**Program to examine the relationship between place of birth, immigrant status, and total income in Canada**

Samuel Nemeroff

Step 0.5: Loading dependencies (Optional, done to condense some code and use easier functions such as %>% pipe operator and whatnot)

1. Make sure to download the tidyverse R package by going to Tools, Install Packages, and type in Tidyverse, and hit install.
2. Once Tidyverse has finished installing, ensure to add tidyverse, and specifically dplyr and ggplot2 to the library for your code.

Part 1: Data cleaning

1. The first step I did to make this significantly easier on myself and my computer is to isolate the variables in excel to bring over to RStudio. For this, I pulled PPSORT, HDGREE, AGEIMM, IMMSTAT, POB, PR, Sex, and TOTINC into a new sheet.
2. First and foremost, as we are looking at Canadians income, we must clean the dataset by removing the invalid responses to the TOTINC column (88,888,888 and 99,999,999).
3. Recode place of birth into a binary variable – either the individual was born in Canada, or they were born outside of Canada. I recoded the name as POB\_Binary.
4. Recode immigration status into a binary variable – either the individual is an immigrant or not. I recoded the name as ImmigrationStat.
5. I now elect to recode the income from a continuous variable into discrete categories – this is optional, as ultimately, we just end up looking at top earners versus non top earners, but I do this for simpler visualization of things like tables. I elected to recode it into intervals of $20,000, with no lower limit under $20k (as some respondents indicated negative income) and no upper limit above $140k
6. I also create the additional binary variable of “top earner”, where if a respondent earned any income above $80k qualifies them as a top earner. This is an arbitrary number for this exercise, and for further testing could be set to another value. 0=not top earner, 1=top earner.
7. I also elected to simplify the HDGREE variable – taking it down to a few less categories. Here, 1=no degree, 2=a degree less than bachelors, 3=bachelors or above, 4=masters, 5=phd, 6=medicine, and 99=not available. I recoded the name as DegreeEarned.

Part 2: Visualizations

1. The first visualization to do is a simple summary table of immigrants versus non-immigrants making above $80,000. Print this table.
2. Time for the logistic regression. This will be a binary logistic regression, as the dependent variable (top earner) is simply a yes/no variable. Here, we will include the variables of POB\_Binary, ImmigrationStat, Sex, DegreeEarned, PR, and AGEIMM. Run this regression, and request a summary of the logit model.