WDS, HW 1

Aaron (team leader) Neil Ben Luke Tanvi

Due: 10:00PM, July 13, 2021

Contents

1 Overview

This is a fast-paced course that covers a lot of material. There will be a large amount of references. It is impossible to learn data science without getting your hands dirty. Please budget your time evenly. Last-minute work ethic will not work here.

Homework in this program is different from your usual homework assignment as a typical student. Most of the time, they are built over real case studies. While you will be applying methods covered in lectures, we often do not ask very specific questions. The focus will be always on the goals of the study, the usefulness of the data gathered, and the limitations in any conclusions you may draw. Always try to challenge your data analysis in a critical way. Frequently, there are no unique solutions.

1.1 Objectives

- Get familiar with `R-studio` and `RMarkdown`
- Hands-on R
- Learn data science essentials
  - gather data
  - clean data
  - summarize data
  - display data
  - conclusion
- Packages
  - `dplyr`
  - `ggplot`

1.2 Instructions

- **Homework assignments are done in a group consisting of 5 members**.
- **All work submitted should be completed in the R Markdown format.** You can find a cheat sheet for R Markdown here.
- **Submit the following files, one submission for each group:** (1) Rmd file, (2) a compiled HTML (or a pdf which might require extra work) version, and (3) all necessary data files if different from our source data. You may directly edit this .rmd file to add your answers. We encourage that you at least attempt each problem by yourself before working with your teammates. Additionally, ensure that you can 'knit' or compile your Rmd file. It is also likely that you need to configure Rstudio to properly convert files to PDF. **These instructions** might be helpful.
- In general, be as concise as possible while giving a fully complete answer to each question. All necessary datasets are available in this homework folder on Canvas. Make sure to document your code with comments (written on separate lines in a code chunk using a hashtag `#` before the comment) so the teaching fellows can follow along. R Markdown is particularly useful because it follows a 'stream of consciousness' approach: as you write code in a code chunk, make sure to explain what you are doing outside of the chunk.
- A few good or solicited submissions will be used as sample solutions. When those are released, make sure to compare your answers and understand the solutions.

2 Review materials

- Study Module 1: DataPreparationEDA_WDS
- Be able to comple DataPreparationEDA_WDS.rmd

3 Case study: Audience Size

How successful is the Wharton Talk Show Business Radio Powered by the Wharton School

Background: Have you ever listened to SiriusXM? Do you know there is a **Talk Show** run by Wharton professors in Sirius Radio? Wharton launched a talk show called Business Radio Powered by the Wharton School through the Sirius Radio station in January of 2014. Within a short period of time the general reaction seemed to be overwhelmingly positive. To find out the audience size for the show, we designed a survey and collected a data set via MTURK in May of 2014. Our goal was to **estimate the audience size**. There were 51.6 million Sirius Radio listeners then. One approach is to estimate the proportion of the Wharton listeners to that of the Sirius listeners, p, so that we will come up with an audience size estimate of approximately 51.6 million times p.

To do so, we launched a survey via Amazon Mechanical Turk (MTurk) on May 24, 2014 at an offered price of $0.10 for each answered survey. We set it to be run for 6 days with a target maximum sample size of 2000 as our goal. Most of the observations came in within the first two days. The main questions of interest are “Have you ever listened to Sirius Radio” and “Have you ever listened to Sirius Business Radio by Wharton?”. A few demographic features used as control variables were also collected; these include Gender, Age and Household Income.

We requested that only people in United States answer the questions. Each person can only fill in the questionnaire once to avoid duplicates. Aside from these restrictions, we opened the survey to everyone in MTurk with a hope that the sample would be more randomly chosen.

The raw data is stored as `Survey_results_final.csv` on Canvas.

3.1 Data preparation

i. We need to clean and select only the variables of interest.

Select only the variables Age, Gender, Education Level, Household Income in 2013, Sirius Listener?, Wharton Listener? and Time used to finish the survey.

Change the variable names to be “age”, “gender”, “education”, “income”, “sirius”, “wharton”, “worktime”.

3.1.1 Load the data from the table

```
survey <- read.csv("data/Survey_results_final.csv", header=T, stringsAsFactors = FALSE)
```

3.1.2 Select the specific variables we want

```
survey_col <- survey %>% select(Answer.Age, Answer.Gender, Answer.Education, Answer.Gender, Answer.Hou
```

3.1.3 Rename the specific variables accordingly

```
survey_rename<-survey_col %>% rename(age = Answer.Age, gender = Answer.Gender, education = Answer.Edu
```

ii. Handle missing/wrongly filled values of the selected variables

As in real world data with user input, the data is incomplete, with missing values, and has incorrect responses. There is no general rule for dealing with these problems beyond “use common sense.” In whatever case, explain what the problems were and how you addressed them. Be sure to explain your rationale for your chosen methods of handling issues with the data. Do not use Excel for this, however tempting it might be.

Since age is not automatically converted to an integer, there must be some unclean/incorrect values in this column.Regaring the other data sets, there are mishaps such as empty cells. In order to address this problem, we need to index each column for the error, then replace the error by assigning it a value through the <- function, whether it is NA or an actual data value.

3.1.4 Clean Age Variable

```
  "Eighteen (18)" <- 18
  "female" <- NA
  "223" <- NA
  "4" <- NA
  "27" <- NA
  "" <- NA
```

Convert column to integer

```
## Here, we're removing the errors that were made when inputting the data. For example, in the numbers
survey_rename$age[survey_rename$age=="Eighteen (18)"] <- 18
survey_rename$age[survey_rename$age=="female"] <- NA
survey_rename$age[survey_rename$age=="223"] <- NA
survey_rename$age[survey_rename$age=="4"] <- NA
survey_rename$age[survey_rename$age == "27'"] <- 27
survey_rename$age[survey_rename$age==""] <- NA
survey_rename$age <- as.integer(survey_rename$age)
str(survey_rename)
```

```
## ’data.frame':    1764 obs. of  7 variables:
##  $ age      : int  21 56 40 52 33 55 24 40 35 62 ...
##  $ gender   : chr  "Female" "Female" "Female" "Female" ...
##  $ education: chr  "Some college, no diploma; or Associate’s degree" "Some college, no diploma; or A
##  $ income   : chr  "$30,000 - $50,000" "$15,000 - $30,000" "$50,000 - $75,000" "Above $150,000" ...
##  $ sirius   : chr  "No" "No" "Yes" "Yes" ...
##  $ wharton  : chr  "No" "No" "No" "Yes" ...
##  $ worktime : int  12 25 13 33 13 24 32 19 20 89 ...
```

3.1.5 Replace Gender Blank with NA

```
## There were blank data points with nothing inputted for gender, and for the other variables
survey_rename$gender[survey_rename$gender ==""] <- NA
```

3.1.6 Replace Education "select one" with NA

```
survey_rename$education[survey_rename$education =="select one"] <- NA
```

3.1.7 Replace Income Blank with NA

```
survey_rename$income[survey_rename$income ==""] <- NA
```

3.1.8 Replace Sirius Blank with NA

```
survey_rename$sirius[survey_rename$sirius ==""] <- NA
```

3.1.9 Replace Wharton Blank with NA

```
survey_rename$wharton[survey_rename$wharton ==""] <- NA
```

```
survey_rename$worktime[survey_rename$worktime ==""] <- NA
```

The most common error was `""` which we fixed with `survey_rename$var[survey_rename$var == ""] <- NA`

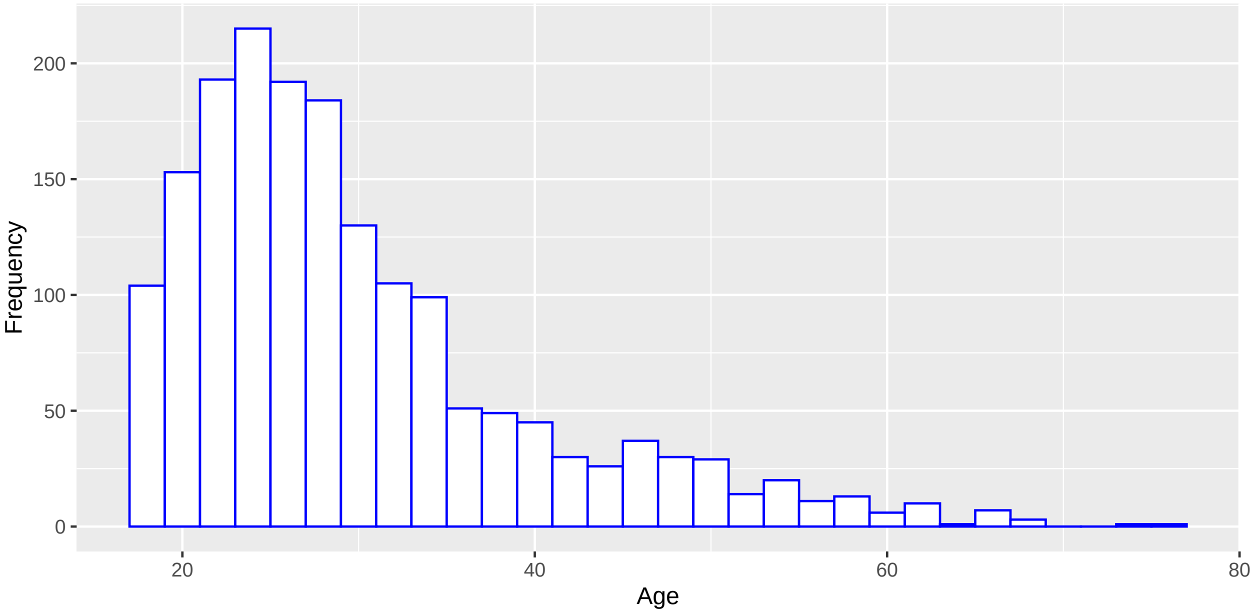
iii. Brief summary

Write a brief report to summarize all the variables collected. Include both summary statistics (including sample size) and graphical displays such as histograms or bar charts where appropriate. Comment on what you have found from this sample. (For example - it's very interesting to think about why would one work for a job that pays only 10cents/each survey? Who are those survey workers? The answer may be interesting even if it may not directly relate to our goal.)

The variables collected and analyzed in the survey were Age, Gender, Education Level, Household Income in 2013, whether the person listens to Wharton and Sirius Radios respectively, and Time elapsed to complete the survey. By plotting the data in histograms, a significant skew to the right can be observed in the `age` variable, showing that the people who took the survey were more likely to be younger (<40). This same phenomenon repeats throughout the `worktime` variable, possibly demonstrating some similarities between younger `age` and a shorter `worktime`. This could be intuitively accounted for by the relatively low survey payout, though there is not enough concrete data to support this hypothesis. However, after running a correlation test, recieving a value of `0.275` renders this correlation weak and with a very minute `p-value`, this null hypothesis is hereby rejected. The `income` samples generally exhibit more `male` responses, whereas the `education` samples are more evenly distributed between `male` and `female` (exempting `other`). Additionally, it can be concluded that a `wharton` listener is likely to listen to `sirius` as well. However, a `sirius` listener is unlikely to listen to `wharton`.

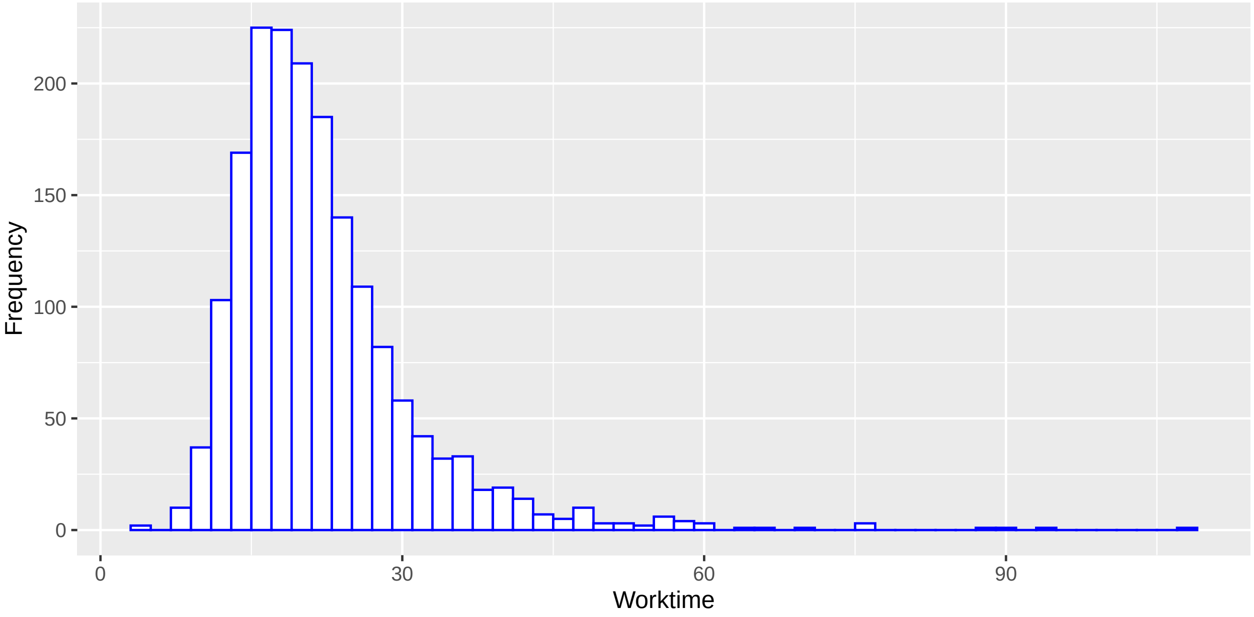
3.1.10 Age Histogram Skews Right

```
survey_rename %>%
  ggplot(aes(x = age)) +
  geom_histogram(binwidth = 2, color = "blue", fill = "white") +
  labs(x = "Age", y="Frequency")
```



3.1.11 Worktime Histogram Skews Right

```
survey_rename %>%
  ggplot(aes(x = worktime)) +
  geom_histogram(binwidth = 2, color = "blue", fill = "white") +
  labs(x = "Worktime", y="Frequency")
```



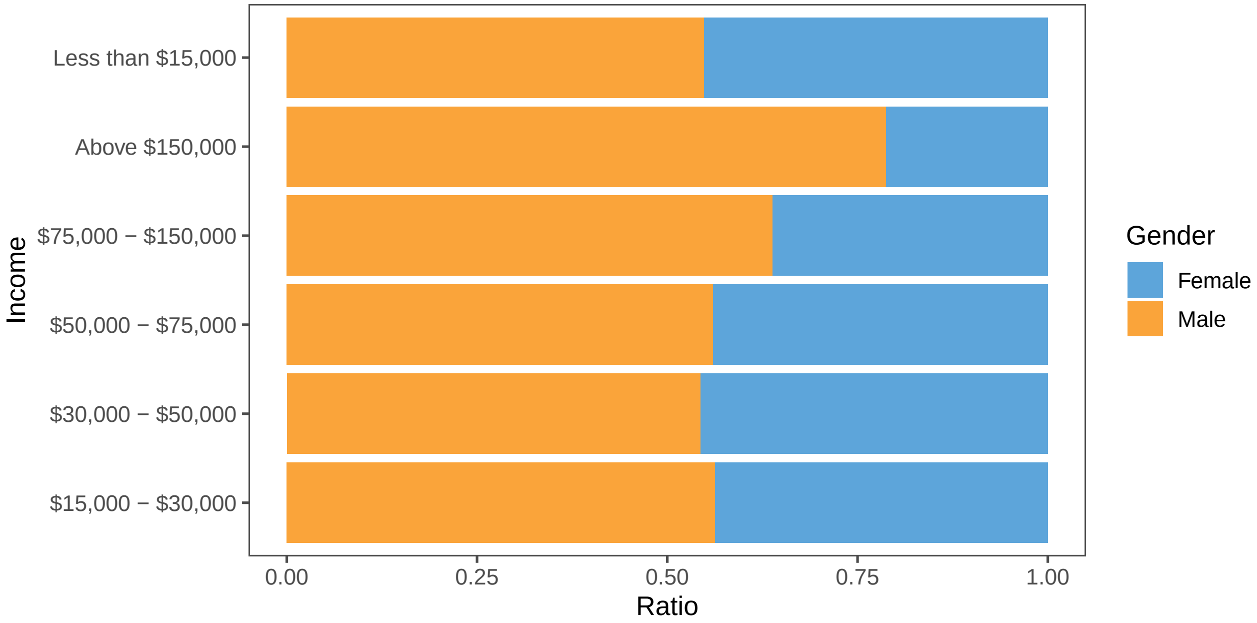
3.1.12 Worktime Age Correlation Test Rejects Null Hypothesis

```
cor.test(formula = ~ worktime + age,
                data = survey_rename)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  worktime and age
## t = 12, df = 1757, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.232 0.318
## sample estimates:
##   cor
## 0.275
```

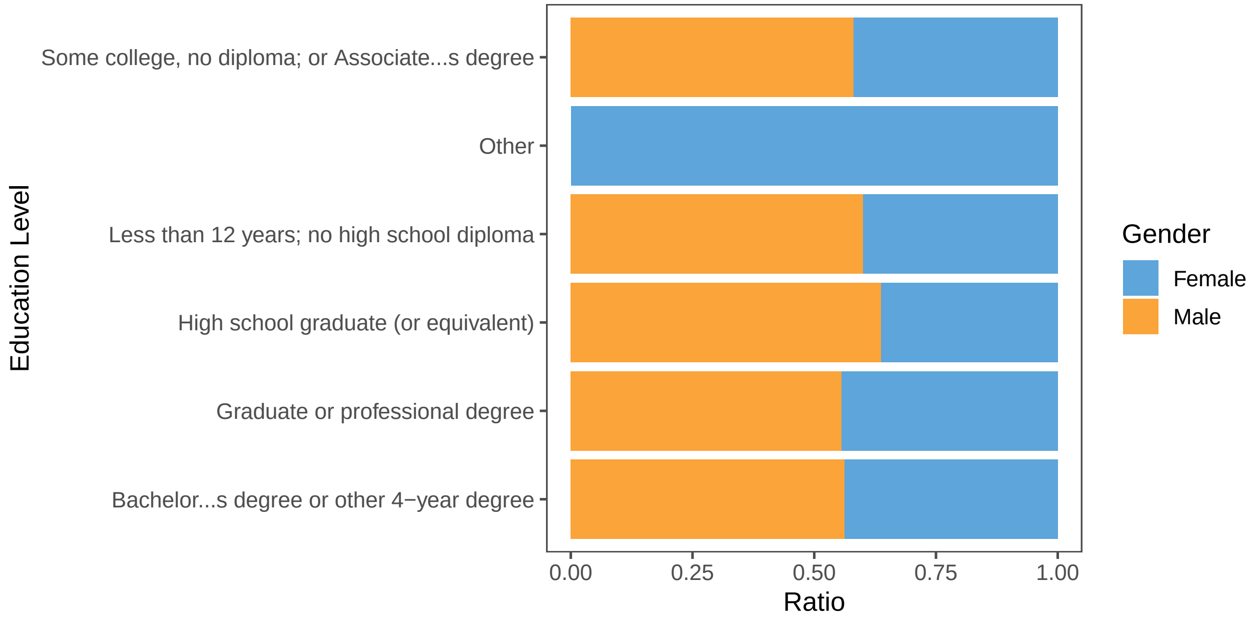
3.1.13 Income Bar Graph Exhibits More Male Responses

```
survey_rename %>%
  select(income, gender) %>%
  na.omit() %>%
  ggplot(aes(, y=income, fill=gender, NA.rm=TRUE)) +
  geom_bar(position = "fill") +
  theme_few() +
  scale_fill_few("Medium") +
  labs(y = "Income", x="Ratio", fill="Gender")
```



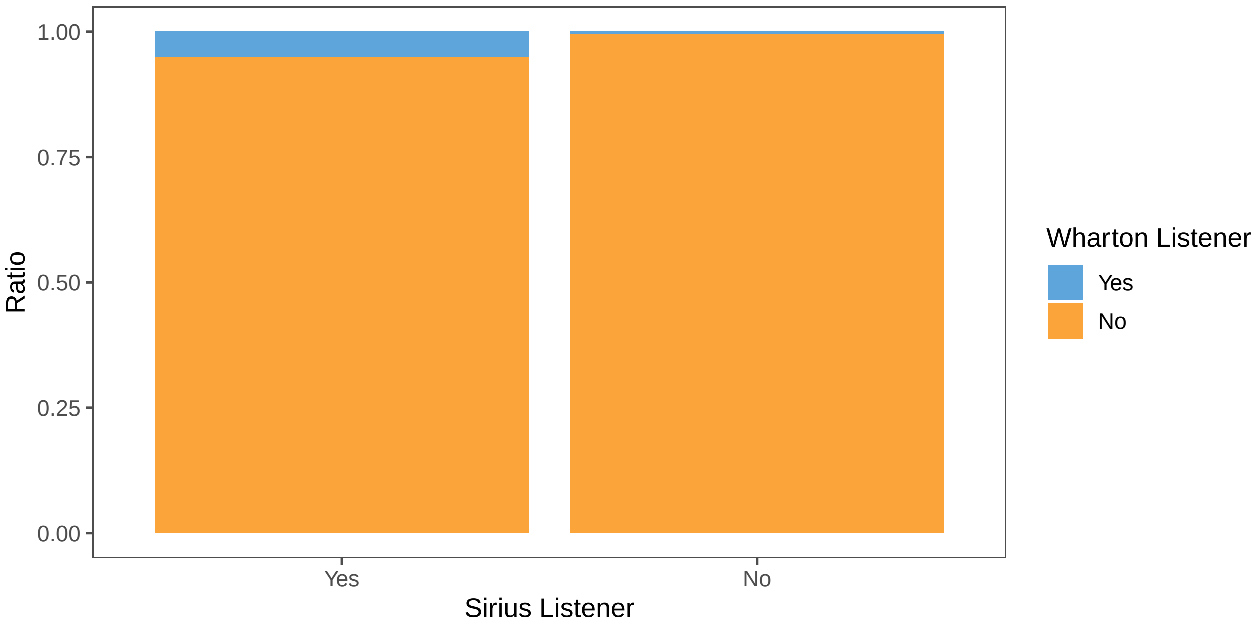
3.1.14 Education Bar Graph Exhibits Roughly Equal Responses

```
survey_rename %>%
  select(education, gender) %>%
  na.omit() %>%
  ggplot(aes(, y=education, fill=gender)) +
  geom_bar(position = "fill") +
  theme_few() +
  scale_fill_few("Medium") +
  labs(y = "Education Level", x="Ratio", fill="Gender")
```



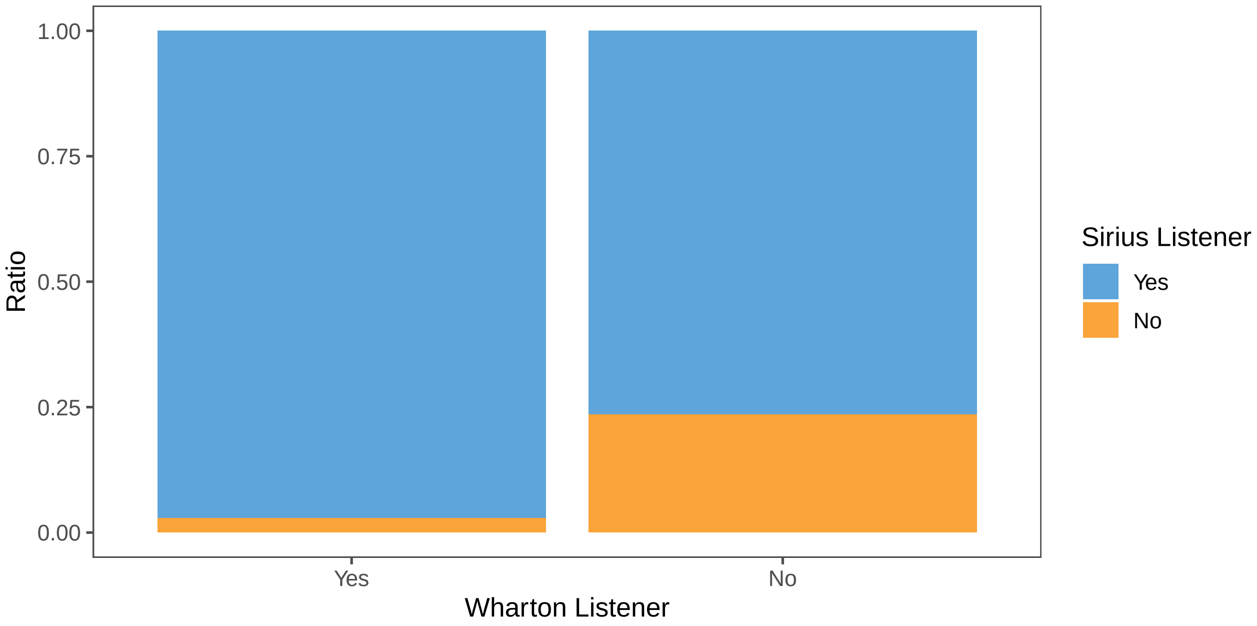
3.1.15 Sirius Bar Graph Shows Low Wharton Ratio

```
survey_rename %>%
  select(sirius, wharton) %>%
  arrange(wharton) %>%
  arrange(sirius) %>%
  mutate(sirius = factor(sirius, levels=c("Yes", "No"))) %>%
  mutate(wharton = factor(wharton, levels=c("Yes", "No"))) %>%
  na.omit() %>%
  ggplot(aes(x= sirius, , fill=wharton)) +
  geom_bar(position = "fill") +
  theme_few() +
  scale_fill_few("Medium") +
  labs(x = "Sirius Listener", y="Ratio", fill="Wharton Listener")
```



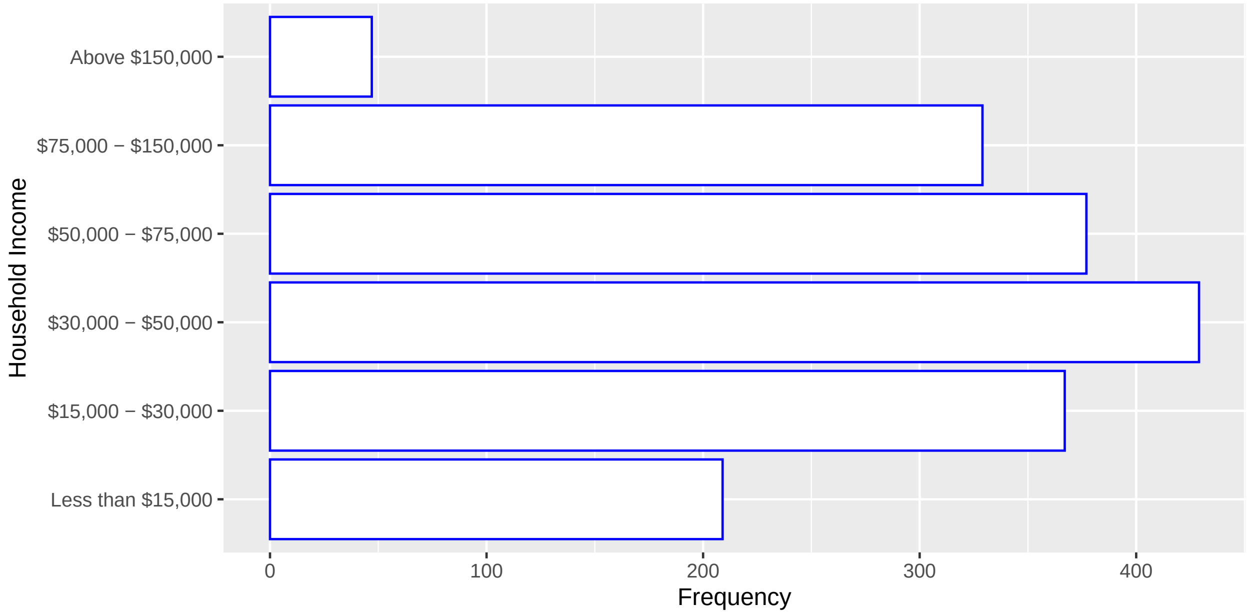
3.1.16 Wharton Bar Graph Shows High Sirius Ratio

```
survey_rename %>%
  select(wharton, sirius) %>%
  arrange(wharton) %>%
  arrange(sirius) %>%
  mutate(sirius = factor(sirius, levels=c("Yes", "No"))) %>%
  mutate(wharton = factor(wharton, levels=c("Yes", "No"))) %>%
  na.omit() %>%
  ggplot(aes(x = wharton, , fill = sirius)) +
  geom_bar(position = "fill", na.rm=TRUE) +
  theme_few() +
  scale_fill_few("Medium") +
  labs(x = "Wharton Listener", y="Ratio", fill="Sirius Listener")
```



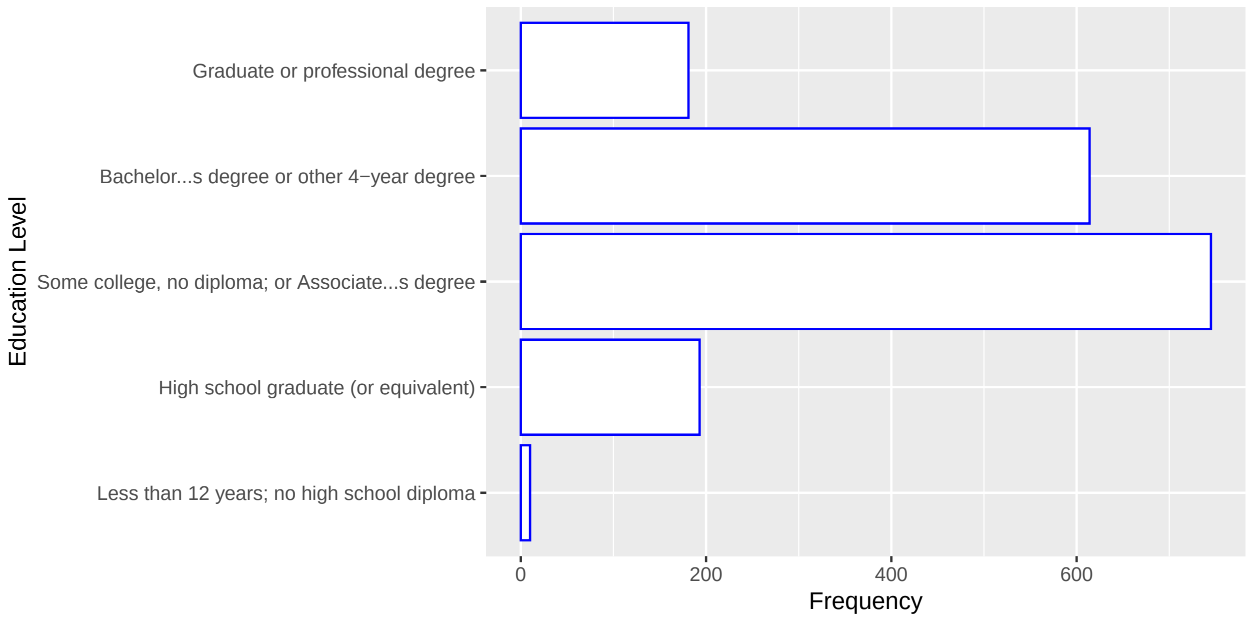
3.1.17 Frequency of Household Income Categories

```
survey_rename %>%
  select(income) %>%
  arrange(income) %>%
  mutate(income = factor(income, levels=c("Less than $15,000", "$15,000 - $30,000", "$30,000 - $50,000"
    "$50,000 - $75,000", "$75,000 - $150,000", "Above $150,000"))) %>%
  na.omit() %>%
  ggplot(aes(y = income)) +
  geom_bar(stat = "count", color = "blue", fill = "white") +
  labs(y = "Household Income", x = "Frequency")
```



3.1.18 Frequency of Education Level Categories

```
survey_rename %>%
  select(education) %>%
  arrange(education) %>%
  mutate(education = factor(education, levels=c("Less than 12 years; no high school diploma",
    "High school graduate (or equivalent)", "Some college, no diploma; or Associate’s degree",
    'Bachelor’s degree or other 4-year degree', "Graduate or professional degree"))) %>%
  na.omit() %>%
  ggplot(aes(y = education)) +
  geom_bar(stat = "count", color = "blue", fill = "white") +
  labs(y = "Education Level", x = "Frequency")
```



3.2 Sample properties

The population from which the sample is drawn determines where the results of our analysis can be applied or generalized. We include some basic demographic information for the purpose of identifying sample bias, if any exists. Combine our data and the general population distribution in age, gender and income to try to characterize our sample on hand.

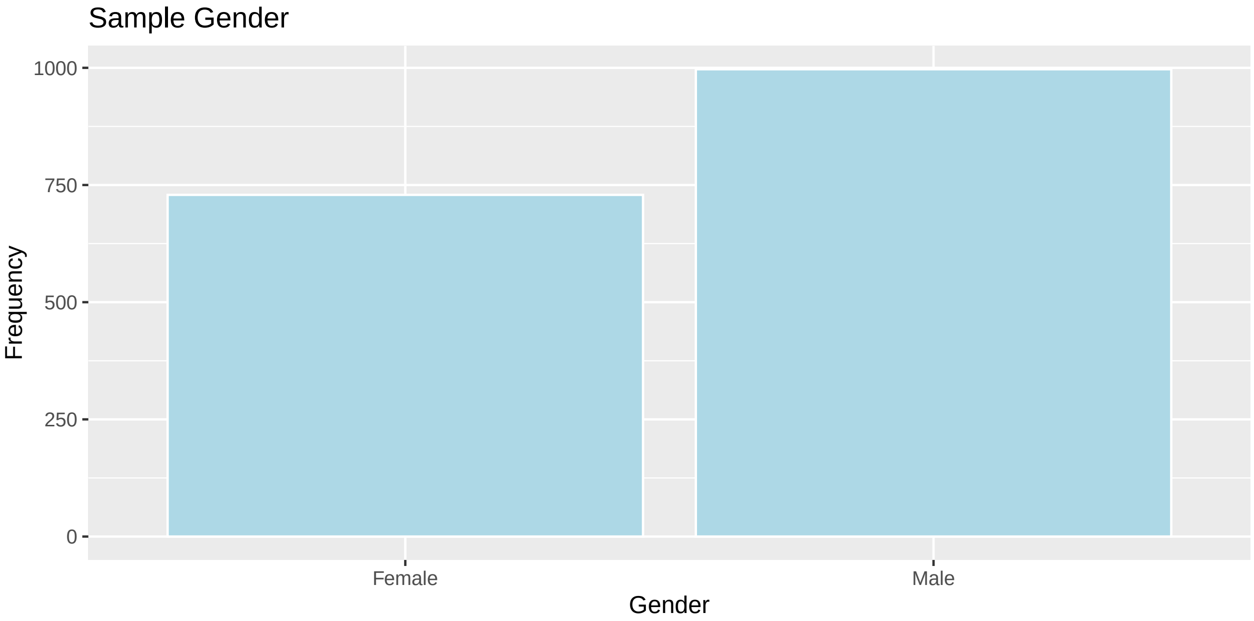
i. Does this sample appear to be a random sample from the general population of the USA?

Null Hypothesis: This sample is a random sample from the general population of the USA.

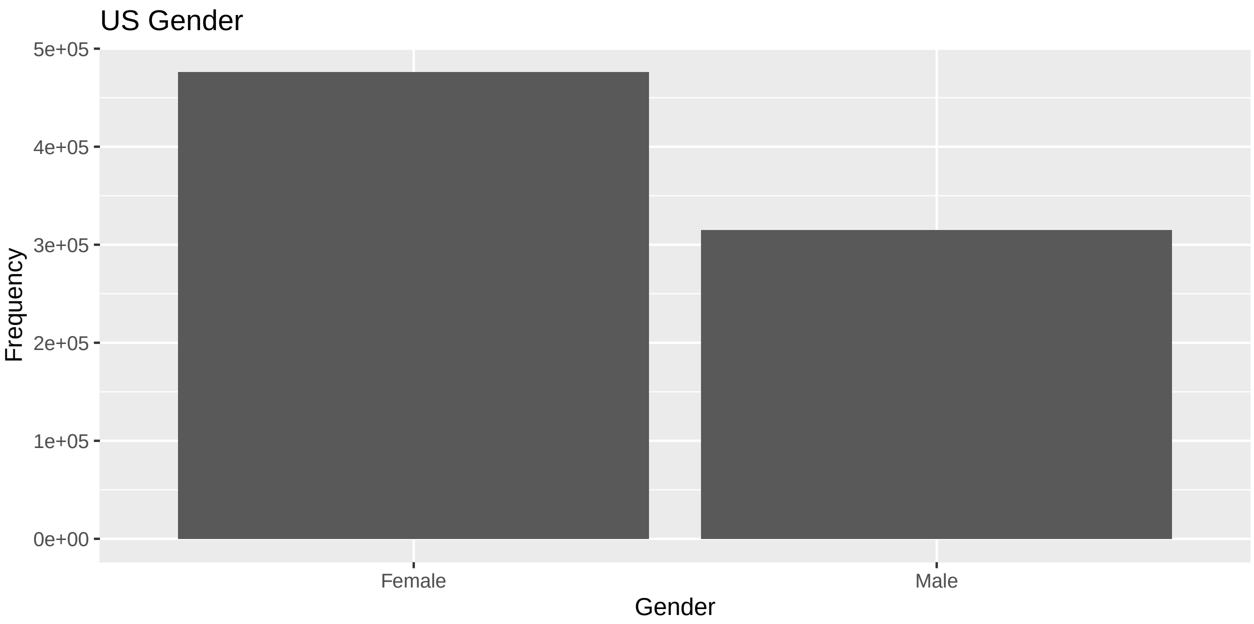
3.2.1 Sample Gender vs. US Gender

The blue graph refers to the data set provided, while the black graph is the total female and male count in the US in 2013.

```
#  survey_rename %>%
#  select(gender) %>%
#  group_by(gender) %>%
#  summarise(n())
survey_rename %>%
  na.omit() %>%
  ggplot(aes(x = gender)) +
  geom_bar(stat = "count", color = "white", fill = "light blue") +
  labs(x = "Gender", y = "Frequency", title="Sample Gender")
```



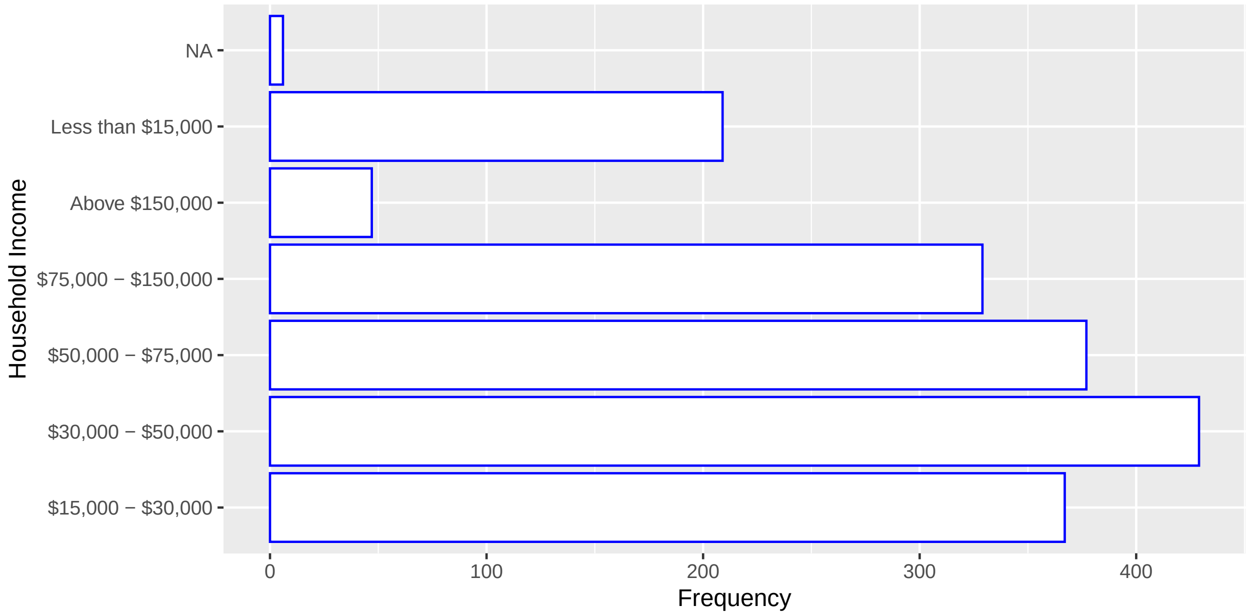
```
df <- data.frame(Gender =c("Male", "Female"),
                Frequency=c(314848 , 476342))


library(ggplot2)
ggplot(data=df, aes(x=Gender, y=Frequency)) +
  geom_bar(stat="identity") +
  labs(x = "Gender", y = "Frequency", title="US Gender")
```



```
#grid.arrange(pp,p , ncol = 2)
```

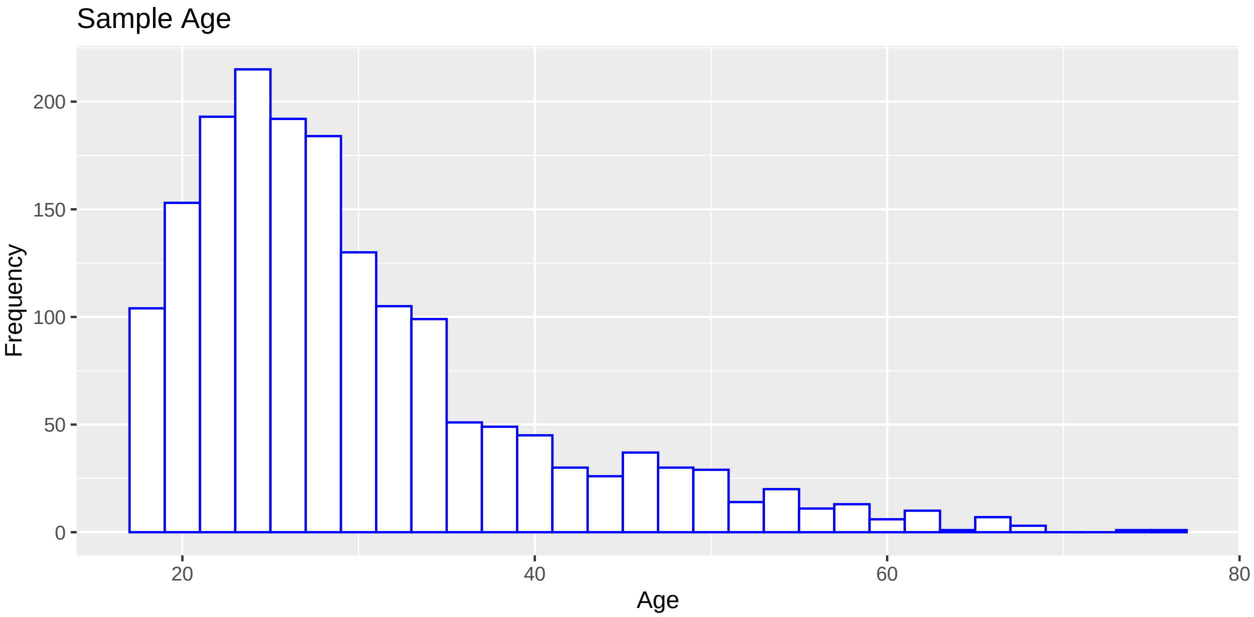
By visual observation, the null hypothesis for `gender` can be rejected.

```
education_income <- read.csv("data/education_income.csv")

education_income_col <- education_income %>%    select(Less.Than.9th.Grade,Some.College.No.Degree..or.a

 education_income_rename <-  education_income_col %>% rename(high_school_or_less = Less.Than.9th.Grade

 education_income_rename$professional_or_doctorate[education_income_rename$professional_or_doctorate==

 survey_rename %>%
 ggplot(aes(y = income)) +
 geom_bar(stat = "count", color = "blue", fill = "white") +
 labs(y = "Household Income", x = "Frequency")
```

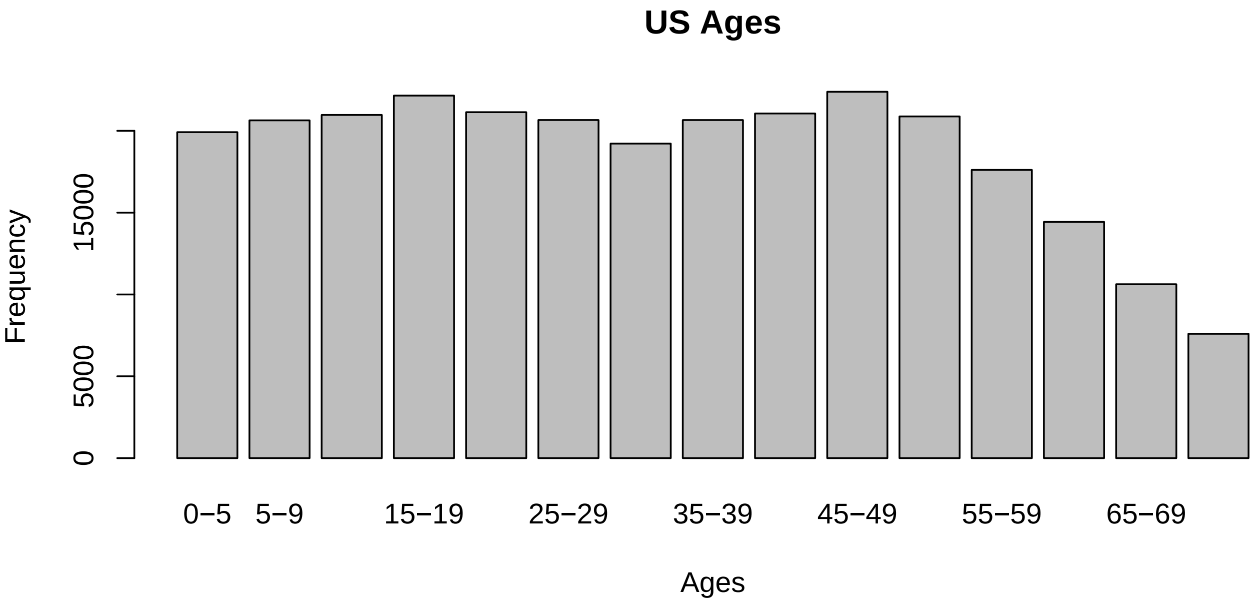


The sample does not show the general population of the USA. The amount of people in each age group is pretty similar in the general population. The sample showed that there was a huge difference between the amount of people in each age group.

```
survey_rename %>%
  ggplot(aes(x = age)) +
  geom_histogram(binwidth = 2, color = "blue", fill = "white") +
  labs(x = "Age", y="Frequency", title="Sample Age")
```



```
UsAge <- read.csv("data/2013age_table1.csv", header = T, stringsAsFactors = FALSE)

Us <- data.frame(Age=c("0-5","5-9","10-14",
                       "15-19","20-24","25-29",
                       "30-34","35-39","40-44",
                       "45-49","50-54","55-59",
                       "60-64","65-69","70-74"),
                 Number =c(19917.20356,20640,20970,22153,21138,20659,19221,20657,
                           21060,22386,20880, 17611,14437,10624,7598))
# add the data
barplot(Us$Number, names.arg=Us$Age, main="US Ages",
        xlab="Ages",
        ylab="Frequency",)
```



```
#creating a graph
```

Note: You can not provide evidence by simply looking at our data here. For example, you need to find distribution of education in our age group in US to see if the two groups match in distribution. Please do not spend too much time gathering evidence.

3.3 Final estimate

Give a final estimate of the Wharton audience size in January 2014. Assume that the sample is a random sample of the MTURK population, and that the proportion of Wharton listeners vs. Sirius listeners in the general population is the same as that in the MTURK population. Write a brief executive summary to summarize your findings and how you came to that conclusion.

```
survey_rename  %>%
    select(wharton) %>%
    group_by(wharton) %>%
    summarise(n())
```

```
## # A tibble: 3 x 2
##   wharton ‘n()‘
##   <chr>   <int>
## 1 No       1690
## 2 Yes        70
## 3 <NA>        4
```

```r
audience <- 70/1770 * 51.6
print(audience)
```

```
## [1] 2.04
```

Final Estimate: 2.04 million listeners in Jan 2014.

To be specific, you should include:

1. Goal of the study:
2. Method used: data gathering, estimation methods:
3. Findings:
4. Limitations of the study.

3.4 Executive Summary

The goal of this study is to predict the size of the audience regarding the “Business Radio Powered by the Wharton School” in terms of the whole 51.6 million Sirius Radio listeners. In order to do so, we picked out specific variables- age, gender, education level, household income, whether one listens to sirius and/or wharton, and worktime. It was necessary to clean the data by assigning unusable values NA or their correct values. The data was gathered using MTurk, at an offered price for $0.10 for each answered survey for 6 days. Through caclulating the proportion of listeners for Wharton to those who don’t listen to Wharton, and multiplying it to the assumed amount of Sirius XM listeners, we get a value of 2.04 million listeners. The limitations of this study are the fact that most of the results came from the first two days, biasing the study towards the people surveyed on those two days. At the same time, this survey is only representative of the people who were willing to partake in the survey, not those who declined, further skewing the data.

3.5 End