SDS322E Project Report 1

Names: Yunseo Hur, Jack Misukanis, Harini Shanmugam

---

1. Title and Introduction

Completed by Jack and Harini

For this project, we wanted to look at income disparity in the US in 2021 and study it from view point of variables not commonly studied.

The specific datasets used are from various sections of the Current Population Survey from the US Census Bureau.

Every dataset used contains the shared variable of Income Level and one additional characteristic variable. The unique characteristic/variables in each data set are are **Level of income**, **Age**, **Level of education attained**, **Marital status**, **Relation to the family household**, **Type of source of income**, and **Tenure type** (whether the individual owns or rents their place of residence).

We chose the variables of Age and Education level because those are commonly studied variables when it comes to this topic and we can see if what we already know about Age and Education level still holds true in 2021. But we chose Marital Status, Relationship to the family household, Type of the source of income, and Tenure status to study in addition to find any new correlation or subset of the population performing well on a specific basis.

Depending on the dataset, a unique row may represent a characteristic, such as whether an individual was living in a renter occupied or owner occupied domicile such as in PINC01. Rows could also represent characteristics regarding specific sources of income streams (like social security or wages and salary) as seen in PINC08. In other datasets, rows could represent an income level, such as in PINC02, where income is broken down into increments $10,000.

We can join the variables based on these income buckets using the full_join function as each of the datasets are split into these income buckets of $10,000.

Although the differing variable in each data set are characteristics (which would look like they are categorical data), they are all still numerical variables because we have a **count** of each characteristic at each income level.

Potential relationships we may expect to see are correlations between individuals having a high level of income and having a higher age, higher education level, being the head of the household, and owning their domicile (tenure).

---

2. Tidying

Completed by Yunseo and Harini

```
# loading all necessary libraries
library(readxl) # to use read_excel() function
library(tidyverse) # to use dplyr and ggplot functions
library(RColorBrewer) # to use scale_fill_brewer() functions

# all data sets loaded in as objects
# age
age <- read_excel("pinc01_1_1_1.xlsx", sheet = 4)

# education level
edu <- read_excel("pinc03_1_1_1_1.xlsx")

# marital status
marital <- read_excel("pinc02_1_1_1.xlsx")

# relationship to family household
relation <- read_excel("pinc01_1_1_1.xlsx", sheet = 5)

# sources of income types
source <- read_excel("pinc08_1_1_1.xlsx")

# tenure type (owner or renter of residence)
tenure <- read_excel("pinc01_1_1_1.xlsx", sheet = 7)
```

Tidying of each variable:
Age

```
# save changes to new object
age2 <- age %>%
  # only keep rows 1, 2, 5, 8, 11, 15, and 18
  slice(1,2,5,8,11,15,18) %>%
  # pivot first row (levels of income) into a column (income levels)
  # -1 excludes the first column from pivoting
  pivot_longer(-1, names_to = "Income Level",
               values_to = "# of People") %>%
  rename("Age" = "Characteristic") %>%
  # display Age column first and then every other column as they were
  select("Age", everything())
```

Educational Level

```
# save changes to new object
edu2 <- edu %>%
  # pivot first row (levels of income) into a column (income levels)
  # -1 excludes the first column from pivoting
  pivot_longer(-1, names_to = "Education Level",
               values_to = "# of People") %>%
```

```
  rename("Income Level" = "Characteristic") %>%
  # display Education Level column first, then every other column as they were
  select("Education Level", everything())
```

Marital Status

```
# save changes to new object
marital2 <- marital %>%
  # change the values from columns 2 to 7 to numeric values
  mutate_at(2:7,as.numeric) %>%
  # only keep rows 1 to 11
  slice(1:11) %>%
  # pivot first row (levels of income) into a column (income levels)
  # -1 excludes the first column from pivoting
  pivot_longer(-1, names_to = "Marital Status",
               values_to = "# of People") %>%
  rename("Income Level" = "Characteristic") %>%
  # display Marital Status column first and then every other column as they were
  select("Marital Status", everything())
```

Relationship to Family Householder

```
# save changes to new object
relation2 <- relation %>%
  # pivot first row (levels of income) into a column (income levels)
  # -1 excludes the first column from pivoting
  pivot_longer(-1, names_to = "Income Level",
               values_to = "# of People") %>%
  rename("Relationship to Family Household" = "Characteristic")
```

Sources of Income

```
# save changes to new object
source2 <- source %>%
  # remove values that read 'NA'
  na_if("NA") %>%
  # change the values from columns 2 to 12 to numeric values
  mutate_at(2:12,as.numeric) %>%
  # only look at the data from rows 2, 5, 13, 17, 24, and 34 in the data set
  slice(2,5,13,17,24,34) %>%
  # pivot first row (levels of income) into a column (income levels)
  # -1 excludes the first column from pivoting
  pivot_longer(-1, names_to = "Income Level",
               values_to = "# of People") %>%
  rename("Source of Income Type" = "Characteristic")
```

Tenure (owner/renter status of residence)

```
# save changes to new object
tenure2 <- tenure %>%
  # pivot first row (levels of income) into a column (income levels)
  # -1 excludes the first column from pivoting
```

```
  pivot_longer(-1, names_to = "Income Level",
               values_to = "# of People") %>%
  rename("Tenure Type" = "Characteristic")
```

---

3. Joining/Merging

Completed by Harini

```
# joining all 6 data sets by the shared variable 'Income Level'
all <- age2 %>%
  full_join(edu2, by = "Income Level") %>%
  full_join(marital2, by = "Income Level") %>%
  full_join(relation2, by = "Income Level") %>%
  full_join(source2, by = "Income Level") %>%
  full_join(tenure2, by = "Income Level")
# using full_join because we want to see the information under all ID variables
# after joining
```

Total observations in each data set (after tidying):

- Age: 77 observations
- Educational Attainment: 99 observations
- Marital Status: 66 observations
- Relationship to Family Household: 55 observations
- Sources of Income: 66 observations
- Tenure (owner or renter of residence): 22 observations

Unique IDs in each dataset: There is only one unique ID variable in each dataset

IDs that appear in one dataset but not the other: 'Age', 'Education Level', 'Marital Status', 'Relationship to Family Household', 'Source of Income Type', and 'Tenure Type'

IDs in common: The only ID in common is 'Income Levels'

IDs that may have been left out: There are no IDs that were left out

Observations/ rows that were dropped/added when joining the data sets: No rows were dropped when joining the data sets since we used the full_join function only, so all observations were kept. One issue, however, is that most of the data are disjoint from each other across datasets, so after joining there are many NA values

---

4. Wrangling

Completed by Yunseo and Jack

Categorical – count of people below the poverty line

```
all %>%
  # select these columns/variables only
  select(`Income Level`,`Tenure Type`,`# of People.y.y.y`) %>%
  # filter for when 'Income Level' is either of these values
  filter(`Income Level`=="$1 to $9,999 or loss"
         |`Income Level`=="$10,000 to $19,999") %>%
  # order the data with 'income level' listed first
  arrange(`Income Level`) %>%
  # remove duplicate rows in the variable and keep all data
  distinct(`# of People.y.y.y`, .keep_all = TRUE) %>%
  # calculate sum of observations/count in the variable
  summarize(sum(`# of People.y.y.y`))
```

```
## # A tibble: 1 x 1
##   `sum(\`# of People.y.y.y\`)`
##                          <dbl>
## 1                        65522
```

There are 65,522,000 of the total income earners in the US who are making an income below the poverty line. This calculation only calculates the sum of those who make less than $20,000 annually, but the poverty line is at around $25,000. So if we had smaller buckets for the income levels, we would find a more accurate count for the number of people living below the poverty line.

Source for where we learned to use ".keep_all = TRUE" : https://www.datanovia.com/en/lessons/identify-and-remove-duplicate-data-in-r/

Categorical – Proportion of people in each income level with income through wages and salary only

```
all %>%
  # group by the source types under 'Source of Income Type'
  group_by(`Source of Income Type`) %>%
  # remove duplicate rows in the variable and keep all data
  distinct(`# of People.x.x.x`, .keep_all = TRUE) %>%
  # filter for when 'Source of Income Type' is "Wages and salary"
  filter(`Source of Income Type` == "Wages and salary") %>%
  # create a new variable calculating proportion of
  mutate(Proportion = `# of People.x.x.x`/ sum(`# of People.x.x.x`)) %>%
  # round previous calculation and save as a new variable
  mutate(Proportion_rounded = round(Proportion, 2)) %>%
  select(`Income Level`, `# of People.x.x.x`, `Proportion_rounded`)
```

```
## # A tibble: 11 x 4
## # Groups:   Source of Income Type [1]
##    `Source of Income Type` `Income Level`    `# of People.x.x.~ Proportion_roun~
##    <chr>                   <chr>                          <dbl>            <dbl>
##  1 Wages and salary        $1 to $9,999 or ~              15492              0.1
##  2 Wages and salary        $10,000 to $19,9~              14237              0.09
##  3 Wages and salary        $20,000 to $29,9~              18492              0.12
##  4 Wages and salary        $30,000 to $39,9~              19463              0.12
##  5 Wages and salary        $40,000 to $49,9~              17380              0.11
##  6 Wages and salary        $50,000 to $59,9~              15070              0.09
##  7 Wages and salary        $60,000 to $69,9~              12217              0.08
##  8 Wages and salary        $70,000 to $79,9~               9273              0.06
##  9 Wages and salary        $80,000 to $89,9~               6894              0.04
## 10 Wages and salary        $90,000 to $99,9~               5064              0.03
## 11 Wages and salary        $100,000 and over              25216              0.16
```

Here is the number of people making an income through wages and salary in each income level. This information is also written as a proportion: the number of earners for a specific income level out of the total number of incomes earned through wages and salary.

The income level earned the most number of times is $100,000 or more, making up 16% of all income earned by wages and salary. Meanwhile, the income level earned the least number of times is $90,000-$99,999 making up on 3% of all income earned by wages and salary.

It is interesting to note, 54% of all income earned by wages and salary are between just $1-$49,999.

Categorical and Numerical – count of non high school degree earners of each income level, min, and max

```
all %>%
  # order the data with 'income level' listed first
  group_by(`Income Level`) %>%
  # remove duplicate rows in the variable and keep all data
  distinct(`# of People.y`, .keep_all = TRUE) %>%
  # filter for when 'Source of Income Type' is "Wages and salary"
  filter(`Education Level` == "Less Than 9th Grade"
         | `Education Level` == "9th to 12th nongrad") %>%
  # sum number of people in each income level
  mutate(`# non hs degree` = sum(`# of People.y`)) %>%
  # remove duplicate rows in the variable and keep all data
  distinct(`# non hs degree`, .keep_all = TRUE) %>%
  # only display this column
  select(`# non hs degree`)
```

```
## # A tibble: 11 x 2
## # Groups:   Income Level [11]
##    `Income Level`       `# non hs degree`
##    <chr>                            <dbl>
##  1 $1 to $9,999 or loss              2285
##  2 $10,000 to $19,999                1820
##  3 $20,000 to $29,999                2076
##  4 $30,000 to $39,999                1931
##  5 $40,000 to $49,999                1086
```

```
##  6 $50,000 to $59,999                669
##  7 $60,000 to $69,999                374
##  8 $70,000 to $79,999                221
##  9 $80,000 to $89,999                120
## 10 $90,000 to $99,999                 62
## 11 $100,000 and over                 238
```

Here is the numerical distribution of non high school degree holding individuals across all income levels. From a brief glance, we can see that most make $49,999 or less.

The income level category with the minimum number of people is $90,000-$99,999 with 62,000 people.

The income level category with the maximum number of people is $1-$9,999 or loss with 2,285,000 people.

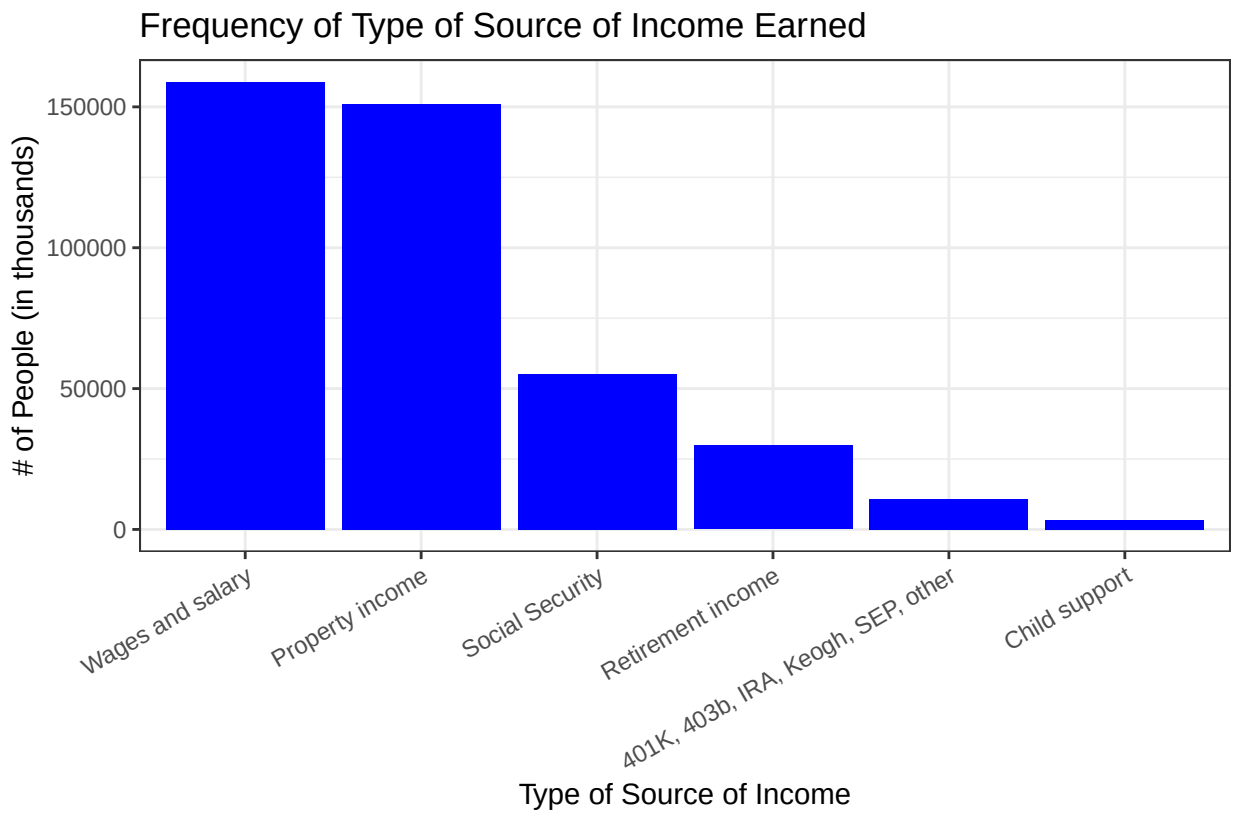
---

5. Visualizing

Completed by Yunseo and Harini and Jack

1. Bar chart of different types of sources of income (one variable)

```
source2 %>%
  # group by types of sources of income
  group_by(`Source of Income Type`) %>%
  # calculate a count for each type in source of income
  summarize(total = sum(`# of People`)) %>%
  # bars in this specific order
  ggplot(aes(x=factor(`Source of Income Type`,
                      level=c("Wages and salary", "Property income",
                              "Social Security", "Retirement income",
                              "401K, 403b, IRA, Keogh, SEP, other",
                              "Child support")),
             y=total)) +
  # keep specified y and color the bars
  geom_bar(stat="identity", fill="blue") +
  labs(title="Frequency of Type of Source of Income Earned",
       caption = "Source: US Census Bureau, 2021",
       x="Type of Source of Income",
       y="# of People (in thousands)") +
  # make y-axis go from 0 to 200000 in increments of 50000
  scale_y_continuous(breaks=seq(0,200000,50000)) +
  theme_bw() +
  # angle labels under each bar
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```

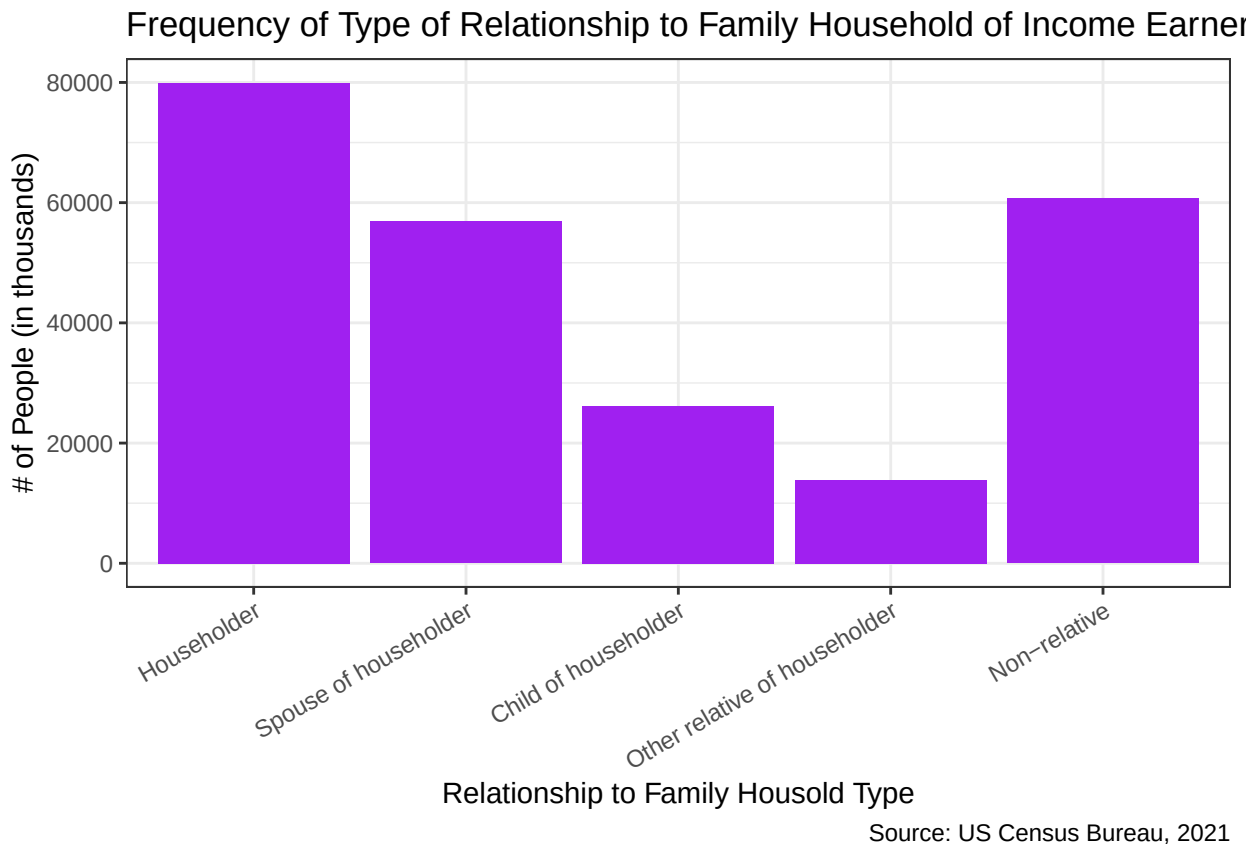


Income earners across all income levels make most of their income from wages and salary followed by property income as the second most frequent method of earning income. On the other hand, income earners receive the least amount of income through child support and retirement accounts like 401K, 403b, IRA, Keough, SEP, and others.

Source where we learned to reorder factor levels: https://www.statology.org/reorder-factor-levels-in-r/

2. Bar chart of different household relations (one variable)

```
relation2 %>%
  # group by types of relations
  group_by(`Relationship to Family Household`) %>%
  # calculate a count for each relation type
  summarize(total = sum(`# of People`)) %>%
  ggplot(aes(x=factor(`Relationship to Family Household`,
                      level=c("Householder", "Spouse of householder",
                              "Child of householder", "Other relative of householder",
                              "Non-relative")),
             y=total)) +
  # keep specified y and color the bars
  geom_bar(stat="identity", fill="purple") +
  labs(title="Frequency of Type of Relationship to Family Household of Income Earners",
       caption = "Source: US Census Bureau, 2021",
       x="Relationship to Family Housold Type",
       y="# of People (in thousands)") +
  # make y axis go from 0 to 100000 in increments of 50000
  scale_y_continuous(breaks=seq(0,100000,20000)) +
  theme_bw() +
  # angle labels under each bar
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```

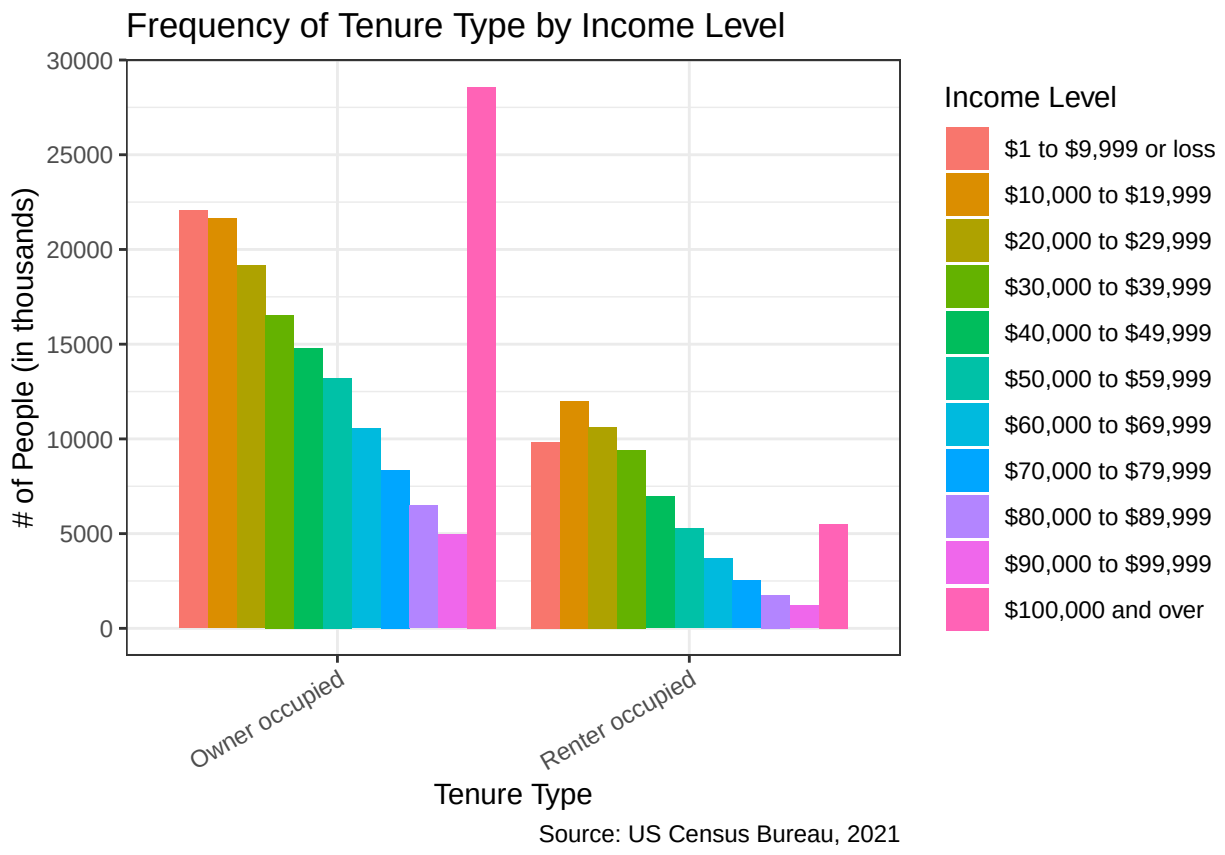


As expected, most income earners are the Householder (about 1/3 of the income earning population). Unex-

pectedly, there is about the same number of Spouse of householder as there are Non-relative (each about 1/4 of the income earning population). There is a category for Child of householder which most likely represents the people age 15-24 who do not work full time but are claimed as dependents and work part time alongside attending school. So the Non-relative demographic may be adult children who are no longer dependents of the householder and simply live in the same domicile as a Householder but operate on their own expenses.

3. Bar chart of tenure type by level of income (two variables)

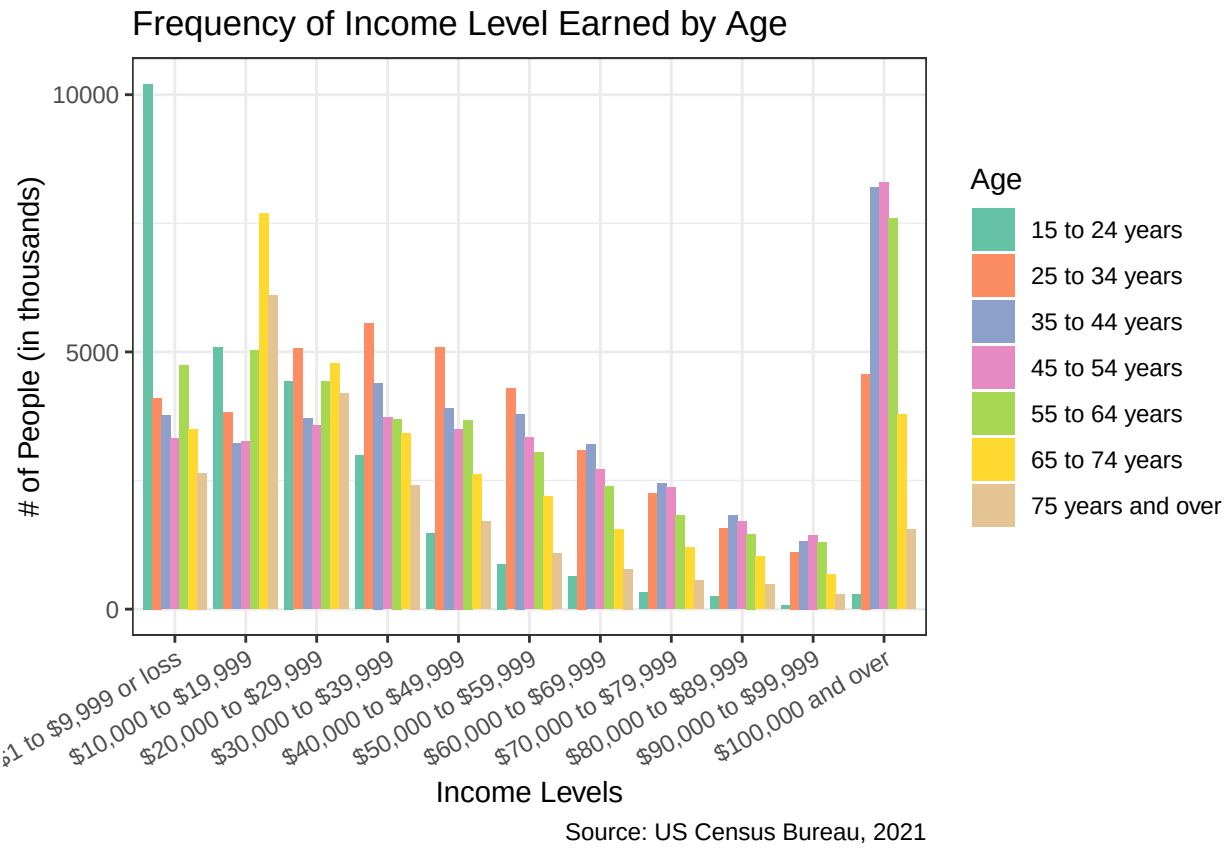
```
tenure2 %>%
  # group by types of tenure and then by income levels
  group_by(`Tenure Type`, `Income Level`) %>%
  # calculate a count for each tenure type and income
  mutate(total = sum(`# of People`)) %>%
  ggplot(aes(x=`Tenure Type`,
             y=total,
             fill=factor(`Income Level`, level=c("$1 to $9,999 or loss",
                                 "$10,000 to $19,999", "$20,000 to $29,999",
                                 "$30,000 to $39,999", "$40,000 to $49,999",
                                 "$50,000 to $59,999", "$60,000 to $69,999",
                                 "$70,000 to $79,999", "$80,000 to $89,999",
                                 "$90,000 to $99,999", "$100,000 and over")))) +
  # keep specified y and stack bars side by side
  geom_bar(stat="identity", position=position_dodge()) +
  labs(title="Frequency of Tenure Type by Income Level",
       caption = "Source: US Census Bureau, 2021",
       x="Tenure Type",
       y="# of People (in thousands)",
       fill="Income Level") +
  scale_y_continuous(breaks=seq(0,30000,5000)) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



The amount of those who own vs rent their residence is proportionate across all income levels expect for those who have an income of $100,000 and over. About 2/3rds of each income level owns their residence instead of renting. However, 5/6th of those who have an income of $100,000 own their residence instead of renting. So if someone is earning \$100,000 or more, it they are more likely to own than rent.

4. Bar chart of level of income by age (two variables)

```
age2 %>%
  # group by types of income levels and then by age
  group_by(`Income Level`, `Age`) %>%
  # calculate a count for each income level and age
  summarize(total = sum(`# of People`)) %>%
  # keep bars in this specific order
  ggplot(aes(x=factor(`Income Level`, level=c("$1 to $9,999 or loss",
                        "$10,000 to $19,999", "$20,000 to $29,999",
                        "$30,000 to $39,999", "$40,000 to $49,999",
                        "$50,000 to $59,999", "$60,000 to $69,999",
                        "$70,000 to $79,999", "$80,000 to $89,999",
                        "$90,000 to $99,999", "$100,000 and over")),
             y=total,
             fill=`Age`)) +
  # keep specified y and stack bars side by side
  geom_bar(stat="identity", position=position_dodge()) +
  labs(title="Frequency of Income Level Earned by Age",
       caption = "Source: US Census Bureau, 2021",
       x="Income Levels",
       y="# of People (in thousands)") +
  # make y axis go from 0 to 10000 in increments of 5000
  scale_y_continuous(breaks=seq(0,15000,5000)) +
  # color theme for the graph
  scale_fill_brewer(palette="Set2") +
  theme_bw() +
  # angle labels under each bar
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



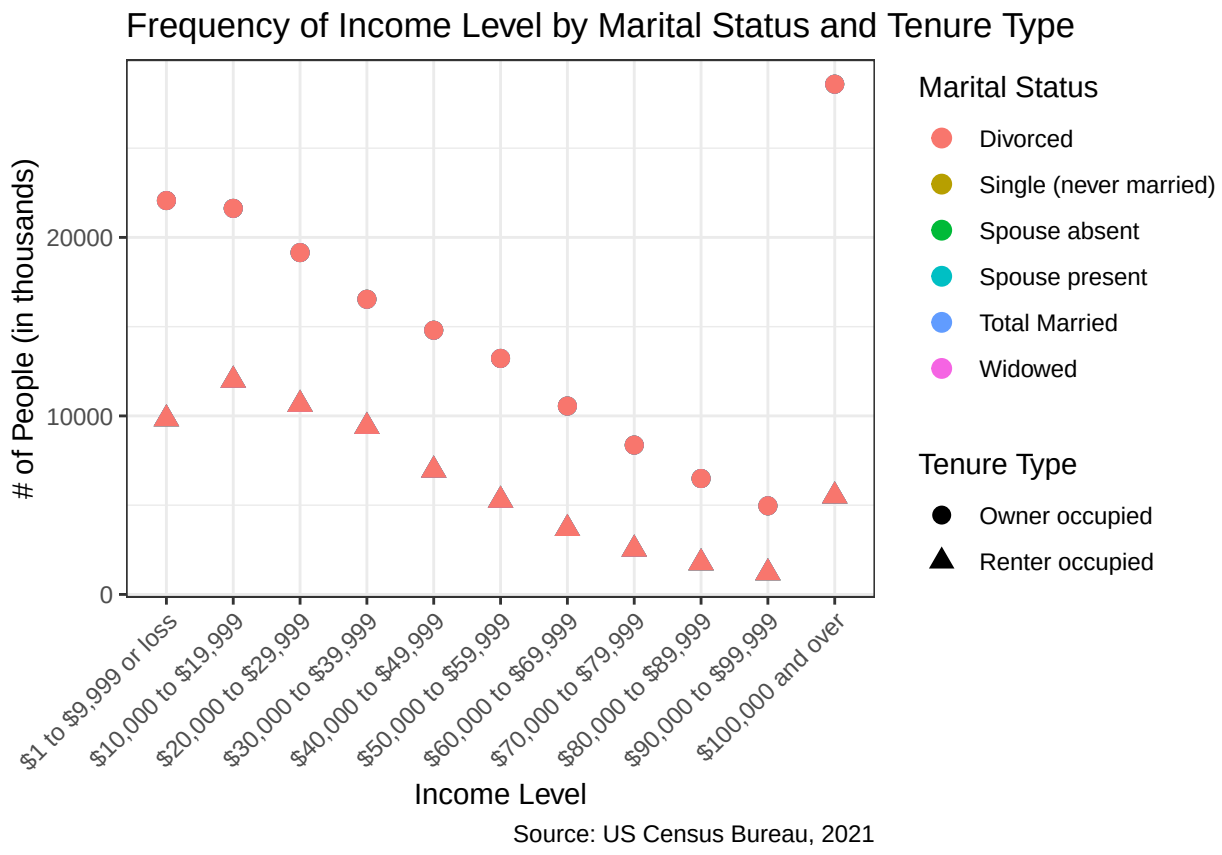
Source: US Census Bureau, 2021

As expected, the majority of 15-24-year-olds make \$1-\$9,999 a year. A large proportion of each age bracket makes \$100,000 or more. Excluding those two skews, each age bracket makes \$30,000-\$39,000 most often.

As expected, the majority of 15-24-year-olds made \$1-\$9,999 or a loss a year. A large proportion of each age bracket made \$100,000 or more. Specifically, those in ages 35-44, 45-55, and 55-64 years made had the most wages and salaries of \$100,00 or more. However, excluding those two skews, the data follows a slightly right skewed distribution centered at around \$40,000-\$49,999. So people of all age levels made *most often* \$30,000-\$39,999 and *on average* \$50,000-\$59,999.

5. Income level by marital status by tenure type (three variables)

```
tenure2 %>%
  # join data sets
  full_join(marital2, by="Income Level") %>%
  # group by the following columns
  group_by(`Tenure Type`, `Marital Status`, `Income Level`) %>%
  # scatter plot
  ggplot(aes(x=factor(`Income Level`, level=c("$1 to $9,999 or loss",
                        "$10,000 to $19,999", "$20,000 to $29,999",
                        "$30,000 to $39,999", "$40,000 to $49,999",
                        "$50,000 to $59,999", "$60,000 to $69,999",
                        "$70,000 to $79,999", "$80,000 to $89,999",
                        "$90,000 to $99,999", "$100,000 and over"))
             , y=`# of People.x`, shape=`Tenure Type`, color=`Marital Status`)) +
  # keep specified y and set point size
  geom_point(stat="identity", size=3) +
  labs(title="Frequency of Income Level by Marital Status and Tenure Type",
       caption = "Source: US Census Bureau, 2021",
       x="Income Level",
       y="# of People (in thousands)") +
  # color theme for the graph
  theme_bw() +
  # angle x-axis labels
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



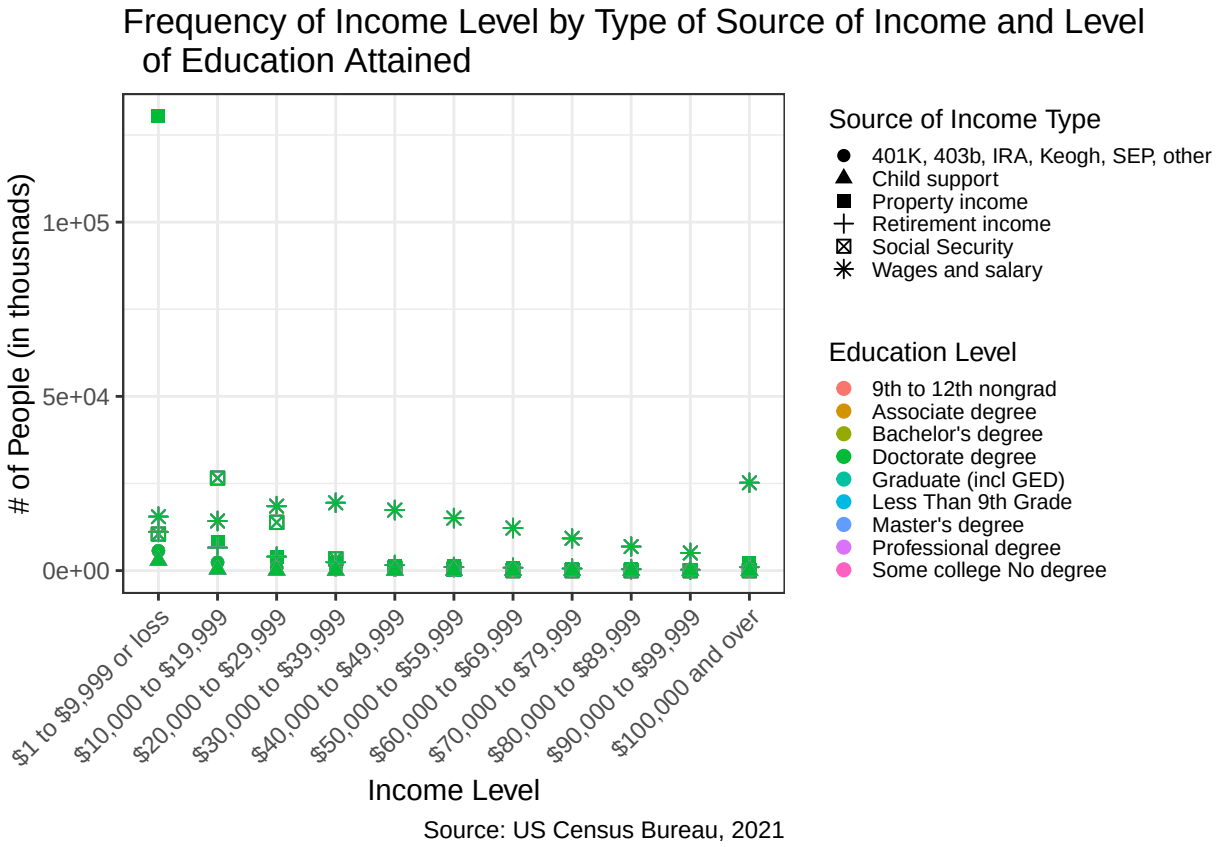
Given the nature of our data (none of the data sets had other variables in common apart from income levels), it was not possible to make an effective visualization demonstrating the relationship between three variables.

However, in the case that our data had more shared information, this is what it would look most like. We would map the frequency of a certain income level for each income level and color the point by marital status and change the shape of the points by marital status. Here we only see one color for the points because we do not have the information on the tenure type of each marital status.

If this graph could give us our desired output, it would build upon the findings in our 3rd graph and show us if a certain marital status owned more than rented and if a certain marital status contributed to the high frequency of people make an income of $100,000 or more and owning their domicile.

6. Income level by source of income by education level (three variables)

```
source2 %>%
  # join data sets
  full_join(edu2, by="Income Level") %>%
  # group by the following columns
  group_by(`Source of Income Type`, `Education Level`, `Income Level`) %>%
  # scatter plot
  ggplot(aes(x=factor(`Income Level`, level=c("$1 to $9,999 or loss",
                        "$10,000 to $19,999", "$20,000 to $29,999",
                        "$30,000 to $39,999", "$40,000 to $49,999",
                        "$50,000 to $59,999", "$60,000 to $69,999",
                        "$70,000 to $79,999", "$80,000 to $89,999",
                        "$90,000 to $99,999", "$100,000 and over"))
             , y=`# of People.x`, shape=`Source of Income Type`, color=`Education Level`)) +
  # keep specified y and set point size
  geom_point(stat="identity", size=2) +
  # color theme for the graph
  scale_fill_brewer(palette="Paired") +
  labs(title="Frequency of Income Level by Type of Source of Income and Level
  of Education Attained",
       caption = "Source: US Census Bureau, 2021",
       x="Income Level",
       y="# of People (in thousnads)") +
  # color theme for the graph
  theme_bw() +
  # adjust legend location and size
  #theme(legend.position = "bottom") +
  theme(legend.key.size = unit(2, 'cm'), #change legend key size
        legend.key.height = unit(0.3, 'cm'), #change legend key height
        legend.key.width = unit(0.4, 'cm'), #change legend key width
        legend.title = element_text(size=10), #change legend title font size
        legend.text = element_text(size=8)) + #change legend text font size
  # angle x-axis labels
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Given the nature of our data (none of the data sets had other variables in common apart from income levels), it was not possible to make an effective visualization demonstrating the relationship between three variables.

However, in the case that our data had more shared information, this is what it would look most like. We would map the frequency of a certain income level for each income level and color the point by education level and change the shape of the points by the type of source of income. Here we only see one color for the points because we do not have the information on the level of education of each type of source of income.

If this graph could give us our desired output, we could confirm if a certain level of education led people to earn more income through a certain source and how much more they earned through those mediums opposed to their counterparts.

Source where we learned to adjust legend size: https://www.statology.org/ggplot2-legend-size/

---

6. Formatting