Final Project

Diego A. Dulanto

05/06/2021

## Final Project Introduction

This project is an objective view into migration patterns from 1999-2000, specifically from Colombia, Peru, and Venezuela.

The data collected is from the national, public database of the United States Department of Justice. Immigration and Naturalization Service. I will be cleaning the data, creating summaries of the data, and displaying the difference between 1999 and 2000 using visual aids.

My hypothesis is that the most common month will be November and December and the average age will be 25-30 in 1999. Additionally, I hypothesize that the most common month will be June and July and the age will increase to 30-35 in 2000.

Using this line of code, I am importing the dataset I will be using.

## [1] "C:/Users/Diego Dulanto/OneDrive/Documents/GitHub/hw-diegodulanto/data/immigrant\_data\_2000.sav"

## # A tibble: 849,386 x 17  
## V1 V2 V3 V4 V5 V6 V7 V8 V9  
## <dbl> <dbl> <dbl+lbl> <dbl> <chr> <dbl+l> <dbl+lb> <dbl+lbl> <dbl+l>  
## 1 7 2000 190 [Cub~ 26 Unkn~ 2 [Mar~ 15 [Flo~ 189 [Mia~ 3 [Not~  
## 2 4 2000 179 [Mex~ 75 Mexi~ 2 [Mar~ 9 [Cal~ 188 [Mer~ 3 [Not~  
## 3 12 1999 179 [Mex~ 75 Mexi~ 4 [Div~ 9 [Cal~ 174 [Los~ 3 [Not~  
## 4 7 2000 179 [Mex~ 75 Mexi~ 3 [Wid~ 55 [Tex~ 260 [San~ 3 [Not~  
## 5 5 2000 179 [Mex~ 75 Mexi~ 3 [Wid~ 9 [Cal~ 263 [San~ 3 [Not~  
## 6 7 2000 62 [Chi~ 15 Chin~ 1 [Sin~ 9 [Cal~ 174 [Los~ 3 [Not~  
## 7 4 2000 83 [Phi~ 31 Phil~ 2 [Mar~ 9 [Cal~ 266 [San~ 3 [Not~  
## 8 8 2000 179 [Mex~ 70 Mexi~ 2 [Mar~ 55 [Tex~ 91 [El ~ 3 [Not~  
## 9 3 2000 52 [Uni~ 75 Unit~ 2 [Mar~ 31 [Mic~ 212 [Non~ 1 [Iss~  
## 10 11 1999 218 [Col~ 71 Colo~ 3 [Wid~ 15 [Flo~ 189 [Mia~ 3 [Not~  
## # ... with 849,376 more rows, and 8 more variables: V10 <dbl+lbl>,  
## # V11 <dbl+lbl>, V12 <chr>, V13 <dbl+lbl>, V14 <dbl+lbl>, V15 <dbl>,  
## # V16 <dbl+lbl>, V17 <dbl+lbl>

## Data Cleaning

First, I will be renaming the variables from the dataset. Right now, the dataset uses variables named V# (i.e., V1, V2, V3, etc.). I will be renaming the variables as such:

V1 - Month of Admission

V2 - Year of Admission

V3 - Country of Birth

V4 - Age

V5 - Country of Last Permanent Residence

V6 - Marital Status

V7 - State of Intended Residence

V8 - Metropolitan Area of Intended Residence

V9 - Labor Certification

v10 - Occupation

V11 - Sex

V12 - Country of Chargeability

v13 - Type of Case

V14 - Nonimmigrant Class of Entry

V15 - Nonimmigrant Year of Entry

V16 - Immigrant Class of Admission

v17 - Prinicipal Vs Derivative Class of Admission

The code for renaming the variables is below.

## # A tibble: 849,386 x 17  
## Month\_of\_Admiss~ Year\_of\_Admissi~ Country\_of\_Birth Age Country\_of\_Last~  
## <dbl> <dbl> <dbl+lbl> <dbl> <chr>   
## 1 7 2000 190 [Cuba] 26 Unknown   
## 2 4 2000 179 [Mexico] 75 Mexico   
## 3 12 1999 179 [Mexico] 75 Mexico   
## 4 7 2000 179 [Mexico] 75 Mexico   
## 5 5 2000 179 [Mexico] 75 Mexico   
## 6 7 2000 62 [China Peop~ 15 China People Re~  
## 7 4 2000 83 [Philippine~ 31 Philippines   
## 8 8 2000 179 [Mexico] 70 Mexico   
## 9 3 2000 52 [United Kin~ 75 United Kingdom   
## 10 11 1999 218 [Colombia] 71 Colombia   
## # ... with 849,376 more rows, and 12 more variables: Marital\_Status <dbl+lbl>,  
## # State\_of\_Intended\_Residence <dbl+lbl>,  
## # Metropolitan\_Area\_of\_Intended\_Residence <dbl+lbl>,  
## # Labor\_Certification <dbl+lbl>, Occupation <dbl+lbl>, Sex <dbl+lbl>,  
## # Country\_of\_Chargeability <chr>, Type\_of\_Case <dbl+lbl>,  
## # Nonimmigrant\_Class\_of\_Entry <dbl+lbl>, Nonimmigrant\_Year\_of\_Entry <dbl>,  
## # Immigrant\_Class\_of\_Admission <dbl+lbl>,  
## # Principal\_Vs\_Derivative\_Class\_of\_Admission <dbl+lbl>

Now that I have renamed the variables appropriately, I can start summarizing the data for analysis. I will be summarizing the data for Colombia, Peru, and Venezuela. I will group these countries by year, so that I can distinguish the differences between 1999 and 2000. The specific differences I will be analyzing is the average age and common month immigrants from these countries enter the US.

# Summarizing Data  
  
colombia\_averages <- immigrant\_data %>%  
 filter(Country\_of\_Birth == 218) %>%  
 group\_by(Year\_of\_Admission) %>%  
 summarise(  
 Age = mean(Age, na.rm = TRUE),  
 Month = Mode(Month\_of\_Admission, na.rm = TRUE))  
  
peru\_averages <- immigrant\_data %>%  
 filter(Country\_of\_Birth == 224) %>%  
 group\_by(Year\_of\_Admission) %>%  
 summarise(  
 Age = mean(Age, na.rm = TRUE),  
 Month = Mode(Month\_of\_Admission, na.rm = TRUE))  
  
venezuela\_averages <- immigrant\_data %>%  
 filter(Country\_of\_Birth == 227) %>%  
 group\_by(Year\_of\_Admission) %>%  
 summarise(  
 Age = mean(Age, na.rm = TRUE),  
 Month = Mode(Month\_of\_Admission, na.rm = TRUE))  
   
  
print(colombia\_averages)

## # A tibble: 2 x 3  
## Year\_of\_Admission Age Month  
## \* <dbl> <dbl> <dbl>  
## 1 1999 30.8 12  
## 2 2000 32.9 5

print(peru\_averages)

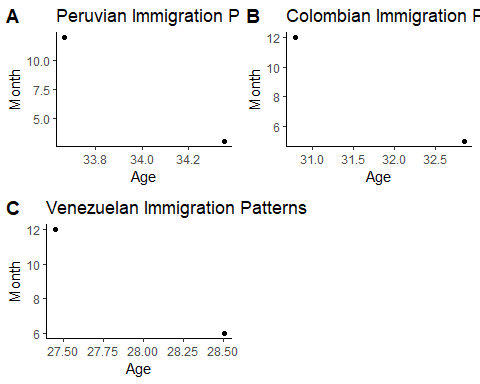
## # A tibble: 2 x 3  
## Year\_of\_Admission Age Month  
## \* <dbl> <dbl> <dbl>  
## 1 1999 33.7 12  
## 2 2000 34.4 3

print(venezuela\_averages)

## # A tibble: 2 x 3  
## Year\_of\_Admission Age Month  
## \* <dbl> <dbl> <dbl>  
## 1 1999 27.4 12  
## 2 2000 28.5 6

With the summary statistics created above, I will now plot the data for analysis. In addition to creating a plot each of the summaries, I also plotted them on the same page and combined them into one object in the code below.

# Plotting Data  
  
peru\_plot <- ggplot(peru\_averages) +  
 aes(x = Age,   
 y = Month) +  
 geom\_point() +  
 scale\_color\_brewer(palette = "Dark2") +  
 labs(title = "Peruvian Immigration Patterns") +  
 theme\_classic()+  
 theme(legend.position = "right")  
  
col\_plot <- ggplot(colombia\_averages) +  
 aes(x = Age,   
 y = Month) +  
 geom\_point() +  
 scale\_color\_brewer(palette = "Dark2") +  
 labs(title = "Colombian Immigration Patterns") +  
 theme\_classic()+  
 theme(legend.position = "right")  
  
venz\_plot <- ggplot(venezuela\_averages) +  
 aes(x = Age,   
 y = Month) +  
 geom\_point() +  
 scale\_color\_brewer(palette = "Dark2") +  
 labs(title = "Venezuelan Immigration Patterns") +  
 theme\_classic()+  
 theme(legend.position = "right")  
  
combined\_plot <- plot\_grid(  
 peru\_plot, col\_plot, venz\_plot,   
 rel\_heights = c(1, 1),   
 labels = c("A", "B", "C"))  
  
combined\_plot



My first hypothesis for the year 1999 was partially correct. As we can see from the plot, for all three countries, the most common month of immigration was December. With the exception of Peru, both Colombia and Venezuela had an average age between 25-30. My secondary hypothesis had similar results. The most common month of immigration was March (Peru), May (Colombia), and June (Venezuela). With the exception of Venezuela, the average age did increase to 30-35.

## Saving 5 x 4 in image