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

AUTHOR
Muskan Dhar

PUBLISHED
July 5, 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.

```r
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.4.4     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.1
✔ 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
```

```r
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”).

```
covid_county <- read_excel("covid-19-dashboard-11-16-23.xlsx",sheet='County Data')
head(covid_county)
```

Season <chr>	Week Start Date <dttm>	Week End Date <dttm>	County <chr>
2023-2024	2023-07-02	2023-07-08	Barnstable
2023-2024	2023-07-02	2023-07-08	Berkshire
2023-2024	2023-07-02	2023-07-08	Bristol
2023-2024	2023-07-02	2023-07-08	Dukes and Nantucket
2023-2024	2023-07-02	2023-07-08	Essex
2023-2024	2023-07-02	2023-07-08	Franklin

6 rows | 1-4 of 14 columns

2. **Answer the following questions.**

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

```
dim(covid_county)
```

```
[1] 285  14
```

   (2) what is the unit of observation? **(2.5%)**

   Each week for each county is the unit of observation.

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%)**

Is there a correlation between weekly testing rates and weekly COVID-19 case rates in counties over the duration of the pandemic?

4. **Based on the research question you proposed above, propose a hypothesis about a possible relationship between two items. (5%)**

Higher weekly testing rates in different counties are associated with more weekly COVID-19 case rates during the same periods, indicating that counties are able to detect more new cases efficiently.

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.

Weekly testing rate:

Reliability: It is consistently computed as the number of tests performed per 100,000 people each week, allowing for comparable and consistent results across different counties and times. Validity: It directly quantifies the intensity of the testing efforts in each county, relative to its population size. This is critical for understanding how widespread and proactive testing is.

Weekly case rate:

Reliability: The weekly case rate is calculated in a standardized way, allowing for reliable comparisons over time and across different population sizes. Validity: This variable is valid for measuring the impact of testing on controlling the virus spread because it reflects the outcome of interest, ie the number of new infections.

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%)**

```
covid_county$`Weekly testing rate` <- as.numeric(as.character(covid_county$`Week
```

```
Warning: NAs introduced by coercion
```

```
covid_county$`Weekly case rate` <- as.numeric(as.character(covid_county$`Weekly
```

```
Warning: NAs introduced by coercion
```

```
testing_stats <- list(
  Range = max(covid_county$`Weekly testing rate`, na.rm = TRUE) - min(covid_coun
  Average = mean(covid_county$`Weekly testing rate`, na.rm = TRUE),
  Standard_Deviation = sd(covid_county$`Weekly testing rate`, na.rm = TRUE),
  Number_of_NAs = sum(is.na(covid_county$`Weekly testing rate`)),
  Number_of_Unique_Values = length(unique(covid_county$`Weekly testing rate`))
```

```r
)

case_stats <- list(
  Range = max(covid_county$`Weekly case rate`, na.rm = TRUE) - min(covid_county$
  Average = mean(covid_county$`Weekly case rate`, na.rm = TRUE),
  Standard_Deviation = sd(covid_county$`Weekly case rate`, na.rm = TRUE),
  Number_of_NAs = sum(is.na(covid_county$`Weekly case rate`)),
  Number_of_Unique_Values = length(unique(covid_county$`Weekly case rate`))
)

print(" Statistics for Weekly Testing Rate:")
```

```
[1] " Statistics for Weekly Testing Rate:"
```

```r
print(testing_stats)
```

```
$Range
[1] 818.4465

$Average
[1] 365.2709

$Standard_Deviation
[1] 158.8836

$Number_of_NAs
[1] 19

$Number_of_Unique_Values
[1] 265
```

```r
print(" Statistics for Weekly Case Rate:")
```

```
[1] " Statistics for Weekly Case Rate:"
```

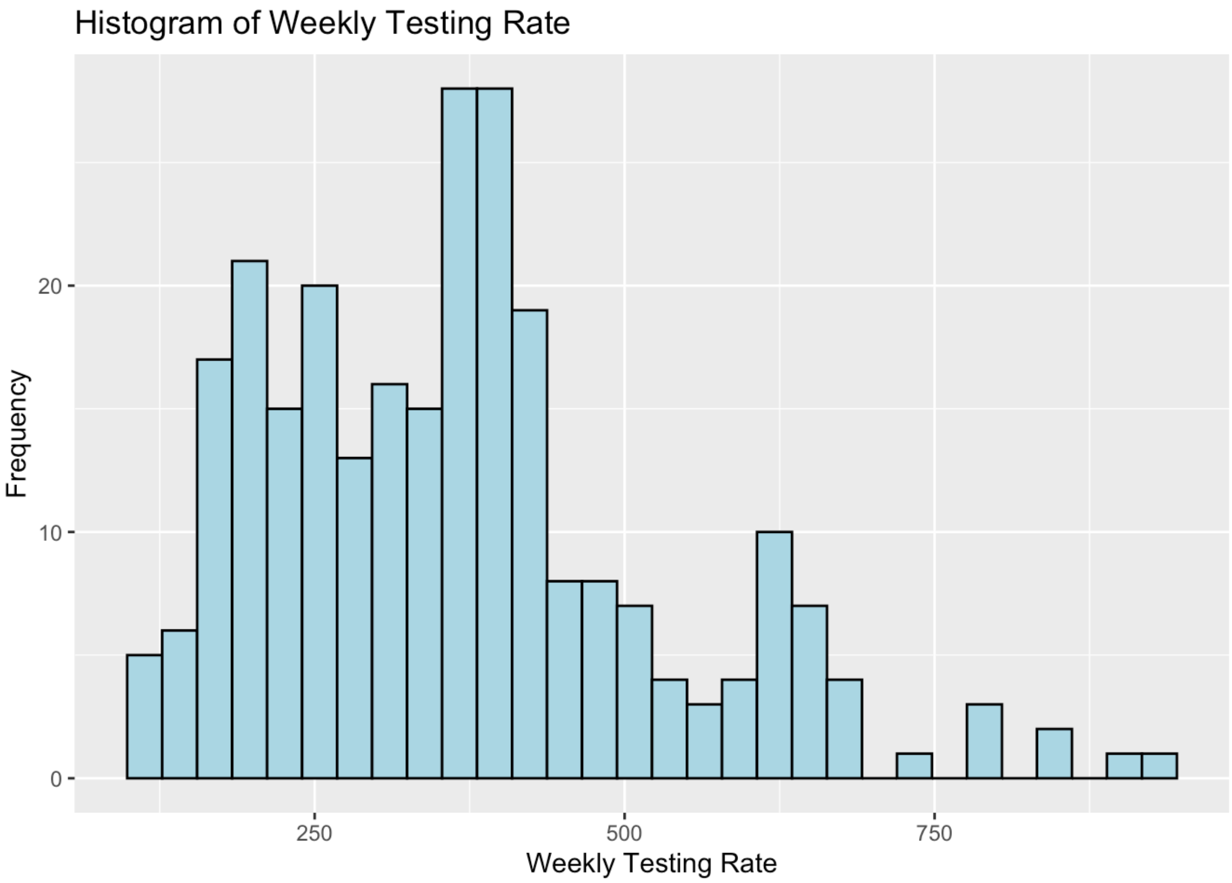
```r
print(case_stats)
```

```
$Range
[1] 76.21738

$Average
[1] 28.173

$Standard_Deviation
[1] 13.9517

$Number_of_NAs
[1] 19

$Number_of_Unique_Values
[1] 244
```

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

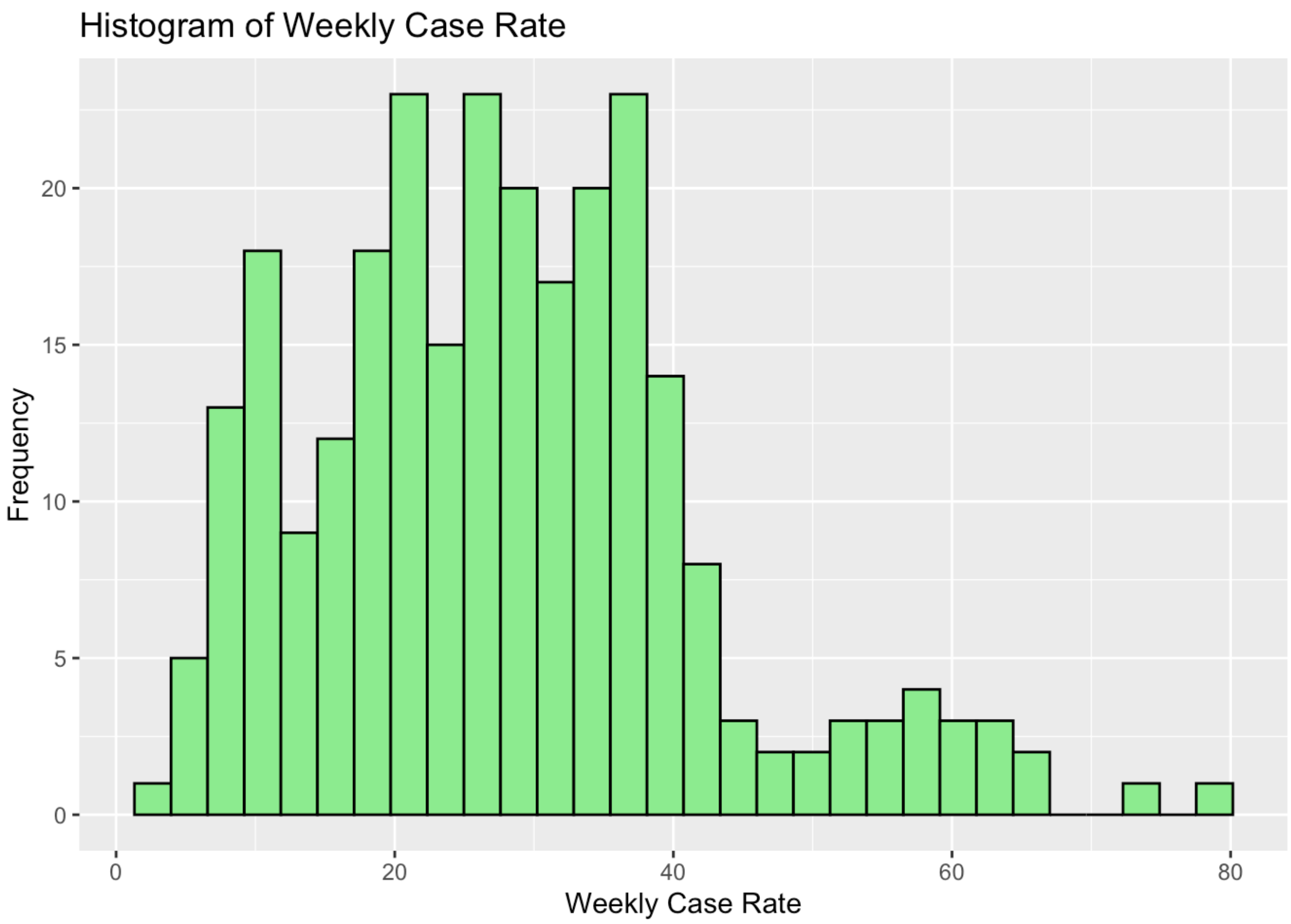
```
ggplot(covid_county, aes(x = `Weekly testing rate`)) +
  geom_histogram(bins = 30, fill = "lightblue", color = "black") +
  ggtitle("Histogram of Weekly Testing Rate") +
  xlab("Weekly Testing Rate") +
  ylab("Frequency")
```

```
Warning: Removed 19 rows containing non-finite values (`stat_bin()`).
```



```
ggplot(covid_county, aes(x = `Weekly case rate`)) +
  geom_histogram(bins = 30, fill = "lightgreen", color = "black") +
  ggtitle("Histogram of Weekly Case Rate") +
  xlab("Weekly Case Rate") +
  ylab("Frequency")
```

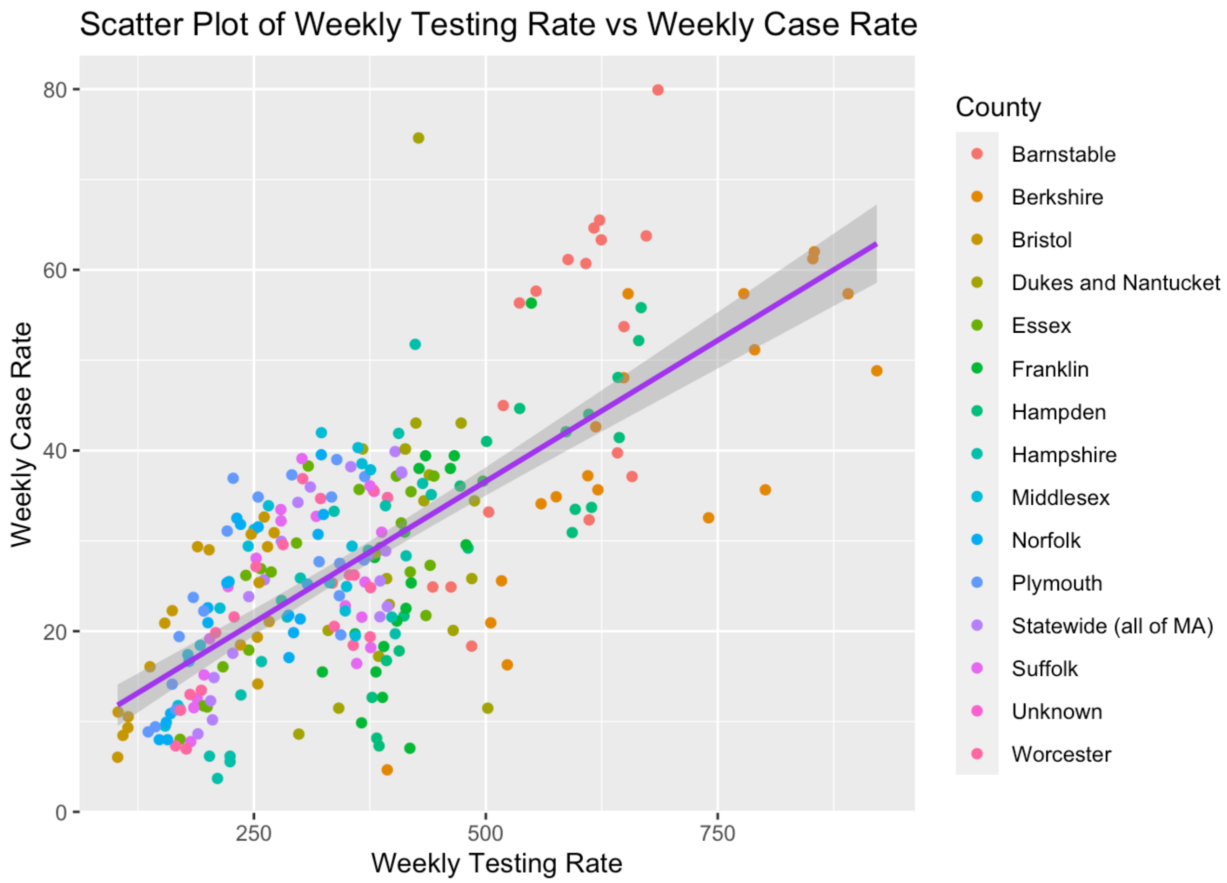
```
Warning: Removed 19 rows containing non-finite values (`stat_bin()`).
```



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%)**

```
ggplot(covid_county, aes(x = `Weekly testing rate`, y = `Weekly case rate`)) +
  geom_point(aes(color = County)) +
  geom_smooth(method = "lm",  color = "purple") +
  labs(title = "Scatter Plot of Weekly Testing Rate vs Weekly Case Rate",
   x = "Weekly Testing Rate",
   y = "Weekly Case Rate")
```

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

Warning: Removed 19 rows containing non-finite values (`stat_smooth()`).

Warning: Removed 19 rows containing missing values (`geom_point()`).
```



```
print(plot)
```

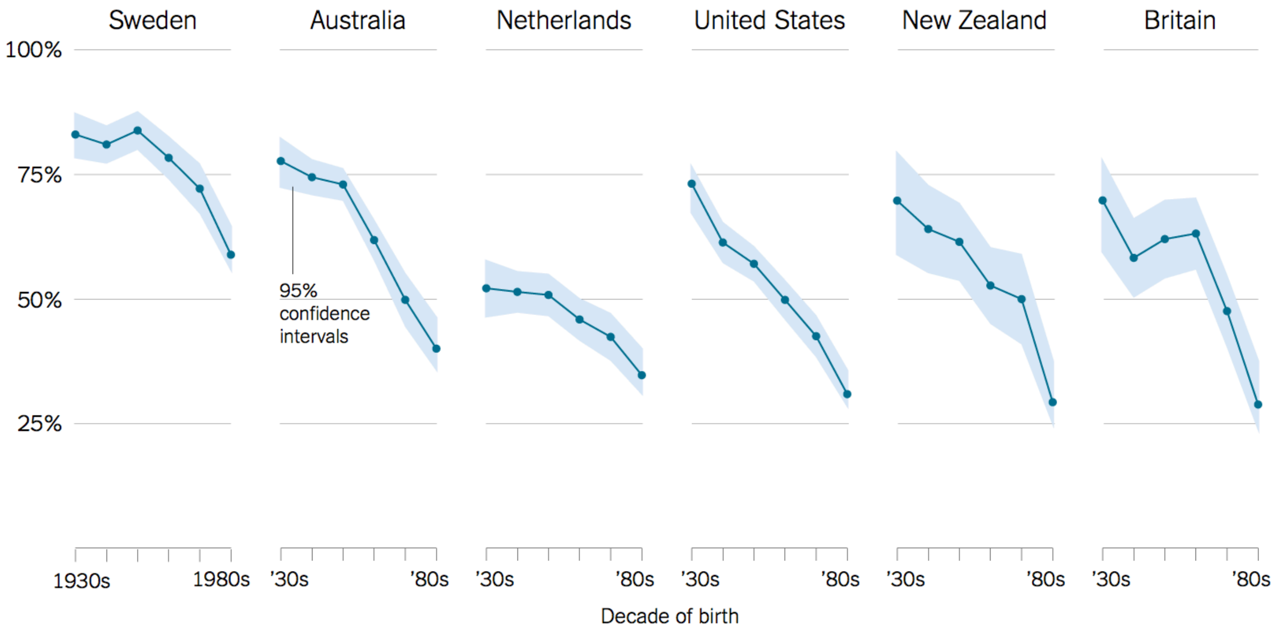
```
function (x, y, ...)
UseMethod("plot")
<bytecode: 0x1094410d8>
<environment: namespace:base>
```

The graph supports the hypothesis that as testing increases, the number of detected cases also tends to increase.

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

%

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%)**

The graph broadly demonstrates a decline across several decades in the percentage of people who consider it “essential” to live in a democracy in six different countries: Sweden, Australia, the Netherlands, the United States, New Zealand, and Britain. For instance, in the United States, the graph shows a sharp decrease from nearly 75% of those born in the 1930s believing democracy is essential to below 30% of those born in the 1980s feeling the same way. New Zealand shows a less dramatic fall compared to the United States, and Britain displays a slight upward tick.

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).

```
wvs_data <- read_rds("WVS5.rds")
wvs_subset <- wvs_data[c("V2", "V236", "V162")]
country_codes <- c(752, 36, 528, 840, 554, 826)
```

```r
wvs_filtered <- wvs_subset[wvs_subset$V2 %in% country_codes, ]
head(wvs_filtered)
```

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

6 rows

(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%)**

```r
unique_countries <- length(unique(wvs_filtered$V2))
na_countries <- sum(is.na(wvs_filtered$V2))

mean_age <- mean(wvs_filtered$V236, na.rm = TRUE)
range_age <- range(wvs_filtered$V236, na.rm = TRUE)
unique_ages <- length(unique(wvs_filtered$V236))
na_ages <- sum(is.na(wvs_filtered$V236))

mean_question <- mean(wvs_filtered$V162, na.rm = TRUE)
range_question <- range(wvs_filtered$V162, na.rm = TRUE)
unique_questions <- length(unique(wvs_filtered$V162))
na_questions <- sum(is.na(wvs_filtered$V162))

cat("Country:\nUnique Values:", unique_countries,", No of NA:", na_countries, "\
```

```
Country:
Unique Values: 6 , No of NA: 0
```

```r
cat("Age:\nMean:", mean_age,", Range:", range_age[1], "-", range_age[2],", Uniqu
```

```
Age:
Mean: 1946.526 , Range: -2 - 1991 , Unique Values: 80 , No. of NA: 0
```

```r
cat("Question:\nMean:", mean_question,", Range:", range_question[1], "-", range_
```

```
Question:
Mean: 6.87139 , Range: -5 - 10 , Unique Values: 14 , No. of NA: 0
```

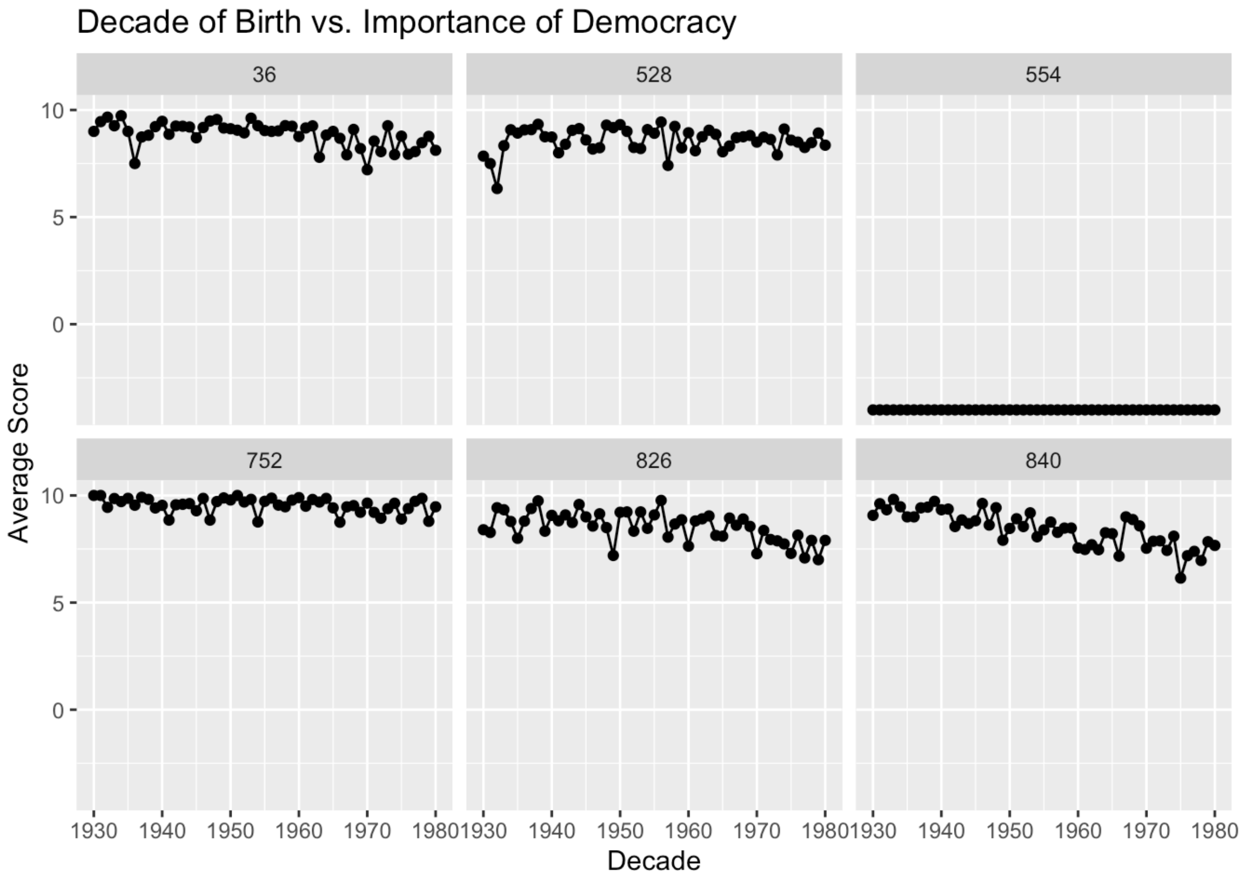
(3) (Optional) Please replicate the graph of the NYT article.

```r
#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%)**

```
  wvs_summary <- wvs_filtered %>%
  group_by(V2, V236) %>%
  summarise(avg_ = mean(V162, na.rm = TRUE), .groups = 'drop')

ggplot(wvs_summary, aes(x = V236, y = avg_, group = V2)) +
  geom_line() +
  geom_point() +
  facet_wrap(~ V2, scales = "fixed") +
  labs(title = "Decade of Birth vs. Importance of Democracy",
   x = "Decade",
   y = "Average Score") +
  scale_x_continuous(breaks = seq(1930, 1980, by = 10), limits = c(1930, 1980))
```

```
Don't know how to automatically pick scale for object of type <labelled>.
Defaulting to continuous.
Don't know how to automatically pick scale for object of type <labelled>.
Defaulting to continuous.

Warning: Removed 134 rows containing missing values (`geom_line()`).

Warning: Removed 134 rows containing missing values (`geom_point()`).
```



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%)**

Here we measures the average score of responses to the question on the importance of democracy whereas NYT graph measures the percentage of people. Both graphs illustrate a downward trend.

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%)**

The concerns regarding the decline in support for democracy, as detailed in the New York Times, might be viewed with less alarm when examining specific data points, such as those from the United States (country code 840). The graph demonstrates that even with the downward trend, average scores in the U.S. rarely dip below 7. This implies that the overall appreciation for democratic values remains robust among the majority.