Project_251

Skyler Hauser

2023-10-01

```
install.packages("tidycensus")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(httr)
library(jsonlite)
library(tidycensus)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()  masks stats::filter()
## x purrr::flatten() masks jsonlite::flatten()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(naniar)
```

```
options(scipen=999)
```

Initiating the Project

Research Question

What are the key differences in occupations and academic paths for men and women who have studied Mathematics and Computer Science?

Explaining the Question, Domain Knowlegde, Concepts and Measurements

I am specifically interested in understanding how gender representation varies within fields, and possibly how it differs between academic and industry settings. I would like to identify discrepancies in gender representation within these areas. This analysis could reveal insights into gender dynamics in Mathematics and Computer Science fields, potentially shedding light on issues of gender equality, career preferences, and the impact of academic backgrounds on career paths. There tends to always be challenges in gender representation in STEM. Most of these historical and contemporary challenges, faced by mainly women, are gender bias and underrepresentation. These occur in the fields of which Mathematics and Computer Science typically withhold, like academia (professors, researchers and lecturers) and industry (software engineers, data scientist, and systems analysts). Understanding the nature of work and career paths in these fields help in contextualizing the data that I will present. Gender, in my opinion, is the primary variable for analyzing differences in representation, gender is represented as Male and Female. The Field of Study represents individuals who majored in Mathematics and Computer Science. Occupation is the classification of the respondents current jobs distinguishing between academic roles and industry roles. Educational Attainment are the levels of education the respondents have achieved. Race is also represented which allows us to view diversity and representation in these fields.

The Data

Data Origin Story

The purpose of the data collection was to gather information on various characteristics of the American population, such as demographic, socioeconomic, and housing characteristics. The ACS aims to provide communities with reliable and detailed data that can be used for planning and decision-making purposes.The data was collected by the American Community Survey (ACS).The ACS Public Use Microdata Sample (PUMS) was used to collect the data. This dataset contains a sample of responses to the ACS and includes variables for nearly every question on the survey. The data is comprised of records representing individual persons or households. The data encompasses not only typical households but also those living in Group Quarters like nursing and college facilities. The data is available at national, state, and Public Use Microdata Area (PUMA) levels. Representativeness is not a concern since it is a survey available for all Americans to answer, ensuring a broad and diverse set of respondents. Since the description, the data aims to be representative as it covers a vast spectrum of the population, including those in households and Group Quarters. Additionally, the PUMA division ensures geographic representation with each unit containing approximately 100,000 people. However, the actual representativeness would depend on the response rate and if certain demographics were more or less likely to respond.The data is divided into PUMAs, ensuring that each state is partitioned into contiguous geographic units containing roughly 100,000 people. This method helps in making sure that data is geographically representative. No specific choices have been mentioned that might limit the representativeness of the data.Potential sources of bias could include non-response bias if certain groups are less likely to complete the survey. Additionally, those without access to the survey medium or with language barriers might be underrepresented. Cultural or social reasons might also deter certain groups from participating. Main potential sources of missing data could include:Non-responses from certain demographics, Technical errors during data collection, People being unaware of the survey and Language or literacy barriers preventing completion. From what is provided, specifics about what is not being measured are not mentioned. However, any variable not directly asked in the survey or derived from the questions would be missing. The main variables in the data set are derived from survey questions and include both raw and derived variables, such as poverty status. Looking at new mexico and Delaware specifically based on two different demographics.

Importing and Concepts

```r
api_key <- "7cbd5ec4997e5ebff25a817162b4ad2069068dc7"

data("pums_variables")

variables <- pums_variables %>%
  filter(year==2021, survey =='acs1', level=="person") %>%
  distinct(var_code, var_label, data_type)


de_pums<- get_pums(
  variables = c("FOD1P", "RAC1P", "SCIENGP", "PINCP", "COW", "SOCP", "SEX", "AGEP", "SCHL"),
  state = "New Mexico",
  survey = "acs1",
  year = 2021
)
```

```
## Getting data from the 2021 1-year ACS Public Use Microdata Sample
```

```
## Downloading: 7.6 kB     Downloading: 7.6 kB     Downloading: 16 kB     Downloading: 16 kB     Downloa
```

```
head(de_pums)
```

```
## # A tibble: 6 x 14
##   SERIALNO  SPORDER  WGTP PWGTP  PINCP  AGEP ST    COW   SCHL  SEX   FOD1P RAC1P
##   <chr>       <dbl> <dbl> <dbl>  <dbl> <dbl> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 2021HU04~       1   135   135      0    20 35    b     12    2     000N  1
## 2 2021HU04~       2   135   217   1500    20 35    1     12    1     000N  1
## 3 2021HU04~       3   135   189 -19999     2 35    b     bb    1     000N  1
## 4 2021HU04~       1    10    10   9300    59 35    b     16    1     000N  1
## 5 2021HU04~       1   145   144  43200    28 35    7     16    1     000N  1
## 6 2021HU04~       2   145   224  20800    30 35    7     18    2     000N  1
## # i 2 more variables: SCIENGP <chr>, SOCP <chr>
```

```
de_pums2 <- get_pums(
  variables = c("FOD1P", "RAC1P", "SCIENGP", "PINCP", "COW", "SOCP", "SEX", "AGEP", "SCHL"),
  state = "Delaware",
  survey = "acs1",
  year = 2021
)
```

```
## Getting data from the 2021 1-year ACS Public Use Microdata Sample
```

```
## Downloading: 16 kB     Downloading: 16 kB     Downloading: 16 kB     Downloading: 16 kB     Download
```

```
head(de_pums2)
```

```
## # A tibble: 6 x 14
##   SERIALNO  SPORDER  WGTP PWGTP  PINCP  AGEP ST    COW   SCHL  SEX   FOD1P RAC1P
##   <chr>       <dbl> <dbl> <dbl>  <dbl> <dbl> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 2021GQ00~       1     0   100      0    21 10    1     19    1     000N  2
## 2 2021GQ00~       1     0    74   1800    22 10    1     16    1     000N  2
## 3 2021GQ00~       1     0    10    230    32 10    1     19    1     000N  1
## 4 2021GQ00~       1     0    79 -19999    13 10    b     10    1     000N  1
## 5 2021GQ00~       1     0    61      0    32 10    1     13    1     000N  9
## 6 2021GQ00~       1     0    59      0    27 10    b     11    1     000N  1
## # i 2 more variables: SCIENGP <chr>, SOCP <chr>
```

To obtain my data, I had to go through the steps for retrieving an API key. Once I was able to attain my API key. The API key allowed me to access the ACD data through an online database. I was then able to download the Public Use Microdata Sample of the American Community Survey into R. The "de_pums" and "de_pums2" are loading a dataset that contains variables from the Public Use Microdata Sample of the American Community Survey. I then used "get_pums" function which allowed me to retrieve data for specific variables for the state of New Mexico and Delaware. Each Row in the datasets "de_pums" and "de_pums2" represents an observation. Each observation represents an individual respondents response from the ACS for the specified year and state. I then began my tidying and cleaning process.My variables are Income, which is a numeric variable. Age which is a numeric variable, representing the respondents Age. Class_Of_Worker which is a categorical variable represented as for example: "State Government", "Self-Employed", "Federal Government" etc. Educational_Attainment which is a categorical variable represented as "Master's Degree", "Bachelor's Degree", "Doctorate Degree" etc. Sex which is a categorical variable and is represented as Male and Female. Field_Of_Degree which is a categorical variable which represents "Mathematics" or "Computer

Science" as a degree. Race which is a categorical variable and is represented as "White", "Two or More", "Asian", etc.. Field_Of_Degree_STEM which is a categorical variable represented as Yes or no. Occupation which is a categorical variable that is categorized into grouped organizations, this will be mentioned during my cleaning process.

Cleaning and Tidying

```
fod1p_labels <- c(
      "3700" = "Mathematics",
      "3702" = "Mathematics",
      "4005" = "Mathematics",
      "2101" = "Computer Science",
      "2100" = "Computer Science",
      "2102" = "Computer Science",
      "2305" =  "Mathematics",
      "N"= "N/A (less than bachelor's degree)",
      "3701"= "Mathematics")
```

```
socp_labels <- c(
    "339021" = "Private Detectives And Investigators",
      "271021" = "Commercial And Industrial Designers",
      "499031"= "Home Appliance Repairers",
      "397010"= "Tour And Travel Guides",
      "1320XX"= "Other Financial Specialists",
      "29203X"= "Nuclear Medicine Technologists And Medical Dosimetrists",
      "291081"= "Podiatrists",
      "272021"= "Athletes And Sports Competitors",
      "29205X"= "Dietetic Technicians And Ophthalmic Medical Technicians",
      "353023"= "Fast Food And Counter Workers",
      "5170XX"= "Other Woodworkers",
      "273043"= "Writers And Authors",
      "519010"= "Chemical Processing Machine Setters, Operators, And Tenders",
      "113013"= "Facilities Managers",
      "113131"= "Training And Development Managers",
      "519041"= "Extruding, Forming, Pressing, And Compacting Machine Setters, Operators, And Tenders",
      "112021"= "Marketing Managers",
      "472061"="Construction Laborers",
      "192041"= "Environmental Scientists And Specialists, Including Health",
      "472152"= "Plumbers, Pipefitters, And Steamfitters",
      "272099"= "Entertainers And Performers, Sports And Related Workers, All Other",
      "519197"= "Tire Builders",
      "274030"= "Television, Video, And Film Camera Operators And Editors",
      "319094"= "Medical Transcriptionists",
      "319095"= "Pharmacy Aides",
      "392011"= "Animal Trainers",
      "172070"= "Electrical And Electronics Engineers",
      "319096"= "Veterinary Assistants And Laboratory Animal Caretakers",
      "492097"= "Audiovisual Equipment Installers And Repairers",
      "173023"= "Electrical and Electronic Engineering Technologists And Technicians",
      "172051"= "Civil Engineers",
      "151252"= "Software Developers",
      "151251"= "Computer Programmers",
      "551010"= "Military Officer Special And Tactical Operations Leaders",
      "332020"= "Fire Inspectors",
      "434051"= "Customer Service Representatives",
      "434171"= "Receptionists And Information Clerks",
      "434081"= "Hotel, Motel, And Resort Desk Clerks",
      "359031"= "Hosts And Hostesses, Restaurant, Lounge, And Coffee Shop",
      "514120"= "Welding, Soldering, And Brazing Workers",
      "119111"= "Medical And Health Services Managers",
      "131141"= "Compensation, Benefits, And Job Analysis Specialists",
      "211013"= "Marriage And Family Therapists",
      "434041"= "Credit Authorizers, Checkers, And Clerks",
      "292034"= "Radiologic Technologists And Technicians",
      "292035"= "Magnetic Resonance Imaging Technologists",
      "536030"= "Transportation Service Attendants",
      "412021"= "Counter And Rental Clerks",
      "514050"= "Metal Furnace Operators, Tenders, Pourers, And Casters",
      "3940XX"= "Embalmers, Crematory Operators And Funeral Attendants",
      "537065"= "Stockers And Order Fillers",
      "516011"= "Laundry And Dry-Cleaning Workers",
      "536061"= "Passenger Attendants",
      "433031"= "Bookkeeping, Accounting, And Auditing Clerks",
      "119021"= "Construction Managers",
      "513099"= "Food Processing Workers, All Other",
      "533051"= "Bus Drivers, School",
      "533052"= "Bus Drivers, Transit And Intercity",
      "132061"= "Financial Examiners",
      "4750YY"= "Derrick, Rotary Drill, And Service Unit Operators, And Roustabouts, Oil And Gas",
      "119070"= "Entertainment And Recreation Managers",
      "533030"= "Driver/Sales Workers And Truck Drivers",
      "475040"= "Underground Mining Machine Operators",
      "292042"= "Emergency Medical Technicians",
      "272030"= "Dancers And Choreographers",
      "353011"= "Bartenders",
      "475020"= "Surface Mining Machine Operators And Earth Drillers",
      "131199"= "Business Operations Specialists, All Other",
      "27102X"= "Other Designers",
      "999920"= "Unemployed",
      "519020"= "Crushing, Grinding, Polishing, Mixing, And Blending Workers",
      "113012"= "Administrative Services Managers",
      "193051"= "Urban And Regional Planners",
      "451011"= "First-Line Supervisors Of Farming, Fishing, And Forestry Workers",
      "1910XX"= "Other Life Scientists",
      "113121"= "Human Resources Managers",
      "474021"= "Elevator And Escalator Installers And Repairers",
      "272041"= "Music Directors And Composers",
      "439041"= "Insurance Claims And Policy Processing Clerks",
      "193033"= "Clinical And Counseling Psychologists",
      "113061"= "Purchasing Managers",
      "472181"= "Roofers",
      "493090"= "Miscellaneous Vehicle  Mechanics",
      "31113X"= "Orderlies And Psychiatric Aides",
      "473010"= "Helpers - Construction Trades",
      "252020"= "Elementary And Middle School Teachers",
      "272091"= "Disc Jockeys",
      "274021"= "Photographers",
      "49209X"= "ElectricalMechanics",
      "413041"= "Travel Agents",
      "5140XX"= "Model Makers",
      "15124X"= "Database Administrators And Architects",
      "151253"= "Software Quality Assurance Analysts And Testers",
      "172081"= "Environmental Engineers",
      "311121"= "Home Health Aides",
      "472231"= "Solar Photovoltaic Installers",
      "172141"= "Mechanical Engineers",
      "172011"= "Aerospace Engineers",
      "371012"= " Landscaping",
      "439XXX"= " Administrative Support Workers",
      "518021"= "Engineers And Boiler Operators",
      "291031"= "Dietitians And Nutritionists",
      "291151"= "Nurse Anesthetists",
      "131022"= "Wholesale And Retail Buyers",
      "291051"= "Pharmacists",
      "211029"= "Social Workers",
      "499094"= "Locksmiths",
      "537062"= "Laborers And Freight",
      "119030"= "Education And Childcare Administrators",
      "419091"= "Door-To-Door Sales Workers",
      "351011"= "Head Cooks",
      "119013"= "Farmers",
      "419099"= "Sales",
      "434XXX"= "Correspondence Clerks",
      "499051"= "Electrical Power-Line Installers",
      "454020"= "Logging Workers",
      "271024"= "Graphic Designers",
      "395092"= "Manicurists and Pedicurists",
      "339011"= "Animal Control Workers",
      "191030"= "onservation Scientists",
      "1191XX"= "Managers",
      "432021"= "Telephone Operators",
      "119051"= "Food Service Managers",
      "132011"= "Accountants And Auditors",
      "439071"= "Office Machine Operators",
      "132020"= "Property Appraisers",
      "132031"= "Budget Analysts",
      "474051"= "Highway Maintenance Workers",
      "272011"= "Actors",
      "272012"= "Producers And Directors",
      "353031"= "Waiters And Waitresses",
      "131070"= "Human Resources",
      "439081"= "Proofreaders",
      "19303X"= "Psychologists",
      "51609X"= "Textile, Apparel, And Furnishings Workers",
      "532010"= "Aircraft Pilots And Flight Engineers",
      "273041"= "Editors",
      "273011"= "Broadcast Announcers",
      "171020"= "Surveyors, Cartographers, And Photogrammetrists",
      "5191XX"= "Production Workers",
      "533011"= "Ambulance Drivers And Attendants",
      "519051"= "Furnace, Kiln, Oven, Drier, And Kettle Operators",
      "413091"= "Sales Representatives Of Services",
      "113051"= "Industrial Production Managers",
      "472031"= "Carpenters",
      "472121"= "Glaziers",
      "2740XX"= "Communication Equipment Workers",
      "435061"= "Production Planning And Expediting Clerks",
      "493050"= "Small Engine Mechanics",
      "516050"= "Tailors",
      "1720XX"= "Biomedical And Agricultural Engineers",
      "518010"= "Power Plant Operators",
      "493022"= "Automotive Glass Installers And Repairers",
      "435031"= "Public Safety Telecommunicators",
      "1930XX"= "Social Scientists",
      "172131"= "Materials Engineers",
      "435021"= "Couriers And Messengers",
      "4330XX"= "Financial Clerks",
      "434071"= "File Clerks",
      "492020"= "Radio Installers And Repairers",
      "151230"= "omputer Support Specialists",
      "131121"= "Event Planners",
      "131023"= "Purchasing Agents, Except Wholesale, Retail, And Farm Products",
      "514033"= "Machine Tool Setters",
      "291291" = "Acupuncturists",
      "291292"="Dental Hygienists",
      "211015"= "Rehabilitation Counselors",
      "515111"= "Prepress Technicians",
      "292031"= "Cardiovascular Technologists",
      "432099"= "Communications Equipment Operators",
      "537051"= "Industrial Truck And Tractor Operators",
      "119161"= "Emergency Management Directors",
      "433011"= "Bill And Account Collectors",
      "291127"= "Speech-Language Pathologists",
      "291124"= "Radiation Therapists",
      "537063"= "Machine Feeders",
      "291131"= "Veterinarians",
      "359099"= "Food Preparation",
      "513092"= "Food Batchmakers",
      "499060"= "Precision Instrument And Equipment Repairers",
      "411012"= "First-Line Supervisors Of Non-Retail Sales Workers",
      "533053"= "Shuttle Drivers and Chauffeurs",
      "292090"= "Health Technologists And Technicians",
      "119081"= "Lodging Managers",
      "132070"= "Credit Counselors And Loan Officers",
      "132081"= "Tax Examiners And Collectors, And Revenue Agents",
      "212099"= "Religious Workers",
      "452041"= "Graders Agricultural Products",
      "292043"= "Paramedics",
      "292056"= "Veterinary Technologists And Technicians",
      "259040"= "Teaching Assistants",
      "132041"= "Credit Analysts",
      "132052"= "Personal Financial Advisors",
      "194010"= "Agricultural And Food Science Technicians",
      "273023"= "News Journalists",
      "439051"= "Mail Clerks ",
      "172121"= "Marine Engineers And Naval Architects",
      "519151"= "Photographic Process Workers ",
      "272042"= "Musicians And Singers",
      "113111"= "Compensation And Benefits Managers",
      "193034"= "School Psychologists",
      "472050"= "Cement Masons, Concrete Finishers, And Terrazzo Workers",
      "518090"= "Plant And System Operators",
      "193011"= "Economists",
      "472040"= "Carpet, Floor, And Tile Installers",
      "311131"= "Nursing Assistants",
      "435071"= "Shipping, Receiving, And Inventory Clerks",
      "492091"= "Avionics Technicians",
      "17301X"= "Drafters",
      "291210"= "Physicians",
      "333021"= "Detectives And Criminal Investigators",
      "273099"= "Media And Communication Workers",
      "472111"= "Electricians",
      "434181"= "Reservation And Transportation Ticket Agents And Travel Clerks",
      "232011"= "Paralegals And Legal Assistants",
      "472211"= "Sheet Metal Workers",
      "435052"= "Postal Service Mail Carriers",
      "413021"= "Insurance Sales Agents",
      "436012"= "Legal Secretaries And Administrative Assistants",
      "413031"= "Securities, Commodities, And Financial Services Sales Agents",
      "436014"= "Secretaries And Administrative Assistants, Except Legal, Medical, And Executive",
      "33909X"= "Other Protective Service Workers",
      "2310XX"= "Lawyers, And Judges, Magistrates, And Other Judicial Workers",
      "433051"= "Payroll And Timekeeping Clerks",
      "339093"= "Transportation Security Screeners",
      "395012"= "Hairdressers",
      "131131"= "Fundraisers",
      "291041"= "Optometrists",
      "331011"= "First-Line Supervisors Of Correctional Officers",
      "211023"= "Mental Health And Substance Abuse Social Workers",
      "291299"= "Healthcare Diagnosing Or Treating Practitioners, All Other",
      "372021"= "Pest Control Workers",
      "514031"= "Press Machine Setters",
      "5340XX"= "Rail Transportation Workers",
      "536021"= "Parking Attendants",
      "434161"= "Human Resources Assistants, Except Payroll And Timekeeping",
      "131151"= "Training And Development Specialists",
      "515112"= "Printing Press Operators",
      "292032"= "Diagnostic Medical Sonographers",
      "291181"= "Audiologists",
      "434031"= "Court, Municipal, And License Clerks",
      "499098"= " Repair Workers",
      "499091"= "Vending Machine Servicers ",
      "291123"= "Physical Therapists",
      "291122"= "Occupational Therapists",
      "537064"= "Packers - Hand",
      "3930XX"= "Entertainment Attendants",
      "119141"= "Real Estate  Managers",
      "4520XX"= "Agricultural Workers",
      "271025"= "Interior Designers",
      "395094"= "Skincare Specialists",
      "454011"= "Forest And Conservation Workers",
      "434YYY"= "Records Clerks",
      "211092"= "Probation Officers And Correctional Treatment Specialists",
      "N"= "N/A ",
      "514XXX"= "Metal And Plastic Workers",
      "533099"= "Motor Vehicle Operators",
      "292055"= "Surgical Technologists",
      "292052"= "Pharmacy Technicians",
      "272022"= "Coaches And Scouts",
      "272023"= "Referees; Sports Officials",
      "453031"= "Fishing And Hunting Workers",
      "39509X"= "Personal Appearance Workers",
      "513011"= "Bakers",
      "475032"= "Explosives Workers",
      "512041"= "Structural Metal Fabricators And Fitters",
      "292081"= "Opticians",
      "131081"= "Logisticians",
      "132051"= "Financial And Investment Analysts",
      "273042"= "Technical Writers",
      "439021"= "Data Entry Keyers",
      "519160"= "Computer Numerically Controlled Tool Operators And Programmers",
      "399032"= "Recreation Workers",
      "519030"= "Cutting Workers",
      "172110"= "Industrial Engineers",
      "252030"= "Secondary School Teachers",
      "273092"= "Court Reporters",
      "3330XX"= "Parking Enforcement Officers",
      "399099"= "Personal Care And Service Workers",
      "192030"= "Chemists And Materials Scientists",
      "532031"= "Flight Attendants",
      "519196"= "Paper Goods Machine Operators",
      "519191"= "Adhesive Bonding Machine Operators",
      "192021"= "Atmospheric And Space Scientists",
      "472130"= "Insulation Workers",
      "254010"= "Archivists, Curators, And Museum Technicians",
      "552010"= "First-Line Enlisted Military Supervisors",
      "17302X"= "Engineering Technologists And Technicians",
      "1721XX"= "Mining And Geological Engineers",
      "319091"= "Dental Assistants",
      "151244"= "Network And Computer Systems Administrators",
      "151241"= "Computer Network Architects",
      "173031"= "Surveying And Mapping Technicians",
      "414010"= "Sales Representatives, Wholesale And Manufacturing",
      "31909X"= "Healthcare Support Workers",
      "172061"= "Computer Hardware Engineers",
      "516021"= "Pressers, Textile, Garment, And Related Materials",
      "492092"= "Electric Motor, Power Tool, And Related Repairers",
      "151254"= "Web Developers",
      "517011"= "Cabinetmakers And Bench Carpenters",
      "516040"= "Shoe And Leather Workers",
      "1721YY"= "Engineers",
      "493023"= "Automotive Service Technicians And Mechanics",
      "434061"= "Eligibility Interviewers, Government Programs",
      "412031"= "Retail Salespersons",
      "393031"= "Ushers, Lobby Attendants, And Ticket Takers",
      "531000"= "Supervisors Of Transportation And Material Moving Workers",
      "413011"= "Advertising Sales Agents",
      "435051"= "Postal Service Clerks",
      "435053"= "Postal Service Mail Sorters",
      "192099"= "Physical Scientists, All Other",
      "359021"= "Dishwashers",
      "436013"= "Medical Secretaries And Administrative Assistants",
      "436011"= "Executive Secretaries And Executive Administrative Assistants",
      "331012"= "First-Line Supervisors Of Police And Detectives",
      "435111"= "Weighers, Measurers, Checkers, And Samplers, Recordkeeping",
      "434141"= "New Accounts Clerks",
      "537021"= "Crane And Tower Operators",
      "472XXX"= "Brickmasons",
      "339091"= "Crossing Guards And Flaggers",
      "211021"= "Child, Family, And School Social Workers",
      "435011"= "Cargo And Freight Agents",
      "2911XX"= "Nurse Practitioners, And Nurse Midwives",
      "433061"= "Procurement Clerks",
      "1520XX"= "Other Mathematical Science Occupations",
      "131161"= "Market Research Analysts And Marketing Specialists",
      "499096"= "Riggers",
      "212011"= "Clergy",
      "119041"= "Architectural And Engineering Managers",
      "212021"= "Directors, Religious Activities And Education",
      "232093"= "Title Examiners, Abstractors, And Searchers",
      "514041"= "Machinists",
      "351012"= "First-Line Supervisors Of Food Preparation And Serving Workers",
      "411011"= "First-Line Supervisors Of Retail Sales Workers",
      "499052"= "Telecommunications Line Installers And Repairers",
      "4750XX"= "Other Extraction Workers",
      "271022"= "Fashion Designers",
      "271026"= "Merchandise Displayers And Window Trimmers",
      "271010"= "Artists And Related Workers",
      "352010"= "Cooks",
      "419041"= "Telemarketers",
      "499021"= "Heating, Air Conditioning, And Refrigeration Mechanics And Installers",
      "432011"= "Switchboard Operators, Including Answering Service",
      "131051"= "Cost Estimators",
      "512020"= "Electrical, Electronics, And Electromechanical Assemblers",
      "439061"= "Office Clerks, General",
      "292061"= "Licensed Practical And Licensed Vocational Nurses",
      "431011"= "First-Line Supervisors Of Office And Administrative Support Workers",
      "474061"= "Rail-Track Laying And Maintenance Equipment Operators",
      "353041"= "Food Servers, Nonrestaurant",
      "132053"= "Insurance Underwriters",
      "171011"= "Architects, Except Landscape And Naval",
      "194021"= "Biological Technicians",
      "2590XX"= "Educational Instruction and Library Workers",
      "452011"= "Agricultural Inspectors",
      "472080"= "Drywall Installers, Ceiling Tile Installers, And Tapers",
      "439111"= "Statistical Assistants",
      "519061"= "Inspectors, Testers, Sorters, Samplers, And Weighers",
      "532020"= "Air Traffic Controllers And Airfield Operations Specialists",
      "519080"= "Dental And Ophthalmic Laboratory Technicians ",
      "519071"= "Jewelers ",
      "519194"= "Etchers And Engravers",
      "113031"= "Financial Managers",
      "319097"= "Phlebotomists",
      "472140"= "Painters ",
      "311122"= "Personal Care Aides",
      "333011"= "Bailiffs",
      "333012"= "Correctional Officers And Jailers",
      "1940XX"= "Environmental Science And Geoscience Technicians, And Nuclear Technicians",
      "517041"= "Sawing Machine Setters, Operators, And Tenders, Wood",
      "493011"= "Aircraft Mechanics And Service Technicians",
      "516093"= "Upholsterers",
      "518031"= "Water And Wastewater Treatment Plant And System Operators",
      "332011"= "Firefighters",
      "359011"= " Cafeteria Attendants And Bartender Helpers",
      "194031"= "Chemical Technicians",
      "393010"= "Gambling Services Workers",
      "519111"= "Packaging And Filling Machine Operators And Tenders",
      "435041"= "Meter Readers",
      "394031"= "Morticians, Undertakers, And Funeral Arrangers",
      "331021"= "First-Line Supervisors Of Firefighting And Prevention Workers",
      "119121"= "Natural Sciences Managers",
      "339094"= "School Bus Monitors",
      "211012"= "Educational, Guidance, And Career Counselors And Advisors",
      "211022"= "Healthcare Social Workers",
      "433071"= "Tellers",
      "131030"= "Claims Adjusters, Appraisers, Examiners, And Investigators",
      "211019"= "Counselors, All Other",
      "5120XX"= "Other Assemblers and Fabricators",
      "491011"= "First-Line Supervisors Of Mechanics, Installers, And Repairers",
      "152011"= "Actuaries",
      "4990XX"= "Other Installation, Maintenance, And Repair Workers",
      "232099"= "Legal Support Workers, All Other",
      "291126"= "Respiratory Therapists",
      "151299"= "Computer Occupations",
      "29112X"= "Therapists",
      "291011"= "Chiropractors",
      "434121"= "Library Assistants",
      "339030"= "Security Guards",
      "291020"= "Dentists",
      "513093"= "Food Cooking Machine Operators And Tenders",
      "434111"= "Interviewers, Except Eligibility And Loan",
      "419020"= "Real Estate Brokers And Sales Agents",
      "534031"= "Railroad Conductors And Yardmasters",
      "352021"= "Food Preparation Workers",
      "191010"= "Agricultural And Food Scientists",
      "112011"= "Advertising And Promotions Managers",
      "4740XX"= "Other Construction And Related Workers",
      "111021"= "General And Operations Managers",
      "535020"= "Ship And Boat Captains And Operators",
      "474031"= "Fence Erectors",
      "113021"= "Computer And Information Systems Managers",
      "474041"= "Hazardous Materials Removal Workers",
      "559830"= "Military, Rank Not Specified",
      "273031"= "Public Relations Specialists",
      "474011"= "Construction And Building Inspectors",
      "19204X"= "Geoscientists And Hydrologists, Except Geographers",
      "399031"= "Exercise Trainers And Group Fitness Instructors",
      "251000"= "Postsecondary Teachers",
      "399041"= "Residential Advisors",
      "312020"= "Physical Therapist Assistants And Aides",
      "472151"= "Pipelayers",
      "519198"= "Helpers--Production Workers",
      "253041"= "Tutors",
      "254022"= "Librarians And Media Collections Specialists",
      "151255"= "Web And Digital Interface Designers",
      "252050"= "Special Education Teachers",
      "471011"= "First-Line Supervisors Of Construction Trades And Extraction Workers",
      "231012"= "Judicial Law Clerks",
      "299000"= "Other Healthcare Practitioners And Technical Occupations",
      "553010"= "Military Enlisted Tactical Operations And Air/Weapons Specialists And Crew Members",
      "2530XX"= "Other Teachers And Instructors",
      "151212"= "Information Security Analysts",
      "151221"= "Computer And Information Research Scientists",
      "173011"= "Architectural And Civil Drafters",
      "319011"= "Massage Therapists",
      "492011"= "Computer, Automated Teller, And Office Machine Repairers",
      "5350XX"= "Sailors And Marine Oilers, And Ship Engineers",
      "131011"= "Agents And Business Managers Of Artists, Performers, And Athletes",
      "395011"= "Barbers",
      "37201X"= "Janitors And Building Cleaners",
      "412010"= "Cashiers",
      "515113"= "Print Binding And Finishing Workers",
      "412022"= "Parts Salespersons",
      "5360XX"= "Other Transportation Workers",
      "373011"= "Landscaping And Groundskeeping Workers",
      "291125"= "Recreational Therapists",
      "119151"= "Social And Community Service Managers",
      "152031"= "Operations Research Analysts",
      "131111"= "Management Analysts",
      "513091"= "Food And Tobacco Roasting, Baking, And Drying Machine Operators And Tenders",
      "396010"= "aggage Porters, Bellhops, And Concierges",
      "533054"= "Taxi Drivers",
      "191020"= "Biological Scientists",
      "271023"= "Floral Designers",
      "511011"= "First-Line Supervisors Of Production And Operating Workers",
      "499043"= "Maintenance Workers, Machinery",
      "211093"= "Social And Human Service Assistants",
      "499044"= "Millwrights",
      "419031"= "Sales Engineers",
      "21109X"= "Other Community And Social Service Specialists",
      "132082"= "Tax Preparers",
      "534010"= "Locomotive Engineers And Operators",
      "499010"= "Control And Valve Installers And Repairers",
      "512031"= "Engine And Other Machine Assemblers",
      "291071"= "Physician Assistants",
      "399011"= "Childcare Workers",
      "292053"= "Psychiatric Technicians",
      "514111"= "Tool And Die Makers",
      "513020"= "Butchers And Other Meat, Poultry, And Fish Processing Workers",
      "292072"= "Medical Records Specialists",
      "131082"= "Project Management Specialists",
      "419010"= "Models, Demonstrators, And Product Promoters",
      "1110XX"= "Chief Executives And Legislators",
      "171012"= "Landscape Architects",
      "439022"= "Word Processors And Typists",
      "472070"= "Construction Equipment Operators",
      "112030"= "Public Relations And Fundraising Managers",
      "112022"= "Sales Managers",
      "391000"= "Supervisors of Personal Care And Service Workers",
      "273091"= "Interpreters And Translators",
      "192010"= "Astronomers And Physicists",
      "472161"= "Plasterers And Stucco Masons",
      "312010"= "Occupational Therapy Assistants And Aides",
      "519195"= "Molders, Shapers, And Casters, Except Metal And Plastic",
      "252010"= "Preschool And Kindergarten Teachers",
      "333050"= "Police Officers",
      "516031"= "Sewing Machine Operators",
      "492098"= "Security And Fire Alarm Systems Installers",
      "319092"= "Medical Assistants",
      "1940YY"= "Other Life, Physical, And Social Science Technicians",
      "392021"= "Animal Caretakers",
      "331090"= "Miscellaneous First-Line Supervisors, Protective Service Workers",
      "472011"= "Boilermakers",
      "517021"= "Furniture Finishers",
      "537070"= "Pumping Station Operators",
      "472221"= "Structural Iron And Steel Workers",
      "493040"= "Heavy Vehicle And Mobile Equipment Service Technicians And Mechanics",
      "537081"= "Refuse And Recyclable Material Collectors",
      "5370XX"= "Conveyor, Dredge, And Hoist And Winch Operators",
      "113071"= "Transportation, Storage, And Distribution Managers",
      "493031"= "Bus And Truck Mechanics And Diesel Engine Specialists",
      "517042"= "Woodworking Machine Setters, Operators, And Tenders, Except Sawing",
      "493021"= "Automotive Body And Related Repairers",
      "435032"= "Dispatchers, Except Police, Fire, And Ambulance",
      "516060"= "Textile Machine Setters, Operators, And Tenders",
      "151211"= "Computer Systems Analysts",
      "371011"= "First-Line Supervisors Of Housekeeping And Janitorial Workers",
      "254031"= "Library Technicians",
      "519120"= "Painting Workers",
      "172041"= "Chemical Engineers",
      "195010"= "Occupational Health And Safety Specialists And Technicians",
      "211011"= "Substance Abuse And Behavioral Disorder Counselors",
      "292010"= "Clinical Laboratory Technologists And Technicians",
      "536051"= "Transportation Inspectors",
      "434131"= "Loan Interviewers And Clerks",
      "131021"= "Buyers And Purchasing Agents, Farm Products",
      "211014"= "Mental Health Counselors",
      "372012"= "Maids And Housekeeping Cleaners",
      "514020"= "Forming Machine Setters, Operators, And Tenders, Metal And Plastic",
      "131041"= "Compliance Officers",
      "373013"= "Tree Trimmers And Pruners",
      "5371XX"= "Other Material Moving Workers",
      "51403X"= "Other Machine Tool Setters, Operators, And Tenders, Metal And Plastic",
      "291240"= "Surgeons",
      "537061"= "Cleaners Of Vehicles And Equipment",
      "499071"= "Maintenance And Repair Workers, General",
      "291141"= "Registered Nurses",
      "37301X"= "Other Grounds Maintenance Workers",
      "433021"= "Billing And Posting Clerks",
      "49904X"= "Industrial And Refractory Machinery Mechanics"
)
```

```
schl_labels <-c(
  "03"= "Kindergarten",
      "16"= "High School",
      "01"= "No schooling completed",
      "07"= "Grade 4",
      "04"= "Grade 1",
      "23"= "Professional degree",
      "22"= "Master's degree",
      "19"= "Some college",
      "10"= "Grade 7",
      "02"= "Preschool",
      "0" = "N/A",
      "20"= "Associate's degree",
      "21"= "Bachelor's degree",
      "08"= "Grade 5",
      "06"= "Grade 3",
      "14"= "Grade 11",
      "24"= "Doctorate degree",
      "12"= "Grade 9",
      "17"= "GED",
      "09"= "Grade 6",
      "11"= "Grade 8",
      "18"= "Some college",
      "15"= "Grade 12 - No Diploma",
      "05"= "Grade 2",
      "13"= "Grade 10"
)
```

```
socp_labels <- c(
  "291080" = "Healthcare Professional",
  "291051" = "Healthcare Professional",
  "291151" = "Healthcare Professional",
  "291291" = "Healthcare Professional",
  "291292" = "Healthcare Professional",
  "292034" =  "Healthcare Professional",
  "292035" =  "Healthcare Professional",
  "292032" = "Healthcare Professional",
  "291181" = "Healthcare Professional",
  "291125" = "Healthcare Professional",
  "291141" = "Healthcare Professional",
  "291240" =  "Healthcare Professional",
  "172070" = "Engineering and Technical Specialists",
  "173023" = "Engineering and Technical Specialists",
  "172051" =  "Engineering and Technical Specialists",
  "151252"=  "Engineering and Technical Specialists",
  "151251" =  "Engineering and Technical Specialists",
  "172081" =  "Engineering and Technical Specialists",
  "172141" =  "Engineering and Technical Specialists",
  "172011" = "Engineering and Technical Specialists",
  "291210" = "Engineering and Technical Specialists",
  "17302X" = "Engineering and Technical Specialists",
  "1721XX" = "Engineering and Technical Specialists",
  "17206"  = "Engineering and Technical Specialists",
  "1721YY" = "Engineering and Technical Specialists",
  "172041" = "Engineering and Technical Specialists",
  "113013" =  "Managers and Administrators",
  "113131" =  "Managers and Administrators",
  "112021" =  "Managers and Administrators",
  "119111 " = "Managers and Administrators",
  "113012" = "Managers and Administrators",
  "193051" = "Managers and Administrators",
  "113121" = "Managers and Administrators",
  "113061" = "Managers and Administrators",
  "113031" =  "Managers and Administrators",
  "119041" = "Managers and Administrators",
  "113021" = "Managers and Administrators",
  "113071" = "Managers and Administrators",
  "472061" =  "Construction and Manual Labor",
  "472152" = "Construction and Manual Labor",
  "472181" = "Construction and Manual Labor",
  "472031" = "Construction and Manual Labor",
  "472121" = "Construction and Manual Labor",
  "472111" = "Construction and Manual Labor",
  "472211" = "Construction and Manual Labor",
  "472221" = "Construction and Manual Labor",
  "472011" =  "Construction and Manual Labor",
 "15124X" = "Computer and IT Specialists",
 "151253" = "Computer and IT Specialists",
  "151244" = "Computer and IT Specialists",
 "151241" = "Computer and IT Specialists",
 "151255" = "Computer and IT Specialists",
 "151212" =  "Computer and IT Specialists",
 "151221" = "Computer and IT Specialists",
 "151211" = "Computer and IT Specialists",
 "252020" =  "Educational Professionals",
 "259040" = "Educational Professionals",
 "252050" = "Educational Professionals",
 "252010" = "Educational Professionals",
 "251000" = "Educational Professionals",
"1320XX" = "Finance and Business Specialists" ,
"132061" = "Finance and Business Specialists" ,
"132011" = "Finance and Business Specialists" ,
"132020" = "Finance and Business Specialists" ,
"132031" = "Finance and Business Specialists" ,
"132053" =  "Finance and Business Specialists" ,
"131021" =  "Finance and Business Specialists" ,
"131111" = "Finance and Business Specialists" ,
"131030" = "Finance and Business Specialists" ,
"131041" = "Finance and Business Specialists" ,
"353023" = "Food Service Workers",
"351011" = "Food Service Workers",
"513011" = "Food Service Workers",
"513091" = "Food Service Workers",
"513092" = "Food Service Workers",
"513093" = "Food Service Workers",
"359011" = "Food Service Workers",
"352010" = "Food Service Workers",
"352021" = "Food Service Workers",
 "2310XX" = "Law and Legal Professionals",
"232011" = "Law and Legal Professionals",
"232093" =  "Law and Legal Professionals",
  "273043" = "Media and Communication",
  "273041" = "Media and Communication",
"273011" = "Media and Communication",
"273023" = "Media and Communication",
"273092" = "Media and Communication",
"273031" = "Media and Communication",
"273099" = "Media and Communication",
"333021" =  "Public Safety and Security",
"339091" = "Public Safety and Security",
"332011" = "Public Safety and Security",
"331011" = "Public Safety and Security",
"331012" = "Public Safety and Security",
"331021" = "Public Safety and Security",
"339093" = "Public Safety and Security",
"339094" = "Public Safety and Security",
"339030" = "Public Safety and Security",
"333050" =  "Public Safety and Security",
  "434051" =  "Retail and Customer Service",
"434171" = "Retail and Customer Service",
"434081" =  "Retail and Customer Service",
"359031" =  "Retail and Customer Service",
"412021" = "Retail and Customer Service",
"412031" = "Retail and Customer Service",
"412022" = "Retail and Customer Service",
"412010" =  "Retail and Customer Service",
"192041" =  "Science and Research",
"1910XX" = "Science and Research",
"194021" = "Science and Research",
"194031" = "Science and Research",
"1940XX" = "Science and Research",
"1940YY" = "Science and Research",
"192030" = "Science and Research",
"192021" = "Science and Research",
"19204X" =  "Science and Research",
"192099" = "Science and Research",
"192010" = "Science and Research",
 "191010" = "Science and Research",
  "536030" = "Transport and Logistics",
"533051" = "Transport and Logistics",
"533052" = "Transport and Logistics",
"533030" = "Transport and Logistics",
"533053" = "Transport and Logistics",
"533054" = "Transport and Logistics",
"532010" = "Transport and Logistics",
"532031" = "Transport and Logistics",
"532020" = "Transport and Logistics",
  "537061" = "Transport and Logistics",
"537070" = "Transport and Logistics",
"537081" = "Transport and Logistics",
  "999920" = "Unemployed",
"N" =  "Miscellaneous",
"4990XX" = "Miscellaneous",
"499098" = "Miscellaneous",
"499091" = "Miscellaneous",
  "499043" = "Miscellaneous",
"499044" = "Miscellaneous",
"49904X" = "Miscellaneous",
"499010" =  "Miscellaneous",
"499051" = "Miscellaneous",
"499052" =  "Miscellaneous",
"499060" = "Miscellaneous",
"499071" =  "Miscellaneous",
"499021" =  "Miscellaneous",
"499031" = "Miscellaneous")
```

```
combined_data <- rbind(de_pums, de_pums2)
combined_data2 <- combined_data %>%
  mutate(SCHL = as.character(SCHL),
         SCHL = schl_labels[SCHL] ) %>%
  rename(Educational_Attainment = SCHL) %>%
  mutate(FOD1P = as.character(FOD1P),
         FOD1P = fod1p_labels[FOD1P]) %>%
  rename(Field_Of_Degree  = FOD1P) %>%
  mutate(SOCP = as.character(SOCP),
         SOCP = socp_labels[SOCP]) %>%
  rename(Occupation = SOCP) %>%
  mutate(SEX = as.character(SEX),
         SEX = case_when(
           SEX == "1" ~ "Male",
           SEX == "2" ~ "Female",
           TRUE ~ NA_character_
         ))%>%
  rename(Age = AGEP) %>%
  mutate(RAC1P = as.character(RAC1P),
         RAC1P = case_when(
           RAC1P == "3" ~ "Native American",
           RAC1P == "1" ~ "White",
           RAC1P == "8" ~ "Other",
           RAC1P == "6" ~ "Asian",
           RAC1P == "9" ~ "Two or More",
           RAC1P == "2" ~ "African American",
           RAC1P == "4" ~ "Asian",
           RAC1P == "7" ~ "Asian",
           RAC1P == "5" ~ "Native American",
           TRUE ~ NA_character_
         ))%>%
  mutate(SCIENGP = as.character(SCIENGP),
         SCIENGP = case_when(
           SCIENGP == "0" ~ "N/A (less than a bachelor's degree)",
           SCIENGP == "2" ~ "No",
           SCIENGP == "1" ~ "Yes",
           TRUE ~ NA_character_
         )) %>%
  rename(Field_Of_Degree_STEM = SCIENGP)%>%
  rename(Income = PINCP) %>%

 select(-SERIALNO) %>%
  select(-SPORDER) %>%
  select(-WGTP)%>%
  select(-PWGTP) %>%
  select(-ST) %>%
  rename(Race = RAC1P) %>%
  mutate(COW = as.character(COW),
         COW = case_when(
           COW == "9" ~ "Unemployed",
           COW == "1" ~ "Private for-profit company",
           COW == "7" ~ "Self-employed",
           COW == "8" ~ "Unpaid Labor",
           COW == "3" ~ "Local government",
           COW == "5" ~ "Federal government",
           COW == "4" ~ "State government",
           COW == "6" ~ "Self-employed",
           COW == "2" ~ "Non-Profit",
           COW == "0" ~ "Unemployed",
           TRUE ~ NA_character_
         ))%>%
  rename(Class_Of_Worker = COW) %>%
  drop_na()


combined_data2
```

```
## # A tibble: 118 x 9
##    Income   Age Class_Of_Worker     Educational_Attainment SEX   Field_Of_Degree
##     <dbl> <dbl> <chr>               <chr>                  <chr> <chr>
##  1  62600    33 Local government    Bachelor's degree      Male  Mathematics
##  2  60000    56 Private for-profit~ Master's degree        Fema~ Mathematics
##  3 129400    69 State government    Doctorate degree       Male  Mathematics
##  4 128100    33 Federal government  Doctorate degree       Fema~ Mathematics
##  5  85000    29 Private for-profit~ Bachelor's degree      Male  Computer Scien~
##  6  75000    40 Private for-profit~ Bachelor's degree      Male  Mathematics
##  7 100700    75 Self-employed       Professional degree    Male  Mathematics
##  8 123200    75 Self-employed       Doctorate degree       Male  Mathematics
##  9 150000    42 Federal government  Bachelor's degree      Fema~ Computer Scien~
## 10  24000    68 Private for-profit~ Bachelor's degree      Fema~ Mathematics
## # i 108 more rows
## # i 3 more variables: Race <chr>, Field_Of_Degree_STEM <chr>, Occupation <chr>
```

I started with combining both de_pums data sets creating combinedf_data2. I had to rename all of the variable names in this data set. I recoded FOD1P to be Field_of_Degree, COW to Employing_Organization, PINCP to Income, SCIENGP to FOD_science_and_engineering, SCHL to Educational_Attainment, AGEP to Age, SOCP to Occupation, RAC1P as Race. After getting the SOCP labels organized I was then able to recode the SOCP labels (Standard Occupational Classification codes) to a more simplified version of: “Healthcare Professional”, “Engineering and Technical Specialists”, “Managers and Administrators”, “Construction and Manual Labor”, “Computer and IT Specialists”, “Educational Professionals”, “Finance and Business Specialists”, “Food Service Workers”, “Law and Legal Professionals’ ’,”Media and Communication”, “Public Safety and Security”, “Retail and Customer Service”, “Science and Research”, “Transport and Logistics”, “Unemployed”, “Miscellaneous”. As well as the SCHL labels to be associated with the correct labeling. I also went through and recoded the FOD1P labels so they only included the two I am interested in: Mathematics and Computer Science. I also made the decision to group some of the Race observations after the dataset was loaded in. I made this decision because their were like groups of Races. I had to drop variables that were automatically inputed when the dataset loaded in, these variales would have been useless in my research which is why I decided to drop them. I used mutate to create new variables. I used this on Class of worker, Race, Occupation labels, Field of Degree labels, Sex labels, and eudcational attainment labels.

Exploratory Data Analysis

Missing Values

The two variables I will look at is Race and Income. Income is numeric data. It is the total person’s income received on an annual basis. The possible values are any non-negative number (as one would think), usually representing the annual income of the individual. It can be 0 (no income) to any large value (a high income). The values seem to be whole numbers in this dataset. Race is categorical data. It represents the Race of the individual in the data set.The possible value are White, Other, Native American, Asian, African American and Two or More

```
missing_table_summary <- miss_var_summary(combined_data2)
print(missing_table_summary)
```

```
## # A tibble: 9 x 3
##   variable               n_miss pct_miss
##   <chr>                   <int>    <dbl>
## 1 Income                      0        0
## 2 Age                         0        0
## 3 Class_Of_Worker             0        0
## 4 Educational_Attainment      0        0
## 5 SEX                         0        0
## 6 Field_Of_Degree             0        0
## 7 Race                        0        0
## 8 Field_Of_Degree_STEM        0         0
## 9 Occupation                  0         0
```

It seems that there is no missing values in this data set.

```
missing_summary <- combined_data2 %>% summarise_all(funs(sum(is.na(.))))
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## i Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
print(missing_summary)
```

```
## # A tibble: 1 x 9
##   Income   Age Class_Of_Worker Educational_Attainment   SEX Field_Of_Degree
##    <int> <int>           <int>                  <int> <int>           <int>
## 1      0     0               0                      0     0               0
## # i 3 more variables: Race <int>, Field_Of_Degree_STEM <int>, Occupation <int>
```

There are no NA values represented in this data set.

```
missing_race <- combined_data2 %>%
  filter(is.na(Race))
print(missing_race)
```

```
## # A tibble: 0 x 9
## # i 9 variables: Income <dbl>, Age <dbl>, Class_Of_Worker <chr>,
## #   Educational_Attainment <chr>, SEX <chr>, Field_Of_Degree <chr>, Race <chr>,
## #   Field_Of_Degree_STEM <chr>, Occupation <chr>
```

```
missing_income <- combined_data2 %>%
  filter(is.na(Income))
missing_income
```

```
## # A tibble: 0 x 9
## # i 9 variables: Income <dbl>, Age <dbl>, Class_Of_Worker <chr>,
## #   Educational_Attainment <chr>, SEX <chr>, Field_Of_Degree <chr>, Race <chr>,
## #   Field_Of_Degree_STEM <chr>, Occupation <chr>
```

For the two variables there is no missing values.

Duplicates

```
unique_dupes <- combined_data2 %>%
  group_by(Race, Income) %>%
  filter(n() == 2) %>%
  ungroup()

nrow(unique_dupes) / 2
```

```
## [1] 5
```

This shows that there are 5 duplicates for the combination of Race and Income. Unique Duplicates will give you the number of unique combinations of Race and Income that are duplicated in the dataset.

```
multi_dupes <- combined_data2 %>%
  group_by(Race, Income) %>%
  filter(n() > 2) %>%
  ungroup()

multi_dupes
```

```
## # A tibble: 8 x 9
##   Income   Age Class_Of_Worker      Educational_Attainment SEX   Field_Of_Degree
##    <dbl> <dbl> <chr>                <chr>                  <chr> <chr>
## 1 100000    27 Private for-profit ~ Master's degree        Male  Computer Scien~
## 2 100000    56 Private for-profit ~ Bachelor's degree      Male  Mathematics
## 3 100000    45 Private for-profit ~ Master's degree        Male  Computer Scien~
## 4  60000    61 Private for-profit ~ Bachelor's degree      Male  Computer Scien~
## 5  60000    61 Private for-profit ~ Bachelor's degree      Male  Computer Scien~
## 6  60000    61 Private for-profit ~ Bachelor's degree      Male  Computer Scien~
## 7 100000    51 Private for-profit ~ Bachelor's degree      Male  Computer Scien~
## 8 100000    40 Private for-profit ~ Bachelor's degree      Male  Computer Scien~
## # i 3 more variables: Race <chr>, Field_Of_Degree_STEM <chr>, Occupation <chr>
```

With Multiple Duplicates you want to know the combinations that appear more than twice (i.e., they have more than one duplicate). There are 8 combinations that appear more than twice for Race and Income.

Inconsistent Values

```
inconsistent_race <- combined_data2[!combined_data2$Race %in% c("White", "Other",
             "Asian", "Two or More", "African American", "Asian",
                  "Asian", "Native American"), c("Race")]
inconsistent_income <- combined_data2[!is.numeric(as.numeric(combined_data2$Income))
                                      | as.numeric(combined_data2$Income) < 0, c("Income")]
inconsistent_race
```

```
## # A tibble: 0 x 1
## # i 1 variable: Race <chr>
```

```
inconsistent_income
```

```
## # A tibble: 1 x 1
##   Income
##    <dbl>
## 1  -2100
```

This shows that there is an inconsistant_income response of -2100. This would indicate that someone responded with a negative income, which orginally I did not think you would find. I thought the lowest would be 0. The inconsistent_race shows that there is no inconsistency with race.

Central Tendencies of Income

```
combined_data2$Income <- as.numeric(combined_data2$Income)
Q1 <- quantile(combined_data2$Income, 0.25, na.rm = TRUE)
print(Q1)
```

```
##   25%
## 41250
```

```r
Q3 <- quantile(combined_data2$Income, 0.75, na.rm = TRUE)
print(Q3)
```

```
##      75%
## 119777.5
```

```r
IQR <- Q3 - Q1
print(IQR)
```

```
##     75%
## 78527.5
```

```r
lower_bound <- Q1 - 1.5 * IQR
print(lower_bound)
```

```
##       25%
## -76541.25
```

```r
upper_bound <- Q3 + 1.5 * IQR
 print(upper_bound)
```

```
##      75%
## 237568.8
```

This shows that 25% of the respondents have an income of $41,250 or below. This shows that 75% of the respondents have an income of $119,777.50 or below. The IQR of income is $78,527.50. The lower bound of the Income is $-76,541.25. The upper bound of Income is $237,568.8

Outliers

```r
outliers_Income <- combined_data2 %>%
  filter(Income < lower_bound | Income > upper_bound)

print(outliers_Income)
```

```
## # A tibble: 3 x 9
##   Income   Age Class_Of_Worker      Educational_Attainment SEX   Field_Of_Degree
##    <dbl> <dbl> <chr>                <chr>                  <chr> <chr>
## 1 340000    63 State government     Doctorate degree       Male  Computer Scien~
## 2 325000    48 Private for-profit ~ Professional degree    Male  Mathematics
## 3 430000    62 Private for-profit ~ Bachelor's degree      Fema~ Mathematics
## # i 3 more variables: Race <chr>, Field_Of_Degree_STEM <chr>, Occupation <chr>
```

This shows all the outliers in the Income variable from less than the lower bound and greater then the upper bound There are 3 observations that represent income outliers in this data set. The values are $340,000, $325,000, and $430,000.

```r
outlier_RACE <- combined_data2 %>%
  filter(!(Race %in% c("White", "Other", "Asian", "Two or More",
                       "African American", "Asian", "Asian",
                       "Native American")))

print(outlier_RACE)
```

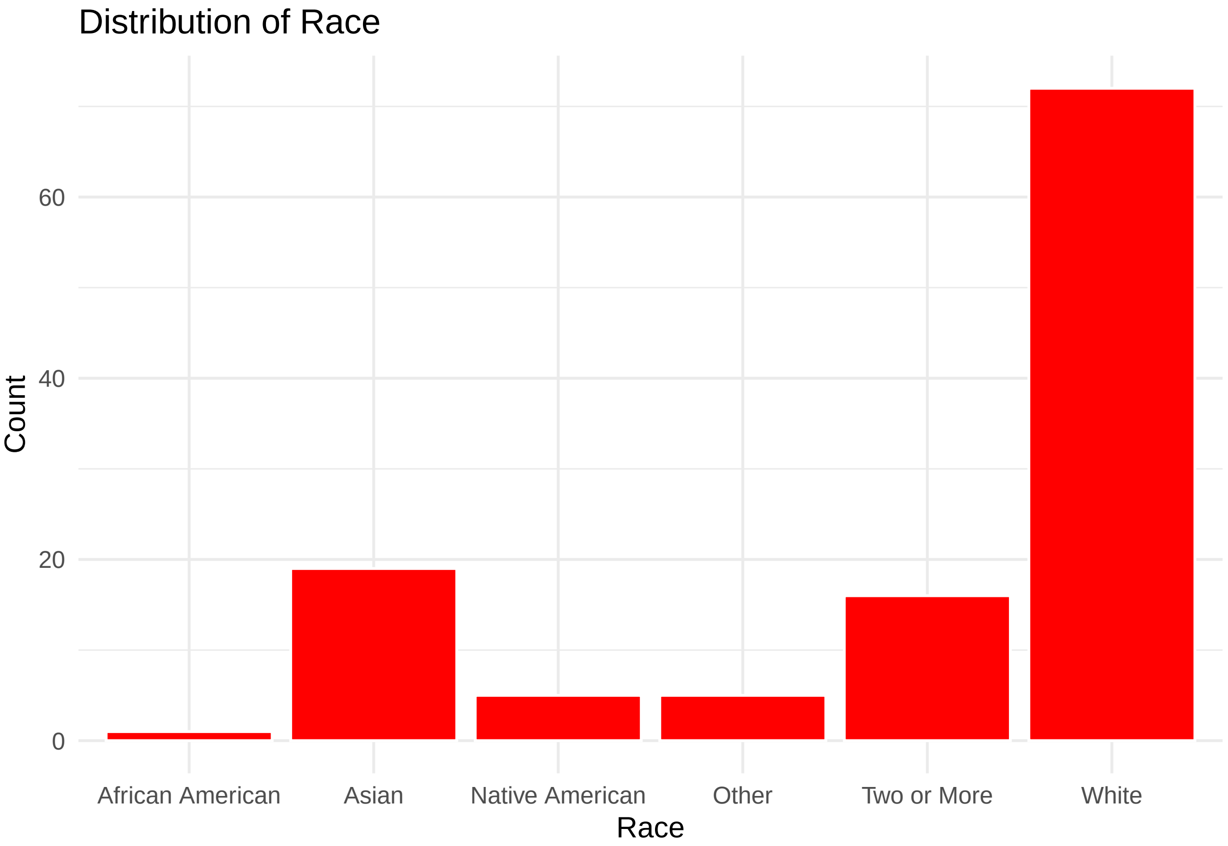
```
## # A tibble: 0 x 9
## # i 9 variables: Income <dbl>, Age <dbl>, Class_Of_Worker <chr>,
## #   Educational_Attainment <chr>, SEX <chr>, Field_Of_Degree <chr>, Race <chr>,
## #   Field_Of_Degree_STEM <chr>, Occupation <chr>
```

This shows that there are no outliers in the race variable.

Univariate Visuals

Distribution of Race

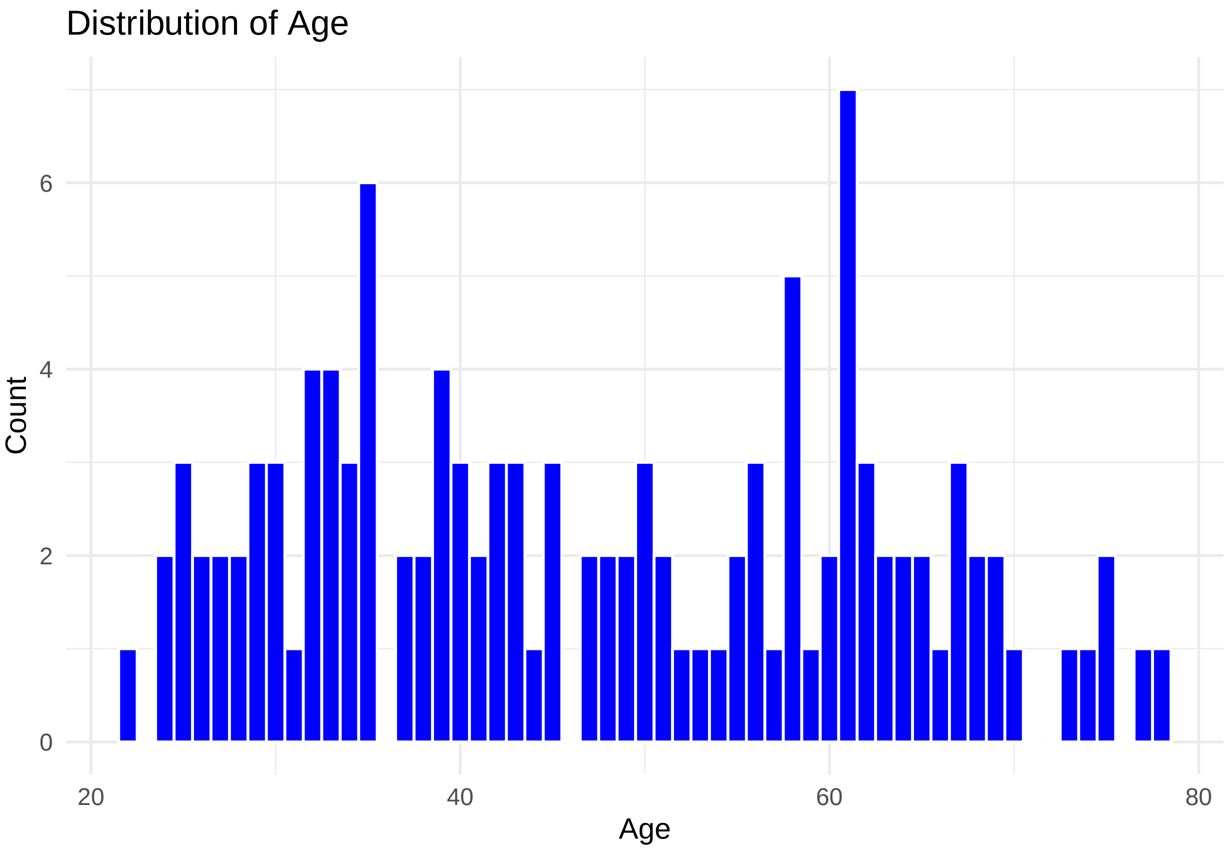
```
ggplot(combined_data2, aes(x = Race)) +
  geom_bar(fill = "red", color = "white") +
  labs(title = "Distribution of Race", x = "Race", y = "Count") +
  theme_minimal()
```



The visualization shows that the category with the highest count is ‘White’, indicating that this is the most common racial identity among the individuals represented in the data.

Distribution of Age

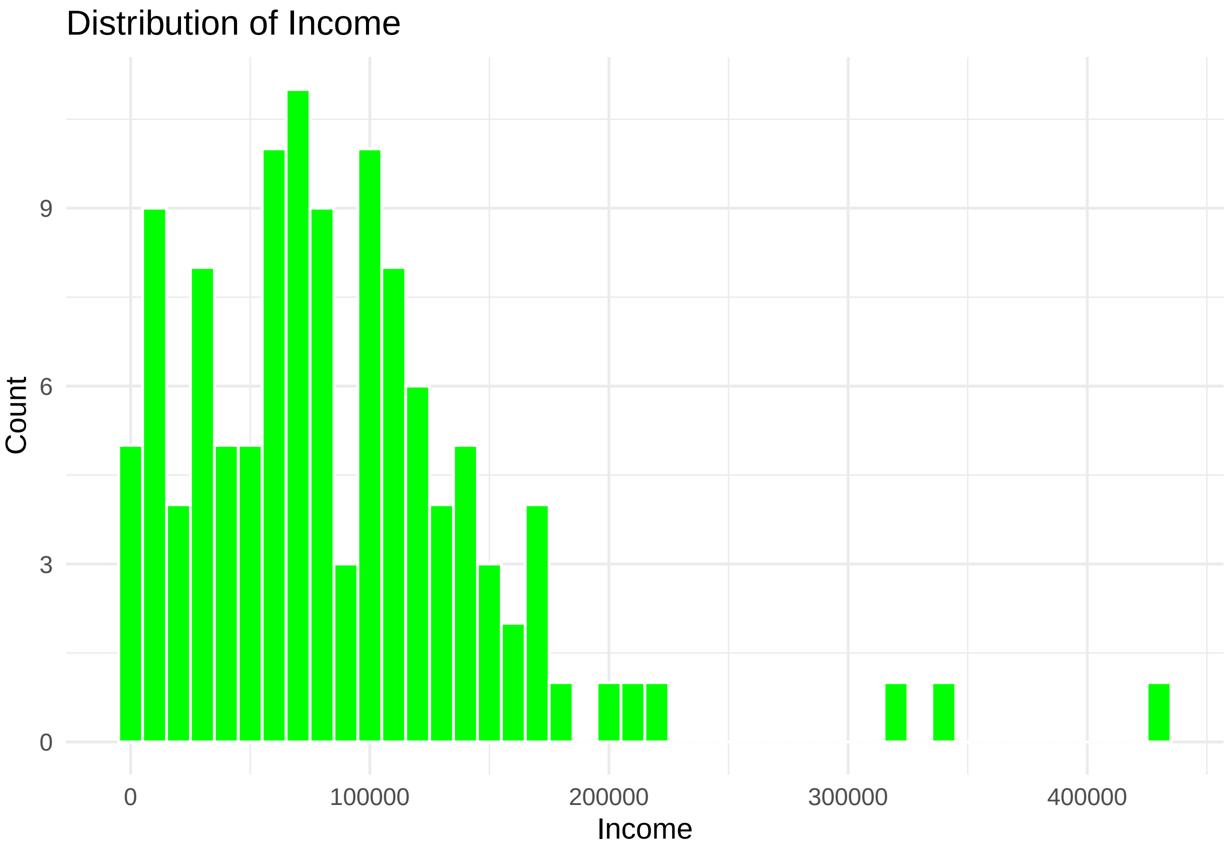
```
combined_data2 <- combined_data2 %>%
  mutate(Age = as.numeric(Age))

ggplot(combined_data2, aes(x = Age)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "white") +
  labs(title = "Distribution of Age", x = "Age", y = "Count") +
  theme_minimal()
```



The histogram of age distribution shows consistency throughout. There does not seem to be more younger or older.

Distribution of Income

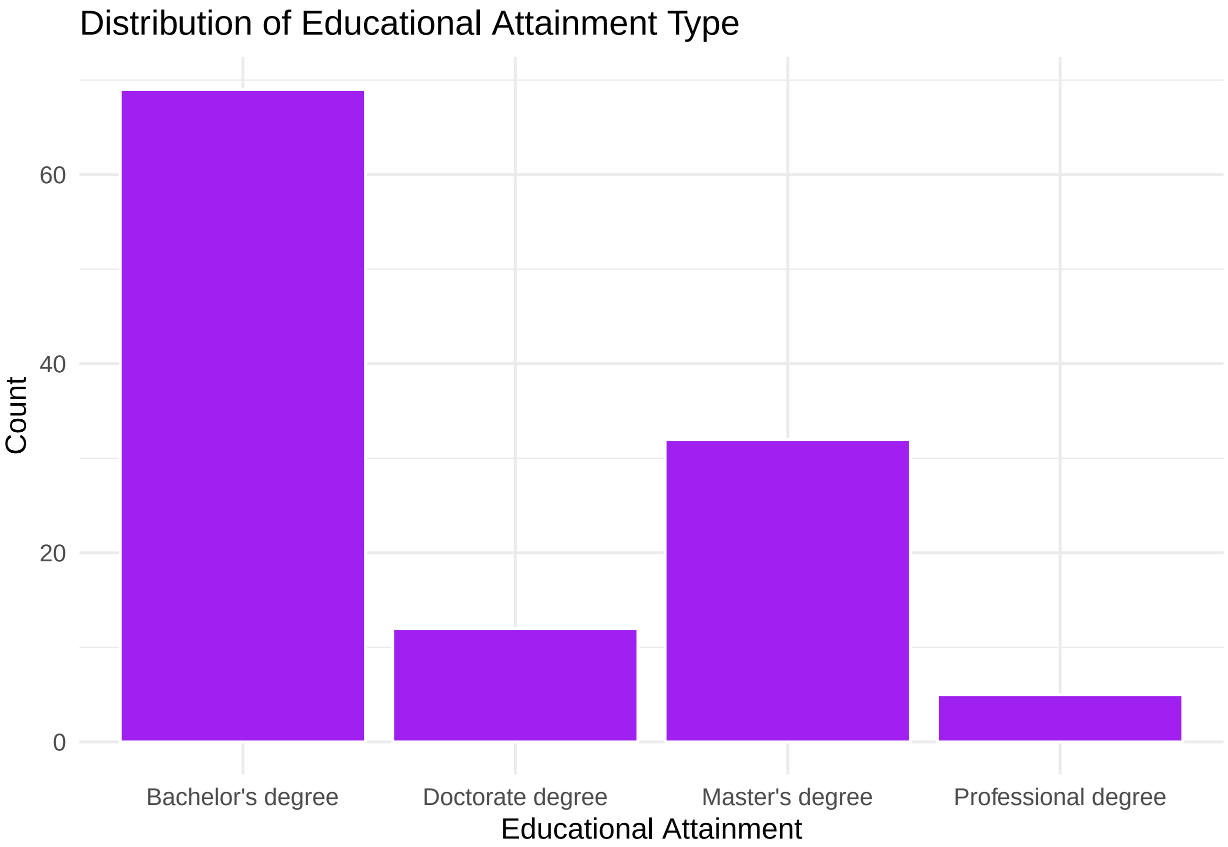
```
ggplot(combined_data2, aes(x = Income)) +
  geom_histogram(binwidth = 10000, fill = "green", color = "white") +
  labs(title = "Distribution of Income", x = "Income", y = "Count") +
  theme_minimal()
```



The income distribution histogram illustrates a highly right-skewed distribution with most counts falling in the lower income brackets and few counts extending into higher incomes.

Distribution of Educational Attainment

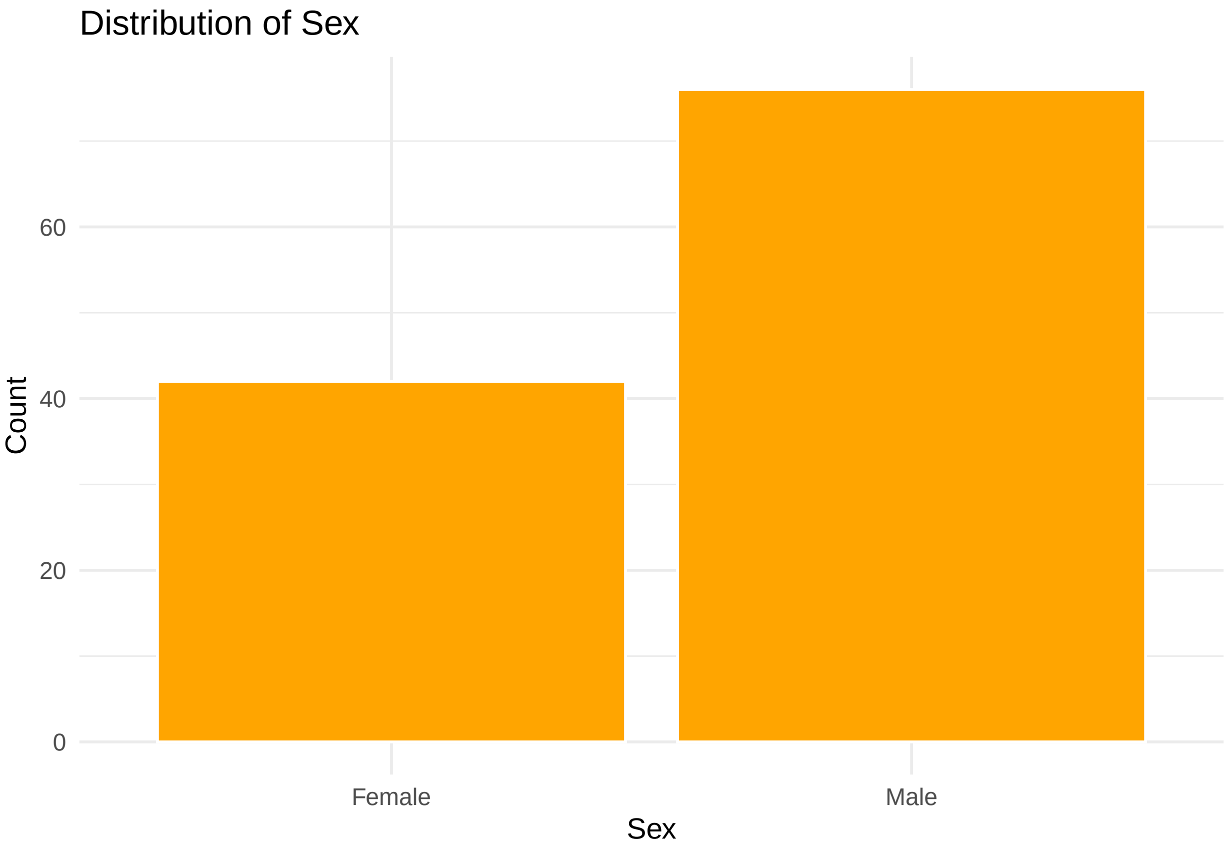
```r
ggplot(combined_data2, aes(y = Educational_Attainment)) +
  geom_bar(fill = "purple", color = "white") +
  labs(title = "Distribution of Educational Attainment Type", y = "Educational Attainment",
       x = "Count") +
  theme_minimal() +
  coord_flip()
```



The bar chart indicates that the majority of individuals are have attained a bachelors degree.

Distribution of Sex

```r
ggplot(combined_data2, aes(x = SEX)) +
  geom_bar(fill = "orange", color = "white") +
  labs(title = "Distribution of Sex", x = "Sex", y = "Count") +
  theme_minimal()
```



This bar plot shows the distribution of sex across the data set. It is visible that there are more Males then Females being represented.

What I want to Further Analyize

Setting up these univariate visuals allowed me to see some of the building blocks of what I would like to continue to research. It is clearly evident that white is a dominant race in this data set and that males dominate women. I would like to research these two variables particularly across educational attainment as well as field of degree and occupation to further analyze the discrepancies within STEM departments.

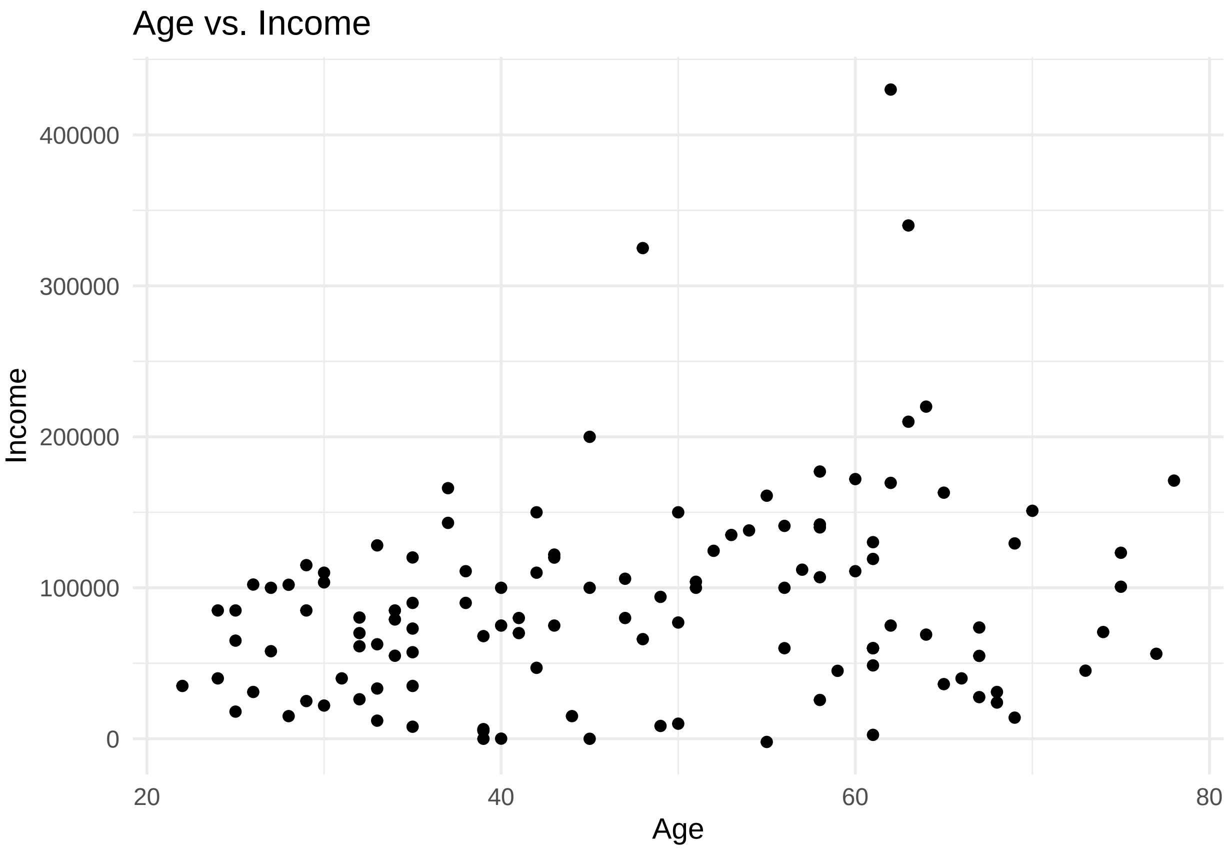
Exploring the Question

So far I have found tha the racial distribution is predominantly White. The age distribution is pretty evenly spread but could also be favoring the younger population.The income distribution is also right-skewed, with most individuals falling into the lower income brackets.The majority of individuals have attained the bachelors educational level. There is a greater representation of males compared to females in the dataset.

Multiple Forms of Visuals

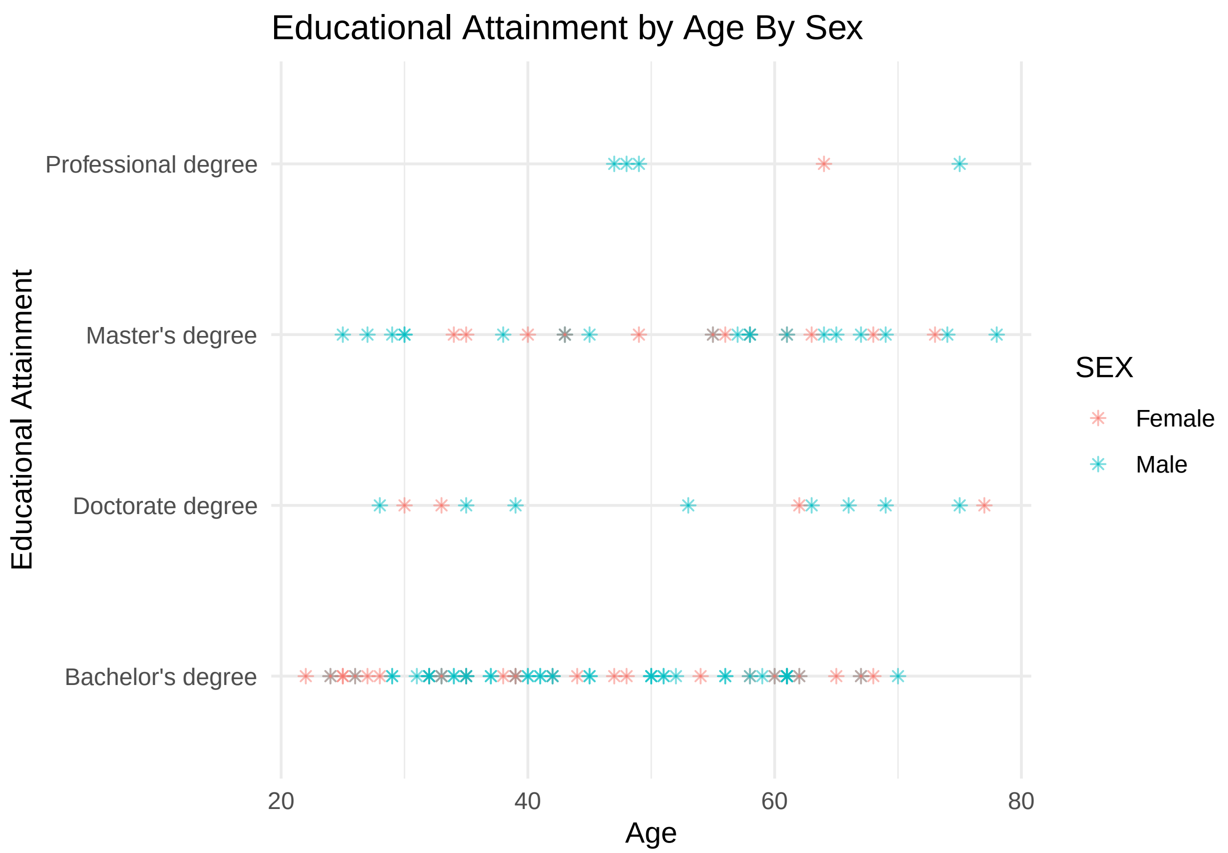
Categorical, Numerical and Comparision Visuals

```
ggplot(combined_data2, aes(x = Age, y = Income)) +
  geom_point() +
  labs(title = "Age vs. Income", x = "Age", y = "Income") +
  theme_minimal()
```



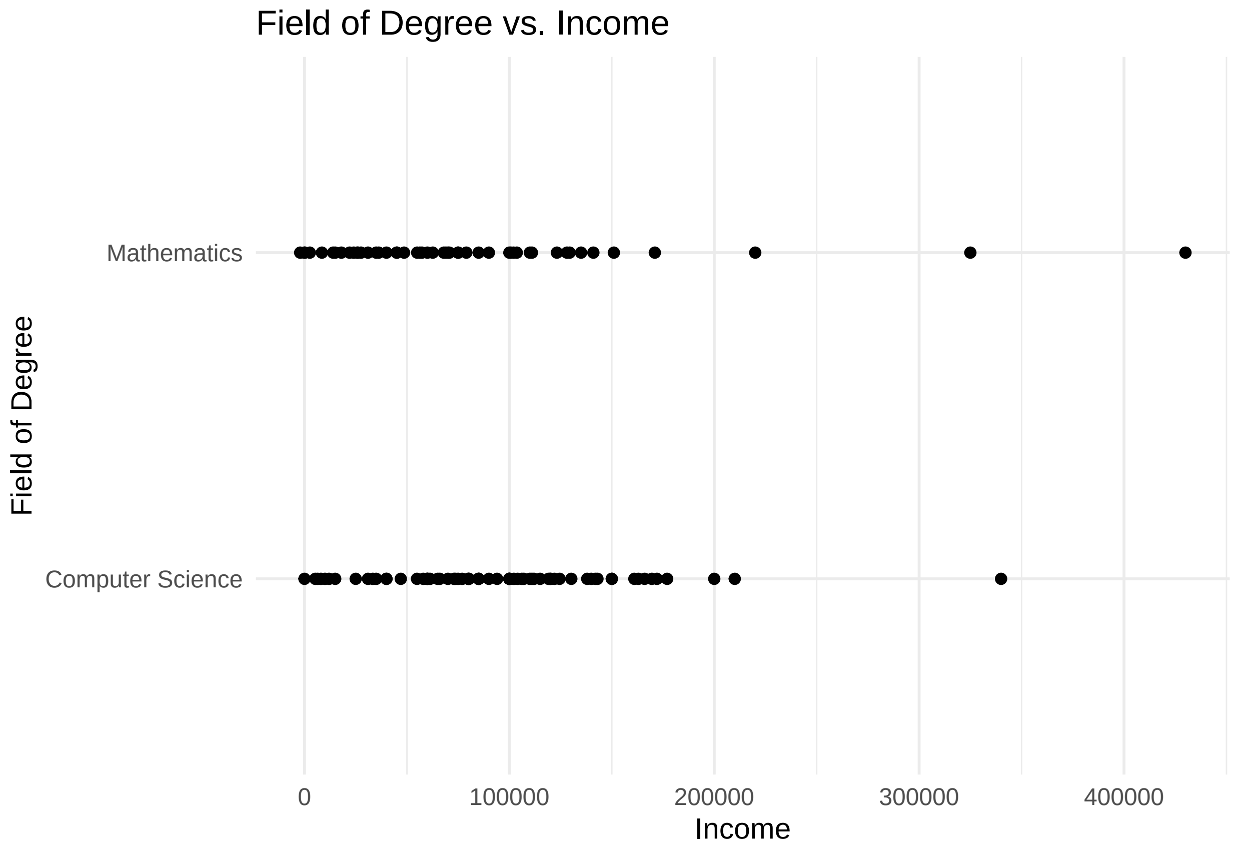
The scatter plot suggests a potential relationship between age and income where income levels seem to cluster around certain age ranges, with some higher income outliers across various ages; this indicates variability in income that could be influenced by factors such as years of experience, seniority, or industry, which would be interesting to explore further.

```
ggplot(combined_data2, aes(x = Age, y = Educational_Attainment, color = SEX)) +
  geom_point(shape = 8, alpha = .5) +
  labs(title = "Educational Attainment by Age By Sex", x = "Age",
       y = "Educational Attainment") +
  theme_minimal()
```



The graph depicts the distribution of educational attainment levels across different ages, suggesting that higher education degrees are achieved by individuals across a broad age range, reflecting both traditional and non-traditional educational paths.

```r
ggplot(combined_data2, aes(x = Field_Of_Degree, y = Income))+
  geom_point()+
  labs(title = "Field of Degree vs. Income", x = "Field of Degree", y = "Income") +
  theme_minimal() +
  coord_flip()
```



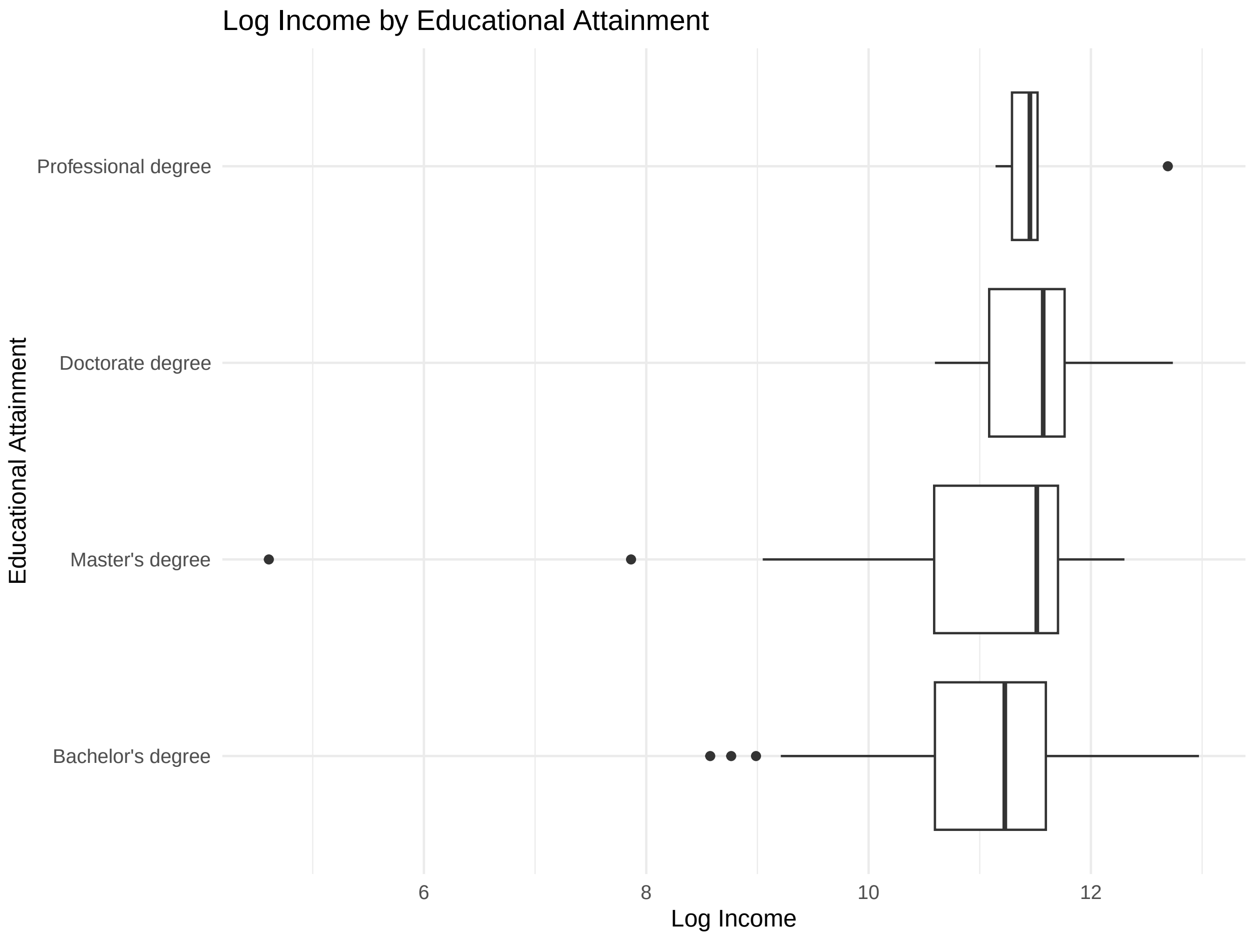
This plot shows the relationship with field of degree (specifically in a mathematics or compsci field). It shows a wide range of incomes within each field of study, with no single field consistently leading to the highest or lowest income. As you can see there are a few more larger outliers in the Mathematics field of degree versus the Computer Science.

```
combined_data2$LogIncome <- log(combined_data2$Income)
```

```
## Warning in log(combined_data2$Income): NaNs produced
```

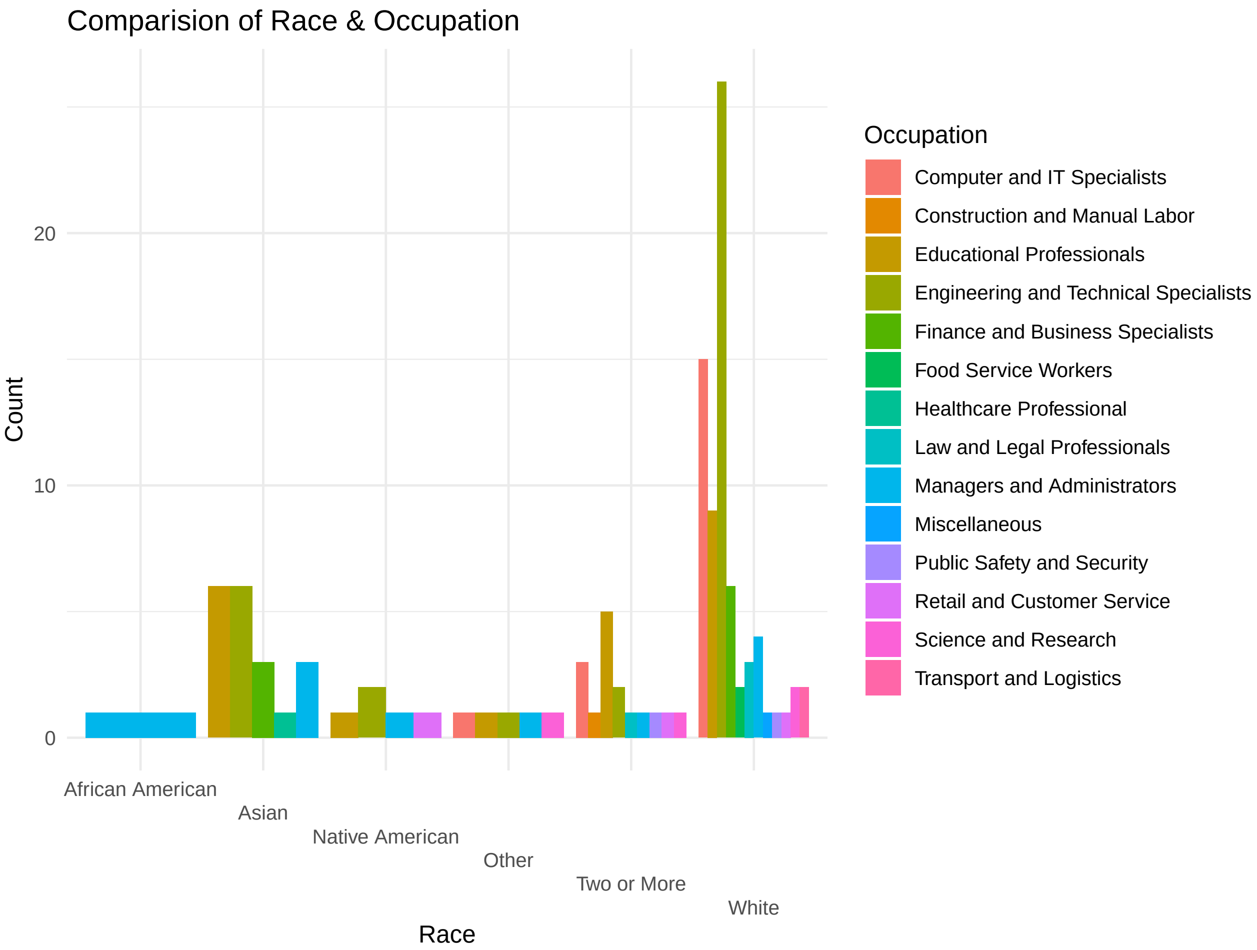
```
EA <- combined_data2$Educational_Attainment <- factor(combined_data2$Educational_Attainment,
                                                      levels = c( "Bachelor's degree",
                                                                  "Master's degree",
                                                                  "Doctorate degree",
                                                                  "Professional degree"))
ggplot(combined_data2, aes(x = EA, y = LogIncome)) +
  geom_boxplot() +
  labs(title = "Log Income by Educational Attainment",
       x = "Educational Attainment",
       y = "Log Income") +
  theme_minimal() +
  coord_flip()
```

```
## Warning: Removed 3 rows containing non-finite values (`stat_boxplot()`).
```



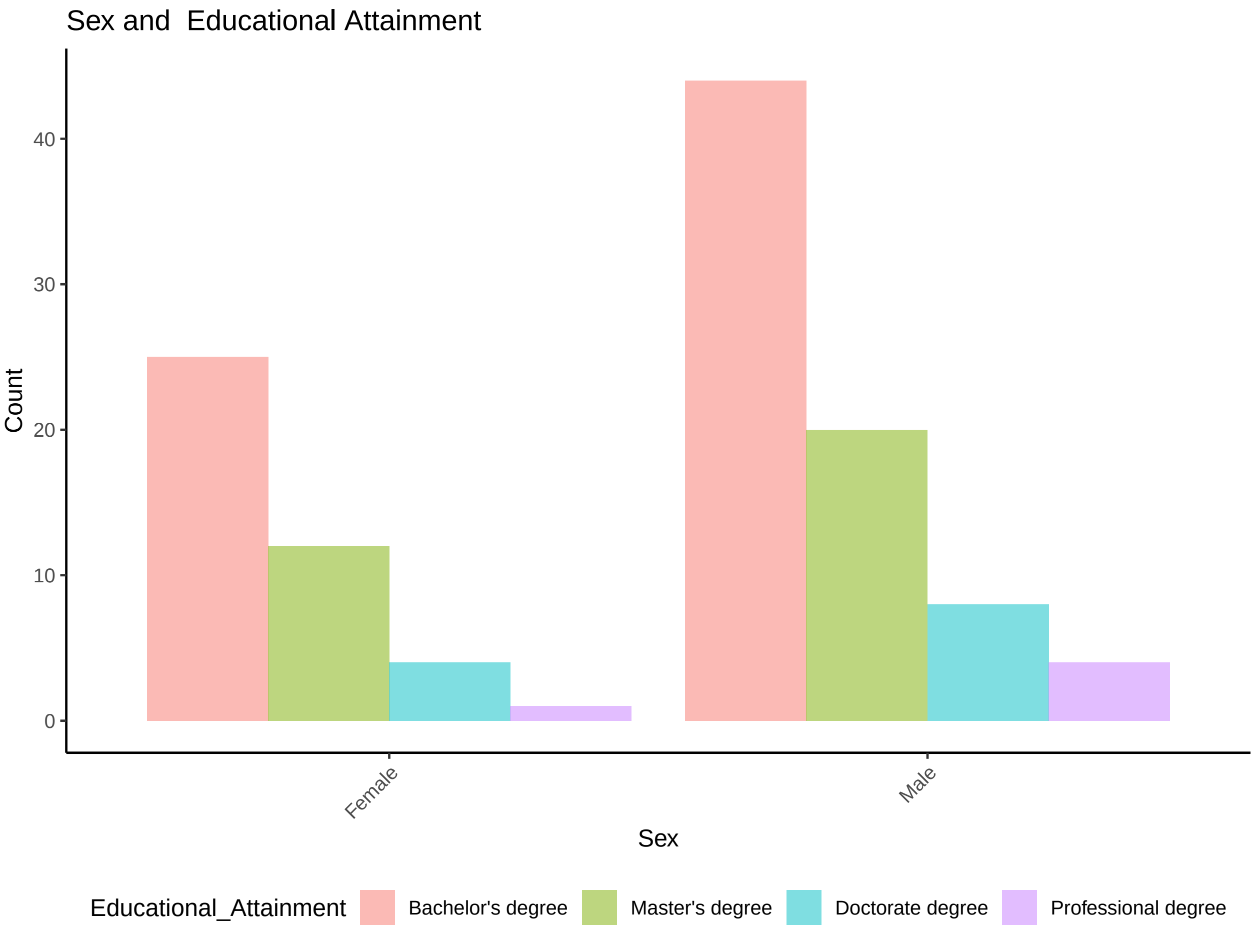
The graph illustrates the distribution of log-transformed income across different levels of educational attainment, indicating a general increase in income with higher educational degrees, yet with a considerable overlap between categories.

```r
ggplot(combined_data2, aes(x = Race, fill = Occupation)) +
  geom_bar(position = "dodge") +
  labs(title = "Comparision of Race & Occupation",
       x = "Race",
       y = "Count") +
  theme_minimal() +
  scale_x_discrete(guide = guide_axis(n.dodge=6))
```



This bar plot shows a comparision of Race and Occupation in this data set. This indicates that white is the most represented in the data set.

```
ggplot(combined_data2, aes(x = SEX, fill = Educational_Attainment)) +
  geom_bar(position = "dodge", alpha = .5) +
  labs(title = "Sex and  Educational Attainment",
       x = "Sex",
       y = "Count") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        strip.text.x = element_text(size = 8),
        legend.position = "bottom")
```



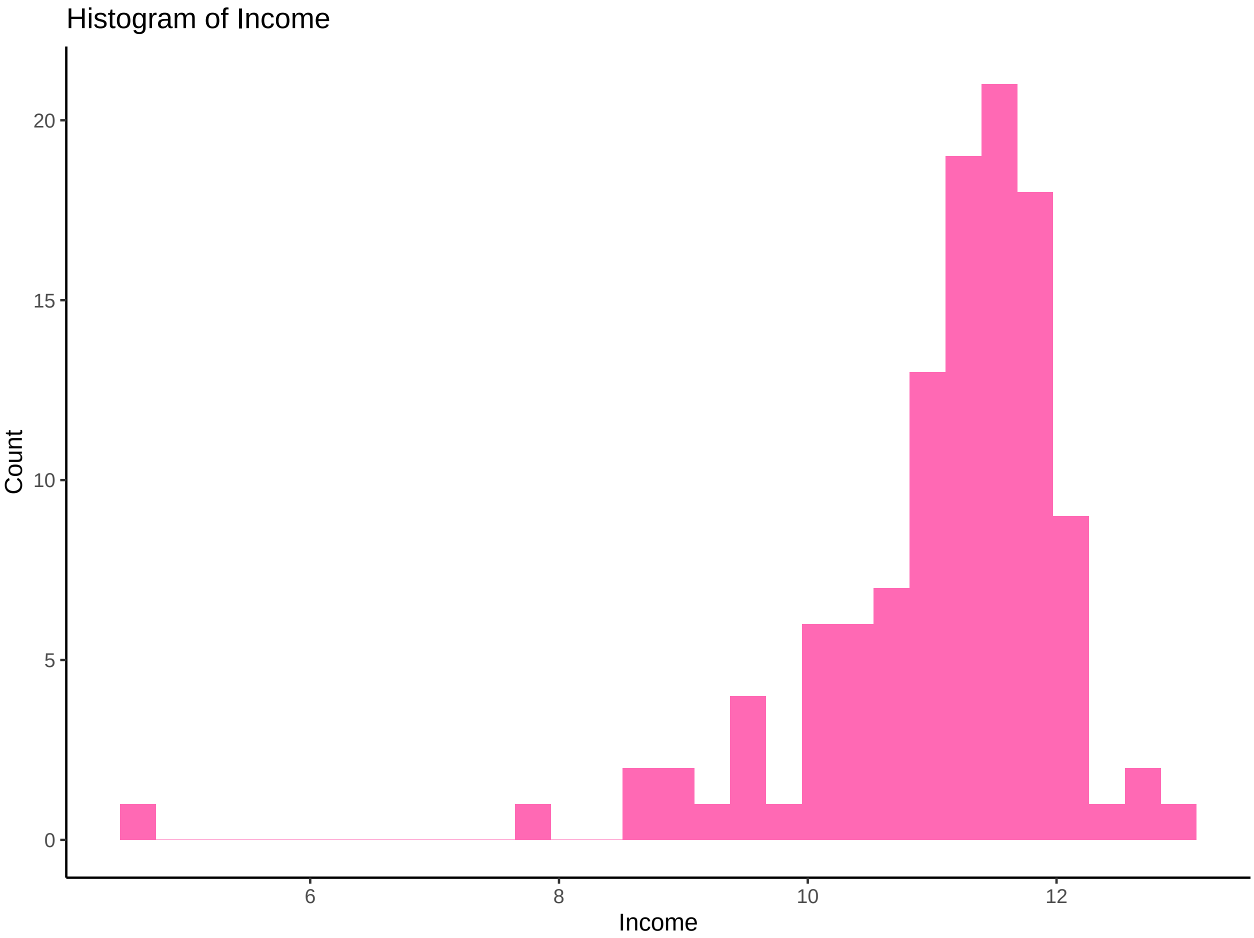
This bar plot shows a comparison of sex and educational attainment. This plot shows that there are more males who have educational attainment compared to women.

Distributions

```
ggplot(combined_data2, aes(x = LogIncome)) +
  geom_histogram(fill = "hotpink") +
  labs(title = "Histogram of Income", x = "Income",
y = "Count") +
  theme_classic()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

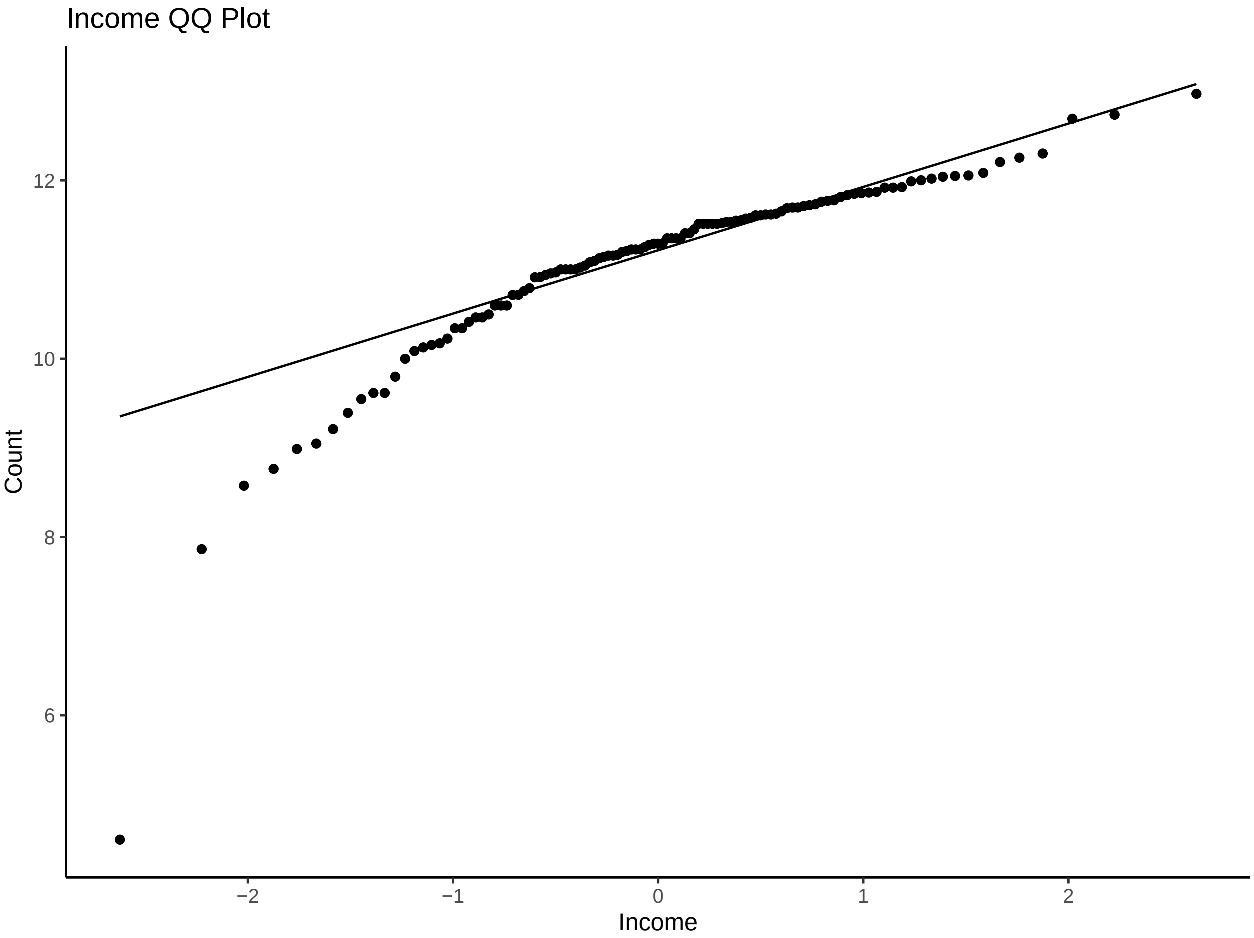
```
## Warning: Removed 3 rows containing non-finite values (`stat_bin()`).
```



The histogram displays a right-skewed distribution of income, with the most common values clustered around 10 and fewer instances of lower and higher incomes.

```
ggplot(combined_data2, aes(sample = LogIncome)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Income QQ Plot", x = "Income", y = "Count") +
  theme_classic()
```

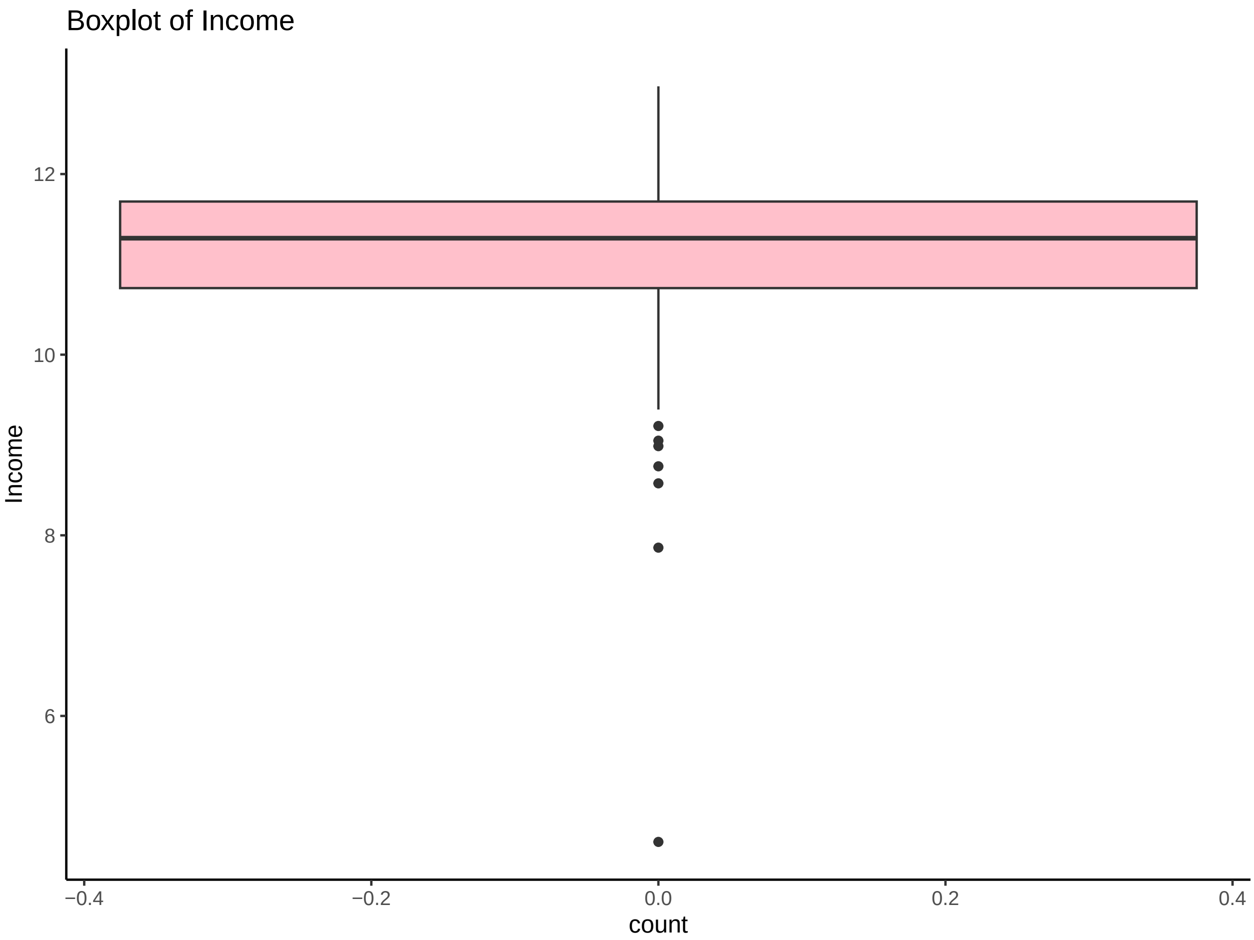
```
## Warning: Removed 3 rows containing non-finite values (`stat_qq()`).

## Warning: Removed 3 rows containing non-finite values (`stat_qq_line()`).
```



This QQ-plot shows that the deviation from the straight line at the ends suggests potential outliers or a departure from normality in the income data distribution.

```r
ggplot(combined_data2, aes(y = LogIncome)) +
  geom_boxplot(fill = "pink") +
  labs(title = "Boxplot of Income", x = "count", y = "Income") +
  theme_classic()
```

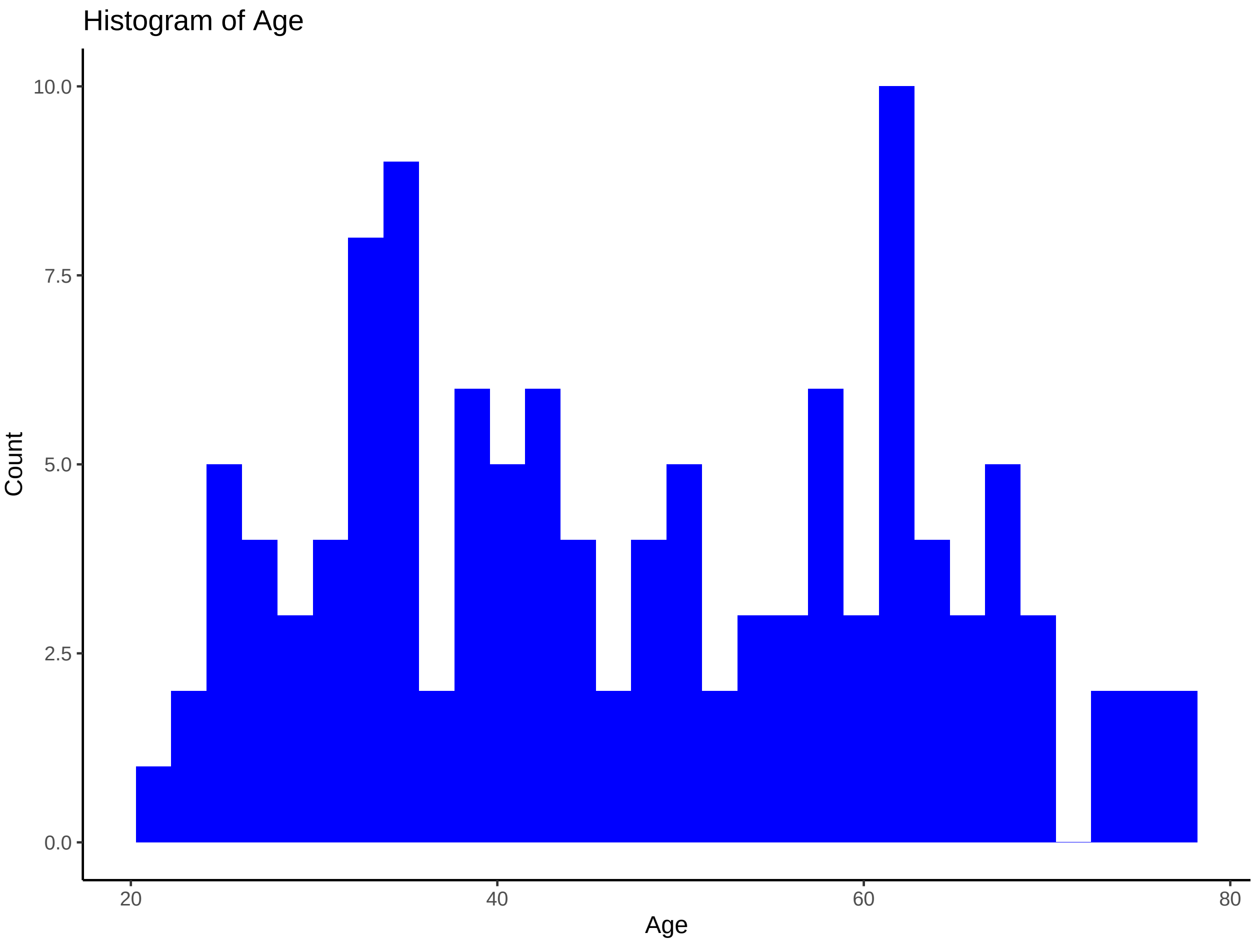
```
## Warning: Removed 3 rows containing non-finite values (`stat_boxplot()`).
```



This boxplot represents the distribution of income, highlighting the median, quartiles, and potential outliers, which indicate a variation in the income data with several points falling significantly below the main cluster of the data.

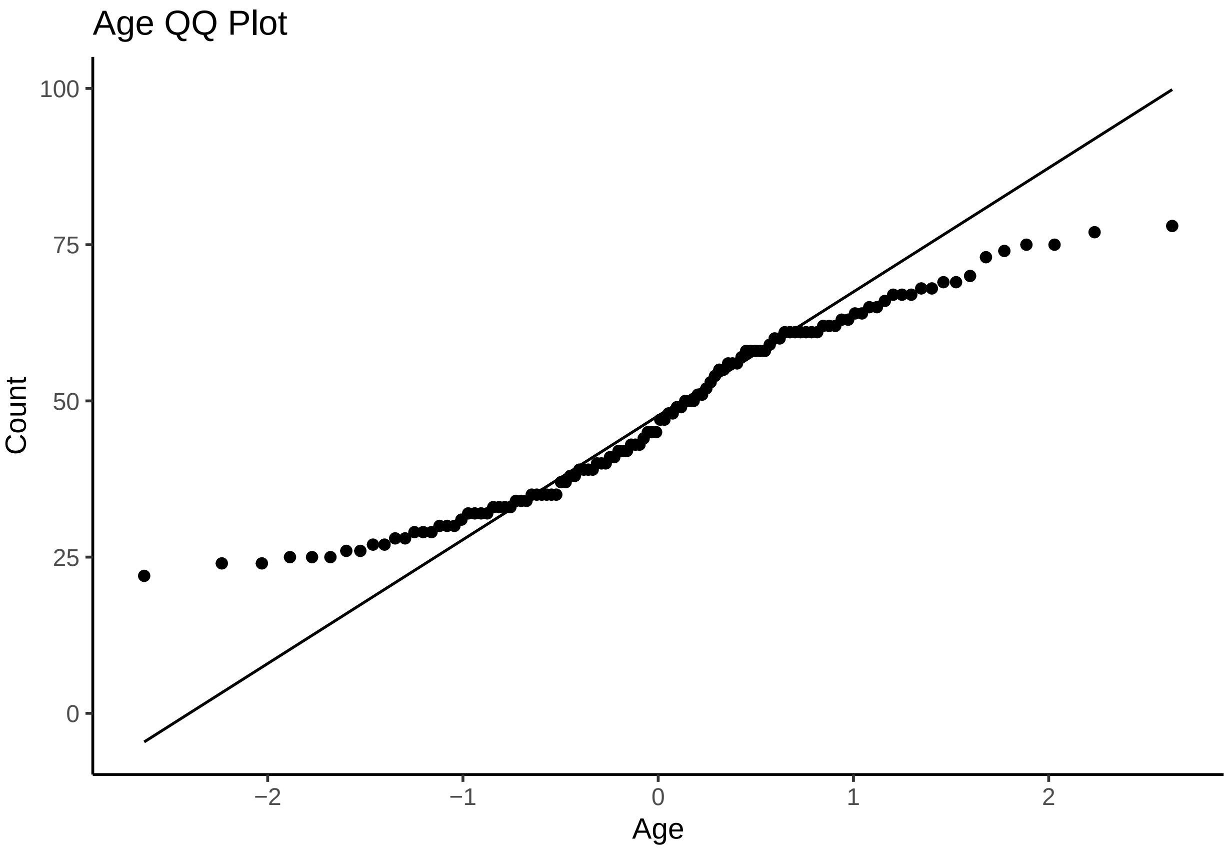
```
ggplot(combined_data2, aes(x = Age)) +
  geom_histogram(fill = "blue") +
  labs(title = "Histogram of Age", x = "Age",
y = "Count") +
  theme_classic()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



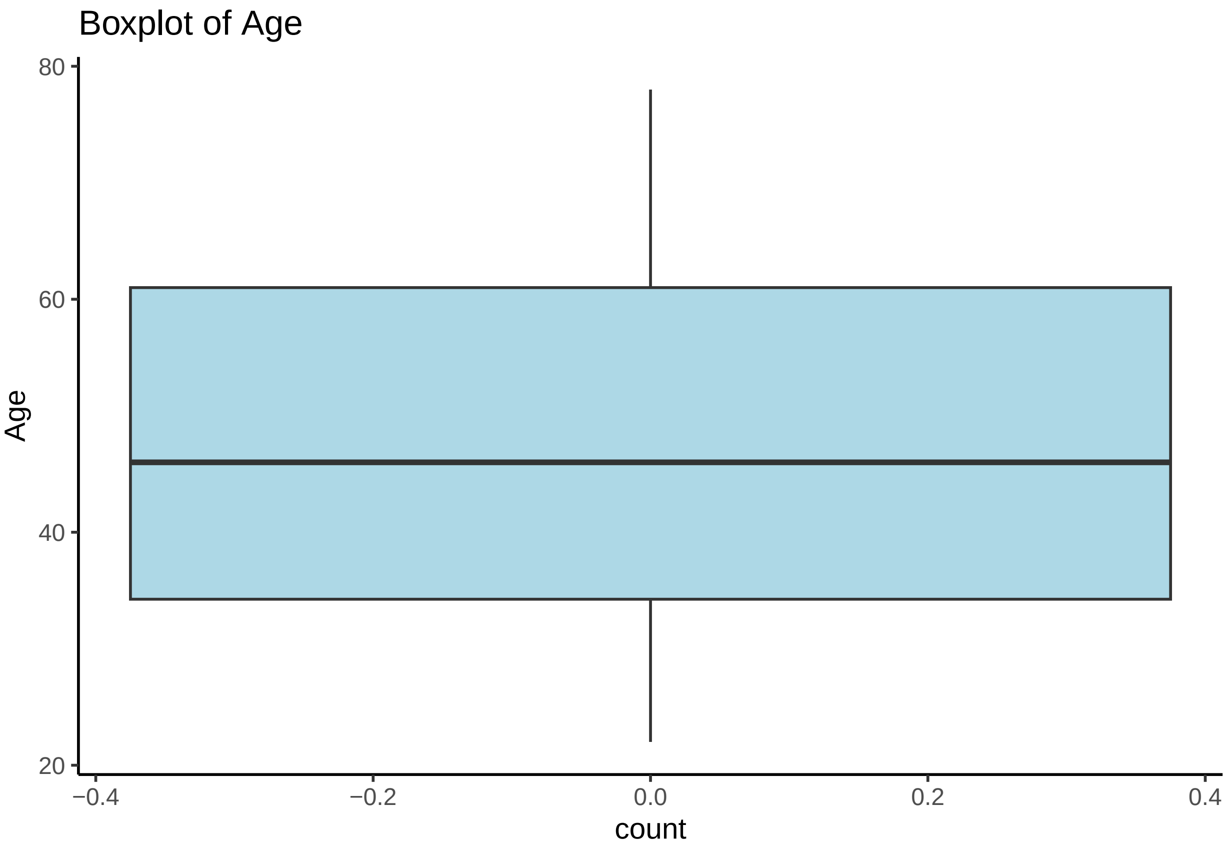
This histogram visualizes the distribution of age within a dataset, showing the frequency of various age groups and suggesting a non-uniform distribution with multiple peaks.

```r
ggplot(combined_data2, aes(sample = Age)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Age QQ Plot", x = "Age", y = "Count") +
  theme_classic()
```



This QQ-Plot compares the distribution of age to a normally distributed dataset, with deviations from the line indicating potential skewness or outliers in the age data.

```r
ggplot(combined_data2, aes(y = Age)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Boxplot of Age", x = "count", y = "Age") +
  theme_classic()
```



This Boxplot provides a summary of the distribution of age data, indicating the median, interquartile range, and potential outliers, with a relatively symmetrical spread around the median.

The Takeaways

Why these visuals?

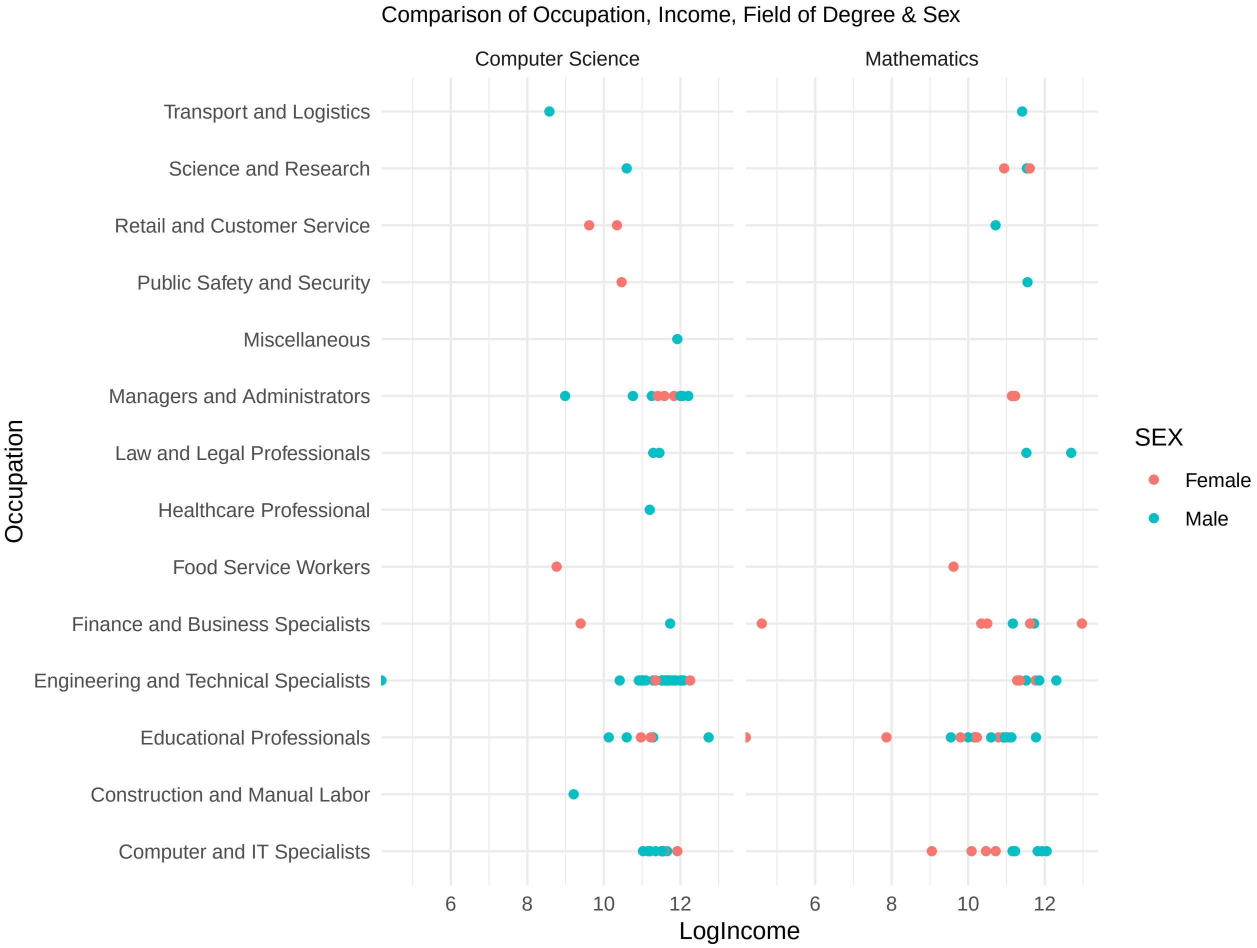
I decided on each of these visuals because they tell the story of the gender differences in Mathematics and Computer Science. The first visual I chose is a dot plot comparing Log income, occupation and sex. Dot plots are selected to compare individual data points across different categories. The dot plot I provided shows the distribution of incomes across various occupations split by gender (blue for male and pink for female). The second visual I chose is a box plot comparing sex, log income, educational attainment and field of degree. Box plots are used for showing the distribution of data based on minimum, first quartile, median, third quartile and maximum. The box plot I used compares the distribution of incomes within field of degrees (Computer Science and Mathematics) by gender and then split by educational attainment. The third visual I chose is a scatter plot with trend lines comparing Income and age with a male (blue) trend line and a female (pink) trend line. A scatter plot effectively shows the relationship between two quantitative variables. The scatter plot I used shows the relationship between age and income, split by gender. This helps identify linear trends or correlations and the shaded area around the lines indicate the confidence interval, giving a sense of. the reliability of the trends. The fourth visual I chose is a stacked bar chart comparing Sex, Race and Educational attainment. A stacked bar chart are helpful when you want to show the total size of groups along with the proportions of sub groups. The stacked bar chart I chose displays the count of individuals by gender and race, then split by educational attainment. This visual helps in understanding the distribution within each demographic category comparing them side by side. The fifth visual I chose is a faceted bar chart comparing Sex, Race, Occupation and field of degree. A faceted bar chart is helpful when you need to compare groups and subgroups within those groups. The faceted bar chart I chose displays the comparison between Occupation within fields of degree by gender and race. This visual allows for direct comparison of subgroups across the main categories.

5 Important Visuals

Visual 1

```
visual1 <- ggplot(combined_data2, aes(x = LogIncome, y = Occupation, color = SEX)) +
  geom_point() +
  labs(title = "Comparison of Occupation, Income, Field of Degree & Sex",
       x = "LogIncome",
       y = "Occupation") +
  facet_grid(. ~ Field_Of_Degree) +
  theme_minimal() +
  theme(plot.title = element_text(size = 10))
visual1
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



```
ggsave("visual1.png", plot = visual1, width = 8, height = 6)
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```

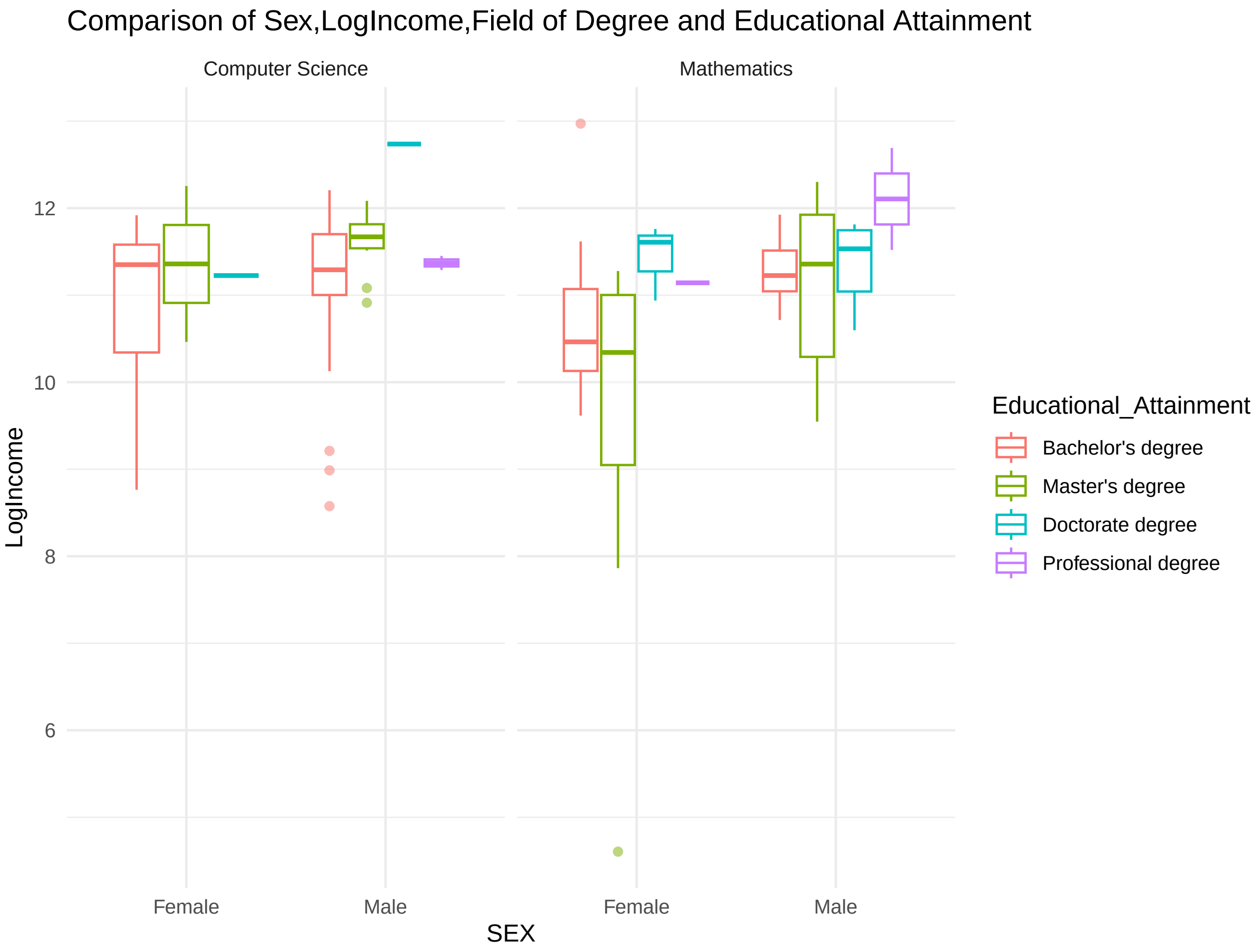
This visual is a dot plot showing the relationship between occupation and log-transformed income, segmented by gender and field of degree (Computer Science and Mathematics). Each dot represents the median income level for a particular occupation, with pink dots representing females and blue dots representing males.

Occupation is a Categorical Variable displayed on the y axis. Log Income is a Numeric variable displayed on the x axis. Field of Degree is a Categorical Variable split into two main categories Computer Science and Mathematics. Sex is a Categorical Variable represented by the color of the dots. Dot plots are appropriate for this type of data as they allow individual income data points to be displayed within occupational categories, providing a clear visual of income distribution within and across these categories. Separating the data by field of degree permits a specialized view that might reveal different income dynamics in each field. The spread of dots within each occupation indicates the variability of income for that profession. A wide spread suggests a significant disparity in how much individuals earn in that occupation. The color differentiation between sexes allows for an analysis of income equity within occupations. From this visualization, one can infer the differences in income within various occupations and how these might correlate with the field of degree and differ between sexes. Such information can be valuable for individuals making career choices, employers setting pay scales, or policymakers focused on labor and economic development. It can also serve as a basis for discussions on gender pay equity across different professional sectors. A specific story from this graph may highlight the challenges that women face in achieving income equally with men, despite having similar levels of education and working within the same fields. It underscores the necessity for ongoing dialogue and intervention regarding equal pay and may inspire action from stakeholders at various levels—from corporate leaders and hiring managers to educators and policymakers—to create more equitable compensation structures.

Visual 2

```r
visual2 <- ggplot(combined_data2, aes(x = SEX, y = LogIncome, color = Educational_Attainment)) +
  geom_boxplot(shape = 17, alpha = .5) +
  labs(title = "Comparison of Sex,LogIncome,Field of Degree and Educational Attainment") +
  theme_minimal() +
  facet_grid(. ~ Field_Of_Degree)
visual2
```

```
## Warning: Removed 3 rows containing non-finite values (`stat_boxplot()`).
```



```
ggsave("visual2.png", plot = visual2, width = 8, height = 6)
```

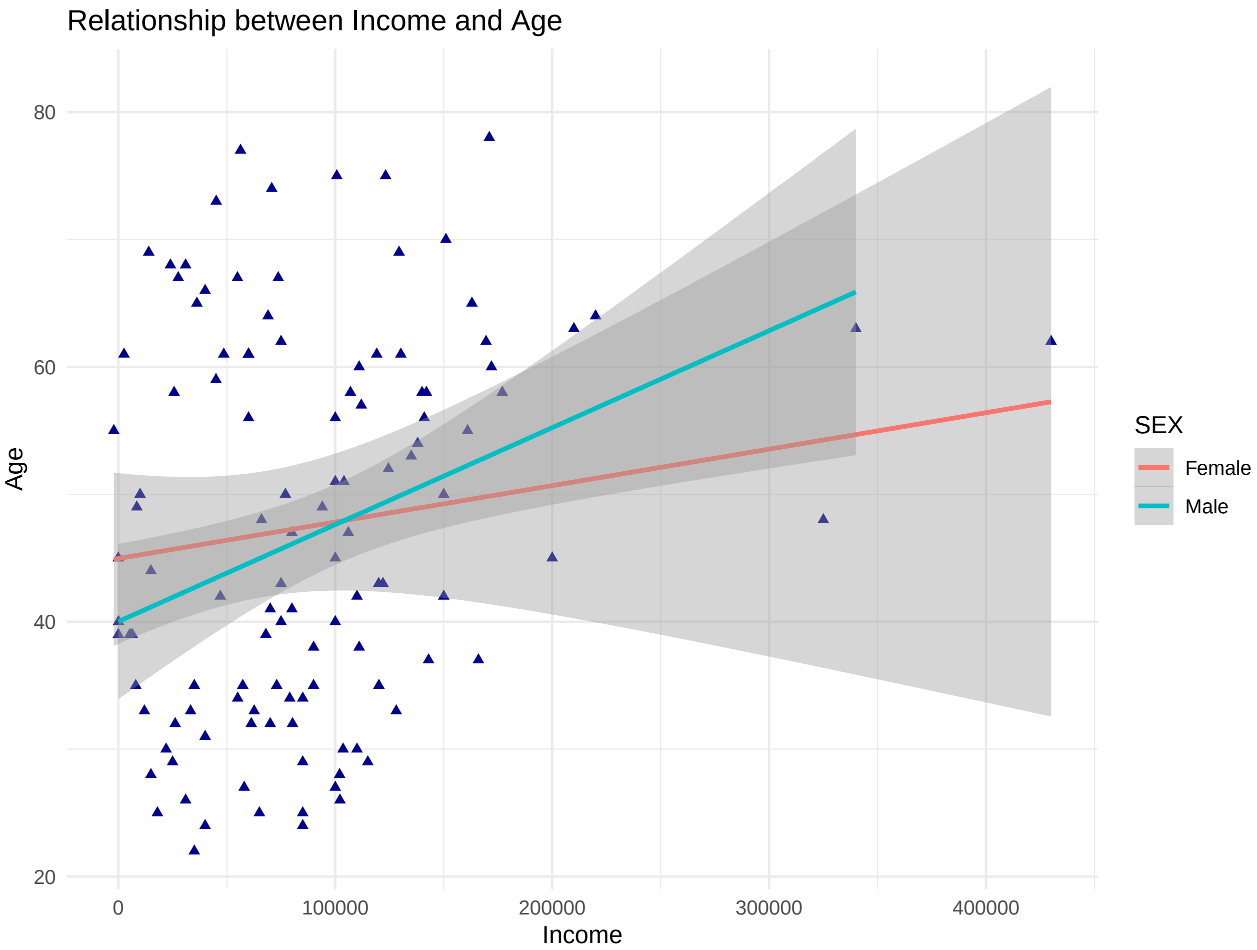
```
## Warning: Removed 3 rows containing non-finite values (`stat_boxplot()`).
```

This visual is a box plot chart comparing the log-transformed income across genders for individuals with different levels of educational attainment in the fields of Computer Science and Mathematics. Each box plot represents the income distribution for the respective category, where the central line indicates the median income, the box spans from the first quartile (Q1) to the third quartile (Q3), and the "whiskers" extend to show the range of the data, excluding outliers, which are plotted as individual points. Sex is a categorical variable displayed on the x-axis. LogIncome is a Quantitative Variable displayed on the y-axis. Educational Attainment is a Categorical Variable represented by the color of the boxes. Field of Degree is a Categorical Variable displayed as two groups on the upper x-axis. Box plots are particularly useful for displaying the distribution of a quantitative variable across different levels of several categorical variables. This makes it possible to compare the median, spread, and skewness of log income across different sexes, levels of educational attainment, and fields of degree. The use of different colors for educational attainment makes it easy to distinguish between the educational categories within each field and sex group.The logarithmic scale for income allows for easier comparison across a wide range of values and can make certain patterns more apparent. By comparing the median lines within the boxes, we can perceive which groups tend to have higher or lower median incomes. For instance, the median lines are typically higher for male across both groups compared to females. The presence of outliers can indicate that there are individuals whose income is significantly different from the norm within their group. Another observation could be that the spread of incomes (as indicated by the lengths of the boxes and whiskers) is different between the genders and degrees, suggesting variability in how education level impacts income within these fields. This graph tells a story about how income correlates with educational attainment in the fields of Computer Science and Mathematics, and how this relationship may vary between males and females. It can highlight issues such as pay equity and the economic value of higher education in different fields. Decision-makers in education and employment might use such data to address gender disparities and to evaluate the return on investment in higher education.

Visual 3

```
visual3 <- ggplot(combined_data2, aes(x = Income, y = Age, color = SEX)) +
  geom_point(color = "darkblue", shape = 17) +
  labs(title = "Relationship between Income and Age",
       x = "Income",
       y = "Age") +
  theme_minimal() +
  geom_smooth(method = "lm")
visual3
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



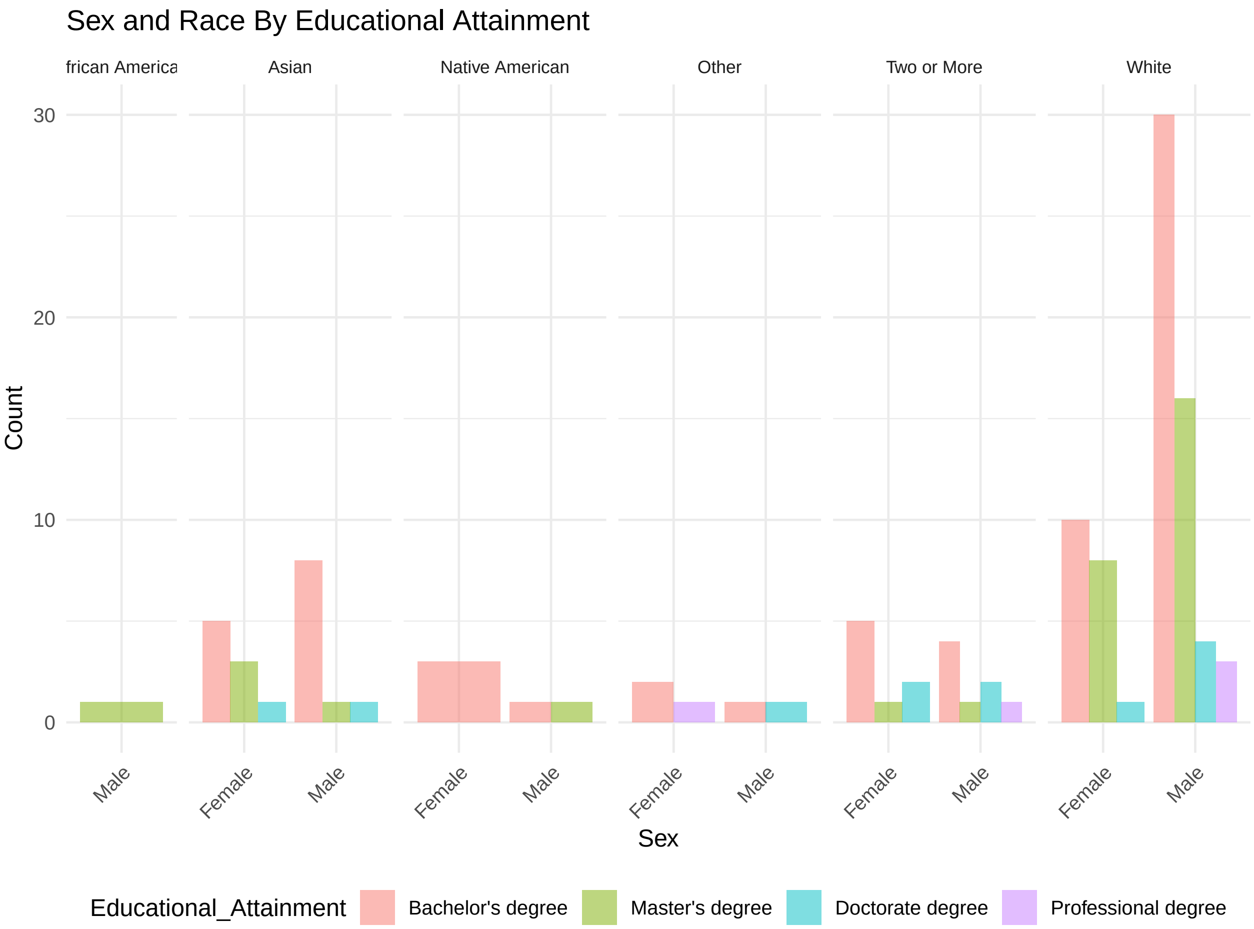
```
ggsave("visual3.png", plot = visual3, width = 8, height = 6)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

This visual is a scatter plot with trend lines, showing the relationship between income and age for two different genders, indicated by the pink line for females and the blue line for males. Age is a quantitative variable represented on the y-axis. Income is a quantitative variable represented on the x-axis. Sex is a categorical variable with two categories—female and male—represented by the color of the trend lines and data points (triangles). The age is measured in years, and the scale is linear, ranging from the lowest to the highest age present in the dataset. The income is likely measured in dollars (or another currency), and the scale appears to be linear as well, showing the range from the lowest to the highest income recorded. The triangles represent individual data points, with the position indicating the age and income of each respondent. There is a positive relationship between age and income for both sexes, as indicated by the upward slope of the trend lines. This suggests that as people get older, their income tends to increase. The separation between the trend lines for males and females indicates that there may be a difference in income with respect to age between the two sexes. The trend line for males is consistently higher than that for females, it could imply that males, on average, have a higher income than females at the same age. A specific story from this graph could focus on the gender income gap as it relates to age, suggesting that while income increases with age for both sexes, there may be systemic differences in how this increase manifests across genders. This kind of insight is valuable for discussions on wage equality and the economic impact of age and gender on earnings.

Visual 4

```r
visual4 <- ggplot(combined_data2, aes(x = SEX, fill = Educational_Attainment)) +
  geom_bar(position = "dodge", alpha = .5) +
  labs(title = "Sex and Race By Educational Attainment",
       x = "Sex",
       y = "Count") +
  facet_grid(. ~ Race, scales = "free_x", space = "free_x") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        strip.text.x = element_text(size = 8),
        legend.position = "bottom")
visual4
```

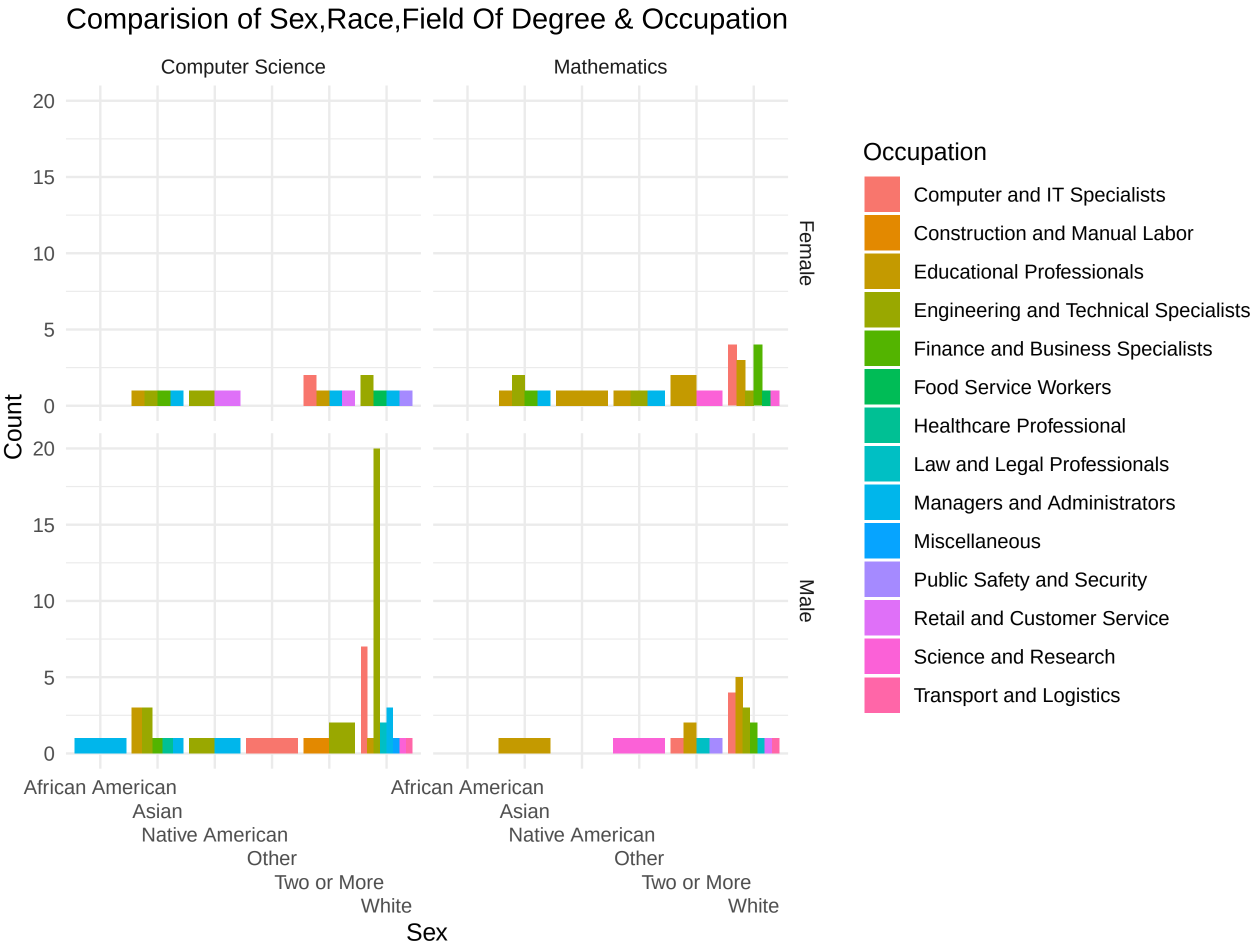


```
ggsave("visual4.png", plot = visual4, width = 8, height = 6)
```

This visual is a bar chart showing the count of individuals by their sex, broken down by Educational Attainment and further categorized by race. Sex is a Categorical variable displayed on the x-axis with two categories represented, Male and Female. Educational Attainment is Categorical variable represented by the colors of the bars. The different colors correspond to different levels of educational attainment: Bachelor's degree, Master's degree, Doctorate degree, and Professional degree. Count is a numerical variable represented on the y-axis. This appears to be the number of individuals who fall into each category of sex, race, and educational attainment. Race is a Categorical variable that segments the data across the top of the chart. The categories include African American, Asian, Native American, Other, Two or more, and White. Each cluster of bars corresponds to a particular race, with separate bars for males and females within that racial group. The height of each bar indicates the count of individuals of a particular sex and race who have attained a specific level of education. The distinct colors within each bar group allow for quick visual comparison between the educational attainments for each sex within a race. The faceted bar chart is appropriate for this data as it allows the viewer to compare educational attainment across two categorical variables (sex and race) simultaneously. It provides a clear visual differentiation between groups, which helps in comparing the relative sizes of these groups. At a glance, the graph reveals patterns of educational attainment across different sexes and races. It shows disparities in educational outcomes; for example; race of white has more representation in male and female compared to all the other races. The specific story one might draw from this graph is that educational attainment is not uniformly distributed across sexes and races. It might reflect socio-economic factors, cultural influences, or the effectiveness of educational policies aimed at different demographic groups. This visualization could be used by educational institutions and policymakers to target interventions where they are most needed to improve equity in educational attainment.

Visual 5

```
visual5 <- ggplot(combined_data2, aes(x = Race, fill = Occupation)) +
  geom_bar(position = "dodge") +
  labs(title = "Comparision of Sex,Race,Field Of Degree & Occupation",
       x = "Sex",
       y = "Count") +
  theme_minimal() +
  scale_x_discrete(guide = guide_axis(n.dodge=6)) +
  facet_grid(SEX ~ Field_Of_Degree)
visual5
```



```
ggsave("visual5.png", plot = visual5, width = 8, height = 6)
```

This visual is a faceted bar chart that depicts the count of individuals by sex (female and male) within two fields of degree (Computer Science and Mathematics), further categorized by race. Each bar represents an occupation. A Categorical variable represented on the x-axis, with two levels (Female and Male). A Categorical variable, which is likely used to create clusters within the sex categories. The specific races included are African American, Asian, Native American, Other, Two or more, and White. Occupation is a categorical variable, with each color representing a different occupation, as indicated in the legend to the right. Occupations include a range from Computer and IT Specialists to Transport and Logistics, among others. Count is a numerical measure represented on the y-axis, indicating the number of individuals or instances within each category. The bar chart displays the count of individuals across different occupations categorized by sex and further divided by race. Each bar's height corresponds to the count of individuals within the respective sex-race-occupation grouping. Different colors represent different occupations, which makes it easy to see the distribution of occupations within each sex and race category. The facted bar chart is appropriate here because it allows for the comparison of multiple categories (race and occupation) within each sex grouping. The graph suggests a disparity in the representation of women, particularly in the fields of Computer Science and Mathematics. Some findings are, the Under representation of Women in STEM. There appears to be a lower count of females across all races in Computer Science and Mathematics compared to males. This could indicate that women are underrepresented in these fields, which is a pattern that has been historically observed in STEM professions. In contrast to females, the bars for males might show higher counts across races, reinforcing the narrative of a gender gap in these occupations. The data speaks to the concept of intersectionality, where the overlap of gender and race creates different experiences for women of various racial backgrounds in accessing and thriving within STEM occupations. The specific story here is that while there may be some progress or representation of women in the Computer Science and Mathematics fields, there is a clear indication that more work needs to be done to achieve gender equality, and this need may be more pronounced for certain racial groups. This graph could be a call to action for educators, policymakers, and industry leaders to address the underlying causes of these disparities and work towards a more equitable and inclusive workforce in STEM fields.

Next Steps and Post-Mortem

As I reflect on my project analyzing data from the American Community Survey (ACS), I realize it's been a journey of discovery and learning. Starting with the acquisition of the data via an API key, my initial steps were focused on cleaning and restructuring the dataset to make it more comprehensible. The process of researching this data revealed patterns and disparities in aspects like income and education levels across various demographics, highlighting the influence of race and gender on economic and educational outcomes. This project taught me a lot about how to get data ready for analysis in the real world. I learned how to fix missing or wrong data and how to arrange the data to make sense of it. I also got better at exploring data to help with my research and making graphs that show complicated information in a simple way. One big finding was that there are noticeable differences in income and education among different groups. However, there are still unanswered questions, especially about why women in STEM fields might earn less. Things like where someone lives or their personal career choices, which weren't in my data, might also affect their income. There were challenges, like getting the API key and cleaning up a lot of code. But I got better at making visuals and staying on track with my project timeline. Next, I want to study more about why these income differences exist, talk to others about what I found, and look at more data over time. I want to look beyond just education and race, and consider things like work experience and economic trends. I plan to discuss these topics with experts and share my findings with schools and online platforms to raise awareness. Also, I'd like to make a model that predicts salaries based on someone's background, using statistics to see how different factors affect income. This model would analyze variables such as educational qualifications, years of experience, industry sector, and demographic details like age, gender, and race. By applying statistical methods and machine learning techniques, the model could uncover how these factors collectively influence salary levels. It would involve testing different algorithms to find the most accurate and reliable predictions, ensuring the model is sensitive to nuances in the data. Additionally, the model would be designed to be adaptable, allowing for the incorporation of new data over time, which is crucial for maintaining its relevance in the ever-changing job market. Through this model, I aim to provide a deeper understanding of income dynamics and offer valuable insights for both individuals and policymakers in addressing wage disparities and promoting equitable employment practices. Despite the insights gained, some questions remain elusive, particularly regarding the gender pay gap in STEM fields. The influence of geographical location and individual career paths, which were absent from my data set, could be pivotal factors in understanding these disparities. Navigating challenges like acquiring the API key and refining a substantial amount of code were part of this journey. However, these hurdles contributed to my growth in data visualization and project management.