Insights into Canada’s Evolving Labor Force

Azad

# Define the file path and Read CSV

full\_dataset\_path <- "C:/820WF/LFS\_2017\_2023/lfs\_17\_23.csv"  
LFS\_data <- read.csv(full\_dataset\_path)

# Select a smaller subset of data and Combine Categories

# Select only the relevant variables  
LFS\_Selected <- LFS\_data[LFS\_data$SURVYEAR >= 2023, c("SURVMNTH","LFSSTAT", "PROV", "AGE\_12", "SEX", "MARSTAT", "EDUC", "IMMIG")]

#Univariate Analysis

library(ggplot2)   
library(dplyr)

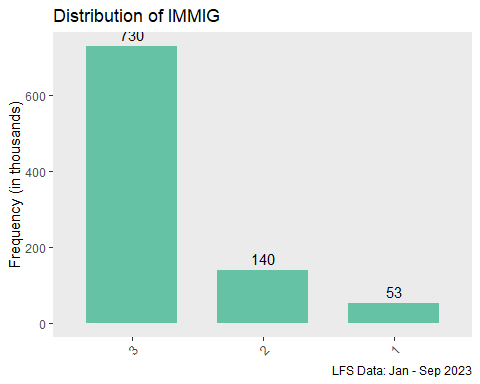
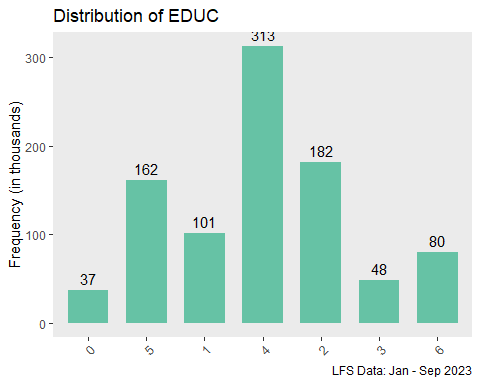
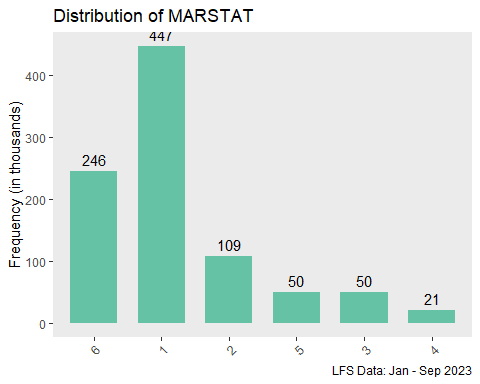
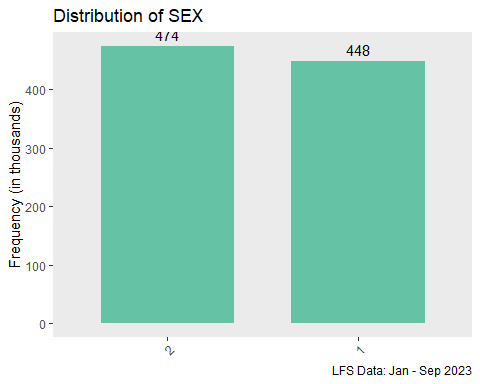
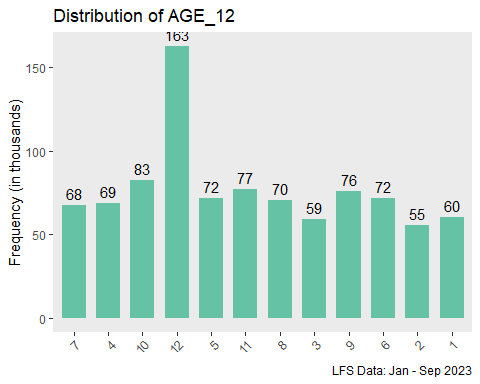
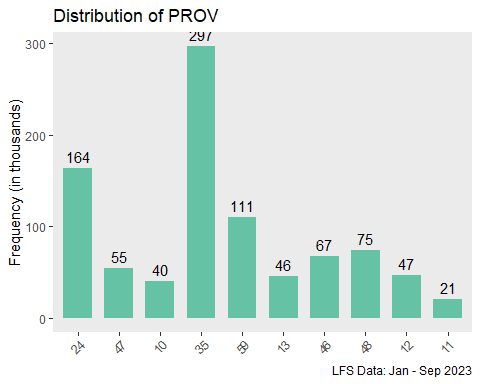
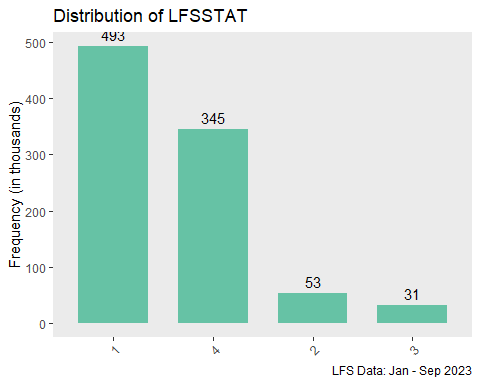
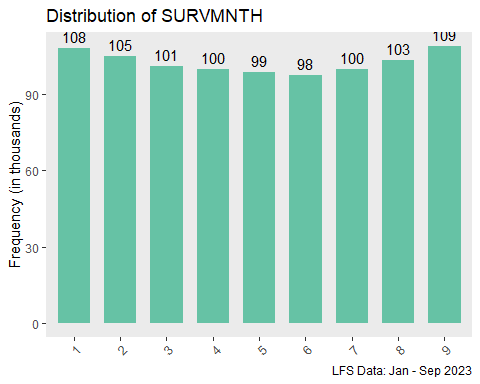
##   
## Attaching package: 'dplyr'

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

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

library(scales)   
  
categorical\_variables <- c("SURVMNTH", "LFSSTAT", "PROV", "AGE\_12", "SEX", "MARSTAT", "EDUC", "IMMIG")  
  
for (variable in categorical\_variables) {  
 # Summary table to show the frequency of each category  
 summary\_table <- table(LFS\_Selected[[variable]])  
  
 # Bar plot to visualize the distribution of categories  
 bar\_plot <- ggplot(LFS\_Selected, aes(x = factor(.data[[variable]], levels = unique(.data[[variable]])))) +  
 geom\_bar(fill = "#66c2a5", width = 0.7) + # Change bar fill color and adjust width  
 geom\_text(stat = "count", aes(label = scales::comma(..count.. / 1000, accuracy = 1)), vjust = -0.5) +  
 labs(title = paste("Distribution of", variable),  
 caption = "LFS Data: Jan - Sep 2023") +  
 xlab(NULL) +   
 ylab("Frequency (in thousands)") +   
 scale\_y\_continuous(labels = scales::comma\_format(scale = 1e-3)) +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1),   
 panel.grid.major = element\_blank(),   
 panel.grid.minor = element\_blank())   
  
 print(bar\_plot)  
}

## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(count)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



str(LFS\_Selected)

## 'data.frame': 922627 obs. of 8 variables:  
## $ SURVMNTH: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ LFSSTAT : int 1 1 1 4 1 1 4 4 4 4 ...  
## $ PROV : int 24 47 10 35 35 35 59 35 10 35 ...  
## $ AGE\_12 : int 7 4 10 12 5 12 10 11 10 8 ...  
## $ SEX : int 2 2 1 2 1 1 2 1 2 2 ...  
## $ MARSTAT : int 6 6 1 1 1 2 1 6 1 1 ...  
## $ EDUC : int 0 5 1 4 2 3 0 6 1 4 ...  
## $ IMMIG : int 3 3 3 3 3 3 2 3 3 2 ...

#Summary of Data

summary(LFS\_Selected)

## SURVMNTH LFSSTAT PROV AGE\_12   
## Min. :1.000 Min. :1.000 Min. :10.00 Min. : 1.000   
## 1st Qu.:3.000 1st Qu.:1.000 1st Qu.:24.00 1st Qu.: 4.000   
## Median :5.000 Median :1.000 Median :35.00 Median : 8.000   
## Mean :4.994 Mean :2.247 Mean :34.57 Mean : 7.319   
## 3rd Qu.:7.000 3rd Qu.:4.000 3rd Qu.:47.00 3rd Qu.:11.000   
## Max. :9.000 Max. :4.000 Max. :59.00 Max. :12.000   
## SEX MARSTAT EDUC IMMIG   
## Min. :1.000 Min. :1.000 Min. :0.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:3.000   
## Median :2.000 Median :2.000 Median :4.000 Median :3.000   
## Mean :1.514 Mean :2.844 Mean :3.414 Mean :2.734   
## 3rd Qu.:2.000 3rd Qu.:6.000 3rd Qu.:5.000 3rd Qu.:3.000   
## Max. :2.000 Max. :6.000 Max. :6.000 Max. :3.000

str(LFS\_Selected)

## 'data.frame': 922627 obs. of 8 variables:  
## $ SURVMNTH: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ LFSSTAT : int 1 1 1 4 1 1 4 4 4 4 ...  
## $ PROV : int 24 47 10 35 35 35 59 35 10 35 ...  
## $ AGE\_12 : int 7 4 10 12 5 12 10 11 10 8 ...  
## $ SEX : int 2 2 1 2 1 1 2 1 2 2 ...  
## $ MARSTAT : int 6 6 1 1 1 2 1 6 1 1 ...  
## $ EDUC : int 0 5 1 4 2 3 0 6 1 4 ...  
## $ IMMIG : int 3 3 3 3 3 3 2 3 3 2 ...

#Detailed statistics

library(psych)

## Warning: package 'psych' was built under R version 4.3.1

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:scales':  
##   
## alpha, rescale

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

# Compute detailed statistics for the entire dataset  
detailed\_statistics <- describe(LFS\_Selected)  
  
# View the detailed statistics  
print(detailed\_statistics)

## vars n mean sd median trimmed mad min max range skew  
## SURVMNTH 1 922627 4.99 2.62 5 4.99 2.97 1 9 8 0.01  
## LFSSTAT 2 922627 2.25 1.42 1 2.18 0.00 1 4 3 0.35  
## PROV 3 922627 34.57 14.60 35 34.47 16.31 10 59 49 -0.03  
## AGE\_12 4 922627 7.32 3.58 8 7.48 4.45 1 12 11 -0.22  
## SEX 5 922627 1.51 0.50 2 1.52 0.00 1 2 1 -0.06  
## MARSTAT 6 922627 2.84 2.16 2 2.68 1.48 1 6 5 0.58  
## EDUC 7 922627 3.41 1.63 4 3.46 1.48 0 6 6 -0.33  
## IMMIG 8 922627 2.73 0.56 3 2.86 0.00 1 3 2 -1.99  
## kurtosis se  
## SURVMNTH -1.26 0.00  
## LFSSTAT -1.80 0.00  
## PROV -0.86 0.02  
## AGE\_12 -1.22 0.00  
## SEX -2.00 0.00  
## MARSTAT -1.47 0.00  
## EDUC -0.86 0.00  
## IMMIG 2.89 0.00

#Correlation Table

correlation\_matrix <- cor(LFS\_Selected)  
correlation\_matrix\_rounded <- round(correlation\_matrix, 2)  
correlation\_table <- as.data.frame(correlation\_matrix\_rounded)  
print(correlation\_table)

## SURVMNTH LFSSTAT PROV AGE\_12 SEX MARSTAT EDUC IMMIG  
## SURVMNTH 1.00 -0.01 -0.01 0.00 0.00 0.00 0.01 -0.01  
## LFSSTAT -0.01 1.00 -0.04 0.40 0.08 -0.01 -0.25 0.05  
## PROV -0.01 -0.04 1.00 -0.05 0.00 0.00 0.02 -0.15  
## AGE\_12 0.00 0.40 -0.05 1.00 0.03 -0.44 0.01 0.05  
## SEX 0.00 0.08 0.00 0.03 1.00 -0.01 0.04 -0.01  
## MARSTAT 0.00 -0.01 0.00 -0.44 -0.01 1.00 -0.19 0.10  
## EDUC 0.01 -0.25 0.02 0.01 0.04 -0.19 1.00 -0.11  
## IMMIG -0.01 0.05 -0.15 0.05 -0.01 0.10 -0.11 1.00

#Section 3: Correlation Visuals

# Load necessary libraries  
library(ggplot2)  
library(reshape2)

## Warning: package 'reshape2' was built under R version 4.3.1

# Create a correlation matrix  
correlation\_matrix <- cor(LFS\_Selected)  
  
# Clean the correlation matrix by replacing NAs with zeros  
correlation\_matrix[is.na(correlation\_matrix)] <- 0  
  
# Ensure that all values are finite (replace Inf with a large value)  
correlation\_matrix[!is.finite(correlation\_matrix)] <- 1e-10  
  
# Calculate hierarchical clustering order  
order <- hclust(as.dist(1 - correlation\_matrix))$order  
  
# Reorder the correlation matrix and create a long format  
correlation\_reordered <- correlation\_matrix[order, order]  
correlation\_long <- melt(correlation\_reordered)  
  
# Create a ggplot2-style correlation matrix plot  
ggplot(data = correlation\_long, aes(x = Var1, y = Var2, fill = value)) +  
 geom\_tile(color = "white", size = 0.5) +  
 geom\_text(aes(label = sprintf("%.2f", value)), vjust = 1, color = "black", size = 3) +  
 scale\_fill\_gradient2(low = "skyblue", high = "darkred", mid = "white", midpoint = 0, name = "Correlation") +  
 theme\_minimal() +  
 theme(  
 axis.text.x = element\_text(angle = 45, hjust = 1),  
 panel.grid = element\_blank(),   
 axis.title = element\_blank()   
 ) +  
 coord\_fixed()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

