Challenge_1: Data Import, Description, and Transformation(1)

AUTHOR
Muskan Dhar

PUBLISHED
June 2, 2024

Make sure you change the author's name.

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     1.1.4     ✔ 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
```

```
library(readr)
library(readxl)
library(haven) #for loading other datafiles (SAS, STATA, SPSS, etc.)
```

Challenge Overview

This first challenge aims to practice the following skill sets:

1. Read datasets in different file types;
2. Describe the datasets;
3. Exploring a few basic functions of data transformation and wrangling and present some descriptive statistics (such as min, max, and median).

There will be coding components (reading datasets and data transformation) and writing components (describing the datasets and some statistical information). Please read the instructions for each part and complete your challenges.

Create your R quarto project and submit the standalone .html file.

Please use Challenge 0 in week 1 as a practice of rendering html files. Find how to make standalone html files in week 1 lecture recordings.

Datasets

There are four datasets provided in this challenge. Please download the following dataset files from Google Classroom and save them to a folder within your project working directory (i.e.: "DACSS601_data"). If you don't have a folder to store the datasets, please create one.

- babynames.csv (Required) ⭐
- ESS_5.dta (Option 1) ⭐
- p5v2018.sav (Option 2)⭐
- railroad.xlsx (Required)⭐⭐

Find the `_data` folder, then use the correct R command to read the datasets.

Part 1(Required). The Baby Names Dataset

1. **Read the dataset "babynames.csv", and check the first few rows:**

```
dir.create("DACSS601_data")
```

```
Warning in dir.create("DACSS601_data"): 'DACSS601_data' already exists
```

```
setwd("DACSS601_data")
baby_names <- read.csv("babynames.csv")
head(baby_names)
```

```
       Name    Sex Occurrences Year
1      Mary Female        7065 1880
2      Anna Female        2604 1880
3      Emma Female        2003 1880
4 Elizabeth Female        1939 1880
5    Minnie Female        1746 1880
6  Margaret Female        1578 1880
```

2. **Data Description: Please use the necessary commands and codes and briefly describe this data with a short writing paragraph answering the following questions.**

```
dim(baby_names)
```

```
[1] 2084710       4
```

```
str(baby_names)
```

```
'data.frame':   2084710 obs. of  4 variables:
 $ Name       : chr  "Mary" "Anna" "Emma" "Elizabeth" ...
 $ Sex        : chr  "Female" "Female" "Female" "Female" ...
 $ Occurrences: int  7065 2604 2003 1939 1746 1578 1472 1414 1320 1288 ...
 $ Year       : int  1880 1880 1880 1880 1880 1880 1880 1880 1880 1880 ...
```

```
colnames(baby_names)
```

```
[1] "Name"        "Sex"         "Occurrences" "Year"
```

(1) What is the dimension of the data (# of rows and columns)? The dataset has 258000 rows and 4 columns.

(2) What do the rows and columns mean in this data? Columns: Name: The name given to babies. Sex: The gender of the babies, either ‘Female’ or ‘Male’. Occurrences: The no. of times the name was given to babies in that year. Year: The year in which the name was given. Each row of the data tells us about a particular babyname and the attributes related to it.

(3) What is the unit of observation? In other words, what does each case mean in this data? Each row in this dataset represents the baby name, its gender, and the year in which it was given.

(4) According to the lecture, is this a “tidy” data? Yes, this is a “tidy” data. Each variable (Name, Sex, Occurrences, Year) forms a column. Each observation (a specific name for a specific gender in a specific year) forms a row and each value has its own cell.

3. **Data Transformation: use necessary commands and codes and answer the following questions.**

```
male_name <- baby_names %>%
  filter(Sex == "Male") %>%
  distinct(Name) %>%
  count() %>%
  pull(n)
print(male_name)
```

```
[1] 43653
```

```
female_name <- baby_names %>%
  filter(Sex == "Female") %>%
  distinct(Name) %>%
  count() %>%
  pull(n)
print(female_name)
```

```
[1] 70225
```

```
all_names <- baby_names %>%
  distinct(Name) %>%
```

```
  count() %>%
  pull(n)
print(all_names)
```

```
[1] 102447
```

```
years <- length(unique(baby_names$Year))
print(years)
```

```
[1] 143
```

```
summary_ <- summary(baby_names$Occurrences)
min_ <- summary_["Min."]
mean_ <- summary_["Mean"]
median_ <- summary_["Median"]
max_ <- summary_["Max."]
print(min_)
```

```
Min.
   5
```

```
print(median_)
```

```
Median
    12
```

```
print(mean_)
```

```
    Mean
175.2112
```

```
print(max_)
```

```
 Max.
99693
```

```
print(summary_)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    5.0     7.0    12.0   175.2    32.0 99693.0
```

\(1\) How many unique male names, unique female names, and total unique names are in the data?
There are 43653 unique male names.
There are 70225 unique female names.
In total, there are 102447 unique names in the dataset.

\(2\) How many years of names does this data record
The data records names over 143 years

```
\(3\) Summarize the min, mean, median, and max of "Occurrence". (Must use
summarize())
The minimum number of occurrences is 5.
The mean number of occurrences is approximately 175.2112.
The median number of occurrences is 12.
The maximum number of occurrences is 99693.

\(4\) (Optional) Summarize the min, mean, median, and max of "Occurrence" by decade.
```

Part 2. Choose One Option of Tasks to Complete

In this part, please choose either of the two datasets to complete the tasks.

Optional 1: The European Social Survey Dataset

The European Social Survey (ESS) is an academically-driven multi-country survey, which has been administered in over 30 countries to date. Its three aims are, firstly - to monitor and interpret changing public attitudes and values within Europe and to investigate how they interact with Europe's changing institutions, secondly - to advance and consolidate improved methods of cross-national survey measurement in Europe and beyond, and thirdly - to develop a series of European social indicators, including attitudinal indicators.

In the fifth round, the survey covers 28 countries and investigates two major topics: Family Work and Wellbeing and Justice.

1. **Read the dataset "ESS_5.dta".**

```
setwd("DACSS601_data")
data <- read_dta("ESS_5.dta")
```

2. **Data Description: Please use the necessary commands and codes and briefly describe this data with a short writing paragraph answering the following questions.**

   (1) What is the dimension of the data (# of rows and columns)? The no. of rows are 52458 and the no. of columns are 696.

   ```
   dim(data)
   ```

   ```
   [1] 52458   696
   ```

   As we can see, this data is very large. We don't want to study the whole data. Let's just reload the following selected columns: "idno, essroud, male, age, edu, income_10, eth_major, media (a standardized measure of the frequency of media consumption), and cntry".

```
data1 <- select(data, idno, essround, male, age, edu, income_10, eth_major, media, c
head(data1)
```

```
# A tibble: 6 × 9
   idno essround  male   age   edu income_10 eth_major  media cntry
  <dbl>    <dbl> <dbl> <dbl> <dbl>     <dbl>     <dbl>  <dbl> <chr>
1 15906        5     0    14     1         2         1 0.312  GR
2 21168        5     0    14     1         2         1 0.438  IE
3    40        5     0    14     1         8        NA 0.375  LT
4  2108        5     0    14     1        NA         1 0.0625 RU
5   519        5     0    14     1        NA         1 0.125  IL
6  2304        5     0    14     1        NA         1 0.25   ES
```

\(2\) For the reloaded/smaller data, what do the rows and columns mean in this data?
Each row represents an individual respondent in the survey.Columns represent the attributes related to the unique respondent such as essround in which they are participating, their gender, age etc.

\(3\) What is the unit of observation? In other words, what does each case mean in this data?
Each row represents an individual respondent's responses to the survey in a specific country during the fifth round of the European Social Survey.

\(4\) According to the lecture, is this a "tidy" data?
Yes because each variable forms a column, each observation forms a row and each type of observational unit forms a table.

3. **Data Transformation: use necessary commands and codes, and answer the following questions.**

```r
countries <- length(unique(data1$cntry))
print(countries)
```

```
[1] 27
```

```r
age_summary <- summary(data1$age)
edu_summary <- summary(data1$edu)
media_summary <- summary(data1$media)

age_range <- range(data1$age, na.rm = TRUE)
edu_range <- range(data1$edu, na.rm = TRUE)
media_range <- range(data1$media, na.rm = TRUE)

age_mean <- mean(data1$age, na.rm = TRUE)
edu_mean <- mean(data1$edu, na.rm = TRUE)
media_mean <- mean(data1$media, na.rm = TRUE)
print(age_range)
```

```
[1]  14 101
```

```r
print(age_mean)
```

```
[1] 47.91529
```

```
print(edu_range)
```

```
[1] 1 4
```

```
print(edu_mean)
```

```
[1] 2.767531
```

```
print(media_range)
```

```
[1] 0 1
```

```
print(media_mean)
```

```
[1] 0.4786802
```

```
eth_major <- sum(is.na(data1$eth_major))
income_10 <- sum(is.na(data1$income_10))
print(eth_major)
```

```
[1] 1310
```

```
print(income_10)
```

```
[1] 12620
```

\(1\) How many unique countries are in the data?
There are 28 unique countries in the dataset.

\(2\) What are the range and average of the following variables: "age", "edu", and "media"? Must use summarize().
Age: Range is 14 to 101 and Mean is 47.91529
Education (edu): Range is 1 to 4 and Mean is 2.767531
Media Consumption (media): Range: 0 to 1 and Mean is 0.4786802

\(3\) How many missing data (NA) are in the following variables: "eth_major" and "income_10"? (tips: use is.na())
The variable eth_major has 1310 missing values. The variable income_10 has 12620 missing values.

Optional 2: Polity V Data

The Polity data series is a data series in political science research. Polity is among prominent datasets that measure democracy and autocracy. The Polity5 dataset covers all major, independent states in the global system over the period 1800-2018 (i.e., states with a total population of 500,000 or more in the most recent year; currently 167 countries with Polity5 refinements completed for about half those countries).

1. **Read the dataset “p5v2018.sav”.**

```
#Type your code here
```

2. **Data Description: Please use the necessary commands and codes and briefly describe this data with a short writing paragraph answering the following questions.**

```
#Type your code here; and write a paragraph answering the questions.
```

(1) What is the dimension of the data (# of rows and columns)?

As we can see, this data contains many columns. We don’t want to study the whole data. Let’s keep the first seven columns and the ninth and the tenth columns.

```
#Type your code here; and write a paragraph answering the questions.
```

(2) For the reloaded data, what do the rows mean in this data? What do the columns (#2-#8) mean? (If you have questions, check out p.11-16 of the User Manual/Codebook of the dataset.

(3) What is the unit of observation? In other words, what does each case mean in this data?

(4) According to the lecture, is this a “tidy” data?

3. **Data Transformation: use necessary commands and codes and answer the following questions.**

```
#Type your code here; and write a paragraph answering the questions.
```

(1) How many unique countries are in the data?

(2) How many years does this data record?

(3) What are the range and average of the following variables: “democ” and “autoc”?

** Noted that in this data, negative integers (-88, -77, and -66) represent special cases. You should exclude them when calculating the range, average, and NAs.

(4) How many missing data (NA) are in the following variables: “democ” and “autoc”? (tips: use is.na())

Part 3. The Railroad Employee Data

1. **Read the dataset “railroads.xlsx”.**

Many government organizations still use Excel spreadsheets to store data. This railroad dataset, published by the Railroad Retirement Board, is a typical example. It records the number of employees in each county and state in 2012.

Please load the data in R in a clean manner. You can start by doing the following things step by step.

(1) Read the first sheet of the Excel file;

(2) Skipping the title rows;

(3) Removing empty columns

(4) Deleting rows that contain the name “total”, e.g. “WI total”

(5) Deleting the row for State “CANADA”

(6) Remove the table notes (the last two rows)

```
setwd("DACSS601_data")
data2 <- read_excel("railroads.xlsx", sheet = 1,  skip = 2)
```

```
New names:
• `` -> `...2`
• `` -> `...4`
```

```
data2 <- data2 %>% select_if(~!all(is.na(.)))
```

```
data2 <- data2 %>%
  filter(!grepl("Total", STATE, ignore.case = TRUE))
data2 <- data2 %>%
  filter(STATE != "CANADA")
   data2 <- data2[1:(nrow(data2) - 2), ]

head(data2)
```

```
# A tibble: 6 × 3
  STATE COUNTY               TOTAL
  <chr> <chr>                <dbl>
1 AE    APO                      2
2 AK    ANCHORAGE                7
3 AK    FAIRBANKS NORTH STAR     2
4 AK    JUNEAU                   3
5 AK    MATANUSKA-SUSITNA        2
6 AK    SITKA                    1
```

2. **Data Description: Please use the necessary commands and codes and briefly describe this data with a short writing paragraph answering the following questions.**

```
dim(data2)
```

```
[1] 2930    3
```

(1) What is the dimension of the data (# of rows and columns)? The number of rows in the data are 2930 and the number of columns are 3.

(2) What do the rows and columns mean? Each row represents a unique observation of a county within a state, detailing the number of railroad employees in that county. Each column represents a variable: the state, county, and number of employees.

(3) What is the unit of observation? In other words, what does each case mean in this data? Each row in this dataset represents a unique observation of a county within a state, detailing the number of railroad employees in that county.

(4) According to the lecture, is this a “tidy” data? Since we had to clean the data before analyzing it, therefore the dataset is not tidy.

3. **Data Transformation: use necessary commands and codes and answer the following questions.**

```
counts <- data2 %>% summarise(across(c(STATE, COUNTY), n_distinct))
print(counts)
```

```
# A tibble: 1 × 2
  STATE COUNTY
  <int>  <int>
1    53   1709
```

```
total <- sum(data2$TOTAL, na.rm = TRUE)
print(total)
```

```
[1] 255432
```

```
print(summary(data2$TOTAL))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    7.00   21.00   87.18   65.00 8207.00
```

```
state_employees <- data2 %>%
  group_by(STATE) %>%
  summarise(total = sum(TOTAL, na.rm = TRUE)) %>%
  arrange(desc(total))
print(state_employees)
```

```
# A tibble: 53 × 2
   STATE total
   <chr> <dbl>
 1 TX    19839
 2 IL    19131
 3 NY    17050
 4 NE    13176
 5 CA    13137
 6 PA    12769
 7 OH     9056
 8 GA     8605
 9 IN     8537
```

```
10 MO     8419
# ℹ 43 more rows
```

```r
counties_employees <- data2 %>%
  group_by(COUNTY) %>%
  summarise(total1 = sum(TOTAL, na.rm = TRUE)) %>%
  arrange(desc(total1))
print(counties_employees)
```

```
# A tibble: 1,709 × 2
   COUNTY           total1
   <chr>             <dbl>
 1 COOK               8211
 2 DOUGLAS            4929
 3 SUFFOLK            4243
 4 TARRANT            4235
 5 INDEPENDENT CITY   4205
 6 JEFFERSON          3723
 7 DUVAL              3074
 8 SAN BERNARDINO     2888
 9 LINCOLN            2861
10 LAKE               2658
# ℹ 1,699 more rows
```

\(1\) How many unique counties and states are in the data? (tips: you can try using the across() function to do an operation on two columns at the same time)
The dataset contains 53 unique states and 1709 unique counties.

\(2\) What is the total number of employees (total_employees) in this data?
The total number of employees in this dataset is 255432.

\(3\) What are the min, max, mean, and median of "total_employees"
The minimum number of employees in a county is 1
The maximum number of employees in a county is 8207.00.
The mean number of employees across counties is 87.18.
The median number of employees across counties is 21.00.

\(4\) Which states have the most employees? And which countries have the most employees? (tips: use group_by() and arrange())
States with maximum employees are TX, IL,NY and so on. Counties with maximum employees are COOK, DOUGLAS, SUFFOLK and so on.