American Community Survey

### Introduction

The objective of this project was analyzing the data of the 2013 American Community Survey (ACS) to figure out whether an employed PhD makes more money compared to a working bachelor or master degree holder. To do this, I concentrated on B.Sc, MSc, and PhD holders data in ACS.

### Total Number of BSc , MSc, and PhD holders

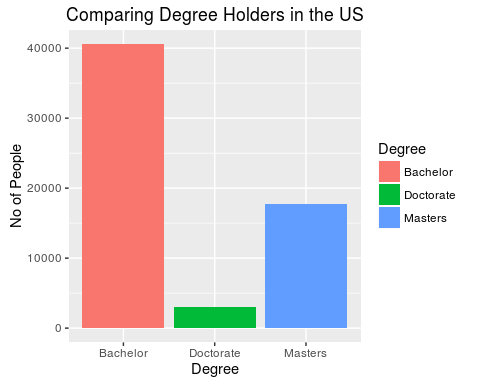
My first analysis was comparing the total number of BSc, MSc, and PhD holders in the US. I first created a clean data set AC\_Survey\_Subset\_Cleaned (I got this cleaned data by removing NA values, and by extracting only the bachelors, masters & PhD's), and next I made use of the [dplyr package] to calculate each number:

# Prepare degree codes  
degree\_codes <- data.frame(SCHL = c(21, 22, 24),   
 Degree = c("Bachelor", "Masters", "Doctorate"))  
  
# Use the pipe operator and chaining   
ac\_survey\_clean <- ac\_survey %>%  
 tbl\_df() %>%  
 na.omit() %>%  
 filter(SCHL %in% c(21,22,24)) %>%  
 inner\_join(degree\_codes)  
  
# Group by SCHL and count each group  
degree\_holders <- ac\_survey\_clean %>%  
 group\_by(Degree) %>%  
 summarize(count = n())  
  
# Print out degree\_holders  
degree\_holders

## Source: local data frame [3 x 2]  
##   
## Degree count  
## <fctr> <int>  
## 1 Bachelor 40561  
## 2 Doctorate 3000  
## 3 Masters 17680

We learnt that there are bachelors (individuals with a bachelor degree), masters (individuals with a masters degree), and PhDs (individuals with a PhD). Visually this gives:

# Visualize the number of Bachelor, Master and PhD holders   
ggplot(degree\_holders, aes(x = Degree, y = count, fill = Degree)) +   
 geom\_bar(stat = "identity") +  
 xlab("Degree") +   
 ylab("No of People") +   
 ggtitle("Comparing Degree Holders in the US")



The visualization of the data was done using the ggplot2 package.

### Do PhD's Earn more?

Next, I needed to figure out whether it's a smart career choice moneywise to pursue a PhD.I created a new data set (named income). Income is created by taking 5000 times a random sample of 1000 observations from The American Community Survey. For each sample min(), max(), median(), and IQR() is calculated.

# Take 5000 random samples of 1000 observations & calculate median income  
over\_thousand <- ac\_survey\_clean %>%  
 filter(PINCP > 1000) %>% # exclude obserations with income under 1000  
 group\_by(Degree)   
   
freq <- 5000 # 5000 samples  
income <- NULL  
for (i in 1:freq) {  
 # Select 1000 observations  
 sample <- sample\_n(over\_thousand, 1000)  
   
 # Calculate stats by degree  
 sample\_stats <- summarise(sample,   
 MinIncome = min(PINCP),   
 MaxIncome = max(PINCP),  
 MedianIncome = median(PINCP),  
 IncomeRange = IQR(PINCP))  
   
 income <- rbind(income, sample\_stats)  
}

income can now be used to create three boxplots of Median income for each degree level:

# Create the boxplots  
ggplot(income, aes(x = Degree, y = MedianIncome, fill = Degree)) +   
 geom\_boxplot() +  
 ggtitle("Comparing Income of Degrees Holders")



The graph clearly shows it is a smart career move to pursue a PhD.