biostats_consulting

2024-10-05

Contents

```r
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(readr)
library(summarytools)
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(stringr)
library(Rtsne)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
# Load the dataset
data_2022 <- read_csv("S:\\biostats_consulting_lab\\cleaned_2022_survey_dta.csv")
```

```
## Rows: 1423 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (14): consent, availability, sec1_q4, sec1_q5, sec1_q6, sec1_q7, sec1_q...
## dbl   (2): caseid, sec11_start
## date  (1): sec1_q1
##
## 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
data_2024 <- read_csv("S:\\biostats_consulting_lab\\cleaned_2024_survey_dta.csv")
```

```
## Rows: 1405 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (27): response_1, response_2, response_3, phone_rel, resp_relationship_...
## dbl   (2): caseid, phone_response
## lgl   (1): religion_oth
## date  (2): birthdate, survey_date
##
## 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
# Rename the specified variables and clean data_2022
data_2022_cleaned <- data_2022 %>%
  rename(
    dob = sec1_q1,
    gender = sec1_q4,
    highest_education = sec1_q5,
    employment_status = sec1_q6,
    marital_status = sec1_q7,
    household_income = sec1_q8,
    residence_area = sec1_q9,
    survey_location = sec1_q10,
    survey_duration = sec11_start,
    religious = sec11_q156,
    religion = sec11_q157,
    specified_other_religion = sec11_q157other,
    science_contradict = sec11_q158,
    science_or_religion = sec11_q159
  ) %>%
  select(-consent, -availability) %>%
  mutate(
    religion = case_when(
      religion %in% c("CCAP", "Traditional African religion") ~ "Other",
      religion %in% c("Seventh Day Adventist") ~ "Other Christian",
      religion == "Prefer not to answer" ~ "Prefer not to answer [do not read aloud]",
      TRUE ~ religion
    ),
    employment_status = if_else(is.na(employment_status), "Missing", employment_status)
  )
```

```
data_2022_cleaned <- data_2022_cleaned %>%
  mutate(across(c(religion,science_or_religion, science_contradict, religious),
                ~ replace_na(., "Missing")))
```

```
# Clean and rename specified variables in data_2024
data_2024_cleaned <- data_2024 %>%
  # Rename variables
  rename(
    caseid = caseid,
    response_status = response_1,
    response_by = response_2,
    dob = birthdate,
    highest_education = educ_level,
    employment_status = employ_status,
    people_speak_to_daily = number_people,
    household_income = hh_income,
    specified_other_religion = religion_oth,
    call_status = call_status
  ) %>%

  # Select only the relevant variables
  select(
    dob, caseid, response_status, response_by, gender, highest_education, marital_status, parent_guardia
    employment_status, work_industry, people_speak_to_daily,
    household_income, residence_area, religion,
    specified_other_religion, call_status, survey_date
  ) %>%

  # Clean data by re-coding and handling missing values
  mutate(
    # Re-code religion variable by grouping similar categories
    religion = case_when(
      religion %in% c("Seventh Day Adventists", "Apostolic/New Apostlic Church", "Church of Christ",
                      "Gospel/NewTestament/Injili Church", "Salvation Army Church", "Assembly of God Chu
                      "Roho Church", "Church of God", "Jehovah's Witness", "Legio Maria Church", "NENO"
                      "Repentance and Holiness", "Pentecostal/ Protestant Church") ~ "Other Christian",
      religion == "Prefer not to answer [do not read aloud]" ~ "Prefer not to answer",
      religion == "Akorino" ~ "Other",
      religion == "Baptist Church" ~ "Baptist",
      TRUE ~ religion
    ),

    # Re-code employment_status variable
    employment_status = case_when(
      employment_status %in% c("Self-employed (includes agribusiness)", "Peasant farmer") ~ "Self-emplo
      TRUE ~ employment_status
    ),
    highest_education = case_when(
      highest_education == "Prefer not to answer" ~ "Prefer not to answer [do not read aloud]",
      TRUE ~ highest_education
    )
  ) %>%

  # Replace NA values in highest_education with "Missing"
  mutate(highest_education = replace_na(highest_education, "Missing")) %>%
  mutate(marital_status = replace_na(marital_status, "Missing"))%>%
  mutate(parent_guardian = replace_na(parent_guardian, "Missing"))%>%
  mutate(work_industry = replace_na(work_industry, "Missing"))%>%
  mutate(people_speak_to_daily = replace_na(people_speak_to_daily, "Missing"))%>%
  mutate(household_income = replace_na(household_income, "Missing"))%>%
  mutate(residence_area = replace_na(residence_area, "Missing"))%>%
  mutate(employment_status = replace_na(employment_status, "Missing"))%>%
  mutate(religion = replace_na(religion, "Missing"))
```

```
data_2024_cleaned <- data_2024_cleaned %>%
  left_join(data_2022_cleaned %>% select(caseid, gender, religion, dob), by = "caseid", suffix = c("_20
  mutate(
    #  2024 gender NA  2022 gender
    gender_2024 = coalesce(gender_2024, gender_2022),
    #  gender NA   "Unknown"
    gender_2024 = replace_na(gender_2024, "Unknown"),

    dob_2024 = coalesce(dob_2024,dob_2022),
    #  religion NA   "Unknown"
    dob_2024 = replace_na(dob_2024, "Unknown")
  ) %>%

  #  gender_2022 religion_2022    gender_2024 religion_2024
  select(-gender_2022, -dob_2022) %>%
  rename(gender = gender_2024, dob = dob_2024)
```

```
# Replace "Prefer not to answer [do not read aloud]" with "Prefer not to answer" across all columns
data_2024_cleaned <- data_2024_cleaned %>%
  mutate(across(everything(), ~str_replace(., "Prefer not to answer \\[do not read aloud\\]", "Prefer no
data_2022_cleaned <- data_2022_cleaned %>%
  mutate(across(everything(), ~str_replace(., "Prefer not to answer \\[do not read aloud\\]", "Prefer no
```

```r
# Display the first few rows of the cleaned datasets
head(data_2022_cleaned)
```

```
## # A tibble: 6 x 15
##   caseid dob        gender highest_education employment_status    marital_status
##   <chr>  <chr>      <chr>  <chr>             <chr>                <chr>
## 1 1012   1993-07-04 Male   Secondary         Casual laborer       Divorced/Sepa~
## 2 1054   1992-02-04 Female Primary           Not employed and no~ Married
## 3 1182   1984-08-17 Female Higher            Not employed but lo~ Married
## 4 1220   1992-06-23 Male   Higher            Self-employed        Married
## 5 1223   1975-01-01 Female Secondary         Self-employed        Married
## 6 1255   1982-09-23 Female Higher            Employed full-time   Married
## # i 9 more variables: household_income <chr>, residence_area <chr>,
## #   survey_location <chr>, survey_duration <chr>, religious <chr>,
## #   religion <chr>, specified_other_religion <chr>, science_contradict <chr>,
## #   science_or_religion <chr>
```

```r
head(data_2024_cleaned)
```

```
## # A tibble: 6 x 18
##   dob        caseid response_status         response_by gender highest_education
##   <chr>      <chr>  <chr>                   <chr>       <chr>  <chr>
## 1 1998-11-09 10003  Answered the phone, co~ <NA>        Female Secondary
## 2 1974-06-06 10048  Answered the phone, co~ <NA>        Female Secondary
## 3 1994-06-30 10077  Answered the phone, co~ <NA>        Female Primary
## 4 1969-07-07 10086  Answered the phone, co~ <NA>        Male   Higher
## 5 1995-08-08 10088  Number does not work (~ <NA>        Male   Missing
## 6 1982-01-01 10119  Answered the phone, co~ <NA>        Female Missing
## # i 12 more variables: marital_status <chr>, parent_guardian <chr>,
## #   employment_status <chr>, work_industry <chr>, people_speak_to_daily <chr>,
## #   household_income <chr>, residence_area <chr>, religion_2024 <chr>,
## #   specified_other_religion <chr>, call_status <chr>, survey_date <chr>,
## #   religion_2022 <chr>
```

```r
# Check for missing values in both datasets
missing_values_2022 <- sapply(data_2022_cleaned, function(x) sum(is.na(x)))
missing_values_2024 <- sapply(data_2024_cleaned, function(x) sum(is.na(x)))
print(missing_values_2022)
```

```
##                   caseid                      dob                   gender
##                        0                        0                        0
##        highest_education        employment_status           marital_status
##                        0                        0                        0
##         household_income           residence_area          survey_location
##                        0                        0                        0
##          survey_duration                religious                 religion
##                       29                        0                        0
## specified_other_religion       science_contradict      science_or_religion
##                     1421                        0                        0
```

```r
print(missing_values_2024)
```

```
##                      dob                   caseid          response_status
##                        0                        0                        0
##              response_by                   gender        highest_education
##                     1367                        0                        0
##           marital_status          parent_guardian        employment_status
##                        0                        0                        0
##            work_industry    people_speak_to_daily         household_income
##                        0                        0                        0
##           residence_area            religion_2024 specified_other_religion
##                        0                        0                     1405
##              call_status              survey_date            religion_2022
##                        0                        1                        0
```

```r
# Get summary statistics for both datasets
summary(data_2022_cleaned)
```

```
##     caseid              dob                gender          highest_education
##  Length:1423        Length:1423        Length:1423        Length:1423
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  employment_status  marital_status     household_income   residence_area
##  Length:1423        Length:1423        Length:1423        Length:1423
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  survey_location    survey_duration     religious           religion
##  Length:1423        Length:1423        Length:1423        Length:1423
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  specified_other_religion science_contradict science_or_religion
##  Length:1423              Length:1423        Length:1423
##  Class :character         Class :character   Class :character
##  Mode  :character         Mode  :character   Mode  :character
```

```r
summary(data_2024_cleaned)
```

```
##      dob               caseid          response_status    response_by
##  Length:1405        Length:1405        Length:1405        Length:1405
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     gender          highest_education  marital_status     parent_guardian
##  Length:1405        Length:1405        Length:1405        Length:1405
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  employment_status  work_industry      people_speak_to_daily household_income
##  Length:1405        Length:1405        Length:1405           Length:1405
##  Class :character   Class :character   Class :character      Class :character
##  Mode  :character   Mode  :character   Mode  :character      Mode  :character
##  residence_area     religion_2024      specified_other_religion
##  Length:1405        Length:1405        Length:1405
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##  call_status        survey_date        religion_2022
##  Length:1405        Length:1405        Length:1405
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```
# Check case IDs in both datasets
caseid_2022 <- data_2022_cleaned$caseid
caseid_2024 <- data_2024_cleaned$caseid

# Filter the 2024 dataset to only include those who successfully followed up
successful_followup_2024 <- data_2024_cleaned %>%
  filter(response_status == "Answered the phone, correct respondent" & call_status == "Completed")

# Extract the case IDs of the successfully followed-up participants
caseid_successful_followup <- successful_followup_2024$caseid

# Identify participants present in both 2022 and successfully followed up in 2024
common_successful_followup <- intersect(caseid_2022, caseid_successful_followup)

# Identify participants in 2022 but not in the successfully followed-up group in 2024 (dropped out)
dropped_participants <- setdiff(caseid_2022, caseid_successful_followup)

# Identify participants in 2024 (successfully followed up) but not in 2022 (new participants)
new_participants <- setdiff(caseid_successful_followup, caseid_2022)

# Output the counts
cat("Number of participants successfully followed up in 2024: ", length(common_successful_followup), "\
```

```
## Number of participants successfully followed up in 2024:  1096
```

```
cat("Number of participants who dropped out after 2022: ", length(dropped_participants), "\n")
```

```
## Number of participants who dropped out after 2022:  327
```

```
cat("Number of new participants who joined in 2024: ", length(new_participants), "\n")
```

```
## Number of new participants who joined in 2024:  0
```

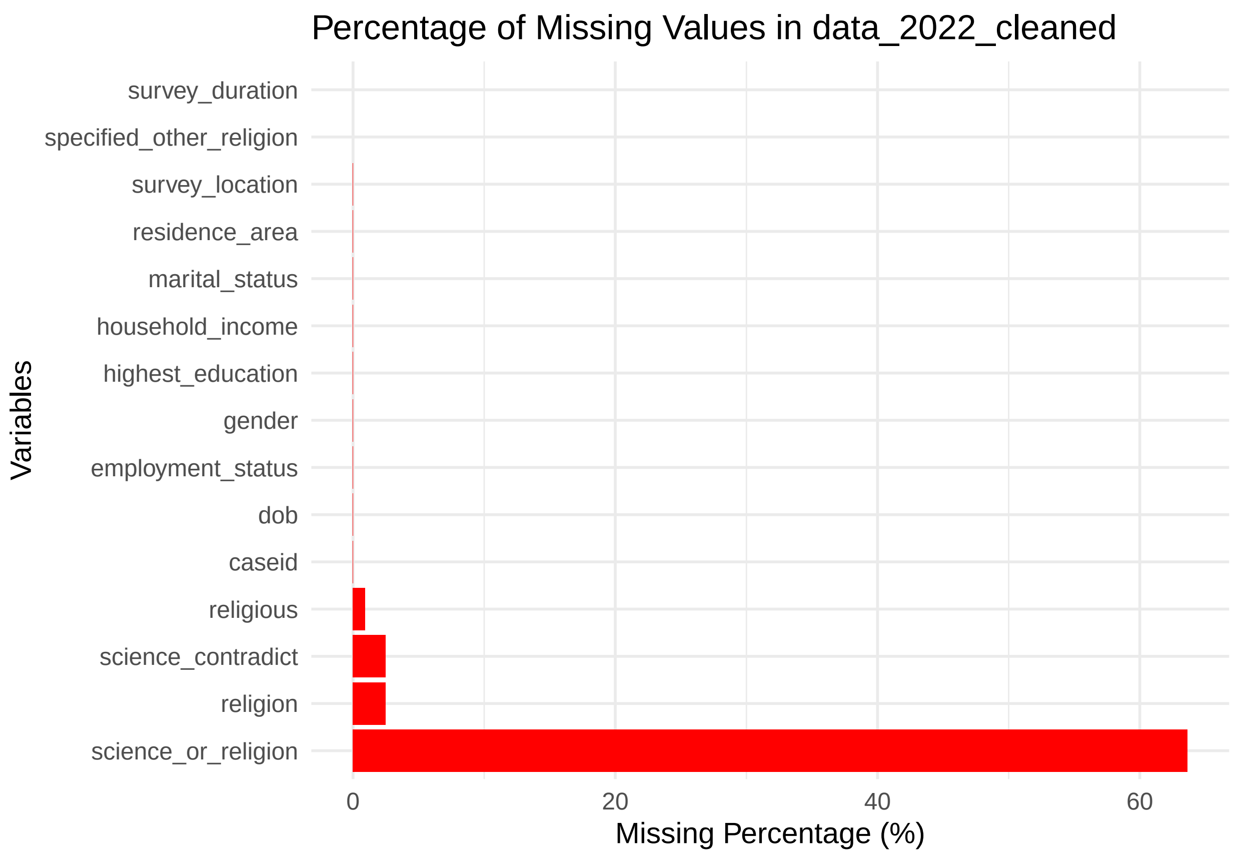
```
# View unique values for key variables across both datasets
list(
  religion_2022 = unique(data_2022_cleaned$religion),
  religion_2024 = unique(data_2024_cleaned$religion_2024),
  highest_education_2022 = unique(data_2022_cleaned$highest_education),
  highest_education_2024 = unique(data_2024_cleaned$highest_education),
  employment_status_2022 = unique(data_2022_cleaned$employment_status),
  employment_status_2024 = unique(data_2024_cleaned$employment_status),
  marital_status_2022 = unique(data_2022_cleaned$marital_status),
  marital_status_2024 = unique(data_2024_cleaned$marital_status)
)
```

```
## $religion_2022
## [1] "Other Christian"      "Anglican"             "Catholic"
## [4] "Muslim"               "Other"                "Missing"
## [7] "Baptist"              "Prefer not to answer"
##
## $religion_2024
## [1] "Other Christian"      "Catholic"             "Missing"
## [4] "Muslim"               "Anglican"             "Prefer not to answer"
## [7] "No Religion"          "Baptist"              "Other"
##
## $highest_education_2022
## [1] "Secondary"                          "Primary"
## [3] "Higher"                             "No school/Did not complete primary"
##
## $highest_education_2024
## [1] "Secondary"                          "Primary"
## [3] "Higher"                             "Missing"
## [5] "No school/Did not complete primary" "Prefer not to answer"
##
## $employment_status_2022
## [1] "Casual laborer"
## [2] "Not employed and not looking for work"
## [3] "Not employed but looking for work"
## [4] "Self-employed"
## [5] "Employed full-time"
## [6] "Employed part-time"
## [7] "Prefer not to answer"
##
## $employment_status_2024
## [1] "Employed part-time"
## [2] "Self-employed"
## [3] "Not employed but looking for work"
## [4] "Employed full-time"
## [5] "Missing"
## [6] "Casual laborer"
## [7] "Not employed and not looking for work"
## [8] "Prefer not to answer"
##
## $marital_status_2022
## [1] "Divorced/Separated"   "Married"              "Single"
## [4] "Widowed"              "Cohabiting/Partnered" "Prefer not to answer"
##
## $marital_status_2024
## [1] "Single"               "Married"              "Missing"
## [4] "Widowed"              "Divorced/Separated"   "Prefer not to answer"
## [7] "Cohabiting/Partnered"
```

```r
data_2022_cleaned <- data_2022_cleaned %>%
  semi_join(data_2024_cleaned, by = "caseid")  #   2024  caseid

data_2024_cleaned <- data_2024_cleaned %>%
  semi_join(data_2022_cleaned, by = "caseid")  #   2022  caseid
```

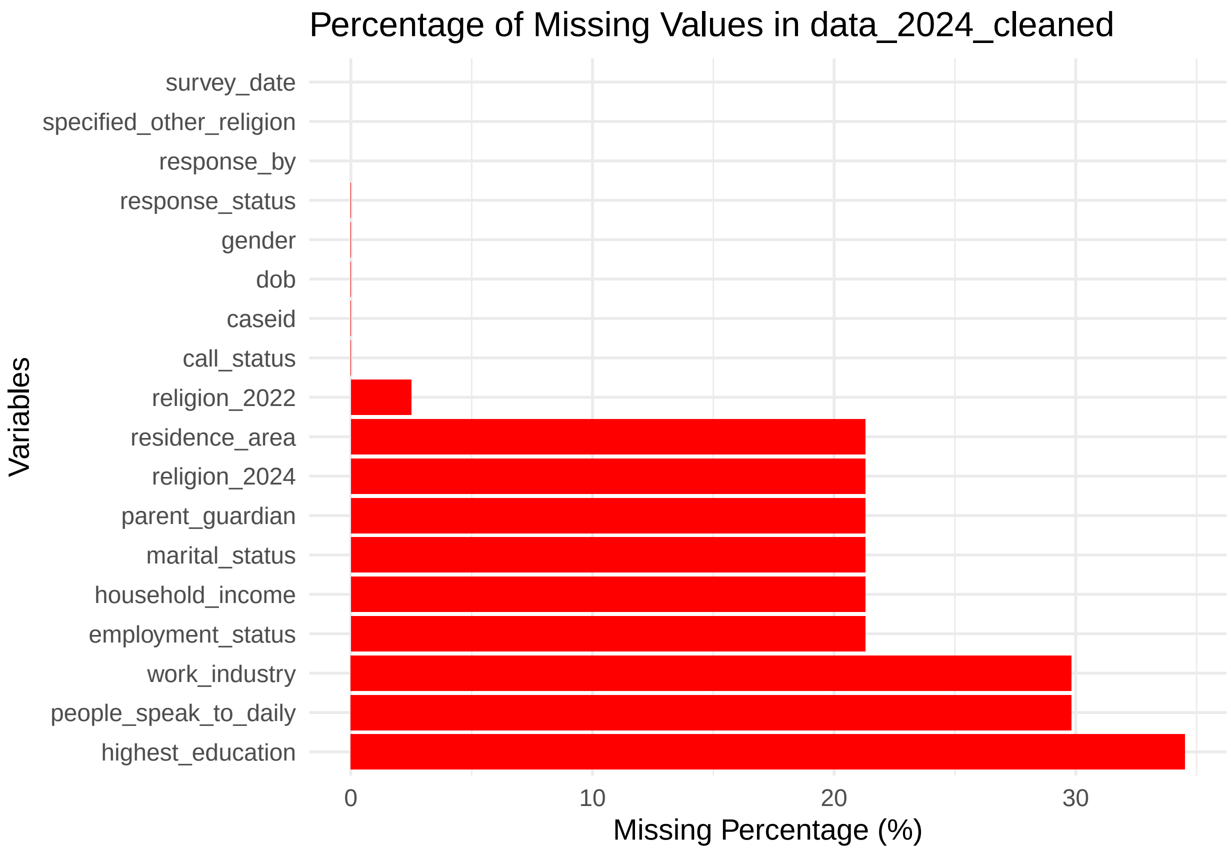
```
# Function to calculate percentage of missing values
calculate_missing_percentage <- function(data) {
  data %>%
    summarise(across(everything(), ~ sum(. == "Missing") / n() * 100)) %>%
    pivot_longer(cols = everything(), names_to = "variable", values_to = "missing_percentage")
}

# Calculate missing percentages for both datasets
missing_2022 <- calculate_missing_percentage(data_2022_cleaned)
missing_2024 <- calculate_missing_percentage(data_2024_cleaned)

# Plot missing values for data_2022_cleaned
plot_2022 <- ggplot(missing_2022, aes(x = reorder(variable, -missing_percentage), y = missing_percentag
  geom_bar(stat = "identity", fill = "red") +
  coord_flip() +
  labs(title = "Percentage of Missing Values in data_2022_cleaned", x = "Variables", y = "Missing Percen
  theme_minimal()

# Plot missing values for data_2024_cleaned
plot_2024 <- ggplot(missing_2024, aes(x = reorder(variable, -missing_percentage), y = missing_percentag
  geom_bar(stat = "identity", fill = "red") +
  coord_flip() +
  labs(title = "Percentage of Missing Values in data_2024_cleaned", x = "Variables", y = "Missing Percen
  theme_minimal()

# Display the plots
plot_2022
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



```
plot_2024
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



```r
# Merge the datasets by caseid and create new variables indicating changes between 2022 and 2024
merged_data <- full_join(data_2022_cleaned, data_2024_cleaned, by = "caseid", suffix = c("_2022", "_202
  mutate(lost = if_else(is.na(response_status) | response_status != "Answered the phone, correct respon
  mutate(
    education_change = if_else(
      is.na(highest_education_2022) | is.na(highest_education_2024) |
      highest_education_2022 == "Missing" | highest_education_2024 == "Missing",
      3,  # Set as 3 when missing in either year
      if_else(highest_education_2022 != highest_education_2024, 1, 0)
    ),

    employment_change = if_else(
      is.na(employment_status_2022) | is.na(employment_status_2024) |
      employment_status_2022 == "Missing" | employment_status_2024 == "Missing",
      3,  # Set as 3 when missing in either year
      if_else(employment_status_2022 != employment_status_2024, 1, 0)
    ),

    income_change = if_else(
      is.na(household_income_2022) | is.na(household_income_2024) |
      household_income_2022 == "Missing" | household_income_2024 == "Missing",
      3,  # Set as 3 when missing in either year
      if_else(household_income_2022 != household_income_2024, 1, 0)
    ),

    residence_change = if_else(
      is.na(residence_area_2022) | is.na(residence_area_2024) |
      residence_area_2022 == "Missing" | residence_area_2024 == "Missing",
      3,  # Set as 3 when missing in either year
      if_else(residence_area_2022 != residence_area_2024, 1, 0)
    ),

    religion_change = if_else(
      is.na(religion_2022) | is.na(religion_2024) |
      religion_2022 == "Missing" | religion_2024 == "Missing",
      3,  # Set as 3 when missing in either year
      if_else(religion_2022 != religion_2024, 1, 0)
    ),

    residence_area_change = if_else(
      is.na(residence_area_2022) | is.na(residence_area_2024) |
      residence_area_2022 == "Missing" | residence_area_2024 == "Missing",
      3,  # Set as 3 when missing in either year
      if_else(residence_area_2022 != residence_area_2024, 1, 0)
    )
  ) %>%
  select(
    caseid,
    dob_2022, gender_2022,
    highest_education_2022, highest_education_2024,
    marital_status_2022, marital_status_2024,
    employment_status_2022, employment_status_2024,
    household_income_2022, household_income_2024,
    residence_area_2022, residence_area_2024, residence_area_change,
    religion_2022, religion_2024,
    specified_other_religion_2022,
    response_status, response_by,
    parent_guardian, science_or_religion,
    lost, religious,
    education_change, employment_change, income_change, residence_change, religion_change
  )

# View the first few rows to verify the new order
head(merged_data)
```

```
## # A tibble: 6 x 28
##   caseid dob_2022   gender_2022 highest_education_2022 highest_education_2024
##   <chr>  <chr>      <chr>       <chr>                  <chr>
## 1 1012   1993-07-04 Male        Secondary              Missing
## 2 1054   1992-02-04 Female      Primary                Missing
## 3 1182   1984-08-17 Female      Higher                 Missing
## 4 1220   1992-06-23 Male        Higher                 Missing
## 5 1223   1975-01-01 Female      Secondary              Missing
## 6 1255   1982-09-23 Female      Higher                 Missing
## # i 23 more variables: marital_status_2022 <chr>, marital_status_2024 <chr>,
## #   employment_status_2022 <chr>, employment_status_2024 <chr>,
## #   household_income_2022 <chr>, household_income_2024 <chr>,
## #   residence_area_2022 <chr>, residence_area_2024 <chr>,
## #   residence_area_change <dbl>, religion_2022 <chr>, religion_2024 <chr>,
## #   specified_other_religion_2022 <chr>, response_status <chr>,
## #   response_by <chr>, parent_guardian <chr>, science_or_religion <chr>, ...
```

```
# Summarize the percentages of each change status (0, 1, 3) for each variable
change_percentages <- merged_data %>%
  summarize(
    education_change_0 = mean(education_change == 0, na.rm = TRUE) * 100,  # No change
    education_change_1 = mean(education_change == 1, na.rm = TRUE) * 100,  # Changed
    education_change_3 = mean(education_change == 3, na.rm = TRUE) * 100,  # Unknown

    employment_change_0 = mean(employment_change == 0, na.rm = TRUE) * 100,
    employment_change_1 = mean(employment_change == 1, na.rm = TRUE) * 100,
    employment_change_3 = mean(employment_change == 3, na.rm = TRUE) * 100,

    income_change_0 = mean(income_change == 0, na.rm = TRUE) * 100,
    income_change_1 = mean(income_change == 1, na.rm = TRUE) * 100,
    income_change_3 = mean(income_change == 3, na.rm = TRUE) * 100,

    residence_change_0 = mean(residence_change == 0, na.rm = TRUE) * 100,
    residence_change_1 = mean(residence_change == 1, na.rm = TRUE) * 100,
    residence_change_3 = mean(residence_change == 3, na.rm = TRUE) * 100,

    religion_change_0 = mean(religion_change == 0, na.rm = TRUE) * 100,
    religion_change_1 = mean(religion_change == 1, na.rm = TRUE) * 100,
    religion_change_3 = mean(religion_change == 3, na.rm = TRUE) * 100,

    residence_area_change_0 = mean(residence_area_change == 0, na.rm = TRUE) * 100,  # No change in res
    residence_area_change_1 = mean(residence_area_change == 1, na.rm = TRUE) * 100,  # Changed residenc
    residence_area_change_3 = mean(residence_area_change == 3, na.rm = TRUE) * 100   # Unknown/missing
  )

change_percentages
```

```
## # A tibble: 1 x 18
##   education_change_0 education_change_1 education_change_3 employment_change_0
##                <dbl>              <dbl>              <dbl>               <dbl>
## 1               47.0               18.5               34.5                43.3
## # i 14 more variables: employment_change_1 <dbl>, employment_change_3 <dbl>,
## #   income_change_0 <dbl>, income_change_1 <dbl>, income_change_3 <dbl>,
## #   residence_change_0 <dbl>, residence_change_1 <dbl>,
## #   residence_change_3 <dbl>, religion_change_0 <dbl>, religion_change_1 <dbl>,
## #   religion_change_3 <dbl>, residence_area_change_0 <dbl>,
## #   residence_area_change_1 <dbl>, residence_area_change_3 <dbl>
```

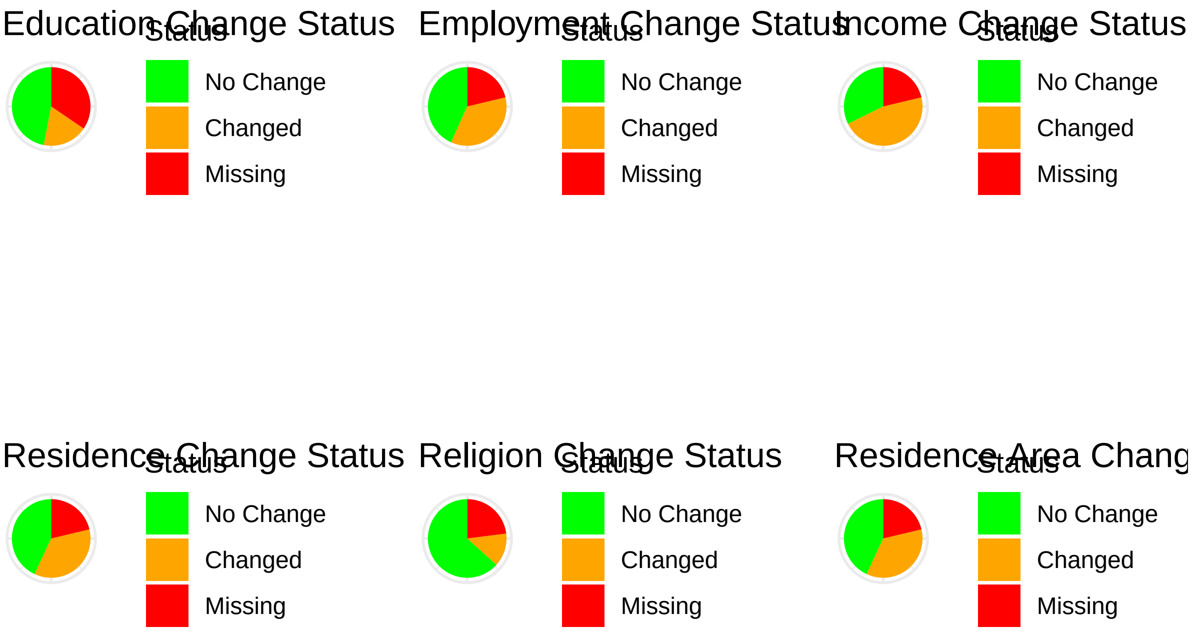
```
dfSummary(merged_data) %>% view()
```

```
## Switching method to 'browser'
```

```
## Output file written: C:\Users\ghlas\AppData\Local\Temp\RtmpGstDKA\file9260515c466d.html
```

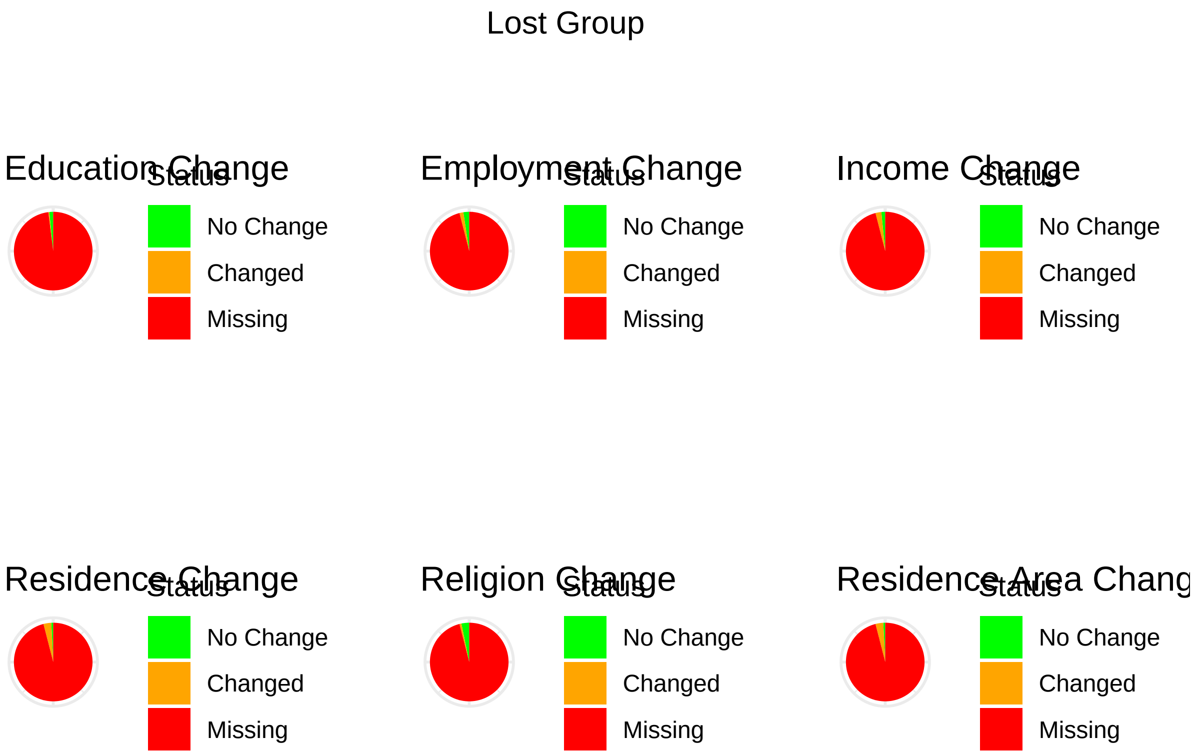
```r
# Prepare data for pie charts
education_data <- merged_data %>%
  count(education_change) %>%
  mutate(percentage = n / sum(n) * 100)

employment_data <- merged_data %>%
  count(employment_change) %>%
  mutate(percentage = n / sum(n) * 100)

income_data <- merged_data %>%
  count(income_change) %>%
  mutate(percentage = n / sum(n) * 100)

residence_data <- merged_data %>%
  count(residence_change) %>%
  mutate(percentage = n / sum(n) * 100)

religion_data <- merged_data %>%
  count(religion_change) %>%
  mutate(percentage = n / sum(n) * 100)

residence_area_data <- merged_data %>%
  count(residence_area_change) %>%
  mutate(percentage = n / sum(n) * 100)

# Create pie charts for each change status
education_pie <- ggplot(education_data, aes(x = "", y = percentage, fill = factor(education_change))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Education Change Status", fill = "Status", y = "", x = "") +
  theme_minimal() +
  theme(axis.text.x = element_blank()) +
  scale_fill_manual(values = c("green", "orange", "red"),
                    labels = c("No Change", "Changed", "Missing"))

employment_pie <- ggplot(employment_data, aes(x = "", y = percentage, fill = factor(employment_change))
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Employment Change Status", fill = "Status", y = "", x = "") +
  theme_minimal() +
  theme(axis.text.x = element_blank()) +
  scale_fill_manual(values = c("green", "orange", "red"),
                    labels = c("No Change", "Changed", "Missing"))

income_pie <- ggplot(income_data, aes(x = "", y = percentage, fill = factor(income_change))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Income Change Status", fill = "Status", y = "", x = "") +
  theme_minimal() +
  theme(axis.text.x = element_blank()) +
  scale_fill_manual(values = c("green", "orange", "red"),
                    labels = c("No Change", "Changed", "Missing"))
```

```
residence_pie <- ggplot(residence_data, aes(x = "", y = percentage, fill = factor(residence_change))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Residence Change Status", fill = "Status", y = "", x = "") +
  theme_minimal() +
  theme(axis.text.x = element_blank()) +
  scale_fill_manual(values = c("green", "orange", "red"),
                    labels = c("No Change", "Changed", "Missing"))

religion_pie <- ggplot(religion_data, aes(x = "", y = percentage, fill = factor(religion_change))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Religion Change Status", fill = "Status", y = "", x = "") +
  theme_minimal() +
  theme(axis.text.x = element_blank()) +
  scale_fill_manual(values = c("green", "orange", "red"),
                    labels = c("No Change", "Changed", "Missing"))

residence_area_pie <- ggplot(residence_area_data, aes(x = "", y = percentage, fill = factor(residence_a
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Residence Area Change Status", fill = "Status", y = "", x = "") +
  theme_minimal() +
  theme(axis.text.x = element_blank()) +
  scale_fill_manual(values = c("green", "orange", "red"),
                    labels = c("No Change", "Changed", "Missing"))

# Arrange the pie charts in a 2 x 3 layout
grid.arrange(education_pie, employment_pie, income_pie, residence_pie, religion_pie, residence_area_pie
```

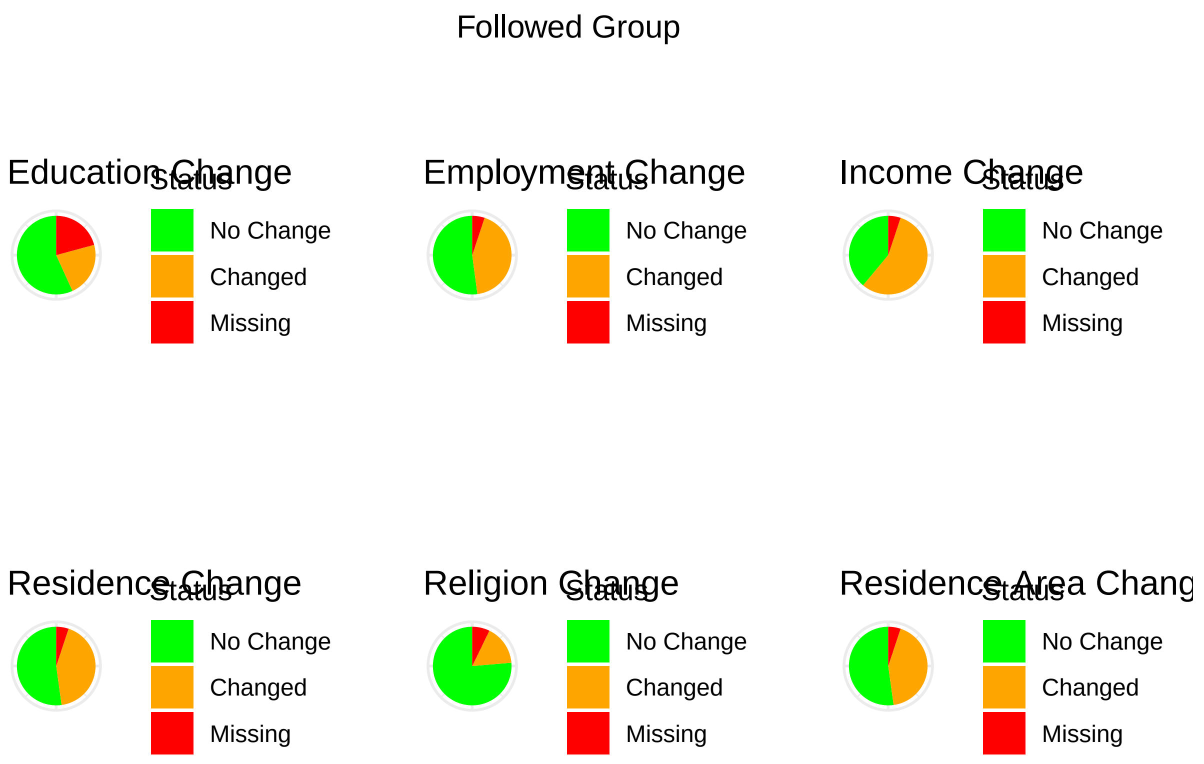


```
# Split data into lost and followed groups
lost_data <- merged_data %>% filter(lost == 1)
followed_data <- merged_data %>% filter(lost == 0)
```

```r
# Function to prepare data for pie chart
prepare_pie_data <- function(data, variable) {
  data %>%
    count({{ variable }}) %>%
    mutate(percentage = n / sum(n) * 100)
}

# Function to create pie chart
create_pie_chart <- function(pie_data, title, variable_name) {
  ggplot(pie_data, aes(x = "", y = percentage, fill = factor({{ variable_name }}))) +
    geom_bar(stat = "identity", width = 1) +
    coord_polar("y", start = 0) +
    labs(title = title, fill = "Status", y = "", x = "") +
    theme_minimal() +
    theme(axis.text.x = element_blank()) +
    scale_fill_manual(values = c("green", "orange", "red"),
                      labels = c("No Change", "Changed", "Missing"))
}

# Prepare data for pie charts for both groups
# For Lost Group
education_lost <- prepare_pie_data(lost_data, education_change)
employment_lost <- prepare_pie_data(lost_data, employment_change)
income_lost <- prepare_pie_data(lost_data, income_change)
residence_lost <- prepare_pie_data(lost_data, residence_change)
religion_lost <- prepare_pie_data(lost_data, religion_change)
residence_area_lost <- prepare_pie_data(lost_data, residence_area_change)

# For Followed Group
education_followed <- prepare_pie_data(followed_data, education_change)
employment_followed <- prepare_pie_data(followed_data, employment_change)
income_followed <- prepare_pie_data(followed_data, income_change)
residence_followed <- prepare_pie_data(followed_data, residence_change)
religion_followed <- prepare_pie_data(followed_data, religion_change)
residence_area_followed <- prepare_pie_data(followed_data, residence_area_change)

# Create pie charts for lost group
education_pie_lost <- create_pie_chart(education_lost, "Education Change", education_change)
employment_pie_lost <- create_pie_chart(employment_lost, "Employment Change", employment_change)
income_pie_lost <- create_pie_chart(income_lost, "Income Change", income_change)
residence_pie_lost <- create_pie_chart(residence_lost, "Residence Change", residence_change)
religion_pie_lost <- create_pie_chart(religion_lost, "Religion Change", religion_change)
residence_area_pie_lost <- create_pie_chart(residence_area_lost, "Residence Area Change", residence_area

# Create pie charts for followed group
education_pie_followed <- create_pie_chart(education_followed, "Education Change", education_change)
employment_pie_followed <- create_pie_chart(employment_followed, "Employment Change", employment_change
income_pie_followed <- create_pie_chart(income_followed, "Income Change", income_change)
residence_pie_followed <- create_pie_chart(residence_followed, "Residence Change", residence_change)
religion_pie_followed <- create_pie_chart(religion_followed, "Religion Change", religion_change)
residence_area_pie_followed <- create_pie_chart(residence_area_followed, "Residence Area Change", reside

# Arrange the pie charts in two sets (Lost and Followed)
# Lost group: 2 rows x 3 columns
grid.arrange(education_pie_lost, employment_pie_lost, income_pie_lost,
             residence_pie_lost, religion_pie_lost, residence_area_pie_lost,
             ncol = 3, top = "Lost Group")
```



```r
# Followed group: 2 rows x 3 columns
grid.arrange(education_pie_followed, employment_pie_followed, income_pie_followed,
             residence_pie_followed, religion_pie_followed, residence_area_pie_followed,
             ncol = 3, top = "Followed Group")
```



```
#
merged_data$highest_education_2022 <- as.factor(merged_data$highest_education_2022)
merged_data$employment_status_2022 <- as.factor(merged_data$employment_status_2022)
merged_data$household_income_2022 <- as.factor(merged_data$household_income_2022)
merged_data$residence_area_2022 <- as.factor(merged_data$residence_area_2022)
merged_data$gender_2022 <- as.factor(merged_data$gender_2022)
merged_data$marital_status_2022 <- as.factor(merged_data$marital_status_2022)
merged_data$religion_2022 <- as.factor(merged_data$religion_2022)
merged_data$lost <- as.factor(merged_data$lost)
merged_data$science_or_religion <- as.factor(merged_data$science_or_religion)

#          "Missing"   "Prefer not to answer"
clean_data <- merged_data %>%
  filter(
    !highest_education_2022 %in% c("Missing") &
    !employment_status_2022 %in% c("Missing") &
    !household_income_2022 %in% c("Missing") &
    !residence_area_2022 %in% c("Missing") &
    !gender_2022 %in% c("Missing") &
    !marital_status_2022 %in% c("Missing") &
    !religion_2022 %in% c("Missing")
    )
#   specified_other_religion_2022   response_by
clean_data <- clean_data %>%
  select(-specified_other_religion_2022, -response_by)

#
print(dim(clean_data))   #
```

```
## [1] 1370   26
```

```
#
clean_model <- glm(lost ~ highest_education_2022 + employment_status_2022 + household_income_2022 + res
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
#
summary(clean_model)
```

```
##
## Call:
## glm(formula = lost ~ highest_education_2022 + employment_status_2022 +
##     household_income_2022 + residence_area_2022 + gender_2022 +
##     marital_status_2022 + religion_2022, family = binomial, data = clean_data)
##
## Coefficients:
##                                                                                         Estimate
## (Intercept)                                                                              0.14683
## highest_education_2022No school/Did not complete primary                                 0.17864
## highest_education_2022Primary                                                            0.13832
## highest_education_2022Secondary                                                          0.25153
## employment_status_2022Employed full-time                                                -0.46860
## employment_status_2022Employed part-time                                                -0.15634
## employment_status_2022Not employed and not looking for work                             -0.15416
## employment_status_2022Not employed but looking for work                                 -0.31022
## employment_status_2022Prefer not to answer                                             -28.62845
## employment_status_2022Self-employed                                                     -0.19386
## household_income_2022Allowed me to save just a little                                  -0.13369
## household_income_2022Only just met my expenses                                          -0.36623
## household_income_2022Prefer not to answer                                               -0.90195
## household_income_2022Was not sufficient, so needed to use savings to meet expenses     -0.18930
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses   -0.72224
## residence_area_2022Trading Center (town)                                                -0.12989
## residence_area_2022Village (rural)                                                      -0.22297
## gender_2022Male                                                                         -0.34320
## marital_status_2022Divorced/Separated                                                   -1.29711
## marital_status_2022Married                                                              -1.08374
## marital_status_2022Prefer not to answer                                                 14.65645
## marital_status_2022Single                                                               -0.84313
## marital_status_2022Widowed                                                              -1.11145
## religion_2022Baptist                                                                    -0.05014
## religion_2022Catholic                                                                    0.27565
## religion_2022Muslim                                                                      0.82240
## religion_2022Other                                                                      -0.28737
## religion_2022Other Christian                                                             0.14829
## religion_2022Prefer not to answer                                                       -0.83134
##                                                                                       Std. Error
## (Intercept)                                                                              1.11207
## highest_education_2022No school/Did not complete primary                                 0.36247
## highest_education_2022Primary                                                            0.20526
## highest_education_2022Secondary                                                          0.18512
## employment_status_2022Employed full-time                                                 0.27555
## employment_status_2022Employed part-time                                                 0.43657
## employment_status_2022Not employed and not looking for work                              0.34606
## employment_status_2022Not employed but looking for work                                 0.29186
## employment_status_2022Prefer not to answer                                            868.34868
## employment_status_2022Self-employed                                                     0.23410
## household_income_2022Allowed me to save just a little                                   0.44519
## household_income_2022Only just met my expenses                                          0.38772
## household_income_2022Prefer not to answer                                               1.13526
## household_income_2022Was not sufficient, so needed to use savings to meet expenses      0.41397
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses    0.39219
## residence_area_2022Trading Center (town)                                                0.19705
## residence_area_2022Village (rural)                                                      0.19138
## gender_2022Male                                                                         0.15658
## marital_status_2022Divorced/Separated                                                   1.03949
## marital_status_2022Married                                                              0.96698
## marital_status_2022Prefer not to answer                                               634.47620
## marital_status_2022Single                                                               0.97936
## marital_status_2022Widowed                                                              1.04607
## religion_2022Baptist                                                                    0.80711
## religion_2022Catholic                                                                   0.27974
## religion_2022Muslim                                                                     0.34868
## religion_2022Other                                                                      0.80681
## religion_2022Other Christian                                                            0.25909
## religion_2022Prefer not to answer                                                    1187.49657
##                                                                                      z value
## (Intercept)                                                                            0.132
## highest_education_2022No school/Did not complete primary                               0.493
## highest_education_2022Primary                                                          0.674
## highest_education_2022Secondary                                                        1.359
## employment_status_2022Employed full-time                                              -1.701
## employment_status_2022Employed part-time                                              -0.358
## employment_status_2022Not employed and not looking for work                           -0.445
## employment_status_2022Not employed but looking for work                               -1.063
## employment_status_2022Prefer not to answer                                            -0.033
## employment_status_2022Self-employed                                                   -0.828
## household_income_2022Allowed me to save just a little                                 -0.300
## household_income_2022Only just met my expenses                                        -0.945
## household_income_2022Prefer not to answer                                             -0.794
## household_income_2022Was not sufficient, so needed to use savings to meet expenses    -0.457
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses  -1.842
## residence_area_2022Trading Center (town)                                              -0.659
## residence_area_2022Village (rural)                                                    -1.165
## gender_2022Male                                                                       -2.192
## marital_status_2022Divorced/Separated                                                 -1.248
## marital_status_2022Married                                                            -1.121
## marital_status_2022Prefer not to answer                                                0.023
## marital_status_2022Single                                                             -0.861
## marital_status_2022Widowed                                                            -1.062
## religion_2022Baptist                                                                  -0.062
## religion_2022Catholic                                                                  0.985
## religion_2022Muslim                                                                    2.359
## religion_2022Other                                                                    -0.356
## religion_2022Other Christian                                                           0.572
## religion_2022Prefer not to answer                                                     -0.001
##                                                                                      Pr(>|z|)
## (Intercept)                                                                            0.8950
## highest_education_2022No school/Did not complete primary                               0.6221
## highest_education_2022Primary                                                          0.5004
## highest_education_2022Secondary                                                        0.1742
## employment_status_2022Employed full-time                                               0.0890
## employment_status_2022Employed part-time                                               0.7203
## employment_status_2022Not employed and not looking for work                            0.6560
## employment_status_2022Not employed but looking for work                                0.2878
## employment_status_2022Prefer not to answer                                             0.9737
## employment_status_2022Self-employed                                                    0.4076
## household_income_2022Allowed me to save just a little                                  0.7639
## household_income_2022Only just met my expenses                                         0.3449
## household_income_2022Prefer not to answer                                              0.4269
## household_income_2022Was not sufficient, so needed to use savings to meet expenses     0.6475
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses   0.0655
## residence_area_2022Trading Center (town)                                               0.5098
## residence_area_2022Village (rural)                                                     0.2440
## gender_2022Male                                                                        0.0284
## marital_status_2022Divorced/Separated                                                  0.2121
## marital_status_2022Married                                                             0.2624
## marital_status_2022Prefer not to answer                                                0.9816
## marital_status_2022Single                                                              0.3893
## marital_status_2022Widowed                                                             0.2880
## religion_2022Baptist                                                                   0.9505
## religion_2022Catholic                                                                  0.3244
## religion_2022Muslim                                                                    0.0183
## religion_2022Other                                                                     0.7217
## religion_2022Other Christian                                                           0.5671
## religion_2022Prefer not to answer                                                      0.9994
##
## (Intercept)
## highest_education_2022No school/Did not complete primary
## highest_education_2022Primary
## highest_education_2022Secondary
## employment_status_2022Employed full-time                                             .
## employment_status_2022Employed part-time
## employment_status_2022Not employed and not looking for work
## employment_status_2022Not employed but looking for work
## employment_status_2022Prefer not to answer
## employment_status_2022Self-employed
## household_income_2022Allowed me to save just a little
## household_income_2022Only just met my expenses
## household_income_2022Prefer not to answer
## household_income_2022Was not sufficient, so needed to use savings to meet expenses
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses .
## residence_area_2022Trading Center (town)
## residence_area_2022Village (rural)
## gender_2022Male                                                                      *
## marital_status_2022Divorced/Separated
## marital_status_2022Married
## marital_status_2022Prefer not to answer
## marital_status_2022Single
## marital_status_2022Widowed
## religion_2022Baptist
## religion_2022Catholic
## religion_2022Muslim                                                                  *
## religion_2022Other
## religion_2022Other Christian
## religion_2022Prefer not to answer
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1274.5  on 1369  degrees of freedom
## Residual deviance: 1234.4  on 1341  degrees of freedom
## AIC: 1292.4
##
## Number of Fisher Scoring iterations: 14
```

```
#
simplified_model <- glm(lost ~ employment_status_2022 + gender_2022 + religion_2022,
                        data = clean_data, family = binomial)
#
summary(simplified_model)
```

```
##
## Call:
## glm(formula = lost ~ employment_status_2022 + gender_2022 + religion_2022,
##     family = binomial, data = clean_data)
##
## Coefficients:
##                                                               Estimate
## (Intercept)                                                   -1.35863
## employment_status_2022Employed full-time                      -0.44703
## employment_status_2022Employed part-time                      -0.20938
## employment_status_2022Not employed and not looking for work  -0.16233
## employment_status_2022Not employed but looking for work      -0.28798
## employment_status_2022Prefer not to answer                  -13.26068
## employment_status_2022Self-employed                           -0.17398
## gender_2022Male                                               -0.36444
## religion_2022Baptist                                          -0.12530
## religion_2022Catholic                                          0.26719
## religion_2022Muslim                                            0.84886
## religion_2022Other                                            -0.23183
## religion_2022Other Christian                                   0.13760
## religion_2022Prefer not to answer                              0.05324
##                                                             Std. Error z value
## (Intercept)                                                    0.31070  -4.373
## employment_status_2022Employed full-time                       0.25369  -1.762
## employment_status_2022Employed part-time                       0.42179  -0.496
## employment_status_2022Not employed and not looking for work    0.34121  -0.476
## employment_status_2022Not employed but looking for work        0.28619  -1.006
## employment_status_2022Prefer not to answer                   440.28389  -0.030
## employment_status_2022Self-employed                            0.22770  -0.764
## gender_2022Male                                                0.14987  -2.432
## religion_2022Baptist                                           0.80035  -0.157
## religion_2022Catholic                                          0.27743   0.963
## religion_2022Muslim                                            0.34042   2.494
## religion_2022Other                                             0.79550  -0.291
## religion_2022Other Christian                                   0.25668   0.536
## religion_2022Prefer not to answer                            763.85068   0.000
##                                                             Pr(>|z|)
## (Intercept)                                                 1.23e-05 ***
## employment_status_2022Employed full-time                      0.0781 .
## employment_status_2022Employed part-time                      0.6196
## employment_status_2022Not employed and not looking for work   0.6343
## employment_status_2022Not employed but looking for work       0.3143
## employment_status_2022Prefer not to answer                    0.9760
## employment_status_2022Self-employed                           0.4448
## gender_2022Male                                               0.0150 *
## religion_2022Baptist                                          0.8756
## religion_2022Catholic                                         0.3355
## religion_2022Muslim                                           0.0126 *
## religion_2022Other                                            0.7707
## religion_2022Other Christian                                  0.5919
## religion_2022Prefer not to answer                             0.9999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1274.5  on 1369  degrees of freedom
## Residual deviance: 1254.5  on 1356  degrees of freedom
## AIC: 1282.5
##
## Number of Fisher Scoring iterations: 13
```

```
chi_square_test <- anova(simplified_model, clean_model, test = "Chisq")
print(chi_square_test)
```

```
## Analysis of Deviance Table
##
## Model 1: lost ~ employment_status_2022 + gender_2022 + religion_2022
## Model 2: lost ~ highest_education_2022 + employment_status_2022 + household_income_2022 +
##     residence_area_2022 + gender_2022 + marital_status_2022 +
##     religion_2022
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1356     1254.5
## 2      1341     1234.4 15    20.08   0.1689
```

```
predicted_probs <- predict(simplified_model, type = "response")
predicted_class <- ifelse(predicted_probs > 0.5, 1, 0)

#
confusion_matrix <- table(Predicted = predicted_class, Actual = clean_data$lost)
print("Confusion Matrix:")
```

```
## [1] "Confusion Matrix:"
```

```r
print(confusion_matrix)
```

```
##          Actual
## Predicted    0    1
##         0 1129  241
```

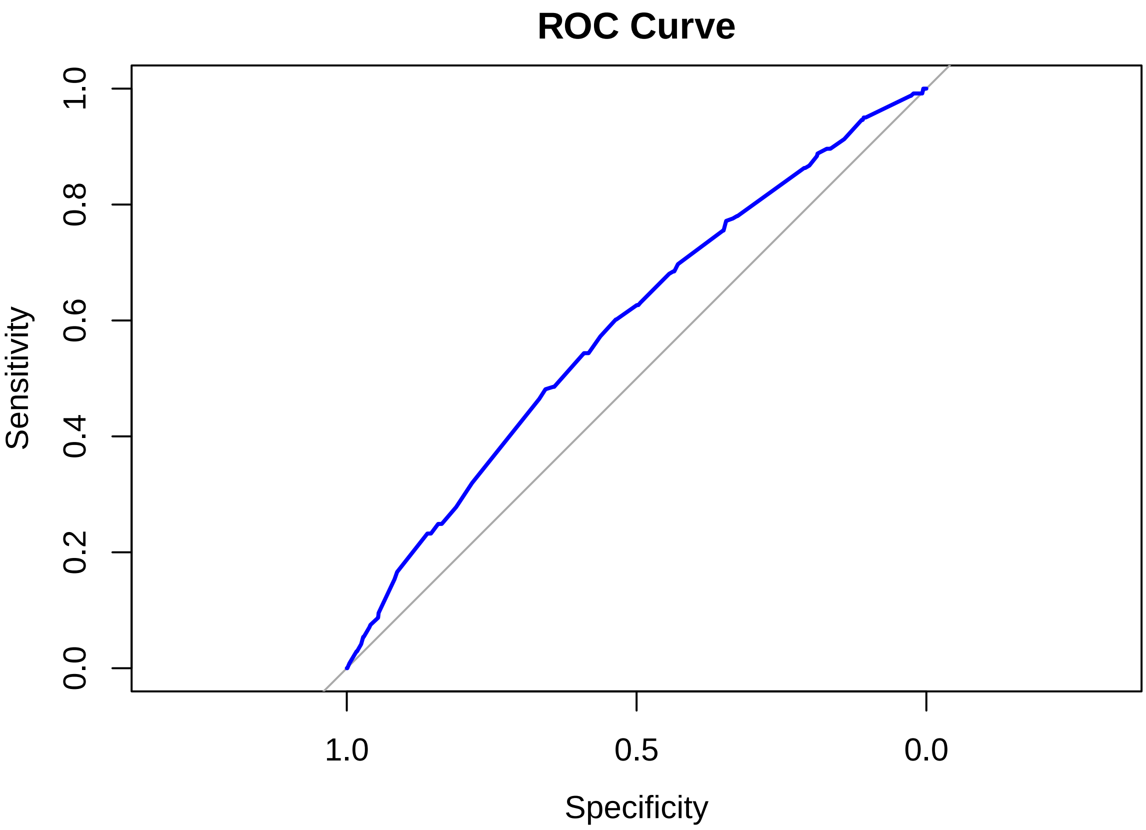
```r
#   ROC     AUC
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
roc_curve <- roc(clean_data$lost, predicted_probs)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```r
#   ROC
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)
```



```
auc_value <- auc(roc_curve)
cat("AUC:", auc_value, "\n")
```

```
## AUC: 0.5909463
```

```
#   McFadden's R^2
null_model <- glm(lost ~ 1, data = clean_data, family = binomial)  #
r2_mcfadden <- 1 - (logLik(clean_model) / logLik(null_model))
cat("McFadden's R^2:", r2_mcfadden, "\n")
```

```
## McFadden's R^2: 0.03146567
```

```
#   dependent variable
library(nnet)
multinom_model <- multinom(lost ~ highest_education_2022 + employment_status_2022 + household_income_20
                   residence_area_2022 + gender_2022 + marital_status_2022 + religion_2022, data = clea
```

```
## # weights:  31 (30 variable)
## initial  value 949.611637
## iter  10 value 626.110543
## iter  20 value 618.256318
## iter  30 value 617.283672
## iter  40 value 617.188491
## final  value 617.188359
## converged
```

```
summary(multinom_model)
```

```
## Warning in sqrt(diag(vc)): NaNs produced
```

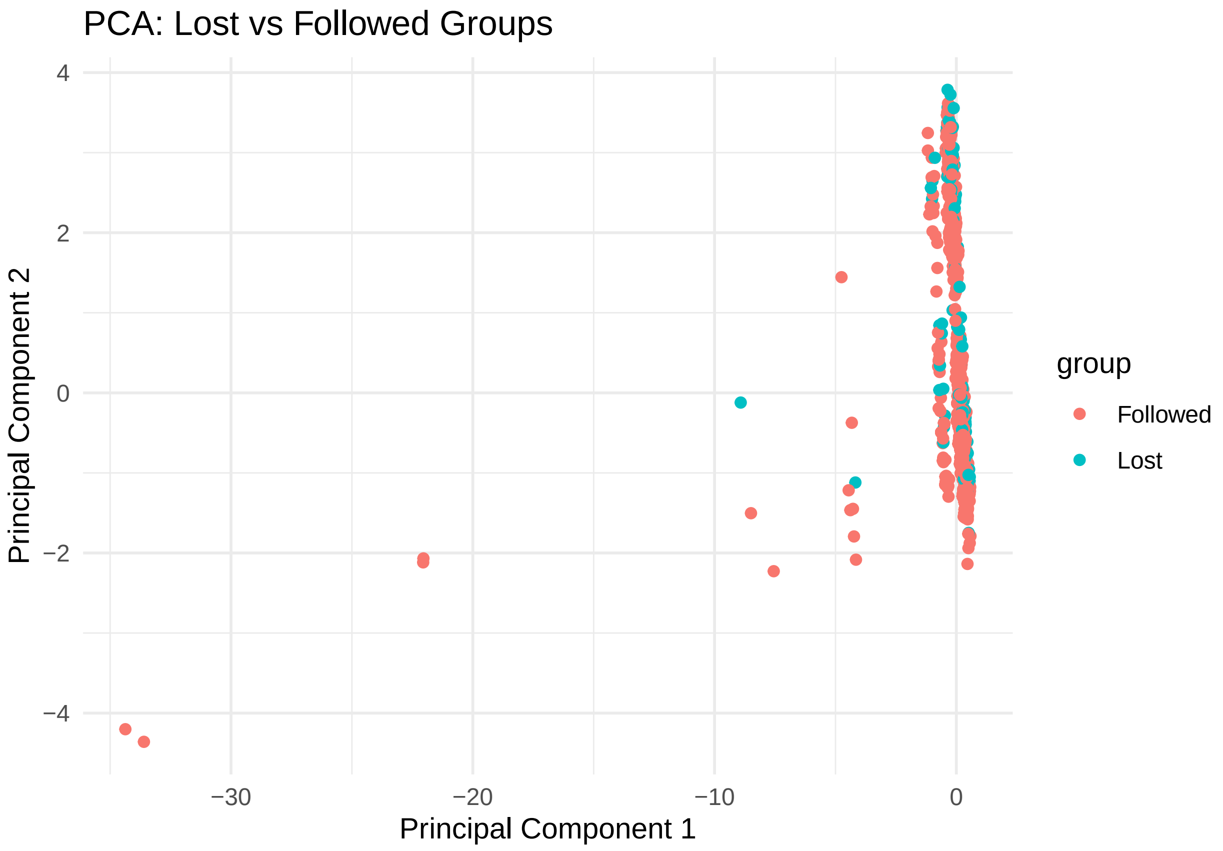
```
## Call:
## multinom(formula = lost ~ highest_education_2022 + employment_status_2022 +
##     household_income_2022 + residence_area_2022 + gender_2022 +
##     marital_status_2022 + religion_2022, data = clean_data)
##
## Coefficients:
##                                                                                             Values
## (Intercept)                                                                             0.14742349
## highest_education_2022No school/Did not complete primary                                0.17863717
## highest_education_2022Primary                                                           0.13830897
## highest_education_2022Secondary                                                         0.25152577
## employment_status_2022Employed full-time                                               -0.46865538
## employment_status_2022Employed part-time                                               -0.15640228
## employment_status_2022Not employed and not looking for work                            -0.15418654
## employment_status_2022Not employed but looking for work                                -0.31025301
## employment_status_2022Prefer not to answer                                            -28.30852483
## employment_status_2022Self-employed                                                    -0.19388660
## household_income_2022Allowed me to save just a little                                  -0.13368377
## household_income_2022Only just met my expenses                                         -0.36625146
## household_income_2022Prefer not to answer                                              -0.90209426
## household_income_2022Was not sufficient, so needed to use savings to meet expenses    -0.18930560
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses  -0.72225588
## residence_area_2022Trading Center (town)                                              -0.12988465
## residence_area_2022Village (rural)                                                    -0.22299043
## gender_2022Male                                                                       -0.34320861
## marital_status_2022Divorced/Separated                                                 -1.29763593
## marital_status_2022Married                                                            -1.08425834
## marital_status_2022Prefer not to answer                                               12.90474401
## marital_status_2022Single                                                             -0.84365666
## marital_status_2022Widowed                                                            -1.11199580
## religion_2022Baptist                                                                  -0.05016383
## religion_2022Catholic                                                                  0.27564002
## religion_2022Missing                                                                   0.00000000
## religion_2022Muslim                                                                    0.82238523
## religion_2022Other                                                                    -0.28742941
## religion_2022Other Christian                                                           0.14827778
## religion_2022Prefer not to answer                                                     -6.48798801
##                                                                                       Std. Err.
## (Intercept)                                                                          1.11203850
## highest_education_2022No school/Did not complete primary                             0.36247104
## highest_education_2022Primary                                                        0.20526193
## highest_education_2022Secondary                                                      0.18512010
## employment_status_2022Employed full-time                                             0.27554931
## employment_status_2022Employed part-time                                             0.43656848
## employment_status_2022Not employed and not looking for work                          0.34606262
## employment_status_2022Not employed but looking for work                              0.29186074
## employment_status_2022Prefer not to answer                                                  NaN
## employment_status_2022Self-employed                                                  0.23409971
## household_income_2022Allowed me to save just a little                                0.44518623
## household_income_2022Only just met my expenses                                       0.38771809
## household_income_2022Prefer not to answer                                            1.13530548
## household_income_2022Was not sufficient, so needed to use savings to meet expenses   0.41397104
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses 0.39218638
## residence_area_2022Trading Center (town)                                             0.19704736
## residence_area_2022Village (rural)                                                   0.19137808
## gender_2022Male                                                                      0.15658003
## marital_status_2022Divorced/Separated                                                1.03945919
## marital_status_2022Married                                                           0.96694511
## marital_status_2022Prefer not to answer                                              0.00002428
## marital_status_2022Single                                                            0.97932726
## marital_status_2022Widowed                                                           1.04604019
## religion_2022Baptist                                                                 0.80711844
## religion_2022Catholic                                                                0.27973636
## religion_2022Missing                                                                        NaN
## religion_2022Muslim                                                                  0.34868337
## religion_2022Other                                                                   0.80681282
## religion_2022Other Christian                                                         0.25908972
## religion_2022Prefer not to answer                                                    0.00000000
##
## Residual Deviance: 1234.377
## AIC: 1292.377
```

```
#   lost
clean_data$lost <- as.integer(clean_data$lost)
```

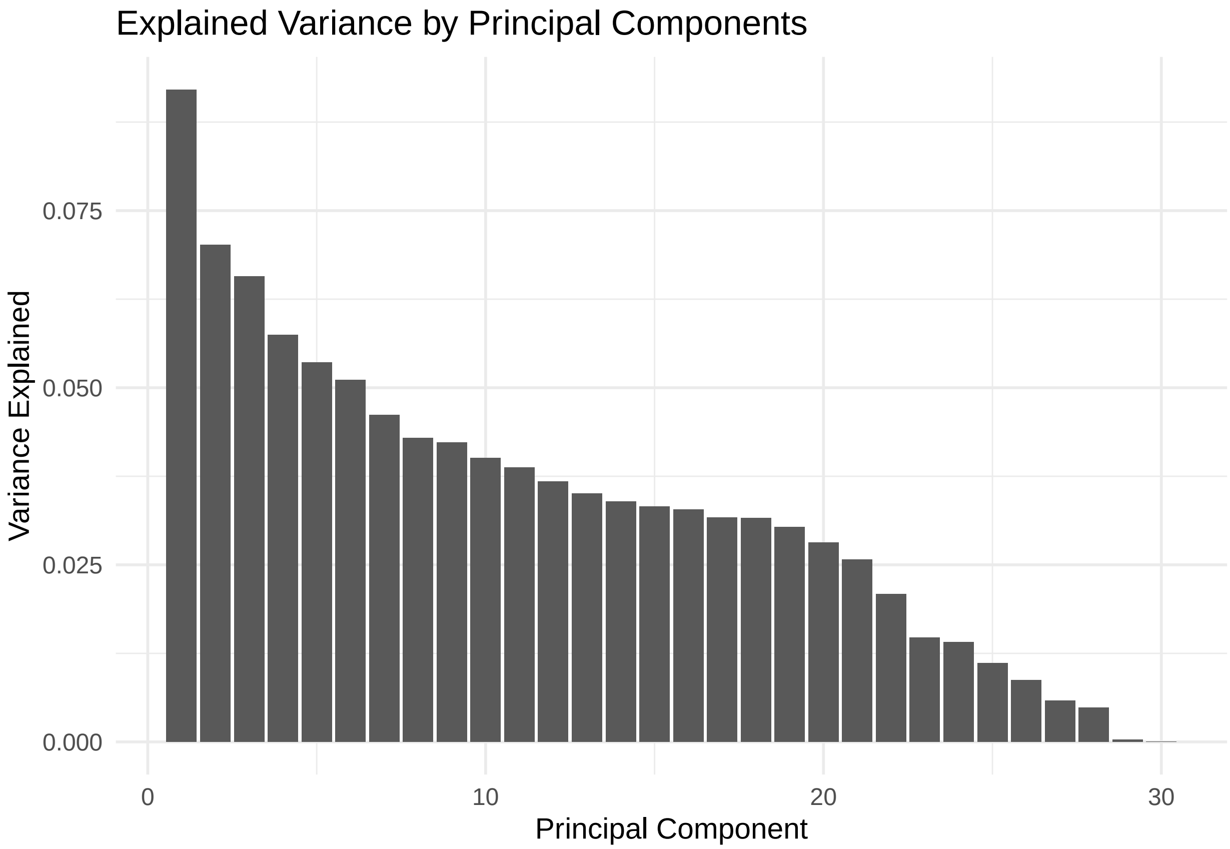
```r
poisson_model <- glm(lost ~ highest_education_2022 + employment_status_2022 + household_income_2022 +
                   residence_area_2022 + gender_2022 + marital_status_2022 + religion_2022,
                     data = clean_data, family = poisson)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = lost ~ highest_education_2022 + employment_status_2022 +
##     household_income_2022 + residence_area_2022 + gender_2022 +
##     marital_status_2022 + religion_2022, family = poisson, data = clean_data)
##
## Coefficients:
##                                                                                        Estimate
## (Intercept)                                                                            0.404624
## highest_education_2022No school/Did not complete primary                               0.017536
## highest_education_2022Primary                                                          0.013851
## highest_education_2022Secondary                                                        0.028802
## employment_status_2022Employed full-time                                             -0.057891
## employment_status_2022Employed part-time                                             -0.022087
## employment_status_2022Not employed and not looking for work                          -0.019981
## employment_status_2022Not employed but looking for work                              -0.039722
## employment_status_2022Prefer not to answer                                           -0.331398
## employment_status_2022Self-employed                                                  -0.024142
## household_income_2022Allowed me to save just a little                                -0.015098
## household_income_2022Only just met my expenses                                       -0.044596
## household_income_2022Prefer not to answer                                            -0.146204
## household_income_2022Was not sufficient, so needed to use savings to meet expenses   -0.020368
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses -0.084610
## residence_area_2022Trading Center (town)                                             -0.014970
## residence_area_2022Village (rural)                                                   -0.026408
## gender_2022Male                                                                      -0.041104
## marital_status_2022Divorced/Separated                                                -0.188995
## marital_status_2022Married                                                           -0.164314
## marital_status_2022Prefer not to answer                                               0.204352
## marital_status_2022Single                                                            -0.132949
## marital_status_2022Widowed                                                           -0.167990
## religion_2022Baptist                                                                 -0.002116
## religion_2022Catholic                                                                 0.032782
## religion_2022Muslim                                                                   0.110256
## religion_2022Other                                                                   -0.032703
## religion_2022Other Christian                                                          0.018042
## religion_2022Prefer not to answer                                                    -0.128150
##                                                                                      Std. Error
## (Intercept)                                                                             0.431344
## highest_education_2022No school/Did not complete primary                                0.128151
## highest_education_2022Primary                                                           0.070048
## highest_education_2022Secondary                                                         0.063864
## employment_status_2022Employed full-time                                                0.095577
## employment_status_2022Employed part-time                                                0.153799
## employment_status_2022Not employed and not looking for work                             0.124460
## employment_status_2022Not employed but looking for work                                 0.104176
## employment_status_2022Prefer not to answer                                              0.637558
## employment_status_2022Self-employed                                                     0.083958
## household_income_2022Allowed me to save just a little                                  0.163852
## household_income_2022Only just met my expenses                                         0.142435
## household_income_2022Prefer not to answer                                              0.356989
## household_income_2022Was not sufficient, so needed to use savings to meet expenses     0.152919
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses   0.142804
## residence_area_2022Trading Center (town)                                               0.069322
## residence_area_2022Village (rural)                                                     0.066792
## gender_2022Male                                                                        0.053472
## marital_status_2022Divorced/Separated                                                  0.405786
## marital_status_2022Married                                                             0.384972
## marital_status_2022Prefer not to answer                                                0.739394
## marital_status_2022Single                                                              0.389340
## marital_status_2022Widowed                                                             0.410603
## religion_2022Baptist                                                                   0.256564
## religion_2022Catholic                                                                  0.093980
## religion_2022Muslim                                                                    0.126133
## religion_2022Other                                                                     0.251828
## religion_2022Other Christian                                                           0.085867
## religion_2022Prefer not to answer                                                      0.914583
##                                                                                      z value
## (Intercept)                                                                            0.938
## highest_education_2022No school/Did not complete primary                               0.137
## highest_education_2022Primary                                                          0.198
## highest_education_2022Secondary                                                        0.451
## employment_status_2022Employed full-time                                              -0.606
## employment_status_2022Employed part-time                                              -0.144
## employment_status_2022Not employed and not looking for work                           -0.161
## employment_status_2022Not employed but looking for work                               -0.381
## employment_status_2022Prefer not to answer                                            -0.520
## employment_status_2022Self-employed                                                   -0.288
## household_income_2022Allowed me to save just a little                                 -0.092
## household_income_2022Only just met my expenses                                        -0.313
## household_income_2022Prefer not to answer                                             -0.410
## household_income_2022Was not sufficient, so needed to use savings to meet expenses    -0.133
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses  -0.592
## residence_area_2022Trading Center (town)                                              -0.216
## residence_area_2022Village (rural)                                                    -0.395
## gender_2022Male                                                                       -0.769
## marital_status_2022Divorced/Separated                                                 -0.466
## marital_status_2022Married                                                            -0.427
## marital_status_2022Prefer not to answer                                                0.276
## marital_status_2022Single                                                             -0.341
## marital_status_2022Widowed                                                            -0.409
## religion_2022Baptist                                                                  -0.008
## religion_2022Catholic                                                                  0.349
## religion_2022Muslim                                                                    0.874
## religion_2022Other                                                                    -0.130
## religion_2022Other Christian                                                           0.210
## religion_2022Prefer not to answer                                                     -0.140
##                                                                                      Pr(>|z|)
## (Intercept)                                                                             0.348
## highest_education_2022No school/Did not complete primary                                0.891
## highest_education_2022Primary                                                           0.843
## highest_education_2022Secondary                                                         0.652
## employment_status_2022Employed full-time                                                0.545
## employment_status_2022Employed part-time                                                0.886
## employment_status_2022Not employed and not looking for work                             0.872
## employment_status_2022Not employed but looking for work                                 0.703
## employment_status_2022Prefer not to answer                                              0.603
## employment_status_2022Self-employed                                                     0.774
## household_income_2022Allowed me to save just a little                                   0.927
## household_income_2022Only just met my expenses                                          0.754
## household_income_2022Prefer not to answer                                               0.682
## household_income_2022Was not sufficient, so needed to use savings to meet expenses      0.894
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses    0.554
## residence_area_2022Trading Center (town)                                                0.829
## residence_area_2022Village (rural)                                                      0.693
## gender_2022Male                                                                         0.442
## marital_status_2022Divorced/Separated                                                   0.641
## marital_status_2022Married                                                              0.670
## marital_status_2022Prefer not to answer                                                 0.782
## marital_status_2022Single                                                               0.733
## marital_status_2022Widowed                                                              0.682
## religion_2022Baptist                                                                    0.993
## religion_2022Catholic                                                                   0.727
## religion_2022Muslim                                                                     0.382
## religion_2022Other                                                                      0.897
## religion_2022Other Christian                                                            0.834
## religion_2022Prefer not to answer                                                       0.889
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 146.09  on 1369  degrees of freedom
## Residual deviance: 141.35  on 1341  degrees of freedom
## AIC: 3087.3
##
## Number of Fisher Scoring iterations: 4
```

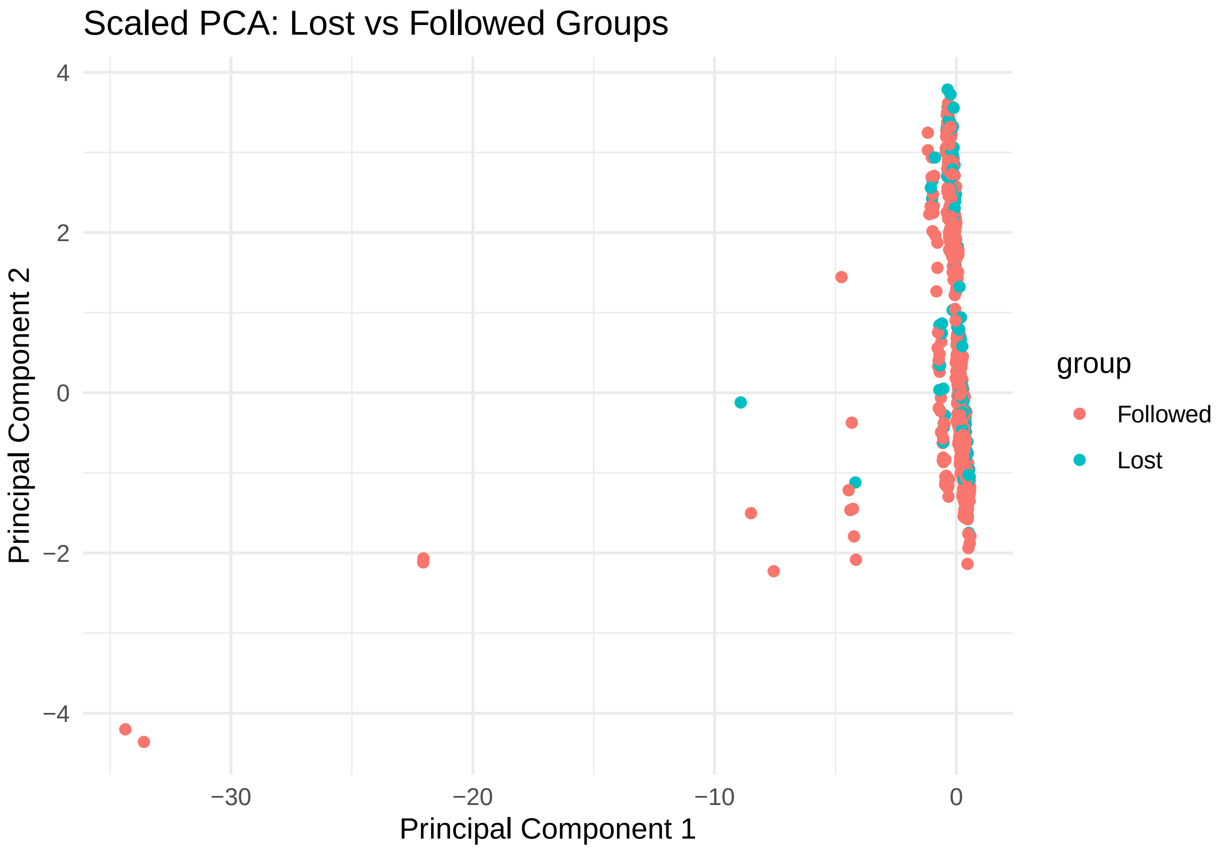
```r
# Load necessary libraries
library(dplyr)
library(ggplot2)

# Define the categorical variables for conversion to dummy variables
categorical_vars <- c("household_income_2022", "highest_education_2022", "employment_status_2022",
                      "residence_area_2022", "gender_2022", "marital_status_2022", "religion_2022")

# Combine Lost and Followed groups first and add group label
combined_data <- merged_data %>%
  mutate(group = if_else(lost == 1, "Lost", "Followed")) %>%
  select(all_of(categorical_vars), group) %>%
  filter(!if_any(all_of(categorical_vars), ~ . == "Missing"))

# Convert categorical variables to factors
combined_data <- combined_data %>%
  mutate(across(all_of(categorical_vars), as.factor))

# Apply model.matrix to the combined dataset (convert categorical variables to dummy variables)
combined_data_clean <- model.matrix(~ . - 1, data = combined_data) %>%
  as.data.frame()

# Add group column back to the cleaned data
combined_data_clean$group <- combined_data$group

# Remove the group column before running PCA
combined_data_for_pca <- combined_data_clean %>%
  select(-group)

# Remove columns with zero variance (constant columns)
combined_data_for_pca <- combined_data_for_pca[, apply(combined_data_for_pca, 2, var) != 0]

# Perform PCA on the cleaned data, scaling the variables
combined_pca <- prcomp(scale(combined_data_for_pca), center = TRUE, scale. = TRUE)

# Extract the first two principal components and add group labels back
pca_df <- as.data.frame(combined_pca$x[, 1:2])
pca_df$group <- combined_data_clean$group

# Plot the PCA results using ggplot2
ggplot(pca_df, aes(x = PC1, y = PC2, color = group)) +
  geom_point() +
  labs(title = "PCA: Lost vs Followed Groups", x = "Principal Component 1", y = "Principal Component 2"
  theme_minimal()
```



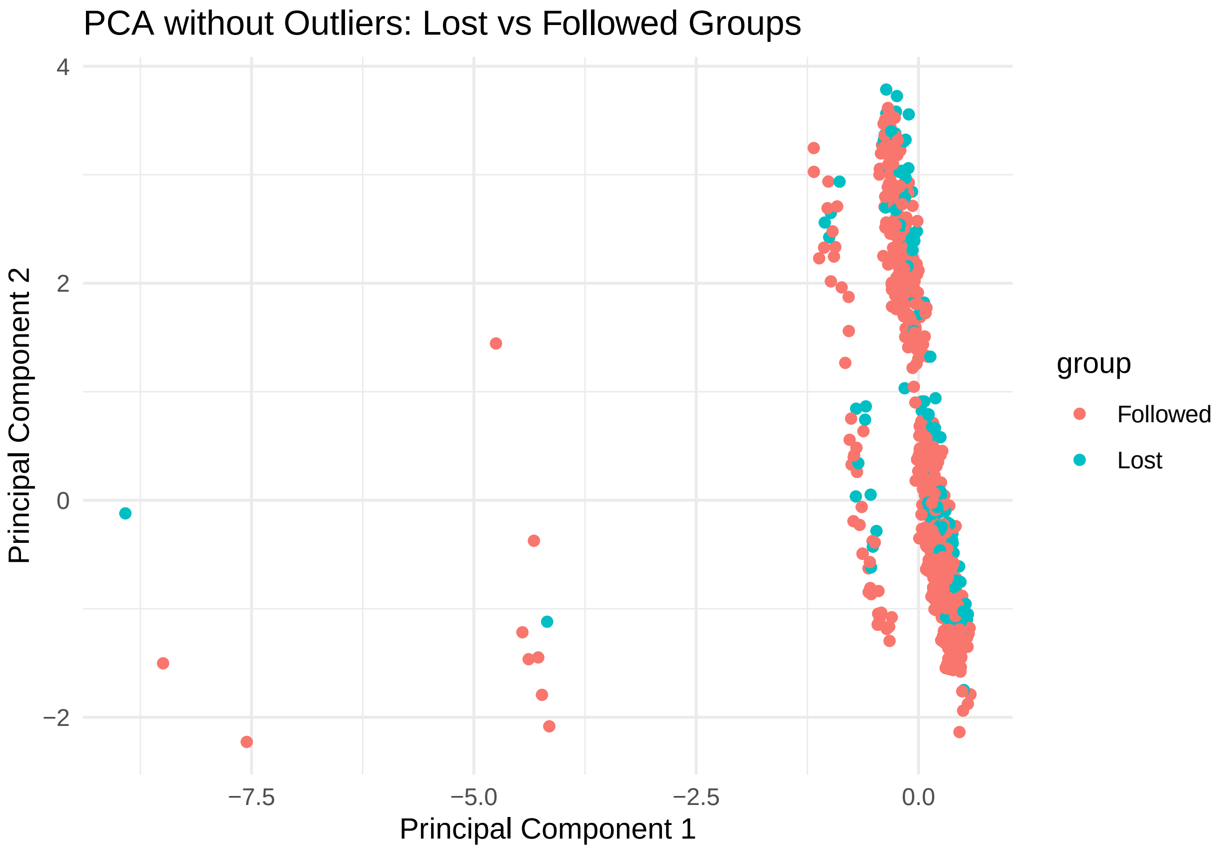
```r
# Initial setup for categorical variables
categorical_vars <- c("household_income_2022", "highest_education_2022", "employment_status_2022",
                      "residence_area_2022", "gender_2022", "marital_status_2022", "religion_2022")

# Combine Lost and Followed groups first and add group label
combined_data <- merged_data %>%
  mutate(group = if_else(lost == 1, "Lost", "Followed")) %>%
  select(all_of(categorical_vars), group) %>%
  filter(!if_any(all_of(categorical_vars), ~ . == "Missing"))

# Convert categorical variables to factors
combined_data <- combined_data %>%
  mutate(across(all_of(categorical_vars), as.factor))

# Apply model.matrix to the combined dataset (with consistent dummy variables for both groups)
combined_data_clean <- model.matrix(~ . - 1, data = combined_data) %>%
  as.data.frame()

# Add group column back to the cleaned data
combined_data_clean$group <- combined_data$group

# Remove the group column before running PCA
combined_data_for_pca <- combined_data_clean %>%
  select(-group)

# Identify and remove columns with zero variance
combined_data_for_pca <- combined_data_for_pca[, apply(combined_data_for_pca, 2, var) != 0]

# 1. Standardize the data
combined_data_for_pca_scaled <- scale(combined_data_for_pca)

# 2. Perform PCA on the scaled data
combined_pca_scaled <- prcomp(combined_data_for_pca_scaled, center = TRUE, scale. = TRUE)

# 3. Calculate explained variance
explained_variance <- combined_pca_scaled$sdev^2 / sum(combined_pca_scaled$sdev^2)

# 4. Plot the explained variance for each principal component
explained_variance_df <- data.frame(
  PC = seq_along(explained_variance),
  Variance = explained_variance
)

ggplot(explained_variance_df, aes(x = PC, y = Variance)) +
  geom_bar(stat = "identity") +
  labs(title = "Explained Variance by Principal Components", x = "Principal Component", y = "Variance E
  theme_minimal()
```



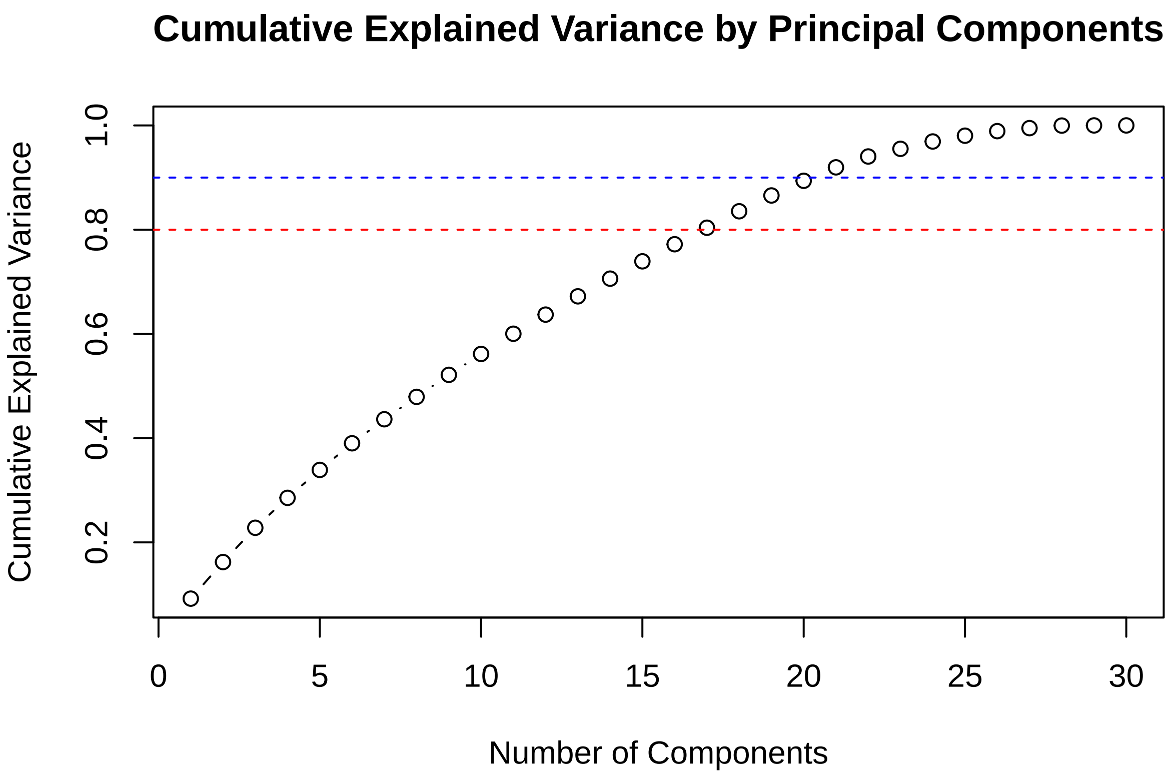
```r
# 5. Extract the first two principal components and plot PCA
pca_df_scaled <- as.data.frame(combined_pca_scaled$x[, 1:2])
pca_df_scaled$group <- combined_data_clean$group

# Plot PCA results
ggplot(pca_df_scaled, aes(x = PC1, y = PC2, color = group)) +
  geom_point() +
  labs(title = "Scaled PCA: Lost vs Followed Groups", x = "Principal Component 1", y = "Principal Compo
  theme_minimal()
```



```
# 6. Identify outliers based on PCA results
outliers <- pca_df_scaled %>%
  filter(PC1 < -10 | PC2 < -10)

# 7. Remove outliers and re-plot PCA without outliers
pca_df_cleaned <- pca_df_scaled %>%
  filter(PC1 > -10 & PC2 > -10)  # Assuming -10 is the threshold for outliers

# Re-plot PCA without outliers
ggplot(pca_df_cleaned, aes(x = PC1, y = PC2, color = group)) +
  geom_point() +
  labs(title = "PCA without Outliers: Lost vs Followed Groups", x = "Principal Component 1", y = "Princi
  theme_minimal()
```



```r
# Calculate the proportion of variance explained by each component
explained_variance <- combined_pca_scaled$sdev^2 / sum(combined_pca_scaled$sdev^2)

# Calculate the cumulative explained variance
cumulative_explained_variance <- cumsum(explained_variance)

# Plot the cumulative explained variance
plot(cumulative_explained_variance, type = "b", xlab = "Number of Components", ylab = "Cumulative Expla
     main = "Cumulative Explained Variance by Principal Components")

# Add a horizontal line for 80% explained variance
abline(h = 0.80, col = "red", lty = 2)
abline(h = 0.90, col = "blue", lty = 2)
```



```r
# Determine how many components explain at least 80% variance
components_80 <- which(cumulative_explained_variance >= 0.80)[1]
components_90 <- which(cumulative_explained_variance >= 0.90)[1]

print(paste("Number of components to retain for 80% variance:", components_80))
```

```
## [1] "Number of components to retain for 80% variance: 17"
```

```r
print(paste("Number of components to retain for 90% variance:", components_90))
```

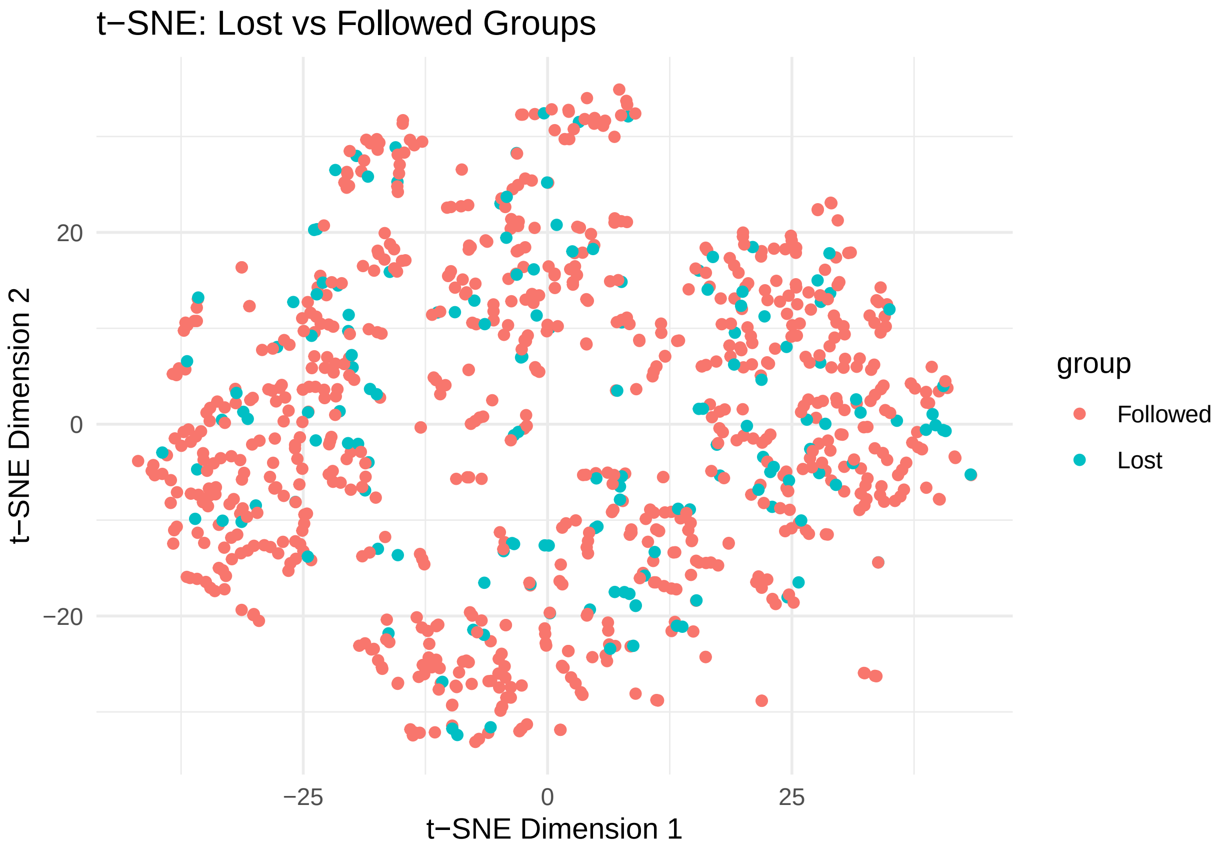
```
## [1] "Number of components to retain for 90% variance: 21"
```

```r
# Convert categorical variables to dummy variables and remove rows containing "Missing"
combined_data <- merged_data %>%
  mutate(group = if_else(lost == 1, "Lost", "Followed")) %>%
  select(all_of(categorical_vars), group) %>%
  filter(!if_any(all_of(categorical_vars), ~ . == "Missing")) %>%
  mutate(across(all_of(categorical_vars), as.factor))

# Apply model.matrix to convert the categorical variables into dummy variables
combined_data_clean <- model.matrix(~ . - 1, data = combined_data) %>%
  as.data.frame()

# Remove duplicate rows before running t-SNE
combined_data_clean <- combined_data_clean %>%
  distinct()
```

```r
# Extract the group information for later plotting
group_labels <- combined_data$group[1:nrow(combined_data_clean)]  # Ensure it matches the reduced datas

# Perform t-SNE on the dummy variables, setting a perplexity value (typically between 5 and 50)
set.seed(42)  # Set seed for reproducibility
tsne_results <- Rtsne(as.matrix(combined_data_clean), dims = 2, perplexity = 10, verbose = TRUE, max_ite
```

```
## Performing PCA
## Read the 912 x 31 data matrix successfully!
## OpenMP is working. 1 threads.
## Using no_dims = 2, perplexity = 10.000000, and theta = 0.500000
## Computing input similarities...
## Building tree...
## Done in 0.04 seconds (sparsity = 0.042960)!
## Learning embedding...
## Iteration 50: error is 82.378375 (50 iterations in 0.05 seconds)
## Iteration 100: error is 82.378373 (50 iterations in 0.06 seconds)
## Iteration 150: error is 82.378331 (50 iterations in 0.05 seconds)
## Iteration 200: error is 82.377128 (50 iterations in 0.05 seconds)
## Iteration 250: error is 82.344514 (50 iterations in 0.05 seconds)
## Iteration 300: error is 2.132151 (50 iterations in 0.05 seconds)
## Iteration 350: error is 1.838930 (50 iterations in 0.04 seconds)
## Iteration 400: error is 1.736508 (50 iterations in 0.04 seconds)
## Iteration 450: error is 1.690701 (50 iterations in 0.04 seconds)
## Iteration 500: error is 1.662501 (50 iterations in 0.04 seconds)
## Iteration 550: error is 1.642442 (50 iterations in 0.04 seconds)
## Iteration 600: error is 1.628258 (50 iterations in 0.04 seconds)
## Iteration 650: error is 1.620069 (50 iterations in 0.04 seconds)
## Iteration 700: error is 1.612360 (50 iterations in 0.04 seconds)
## Iteration 750: error is 1.606821 (50 iterations in 0.04 seconds)
## Iteration 800: error is 1.601778 (50 iterations in 0.04 seconds)
## Iteration 850: error is 1.597690 (50 iterations in 0.04 seconds)
## Iteration 900: error is 1.594818 (50 iterations in 0.04 seconds)
## Iteration 950: error is 1.591143 (50 iterations in 0.04 seconds)
## Iteration 1000: error is 1.589428 (50 iterations in 0.04 seconds)
## Fitting performed in 0.87 seconds.
```

```r
# Convert t-SNE results into a data frame for plotting
tsne_df <- as.data.frame(tsne_results$Y)
colnames(tsne_df) <- c("Dim1", "Dim2")
tsne_df$group <- group_labels

# Plot the t-SNE results using ggplot2
ggplot(tsne_df, aes(x = Dim1, y = Dim2, color = group)) +
  geom_point() +
  labs(title = "t-SNE: Lost vs Followed Groups", x = "t-SNE Dimension 1", y = "t-SNE Dimension 2") +
  theme_minimal()
```


t−SNE: Lost vs Followed Groups
t−SNE Dimension 2
20
0
−20
−25
0
25
t−SNE Dimension 1
group
● Followed
● Lost