Profiling Youth Sexual Risk in Nigeria: A Machine Learning Approach Using Multiple Indicator Cluster Survey (MICS6) Data

Nasir Umar

2025-05-15

Contents

Introduction

```
This report presents a machine learning analysis of youth sexual risk and media
engagement patterns in Nigeria using data from the Multiple Indicator Cluster Survey
(MICS6). Young people aged 15-24 were selected from the full MICS dataset
```

```
and analysed across two thematic tracks:

Track 1: Patterns of access to digital and traditional media platforms
(e.g., radio, TV, internet, mobile phones).

Track 2: Sexual risk behaviors, focusing on condom use and nature of last sexual partner
(spousal vs. non-spousal).

We apply unsupervised learning models to identify naturally occurring
clusters within each track:
K-Means clustering is used for Track 1 (Media Engagement),
Partitioning Around Medoids (PAM) with Gower Distance is used for Track 2
(Sexual Risk Profiling), to accommodate mixed data types.

We explore how demographic characteristics such as sex and wealth intersect with
these clusters - and whether digitally engaged youth are more at risk.
```

Load & Prepare Data

- Load needed libraries

```r
#   Phase 1: Project Setup & Data Loading
library(haven)              # mics is in spss (.sav) file, read raaw data
#   Phase 2: Data Cleaning, Wrangling, and Preparation
library(tidyverse)          # data cleaning & wrangling
library(data.table)         # data dataframe object handling & manipulation
library(labelled)           # labels & values
library(Hmisc)              # Initial data view - values and label
library(mice)               # to handle missing data, imputation
library(dplyr)              # data prep
library(tidyr)              # data prep
#   Phase 3: Exploratory Data Analysis (EDA) / Descriptives
library(modelsummary)       # descriptive tables
library(kableExtra)         # custom tables
library(sjPlot)             # model visualizations or descriptive plots
library(ggplot2)            # plots, graphs
library(patchwork)          # to combine multiple ggplot2
library(gt)                 # tables spec
#   Phase 4: Main Machine Learning Analysis & Model Evaluation
library(tidymodels)         # machine learning
library(caret)              # machine learning model training
library(yardstick)          # calculate performance metrics for machine learning models
#   Phase 5: Reporting and Output Generation
library(flextable)          # custom tables with r markdown
library(knitr)
```

```r
dir.create("figures", showWarnings = FALSE)
dir.create("tables", showWarnings = FALSE)
dir.create("output", showWarnings = FALSE)
```

Load, Subset, Clean & Organise Data for Analysis

- Data Setup & Loading

```
# Data source: MICS6 2023 (Year of interview 2021)
# Data files:
    # mn.sav (males)
    # wm.sav (females)

mics6_mn <- read_sav("Data/mn.sav")
mics6_wn <- read_sav("Data/wm.sav")
```

```
view_df(mics6_mn)
glimpse(mics6_mn)
label(mics6_mn)

view_df(mics6_wn)
view(mics6_wn)
```

Select Var (males) for Analysis

```
mics6_mn_selected <- mics6_mn %>%
  mutate(
    cluster_id = MWM1,
    household_id = MWM2,
    man_id = MWM3,
    household_uid = paste(cluster_id,household_id, sep = "_"),
    man_uid = paste(cluster_id, household_id, man_id, sep = "_"),
    age_male = MWB4,
    education_ever = MWB5,
    education_level = MWB6A,
    sch_now = MWB9,
    sch_now_level = MWB10A,
    read_can = MWB14,
    state = HH7,
    age_cat = MWAGE,
    newspaper = MMT1,
    radio = MMT2,
    tv = MMT3,
    computer_tab = MMT5,
    internet = MMT10,
    mphone_own = MMT11,
    mphone = MMT12,
    sex_last12months = MSB7,
    condom_use = MSB8,
    relations_sexpartner = MSB9,
    urban_rural = HH6,
    education_male = mwelevel,
    survey_weight = mnweight,
    strata = stratum,
    wealth_score = wscore,
```

```
    wealth_quintile = windex5) %>%
  select(cluster_id, household_id, man_id, household_uid,
         man_uid, age_male, education_ever, education_level,
         sch_now, sch_now_level, read_can, state, age_cat,
         newspaper, radio, tv, computer_tab,  internet,
         mphone_own,    mphone, sex_last12months, condom_use,
         relations_sexpartner, urban_rural, education_male,
         survey_weight, strata, wealth_score, wealth_quintile)
```

Data Wrangling Male Subset Dataset

```
mn_clean <- mics6_mn_selected %>%
  mutate(
    newspaper = na_if(newspaper, 9),
    radio = na_if(radio, 9),
    tv = na_if(tv, 9),
    computer_tab = na_if(computer_tab, 9),
    internet = na_if(internet, 9),
    mphone = na_if(mphone, 9),

    mphone_own = case_when(
      mphone_own == 1 ~ 1,
      mphone_own == 2 ~ 0,
      TRUE ~ NA_real_
    ),

    education_ever = na_if(education_ever, 9),
    sch_now = na_if(sch_now, 9),
    sch_now_level = na_if(sch_now_level, 99),
    read_can = na_if(read_can, 9),

    urban_rural = case_when(
      urban_rural == 1 ~ "Urban",
      urban_rural == 2 ~ "Rural",
      TRUE ~ NA_character_
    ),

    sex_last12months = case_when(
      sex_last12months == 1 ~ 1,
      sex_last12months == 2 ~ 0,
      TRUE ~ NA_real_
    ),

    condom_use = case_when(
      condom_use == 1 ~ 0,
      condom_use == 2 ~ 1,
      TRUE ~ NA_real_
    ),

    relations_sexpartner = case_when(
      relations_sexpartner %in% c(4, 5, 6) ~ 1,
```

```
      relations_sexpartner %in% c(1, 2, 3) ~ 0,
      TRUE ~ NA_real_
    ),


    education_male = factor(education_male,
                            levels = c(0, 1, 2, 3, 4, 9),
                            labels = c("None", "Primary", "Jnr Secondary", "Snr Secondary", "Higher", "

    wealth_quintile = factor(wealth_quintile,
                             levels = 1:5,
                             labels = c("Poorest", "Second", "Middle", "Fourth", "Richest"))
  )

mn_clean <- mn_clean %>%
  filter(age_cat %in% 1:2) %>%
  mutate(sex = "Male")
```

Select Var (females) for Analysis

```
mics6_wn_selected <- mics6_wn %>%
  mutate(
    cluster_id = WM1,
    household_id = WM2,
    woman_id = WM3,
    household_uid = paste(cluster_id,household_id, sep = "_"),
    man_uid = paste(cluster_id, household_id, woman_id, sep = "_"),
    age_female = WB4,
    education_ever = WB5,
    education_level = WB6A,
    sch_now = WB9,
    sch_now_level = WB10A,
    read_can = WB14,
    state = HH7,
    age_cat = WAGE,
    newspaper = MT1,
    radio = MT2,
    tv = MT3,
    computer_tab = MT5,
    internet = MT10,
    mphone_own = MT11,
    mphone = MT12,
    sex_last12months = SB7,
    condom_use = SB8,
    relations_sexpartner = SB9,
    urban_rural = HH6,
    education_male = welevel,
    survey_weight = wmweight,
    strata = stratum,
    wealth_score = wscore,
    wealth_quintile = windex5) %>%
```

```
  select(cluster_id,  household_id, woman_id, household_uid,
         man_uid, age_female, education_ever, education_level,
         sch_now, sch_now_level, read_can, state, age_cat, newspaper,
         radio,  tv, computer_tab, internet, mphone_own, mphone,
         sex_last12months, condom_use, relations_sexpartner,
         urban_rural, education_male, survey_weight, strata,
         wealth_score, wealth_quintile)
```

Data Wrangling Female Subset Dataset

```
wn_clean <- mics6_wn_selected %>%
  mutate(sex = "Female") %>%

  mutate(
    newspaper = na_if(newspaper, 9),
    radio = na_if(radio, 9),
    tv = na_if(tv, 9),
    computer_tab = na_if(computer_tab, 9),
    internet = na_if(internet, 9),
    mphone = na_if(mphone, 9),

    mphone_own = case_when(
      mphone_own == 1 ~ 1,
      mphone_own == 2 ~ 0,
      TRUE ~ NA_real_
    ),

    education_ever = na_if(education_ever, 9),
    sch_now = na_if(sch_now, 9),
    sch_now_level = na_if(sch_now_level, 99),
    read_can = na_if(read_can, 9),

    urban_rural = case_when(
      urban_rural == 1 ~ "Urban",
      urban_rural == 2 ~ "Rural",
      TRUE ~ NA_character_
    ),

    sex_last12months = case_when(
      sex_last12months == 1 ~ 1,
      sex_last12months == 2 ~ 0,
      TRUE ~ NA_real_
    ),

    condom_use = case_when(
      condom_use == 1 ~ 0,
      condom_use == 2 ~ 1,
      TRUE ~ NA_real_
    ),

    relations_sexpartner = case_when(
```

```
      relations_sexpartner %in% c(4, 5, 6) ~ 1,
      relations_sexpartner %in% c(1, 2, 3) ~ 0,
      TRUE ~ NA_real_
    ),

    age_female = as.numeric(age_female),

    is_youth = age_cat %in% 1:2  # 1 = 15-19, 2 = 20-24
  ) %>%

  filter(is_youth) %>%

  mutate(
    education_male = factor(education_male,
                            levels = c(0, 1, 2, 3, 4, 9),
                            labels = c("None", "Primary", "Jnr Secondary", "Snr Secondary", "Higher", "

    wealth_quintile = factor(wealth_quintile,
                             levels = 1:5,
                             labels = c("Poorest", "Second", "Middle", "Fourth", "Richest"))
  )
```

- Merge Cleaned Male and Female Datasets

```
combined_data <- bind_rows(mn_clean, wn_clean)
```

```
view_df(combined_data)
glimpse(combined_data)
label(combined_data)
```

Explore & Visualise Missingness Combined_data Dataset

```
library(naniar)
library(dplyr)

miss_var_summary(combined_data)

print(miss_var_summary(combined_data), n = Inf)

combined_data %>%
  select(sex, age_male, age_female, age_cat, newspaper, radio, tv, internet, mphone,
         sex_last12months, condom_use, relations_sexpartner) %>%
  summarise_all(~ sum(is.na(.)))

vis_miss(combined_data)
```

```
combined_data %>%
  summarise(
    age = mean(ifelse(is.na(age_male), age_female, age_male), na.rm = TRUE),
```

```
    internet_mean = mean(internet, na.rm = TRUE),
    tv_mean = mean(tv, na.rm = TRUE),
    condom_use_mean = mean(condom_use, na.rm = TRUE)
  )

combined_data %>%
  group_by(sex) %>%
  summarise(
    n = n(),
    prop_newspaper = mean(!is.na(newspaper), na.rm = TRUE),
    sexually_active = mean(sex_last12months == 1, na.rm = TRUE),
    used_condom = mean(condom_use == 0, na.rm = TRUE)
  )
```

- Filter youth (15–24)

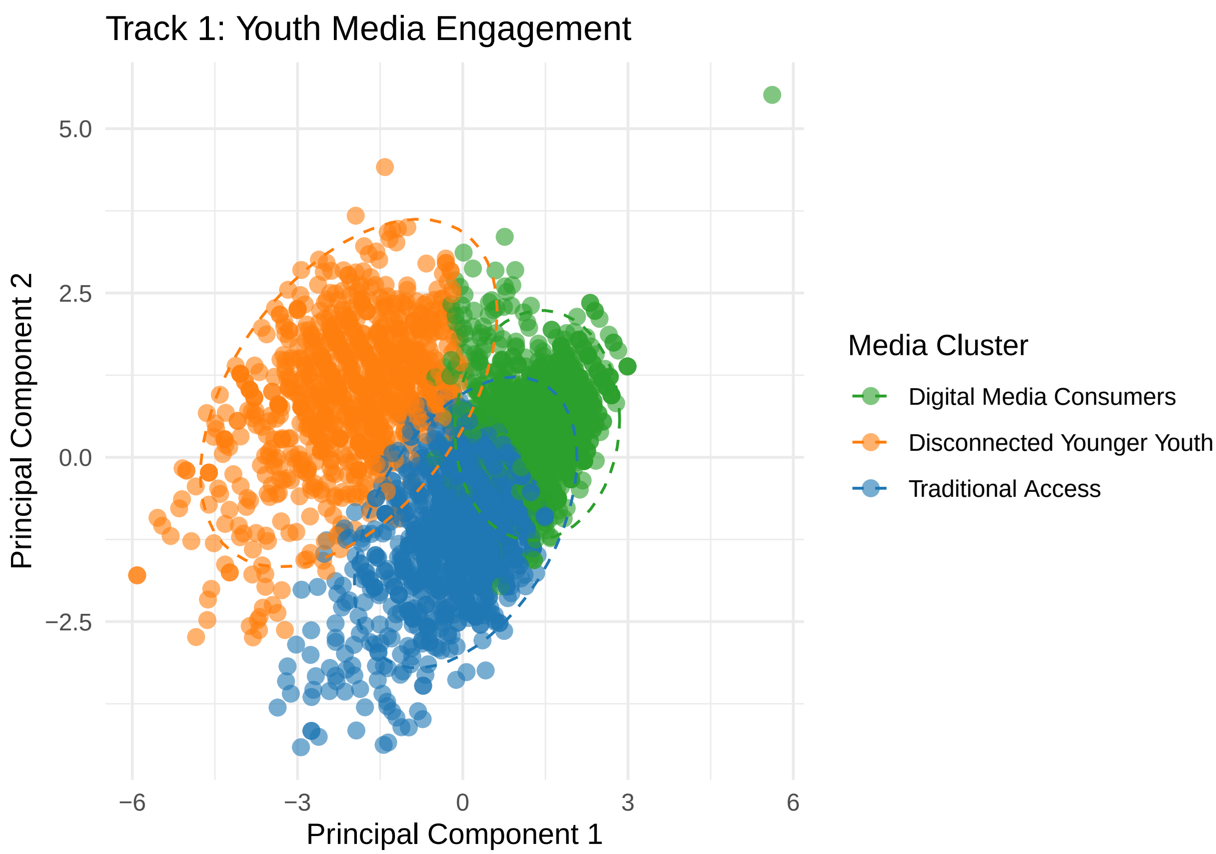
```
combined_data <- combined_data %>%
  filter(age_cat %in% 1:2) %>%
  mutate(
    age = ifelse(is.na(age_male), age_female, age_male),
    sex = as.character(sex)
  )
```

Track 1 — Media Clustering (K-Means)

```
youth_media_data <- combined_data %>%
  select(man_uid, sex, age, newspaper, radio, tv, internet, mphone, mphone_own,
         education_level, wealth_quintile, read_can) %>%
  drop_na()

media_vars_scaled <- youth_media_data %>%
  select(age, newspaper, radio, tv, internet, mphone, mphone_own, education_level, read_can) %>%
  scale()

set.seed(42)
k3 <- kmeans(media_vars_scaled, centers = 3, nstart = 25)
youth_media_data$media_cluster <- factor(k3$cluster)

youth_media_data <- youth_media_data %>%
  mutate(
    media_cluster_label = case_when(
      media_cluster == 1 ~ "Traditional Access",
      media_cluster == 2 ~ "Disconnected Younger Youth",
      media_cluster == 3 ~ "Digital Media Consumers"
    )
  )
```

PCA visualisation (Track 1)

```r
pca_result <- prcomp(media_vars_scaled, center = TRUE, scale. = TRUE)

pca_df_track1 <- as.data.frame(pca_result$x[, 1:2]) %>%
  mutate(media_cluster_label = youth_media_data$media_cluster_label)

cluster_colors <- c(
  "Traditional Access" = "#1f77b4",
  "Disconnected Younger Youth" = "#ff7f0e",
  "Digital Media Consumers" = "#2ca02c",
  "Low-Risk, Low-Media" = "#1f77b4",
  "Cautious Digital Adopters" = "#ff7f0e",
  "Digitally Active, High-Risk" = "#2ca02c"
)

# Plot with ellipses
library(ggplot2)

ggplot(pca_df_track1, aes(x = PC1, y = PC2, color = media_cluster_label)) +
  geom_point(alpha = 0.6, size = 2.5) +
  stat_ellipse(level = 0.95, type = "norm", linetype = "dashed") +
  scale_color_manual(values = cluster_colors) +
  labs(
    title = "Track 1: Youth Media Engagement",
    x = "Principal Component 1",
    y = "Principal Component 2", color = "Media Cluster"
  ) +
  theme_minimal()
```



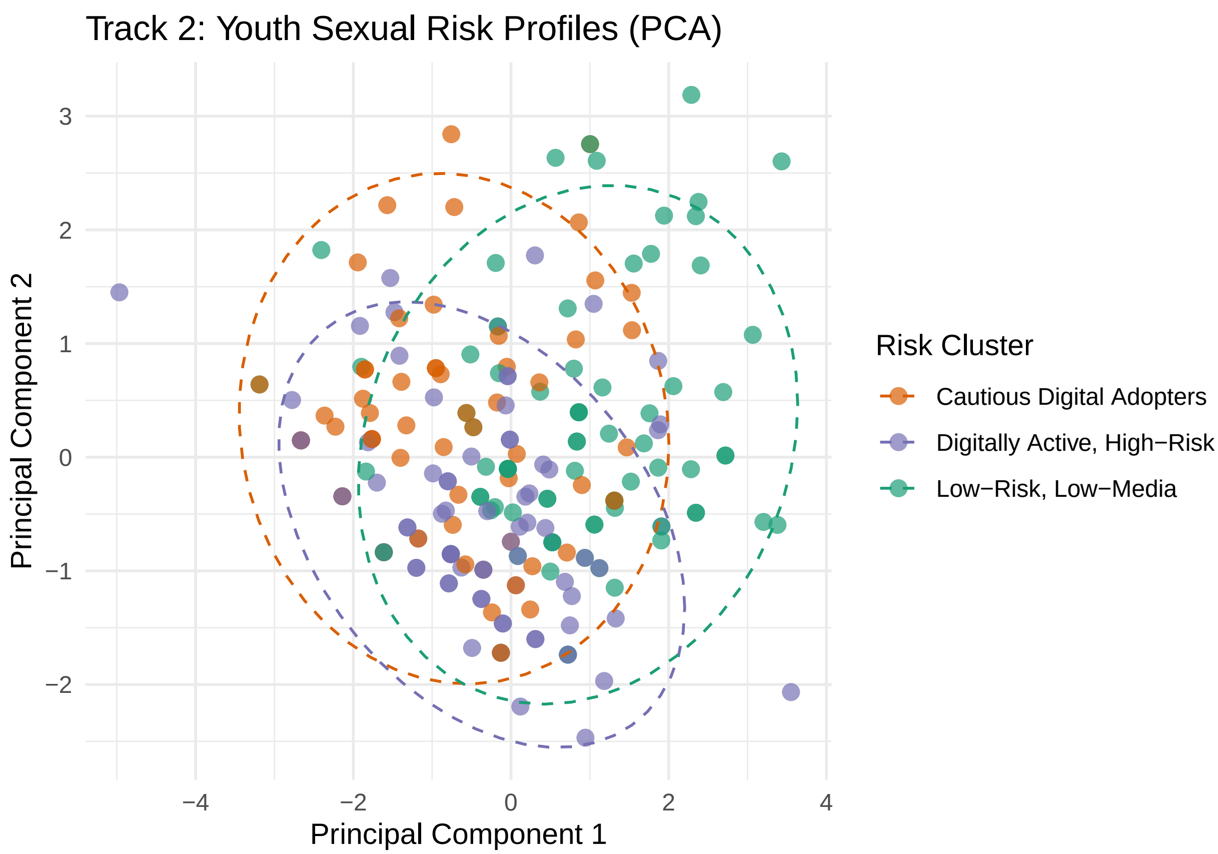
Track 2 — Sexual Risk Clustering (PAM)

```
track2_data <- combined_data %>%
  filter(sex_last12months == 1) %>%
  select(man_uid, sex, age, condom_use, relations_sexpartner,
         newspaper, radio, tv, internet, mphone, mphone_own,
         education_level, read_can, wealth_quintile) %>%
  drop_na()

track2_clean <- track2_data %>%
  mutate(
    condom_use = as.factor(condom_use),
    relations_sexpartner = as.factor(relations_sexpartner),
    mphone_own = as.factor(mphone_own),
    sex = as.factor(sex),
    wealth_quintile = as.factor(wealth_quintile),
    newspaper = as.numeric(newspaper),
    radio = as.numeric(radio),
    tv = as.numeric(tv),
    internet = as.numeric(internet),
    mphone = as.numeric(mphone),
    education_level = as.numeric(education_level),
    read_can = as.numeric(read_can)
```

```
  )

library(cluster)
gower_dist <- daisy(track2_clean %>% select(-man_uid, -sex, -age), metric = "gower")
pam_fit <- pam(gower_dist, k = 3)

track2_data$risk_cluster <- as.factor(pam_fit$clustering)

track2_data <- track2_data %>%
  mutate(
    risk_cluster_label = case_when(
      risk_cluster == "1" ~ "Low-Risk, Low-Media",
      risk_cluster == "2" ~ "Cautious Digital Adopters",
      risk_cluster == "3" ~ "Digitally Active, High-Risk"
    )
  )
```

Visualise PAM Clusters (Track 2)

```
track2_pca_data <- track2_data %>%
  select(newspaper, radio, tv, internet, mphone, education_level, read_can)

track2_pca_scaled <- scale(track2_pca_data)
pca_track2 <- prcomp(track2_pca_scaled)

pca_df_track2 <- as.data.frame(pca_track2$x[, 1:2])
pca_df_track2$risk_cluster_label <- track2_data$risk_cluster_label

cluster_colors <- c(
  "Low-Risk, Low-Media" = "#1b9e77",
  "Cautious Digital Adopters" = "#d95f02",
  "Digitally Active, High-Risk" = "#7570b3"
)

# Plot with ellipses
library(ggplot2)

ggplot(pca_df_track2, aes(x = PC1, y = PC2, color = risk_cluster_label)) +
  geom_point(size = 2.5, alpha = 0.7) +
  stat_ellipse(level = 0.95, linetype = "dashed") +
  scale_color_manual(values = cluster_colors) +
  labs(
    title = "Track 2: Youth Sexual Risk Profiles (PCA)",
    x = "Principal Component 1",
    y = "Principal Component 2",
    color = "Risk Cluster"
  ) +
  theme_minimal()
```




```
- Merge Tracks
```

```
track1_data <- youth_media_data %>%
  select(man_uid, media_cluster_label)

track1_vs_2 <- left_join(track2_data, track1_data, by = "man_uid")
```

Comparison of Clusters (media & risk)

Side-by-Side PCA Cluster Visualisation

```
library(ggplot2)
library(patchwork)

cluster_colors <- c(
  "Traditional Access" = "#1f77b4",
  "Disconnected Younger Youth" = "#ff7f0e",
  "Digital Media Consumers" = "#2ca02c",
  "Low-Risk, Low-Media" = "#1f77b4",
  "Cautious Digital Adopters" = "#ff7f0e",
  "Digitally Active, High-Risk" = "#2ca02c"
)
```

```r
pca_df_track1 <- as.data.frame(pca_result$x[, 1:2]) %>%
  mutate(cluster = youth_media_data$media_cluster_label)

p1 <- ggplot(pca_df_track1, aes(x = PC1, y = PC2, color = cluster)) +
  geom_point(alpha = 0.6, size = 2.5) +
  stat_ellipse(level = 0.95, type = "norm", linetype = "dashed") +
  scale_color_manual(values = cluster_colors) +
  labs(
    title = "Track 1: Youth Media Engagement",
    x = "PC1", y = "PC2", color = "Media Cluster"
  ) +
  theme_minimal()

p2 <- ggplot(pca_df_track2, aes(x = PC1, y = PC2, color = risk_cluster_label)) +
  geom_point(alpha = 0.6, size = 2.5) +
  stat_ellipse(level = 0.95, type = "norm", linetype = "dashed") +
  scale_color_manual(values = cluster_colors) +
  labs(
    title = "Track 2: Youth Sexual Risk Profiles",
    x = "PC1", y = "PC2", color = "Risk Cluster"
  ) +
  theme_minimal()

combined_plot <- p1 + p2 +
  plot_layout(ncol = 2, guides = "collect") &
  theme(legend.position = "bottom")

combined_plot
```

- Export Plot (PNG and PDF)

```r
ggsave(
  filename = "figures/track1_track2_side_by_side.pdf",
  plot = combined_plot,
  width = 14,
  height = 7
)

ggsave(
  filename = "figures/track1_track2_side_by_side.png",
  plot = combined_plot,
  width = 14,
  height = 7,
  dpi = 300
)
```

Figure 1: Figure: Side-by-Side PCA Plots of Youth Media (Track 1) and Sexual Risk (Track 2) Clusters

Clusters by Demographics

Track 1 – Media Cluster by sex & wealth

```
track1_summary <- youth_media_data %>%
  group_by(media_cluster_label, sex, wealth_quintile) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(media_cluster_label) %>%
  mutate(pct = round(100 * n / sum(n), 1)) %>%
  ungroup()

track1_table <- track1_summary %>%
  unite("Group", sex, wealth_quintile, sep = " - ") %>%
  select(media_cluster_label, Group, pct) %>%
  tidyr::pivot_wider(names_from = Group, values_from = pct)

track1_table %>%
  gt() %>%
  tab_header(
    title = "Demographic Distribution by Media Cluster (Track 1)",
    subtitle = "Percentage of respondents by sex and wealth within each media cluster"
  ) %>%
  fmt_percent(columns = everything(), scale_values = FALSE)
```

Demographic Distribution by Media Cluster (Track 1)
Percentage of respondents by sex and wealth within each media cluster

media_cluster_label	Female - Poorest	Female - Second	Female - Middle	Female - Fourth
Digital Media Consumers	0.50%	2.60%	9.60%	19.20%
Disconnected Younger Youth	4.00%	8.20%	14.40%	14.30%
Traditional Access	2.80%	8.00%	14.80%	15.90%

Table 1: Demographic Distribution by Sexual Risk Cluster (Track 2)

risk_cluster_label	Female - Second	Female - Richest	Male - Second	Male - Middle	Male - Fourth	Ma
Cautious Digital Adopters	8.8	12.3	8.8	21.1	19.3	
Digitally Active, High-Risk	NA	1.2	15.0	3.8	38.8	
Low-Risk, Low-Media	5.3	NA	26.3	30.3	1.3	

Track 2 – Risk Cluster by sex & wealth

```
track2_summary <- track2_data %>%
  group_by(risk_cluster_label, sex, wealth_quintile) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(risk_cluster_label) %>%
  mutate(pct = round(100 * n / sum(n), 1)) %>%
  ungroup()

track2_table <- track2_summary %>%
  unite("Group", sex, wealth_quintile, sep = " - ") %>%
  select(risk_cluster_label, Group, pct) %>%
  tidyr::pivot_wider(names_from = Group, values_from = pct)

track2_table %>%
  kbl(caption = "Demographic Distribution by Sexual Risk Cluster (Track 2)") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), full_width = FALSE)
```

```
Discussion and Implications

This analysis revealed three distinct media engagement clusters and three
sexual risk profiles among Nigerian youth:

Digital Media Consumers tend to have higher access to internet and mobile phones,
and show some overlap with higher-risk sexual behavior.

Disconnected Younger Youth show low digital engagement and tend to be less sexually
active or lower-risk.

Cautious Digital Adopters and Low-Risk, Low-Media profiles suggest that access
alone does not equate to risk.

The integration of Track 1 and Track 2 clusters reveals important overlaps - but also
highlights that digital engagement alone does not uniformly predict risk. Instead,
demographic context (such as sex, wealth) remains crucial.
```

Policy Relevance

Targeted Sexual Health Interventions: Programs may need to differentiate between digitally connected yet cautious youth versus those who are digitally active and high-risk.

Digital Literacy and Safe Behavior Campaigns:** Especially for those with high internet/mobile access.

Addressing Inequity: Disconnected youth may lack both sexual health information and general digital opportunities.

Limitations

Survey weights were not applied; hence findings may not be nationally representative.
Unmeasured variables (e.g., peer influence, religious beliefs) may affect behaviors.

Next Steps

- Extend to rural vs. urban comparisons.
- Include additional behavioral indicators if available.
- Apply supervised machine learning for prediction (e.g., classification models).