairly assuming that the household and earning members cannot be more than 30, we will remove th
e outlier value by filtering out using this criteria.

aser_child_household_tib <- aser_child_household_tib %>%
  filter(HouseholdCounter < 30) %>% # Removing the outlier value
  filter(EarningMembers < 30) # Removing the outlier value

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

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




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



Final check for NA values in the dataset.

```
aser_child_household_tib %>%
  summarize(total_na = sum(across(everything(), is.na))) # 0 NA values found in the dataset
```

```
## # A tibble: 1 × 1
##   total_na
##      <int>
## 1        0
```

```
# Making a new tibble with the columns that will be used for clustering analysis.

aser_child_household_data_excluded <- aser_child_household_tib %>%
  select(-c(Id, IsInternetAvailable, InstitutionType)) # Excluding the columns for further analysis, we will get back to them later
```

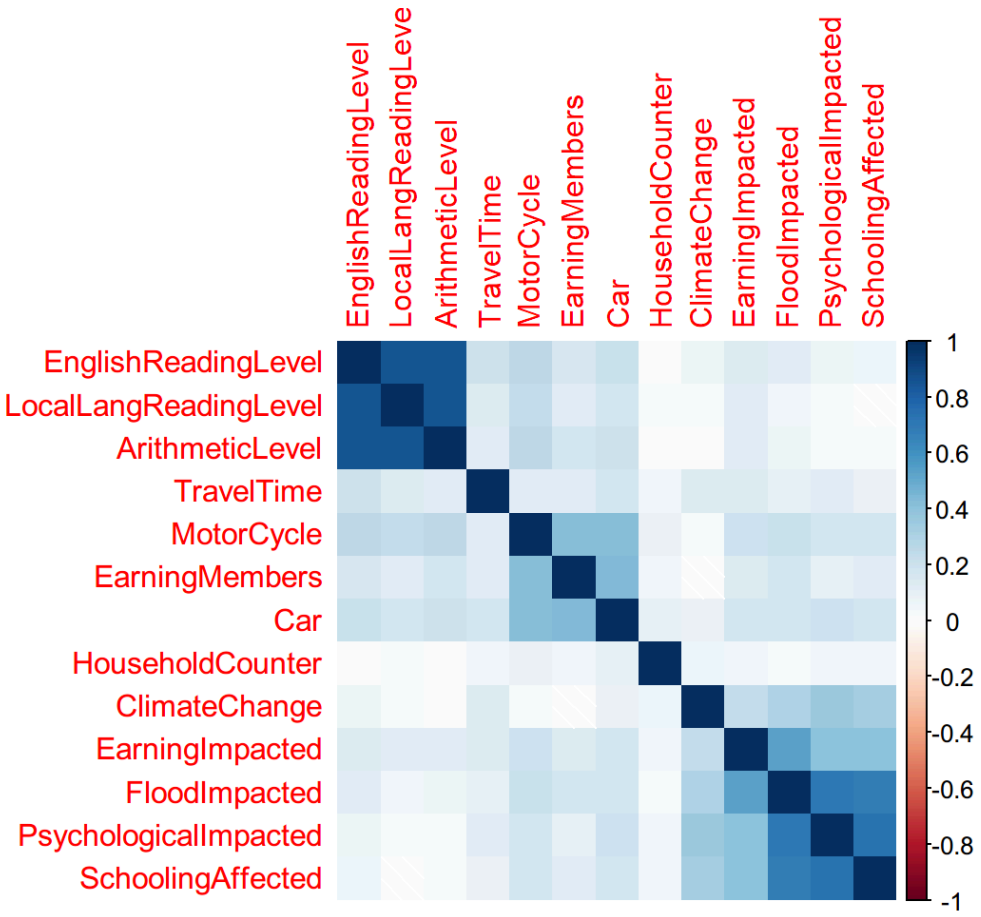
Multicollinearity Check

Checking for multicollinearity between the variables using the correlation matrix and the correlation plot.

```
# Checking for multicollinearity between the variables using the correlation matrix and the correlation plot. Setting the threshold for multicollinearity at r greater than .899

corr_aser_cluster <- round(cor(aser_child_household_data_excluded), 2)

corr_aser_cluster
```

```
##                       HouseholdCounter EarningMembers  Car MotorCycle
## HouseholdCounter                  1.00           0.04 0.10       0.08
## EarningMembers                    0.04           1.00 0.43       0.41
## Car                               0.10           0.43 1.00       0.41
## MotorCycle                        0.08           0.41 0.41       1.00
## TravelTime                        0.05           0.13 0.17       0.13
## ClimateChange                     0.06          -0.01 0.08       0.02
## FloodImpacted                     0.03           0.17 0.17       0.21
## EarningImpacted                   0.04           0.14 0.17       0.19
## PsychologicalImpacted             0.05           0.09 0.19       0.18
## SchoolingAffected                 0.05           0.12 0.17       0.17
## LocalLangReadingLevel             0.02           0.13 0.17       0.23
## ArithmeticLevel                   0.01           0.18 0.20       0.25
## EnglishReadingLevel               0.01           0.16 0.21       0.25
##                       TravelTime ClimateChange FloodImpacted EarningImpacted
## HouseholdCounter            0.05          0.06          0.03            0.04
## EarningMembers              0.13         -0.01          0.17            0.14
## Car                         0.17          0.08          0.17            0.17
## MotorCycle                  0.13          0.02          0.21            0.19
## TravelTime                  1.00          0.14          0.10            0.15
## ClimateChange               0.14          1.00          0.30            0.23
## FloodImpacted               0.10          0.30          1.00            0.53
## EarningImpacted             0.15          0.23          0.53            1.00
## PsychologicalImpacted       0.11          0.36          0.70            0.40
## SchoolingAffected           0.08          0.33          0.68            0.40
## LocalLangReadingLevel       0.14          0.02          0.04            0.11
## ArithmeticLevel             0.12          0.00          0.07            0.11
## EnglishReadingLevel         0.20          0.07          0.12            0.15
##                       PsychologicalImpacted SchoolingAffected
## HouseholdCounter                       0.05              0.05
## EarningMembers                         0.09              0.12
## Car                                    0.19              0.17
## MotorCycle                             0.18              0.17
## TravelTime                             0.11              0.08
## ClimateChange                          0.36              0.33
## FloodImpacted                          0.70              0.68
## EarningImpacted                        0.40              0.40
## PsychologicalImpacted                  1.00              0.73
## SchoolingAffected                      0.73              1.00
## LocalLangReadingLevel                  0.02             -0.02
## ArithmeticLevel                        0.03              0.02
## EnglishReadingLevel                    0.07              0.06
##                       LocalLangReadingLevel ArithmeticLevel EnglishReadingLevel
## HouseholdCounter                       0.02            0.01                0.01
## EarningMembers                         0.13            0.18                0.16
## Car                                    0.17            0.20                0.21
## MotorCycle                             0.23            0.25                0.25
## TravelTime                             0.14            0.12                0.20
## ClimateChange                          0.02            0.00                0.07
## FloodImpacted                          0.04            0.07                0.12
## EarningImpacted                        0.11            0.11                0.15
## PsychologicalImpacted                  0.02            0.03                0.07
## SchoolingAffected                     -0.02            0.02                0.06
## LocalLangReadingLevel                  1.00            0.85                0.85
## ArithmeticLevel                        0.85            1.00                0.85
## EnglishReadingLevel                    0.85            0.85                1.00
```

```
corrplot::corrplot(corr_aser_cluster, method = "shade", order = "hclust")
```




```
# Setting r > .899 as the threshold for multicollinearity. Hence, not removing any variables as th
ey do not meet the threshold.
```

The variables `LocalLangReadingLevel` , `EnglishReadingLevel` and `ArithmeticLevel` are highly correlated but not meeting the threshold r greater than .899 for removal.

Scale the data

Scaling the variables is important for cluster analysis to ensure that all variables are on the same scale meaning the mean is 0 and the standard deviation is 1.

```
#scale the variables (important for cluster analysis)

aser_child_household_tib_scaled <- aser_child_household_data_excluded %>%
  mutate_at(c(1:13), ~(scale(.) %>% as.vector))

glimpse(aser_child_household_tib_scaled)
```

```
## Rows: 2,454
## Columns: 13
## $ HouseholdCounter      <dbl> -0.9280995, -1.2648417, -1.6015840, 1.4290962, -…
## $ EarningMembers        <dbl> -0.08788457, -0.08788457, -0.08788457, -1.029669…
## $ Car                   <dbl> 0.3272828, 0.3272828, 0.3272828, -0.8896143, 0.3…
## $ MotorCycle            <dbl> 0.1928229, 0.1928229, 0.1928229, -1.6132358, 0.1…
## $ TravelTime            <dbl> -0.7133510, -0.7133510, 0.7515556, 0.7515556, -0…
## $ ClimateChange         <dbl> 1.26933991, 1.26933991, 0.02831993, 1.26933991, …
## $ FloodImpacted         <dbl> 0.6232690, 0.6232690, -2.0344238, -2.0344238, 0.…
## $ EarningImpacted       <dbl> 0.70141091, 0.70141091, -0.02979656, -0.76100402…
## $ PsychologicalImpacted <dbl> 0.6564006, 0.6564006, -1.1951018, -1.1951018, 0.…
## $ SchoolingAffected     <dbl> 0.6062312, 0.6062312, -0.3875506, -1.3813324, 0.…
## $ LocalLangReadingLevel <dbl> 0.4213838, -1.7710031, -1.7710031, -1.0402075, -…
## $ ArithmeticLevel       <dbl> 1.1567245, -1.8114895, -0.8220848, -1.3167871, -…
## $ EnglishReadingLevel   <dbl> 1.008715, -1.890713, -1.890713, -1.165856, -0.44…
```

```
psych::describe(aser_child_household_tib_scaled) # All variables are now scaled with mean 0 and SD 1
```

```
##                       vars    n mean sd median trimmed  mad   min   max range
## HouseholdCounter         1 2454    0  1   0.08    0.00 1.25 -1.77  2.61  4.38
## EarningMembers           2 2454    0  1  -0.09   -0.13 1.40 -1.03  6.50  7.53
## Car                      3 2454    0  1   0.33   -0.07 0.00 -0.89 10.06 10.95
## MotorCycle               4 2454    0  1   0.19   -0.03 0.00 -1.61  7.42  9.03
## TravelTime               5 2454    0  1  -0.71   -0.19 0.00 -0.71  2.22  2.93
## ClimateChange            6 2454    0  1   0.03   -0.01 1.84 -1.21  1.27  2.48
## FloodImpacted            7 2454    0  1   0.62    0.18 0.00 -2.03  0.62  2.66
## EarningImpacted          8 2454    0  1   0.70    0.18 0.00 -2.22  0.70  2.92
## PsychologicalImpacted    9 2454    0  1   0.66    0.18 0.00 -2.12  0.66  2.78
## SchoolingAffected       10 2454    0  1   0.61    0.21 0.00 -2.38  0.61  2.98
## LocalLangReadingLevel   11 2454    0  1   0.42    0.08 1.08 -1.77  1.15  2.92
## ArithmeticLevel         12 2454    0  1   0.17    0.06 1.47 -1.81  1.16  2.97
## EnglishReadingLevel     13 2454    0  1   0.28    0.11 1.07 -1.89  1.01  2.90
##                        skew kurtosis   se
## HouseholdCounter      -0.01    -1.24 0.02
## EarningMembers         2.18     8.89 0.02
## Car                    4.72    40.70 0.02
## MotorCycle             0.50     3.56 0.02
## TravelTime             1.07    -0.15 0.02
## ClimateChange          0.04    -1.46 0.02
## FloodImpacted         -1.22    -0.14 0.02
## EarningImpacted       -1.13    -0.14 0.02
## PsychologicalImpacted -1.12    -0.34 0.02
## SchoolingAffected     -1.35     0.34 0.02
## LocalLangReadingLevel -0.28    -1.23 0.02
## ArithmeticLevel       -0.25    -1.28 0.02
## EnglishReadingLevel   -0.55    -1.03 0.02
```

Distance Matrix

The distance matrix is important in cluster analysis because it helps us measure how similar or different data points are from each other.

The distance matrix is calculated using the `daisy` function from the `cluster` package. The dissimilarity matrix is visualized using the `fviz_dist` function from the `factoextra` package.

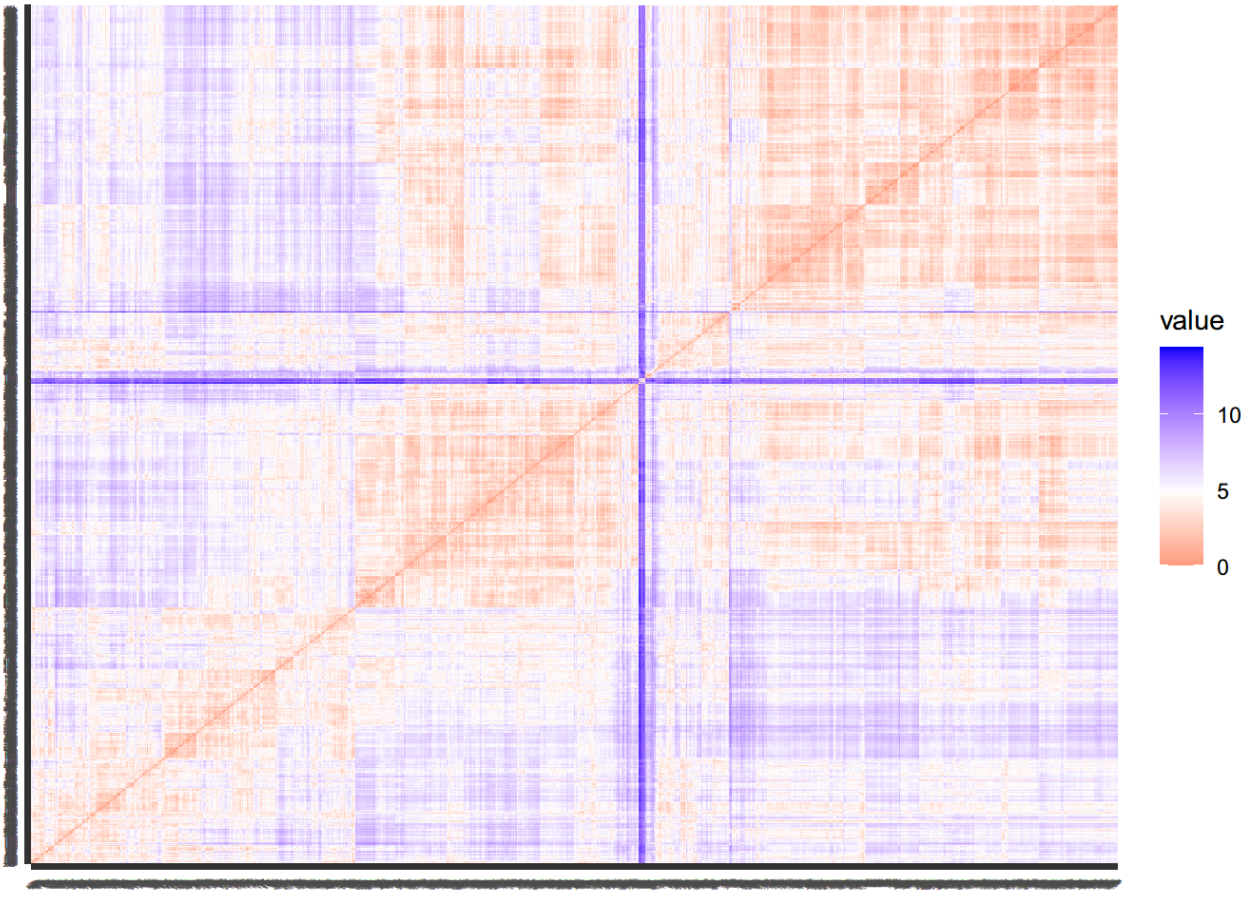
```
library(cluster)
library(factoextra)
```

```
## 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
```

```
#Using the daisy function to calculate the dissimilarity matrix

data_daisy <- daisy(aser_child_household_tib_scaled, metric = "euclidean")

round(as.matrix(data_daisy), 2) #rounding to 2 decimal places

#eucl_dist <- get_dist(aser_child_household_tib_scaled, method = "euclidean")
#head(round(as.matrix(eucl_dist), 2))
```

```
# Visualize the dissimilarity matrix (which will be a mess with  many observations)

fviz_dist(data_daisy, lab_size = 2) #set label size to 2 so it's readable
```




```
# The red color indicates high similarity
# The blue color indicates low similarity
# The color level is proportional to the value of dissimilarity between observations (pure red rep
resents zero and pure blue represents one)
# There are some clusters in there
```

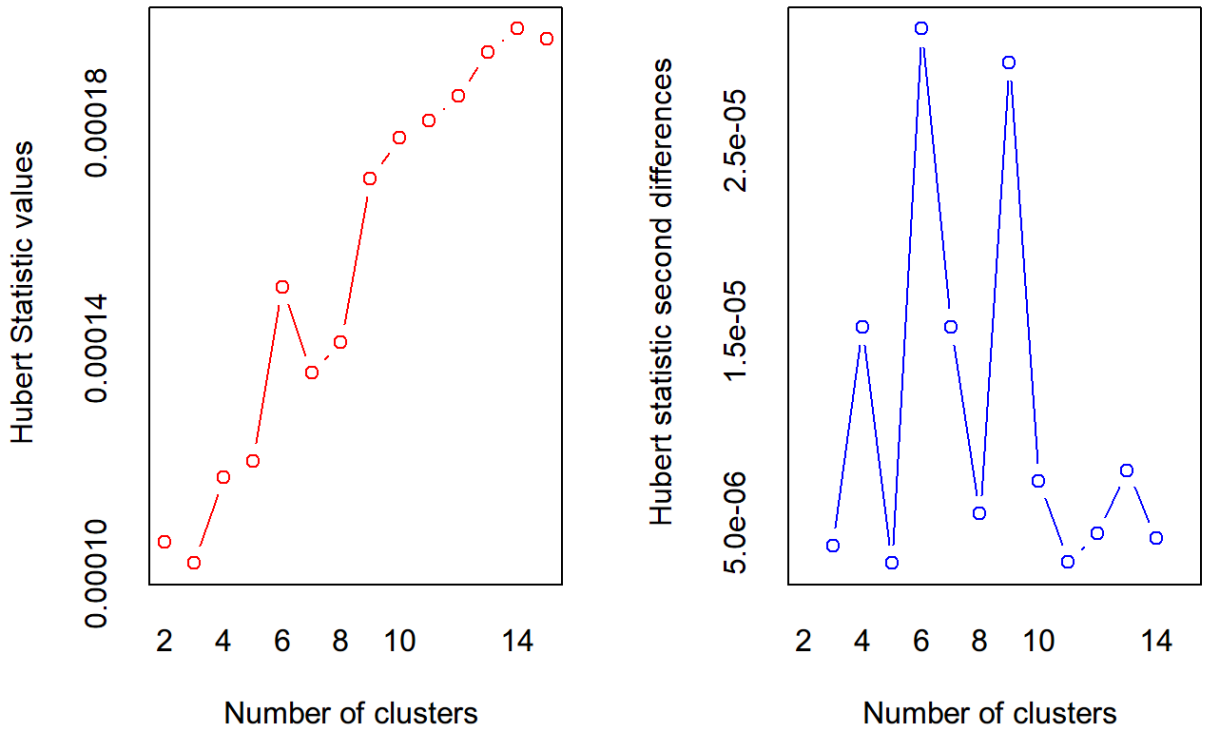
Partitioning Clustering

Partitioning clustering is a type of clustering that divides the data into non-overlapping subsets. The most popular partitioning clustering method is K-means clustering which we will be using for this project. This is because K-means clustering can handle larger datasets.

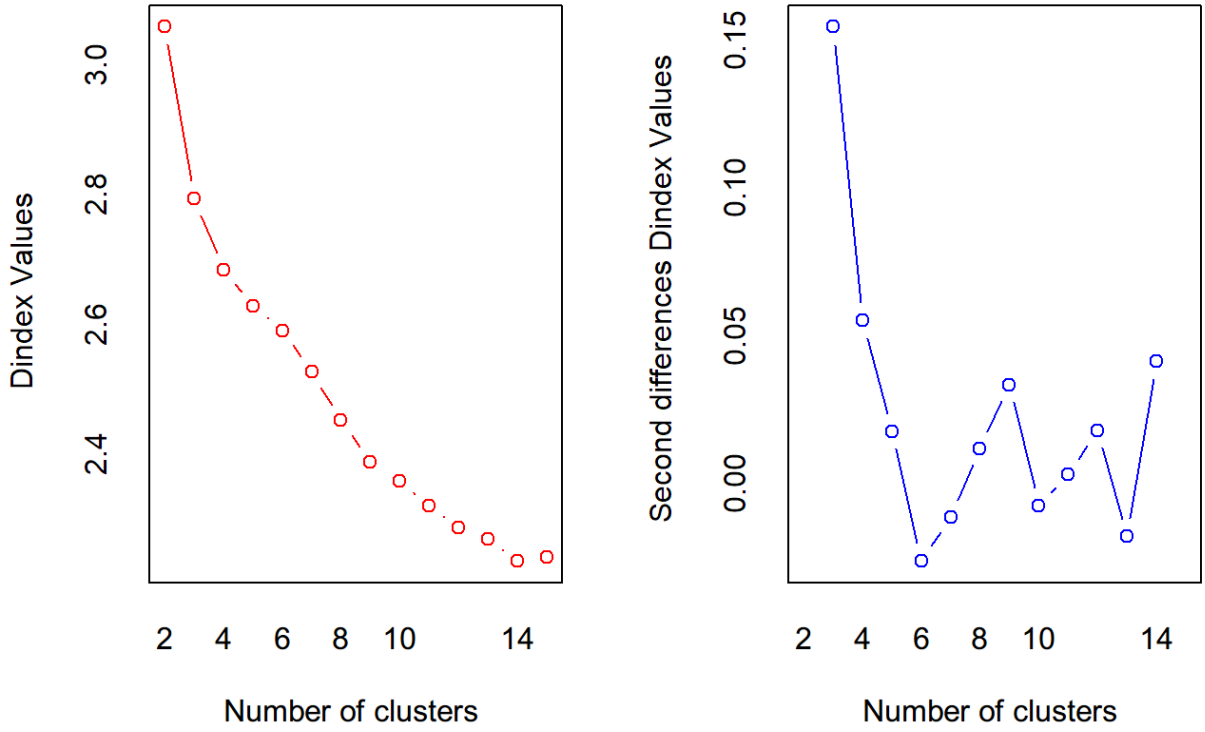
```
library(ggplot2) #for plotting
library(ggdendro) #for dendrograms
```

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

```
library(cluster) #for clustering
library(NbClust) #for finding the optimal number of clusters

# K-Means Clustering

set.seed(1234)

num_clust_k_mean <- NbClust(aser_child_household_tib_scaled, min.nc=2, max.nc=15, method="kmeans")
# 2-15 clusters, which is convention
```




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hu
bert
##                 index second differences plot.
##
```




```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Din
dex
##                 second differences plot) that corresponds to a significant increase of the valu
e of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 6 proposed 2 as the best number of clusters
## * 7 proposed 3 as the best number of clusters
## * 4 proposed 6 as the best number of clusters
## * 1 proposed 8 as the best number of clusters
## * 3 proposed 9 as the best number of clusters
## * 2 proposed 14 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************
```

```
#According to the majority rule, the best number of clusters is  3

num_clust_k_mean$Best.nc #what are the best number of clusters according to results
```

```
##                     KL       CH Hartigan     CCC    Scott      Marriot  TrCovW
## Number_clusters 9.0000   2.0000   3.0000 14.0000    9.000 6.000000e+00       3
## Value_Index     6.1254 638.9234 237.6244 28.3102 2455.347 1.357332e+41 2166595
##                   TraceW Friedman   Rubin Cindex     DB Silhouette   Duda
## Number_clusters    3.000   3.0000  9.0000 8.0000 6.0000     2.0000 2.0000
## Value_Index     2209.674   7.4259 -0.1089 0.2385 1.6455     0.2362 0.9945
##                  PseudoT2  Beale Ratkowsky    Ball PtBiserial Frey McClain
## Number_clusters   2.0000 2.0000    3.0000    3.00     6.0000    1  2.0000
## Value_Index       9.6952 0.0495    0.2965 5482.96     0.5106   NA  0.5944
##                   Dunn Hubert SDindex Dindex    SDbw
## Number_clusters 6.0000      0  3.0000      0 14.0000
## Value_Index     0.0446      0  1.6495      0  0.6921
```

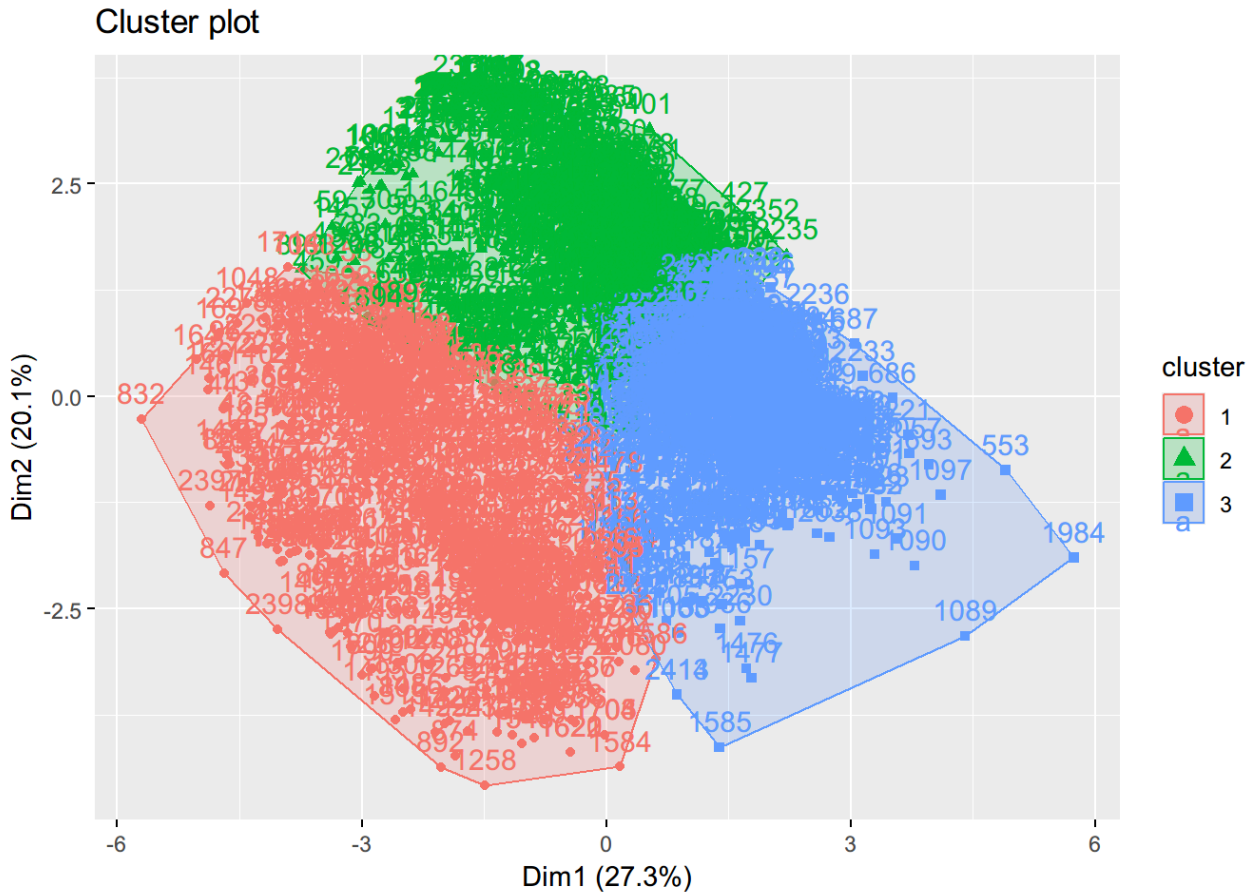
```
#show this in a histogram...
hist(num_clust_k_mean$Best.nc[1,],breaks = 15)  #3 seems to be the best number of clusters

# Final cluster
set.seed(1234)

final_km <- kmeans(aser_child_household_tib_scaled, 3, nstart=25)

final_km$size #how many observations in each cluster?
```

```
## [1]  691  714 1049
```

```
final_km$centers # coordinates of the final cluster centers (the final locations of the cluster ce
ntroids)
```

```
##   HouseholdCounter EarningMembers         Car MotorCycle TravelTime
## 1      -0.10305663     -0.3604706 -0.46871932 -0.4213939 -0.2193955
## 2      -0.01361321     -0.0496328 -0.06301049 -0.1056574 -0.1552914
## 3       0.07715154      0.2712326  0.35164398  0.3494972  0.2502196
##   ClimateChange FloodImpacted EarningImpacted PsychologicalImpacted
## 1   -0.53022885    -1.3151929      -0.8975102            -1.2540498
## 2    0.08741612     0.4892677       0.2456864             0.4800671
## 3    0.28977410     0.5333280       0.4239842             0.4993141
##   SchoolingAffected LocalLangReadingLevel ArithmeticLevel EnglishReadingLevel
## 1        -1.2375145           -0.04924437     -0.09470771          -0.1672107
## 2         0.5171527           -1.00847826     -0.97659268          -0.9689060
## 3         0.4631797            0.71885733      0.72710220           0.7696297
```

```
#get a cluster plot

fviz_cluster(final_km, data=aser_child_household_tib_scaled)
```



Analyzing The Clusters

The next step is to analyze the clusters based on the mean values of the variables in each cluster. This will help us understand the characteristics of each cluster and how they differ from each other. Note that while our original dataset contains a mix of continuous and ordinal variables, our interpretation of the clusters will be based on the mean values of these variables that will still be informative.

We begin by adding the cluster assignment to the original dataset.

```
# Adding the cluster assignment to the original dataset

final_km_clusters <- final_km$cluster

final_km_clusters <- as.data.frame(final_km_clusters) %>%
  rename(cluster = final_km_clusters)

aser_child_household_tib$cluster <- final_km$cluster
```

Creating a new column to see the sample size of each cluster. We find the following.

Cluster	Count
Cluster 1	691
Cluster 2	714
Cluster 3	1049

Across all the variables, we are summarizing the characteristics of variables by mean only.

```
# Analyzing the clusters based on the mean values of the variables in each cluster

aser_child_household_tib_test <- aser_child_household_tib %>%
  group_by(cluster) %>%
  mutate(sample_size = n()) %>%
  summarise(across(-c(IsInternetAvailable, InstitutionType, sample_size), ~ mean(.)),
              sample_size = mean(sample_size))

# Printing the output of the first analysis tibble.

print(aser_child_household_tib_test)
```

```
## # A tibble: 3 × 16
##   cluster       Id HouseholdCounter EarningMembers   Car MotorCycle TravelTime
##     <int>    <dbl>            <dbl>          <dbl> <dbl>      <dbl>      <dbl>
## 1       1 1561810.             9.90          0.711 0.346      0.660       1.34
## 2       2 1567553.            10.4           1.04  0.679      0.835       1.38
## 3       3 1567750.            11.0           1.38  1.02       1.09        1.66
## # ℹ 9 more variables: ClimateChange <dbl>, FloodImpacted <dbl>,
## #   EarningImpacted <dbl>, PsychologicalImpacted <dbl>,
## #   SchoolingAffected <dbl>, LocalLangReadingLevel <dbl>,
## #   ArithmeticLevel <dbl>, EnglishReadingLevel <dbl>, sample_size <dbl>
```

Making a second tibble to provide results of cluster based count and proportion for our grouping variables: IsInternetAvailable and InstitutionType to see how they differentiate across the three clusters.

```
# Making a second tibble to provide results of cluster based count and proportion for our grouping variables: IsInternetAvailable and InstitutionType to see how they differentiate across the three clusters.

aser_child_household_tib_test_2 <- aser_child_household_tib %>%
  mutate(IsInternetAvailable = factor(IsInternetAvailable, levels = c(0, 1), labels = c("No", "Yes")),
         InstitutionType = factor(InstitutionType, levels = c(1, 2, 3, 4), labels = c("Government", "Private", "Madrassah", "NFE/Other"))) %>%
  group_by(cluster, IsInternetAvailable, InstitutionType) %>%
  summarize(count = n()) %>%
  mutate(proportion = count / sum(count)) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'cluster', 'IsInternetAvailable'. You can
## override using the `.groups` argument.
```

```
print(aser_child_household_tib_test_2)
```

```
## # A tibble: 18 × 5
##    cluster IsInternetAvailable InstitutionType count proportion
##      <int> <fct>               <fct>           <int>      <dbl>
##  1       1 No                  Government        429    0.796
##  2       1 No                  Private            74    0.137
##  3       1 No                  NFE/Other          36    0.0668
##  4       1 Yes                 Government         71    0.467
##  5       1 Yes                 Private            81    0.533
##  6       2 No                  Government        360    0.721
##  7       2 No                  Private           132    0.265
##  8       2 No                  Madrassah           7    0.0140
##  9       2 Yes                 Government         78    0.363
## 10       2 Yes                 Private           135    0.628
## 11       2 Yes                 Madrassah           2    0.00930
## 12       3 No                  Government        368    0.626
## 13       3 No                  Private           204    0.347
## 14       3 No                  Madrassah          11    0.0187
## 15       3 No                  NFE/Other           5    0.00850
## 16       3 Yes                 Government        218    0.473
## 17       3 Yes                 Private           239    0.518
## 18       3 Yes                 Madrassah           4    0.00868
```

Interpretation of Clusters

The clusters are interpreted based on the mean values of the variables in each cluster.

Variables such as **Climate Change** (whether the household reports being informed about climate change or not) and **Schooling Affected** (whether the household reports schooling being affected during the survey year due to environmental factors) require inverse interpretation. For example, a **higher** mean score for 'Climate Change' within a cluster indicates lower reported understanding of climate-related challenges, while a **lower** score for 'Schooling Affected' corresponds to a higher extent of educational disruption.

For **FloodImpacted** (whether the household reports being impacted during the 2022-23 floods and to what extent), **lower** mean scores indicate **significant** impact from natural disasters and vice versa.

For **EarningImpacted** (whether the household reports their earning being impacted during the 2022-23 floods and to what extent), **lower** mean scores indicate that the cluster experienced **minor** earnings impact and vice versa.

For **PsychologicalImpact** (whether the household reports being psychologically impacted during the 2022-23 floods and to what extent), **lower** mean scores indicate **substantial** psychological impact from natural disasters and vice versa.

The variables related to academic achievement: **LocalLangReadingLevel**, **ArithmeticLevel**, **EnglishReadingLevel** can be interpreted as lower mean score indicating low proficiency and higher mean score indicating higher proficiency.

The variables related to **HouseholdCounter**, **EarningMembers**, **Car**, **Motorcycle**, and **TravelTime** are continuous and hence can be interpreted at face value.

Cluster 1: Low Socioeconomic Status, Climate-Conscious, Significantly Impacted, Moderate Academic Achievement

Cluster 1 represents a group that is aware of and significantly affected by climate change, particularly floods. They experience educational disruption and psychological impacts. Despite these challenges, they demonstrate moderate academic achievement. This group might have limited resources, as indicated by the lower number of earning members and car ownership. Their relatively shorter travel time to school could be a contributing factor to their moderate academic performance despite the climate-related challenges.

Cluster 2: Moderate Socioeconomic Status, Lower Climate Change Awareness, Moderately Impacted, Low Academic Achievement

Cluster 2 exhibits lower awareness of climate change and experiences a lower impact from floods compared to Cluster 1. They also report less disruption to earnings, education, and psychological well-being due to environmental factors. However, this cluster has the lowest academic achievement across all three subjects. This group might have slightly more resources than Cluster 1, but their academic performance is notably poor.

Cluster 3: High Socioeconomic Status, Lower Climate Change Awareness, Moderately Impacted, High Academic Achievement

Cluster 3 represents a group with lower reported understanding of climate change but also experiences a lower impact from floods and other environmental factors compared to Cluster 1. They have the highest academic achievement across all subjects, despite having the longest travel time to school. This cluster appears to be the most well-off in terms of household size, earning members, and asset ownership (cars and motorcycles).

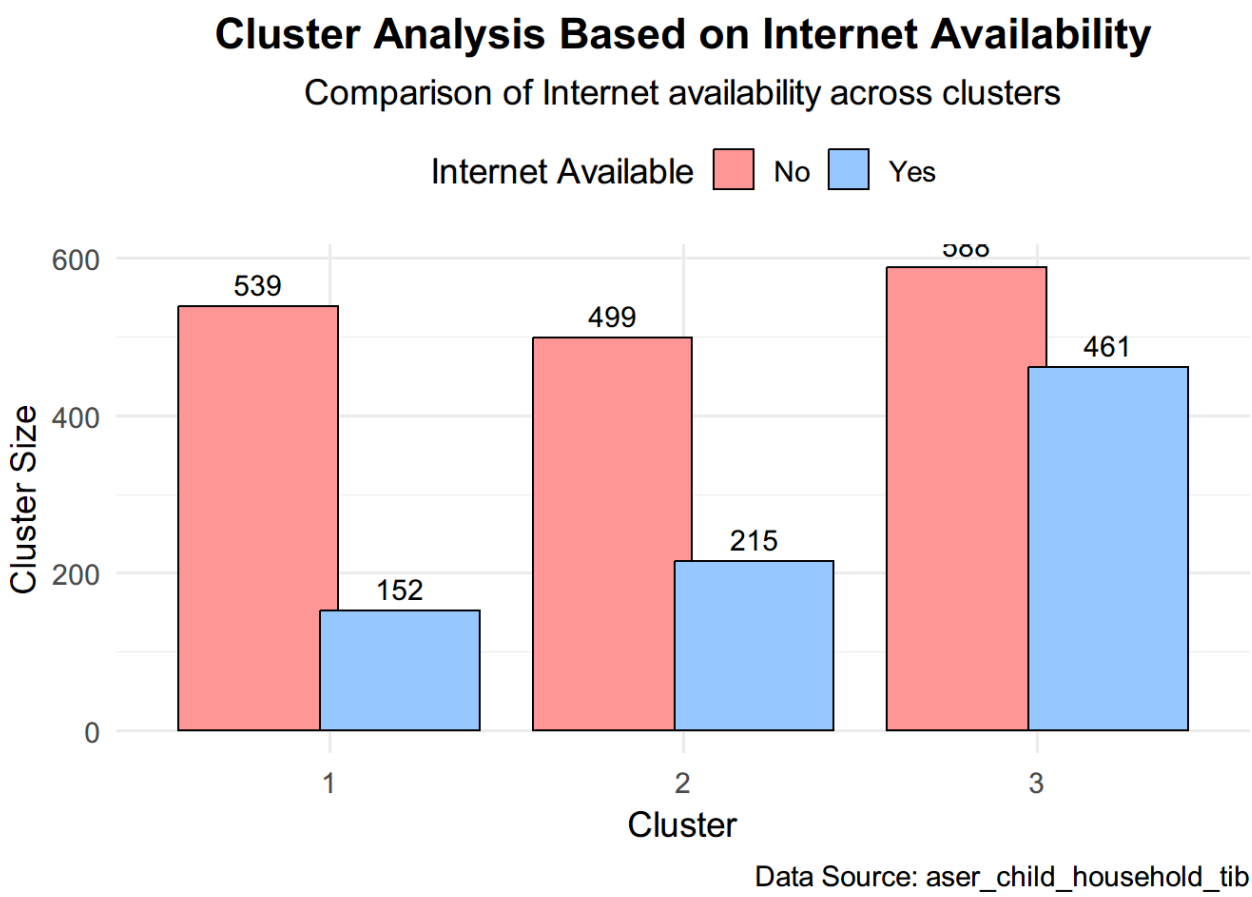
Visualizations of Clusters

Internet Availability Across the Clusters (Frequency)

The following bar plot shows the distribution of households with and without internet availability across the three clusters. Cluster 3 has the highest proportion of households with internet access, followed by Cluster 2 and Cluster 1. This suggests that internet access is predominant in households with higher socioeconomic status and there may be an underlying pattern resulting in higher academic achievement for the students in the said households compared to the households with no access.

```
# Analyzing The Clusters Based On IsInternetAvailable

ggplot(aser_child_household_tib %>%
         group_by(cluster, IsInternetAvailable) %>%
         summarise(sample_size = n()),
       aes(x = factor(cluster), y = sample_size, fill = factor(IsInternetAvailable))) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), color = "black") +
  scale_fill_manual(values = c("0" = "#FF9999", "1" = "#99CCFF"),
                    name = "Internet Available",
                    labels = c("No", "Yes")) +
  labs(
    title = "Cluster Analysis Based on Internet Availability",
    subtitle = "Comparison of Internet availability across clusters",
    x = "Cluster",
    y = "Cluster Size",
    caption = "Data Source: aser_child_household_tib"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "top"
  ) +
  geom_text(aes(label = sample_size),
            position = position_dodge(width = 0.8),
            vjust = -0.5, size = 4, color = "black")
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```

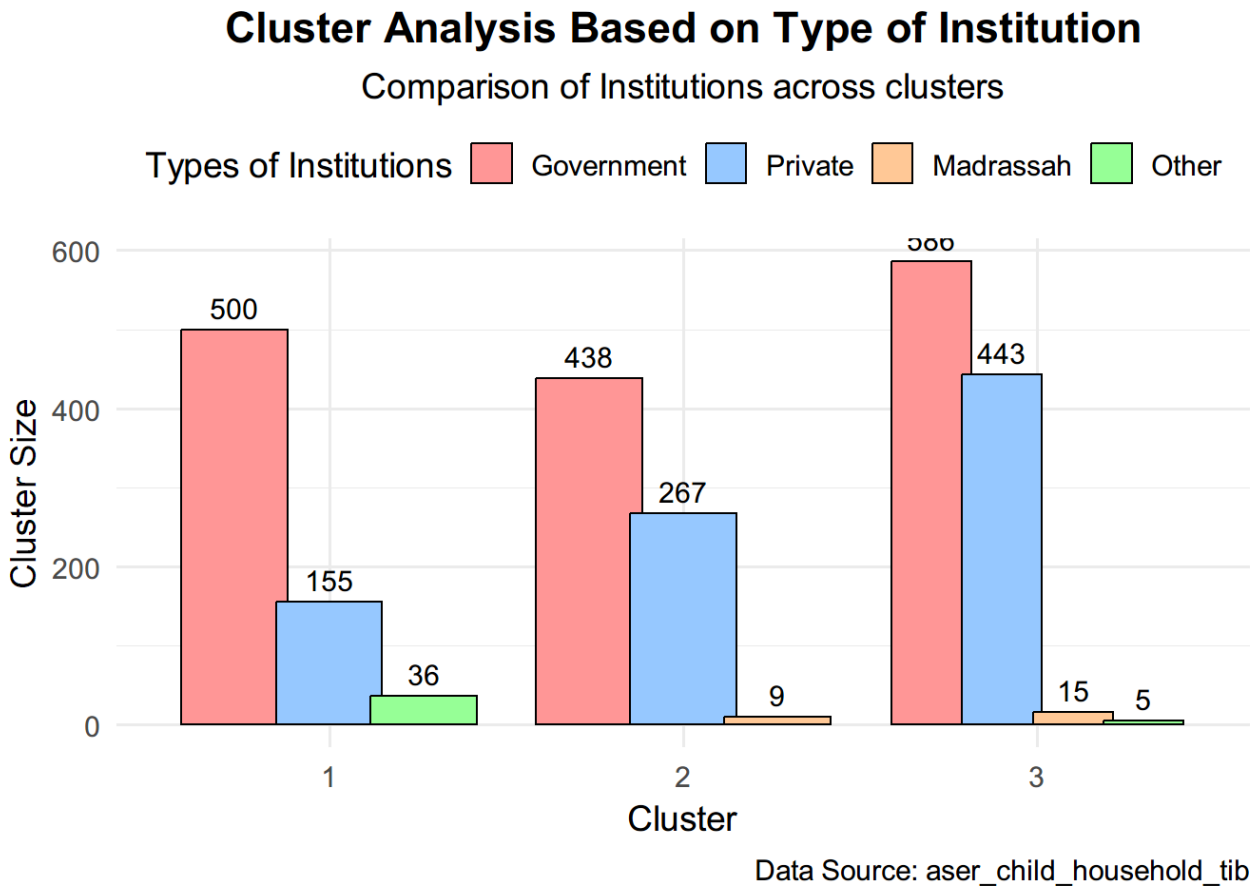


Types of School Across the Clusters (Frequency)

The following bar plot shows the distribution of different types of institutions that students are enrolled in across the three clusters. Cluster 3 has the highest proportion of students enrolled in private institutions, followed by Cluster 2 and Cluster 1. This suggests that students from households with higher socioeconomic status are more likely to attend private schools, which may contribute to their higher academic achievement compared to students in government or other types of institutions.

```
# Analyzing The Clusters Based On InstitutionType

ggplot(aser_child_household_tib %>%
         group_by(cluster, InstitutionType) %>%
         summarise(sample_size = n()),
       aes(x = factor(cluster), y = sample_size, fill = factor(InstitutionType))) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), color = "black") +
  scale_fill_manual(values = c("1" = "#FF9999", "2" = "#99CCFF", "3" = "#FFCC99", "4" = "#99FF9
9"),
                    name = "Types of Institutions",
                    labels = c("Government", "Private", "Madrassah", "Other")) +
  labs(
    title = "Cluster Analysis Based on Type of Institution",
    subtitle = "Comparison of Institutions across clusters",
    x = "Cluster",
    y = "Cluster Size",
    caption = "Data Source: aser_child_household_tib"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "top"
  ) +
  geom_text(aes(label = sample_size),
            position = position_dodge(width = 0.8),
            vjust = -0.5, size = 4, color = "black")
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```

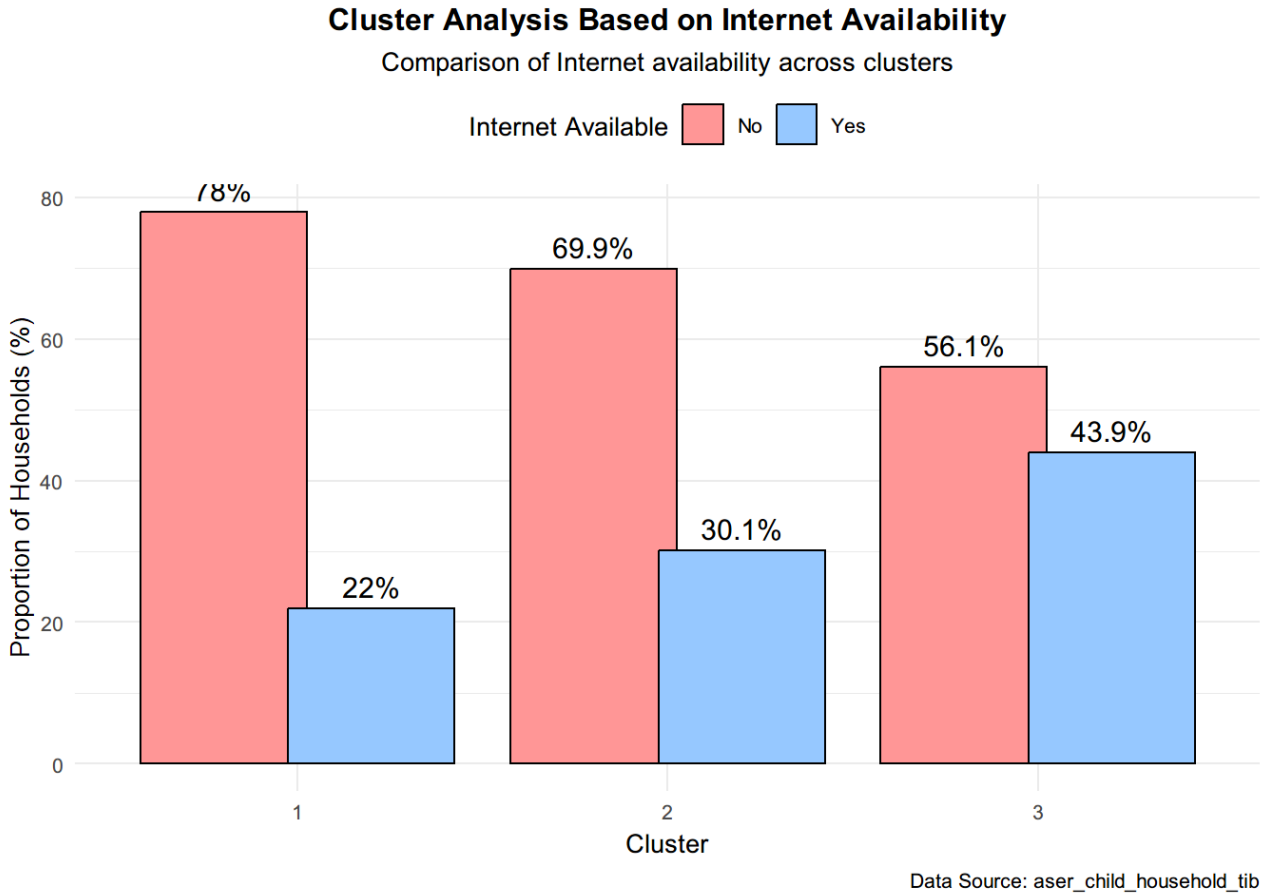


Internet Availability Across the Clusters (Percentage)

```
# Calculate proportions within each cluster
proportions_df <- aser_child_household_tib %>%
  group_by(cluster, IsInternetAvailable) %>%
  summarise(sample_size = n()) %>%
  group_by(cluster) %>%
  mutate(proportion = (sample_size / sum(sample_size)) * 100)
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```

```
# Plot the proportions
ggplot(proportions_df,
       aes(x = factor(cluster), y = proportion, fill = factor(IsInternetAvailable))) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), color = "black") +
  scale_fill_manual(values = c("0" = "#FF9999", "1" = "#99CCFF"),
                        name = "Internet Available",
                        labels = c("No", "Yes")) +
  labs(
    title = "Cluster Analysis Based on Internet Availability",
    subtitle = "Comparison of Internet availability across clusters",
    x = "Cluster",
    y = "Proportion of Households (%)",
    caption = "Data Source: aser_child_household_tib"
  ) +
  theme_minimal(base_size = 10) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "top"
  ) +
  geom_text(aes(label = paste0(round(proportion, 1), "%")),
              position = position_dodge(width = 0.8),
              vjust = -0.5, size = 4, color = "black")
```

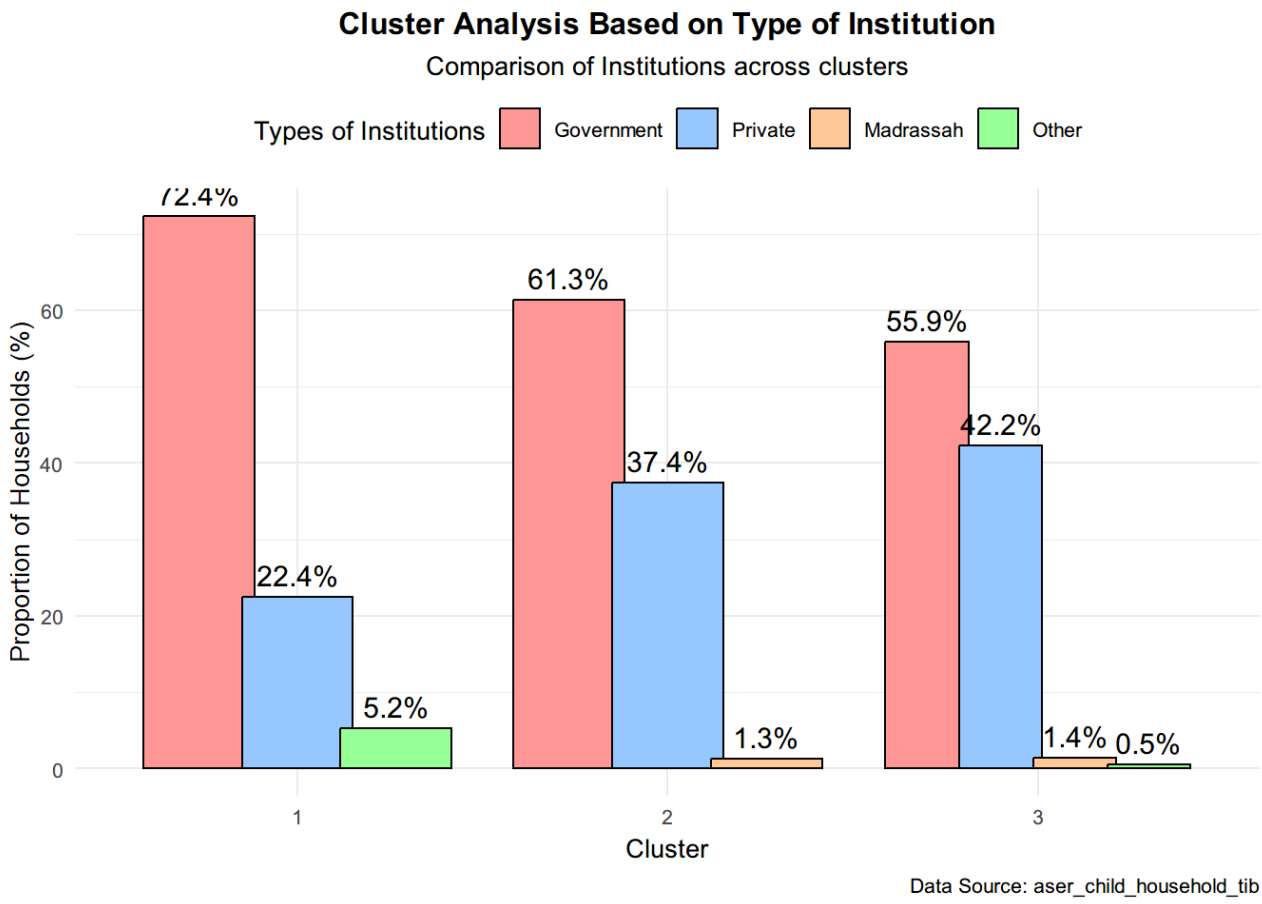


Types of School Across the Clusters (Percentage)

```
# Calculate proportions within each cluster
proportions_df_in <- aser_child_household_tib %>%
  group_by(cluster, InstitutionType) %>%
  summarise(sample_size = n()) %>%
  group_by(cluster) %>%
  mutate(proportion = (sample_size / sum(sample_size)) * 100)
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```

```
# Plot the proportions
ggplot(proportions_df_in,
       aes(x = factor(cluster), y = proportion, fill = factor(InstitutionType))) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), color = "black") +
  scale_fill_manual(values = c("1" = "#FF9999", "2" = "#99CCFF", "3" = "#FFCC99", "4" = "#99FF9
9"),
                    name = "Types of Institutions",
                    labels = c("Government", "Private", "Madrassah", "Other")) +
  labs(
    title = "Cluster Analysis Based on Type of Institution",
    subtitle = "Comparison of Institutions across clusters",
    x = "Cluster",
    y = "Proportion of Households (%)",
    caption = "Data Source: aser_child_household_tib"
  ) +
  theme_minimal(base_size = 10) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "top"
  ) +
  geom_text(aes(label = paste0(round(proportion, 1), "%")),
            position = position_dodge(width = 0.8),
            vjust = -0.5, size = 4, color = "black")
```



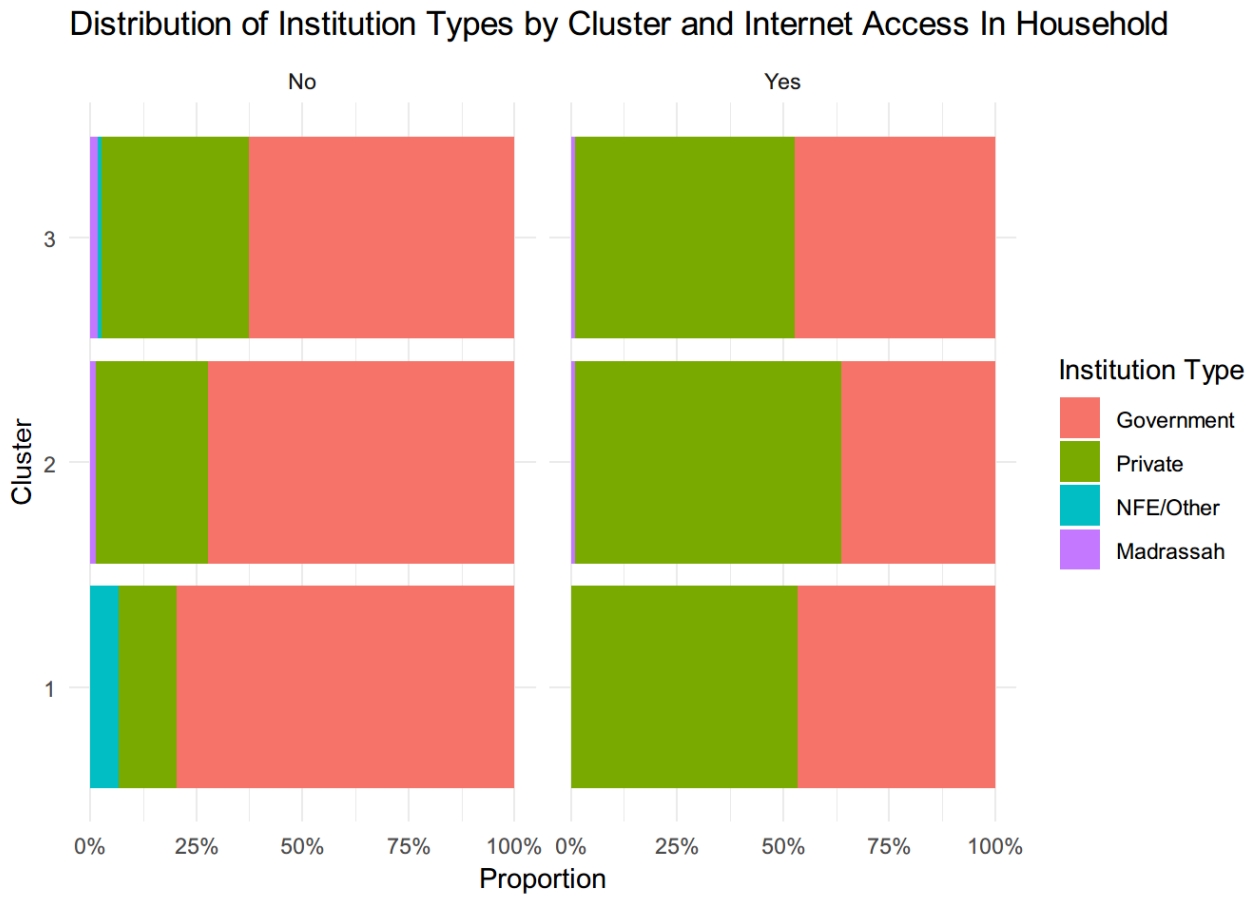
Clustering Rural Households by Internet Access and Institution Type

The following plot shows the distribution of different types of institutions that students are enrolled in across the three clusters, with separate panels for households with and without internet access. This visualization provides a comprehensive view of how internet availability and institution types are distributed across the clusters.

As we can see, households with internet access are more likely to have students enrolled in private institutions, followed by government institutions. In contrast, households without internet access have a higher proportion of students enrolled in government institutions, followed by private institutions.

This suggests that internet access may be associated with higher socioeconomic status, leading to a higher likelihood of students attending private schools leading to better academic outcomes.

```
ggplot(aser_child_household_tib_test_2, aes(x = factor(cluster), y = proportion, fill = reorder(In
stitutionType, -proportion))) +
  geom_bar(stat = "identity") +
  coord_flip() + # Flip the x and y axes for improved readability
  facet_wrap(~IsInternetAvailable) + # Create separate panels for "Yes" and "No" internet access
  labs(title = "Distribution of Institution Types by Cluster and Internet Access In Household",
       x = "Cluster",
       y = "Proportion",
       fill = "Institution Type") +
  theme_minimal() +
  scale_y_continuous(labels = scales::percent_format()) # Format y-axis as percentages
```



Conclusion

This project identified three distinct rural household profiles in Pakistan based on socioeconomic factors, climate change impacts, and academic achievements of the students belonging to those surveyed households.

- **Cluster 1** represents households with low socioeconomic status, high climate change awareness, significant impact from floods, moderate academic achievement, and limited resources.
- **Cluster 2** represents households with moderate socioeconomic status, lower climate change awareness, moderate impact from floods, low academic achievement, and slightly more resources than Cluster 1.
- **Cluster 3** represents households with high socioeconomic status, lower climate change awareness, moderate impact from floods, high academic achievement, and the most resources among the three clusters.

The analysis also revealed that across the three clusters households with internet access are more likely to have students enrolled in private institutions, while households without internet access have a higher proportion of students enrolled in government institutions. This suggests that internet access may be associated with higher socioeconomic status, leading to better academic outcomes for students attending private schools.

Another way to interpret these results would be to consider the specific case of Cluster 3, which shows resilience to climate change impact despite showing low awareness about climate change compared to Cluster 1. Yet, due to their high socioeconomic status through Internet and/or digital access to communication and information, they are able to provide better educational opportunities for their children, leading to higher academic achievement.

Discussion and Limitations

The findings of this project have certain limitations beginning with the nature of the data itself. The data is based on self-reported responses from households, which may be subject to recall bias or social desirability bias. Additionally, the survey was conducted in rural areas of Pakistan, which, despite the label, is not a homogeneous group. There may be significant variations within rural areas that are not captured by the data.

Furthermore, this study did not conduct analysis of variance (ANOVA) or other statistical tests to determine the statistical significance of the observed differences in internet access and academic achievement across the clusters. Therefore, the findings should be interpreted as descriptive and exploratory rather than inferential.

Furthermore, the clustering analysis is based on a limited set of variables, which may not fully capture the complexity of rural households in Pakistan.