Data Promperú Clustering PAM FAMD

Gabriel Gonzalo Ojeda Cárcamo

26/11/2025

Contents

Introduction

This project applies mixed-data clustering techniques, combining Factor Analysis of Mixed Data (FAMD) for dimensionality reduction and the Partitioning Around Medoids (PAM) algorithm for cluster formation. This methodological approach makes it possible to identify latent patterns within the population of international tourists by simultaneously incorporating sociodemographic characteristics, travel motivations, consumption preferences, and spending habits.

The resulting segmentation reveals five clearly differentiated tourist profiles, providing valuable insights for marketing strategies, product development, and destination management. Identifying these groups enhances understanding of the heterogeneity among international visitors and supports evidence-based decision-making to strengthen the competitiveness and sustainability of the tourism sector.

Study Objective

- Identify groups of tourists who share similar characteristics
- Understand behavioral patterns related to travel, consumption, and personal attributes
- Support decision-making in tourism promotion, product design, and destination planning

Data Source

The analysis is based on the 2024 Foreign Tourist Profile Survey, conducted in February, May, August, and November of 2024. The dataset includes 5,268 international tourists, aged 15 years and older, surveyed at the Jorge Chávez International Airport upon entering the country. The survey collects detailed information on motivations, expectations, sociodemographic characteristics, and travel behavior, making it a robust source for statistical analysis and clustering.

Loading Packages

```
#Loading packages
required_packages <- c(
  "haven",        # read .sav (SPSS) files
  "dplyr",        # data manipulation
  "tidyr",        # tidy data tools
  "purrr",        # functional programming
  "janitor",      # clean tables, crosstabs
  "sjlabelled",   # convert SPSS labelled → factor
  "ggplot2",      # visualization
  "GGally",       # ggpairs, exploratory plots
  "naniar",       # missing data visualization
  "VIM",          # alternative missing data plots
  "missForest",   # mixed-type imputation
  "FactoMineR",   # FAMD (mixed data PCA)
  "factoextra",   # visualizations for FAMD/clustering
  "cluster",      # PAM, silhouette
  "clusterCrit",  # CH, Dunn indices
  "clustertend",  # Hopkins, VAT (clusterability)
  "fpc",          # cluster stability (clusterboot)
  "ggrepel",      # nicer text labels
```

```
  "scales",       # formatting (percentages, numbers)
  "stringr",      # string operations
  "forcats",      # factor manipulation
  "knitr",
  "kableExtra"
)

# Install missing packages and load them
for (pkg in required_packages) {
  if (!require(pkg, character.only = TRUE)) {
    install.packages(pkg, repos = "http://cran.us.r-project.org")
    library(pkg, character.only = TRUE)
  } else {
    library(pkg, character.only = TRUE)
  }
}
```

Loading Data

```
# Loading Data

url_sav  <- "https://github.com/gabdmns/Capstone_project_2/raw/main/PTE2024PROMPERU.sav"
tmp_file <- tempfile(fileext = ".sav")

download.file(
  url    = url_sav,
  destfile = tmp_file,
  mode   = "wb",
  method = "libcurl"  # <- clave en Windows moderno
)

data_sav <- read_sav(tmp_file)

nrow(data_sav)  # Number of observations
```

```
## [1] 5268
```

```
ncol(data_sav)  # Number of variables
```

```
## [1] 5552
```

Data Wrangling

The selection of variables was based on the indicators reported by TurismoIN of PROMPERÚ, specifically on the set of summary measures from the survey. Variables related to the main purpose of travel, sociodemographic characteristics (age, gender, marital status, educational level, employment sector, and generational group), travel-planning aspects, and total spending were included.

Each variable was converted to its appropriate type (factor or numeric). For variables P59 and P43_1, the category "NS/NR" was assigned to missing values in P59, while P43_1 was log-transformed and its missing

values were imputed using the missForest algorithm, which applies Random Forest models iteratively to handle mixed-type data.

```
# NA percentage in each variable
na_pct <- sapply(data_sav, function(x) mean(is.na(x)) * 100)
# Filter variables >90% NA's
vars_mas_90_na <- na_pct[na_pct > 90]
# Count of variables with more than 90% NA'S
length(vars_mas_90_na)
```

```
## [1] 5257
```

```
# Variable selection

vars_final <- c("P01","P53_RANGO2","P54","P56","P57_1","P58","P59","P33","P34",
                "P33_34_RNG","P60","P61","P53_GENERACION","P43_1")
data_filtrada <- data_sav[, vars_final, drop = FALSE]

info_compacto <- tibble(
  variable  = vars_final,
  pregunta  = map_chr(
    vars_final,
    ~ {
      lab <- attr(data_sav[[.x]], "label")
      if (is.null(lab)) NA_character_ else lab
    }
  ),
  categorias = map_chr(
    vars_final,
    ~ {
      labs <- attr(data_sav[[.x]], "labels")
      if (is.null(labs)) return(NA_character_)
      paste(names(labs), collapse = " | ")
    }
  ),
  na_pct = na_pct[vars_final]
)

print(info_compacto)
```

```
## # A tibble: 14 x 4
##    variable       pregunta                                    categorias na_pct
##    <chr>          <chr>                                       <chr>       <dbl>
##  1 P01            P01. Según la siguiente tarjeta ¿Cuál fue s~ Vacacione~    0
##  2 P53_RANGO2     ¿Qué edad tiene usted? (RANGO AGRUPADO)      De 15 a 2~    0
##  3 P54            P54. Género                                 Masculino~    0
##  4 P56            P56. Según la siguiente tarjeta, ¿Cuál es s~ Soltero |~    0
##  5 P57_1          P57. ¿En cuál de las siguientes situaciones~ No tengo ~    0
##  6 P58            P58. Según la siguiente tarjeta, ¿En cuál d~ Trabajado~    0
##  7 P59            P59. ¿En qué rubro trabaja?                 Industria~   18.7
##  8 P33            P33. ¿En qué año compró su pasaje y /o paqu~ Año 2023 ~    0
##  9 P34            P34. ¿En qué mes compró su pasaje y /o paqu~ Enero | F~    0
## 10 P33_34_RNG     P34. ¿En qué mes compró su pasaje y /o paqu~ Menos de ~    0
```

```
## 11 P60            P60. Según la siguiente tarjeta, ¿Cuál es e~ Primaria ~    0
## 12 P61            P61. Según la siguiente tarjeta ¿cuál es su~ Menos de ~    0
## 13 P53_GENERACION P53 Grupo Generacional                       Centennia~    0
## 14 P43_1          P43_1. (Monto en Dólares americanos) ¿Cuánt~ <NA>         43.8
```

```
# Correct Data type

data_fixed <- data_filtrada %>%
  # 1) labelled → factor
  mutate(across(
    where(~ sjlabelled::is_labelled(.x)),
    ~ sjlabelled::as_factor(.x, levels = "labels")
  )) %>%
  # 2) character → factor
  mutate(across(
    where(is.character),
    as.factor
  )) %>%
  # 3) Date → numeric
  mutate(across(
    where(~ inherits(.x, "Date")),
    ~ as.numeric(.x)
  )) %>%
  # 4) POSIXct → numeric
  mutate(across(
    where(~ inherits(.x, "POSIXt")),
    ~ as.numeric(.x)
  ))

# P59: add NS/NR category
if ("P59" %in% names(data_fixed) && is.factor(data_fixed[["P59"]])) {
  tmp <- as.character(data_fixed[["P59"]])
  tmp[is.na(tmp)] <- "NS/NR"
  data_fixed[["P59"]] <- factor(tmp)
}

# P43_1: log-transform and set invalids to NA
if ("P43_1" %in% names(data_fixed) && is.numeric(data_fixed[["P43_1"]])) {
  data_fixed <- data_fixed %>%
    mutate(
      P43_1 = ifelse(P43_1 <= 0, NA_real_, P43_1),
      P43_1 = log(P43_1)
    )
}

data_fixed_df <- as.data.frame(data_fixed)

colSums(is.na(data_fixed_df))
```

```
##            P01     P53_RANGO2            P54            P56          P57_1
##              0              0              0              0              0
##            P58            P59            P33            P34     P33_34_RNG
##              0              0              0              0              0
##            P60            P61 P53_GENERACION         P43_1
```

```
##              0              0              0           2306
```

```
# Validation
which(sapply(data_fixed, is.list))
```

```
## named integer(0)
```

Missing value imputation

```
# Handling Missing Values


# Add a helper numeric variable from a factor (e.g., age range)
data_fixed_df$helper_num <- as.numeric(data_fixed_df$P53_RANGO2)

set.seed(123)
mf_res <- missForest(data_fixed_df)

# Check OOB error
mf_res$OOBerror
```

```
##     NRMSE       PFC
## 0.1915412 0.0000000
```

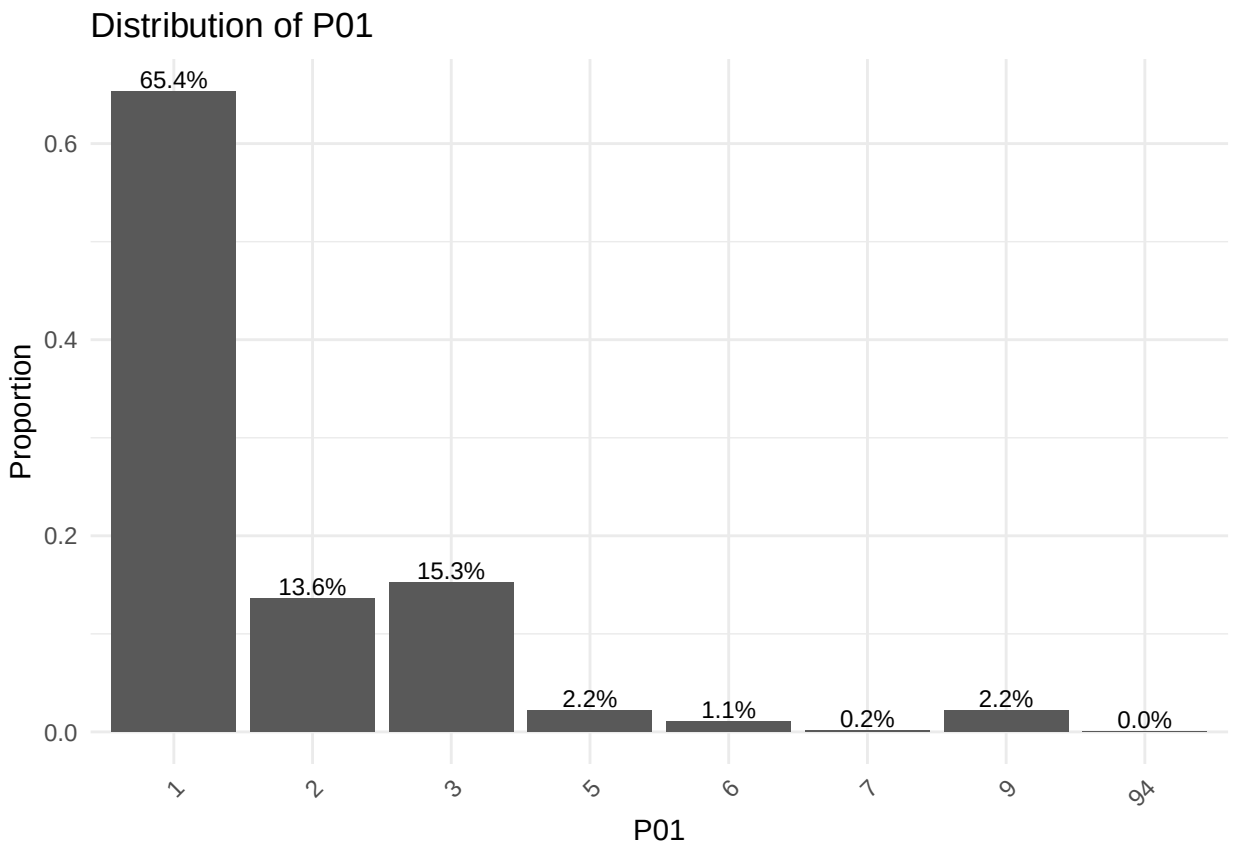
```
# Update data

data_fixed_df_imputed <- mf_res$ximp

# Drop helper variable
data_fixed_df_imputed$helper_num <- NULL

# Final imputed dataset
data_fixed <- data_fixed_df_imputed
```

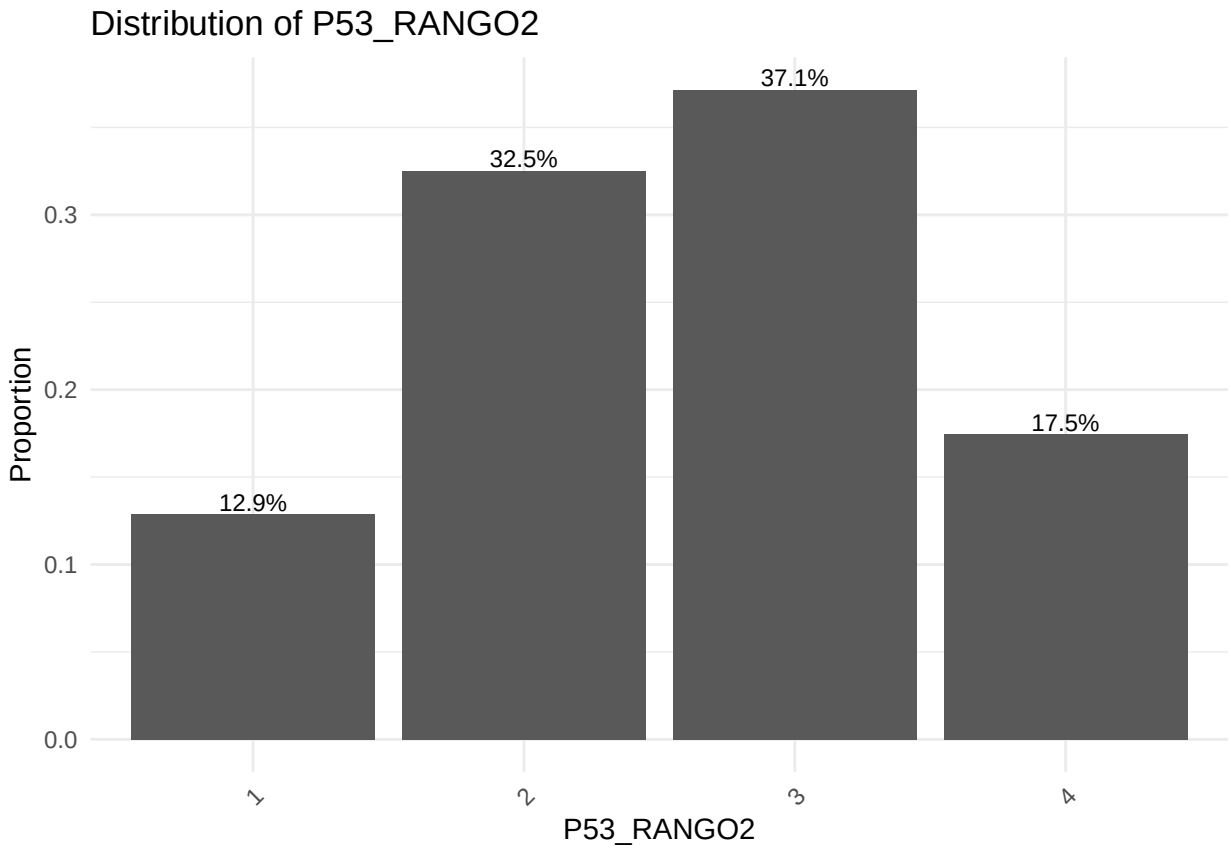
Missing values were imputed using the missForest algorithm, which applies Random Forest regression and classification to iteratively estimate missing entries in mixed-type datasets. A helper numeric variable derived from the age-range factor was included to stabilize the procedure, and the spending variable was log-transformed prior to imputation. The algorithm fit Random Forest models for each incomplete variable using all others as predictors, updating imputed values until convergence. The out-of-bag error indicated good performance (NRMSE ≈ 0.19 for numeric variables, PFC $= 0.00$ for categorical variables), providing evidence that the imputation preserved the underlying structure of the data. The auxiliary variable was then removed, and the fully imputed dataset was used for subsequent FAMD and clustering analyses.

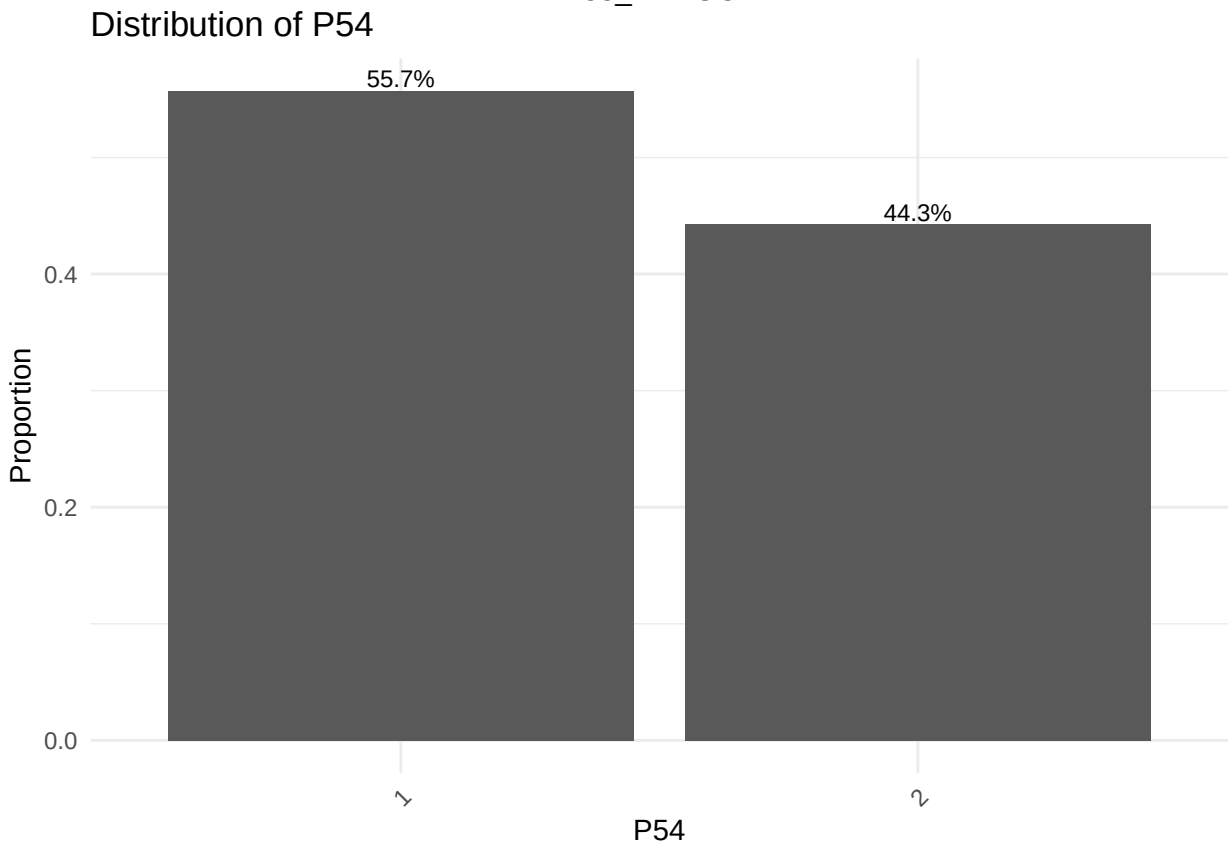
Exploratory Data Analysis

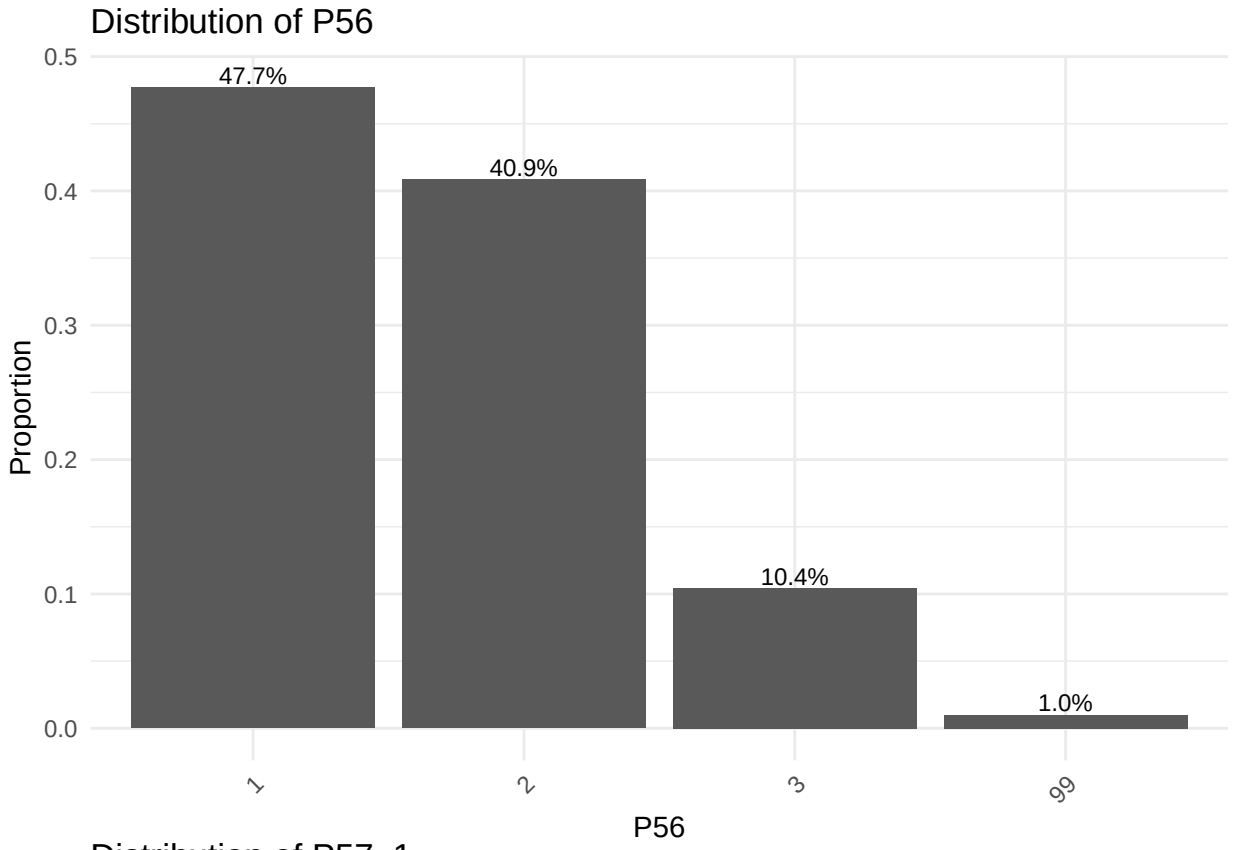
```
# Exploratory Data Analysis

# Distributions of key variables
```

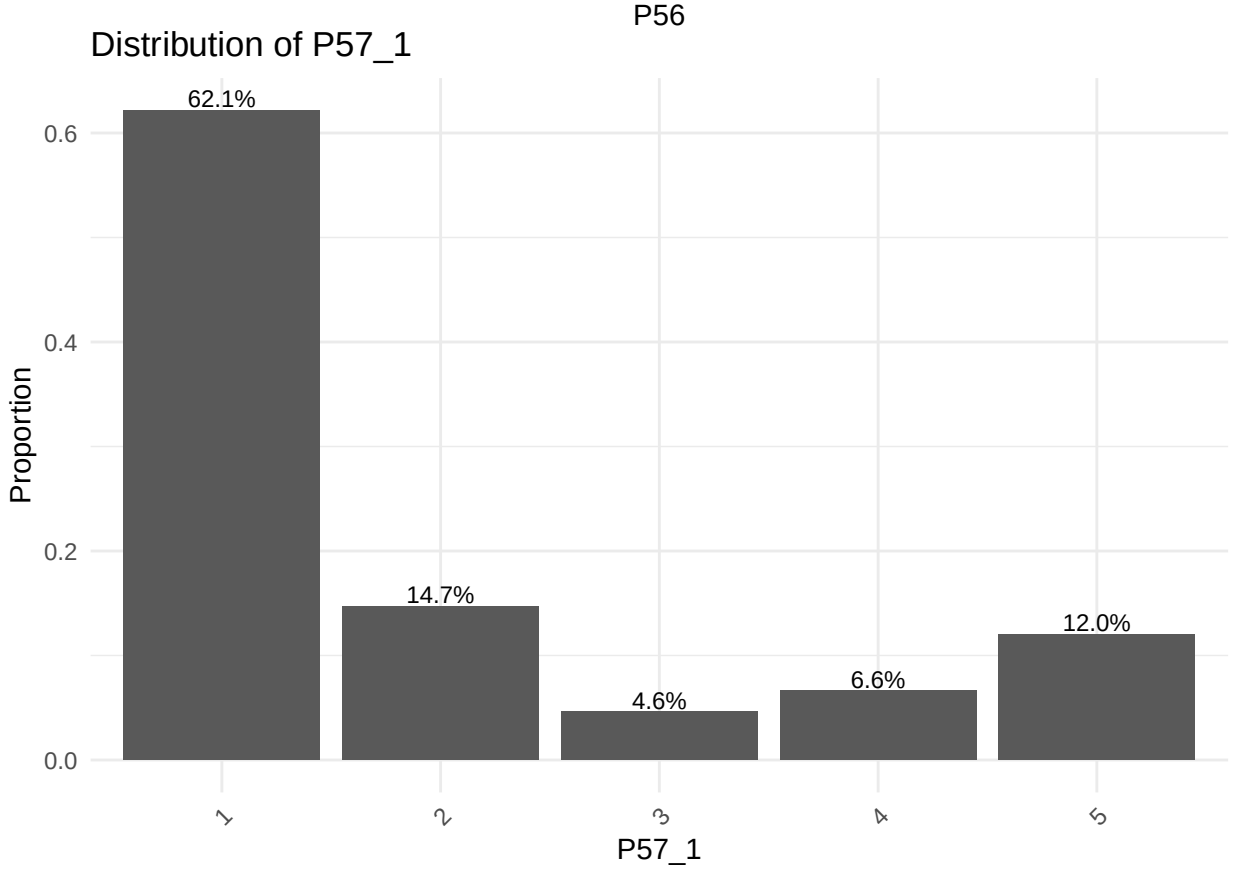
```r
# We work on data_fixed (already cleaned, 14 vars)
df <- data_fixed

# 1.1 Categorical variables: barplots in a loop --------------------------
cat_vars <- c("P01", "P53_RANGO2", "P54", "P56", "P57_1",
              "P58", "P59", "P33", "P34", "P33_34_RNG",
              "P60", "P61", "P53_GENERACION")

for (v in cat_vars) {
  p <- df %>%
    count(.data[[v]]) %>%
    mutate(prop = n / sum(n)) %>%
    ggplot(aes(x = .data[[v]], y = prop)) +
    geom_col() +
    geom_text(aes(label = scales::percent(prop, accuracy = 0.1)),
              vjust = -0.2, size = 3) +
    labs(
      title = paste("Distribution of", v),
      x = v, y = "Proportion"
    ) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  print(p)
}
```

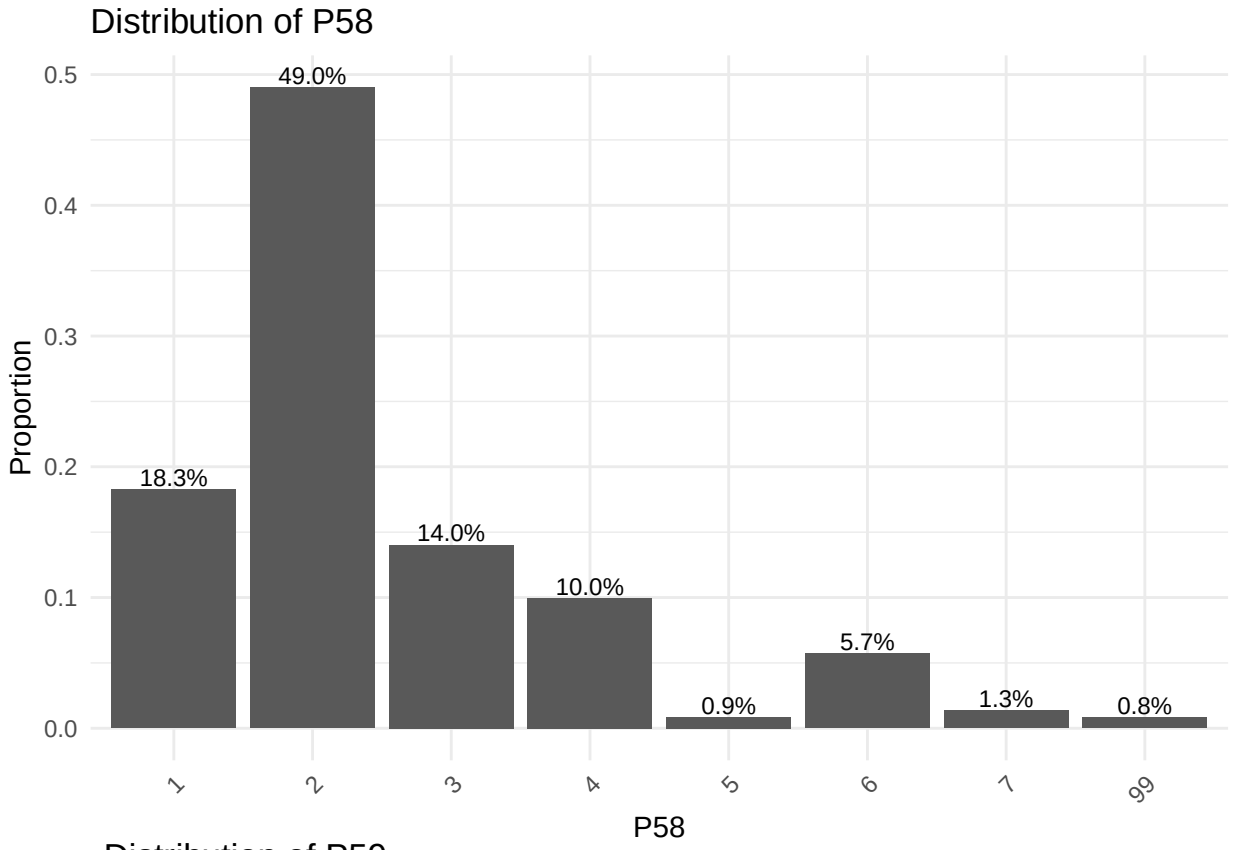


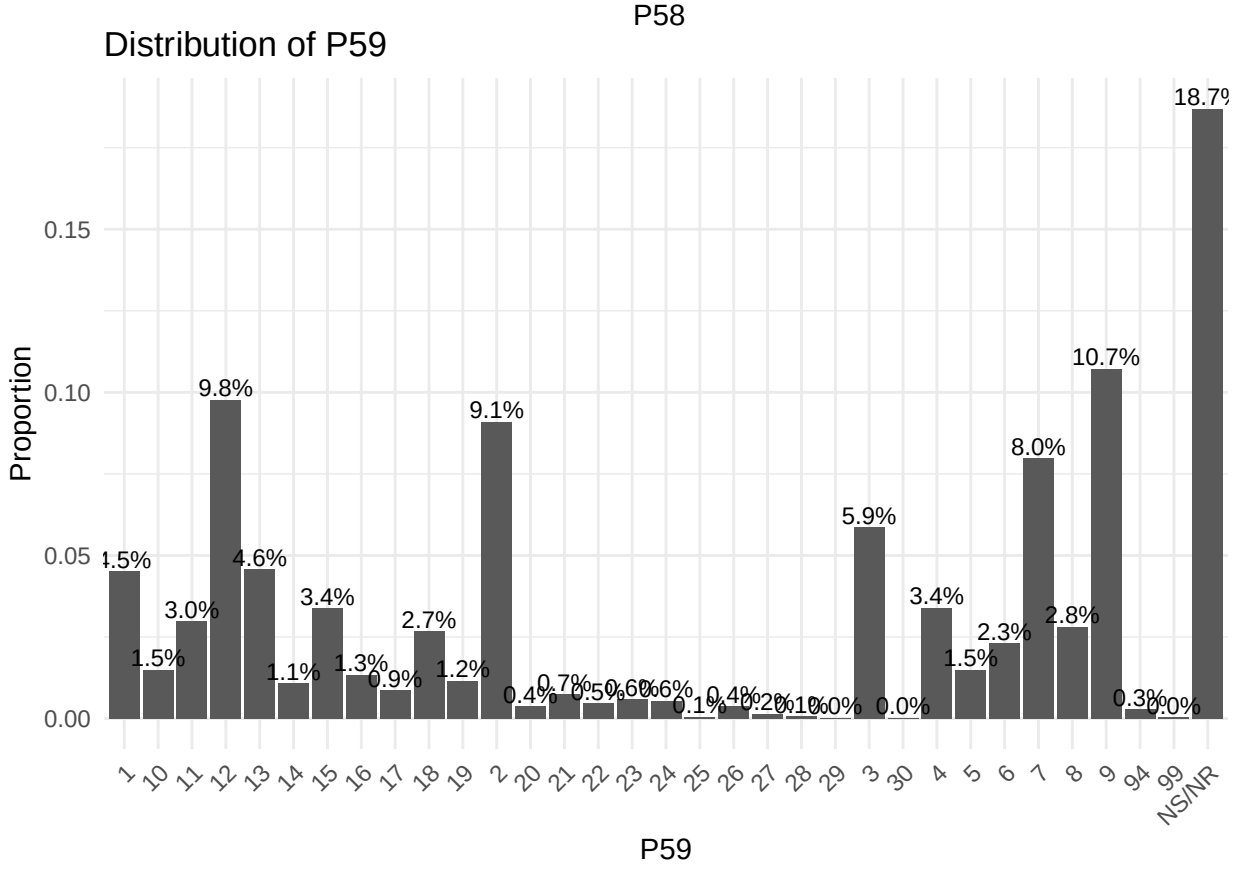

Distribution of P53_RANGO2
37.1%
32.5%
17.5%
12.9%
0.3
0.2
0.1
0.0
Proportion
1
2
3
4
P53_RANGO2

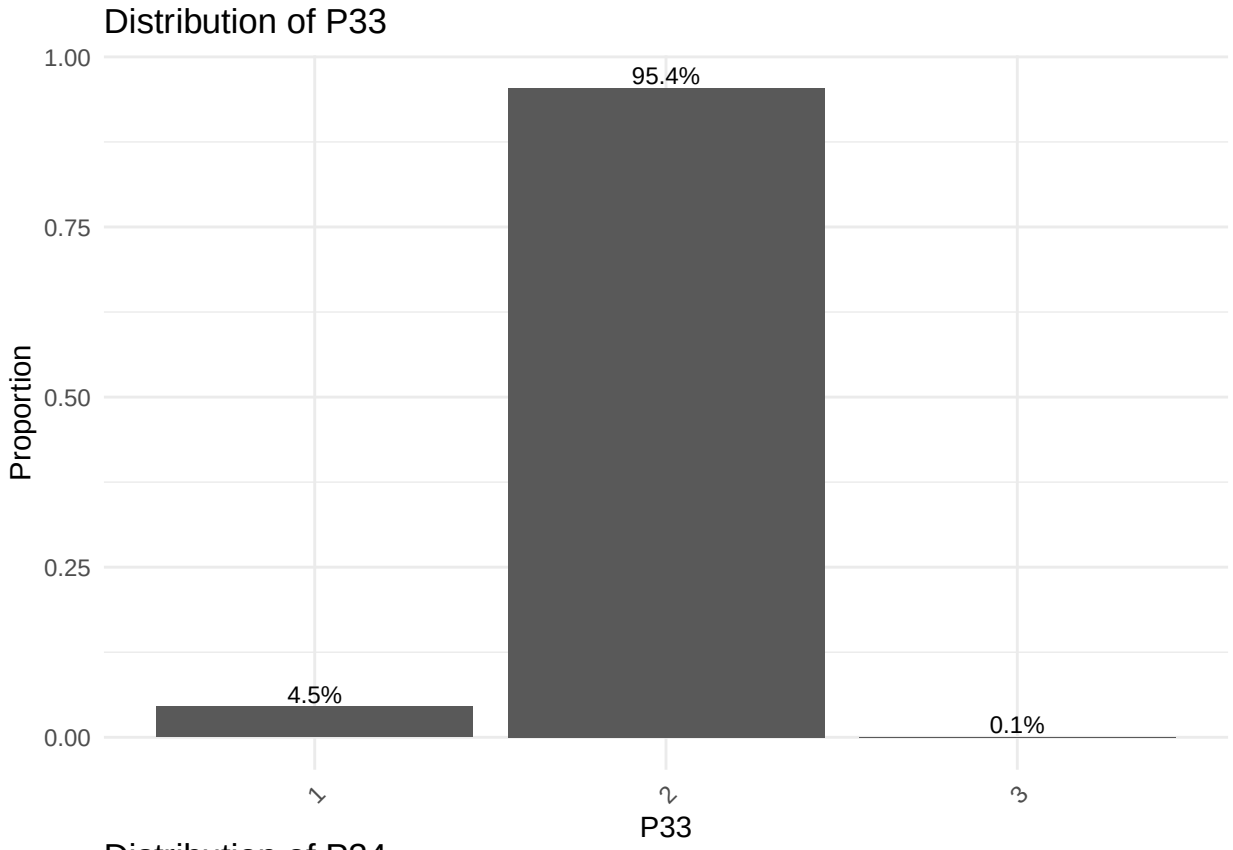

Distribution of P54
55.7%
44.3%
0.4
0.2
0.0
Proportion
1
2
P54

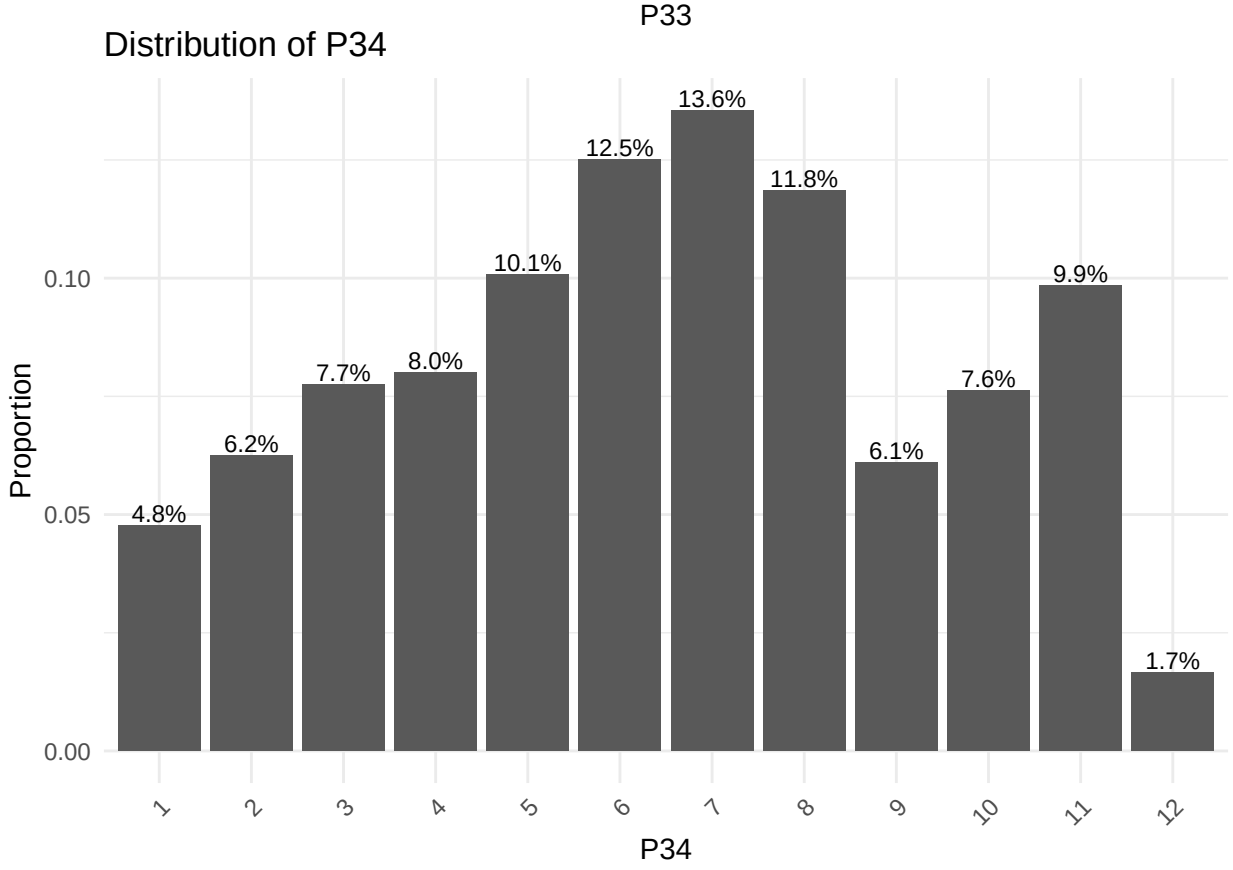

Distribution of P56
0.5
0.4
0.3
0.2
0.1
0.0
Proportion
47.7%
40.9%
10.4%
1.0%
1
2
3
99
P56

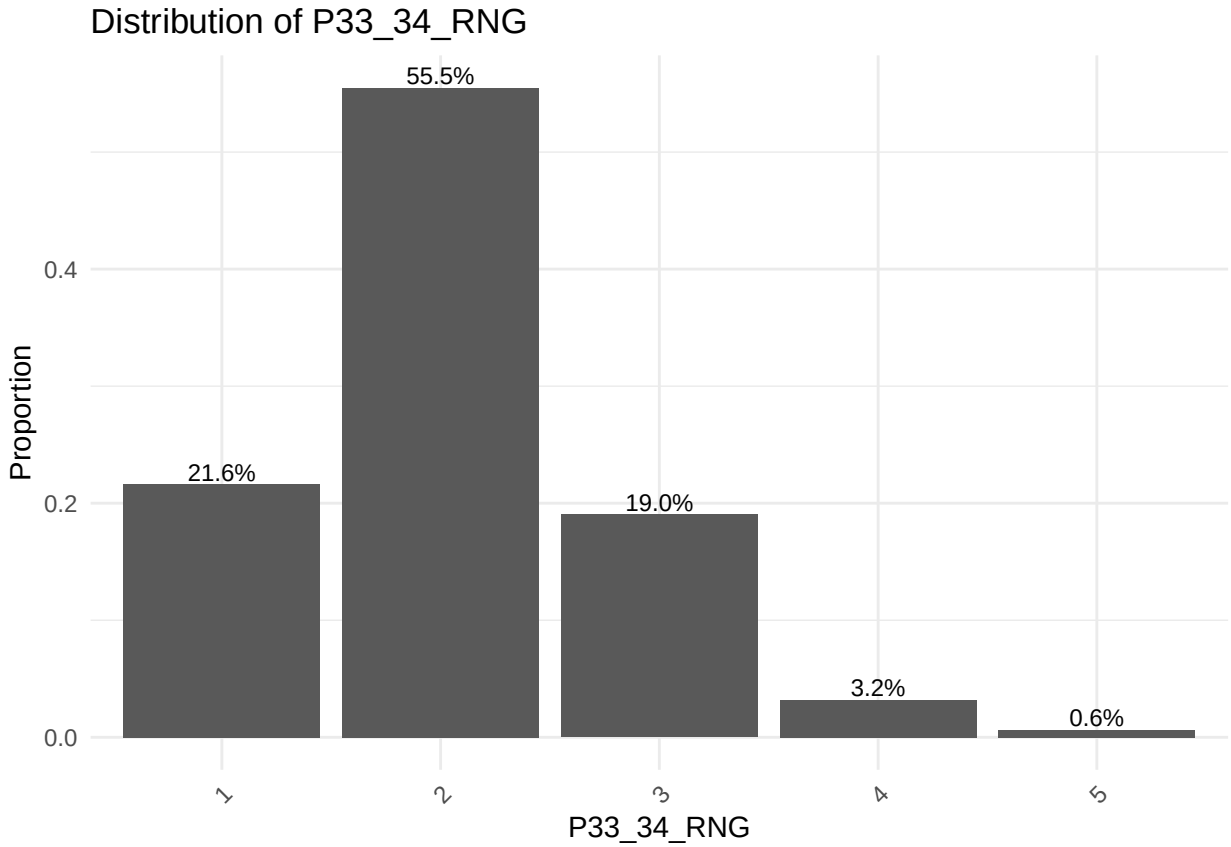

Distribution of P57_1
0.6
0.4
0.2
0.0
Proportion
62.1%
14.7%
4.6%
6.6%
12.0%
1
2
3
4
5
P57_1

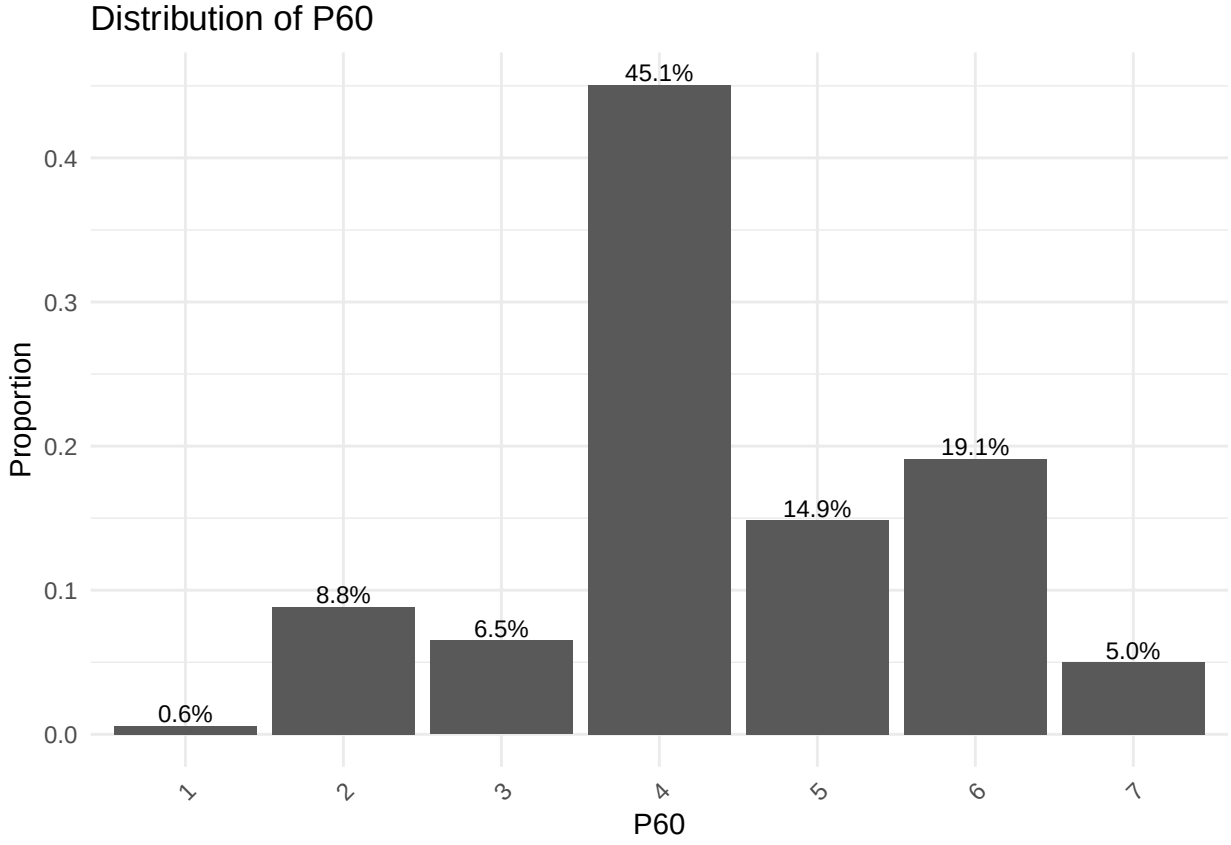

Distribution of P58
Proportion
0.5
0.4
0.3
0.2
0.1
0.0
49.0%
18.3%
14.0%
10.0%
0.9%
5.7%
1.3%
0.8%
1
2
3
4
5
6
7
99
P58

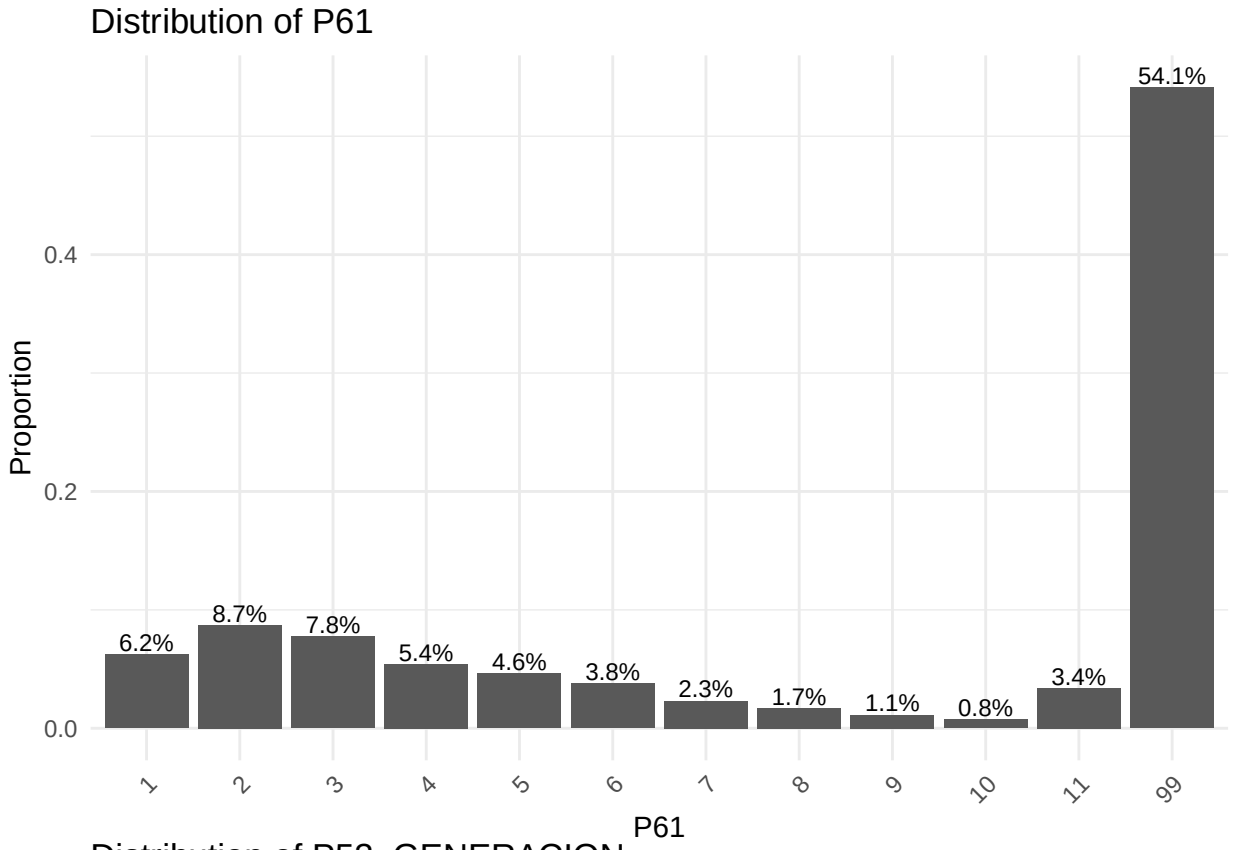

Distribution of P59
Proportion
0.15
0.10
0.05
0.00
18.7%
10.7%
9.8%
9.1%
8.0%
5.9%
4.6%
3.4%
3.4%
3.0%
2.8%
2.7%
2.3%
1.5%
1.5%
1.3%
1.2%
1.1%
0.9%
0.7%
0.4%
0.3%
0.0%
1
10
11
12
13
14
15
16
17
18
19
2
20
21
22
23
24
25
26
27
28
29
3
30
4
5
6
7
8
9
94
99
NS/NR
P59

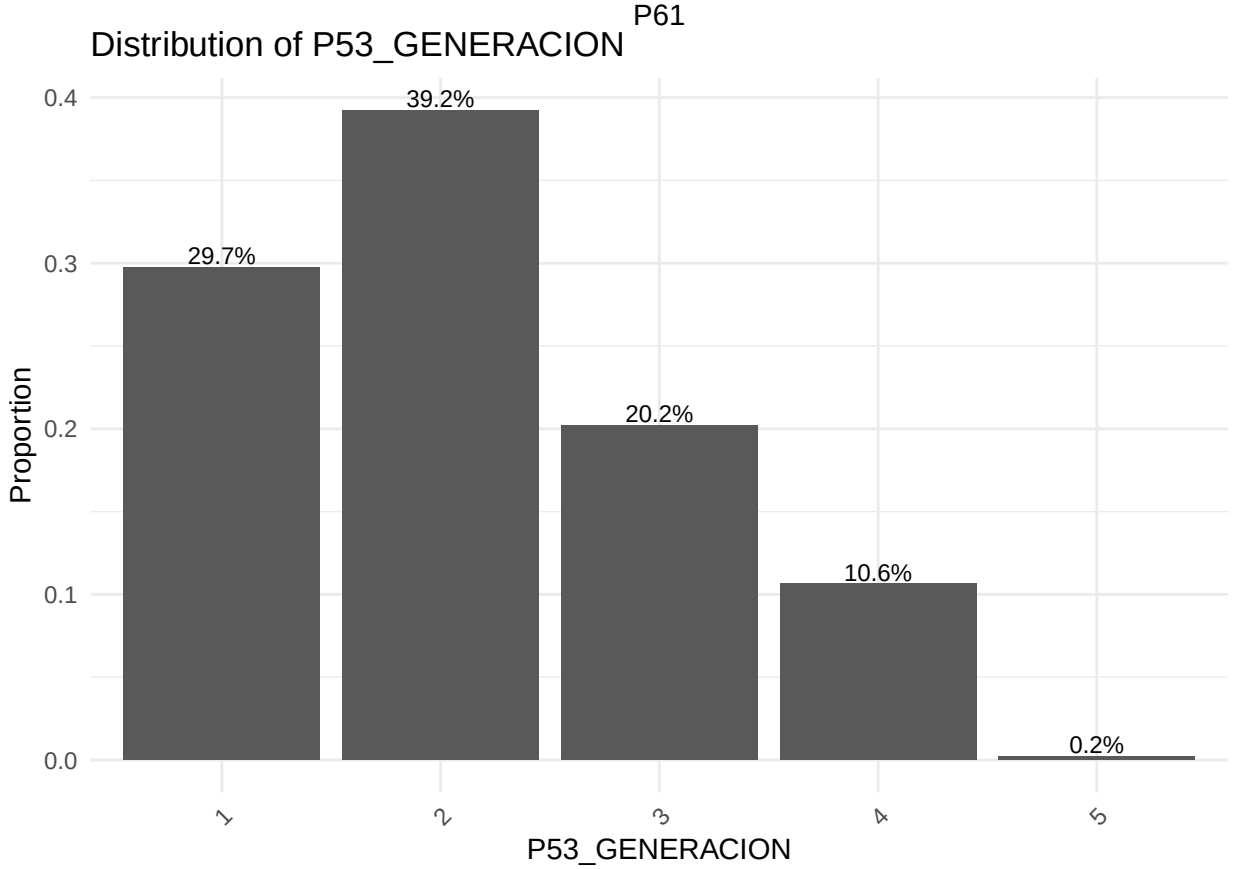

Distribution of P33
1.00
0.75
0.50
0.25
0.00
Proportion
95.4%
4.5%
0.1%
1
2
3
P33


Distribution of P34
0.10
0.05
0.00
Proportion
4.8%
6.2%
7.7%
8.0%
10.1%
12.5%
13.6%
11.8%
6.1%
7.6%
9.9%
1.7%
1
2
3
4
5
6
7
8
9
10
11
12
P34

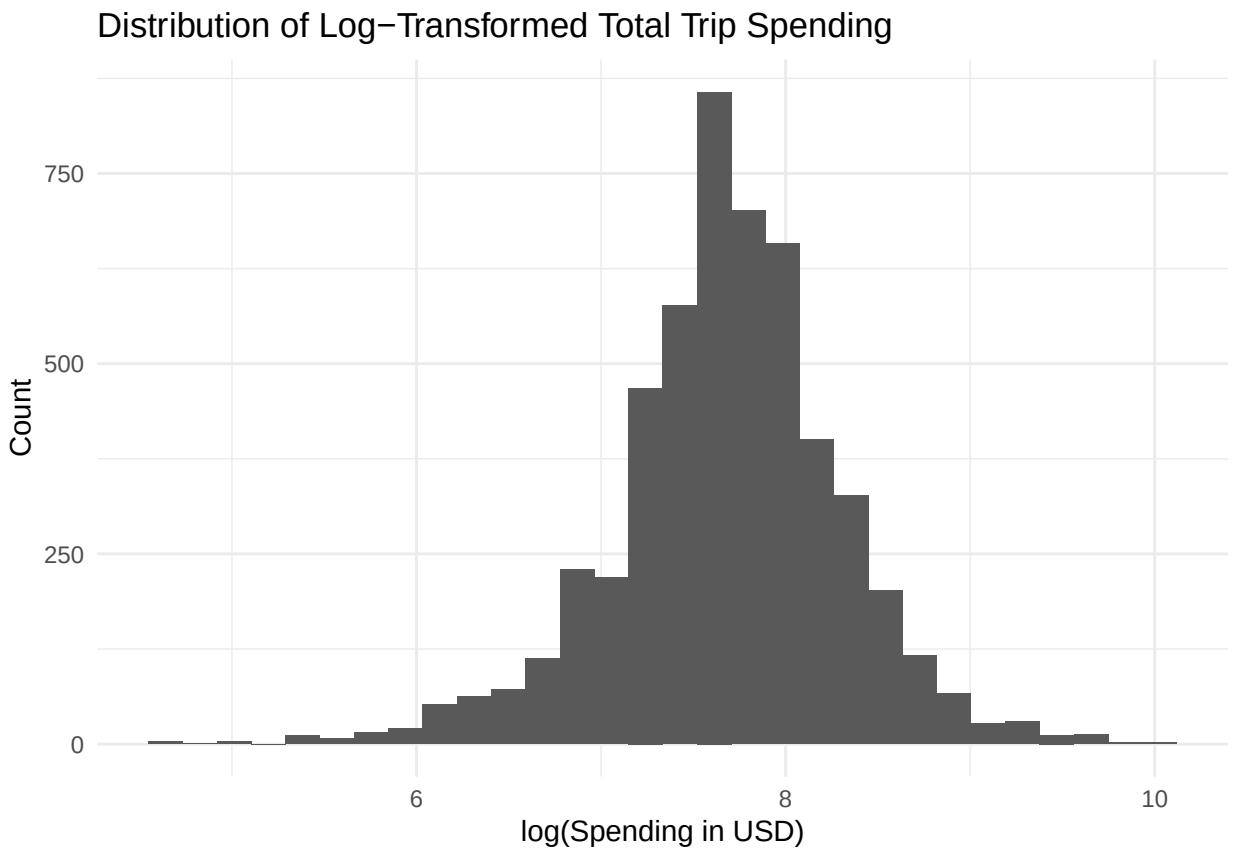

Distribution of P33_34_RNG
55.5%
21.6%
19.0%
3.2%
0.6%
0.4
0.2
0.0
Proportion
1
2
3
4
5
P33_34_RNG


Distribution of P60
45.1%
19.1%
14.9%
8.8%
6.5%
5.0%
0.6%
0.4
0.3
0.2
0.1
0.0
Proportion
1
2
3
4
5
6
7
P60

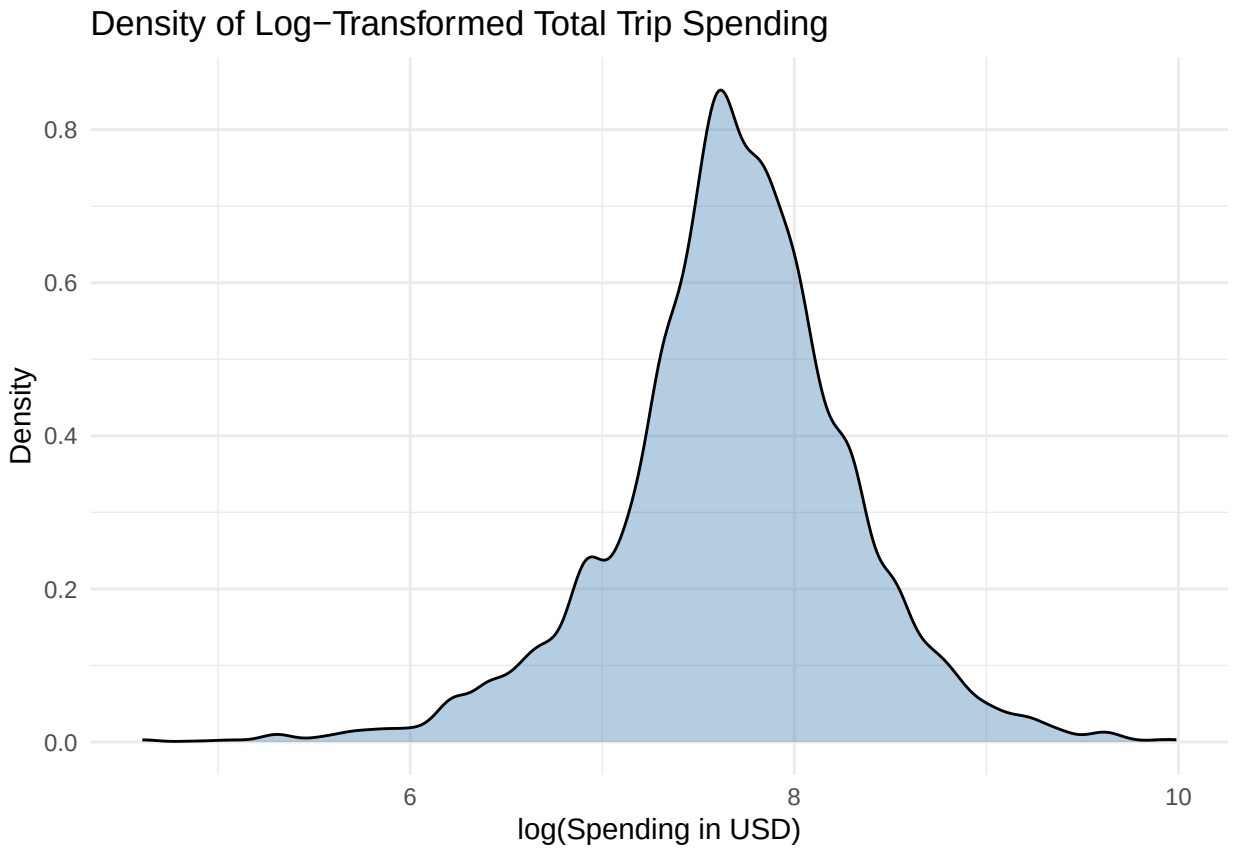

Distribution of P61
Proportion
0.0
0.2
0.4
6.2%
8.7%
7.8%
5.4%
4.6%
3.8%
2.3%
1.7%
1.1%
0.8%
3.4%
54.1%
1
2
3
4
5
6
7
8
9
10
11
99
P61


Distribution of P53_GENERACION
Proportion
0.0
0.1
0.2
0.3
0.4
29.7%
39.2%
20.2%
10.6%
0.2%
1
2
3
4
5
P53_GENERACION

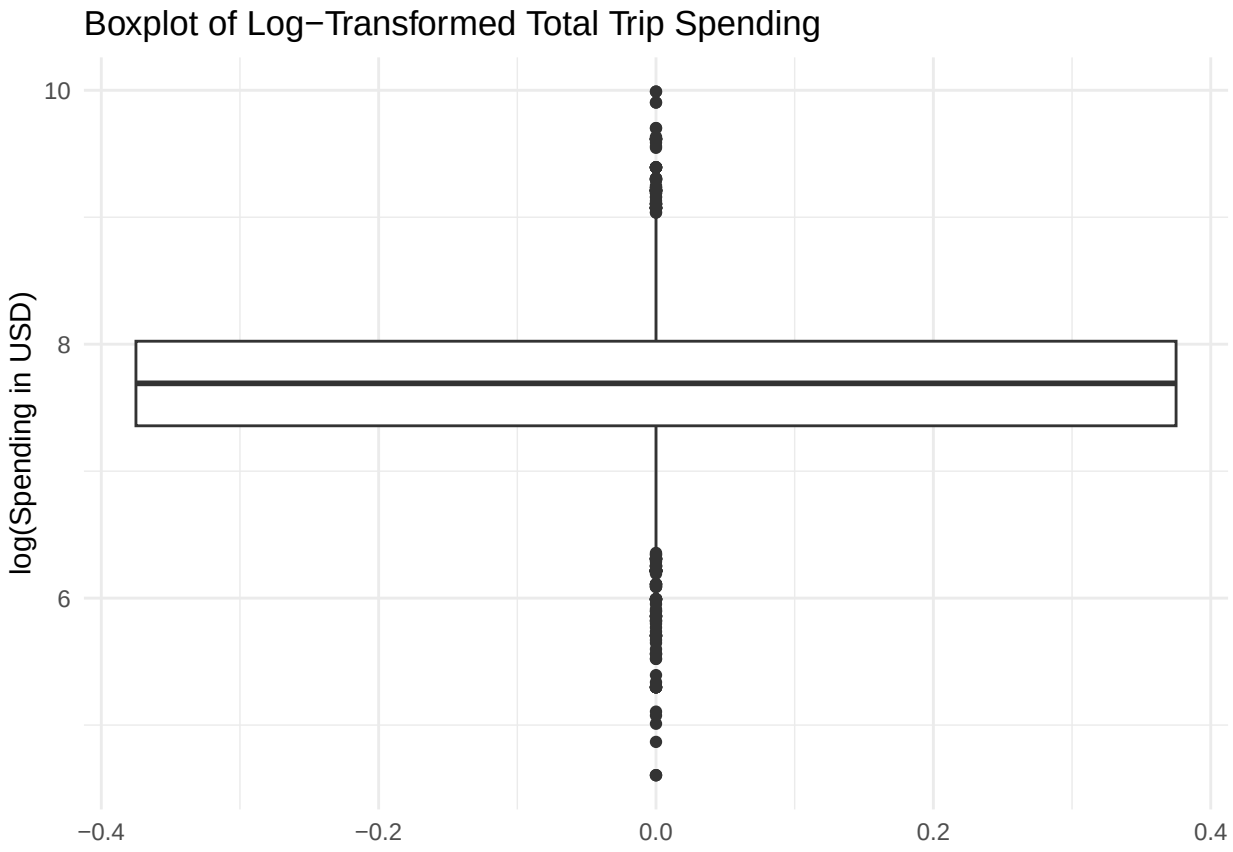
```r
# Numeric variable: log-transformed total spending (P43_1)

# Histogram of log-spending
ggplot(df, aes(x = P43_1)) +
  geom_histogram(bins = 30) +
  labs(
    title = "Distribution of Log-Transformed Total Trip Spending",
    x = "log(Spending in USD)",
    y = "Count"
  ) +
  theme_minimal()
```




```r
# Density plot (better for log data)
ggplot(df, aes(x = P43_1)) +
  geom_density(fill = "steelblue", alpha = 0.4) +
  labs(
    title = "Density of Log-Transformed Total Trip Spending",
    x = "log(Spending in USD)",
    y = "Density"
  ) +
  theme_minimal()
```




```
# Boxplot of log-spending
ggplot(df, aes(y = P43_1)) +
  geom_boxplot() +
  labs(
    title = "Boxplot of Log-Transformed Total Trip Spending",
    y = "log(Spending in USD)"
  ) +
  theme_minimal()
```



Modelling

FAMD

```
# Prepare data for FAMD (ensure unique rownames)

data_famd_use <- as.data.frame(data_fixed)
rownames(data_famd_use) <- paste0("ind_", seq_len(nrow(data_famd_use)))

# FAMD on data without NA
famd_res <- FAMD(data_famd_use, graph = FALSE)

# Individual coordinates (for clustering)
head(famd_res$ind$coord)
```

```
##           Dim.1       Dim.2      Dim.3       Dim.4      Dim.5
## ind_1 -3.441408  0.08468887  0.9397971  0.85015552 -0.3492249
## ind_2  1.424451 -1.72827325  0.1818857  0.07824803  0.3877743
## ind_3 -4.651124  0.56141428 -1.4756781 -0.32075472  2.0196949
## ind_4  2.486403  0.25063633  0.4889927 -1.16405669  2.5536124
## ind_5  1.758875 -0.11674682  2.6529358  3.46614580  2.1168386
## ind_6 -2.070407 -0.48229963  0.6777593 -0.92040027 -0.1161030
```

```r
# Contribution of variables to each dimension
var_contrib <- famd_res$var$contrib
var_contrib
```

```
##                       Dim.1      Dim.2      Dim.3        Dim.4      Dim.5
## P43_1           0.04189280  2.2762178 10.7171439 7.196885e-06  0.2062898
## P01             6.06179698  1.8690775  7.6546283 4.540816e+00  5.1463083
## P53_RANGO2     18.90836424 20.7970702  4.1467490 8.845771e-01 21.6424447
## P54             1.62690725  0.2767977  1.7025845 3.143071e-02  0.3823124
## P56             9.55379067  2.3535688  0.4505913 2.971350e-01  2.1581823
## P57_1          11.02724114 16.9453046  1.3388330 5.064192e-01 13.4982140
## P58            15.24980262 17.8946865  9.3087307 1.508990e+00 15.6216753
## P59             9.80209893  9.5548506 11.8451086 3.777636e+00  8.9994099
## P33             0.06609079  0.9374244  5.2603761 2.127422e+01  0.1369904
## P34             2.52968635  1.0447274 18.1714582 3.270333e+01  1.3483722
## P33_34_RNG      2.51132396  2.2851080 21.4788744 3.299841e+01  1.2428152
## P60             4.44288993  0.8768156  4.4744441 8.084588e-01  8.2780148
## P61             1.48493245  0.8388762  1.8464378 2.820653e-01  2.5273973
## P53_GENERACION 16.69318189 22.0494748  1.6040401 3.864993e-01 18.8115735
```

Dimension 1. This dimension is mainly defined by **age range (P53_RANGO2)**, **generational group (P53_GENERACION)**, and **trip anticipation (P58)**, with additional influence from **travel-companion type (P57_1)**, **employment status (P56)**, and **work sector (P59)**. Together, these variables form a sociodemographic axis that separates younger, spontaneous travelers from older and more structured visitors.

Dimension 2. Dimension 2 is again driven by **generation (P53_GENERACION)**, **age range (P53_RANGO2)**, **trip anticipation (P58)**, and **travel-companion characteristics (P57_1)**. Unlike Dimension 1, it emphasizes differences in planning behavior within demographic groups, distinguishing early versus late planners across age cohorts.

Dimension 3. This dimension is dominated by **trip-month variables**, particularly **P33_34_RNG** and **P34**, followed by **sector of work (P59)** and the **log-transformed spending variable (P43_1)**. These contributions indicate that seasonal timing and expenditure levels are closely related in the structure of the data.

Dimension 4. Dimension 4 reflects a pure seasonality structure, defined almost entirely by **trip-month (P34)**, **combined trip timing (P33_34_RNG)**, and **planning month (P33)**. This axis captures temporal clustering patterns that are independent from demographic or socioeconomic characteristics.

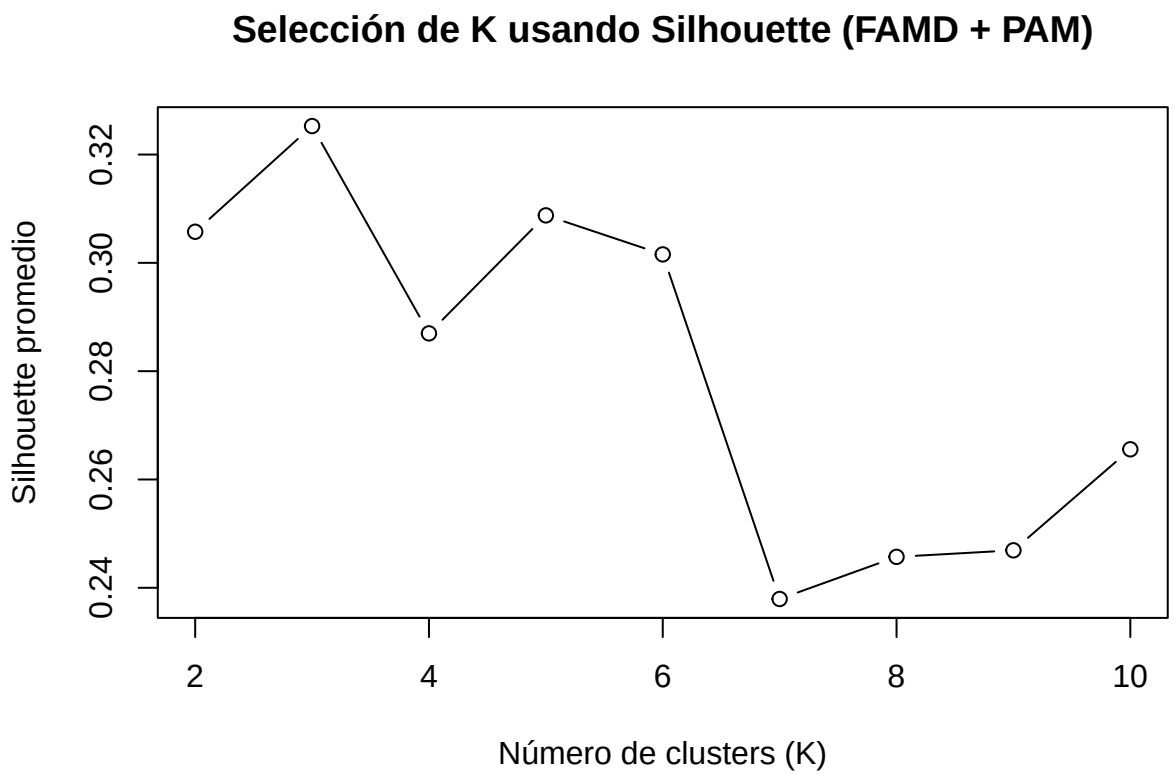
Dimension 5. This dimension reintroduces sociodemographic variation through **age range (P53_RANGO2)**, **generation (P53_GENERACION)**, **trip anticipation (P58)**, and **travel-companion patterns (P57_1)**. It highlights subtle behavioral differences within demographic groups that are not explained by the earlier dimensions.

```r
#  Distance + PAM + Silhouette (k = 2 to 10)
dist_famd <- dist(famd_res$ind$coord)

max_k <- 10
sil_famd <- numeric(max_k)
```

```r
for (k in 2:max_k) {
  pam_fit     <- pam(dist_famd, k = k, diss = TRUE)
  sil_vals    <- silhouette(pam_fit$clustering, dist_famd)
  sil_famd[k] <- mean(sil_vals[, 3])
  cat("k =", k, "| Silhouette (FAMD+PAM) =", round(sil_famd[k], 4), "\n")
}
```

```
## k = 2 | Silhouette (FAMD+PAM) = 0.3057
## k = 3 | Silhouette (FAMD+PAM) = 0.3253
## k = 4 | Silhouette (FAMD+PAM) = 0.287
## k = 5 | Silhouette (FAMD+PAM) = 0.3088
## k = 6 | Silhouette (FAMD+PAM) = 0.3016
## k = 7 | Silhouette (FAMD+PAM) = 0.2379
## k = 8 | Silhouette (FAMD+PAM) = 0.2457
## k = 9 | Silhouette (FAMD+PAM) = 0.2469
## k = 10 | Silhouette (FAMD+PAM) = 0.2656
```

```r
plot(2:max_k, sil_famd[2:max_k], type = "b",
     xlab = "Número de clusters (K)",
     ylab = "Silhouette promedio",
     main = "Selección de K usando Silhouette (FAMD + PAM)")
```



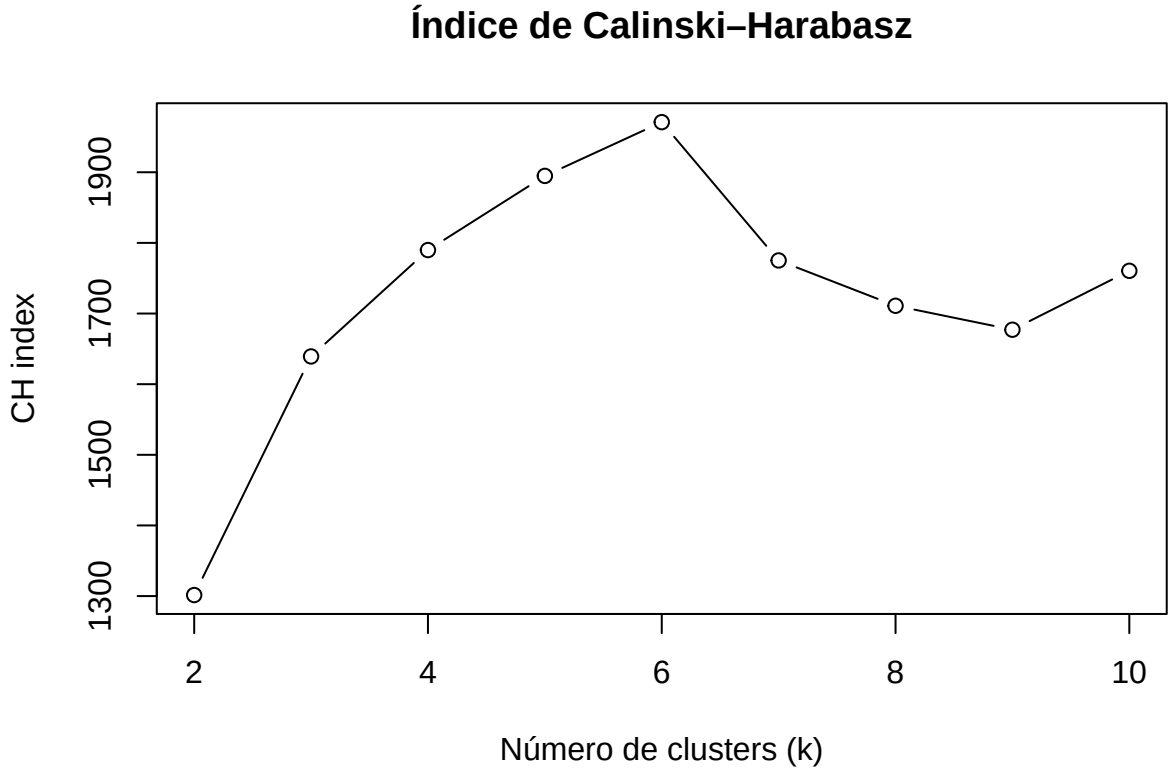

```r
# Select optimal k (maximum silhouette between 2 and max_k)
k_opt <- which.max(sil_famd[2:max_k]) + 1  # +1 porque el vector empieza en k=2
cat("k óptimo elegido:", k_opt, "\n")
```

```
## k óptimo elegido: 3
```

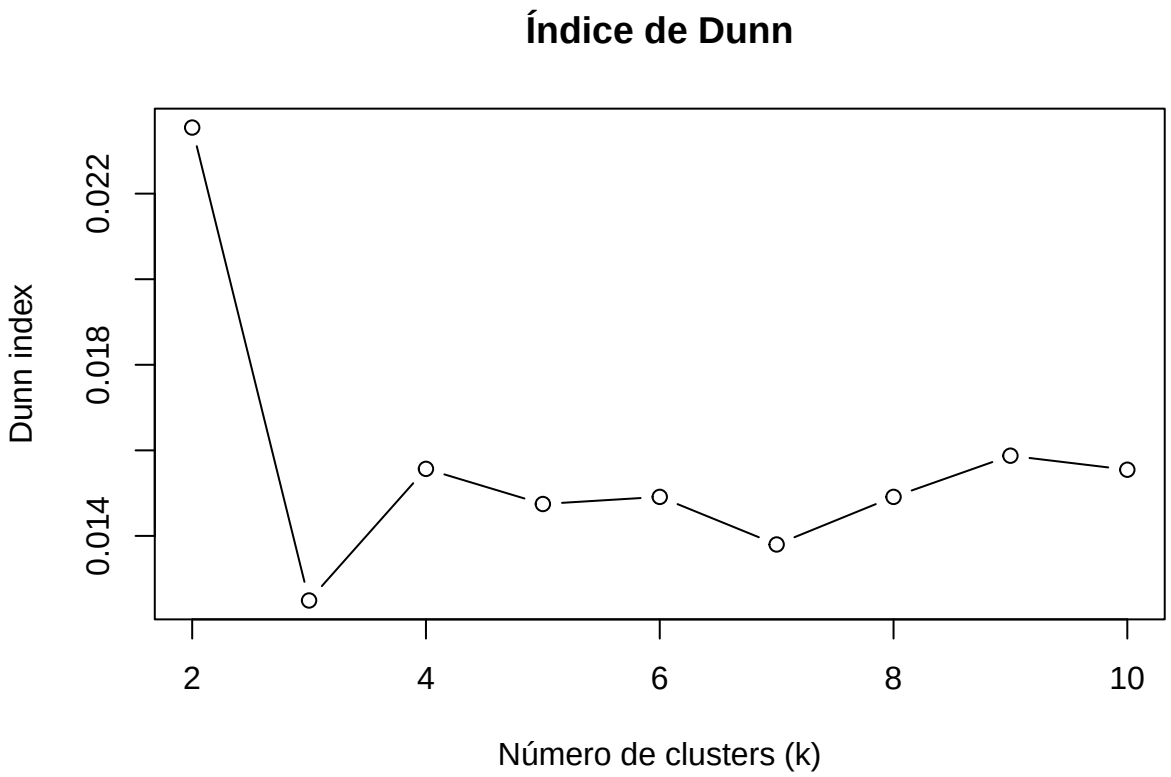
```r
# CALINSKI-HARABASZ INDEX
install.packages("clusterCrit")
```

```
## Warning: package 'clusterCrit' is in use and will not be installed
```

```r
library(clusterCrit)

# Convert FAMD coordinates to numeric matrix
X <- as.matrix(famd_res$ind$coord)

# Evaluate CH for k = 2 to 10
ch_values <- numeric(max_k)

for (k in 2:max_k) {
  pam_fit <- pam(dist_famd, k = k, diss = TRUE)

  ch_values[k] <- intCriteria(
    X,
    as.integer(pam_fit$clustering),
    c("calinski_harabasz")
  )$calinski_harabasz
}

plot(2:max_k, ch_values[2:max_k], type = "b",
     main = "Índice de Calinski-Harabasz",
     xlab = "Número de clusters (k)",
     ylab = "CH index")
```



```r
# DUNN INDEX

dunn_values <- numeric(max_k)

for (k in 2:max_k) {
  pam_fit <- pam(dist_famd, k = k, diss = TRUE)

  dunn_values[k] <- intCriteria(
    X,
    as.integer(pam_fit$clustering),
    c("dunn")
  )$dunn
}

plot(2:max_k, dunn_values[2:max_k], type = "b",
     main = "Índice de Dunn",
     xlab = "Número de clusters (k)",
     ylab = "Dunn index")
```



Selection of the number of clusters. The Dunn index attains its highest value at $K = 2$, indicating excellent compactness and separation at this level; however, it drops sharply from $K = 3$ onward and remains low and stable across larger values, suggesting increased overlap between clusters when $K > 2$. In contrast, the Calinski–Harabasz index increases steadily and reaches its maximum at $K = 6$, after which it declines, indicating that overall between-group separation relative to within-group cohesion is strongest around six clusters. The Silhouette index achieves its best value at $K = 3$, with a secondary local improvement at $K = 5$, while solutions from $K = 6$ to $K = 9$ show considerably weaker structure.

Taken together, these indices suggest that no single value of K is optimal across all criteria: $K = 2$ maximizes

Dunn but produces overly coarse segmentation; $K = 3$ maximizes Silhouette but may overlook relevant subgroup variation; and $K = 6$ maximizes the Calinski–Harabasz index but risks generating less interpretable clusters. Considering the objective of obtaining differentiated, interpretable, and actionable tourist profiles, and given that both Silhouette and Calinski–Harabasz show local support for intermediate solutions, selecting $K = 5$ represents a balanced and defensible compromise, even if it is not the strictly optimal solution for any single index.

Final clustering with PAM

```
#  Final clustering with PAM

pam_famd <- pam(dist_famd, k = 5, diss = TRUE)
clusters_famd <- factor(pam_famd$clustering)

table(clusters_famd)
```

```
## clusters_famd
##    1    2    3    4    5
##  615  919 1076 1983  675
```

```
# Final average silhouette
sil_famd_kopt <- silhouette(pam_famd$clustering, dist_famd)
mean_sil_kopt <- mean(sil_famd_kopt[, 3])
cat("Silhouette promedio (k_opt) =", round(mean_sil_kopt, 4), "\n")
```

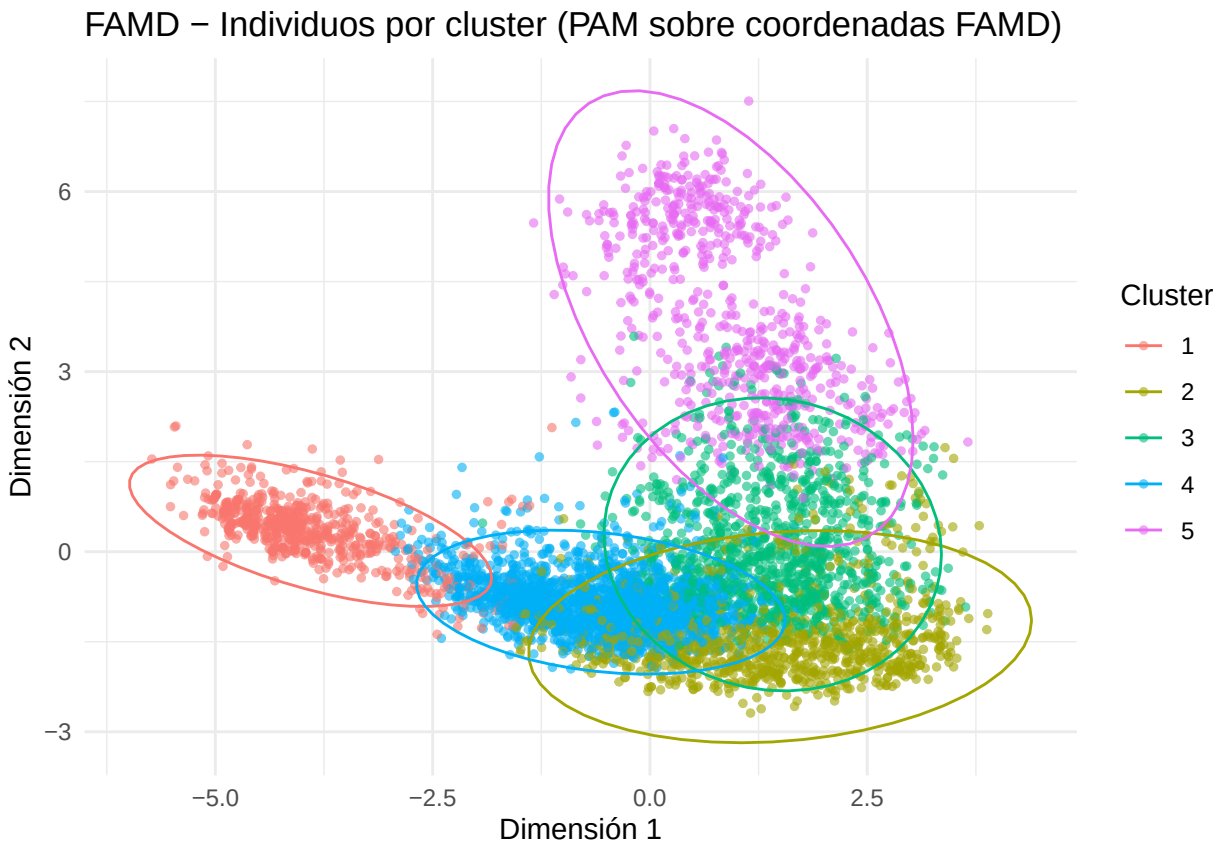
```
## Silhouette promedio (k_opt) = 0.3088
```

```
# Add clusters to dataset
data_clust <- data_famd_use %>%
  mutate(cluster = clusters_famd)

table(data_clust$cluster)
```

```
##
##    1    2    3    4    5
##  615  919 1076 1983  675
```

```
# FAMD plot: individuals by cluster (with ggplot2)

# Using coordinates of the first two dimensions:
ind_coords <- as.data.frame(famd_res$ind$coord[, 1:2])
colnames(ind_coords) <- c("Dim1", "Dim2")
ind_coords$cluster <- clusters_famd

ggplot(ind_coords, aes(x = Dim1, y = Dim2, color = cluster)) +
  geom_point(alpha = 0.6, size = 1) +
  stat_ellipse(aes(group = cluster), type = "norm", level = 0.95) +
  labs(
    title = "FAMD - Individuos por cluster (PAM sobre coordenadas FAMD)",
    x = "Dimensión 1",
```

```
    y = "Dimensión 2",
    color = "Cluster"
  ) +
  theme_minimal()
```



Interpretation of the Final PAM Clustering ($k = 5$). After selecting $k = 5$ as the preferred number of clusters, the PAM algorithm was applied to the FAMD distance matrix. The resulting clusters show a balanced and clearly differentiated structure, with the following group sizes: Cluster~1: 615 individuals; Cluster~2: 919 individuals; Cluster~3: 1{,}076 individuals; Cluster~4: 1{,}983 individuals; and Cluster~5: 675 individuals. Cluster~4 is noticeably the largest group, while Clusters~1 and~5 are smaller but still substantial, indicating that none of the clusters is excessively small or unstable due solely to sample size.

The average Silhouette value for the selected solution is Silhouette($k = 5$) = 0.3088. A value around 0.30 . This supports the conclusion that the five-cluster solution provides meaningful, interpretable, and statistically sound segmentation without imposing overly rigid cluster boundaries.

Cluster stability analysis

```
# Cluster Stability Analysis

# Wrapper function for PAM (required by clusterboot)
pam_5 <- function(x, k) {
  pam(x, k = k, diss = FALSE)$clustering
}
```

```r
# Cluster stability analysis with bootstrap
set.seed(123)
invisible(
  capture.output(
    cb <- clusterboot(
      X,
      B            = 300,         # number of bootstrap samples
      bootmethod   = "boot",      # resample rows with replacement
      clustermethod = pamkCBI,    # PAM method compatible with clusterboot
      k            = 5,           # selected number of clusters
      seed         = 123,
      showplots    = FALSE
    )
  )
)

cb
```

```
## * Cluster stability assessment *
## Cluster method:  pam/estimated k
## Full clustering results are given as parameter result
## of the clusterboot object, which also provides further statistics
## of the resampling results.
## Number of resampling runs:  300
##
## Number of clusters found in data:  5
##
##  Clusterwise Jaccard bootstrap (omitting multiple points) mean:
## [1] 0.9823544 0.8757553 0.8379446 0.9127500 0.9324671
## dissolved:
## [1]  0  0 23  0  0
## recovered:
## [1] 300 264 237 279 296
```

Cluster stability was evaluated through bootstrap resampling using the `clusterboot` procedure. The resulting Jaccard similarity coefficients—[0.98, 0.88, 0.84, 0.91, 0.93]—indicate strong and consistent recovery of the five clusters across resampled datasets. Clusters~1, 4, and~5 exhibit extremely high stability (Jaccard > 0.90), Cluster~2 shows very high stability (0.88), and Cluster~3 reaches acceptable stability (0.84), despite showing a small number of dissolutions (23 out of 300 runs). All other clusters experienced zero dissolutions, confirming excellent reproducibility. Overall, these results demonstrate that the five-cluster solution is empirically robust and that the identified segments persist under resampling rather than arising from random variation.

Results

```r
# Cluster profiles (categorical variables)

perfil_var_cluster <- function(var_name) {
  tab <- table(data_clust[[var_name]], data_clust$cluster)
  prop_clust <- prop.table(tab, margin = 2)  # % within each cluster
  prop_total <- prop.table(tab)              # % overall
```

```
  list(
    tabla            = tab,
    prop_por_cluster = round(prop_clust * 100, 1),
    prop_global      = round(prop_total * 100, 1)
  )
}

# Variables used in FAMD (same vars_final),
# excluding numeric P43_1 from categorical profiling:
vars_perfil <- intersect(
  c("P01", "P53_RANGO2", "P54", "P56", "P57_1", "P58",
    "P59", "P33", "P34", "P33_34_RNG", "P60", "P61", "P53_GENERACION"),
  names(data_clust)
)

# Create a list of profiles per variable
lista_perfiles <- lapply(vars_perfil, perfil_var_cluster)
names(lista_perfiles) <- vars_perfil

# --- FUNCTION TO EXTRACT DETAILED VARIABLE INFORMATION ---
info_variable <- function(v) {
  etiqueta   <- attr(data_sav[[v]], "label")
  categorias <- attr(data_sav[[v]], "labels")

  cat("\n============================\n")
  cat("Variable:", v, "\n")
  cat("Question:", ifelse(is.null(etiqueta), "NO LABEL AVAILABLE", etiqueta), "\n")

  if (!is.null(categorias)) {
    cat("Categories (original SPSS labels):\n")
    print(categorias)
  } else {
    cat("Categories: NOT APPLICABLE (numeric or no labels)\n")
  }

  cat("----------------------------\n")
}

# --- DISPLAY CLUSTER PROFILES FOR CATEGORICAL VARIABLES ---
for (v in vars_perfil) {
  info_variable(v)   # show variable name + question + categories

  cat("Percentage distribution by cluster (%):\n")
  print(lista_perfiles[[v]]$prop_por_cluster)

  cat("\n")
}
```

```
##
## ============================
## Variable: P01
## Question: P01. Según la siguiente tarjeta ¿Cuál fue su principal motivo de visita al Perú? [AGRUPADO
## Categories (original SPSS labels):
```

```
##                                Vacaciones, recreación u ocio
##                                                            1
##                               Visitar a familiares o amigos.
##                                                            2
##                                                     Negocios
##                                                            3
## Asistir a seminarios, conferencias, convenciones o congresos
##                                                            5
##                                                    Educación
##                                                            6
##                Salud (tratamiento médico) / comprar medicina
##                                                            7
##                  Misiones / trabajo religioso / voluntariado
##                                                            9
##                                                       Otros
##                                                          94
## ----------------------------
## Percentage distribution by cluster (%):
##
##         1    2    3    4    5
##   1  63.6 28.9 63.8 82.7 68.1
##   2  19.8  8.3 14.8 12.3 17.3
##   3   0.5 60.3 11.8  2.8  9.6
##   5   1.5  1.4  6.1  0.7  2.4
##   6   4.9  0.2  1.0  0.6  0.4
##   7   0.3  0.4  0.1  0.1  0.3
##   9   9.3  0.4  2.4  0.9  1.8
##   94  0.2  0.0  0.0  0.0  0.0
##
##
## ============================
## Variable: P53_RANGO2
## Question: ¿Qué edad tiene usted? (RANGO AGRUPADO)
## Categories (original SPSS labels):
##  De 15 a 24 años  De 25 a 34 años  De 35 a 54 años De 55 años a más
##                1                2                3                4
## ----------------------------
## Percentage distribution by cluster (%):
##
##        1    2    3    4    5
##   1 80.8  3.4  0.2  7.5  0.0
##   2 18.4 30.9  2.0 65.3  0.0
##   3  0.7 60.6 79.2 26.6  2.5
##   4  0.2  5.1 18.6  0.7 97.5
##
##
## ============================
## Variable: P54
## Question: P54. Género
## Categories (original SPSS labels):
## Masculino  Femenino
##         1         2
## ----------------------------
## Percentage distribution by cluster (%):
```

```
##
##        1    2    3    4    5
##   1 43.6 74.8 53.1 51.0 58.8
##   2 56.4 25.2 46.9 49.0 41.2
##
##
## ============================
## Variable: P56
## Question: P56. Según la siguiente tarjeta, ¿Cuál es su condición actual?
## Categories (original SPSS labels):
##                                       Soltero
##                                             1
##                           Casado o conviviente
##                                             2
## Parte de una pareja no casado ni conviviente
##                                             3
##                                     No precisa
##                                            99
## ----------------------------
## Percentage distribution by cluster (%):
##
##         1    2    3    4    5
##   1  82.1 41.9 23.0 60.1 27.3
##   2   2.6 51.5 71.8 21.9 67.4
##   3  14.8  6.3  3.6 17.4  2.4
##   99  0.5  0.3  1.6  0.5  3.0
##
##
## ============================
## Variable: P57_1
## Question: P57. ¿En cuál de las siguientes situaciones se encuentra actualmente?
## Categories (original SPSS labels):
##                                                        No tengo hijos
##                                                                     1
##                                         Tengo hijos de 0 a 14 años
##                                                                     2
##                                        Tengo hijos de 15 a 18 años
##                                                                     3
##                    Tengo hijos mayores de 18 años que viven en casa
##                                                                     4
## Tengo hijos mayores de 18 años que viven de manera independiente
##                                                                     5
## ----------------------------
## Percentage distribution by cluster (%):
##
##        1    2    3    4    5
##   1 98.7 49.4 23.5 94.4 13.2
##   2  1.0 37.3 29.3  4.9  1.5
##   3  0.2  4.9 16.9  0.1  1.8
##   4  0.2  4.4 20.3  0.1 12.9
##   5  0.0  4.0 10.0  0.5 70.7
##
##
## ============================
```

```
## Variable: P58
## Question: P58. Según la siguiente tarjeta, ¿En cuál de las siguientes categorías se encuentra actualm
## Categories (original SPSS labels):
## Trabajador del sector público Trabajador del sector privado
##                             1                             2
##      Trabajador independiente                    Estudiante
##                             3                             4
##                   Ama de casa                      Jubilado
##                             5                             6
##                   Desempleado                    No precisa
##                             7                            99
## ----------------------------
## Percentage distribution by cluster (%):
##
##         1    2    3    4    5
##   1   2.9  5.4 35.5 21.8 11.9
##   2   2.9 77.3 47.3 59.1 25.6
##   3   2.6 16.2 13.9 16.4 14.7
##   4  83.6  0.0  0.3  0.4  0.0
##   5   0.7  0.3  2.0  0.1  2.2
##   6   0.0  0.0  0.1  0.0 44.4
##   7   5.9  0.3  0.5  1.2  0.4
##   99  1.5  0.4  0.4  1.1  0.7
##
##
## ============================
## Variable: P59
## Question: P59. ¿En qué rubro trabaja?
## Categories (original SPSS labels):
##                   Industria manufacturera
##                                         1
##                                  Servicios
##                                         2
##                                  Comercio
##                                         3
##                               Construcción
##                                         4
##                  Agricultura y ganadería
##                                         5
##                                    Energía
##                                         6
##              Tecnología y comunicaciones
##                                         7
##                      Turismo y hotelería
##                                         8
##                                     Salud
##                                         9
## Medios de comunicación y entretenimiento
##                                        10
##                   Transporte y logística
##                                        11
##                                 Educación
##                                        12
##                       Finanzas y seguros
```

```
##                                     13
##                               Inmobiliaria
##                                     14
##    Consultoría y servicios profesionales
##                                     15
##                                  Deporte
##                                     16
##                                  Minería
##                                     17
##                                Ingeniería
##                                     18
##                                 Marketing
##                                     19
##                                  Biología
##                                     20
##                           Policia/militar
##                                     21
##                       Ecología / Ambiental
##                                     22
##                  Arte / Diseño /Fotografía
##                                     23
##                        Trabajo Social /ONG
##                                     24
##                                Arqueología
##                                     25
##                                Gastronomía
##                                     26
##                   Fashion /moda/cosméticos
##                                     27
##                   Legal / Gobierno / Leyes
##                                     28
##                                Psicología
##                                     29
##                 Representante sindicalista
##                                     30
##                                   Otros
##                                     94
##                                No precisa
##                                     99
## ----------------------------
## Percentage distribution by cluster (%):
##
##            1    2    3    4    5
##   1      0.7 11.8  3.5  3.5  2.8
##   10     0.3  1.6  1.0  2.3  0.7
##   11     0.2  5.0  3.1  3.3  1.8
##   12     1.1  1.7 22.5 10.9  4.9
##   13     0.3  3.7  4.9  6.6  3.3
##   14     0.2  1.3  1.3  1.2  1.0
##   15     0.5  3.8  4.0  4.4  1.3
##   16     0.3  3.0  1.3  1.1  0.7
##   17     0.0  2.9  0.4  0.7  0.3
##   18     0.0  4.9  1.5  3.2  2.5
##   19     0.5  1.1  0.2  2.2  0.4
```

```
##   2      1.5  7.5 12.1 11.0  7.9
##   20     0.3  0.2  0.7  0.3  0.4
##   21     0.3  0.4  1.3  0.9  0.3
##   22     0.0  0.8  0.8  0.5  0.0
##   23     0.2  0.7  0.2  0.9  0.7
##   24     0.0  0.7  1.1  0.5  0.3
##   25     0.0  0.0  0.0  0.2  0.0
##   26     0.2  0.8  0.2  0.5  0.0
##   27     0.0  0.7  0.0  0.1  0.1
##   28     0.0  0.0  0.3  0.1  0.0
##   29     0.0  0.0  0.0  0.1  0.0
##   3      0.7  8.8  5.9  6.7  4.1
##   30     0.0  0.0  0.0  0.1  0.0
##   4      0.0  5.1  2.1  4.5  2.8
##   5      0.2  2.9  1.8  0.9  2.1
##   6      0.0  4.8  1.3  3.0  0.7
##   7      0.2 15.1  6.5  9.6  3.0
##   8      0.0  3.5  1.5  4.2  2.4
##   9      1.0  5.8 17.0 14.0  6.8
##   94     0.0  0.3  0.3  0.4  0.3
##   99     0.0  0.0  0.0  0.0  0.3
##   NS/NR 91.5  1.1  3.3  2.7 47.9
##
##
## ============================
## Variable: P33
## Question: P33. ¿En qué año compró su pasaje y /o paquete para realizar este viaje?
## Categories (original SPSS labels):
## Año 2023 Año 2024 Año 2022    Otros
##        1        2        3       94
## ----------------------------
## Percentage distribution by cluster (%):
##
##        1    2    3    4    5
##   1  5.4  3.2  4.5  3.9  7.3
##   2 94.6 96.8 95.4 96.0 92.7
##   3  0.0  0.0  0.2  0.1  0.0
##
##
## ============================
## Variable: P34
## Question: P34. ¿En qué mes compró su pasaje y /o paquete para realizar este viaje?
## Categories (original SPSS labels):
##     Enero   Febrero     Marzo     Abril      Mayo     Junio     Julio    Agosto
##         1         2         3         4         5         6         7         8
## Setiembre   Octubre Noviembre Diciembre
##         9        10        11        12
## ----------------------------
## Percentage distribution by cluster (%):
##
##         1    2    3    4    5
##   1   5.2  0.2  5.2  6.9  3.6
##   2   5.9  0.3  5.8  9.4  6.1
##   3  11.2  0.2  9.6  9.3  7.4
```

```
##   4  11.2  0.4  8.7  9.9  8.6
##   5  10.4  0.8 10.5 13.3 12.4
##   6  14.8  2.1 15.1 15.5 11.7
##   7  13.7 18.4 13.8 11.1 13.5
##   8  10.4 29.7  8.2  5.8 12.4
##   9   4.2  1.2  7.6  8.1  6.4
##   10  6.7  4.5 10.3  7.7  8.3
##   11  4.7 40.6  4.1  1.0  8.0
##   12  1.6  1.6  1.1  2.0  1.6
##
##
## ============================
## Variable: P33_34_RNG
## Question: P34. ¿En qué mes compró su pasaje y /o paquete para realizar este viaje? [EN RANGO]
## Categories (original SPSS labels):
##  Menos de 1 mes  De 1 a 4 meses  De 5 a 8 meses De 9 a 12 meses Más de 12 meses
##               1               2               3               4               5
##     No Responde
##              99
## ----------------------------
## Percentage distribution by cluster (%):
##
##        1    2    3    4    5
##   1 15.1 82.7 10.3  3.4 16.0
##   2 61.6 13.2 66.4 66.3 58.4
##   3 19.5  1.7 19.9 26.5 18.8
##   4  2.9  2.4  2.0  3.5  5.6
##   5  0.8  0.0  1.4  0.3  1.2
##
##
## ============================
## Variable: P60
## Question: P60. Según la siguiente tarjeta, ¿Cuál es el grado de instrucción más alto concluido por us
## Categories (original SPSS labels):
##      Primaria    Secundaria       Técnica Universitaria    Post Grado
##             1             2             3             4             5
##      Maestría     Doctorado
##             6             7
## ----------------------------
## Percentage distribution by cluster (%):
##
##        1    2    3    4    5
##   1  0.7  0.1  1.1  0.2  1.5
##   2 36.9  4.4  5.7  4.9  6.1
##   3  4.6  6.1  7.4  5.5 10.2
##   4 44.9 50.7 30.5 50.7 44.3
##   5  4.7 19.7 16.3 15.7 12.7
##   6  7.6 18.1 23.5 21.2 17.8
##   7  0.7  1.0 15.5  1.7  7.4
##
##
## ============================
## Variable: P61
## Question: P61. Según la siguiente tarjeta ¿cuál es su ingreso familiar anual?
```

```
## Categories (original SPSS labels):
##          Menos de US$ 20,000   De US$ 20,000 a US$ 39,999
##                            1                            2
##   De US$ 40,000 a US$ 59,999   De US$ 60,000 a US$ 79,999
##                            3                            4
##   De US$ 80,000 a US$ 99,999 De US$ 100,000 a US$ 119,999
##                            5                            6
## De US$ 120,000 a US$ 139,999 De US$ 140,000 a US$ 159,999
##                            7                            8
## De US$ 160,000 a US$ 179,999 De US$ 180,000 a US$ 199,999
##                            9                           10
##            US$ 200,000 o más                   No precisa
##                           11                           99
## ----------------------------
## Percentage distribution by cluster (%):
##
##         1    2    3    4    5
##   1  10.2  6.2  3.7  7.1  4.3
##   2   8.0 11.1  6.7 10.5  4.1
##   3   6.5  8.2  5.1 10.2  5.5
##   4   2.4  4.1  5.9  7.0  4.4
##   5   2.4  4.8  5.9  4.5  4.9
##   6   1.8  5.0  4.6  2.9  5.3
##   7   0.7  2.6  3.2  1.9  3.4
##   8   0.5  1.3  2.2  1.7  2.5
##   9   1.5  1.0  3.0  0.4  0.3
##   10  0.3  0.4  2.4  0.2  0.6
##   11  2.3  3.8  5.9  1.2  6.2
##   99 63.4 51.5 51.3 52.5 58.4
##
##
## ============================
## Variable: P53_GENERACION
## Question: P53 Grupo Generacional
## Categories (original SPSS labels):
##           Centennials           Millennials          Generación X
##                     1                     2                     3
##           Baby Boomer Generación Silenciosa
##                     4                     5
## ----------------------------
## Percentage distribution by cluster (%):
##
##        1    2    3    4    5
##   1 96.9 14.9  0.6 41.7  0.0
##   2  2.6 65.0 32.0 55.9  0.1
##   3  0.3 19.5 66.6  2.3 17.8
##   4  0.2  0.7  0.8  0.1 80.3
##   5  0.0  0.0  0.0  0.0  1.8
```

```
# Numerical summary per cluster


# FUNCTION TO SHOW NUMERIC VARIABLE INFO (NAME + QUESTION)
show_numeric_info <- function(var) {
```

```
  etiqueta <- attr(data_sav[[var]], "label")

  cat("\n============================\n")
  cat("Numeric variable:", var, "\n")
  cat("Question:", ifelse(is.null(etiqueta), "NO LABEL AVAILABLE", etiqueta), "\n")
  cat("----------------------------\n")
}

# Add original-scale spending (USD) for reporting
data_clust <- data_clust %>%
  mutate(
    P43_1_usd = exp(P43_1)  # back-transform from log
  )

# Optional: add a descriptive label for the back-transformed variable
attr(data_clust$P43_1_usd, "label") <-
  "Total trip spending in USD (back-transformed from log)"

# Identify numeric variables in data_clust
num_vars <- names(data_clust)[sapply(data_clust, is.numeric)]

# Remove the log version of P43_1 from the summary (we only want original-scale USD)
num_vars <- setdiff(num_vars, "P43_1")

# Show info for numeric variables (only those that exist in data_sav or are P43_1_usd)
for (var in num_vars) {
  if (var %in% names(data_sav)) {
    show_numeric_info(var)
  } else if (var == "P43_1_usd") {
    cat("\n============================\n")
    cat("Numeric variable: P43_1_usd\n")
    cat("Question: Total trip spending in USD (back-transformed from log)\n")
    cat("----------------------------\n")
  }
}
```

```
##
## ============================
## Numeric variable: P43_1_usd
## Question: Total trip spending in USD (back-transformed from log)
## ----------------------------
```

```
# COMPUTE NUMERIC SUMMARY PER CLUSTER (using original-scale spending)
resumen_numerico <- data_clust %>%
  select(cluster, all_of(num_vars)) %>%
  group_by(cluster) %>%
  summarise(
    across(
      .cols = where(is.numeric),
      .fns  = list(
        mean = ~ mean(.x, na.rm = TRUE),
        sd   = ~ sd(.x, na.rm = TRUE)
      ),
```

```
      .names = "{.col}_{.fn}"
    ),
    .groups = "drop"
  )

cat("\nNUMERIC SUMMARY BY CLUSTER (original scale for spending):\n")
```

```
##
## NUMERIC SUMMARY BY CLUSTER (original scale for spending):
```

```
print(resumen_numerico)
```

```
## # A tibble: 5 x 3
##   cluster P43_1_usd_mean P43_1_usd_sd
##   <fct>            <dbl>        <dbl>
## 1 1                2410.        1864.
## 2 2                1820.        1141.
## 3 3                3070.        2131.
## 4 4                2576.        1490.
## 5 5                3186.        2170.
```

Main Defining Variables by Cluster (with Percentages)

- **Cluster 1 – Very young student leisure travelers**
  Dominant characteristics:
  - **Age (P53_RANGO2):** 15–24 years (**80.8%**); 25–34 years (18.4%).
  - **Generation (P53_GENERACION):** Centennials (**96.9%**).
  - **Marital Status (P56):** Single (**82.1%**).
  - **Children (P57_1):** No children (**98.7%**).
  - **Occupation (P58):** Students (**83.6%**).
  - **Main Motive (P01):** Vacations / leisure (**63.6%**).
  - **Education (P60):** University completed (44.9%), secondary (36.9%).
  - **Planning Horizon (P33_34_RNG):** 1–4 months (**61.6%**).
  - **Spending (P43_1_usd):** Mean USD **2,410**.
- **Cluster 2 – Business-oriented working Millennials**
  Dominant characteristics:
  - **Main Motive (P01):** Business (**60.3%**).
  - **Age (P53_RANGO2):** 35–54 years (60.6%), 25–34 years (30.9%).
  - **Generation:** Millennials (65.0%), Gen X (19.5%).
  - **Marital Status:** Married/partnered (**51.5%**).
  - **Children:** Children under 14 (**37.3%**).
  - **Occupation:** Private-sector workers (**77.3%**).
  - **Sector (P59):** Services (7.5%), transport (15.1%), health (5.8%).
  - **Planning Horizon:** <1 month (**82.7%**).
  - **Education:** University completed (**50.7%**); postgraduate (18.1%).

  - **Spending:** Mean USD **1,820**.

- **Cluster 3 – Mid-life families with higher spending**
  Dominant characteristics:
  - **Age:** 35–54 years (**79.2%**); 55+ (18.6%).
  - **Generation:** Gen X (**66.6%**).
  - **Marital Status:** Married (**71.8%**).
  - **Children:** Children 0–14 (**29.3%**); children >18 (**20.3%**).
  - **Occupation:** Private-sector (47.3%), public-sector (35.5%).
  - **Sector:** Education (**22.5%**), health (17.0%), services (12.1%).
  - **Motive:** Vacations (63.8%).
  - **Planning Horizon:** 1–4 months (66.4%).
  - **Education:** University (30.5%), master's (23.5%), doctorate (15.5%).
  - **Spending:** Mean USD **3,070**.

- **Cluster 4 – Young adult leisure travelers**
  Dominant characteristics:
  - **Age:** 25–34 years (**65.3%**); 15–24 years (7.5%).
  - **Generation:** Millennials (**55.9%**), Centennials (41.7%).
  - **Marital Status:** Single (**60.1%**).
  - **Children:** No children (**94.4%**).
  - **Motive:** Vacations (**82.7%**).
  - **Occupation:** Private-sector (59.1%), students (21.8%).
  - **Planning Horizon:** 1–4 months (**66.3%**).
  - **Education:** Mainly university (50.7%).
  - **Spending:** Mean USD **2,576**.

- **Cluster 5 – Senior, high-income, high-spending travelers**
  Dominant characteristics:
  - **Age:** 55+ years (**97.5%**).
  - **Generation:** Baby Boomers (**80.3%**).
  - **Marital Status:** Married (**67.4%**).
  - **Children:** Adult children living independently (**70.7%**).
  - **Occupation:** Retired (**44.4%**).
  - **Motive:** Vacations (**68.1%**); visit family (17.3%).
  - **Planning Horizon:** 1–4 months (**58.4%**).
  - **Education:** University (44.3%), master's (17.8%), doctorate (7.4%).
  - **Spending:** Highest mean: USD **3,186**.

Conclusion

This study demonstrates the effectiveness of unsupervised machine-learning methods for uncovering latent behavioral and demographic structures within the 2024 Foreign Tourist Profile Survey. By applying a rigorous workflow centered on dimensionality reduction (FAMD) and partitioning clustering (PAM), the analysis successfully identified complex patterns in a dataset characterized by mixed variable types, nonlinear relationships, and high heterogeneity—conditions for which unsupervised learning is particularly powerful. These methods enabled the discovery of natural groupings without imposing prior assumptions or predefined categories, allowing the data itself to reveal the underlying segmentation of international visitors to Peru.

The FAMD analysis, an unsupervised multivariate technique, reduced the complexity of the dataset while preserving key sources of variance. This step ensured that the clustering did not rely on raw, noisy, or redundant variables, but on a condensed representation that captures the true structural relationships among travelers.

Using this FAMD space, the PAM clustering algorithm, a classic unsupervised method designed for mixed-data applications, was used to partition the tourist population. Multiple internal validation metrics—Silhouette, Dunn, and Calinski–Harabasz—were evaluated to avoid relying on a single criterion. Although the indices pointed to different optimal k values (a frequent occurrence in unsupervised learning), their combined interpretation and practical considerations identified five clusters as the most balanced and actionable solution.

Crucially, the reliability of the unsupervised segmentation was assessed through bootstrap-based stability analysis using clusterboot. The resulting Jaccard coefficients (0.84–0.98) confirmed that the five clusters are not artifacts of noise or sampling variation but stable structural components consistently reproduced across resampled datasets. This step is essential in unsupervised learning, where model validation does not rely on ground-truth labels but on the reproducibility of discovered patterns.

The resulting five-segment structure—ranging from very young student leisure travelers to senior high-income visitors—captures meaningful and interpretable market segments that emerge organically from the data. Differences observed across age, generation, household structure, planning behavior, occupational status, travel motivation, and spending levels illustrate the power of unsupervised methods to reveal hidden behavioral profiles that supervised models or traditional descriptive statistics would not detect.

Overall, the study provides strong evidence that unsupervised learning is a robust and insightful approach for segmenting heterogeneous populations in tourism analytics. Through careful preprocessing, dimensionality reduction, clustering, and stability validation, the analysis delivered a statistically sound and practically useful segmentation. The five-cluster solution offers a comprehensive, data-driven understanding of international tourist behavior—supporting more precise marketing strategies, product differentiation, and policy planning in Peru's tourism sector.