biostats_consulting

2024-10-05

Contents

```
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
```

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

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

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

```
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 n
data_2022_cleaned <- data_2022_cleaned %>%
  mutate(across(everything(), ~str_replace(., "Prefer not to answer \\[do not read aloud\\]", "Prefer n
```

```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), "\n
```

```
## 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
```

```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", "_2024
  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,
    lost,
    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 26
##   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 21 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>, lost <dbl>, ...
```

```
# 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>
```

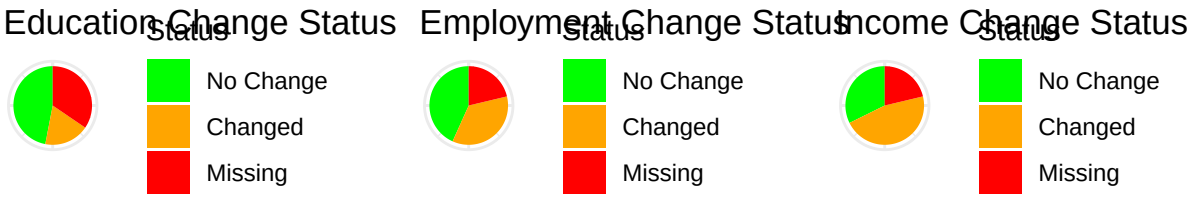
```
dfSummary(merged_data) %>% view()
```

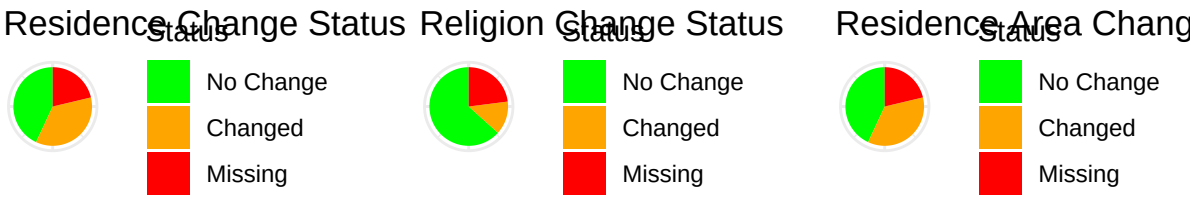
```
## Switching method to 'browser'
```

```
## Output file written: C:\Users\ghlas\AppData\Local\Temp\RtmpaKV2K9\file726431784a2a.html
```

```
# 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
```






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

# 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
```

```r
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")
```


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

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

```
## [1] 1355   24
```

```r
#
clean_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 = binomial)

#
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)                                                                      0.16260
## highest_education_2022No school/Did not complete primary                         0.15757
## highest_education_2022Primary                                                    0.10408
## highest_education_2022Secondary                                                  0.24229
## employment_status_2022Employed full-time                                        -0.50641
## employment_status_2022Employed part-time                                        -0.17216
## employment_status_2022Not employed and not looking for work                     -0.17032
## employment_status_2022Not employed but looking for work                         -0.31643
## employment_status_2022Self-employed                                             -0.20427
## household_income_2022Allowed me to save just a little                           -0.13422
## household_income_2022Only just met my expenses                                  -0.36793
## household_income_2022Was not sufficient, so needed to use savings to meet expenses  -0.19081
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses -0.72285
## residence_area_2022Trading Center (town)                                        -0.10725
## residence_area_2022Village (rural)                                              -0.20211
## gender_2022Male                                                                 -0.33725
## marital_status_2022Divorced/Separated                                           -1.29280
## marital_status_2022Married                                                      -1.08380
## marital_status_2022Single                                                       -0.83922
## marital_status_2022Widowed                                                      -1.10961
## religion_2022Baptist                                                            -0.05864
## religion_2022Catholic                                                            0.25375
## religion_2022Muslim                                                              0.81719
## religion_2022Other                                                              -0.28386
## religion_2022Other Christian                                                     0.14611
##                                                                                 Std. Error
## (Intercept)                                                                        1.11331
## highest_education_2022No school/Did not complete primary                           0.36237
## highest_education_2022Primary                                                      0.20618
## highest_education_2022Secondary                                                    0.18509
## employment_status_2022Employed full-time                                           0.27678
```

```
## employment_status_2022Employed part-time                                    0.43655
## employment_status_2022Not employed and not looking for work                  0.34615
## employment_status_2022Not employed but looking for work                      0.29209
## employment_status_2022Self-employed                                          0.23436
## household_income_2022Allowed me to save just a little                        0.44529
## household_income_2022Only just met my expenses                               0.38783
## household_income_2022Was not sufficient, so needed to use savings to meet expenses     0.41409
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses    0.39229
## residence_area_2022Trading Center (town)                                     0.19803
## residence_area_2022Village (rural)                                           0.19222
## gender_2022Male                                                              0.15676
## marital_status_2022Divorced/Separated                                        1.04053
## marital_status_2022Married                                                   0.96814
## marital_status_2022Single                                                    0.98052
## marital_status_2022Widowed                                                   1.04714
## religion_2022Baptist                                                         0.80724
## religion_2022Catholic                                                        0.28044
## religion_2022Muslim                                                          0.34880
## religion_2022Other                                                           0.80672
## religion_2022Other Christian                                                 0.25927
##                                                                           z value
## (Intercept)                                                                 0.146
## highest_education_2022No school/Did not complete primary                    0.435
## highest_education_2022Primary                                               0.505
## highest_education_2022Secondary                                             1.309
## employment_status_2022Employed full-time                                   -1.830
## employment_status_2022Employed part-time                                   -0.394
## employment_status_2022Not employed and not looking for work                -0.492
## employment_status_2022Not employed but looking for work                    -1.083
## employment_status_2022Self-employed                                        -0.872
## household_income_2022Allowed me to save just a little                      -0.301
## household_income_2022Only just met my expenses                             -0.949
## household_income_2022Was not sufficient, so needed to use savings to meet expenses   -0.461
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses -1.843
## residence_area_2022Trading Center (town)                                   -0.542
## residence_area_2022Village (rural)                                         -1.051
## gender_2022Male                                                            -2.151
## marital_status_2022Divorced/Separated                                      -1.242
## marital_status_2022Married                                                 -1.119
## marital_status_2022Single                                                  -0.856
## marital_status_2022Widowed                                                 -1.060
## religion_2022Baptist                                                       -0.073
## religion_2022Catholic                                                       0.905
## religion_2022Muslim                                                         2.343
## religion_2022Other                                                         -0.352
## religion_2022Other Christian                                                0.564
##                                                                         Pr(>|z|)
## (Intercept)                                                               0.8839
## highest_education_2022No school/Did not complete primary                  0.6637
## highest_education_2022Primary                                             0.6137
## highest_education_2022Secondary                                           0.1905
## employment_status_2022Employed full-time                                  0.0673
## employment_status_2022Employed part-time                                  0.6933
## employment_status_2022Not employed and not looking for work               0.6227
```

```
## employment_status_2022Not employed but looking for work                                0.2787
## employment_status_2022Self-employed                                                     0.3834
## household_income_2022Allowed me to save just a little                                   0.7631
## household_income_2022Only just met my expenses                                          0.3428
## household_income_2022Was not sufficient, so needed to use savings to meet expenses     0.6449
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses   0.0654
## residence_area_2022Trading Center (town)                                                0.5881
## residence_area_2022Village (rural)                                                      0.2930
## gender_2022Male                                                                         0.0314
## marital_status_2022Divorced/Separated                                                   0.2141
## marital_status_2022Married                                                              0.2629
## marital_status_2022Single                                                               0.3921
## marital_status_2022Widowed                                                              0.2893
## religion_2022Baptist                                                                    0.9421
## religion_2022Catholic                                                                   0.3656
## religion_2022Muslim                                                                     0.0191
## religion_2022Other                                                                      0.7249
## religion_2022Other Christian                                                            0.5731
##
## (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_2022Self-employed
## household_income_2022Allowed me to save just a little
## household_income_2022Only just met my expenses
## 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_2022Single
## marital_status_2022Widowed
## religion_2022Baptist
## religion_2022Catholic
## religion_2022Muslim                                                                    *
## religion_2022Other
## religion_2022Other Christian
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1262.5  on 1354  degrees of freedom
## Residual deviance: 1228.4  on 1330  degrees of freedom
## AIC: 1278.4
##
## Number of Fisher Scoring iterations: 4
```

```r
#
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 Std. Error
## (Intercept)                                         -1.3379     0.3117
## employment_status_2022Employed full-time            -0.4745     0.2549
## employment_status_2022Employed part-time            -0.2226     0.4219
## employment_status_2022Not employed and not looking for work  -0.1757     0.3414
## employment_status_2022Not employed but looking for work      -0.2913     0.2865
## employment_status_2022Self-employed                 -0.1939     0.2282
## gender_2022Male                                     -0.3529     0.1503
## religion_2022Baptist                                -0.1386     0.8005
## religion_2022Catholic                                0.2488     0.2781
## religion_2022Muslim                                  0.8379     0.3405
## religion_2022Other                                  -0.2427     0.7957
## religion_2022Other Christian                         0.1281     0.2570
##                                                    z value Pr(>|z|)
## (Intercept)                                         -4.293 1.76e-05
## employment_status_2022Employed full-time            -1.862   0.0626
## employment_status_2022Employed part-time            -0.528   0.5978
## employment_status_2022Not employed and not looking for work  -0.515   0.6067
## employment_status_2022Not employed but looking for work      -1.017   0.3092
## employment_status_2022Self-employed                 -0.850   0.3954
## gender_2022Male                                     -2.348   0.0189
## religion_2022Baptist                                -0.173   0.8625
## religion_2022Catholic                                0.895   0.3709
## religion_2022Muslim                                  2.460   0.0139
## religion_2022Other                                  -0.305   0.7603
## religion_2022Other Christian                         0.499   0.6180
##
## (Intercept)                                        ***
## 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_2022Self-employed
## gender_2022Male                                    *
## religion_2022Baptist
## religion_2022Catholic
## religion_2022Muslim                                *
## religion_2022Other
## religion_2022Other Christian
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1262.5  on 1354  degrees of freedom
## Residual deviance: 1244.9  on 1343  degrees of freedom
## AIC: 1268.9
##
## Number of Fisher Scoring iterations: 4
```

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

#
summary(interaction_model)
```

```
##
## Call:
## glm(formula = lost ~ gender_2022 * employment_status_2022 + religion_2022,
##     family = binomial, data = clean_data)
##
## Coefficients:
##                                                                    Estimate
## (Intercept)                                                        -1.35608
## gender_2022Male                                                    -0.29340
## employment_status_2022Employed full-time                           -0.48519
## employment_status_2022Employed part-time                           -0.51987
## employment_status_2022Not employed and not looking for work        -0.03423
## employment_status_2022Not employed but looking for work            -0.26067
## employment_status_2022Self-employed                                -0.14170
## religion_2022Baptist                                               -0.12398
## religion_2022Catholic                                               0.23753
## religion_2022Muslim                                                 0.80447
## religion_2022Other                                                 -0.24950
## religion_2022Other Christian                                        0.11865
## gender_2022Male:employment_status_2022Employed full-time            0.02154
## gender_2022Male:employment_status_2022Employed part-time            0.62007
## gender_2022Male:employment_status_2022Not employed and not looking for work -0.72602
## gender_2022Male:employment_status_2022Not employed but looking for work     -0.06141
## gender_2022Male:employment_status_2022Self-employed                -0.11466
##                                                                    Std. Error
## (Intercept)                                                           0.35948
## gender_2022Male                                                       0.40078
## employment_status_2022Employed full-time                              0.35529
## employment_status_2022Employed part-time                              0.61141
## employment_status_2022Not employed and not looking for work           0.40679
## employment_status_2022Not employed but looking for work               0.35764
## employment_status_2022Self-employed                                   0.31280
## religion_2022Baptist                                                  0.80124
## religion_2022Catholic                                                 0.27854
## religion_2022Muslim                                                   0.34288
## religion_2022Other                                                    0.79787
## religion_2022Other Christian                                          0.25739
```

```
## gender_2022Male:employment_status_2022Employed full-time                      0.50991
## gender_2022Male:employment_status_2022Employed part-time                      0.84639
## gender_2022Male:employment_status_2022Not employed and not looking for work   0.89701
## gender_2022Male:employment_status_2022Not employed but looking for work       0.63854
## gender_2022Male:employment_status_2022Self-employed                           0.45900
##                                                                             z value
## (Intercept)                                                                  -3.772
## gender_2022Male                                                              -0.732
## employment_status_2022Employed full-time                                     -1.366
## employment_status_2022Employed part-time                                     -0.850
## employment_status_2022Not employed and not looking for work                  -0.084
## employment_status_2022Not employed but looking for work                      -0.729
## employment_status_2022Self-employed                                          -0.453
## religion_2022Baptist                                                         -0.155
## religion_2022Catholic                                                         0.853
## religion_2022Muslim                                                           2.346
## religion_2022Other                                                           -0.313
## religion_2022Other Christian                                                  0.461
## gender_2022Male:employment_status_2022Employed full-time                      0.042
## gender_2022Male:employment_status_2022Employed part-time                      0.733
## gender_2022Male:employment_status_2022Not employed and not looking for work  -0.809
## gender_2022Male:employment_status_2022Not employed but looking for work      -0.096
## gender_2022Male:employment_status_2022Self-employed                          -0.250
##                                                                            Pr(>|z|)
## (Intercept)                                                                0.000162
## gender_2022Male                                                            0.464125
## employment_status_2022Employed full-time                                   0.172061
## employment_status_2022Employed part-time                                   0.395168
## employment_status_2022Not employed and not looking for work                0.932939
## employment_status_2022Not employed but looking for work                    0.466093
## employment_status_2022Self-employed                                        0.650538
## religion_2022Baptist                                                       0.877033
## religion_2022Catholic                                                      0.393800
## religion_2022Muslim                                                        0.018965
## religion_2022Other                                                         0.754507
## religion_2022Other Christian                                               0.644829
## gender_2022Male:employment_status_2022Employed full-time                   0.966303
## gender_2022Male:employment_status_2022Employed part-time                   0.463801
## gender_2022Male:employment_status_2022Not employed and not looking for work 0.418294
## gender_2022Male:employment_status_2022Not employed but looking for work    0.923379
## gender_2022Male:employment_status_2022Self-employed                        0.802746
##
## (Intercept)                                                                ***
## gender_2022Male
## 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_2022Self-employed
## religion_2022Baptist
## religion_2022Catholic
## religion_2022Muslim                                                        *
## religion_2022Other
## religion_2022Other Christian
```

```
## gender_2022Male:employment_status_2022Employed full-time
## gender_2022Male:employment_status_2022Employed part-time
## gender_2022Male:employment_status_2022Not employed and not looking for work
## gender_2022Male:employment_status_2022Not employed but looking for work
## gender_2022Male:employment_status_2022Self-employed
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1262.5  on 1354  degrees of freedom
## Residual deviance: 1243.2  on 1338  degrees of freedom
## AIC: 1277.2
##
## Number of Fisher Scoring iterations: 4
```

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

#
table(Predicted = predicted_class, Actual = clean_data$lost)
```

```
##          Actual
## Predicted    0    1
##         0 1115  239
##         1    1    0
```

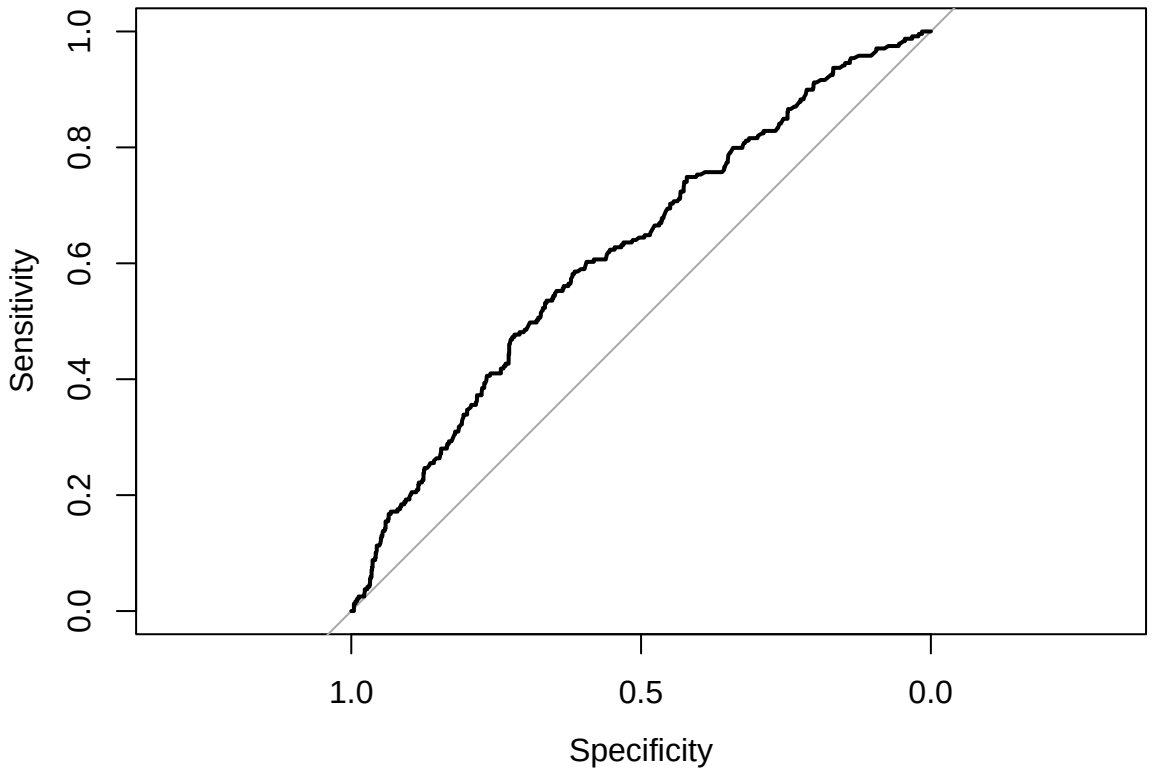
```
#   ROC     AUC
library(pROC)
```

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

##
## Attaching package: 'pROC'

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

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

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

## Setting direction: controls < cases
```

```
plot(roc_curve)
```




```
auc(roc_curve)
```

```
## Area under the curve: 0.6211
```

```
#   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 939.214430
## iter  10 value 621.340835
## iter  20 value 614.317828
## iter  30 value 614.188971
## iter  30 value 614.188966
## iter  30 value 614.188966
## final  value 614.188966
## 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)                                                                  0.16295972
## highest_education_2022No school/Did not complete primary                     0.15758613
## highest_education_2022Primary                                                0.10407742
## highest_education_2022Secondary                                              0.24228313
## employment_status_2022Employed full-time                                    -0.50643993
## employment_status_2022Employed part-time                                    -0.17219708
## employment_status_2022Not employed and not looking for work                 -0.17033527
## employment_status_2022Not employed but looking for work                     -0.31644164
## employment_status_2022Prefer not to answer                                   0.00000000
## employment_status_2022Self-employed                                         -0.20425349
## household_income_2022Allowed me to save just a little                       -0.13420993
## household_income_2022Only just met my expenses                              -0.36792811
## household_income_2022Prefer not to answer                                    0.00000000
## household_income_2022Was not sufficient, so needed to use savings to meet expenses   -0.19079327
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses -0.72285485
## residence_area_2022Trading Center (town)                                    -0.10726892
## residence_area_2022Village (rural)                                          -0.20207390
## gender_2022Male                                                             -0.33725979
## marital_status_2022Divorced/Separated                                       -1.29319128
## marital_status_2022Married                                                  -1.08415441
## marital_status_2022Prefer not to answer                                      0.00000000
## marital_status_2022Single                                                   -0.83961175
## marital_status_2022Widowed                                                  -1.10998621
## religion_2022Baptist                                                        -0.05825361
## religion_2022Catholic                                                        0.25374674
## religion_2022Missing                                                         0.00000000
## religion_2022Muslim                                                          0.81717549
## religion_2022Other                                                          -0.28443478
## religion_2022Other Christian                                                 0.14611230
## religion_2022Prefer not to answer                                            0.00000000
##                                                                            Std. Err.
## (Intercept)                                                             1.113310e+00
## highest_education_2022No school/Did not complete primary                3.623686e-01
## highest_education_2022Primary                                           2.061766e-01
## highest_education_2022Secondary                                         1.850948e-01
## employment_status_2022Employed full-time                                2.767788e-01
## employment_status_2022Employed part-time                                4.365552e-01
## employment_status_2022Not employed and not looking for work             3.461481e-01
## employment_status_2022Not employed but looking for work                 2.920866e-01
## employment_status_2022Prefer not to answer                              1.275556e-14
## employment_status_2022Self-employed                                     2.343646e-01
## household_income_2022Allowed me to save just a little                   4.452875e-01
## household_income_2022Only just met my expenses                          3.878314e-01
## household_income_2022Prefer not to answer                               2.447367e-15
## household_income_2022Was not sufficient, so needed to use savings to meet expenses   4.140911e-01
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses 3.922920e-01
## residence_area_2022Trading Center (town)                                1.980283e-01
## residence_area_2022Village (rural)                                      1.922163e-01
## gender_2022Male                                                         1.567609e-01
## marital_status_2022Divorced/Separated                                   1.040531e+00
```

```
## marital_status_2022Married                                                    9.681430e-01
## marital_status_2022Prefer not to answer                                                NaN
## marital_status_2022Single                                                     9.805222e-01
## marital_status_2022Widowed                                                    1.047139e+00
## religion_2022Baptist                                                          8.071520e-01
## religion_2022Catholic                                                         2.804434e-01
## religion_2022Missing                                                          0.000000e+00
## religion_2022Muslim                                                           3.488047e-01
## religion_2022Other                                                            8.068714e-01
## religion_2022Other Christian                                                  2.592672e-01
## religion_2022Prefer not to answer                                             0.000000e+00
##
## Residual Deviance: 1228.378
## AIC: 1278.378
```

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

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)
```

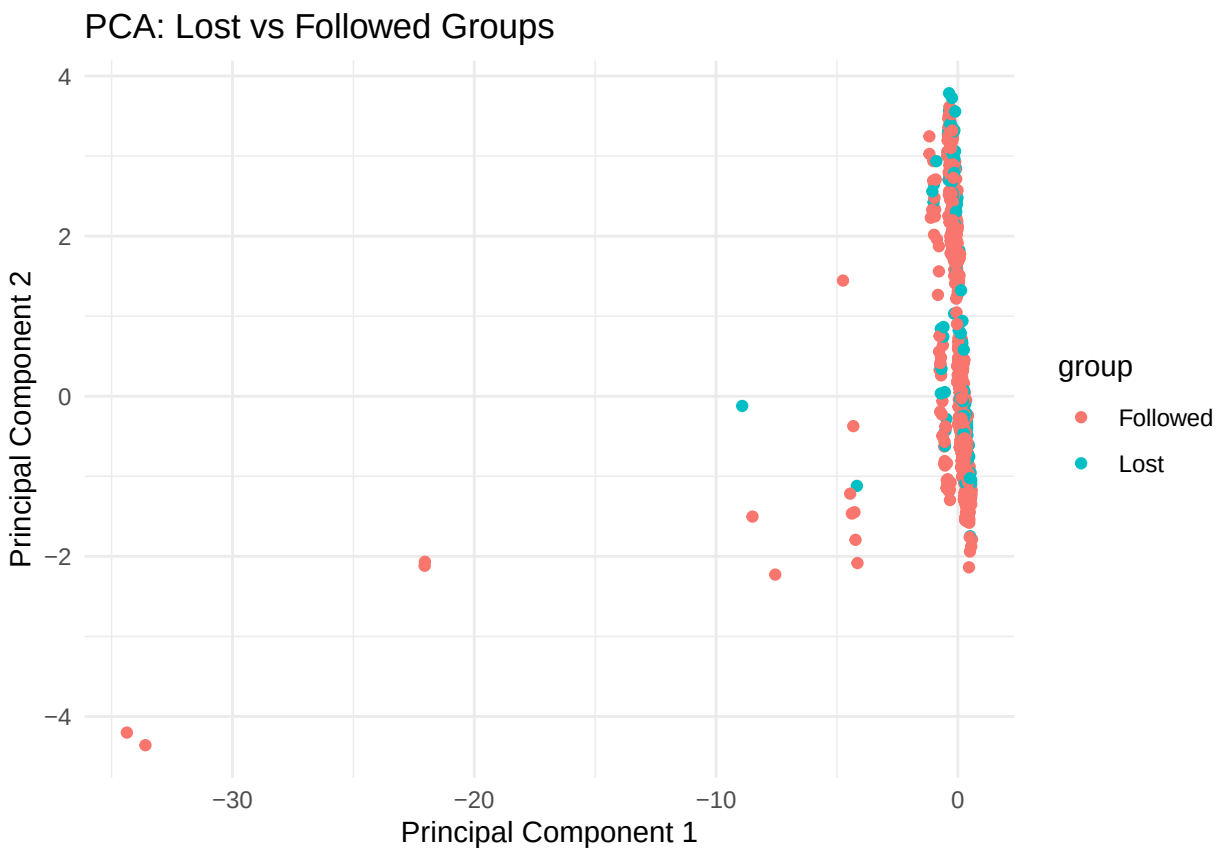
```
##
## Call:
## glm(formula = lost ~ highest_education_2022 + employment_status_2022 +
##     household_income_2022 + residence_area_2022 + gender_2022 +
##     marital_status_2022 + religion_2022, family = poisson, data = clean_data)
##
## Coefficients:
##                                                                                Estimate
## (Intercept)                                                                    0.406132
## highest_education_2022No school/Did not complete primary                       0.019448
## highest_education_2022Primary                                                  0.011270
## highest_education_2022Secondary                                                0.029129
## employment_status_2022Employed full-time                                      -0.061389
## employment_status_2022Employed part-time                                      -0.022701
## employment_status_2022Not employed and not looking for work                   -0.021388
## employment_status_2022Not employed but looking for work                       -0.040076
## employment_status_2022Self-employed                                           -0.025398
## household_income_2022Allowed me to save just a little                         -0.015212
## household_income_2022Only just met my expenses                                -0.045056
## household_income_2022Was not sufficient, so needed to use savings to meet expenses   -0.021028
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses -0.086325
## residence_area_2022Trading Center (town)                                      -0.013402
## residence_area_2022Village (rural)                                            -0.024416
## gender_2022Male                                                               -0.040160
## marital_status_2022Divorced/Separated                                         -0.188213
## marital_status_2022Married                                                    -0.164394
## marital_status_2022Single                                                     -0.131953
## marital_status_2022Widowed                                                    -0.167709
## religion_2022Baptist                                                          -0.003128
## religion_2022Catholic                                                          0.030158
## religion_2022Muslim                                                            0.109352
```

```
## religion_2022Other                                                          -0.032729
## religion_2022Other Christian                                                 0.017502
##                                                                           Std. Error
## (Intercept)                                                                 0.431507
## highest_education_2022No school/Did not complete primary                    0.129192
## highest_education_2022Primary                                               0.070392
## highest_education_2022Secondary                                             0.064078
## employment_status_2022Employed full-time                                    0.096090
## employment_status_2022Employed part-time                                    0.153965
## employment_status_2022Not employed and not looking for work                 0.124674
## employment_status_2022Not employed but looking for work                     0.104549
## employment_status_2022Self-employed                                         0.084324
## household_income_2022Allowed me to save just a little                       0.163861
## household_income_2022Only just met my expenses                              0.142455
## household_income_2022Was not sufficient, so needed to use savings to meet expenses    0.152950
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses   0.142879
## residence_area_2022Trading Center (town)                                    0.069792
## residence_area_2022Village (rural)                                          0.067058
## gender_2022Male                                                             0.053617
## marital_status_2022Divorced/Separated                                       0.405791
## marital_status_2022Married                                                  0.384979
## marital_status_2022Single                                                   0.389363
## marital_status_2022Widowed                                                  0.410620
## religion_2022Baptist                                                        0.256657
## religion_2022Catholic                                                       0.094339
## religion_2022Muslim                                                         0.126301
## religion_2022Other                                                          0.251908
## religion_2022Other Christian                                                0.086132
##                                                                             z value
## (Intercept)                                                                   0.941
## highest_education_2022No school/Did not complete primary                      0.151
## highest_education_2022Primary                                                 0.160
## highest_education_2022Secondary                                               0.455
## employment_status_2022Employed full-time                                     -0.639
## employment_status_2022Employed part-time                                     -0.147
## employment_status_2022Not employed and not looking for work                  -0.172
## employment_status_2022Not employed but looking for work                      -0.383
## employment_status_2022Self-employed                                          -0.301
## household_income_2022Allowed me to save just a little                        -0.093
## household_income_2022Only just met my expenses                               -0.316
## household_income_2022Was not sufficient, so needed to use savings to meet expenses    -0.137
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses  -0.604
## residence_area_2022Trading Center (town)                                     -0.192
## residence_area_2022Village (rural)                                           -0.364
## gender_2022Male                                                              -0.749
## marital_status_2022Divorced/Separated                                        -0.464
## marital_status_2022Married                                                   -0.427
## marital_status_2022Single                                                    -0.339
## marital_status_2022Widowed                                                   -0.408
## religion_2022Baptist                                                         -0.012
## religion_2022Catholic                                                         0.320
## religion_2022Muslim                                                           0.866
## religion_2022Other                                                           -0.130
## religion_2022Other Christian                                                  0.203
```

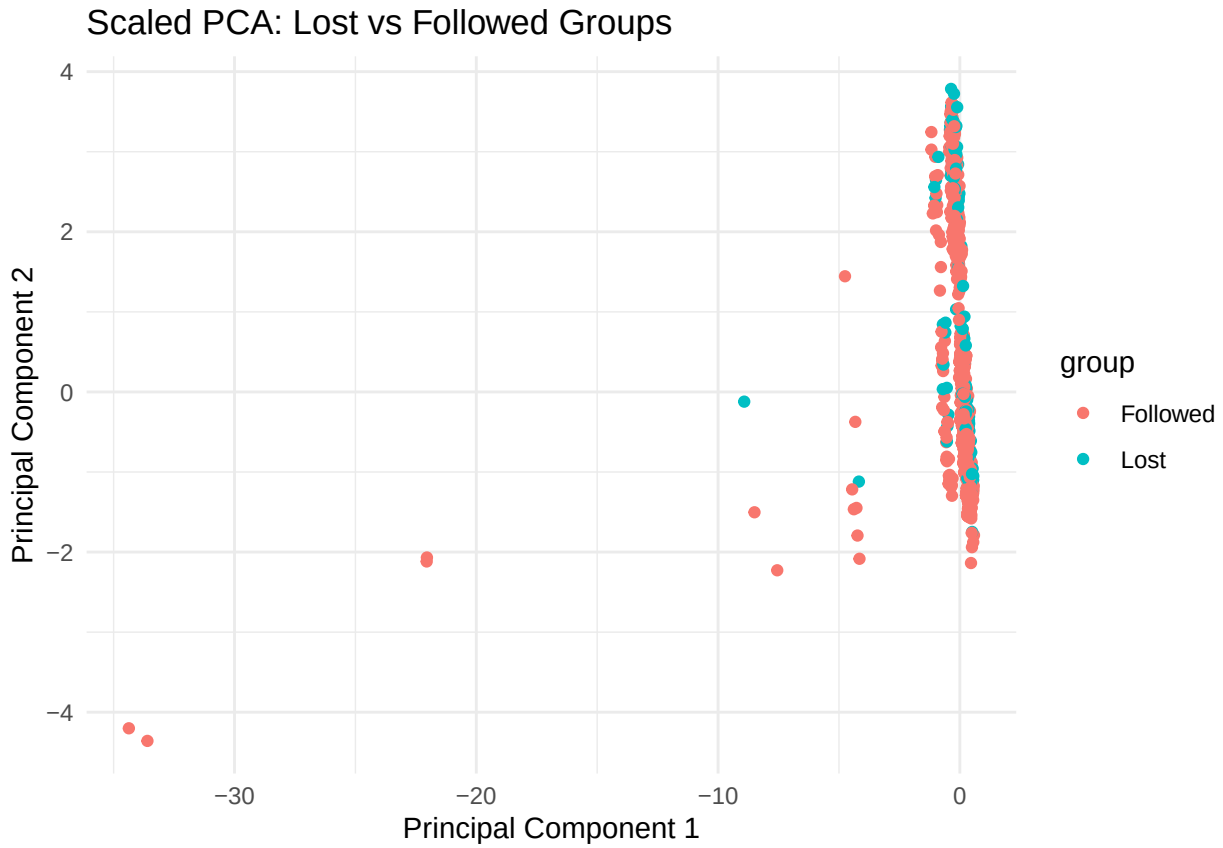
```
##                                                                                 Pr(>|z|)
## (Intercept)                                                                        0.347
## highest_education_2022No school/Did not complete primary                          0.880
## highest_education_2022Primary                                                      0.873
## highest_education_2022Secondary                                                    0.649
## employment_status_2022Employed full-time                                           0.523
## employment_status_2022Employed part-time                                           0.883
## employment_status_2022Not employed and not looking for work                        0.864
## employment_status_2022Not employed but looking for work                            0.701
## employment_status_2022Self-employed                                                0.763
## household_income_2022Allowed me to save just a little                              0.926
## household_income_2022Only just met my expenses                                     0.752
## household_income_2022Was not sufficient, so needed to use savings to meet expenses 0.891
## household_income_2022Was really not sufficient, so needed to borrow to meet expenses 0.546
## residence_area_2022Trading Center (town)                                           0.848
## residence_area_2022Village (rural)                                                 0.716
## gender_2022Male                                                                    0.454
## marital_status_2022Divorced/Separated                                              0.643
## marital_status_2022Married                                                         0.669
## marital_status_2022Single                                                          0.735
## marital_status_2022Widowed                                                         0.683
## religion_2022Baptist                                                               0.990
## religion_2022Catholic                                                              0.749
## religion_2022Muslim                                                                0.387
## religion_2022Other                                                                 0.897
## religion_2022Other Christian                                                       0.839
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 144.77  on 1354  degrees of freedom
## Residual deviance: 140.49  on 1330  degrees of freedom
## AIC: 3047.2
##
## 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)
```

```r
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()
```

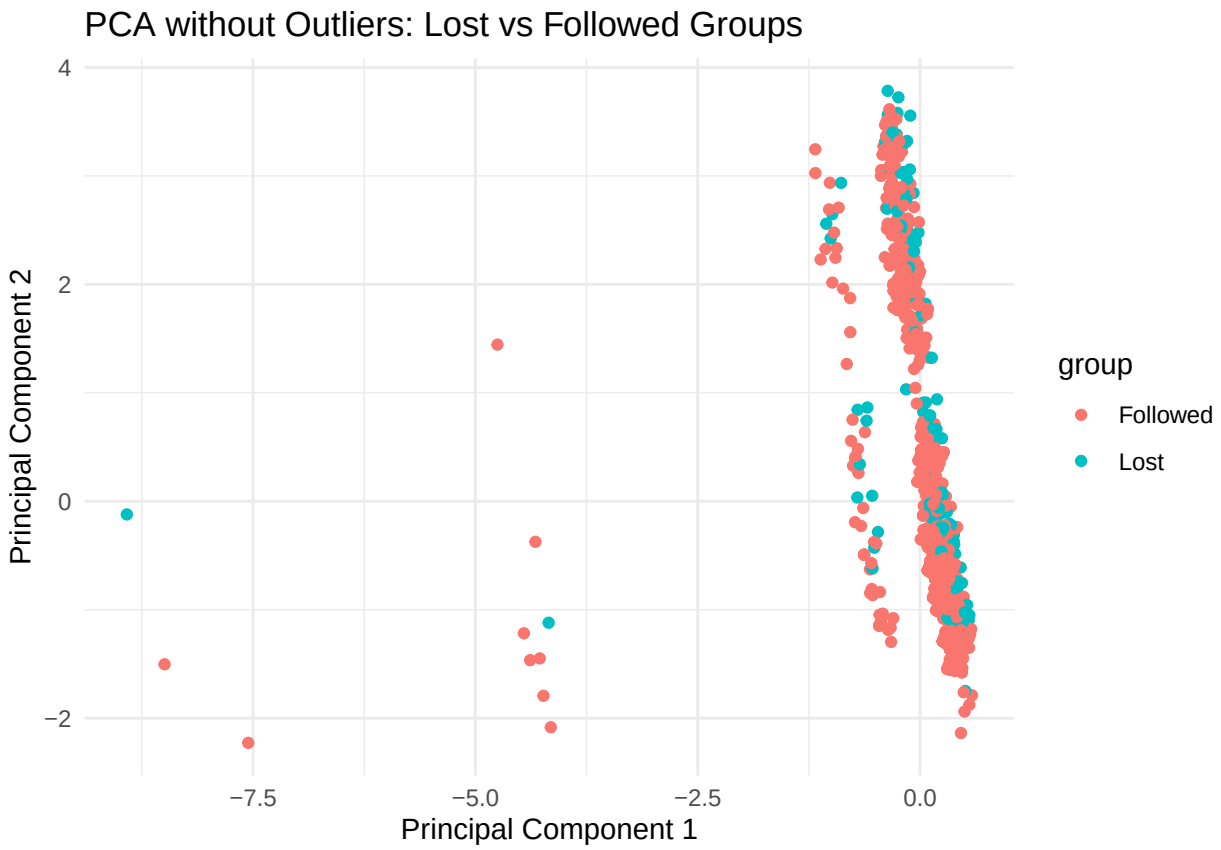


```r
# 1.
combined_data_for_pca_scaled <- scale(combined_data_for_pca)

# 2.   PCA
combined_pca_scaled <- prcomp(combined_data_for_pca_scaled, center = TRUE, scale. = TRUE)

# 3.        PCA
pca_df_scaled <- as.data.frame(combined_pca_scaled$x[, 1:2])
pca_df_scaled$group <- combined_data_clean$group

#     PCA
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()
```




```r
# 4.
outliers <- pca_df_scaled %>%
  filter(PC1 < -10 | PC2 < -10)

# 5.           PCA
pca_df_cleaned <- pca_df_scaled %>%
  filter(PC1 > -10 & PC2 > -10)  #   -10

#       PCA
ggplot(pca_df_cleaned, aes(x = PC1, y = PC2, color = group)) +
```

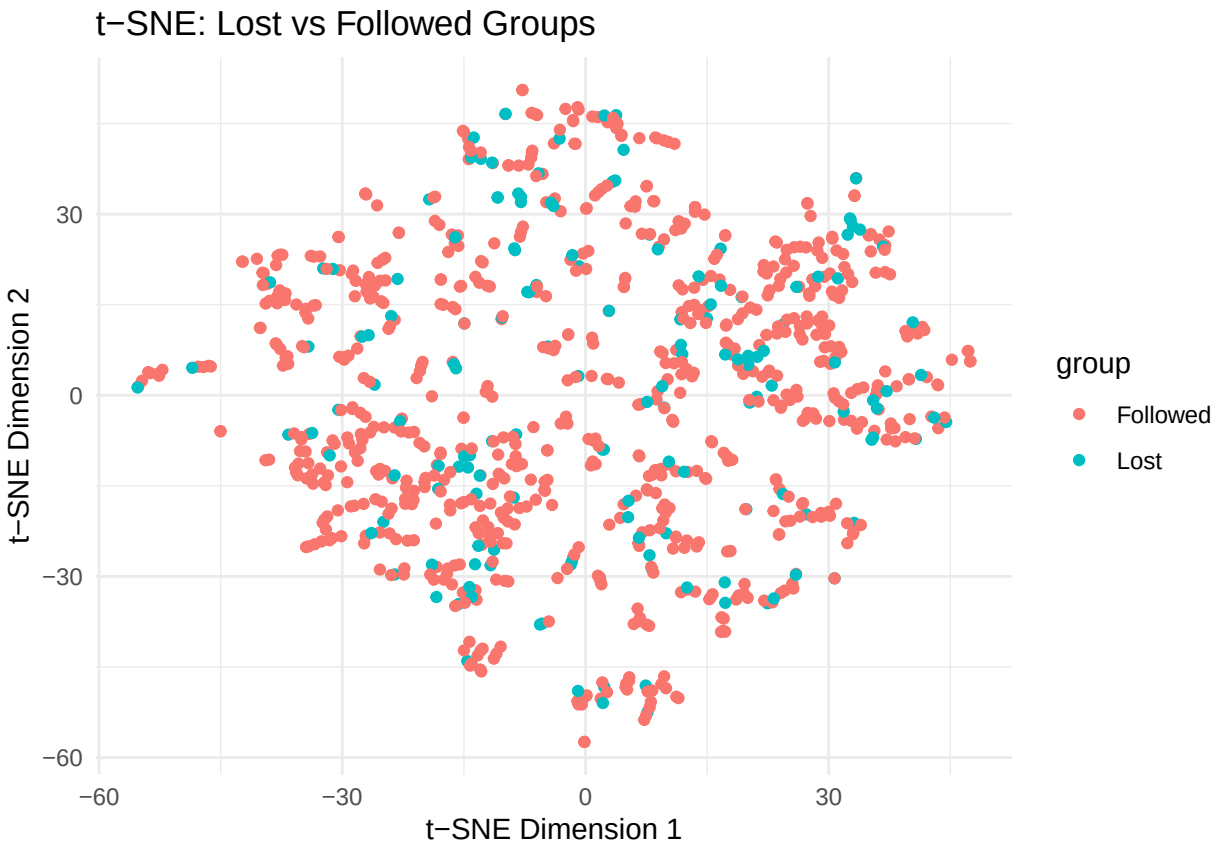
```
  geom_point() +
  labs(title = "PCA without Outliers: Lost vs Followed Groups", x = "Principal Component 1", y = "Princi
  theme_minimal()
```




```
# 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()

# 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 = 5, verbose = TRUE, max_ite
```

```
## Performing PCA
## Read the 912 x 31 data matrix successfully!
## OpenMP is working. 1 threads.
## Using no_dims = 2, perplexity = 5.000000, and theta = 0.500000
## Computing input similarities...
## Building tree...
## Done in 0.03 seconds (sparsity = 0.021723)!
## Learning embedding...
## Iteration 50: error is 90.078650 (50 iterations in 0.05 seconds)
## Iteration 100: error is 87.183430 (50 iterations in 0.04 seconds)
## Iteration 150: error is 85.864474 (50 iterations in 0.04 seconds)
## Iteration 200: error is 85.308474 (50 iterations in 0.04 seconds)
## Iteration 250: error is 85.007271 (50 iterations in 0.03 seconds)
## Iteration 300: error is 2.199073 (50 iterations in 0.03 seconds)
## Iteration 350: error is 1.846683 (50 iterations in 0.04 seconds)
## Iteration 400: error is 1.738662 (50 iterations in 0.04 seconds)
## Iteration 450: error is 1.695251 (50 iterations in 0.04 seconds)
## Iteration 500: error is 1.667561 (50 iterations in 0.04 seconds)
## Iteration 550: error is 1.647817 (50 iterations in 0.04 seconds)
## Iteration 600: error is 1.632739 (50 iterations in 0.04 seconds)
## Iteration 650: error is 1.618047 (50 iterations in 0.04 seconds)
## Iteration 700: error is 1.607665 (50 iterations in 0.04 seconds)
## Iteration 750: error is 1.598405 (50 iterations in 0.04 seconds)
## Iteration 800: error is 1.590489 (50 iterations in 0.04 seconds)
## Iteration 850: error is 1.586424 (50 iterations in 0.04 seconds)
## Iteration 900: error is 1.581909 (50 iterations in 0.04 seconds)
## Iteration 950: error is 1.578473 (50 iterations in 0.04 seconds)
## Iteration 1000: error is 1.576210 (50 iterations in 0.04 seconds)
## Fitting performed in 0.76 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
30
0
−30
−60
−60
−30
0
30
t−SNE Dimension 1
group
Followed
Lost