Challenge_7: Concepts and Practices of Research Design for a Data Science Project

AUTHOR
Mehak Nargotra

PUBLISHED
April 26, 2024

Make sure you change the author's name in the above YAML header.

Setup

If you have not installed the following packages, please install them before loading them.

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.3     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.4     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.0
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(readxl)
library(haven) #for loading other datafiles (SAS, STATA, SPSS, etc.)
library(stringr) # if you have not installed this package, please install it.
library(ggplot2) # if you have not installed this package, please install it.
```

Challenge Overview

In this challenge, we will apply the knowledge about research design and other topics covered in lectures so far to the dataset presented.

There will be coding components and writing components. Please read the instructions for each part and complete your challenges.

Part 1. Choose one of the following datasets to do a simple practice of research design and hypothesis testing (50%)

Dataset 1: The General Social Survey (2022). You can find more information about this data project at https://gss.norc.org/About-The-GSS. A codebook explaining the definition of each variable and column is also included.

Dataset 2: The Covid-19 Reports in Massachusetts. The datasets are stored in an Excel file of multiple sheets. You can find more information about this data project in the “Introduction”, “Definition”, “Notes”, and “Data Dictionary” tabs in the Excel file.

1. **Read the data you choose in R. (5%)**

   For GSS, there is only one data sheet (.dta).

   For the MA Covid-19 reports, you can choose **one of the four datasheets(tabs in Excel)** to read (“Weekly Cases and Deaths”, “Case and Death Demographics”, “County Data”, and “City and Town Data”).

```
#type your code here
gss_data <- read_dta("~/Desktop/DACSS 601/DACSS_601_datasets/GSS2022.dta")

head(gss_data)
```

year	id	wrkstat	hrs1	hrs2	evwork	wrkslf	occ10	prestg1
<dbl+lbl>	<dbl>	<dbl+lbl>	<dbl+lbl>	<dbl+lbl>	<dbl+lbl>	<dbl+lbl>	<dbl+lbl>	<dbl+lb
2022	1	1	40	NA(i)	NA(i)	2	430	3
2022	2	5	NA(i)	NA(i)	1	2	50	5
2022	3	1	52	NA(i)	NA(i)	2	4610	4
2022	4	3	NA(i)	25	NA(i)	2	4120	3
2022	5	8	NA(i)	NA(i)	1	2	7330	3
2022	6	1	50	NA(i)	NA(i)	2	4610	4

6 rows | 1-9 of 879 columns

2. **Answer the following questions.**

   (1) what is the structure (dimension) of the data? **(2.5%)**

```
#type your code here
print('GSS DATA')
```

```
[1] "GSS DATA"
```

```
dim(gss_data)
```

```
[1] 3544  879
```

```
print('Number of rows :')
```

```
[1] "Number of rows :"
```

```
print(dim(gss_data)[1])
```

```
[1] 3544
```

```
print('Number of columns : ')
```

```
[1] "Number of columns : "
```

```
print(dim(gss_data)[2])
```

```
[1] 879
```

(2) what is the unit of observation? **(2.5%)** Answer: For the GSS data, each row represents every unique respondent and their details, who participated in the survey. These respondents answered question related to attitudes, demographic characteristics and behaviors in the survey by providing unique response.

3. **Read the overview introduction, codebook (for the GSS data), and other related information about the data (for the Covid-19 data). Now browse the data loaded in R, it seems like there are many different questions this data can answer. Based on the class lecture and KKV's reading about "good research questions", please propose ONE research question that can be answered using this data. (5%)** Answer: Research Question: What is the average income of respondents belonging to a specific race?

4. **Based on the research question you proposed above, propose a hypothesis about a possible relationship between two items. (5%)** Answer: Hypothesis: If the race of the respondent is White, then their average income is expected to be higher compared to respondents of Black or other races.

5. **Based on the hypothesis proposed, please select variables/columns in the data to measure the corresponding concepts in the hypothesis statement. You should select at least one variable/column to measure each concept.**

   You should also specify which variables/columns you choose and explain why they are the proper ones to measure the concepts. (10%)

   Instruction: Don't just answer, "They are reliable and valid". Instead, you should discuss more why they are reliable (can consistently produce the same results regardless of the same results regardless different times and contexts) and valid (why it is better than other possible or alternative variables/columns). You can find the concepts of validity and reliability in the Nov 20 lecture and the slides (p23-25). There are also more in-depth introductions online, such as this page.

rincome - The values of rincome varies from 1 to 20 which represents different levels of income starting from below 1000 dollars to more than 170k dollars. Since, we are comparing the income of a particular race according to our hypothesis, having these numerical values will help us to easily determine the which race has a higher expected income. Hence, this column is valid for testing our hypothesis. race - The values of race varies from 1 to 3. 1 is white, 2 is black and 3 is other. The different values of race tells us how we can group the different groups of respondents based on their race which will help us to determine their relationship with the rincome.

6. **Use the code we learned in the previous week to conduct descriptive statistics for the two variables/columns you selected above. You should present the following information**

in your descriptive statistics: range, average, standard deviation, the number of NAs, and the number of unique values. (5%)

```
#type your code here
gss_statistics_rincome <- gss_data %>%
    summarise(
    Range = max(rincome, na.rm = TRUE) - min(rincome, na.rm = TRUE),
    Average = mean(rincome, na.rm = TRUE),
    SD = sd(rincome, na.rm = TRUE),
    NAs = sum(is.na(rincome)),
    Unique_Values = n_distinct(rincome)
      )
  print(gss_statistics_rincome)
```

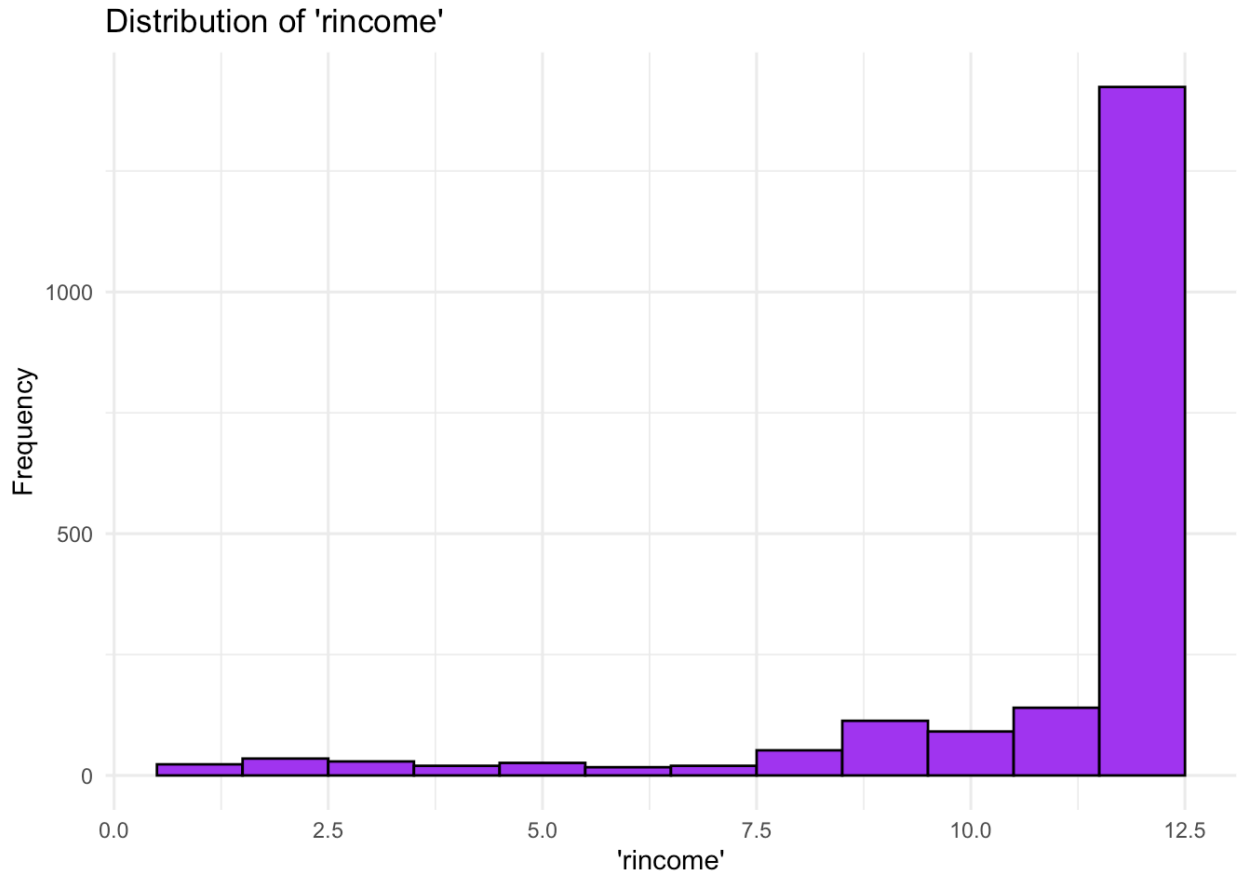
```
# A tibble: 1 × 5
  Range Average    SD   NAs Unique_Values
  <dbl>   <dbl> <dbl> <int>         <int>
1    11    10.9  2.46  1554            13
```

```
gss_statistics_race <- gss_data %>%
    summarise(
    Range = max(race, na.rm = TRUE) - min(race, na.rm = TRUE),
    Average = mean(race, na.rm = TRUE),
    SD = sd(race, na.rm = TRUE),
    NAs = sum(is.na(race)),
    Unique_Values = n_distinct(race)
      )
  print(gss_statistics_race)
```

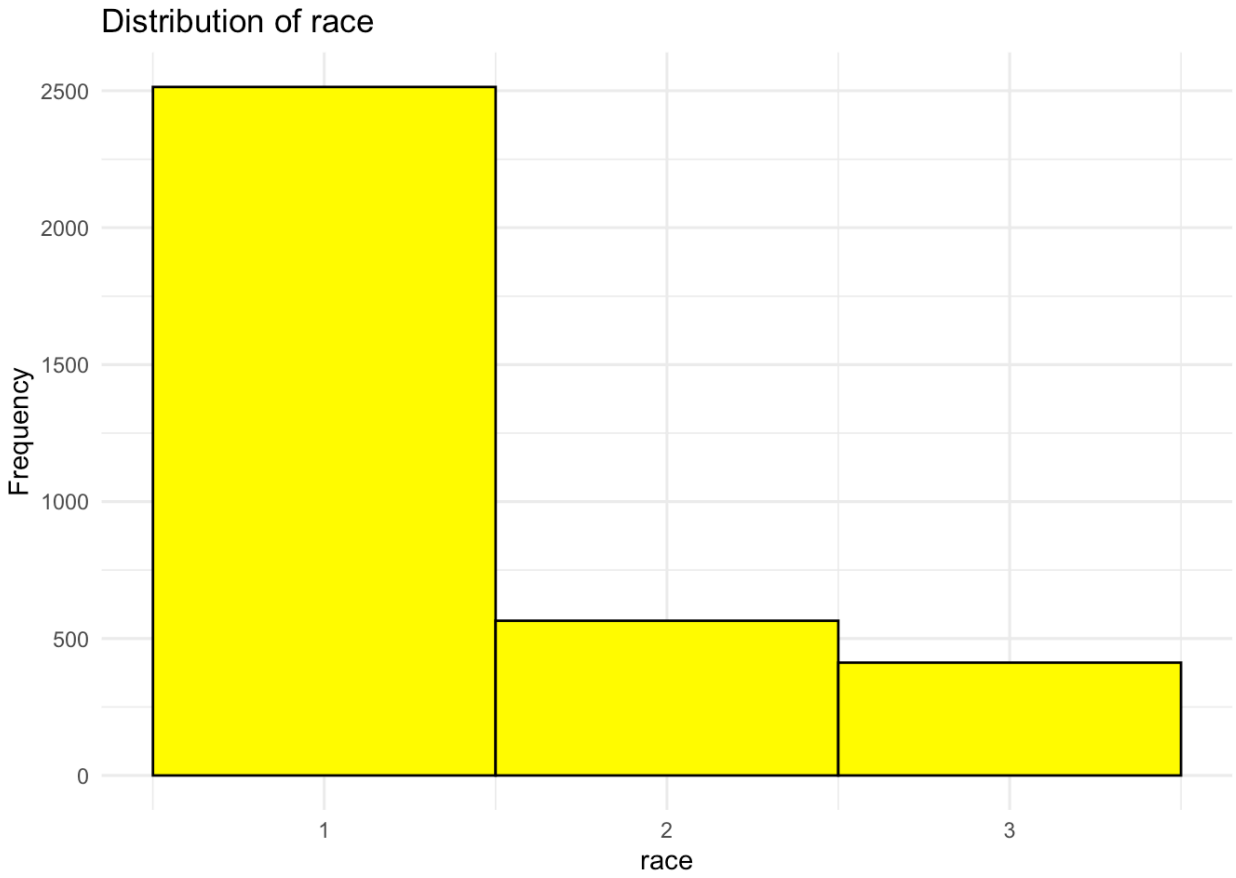
```
# A tibble: 1 × 5
  Range Average    SD   NAs Unique_Values
  <dbl>   <dbl> <dbl> <int>         <int>
1     2    1.40 0.690    53             4
```

7. **Plot one univariate graph for each of the variables/columns. (5%)**

```
#type your code here
gss_data %>%
  filter(!is.na(rincome) & !is.infinite(covid12)) %>%
  ggplot(aes(x = rincome)) +
  geom_histogram(binwidth = 1, fill = "purple", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of 'rincome'", x = "'rincome'", y = "Frequency")
```



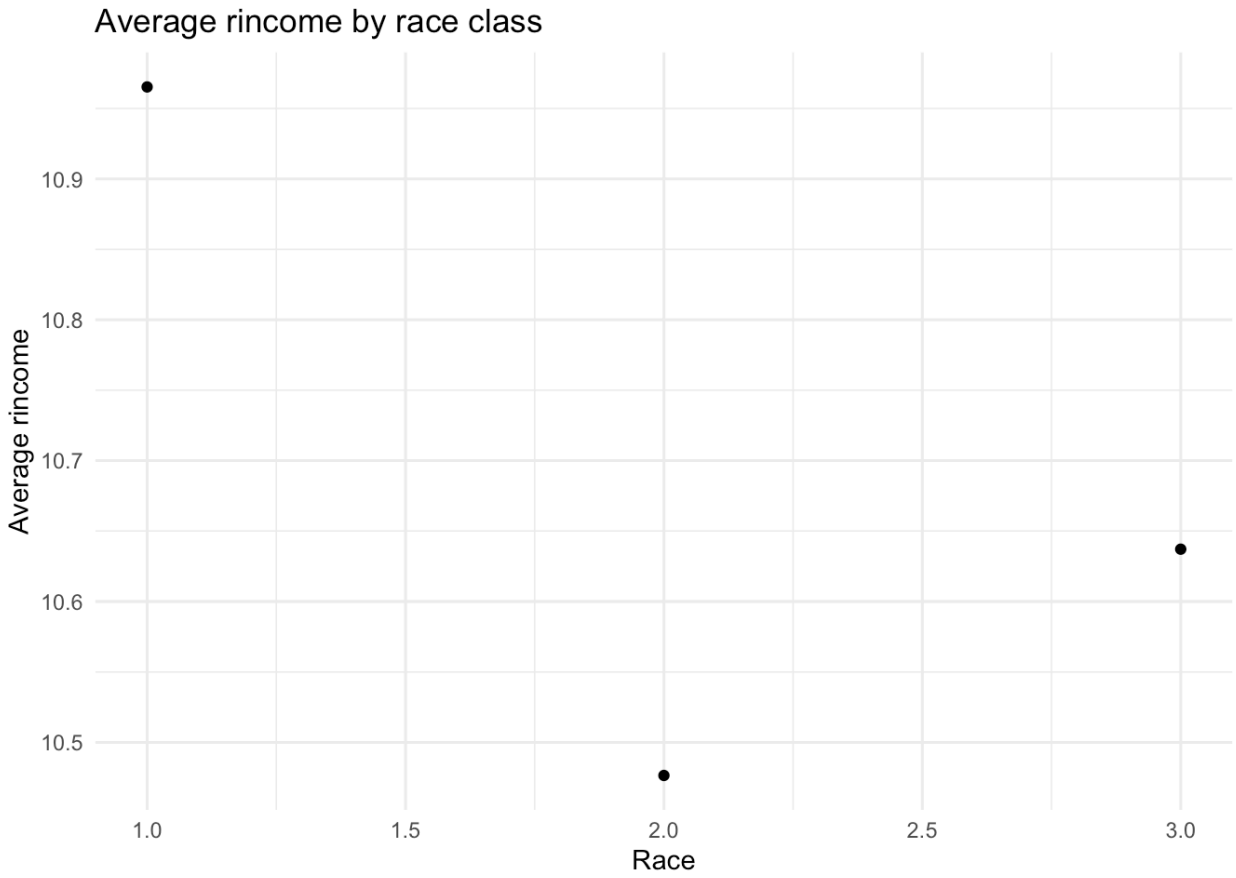

```r
gss_data %>%
  filter(!is.na(race) & !is.infinite(race)) %>%
  ggplot(aes(x = race)) +
  geom_histogram(binwidth = 1, fill = "yellow", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of race", x = "race", y = "Frequency")
```



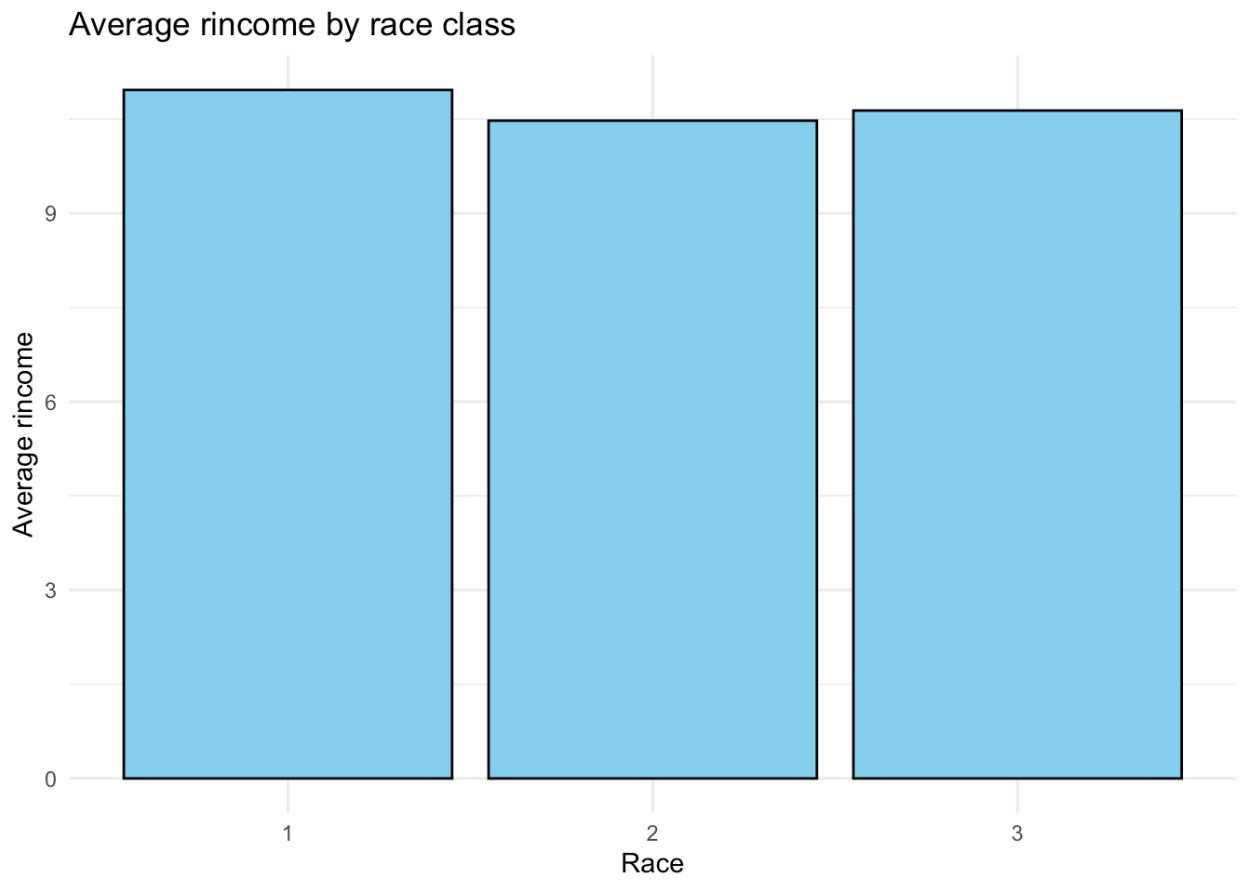
8. **Finally, plot a graph to visually test the hypothesis you propose. Based on the visual evidence, do you see any potential correlation between the two variables? (10%)**

```
avg_rincome <- gss_data %>%
  filter(!is.na(rincome) & !is.na(race) & !is.infinite(rincome) & !is.infinite(r
  group_by(race) %>%
  summarise(avg_rincome = mean(rincome))

# Plot bivariate graph
ggplot(avg_rincome, aes(x = race, y = avg_rincome)) +
  geom_point() +
  labs(title = "Average rincome by race class", x = "Race", y = "Average rincome
  theme_minimal()
```




```
# Plot bivariate bar graph
ggplot(avg_rincome, aes(x = factor(race), y = avg_rincome)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "Average rincome by race class", x = "Race", y = "Average rincome
  theme_minimal()
```



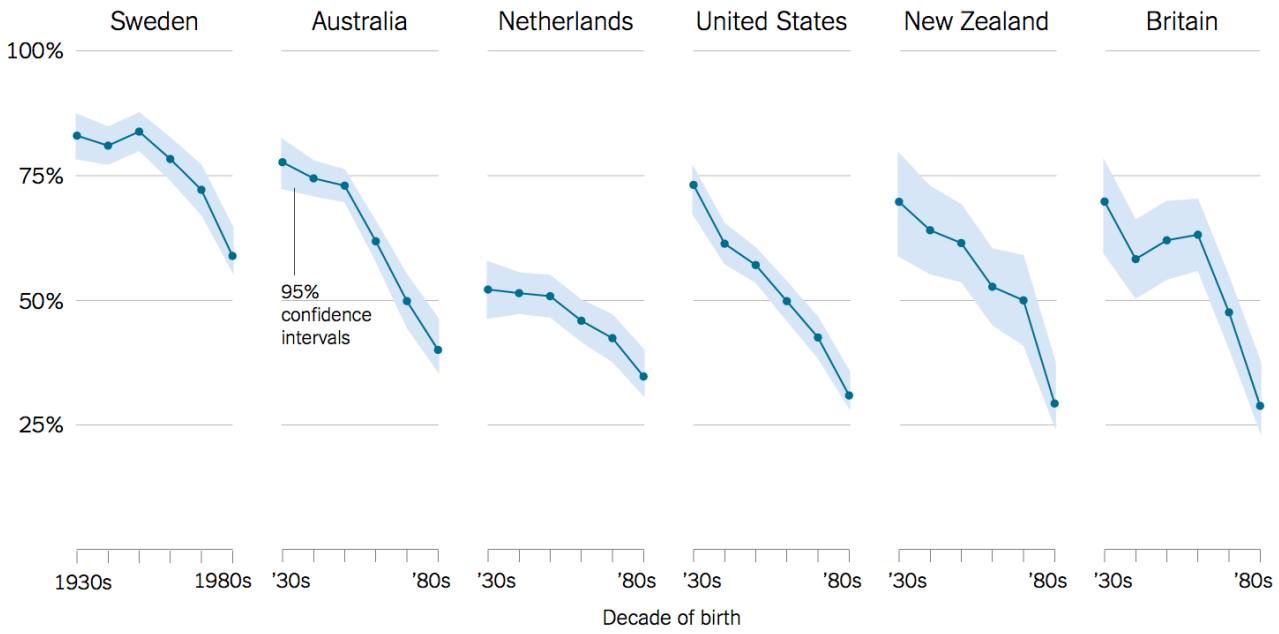

```
print('There is correlation observed between the 2  variables. As we can see from th
```

```
[1] "There is correlation observed between the 2  variables. As we can see from the
graphs, the average of white race (1) is more than black race (2) and the average of
other race is also more than black race. This satisfies our hypothesis. We also have
a point graph to show this relation."
```

Part 2. Reviewing the findings of a graph by examining the raw data. (50%)

This part of the challenge is based on a scenario. Suppose you are a data scientist who provides consulting services to the government. One day, your client asks you to investigate an article by the New York Times that reported on some research on people's confidence in the institutions of democracy. It had been published in an academic journal. The headline in the Times ran, "How Stable Are Democracies? 'Warning Signs Are Flashing Red'" (Taub, 2016). The graph accompanying the article, as shown below, plots people's responses to a question in the World Value Survey (WVS) (V162-Importance of democracy). The graph certainly seemed to show an alarming decline. The graph was widely circulated on social media. It's an elegant small-multiple that, in addition to the point ranges it identifies, also shows an error range (labeled as such for people who might not know what it is), and the story told across the panels for each country is pretty consistent.

Percentage of people who say it is "essential" to live in a democracy



Source: Yascha Mounk and Roberto Stefan Foa, "The Signs of Democratic Deconsolidation," Journal of Democracy | By The New York Times

1. **Please briefly describe the major findings of this graph. (5%)** Answer : The graph shows that the percentage of people who consider living in a democracy as essential from the 1930s to the 1980s across six countries. In Sweden, approx. 80% of people feel that it is important to live in democracy in the 1930s, which was the highest among all countries. However, this percentage declined until the 1940s, followed by an increase until the 1950s, and then a continuous decline to around 55-60% by the 1980s. Australia also experienced a steady decline in the percentage from the 1930s, starting at around 77-78%. The decline was less steep until the 1950s but became more steep afterwards, reaching almost 40% by the end of the 1980s. In the Netherlands, there was a slight decline from close to 50% in the 1930s, which looked constant until the 1950s. However, there was a decline observed from the 1950s onwards, with the percentage dropping to almost 35-40% by the 1980s. The United States saw a sharp decline in the percentage from the 1930s to the 1940s, followed by a gradual decline until the 1980s, where almost 30% of people believed in the essentialness of democracy. In New Zealand, there was an overall decline from around 70% in the 1930s to almost 27% in the 1980s. The decline was particularly steep from the 1970s to the 1980s compared to other periods. This shows that people started thinking a bit differently during that decade. Finally, in Britain, the percentage of people who believed that it was essential to live in a democracy started around 70% in the 1930s, dropping to almost 60% by the 1940s. There was then an increase from the 1940s to the 1960s to almost 65%, followed by a steep decline from the 1960s to the 1980s, reaching the lowest point at 28%.

2. **Your client is concerned about the findings of this graph.** On the one hand, they are surprised and worried by the "crisis of democracy" presented in this graph.**On the other hand, they also doubt the argument of the NYT article and the validity of the findings of this graph.** Before deciding on making any policy to respond, they ask you to conduct some additional research with the original data.

(1) Read the provided WVS data. The dataset is large, so you must subset it before analyzing it. **Please keep only the following columns: respondents' country(V2), age(V236), and the question for plotting (V162).** You also need to filter only the observations in the six countries mentioned above: Sweden, Australia, Netherlands, United States, New Zealand, and Britain/United Kingdom. **(10%)**

Note: all the columns, including those that are measured categorically, are represented by numbers. You must check out the WVS5 codebook to identify what the numerical values mean (especially for V2-country, see p57 of the codebook).

```
#type your code here
wvs_data <- readRDS("~/Desktop/DACSS 601/DACSS_601_datasets/WVS5.rds")
wvs_subset <- wvs_data %>%
    select(V2, V236, V162)
countr_codes <- c(752, 36, 528, 840, 554, 826)
wvs_subset_updated <- wvs_subset %>%
  filter(V2 %in% countr_codes)

wvs_subset_updated
```

V2 <labelled>	V236 <labelled>	V162 <labelled>
36	1921	10
36	1939	10
36	1954	10
36	1947	10
36	1965	9
36	1980	4
36	1934	10
36	1915	10
36	1959	10
36	1960	10

1-10 of 6,718 rows Previous **1** 2 3 4 5 6 … 672 Next

(2) Conduct descriptive statistics to show these three columns' unique values, means, ranges, and numbers of NA. You can plot univariate graphs as we did in challenge#4 or apply the summary statistics function as in challenge#3. Just do either approach. **(10%)**

```
#type your code here

unique_values_V2 <- length(unique(wvs_subset_updated$V2))
unique_values_V236 <- length(unique(wvs_subset_updated$V236))
unique_values_V162 <- length(unique(wvs_subset_updated$V162))

mean_V2 <- mean(wvs_subset_updated$V2, na.rm = TRUE)
mean_V236 <- mean(wvs_subset_updated$V236, na.rm = TRUE)
mean_V162 <- mean(wvs_subset_updated$V162, na.rm = TRUE)
```

```
range_V2 <- range(wvs_subset_updated$V2, na.rm = TRUE)
range_V236 <- range(wvs_subset_updated$V236, na.rm = TRUE)
range_V162 <- range(wvs_subset_updated$V162, na.rm = TRUE)

na_count_V2 <- sum(is.na(wvs_subset_updated$V2))
na_count_V236 <- sum(is.na(wvs_subset_updated$V236))
na_count_V162 <- sum(is.na(wvs_subset_updated$V162))

statistics_summary <- data.frame(
  Column = c("V2", "V236", "V162"),
  Unique_Values = c(unique_values_V2, unique_values_V236, unique_values_V162),
  Mean = c(mean_V2, mean_V236, mean_V162),
  Range = c(paste(range_V2, collapse = " - "), paste(range_V236, collapse = " -
  NA_Count = c(na_count_V2, na_count_V236, na_count_V162)
)

print(statistics_summary)
```

```
  Column Unique_Values       Mean     Range NA_Count
1     V2             6  565.25067  36 - 840        0
2   V236            80 1946.52560 -2 - 1991        0
3   V162            14    6.87139   -5 - 10        0
```

(3) (Optional) Please replicate the graph of the NYT article.

```
#type your code here
```

(4) Now, please plot a graph to show the relationship between the decades of birth (x-axis) and the average level of the response scores to the question “importance of democracy” (y-axis) for each of the six countries. You can use facet_grid or facet_wrap to combine multiple graphs into a matrix of panels. **(15%)**

```
#type your code here
  average_response_scores <- wvs_subset_updated %>%
  group_by(V2, Decade = cut(V236, breaks = seq(1900, 2010, by = 10), labels = se
  summarise(Avg_Score = mean(V162, na.rm = TRUE))
```

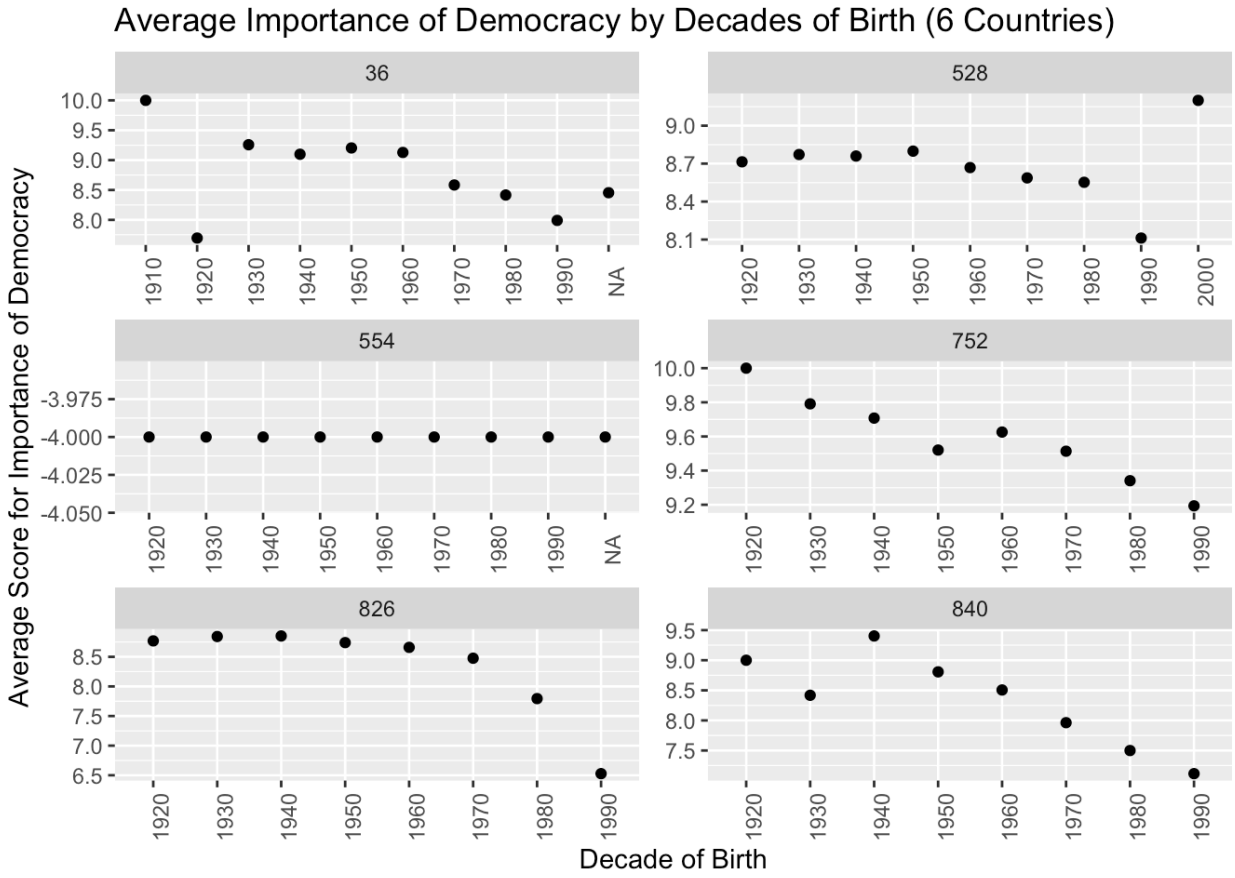
```
`summarise()` has grouped output by 'V2'. You can override using the `.groups`
argument.
```

```
plot <- ggplot(average_response_scores, aes(x = Decade, y = Avg_Score)) +
  geom_line() +
  geom_point() +
  facet_wrap(~ V2, scales = "free", nrow = 3) +  # Facet by country with 2 rows,
  labs(title = "Average Importance of Democracy by Decades of Birth (6 Countries
       x = "Decade of Birth",
       y = "Average Score for Importance of Democracy") +

  theme(axis.text.x = element_text(angle = 90))

print(plot)
```

```
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```



3. **Describe what you find from the graph you made above. Compared to the graph on NYT, what's in common, or what's different? Please type your answer below. (5%)** Answer: The graph shows how people from six different countries feel about democracy over time. It is different from the graph in the NYT because it shows average responses on a scale from 1 to 10 or -1 to -5. In Australia, many people support democracy, but their believe in democracy fluctuates between up and down over the years. In the Netherlands, the average response increases a little, then decreases from the 1920s to the 1990s, and suddenly increases again in the 2000s. For New Zealand, the average response is mostly negative. In Sweden, people's feelings about democracy go down from the 1920s to the 1950s, then up from the 1950s to the 1960s, and down again until the 1990s. In the UK, people's belief seem steady from the 1920s to the 1960s, then drop slowly from the 1960s to the 1980s, and then a lot from the 1980s to the 1990s. If we look at the United States, people feel worse about democracy from the 1920s to the 1930s, better from the 1930s to the 1940s, and then worse again from the 1940s to the 1990s.

4. **Your client wants to hear your conclusion. Do you agree with the argument presented by the graph and the NYT article? Should we really worry about the decline? This is an op-ed question. Please type your answer below. (5%)** Answer: No, we can't agree with the argument in the NYT article because we don't know how the percentage was calculated. Without understanding the methodology and calculations behind the percentage calculation, we can't draw conclusions or worry about any decline. Hence, the whole argument is ambiguous.