WDS, HW 1

Group Member Andrew Group Member Sahej Group Member Adithya
Group Member Raymond Group Member Kamron

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".

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.

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.)

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?

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.

To be specific, you should include:

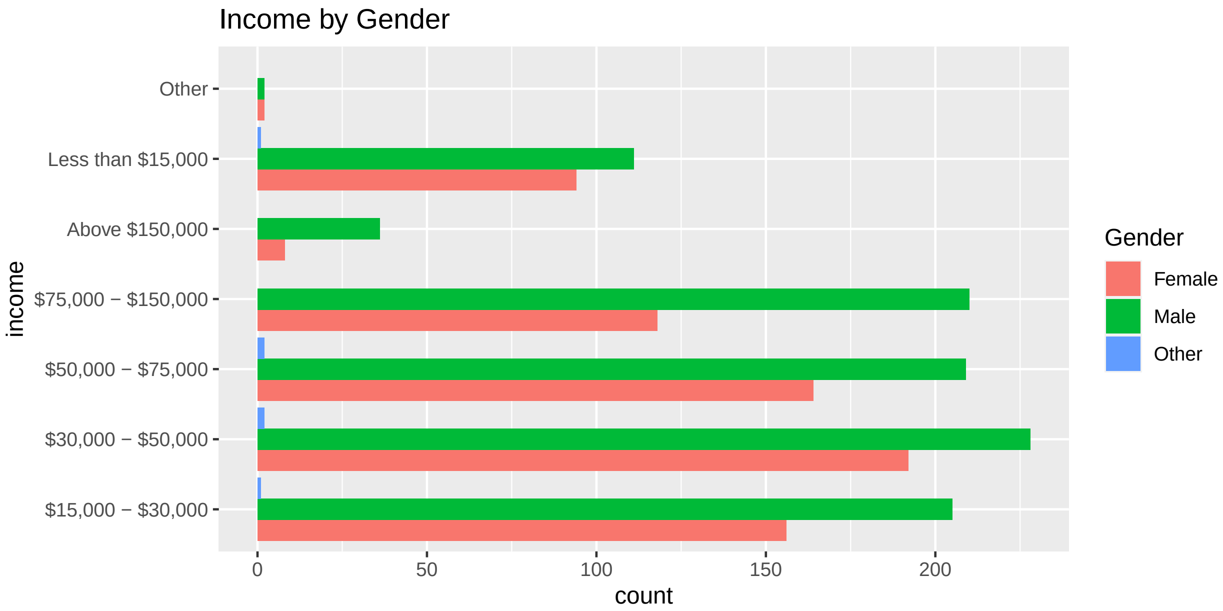
1. Goal of the study How successful is the Wharton Talk Show
2. Method used: data gathering, estimation methods
3. Findings
4. Limitations of the study.

```r
# Read in the data with correct cols
# We need to set the encoding to get ’ to render properly
# We will later remove rows with null cols, so fill = TRUE will be dealt with

sirius_data <- fread("/Users/adiay/Downloads/Wharton/HW1/data/Survey_results_final.csv", select = c("Ans


# Rename Columns
sirius_data <- sirius_data %>%
  rename(
    age = "Answer.Age",
    education = "Answer.Education",
    gender = "Answer.Gender",
    income = "Answer.HouseHoldIncome",
    sirius = "Answer.Sirius Radio",
    wharton = "Answer.Wharton Radio",
    worktime = "WorkTimeInSeconds"
  )

# Data Cleanup
# Age Cleanup: Remove any ages that are not between 1 and 122 (the oldest anyone has ever been) as well
# NB: There is one oddly formatted input for 18 that could be reformatted here
sirius_data <- sirius_data[sirius_data$age %in% 1:122]
```

```
# Education Cleanup: Remove things that are not one of several choices
sirius_data[sirius_data == ""] <- "Other"
sirius_data <- sirius_data[sirius_data$education %in% c("Some college, no diploma; or Associate’s degre
                                                        "Graduate or professional degree",
                                                        "Bachelor’s degree or other 4-year degree",
                                                        "High school graduate (or equivalent)",
                                                        "Less than 12 years; no high school diploma",
                                                        "Other")]

# Data Cleanup: Remove nulls from the data
sirius_data.gender <- sirius_data[sirius_data$gender != ""]
sirius_data.income <- sirius_data[sirius_data$income != ""]
sirius_data.sirius <- sirius_data[sirius_data$sirius != ""]
sirius_data.wharton <- sirius_data[sirius_data$wharton != ""]

str(sirius_data)

summary(sirius_data)
```

#The summary shows sample size of 1741 objects of 7 variables # Plotting using gglot income vs Gender

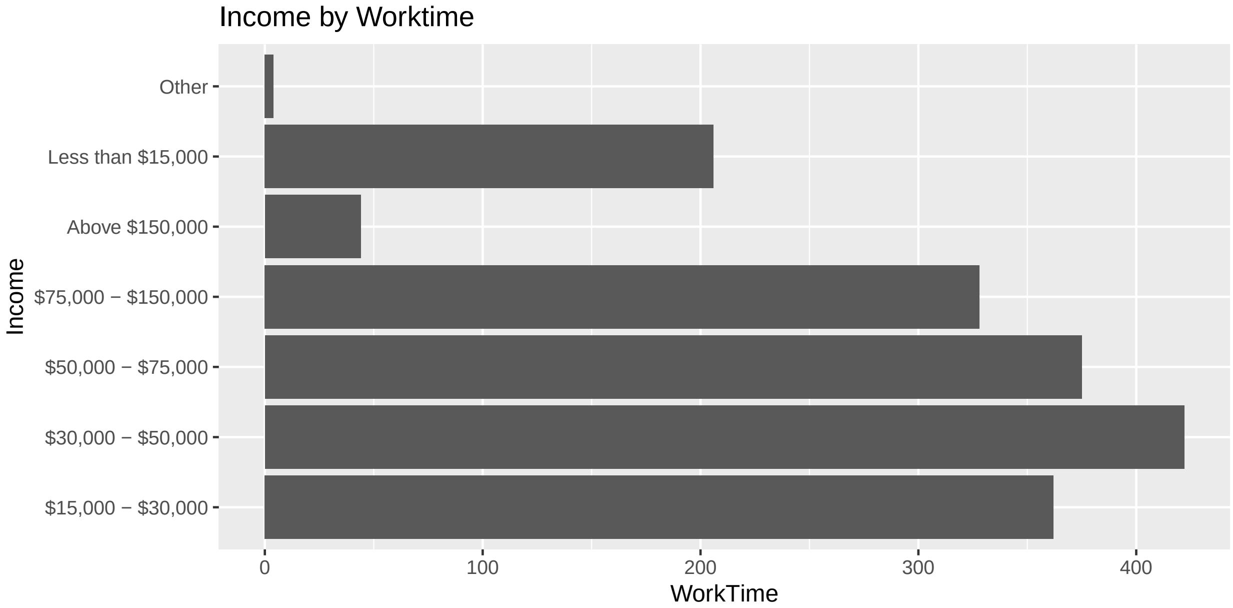
```
ggplot(sirius_data, aes(y=income, fill=gender)) +
  geom_bar(position = position_dodge(preserve = "single"))+
labs(
       fill = "Gender",
       x = "count",
       title = "Income by Gender")
```



# From the income by gender graph, we know that men earn more than women.

4 Plotting using gglot income vs Worktime

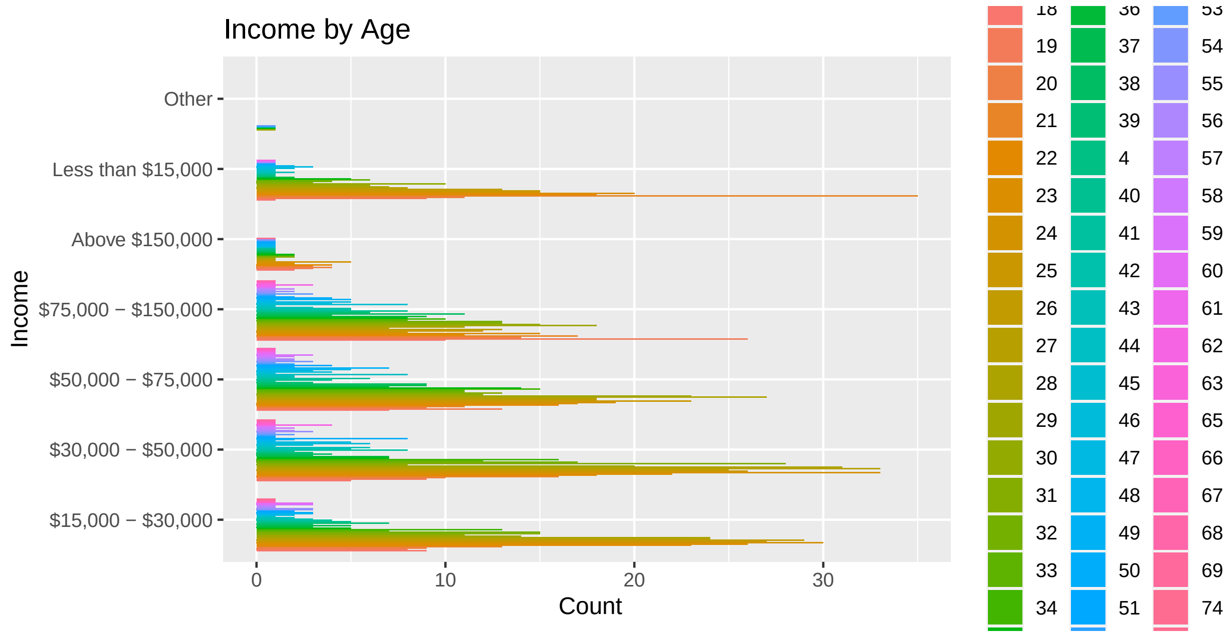
```
ggplot(sirius_data, aes(y=income, fill=worktime)) +
  geom_bar(position = position_dodge(preserve = "single"))+
labs( y="Income",
       fill = "Income",
       x = "WorkTime",
       title = "Income by Worktime")
```



# In this income by work time graph, the worktime increases initially, when income increases but at a certain point, in this case #50,000, the worktime will decrease.

5 Plotting using gglot income vs age

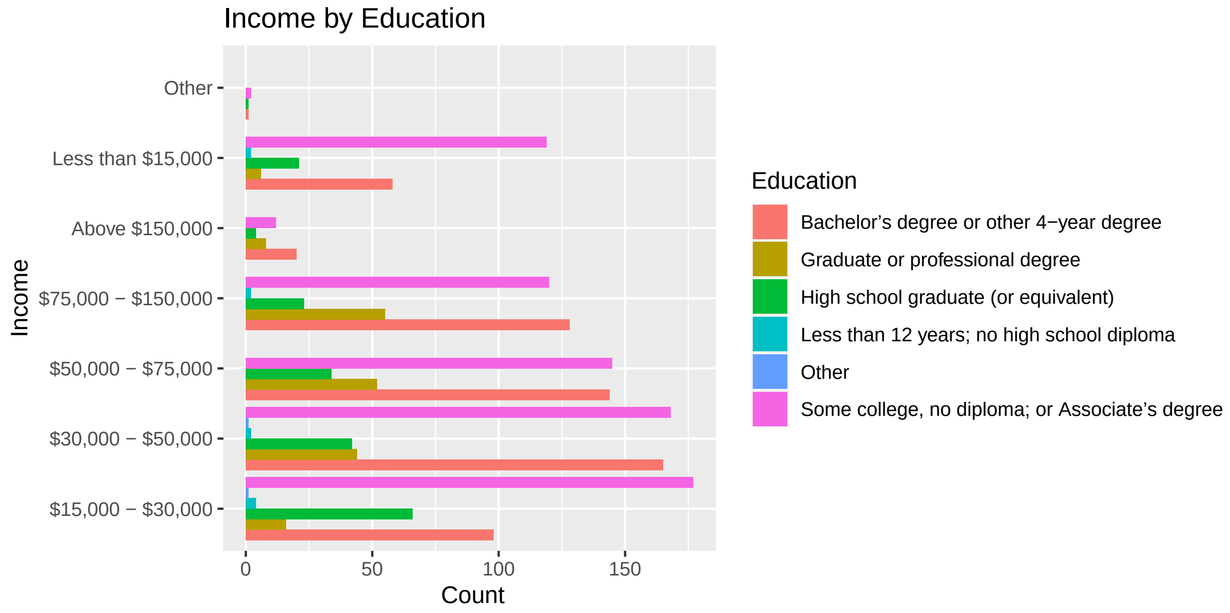
```
ggplot(sirius_data, aes(y=income, fill=age)) +
geom_bar(position = position_dodge(preserve = "single"))+
labs( y="Income",
      fill = "Age",
      x = " Count",
      title = "Income by Age")
```



# Because there are more young people in the workforce, they are overrepresented in all of the income categories in contrast to the #older age groups.

6 Plotting using gglot income vs education

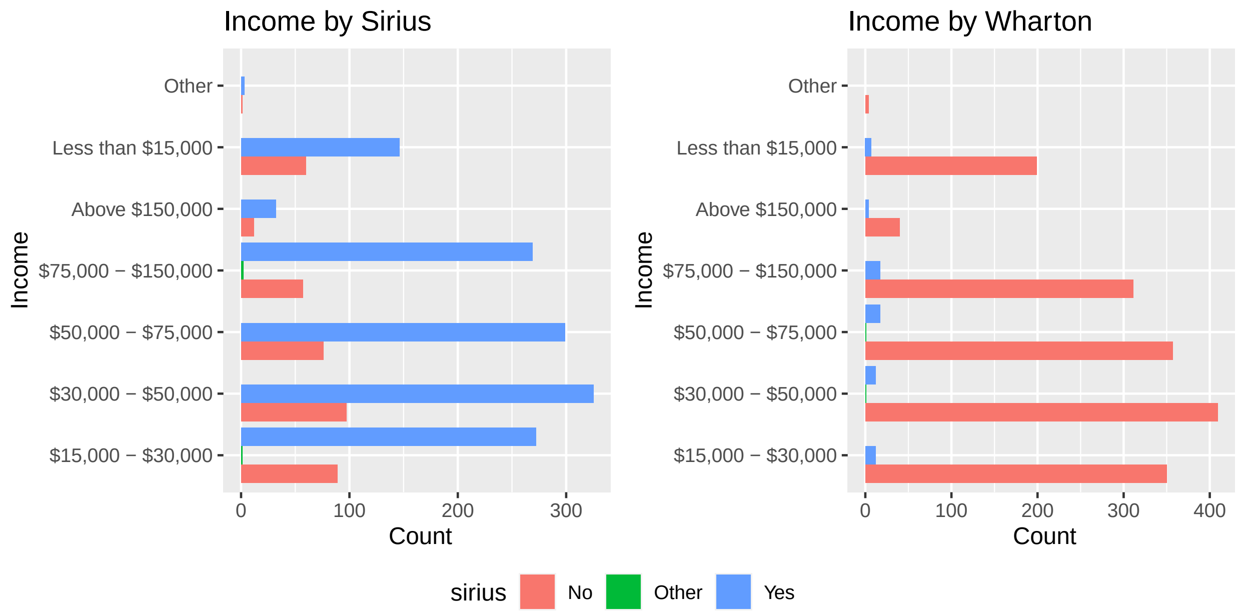
```
ggplot(sirius_data, aes(y= income, fill = education)) +
 geom_bar(position = position_dodge(preserve = "single"))+
labs( y="Income",
       fill = "Education",
       x = " Count",
       title = "Income by Education")
```



# People with some college education and more are more likely to earn more and are overrepresented in all the income categories.

7 Bivariate Relationships comparing Income in Sirius vs Wharton

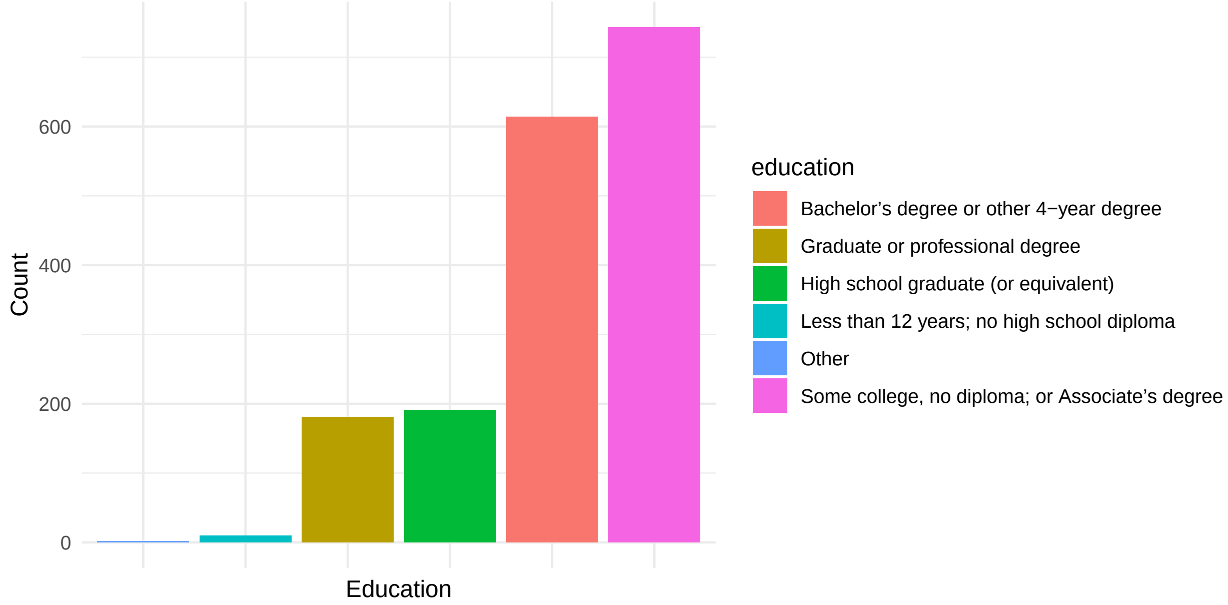
```
sirius_plot = ggplot(sirius_data, aes(y= income, fill = sirius)) +
 geom_bar(position = position_dodge(preserve = "single"))+
labs( y="Income",
       fill = "sirius",
       x = " Count",

       title = "Income by Sirius")

wharton_plot = ggplot(sirius_data, aes(y= income, fill = wharton)) +
 geom_bar(position = position_dodge(preserve = "single"))+
labs( y="Income",
       fill = "wharton",
       x = " Count",

       title = "Income by Wharton")
p <- ggpubr::ggarrange(sirius_plot, wharton_plot, common.legend = T, legend = "bottom", ncol = 2)
p
```



# People who listen to Sirius and Wharton make roughly the same amount of money.

8 Education

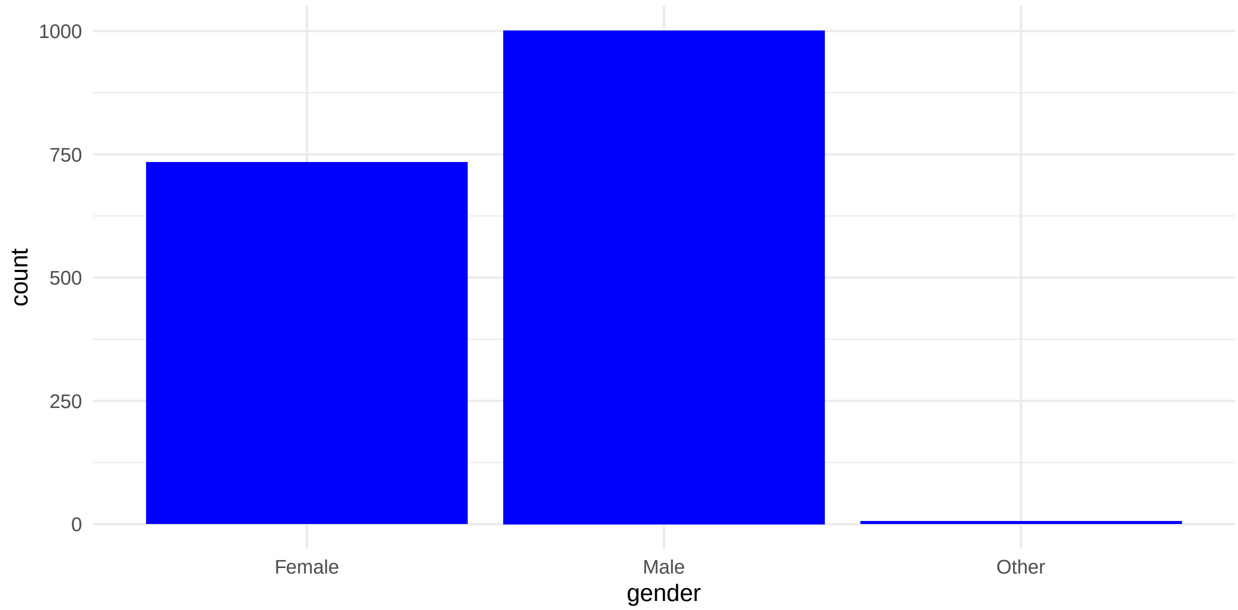
```
count(sirius_data, education) %>%
      ggplot(aes(x = reorder(education, n), y = n, fill=education)) +
      geom_col() +
      theme_minimal() +
      theme(axis.text.x=element_blank(),
            axis.ticks.x=element_blank()) +
      labs(y = "Count", x="Education")
```



# People are more likely to have some college and no diploma or associate degree or bachelor's degree than other qualifications.

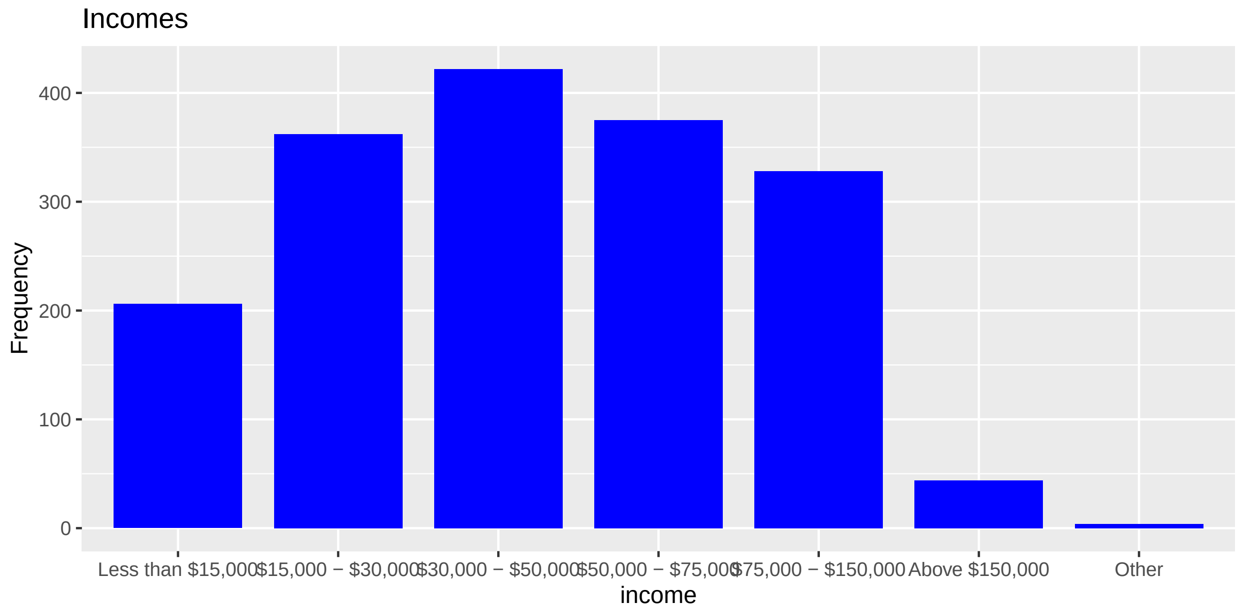
9 Income

```
ggplot(sirius_data) +
  geom_bar(aes(x = gender), fill = "blue") +
  theme_minimal()
```



#This graph shows that there are more men listeners than women.

```
sirius_data$income <- factor(sirius_data$income, levels = c("Less than $15,000", "$15,000 - $30,000", "$
ggplot(sirius_data) +
  geom_bar(aes(x = income), fill = "blue", width = 0.8) +
  labs( title = "Incomes", x = "income" , y = "Frequency")
```



#The graph is unimodal and the median is around 30000-50000 and people are most likely to have incomes from 30000 to 50000.

10 Final Answer

Set tables to get the amt of people who listen to Wharton relative to the num of people who listen to sirius

```
temp_wharton <- (sirius_data %>% count(wharton))
temp_sirius <- (sirius_data %>% count(sirius))

percent_wharton <- (temp_wharton[3,2]/temp_sirius[3,2])
percent_wharton

total_listeners <- percent_wharton*25800000
```

11 Formatted Output

```
dd=data.frame(WhartonPercent =percent_wharton, TotalWhartonListeners = total_listeners)
names(dd)[names(dd)=="n"] <- "Wharton Percentage"
names(dd)[names(dd)=="n.1"] <- "Total Listeners"
knitr::kable(head(dd))
```

12 Conclusion Summary and Inferences

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

# Age Cleanup: Remove any ages that are not between 1 and 122 (the oldest anyone has ever been) as well as null values and badly formatted data # NB: There is one oddly formatted input for 18 that could be reformatted here # Work time is collected by the polling site, removing the ability for user error. # Education Cleanup: Remove things that are not one of several choices # Data Cleanup: Remove nulls from the data

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.

We found 1741 obs of 7 variables such as age, education, gender, income, sirius, wharton, worktime. Many of them are part time workers who have done some college and are doing a job to gain experience.

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

#When we compare the US population demographics with our data the US data has the following percentages related to Education, age, gender 1. High school graduate or higher over 25 yrs is 88%. 2. Bachelors Degree , percent of person's age over 25yrs is 32.1%. 3. Female (50.8%) to Male Ratio is almost 1:1 in the years 2015- 2019. #Our data 1.Female to Male ratio is 750:1000 which is 3:4. 2.High School graduate or higher 1550/1741 which is 89%. 3.Bachelors degree is 610/1741 which is 35%.

#Since both are similar we conclude that this sample appears to be a random sample from the general population of the USA. #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.)

#Final Answer Wharton Percentage:0.051 Total Listeners:1322585

#Write a brief executive summary to summarize your findings and how you came to that conclusion. #We noticed that there were not much differences between the sample auidence and the US population with respect to Education and #gender. #We arrived at this conclusion by comparing our data demographics and by using ggplot to verify our findings against the #US population demographics.(https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2019/)