# PS 138Z - The Politics of Immigration: Section 7

2024-03-7

## Introduction

Today, we will continue using `ggplot` to analyze data from the **Mexican Migration Project (MMP) Database** to produce.

## Cleaning data

First, let's load the data using the `read.csv()` function. This time, we will assign the dataset to an object called `mmp.data`.

```
mmp.data <- read.csv("mmp_subset.csv")
```

Let's upload `tidyverse` package only. You can upload a (pre-installed) package with the `library()` function.

```
library(tidyverse)
```

```
## Warning: package 'tibble' was built under R version 4.2.3
```

In this exercise, we will work with four variables: "Currently on last migration" (`usyrl`), "age" (`age`), "sex" (`sex`), and "education years" (`edyrs`). Remember that you can see the description of these and other variables in the dataset in the codebook. Let's start by creating a dataset with only these variables. Complete the code below:

```
no.miss.data <- mmp.data
no.miss.data <- cbind(mmp.data$uscurtrp, no.miss.data$age,
                      no.miss.data$sex, no.miss.data$edyrs)
no.miss.data <- as.data.frame(no.miss.data )
names(no.miss.data) <- c("uscurtrp", "age", "sex", "edyrs")
```

Now, let's examine if any of these four variables has any numerical missing values. You can do this in multiple ways. This time, we will use the `range()` function. Complete the code below (Hint: remember that you should use the argument `na.rm = TRUE` because we do not know if the variables have missing values coded as `NA` as well):

```
range(no.miss.data$uscurtrp, na.rm = TRUE)
```

```
## [1]    1 8888
```

```
range(no.miss.data$age, na.rm = TRUE)
```

```
## [1]    0 8888
```

```
range(no.miss.data$sex, na.rm = TRUE)
```

```
## [1] 1 2
```

```
range(no.miss.data$edyrs, na.rm = TRUE)
```

```
## [1]    0 8888
```

*What can you conclude? Which variables have numerical missing values?*

Remember that for some functions, such as `mean()` and `range()`, we can ignore missing values coded as `NA` with the argument `na.rm = TRUE`. Also, remember that we can easily drop all the rows with at least one `NA` using the function `omit.na()`. For these reasons, we may prefer to replace the numerical missing values with `NA`s. Complete the code below:

```
no.miss.data$uscurtrp[no.miss.data$uscurtrp == 8888] <- NA
no.miss.data$age[no.miss.data$age == 8888] <- NA
no.miss.data$edyrs[no.miss.data$edyrs == 8888] <- NA
```

Now that we have replaced all the numerical missing values with `NA`s, we can drop any rows with `NA`s.

```
no.miss.data <- na.omit(no.miss.data)
```

To this point, we have a "clean" dataset with only the variables we are interested in and no missing values. As a final step, let's recode the `sex` and `uscurtrp` variables to facilitate plotting. Run the code below:

```
no.miss.data$uscurtrp[no.miss.data$uscurtrp == 1] <- "Yes"
no.miss.data$uscurtrp[no.miss.data$uscurtrp == 2] <- "No"

no.miss.data$sex[no.miss.data$sex == 1] <- "Male"
no.miss.data$sex[no.miss.data$sex == 2] <- "Female"
```

Use the `head()` function to look at the first 15 rows in the dataset.

```
head(no.miss.data, 15)
```

```
##    uscurtrp age    sex edyrs
## 1       Yes  26   Male     9
## 7        No  59   Male     6
## 8       Yes  27   Male     9
## 12       No  52   Male     0
## 15      Yes  27   Male     9
## 18       No  55   Male     6
## 20      Yes  31   Male     9
## 25      Yes  40   Male    11
## 26      Yes  41 Female     4
## 31      Yes  26   Male     7
## 35       No  30   Male     6
## 36       No  34   Male     6
## 38      Yes  30   Male     6
## 39      Yes  30 Female     9
## 40      Yes   8   Male     3
```

# Characterizing recent immigrants

Complete the code below to calculate the average age and educational level by immigration status.

```
no.miss.data %>%
  group_by(uscurtrp) %>%
  summarise_at(vars(age, edyrs), c(mean), na.rm = TRUE)
```

```
## # A tibble: 2 x 3
##   uscurtrp   age edyrs
##   <chr>    <dbl> <dbl>
## 1 No        46.6  6.57
## 2 Yes       37.7  7.70
```

Complete the code below to cross-tabulate immigration status and sex.

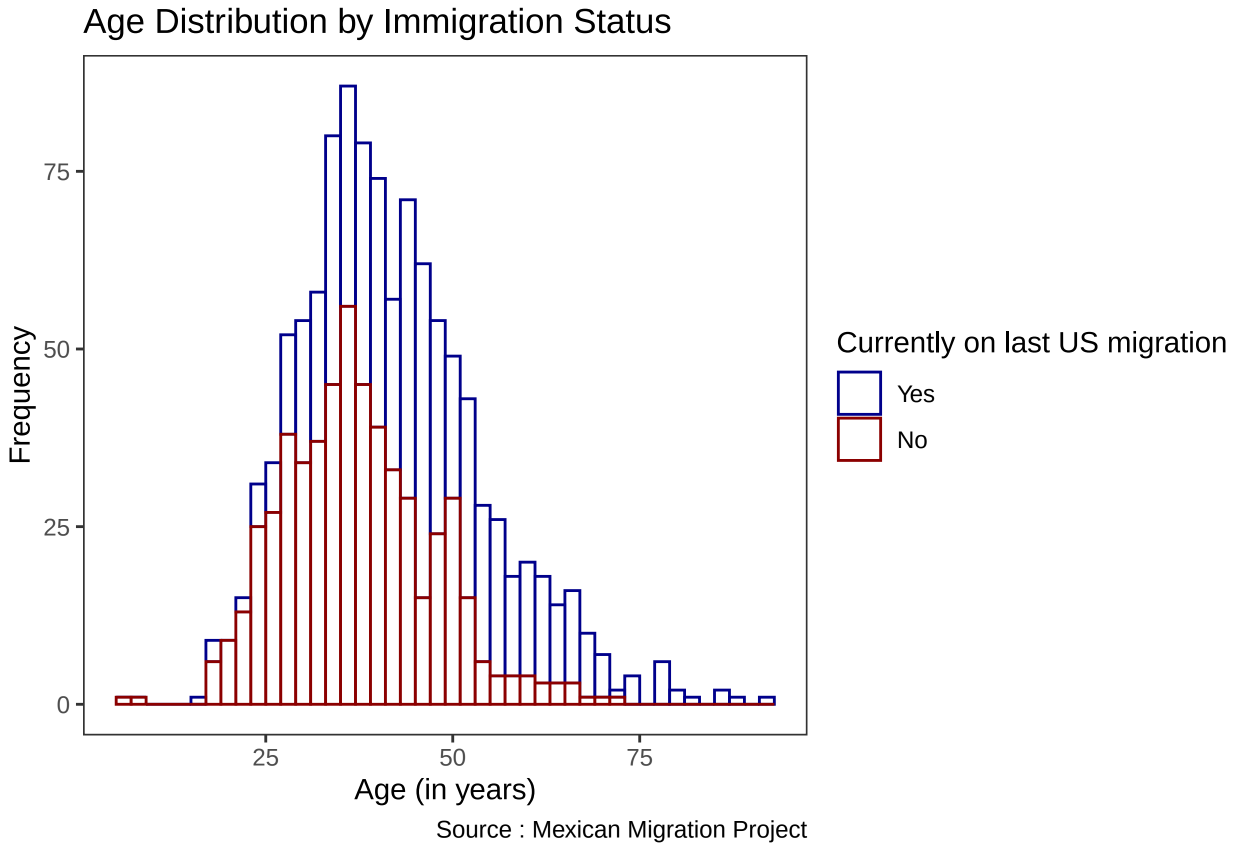
```
table(no.miss.data$sex, no.miss.data$uscurtrp)
```

```
##
##           No Yes
##   Female  76 141
##   Male   470 410
```

*What can you conclude from these descriptive statistics?*

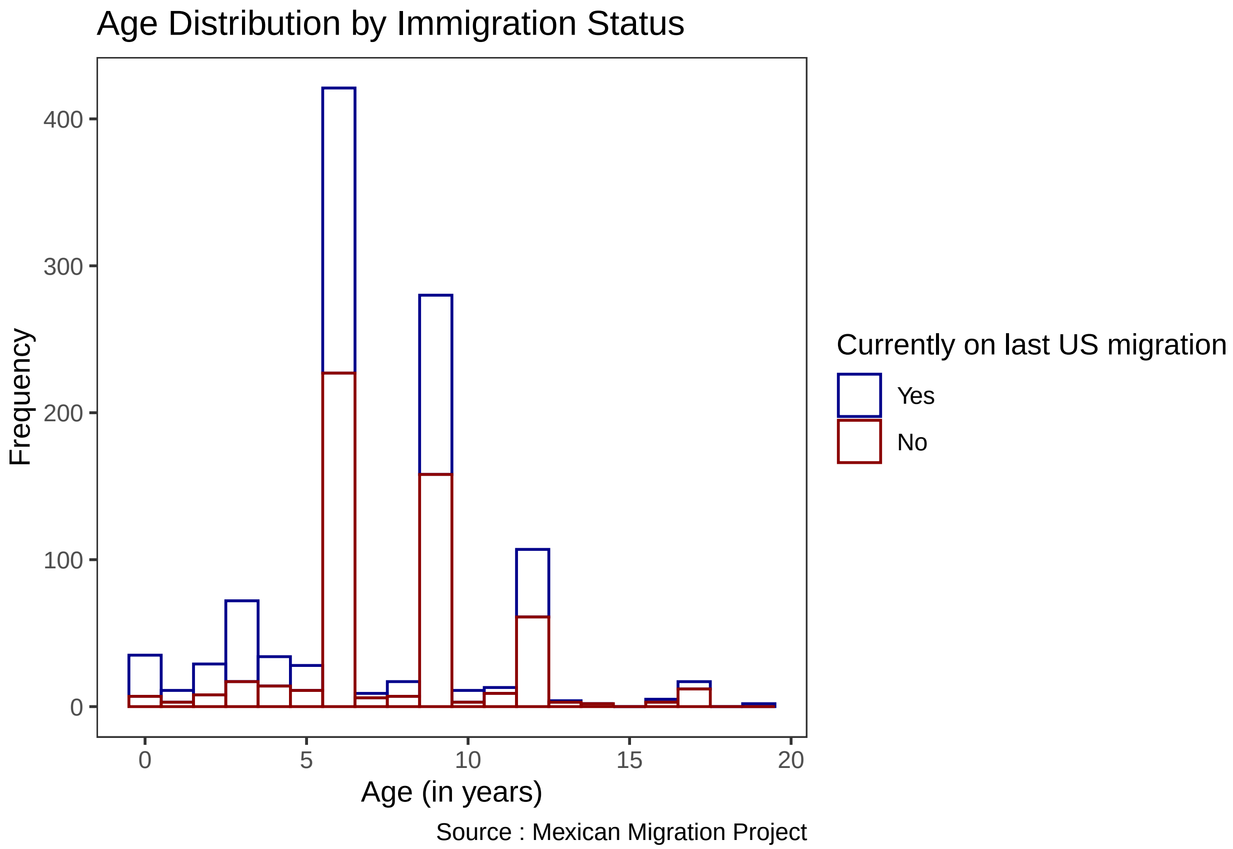
Now, let's use the ggplot package to plot a grouped histogram showing the age distribution by immigration status. Complete the code below:

```
plot1 <- ggplot(no.miss.data, aes(x=age, color= factor(uscurtrp))) +
  geom_histogram(binwidth = 2, fill="white", alpha=0.5) +
  labs(
    title = "Age Distribution by Immigration Status",
    caption = "Source : Mexican Migration Project",
    x = "Age (in years)",
    y = "Frequency",
    color = "Currently on last US migration") +
  scale_color_manual(labels = c("Yes", "No"), values = c("darkblue", "darkred")) +
  theme_test()
plot1
```



Similarly, let's use the ggplot package to plot a grouped histogram showing the distribution of school years completed by immigration status. Complete the code below:

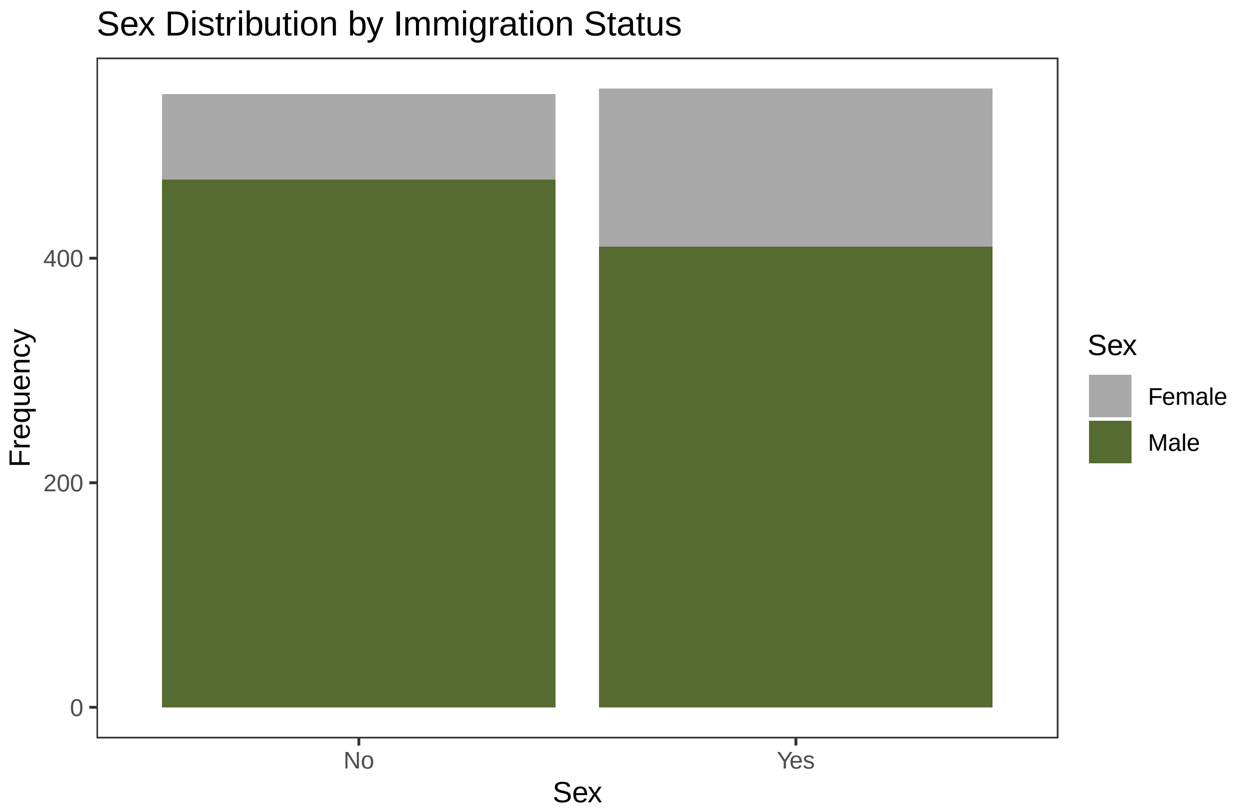
```
plot2 <- ggplot(no.miss.data, aes(x=edyrs, color= factor(uscurtrp))) +
  geom_histogram(binwidth = 1, fill="white", alpha=0.5) +
  labs(
    title = "Age Distribution by Immigration Status",
    caption = "Source : Mexican Migration Project",
    x = "Age (in years)",
    y = "Frequency",
    color = "Currently on last US migration") +
  scale_color_manual(labels = c("Yes", "No"), values = c("darkblue", "darkred")) +
  theme_test()
plot2
```



Finally, let's use `geom_bar()` to create a grouped bar plot showing how individuals with different immigration statuses are distributed by sex. Complete the code below:

```
plot2 <- ggplot(no.miss.data, aes(x = uscurtrp, fill = factor(sex))) +
  geom_bar() +
  labs(
    title = "Sex Distribution by Immigration Status",
    caption = "Source : Mexican Migration Project",
    x = "Sex",
    y = "Frequency",
    legend = "Sex") +
  scale_fill_manual(values = c("darkgray", "darkolivegreen"),
                    name = "Sex") +
  theme_test()
```

```
plot2
```



Source : Mexican Migration Project

*How do the descriptive tables and figures we produced relate to the theories we have discussed in class?*