Challenge_1: Data Import, Description, and Transformation

AUTHOR
Mehak Nargotra

PUBLISHED
February 1, 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     1.1.3     ✔ 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(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:**

```
#Type your code here
babynames <- read_csv("~/Desktop/DACSS 601/DACSS_601_datasets/babynames.csv")
```

```
Rows: 2084710 Columns: 4
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (2): Name, Sex
dbl (2): Occurrences, Year

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(babynames)
```

```
# A tibble: 6 × 4
  Name      Sex    Occurrences  Year
  <chr>     <chr>        <dbl> <dbl>
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.**

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

   (2) What do the rows and columns mean in this data?

   (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?

```
#Type your code in the code chunk; then write a paragraph answering the questior
# \(1\) What is the dimension of the data (# of rows and columns)?
dim_babynames <- dim(babynames)
cat("The babynames.csv contains:",dim_babynames[1]," rows and ", dim_babynames[2
```

```
The babynames.csv contains: 2084710  rows and  4  columns.
```

```
#\(2\) What do the rows and columns mean in this data?
colnames_babynames <- colnames(babynames)
cat("The columns are:",colnames_babynames,"\n.
    They are the variables which contain the data.\n")
```

```
The columns are: Name Sex Occurrences Year
.
        They are the variables which contain the data.
```

```
cat("The rows represent the organizational unit of the dataset.
Column 'Name' contains the names given to a baby.
Column 'Sex' contains information about whether the baby name represents a male
Column 'Occurrences' contains the number of occurrences of the given name in tha
Column 'Year' contains the year in which the name was recorded or registered.\n
In this considering the 1st row, 'Name: Mary' is given to a  'Sex: Female' baby
'Year:1880'.")
```

```
The rows represent the organizational unit of the dataset.
    Column 'Name' contains the names given to a baby.
    Column 'Sex' contains information about whether the baby name represents a male
or a female.
    Column 'Occurrences' contains the number of occurrences of the given name in
that specific year.
    Column 'Year' contains the year in which the name was recorded or registered.

    In this considering the 1st row, 'Name: Mary' is given to a  'Sex: Female' baby
which was given to 'Occurences: 7065' babies in the
    'Year:1880'.
```

```
#\(3\) What is the unit of observation? In other words, what does each case mear
uniquevalues <- nrow(unique(babynames[, c("Name", "Year")]))
cat("Number of unique combinations of 'Name' and 'Year' i.e., obseravtional unit
```

```
Number of unique combinations of 'Name' and 'Year' i.e., obseravtional unit:
1903046
```

```
    #\(4\) According to the lecture, is this a "tidy" data?
    print("Yes, this is a tidy data as the data is organized into proper rows and col
```

```
[1] "Yes, this is a tidy data as the data is organized into proper rows and columns.
Also there are no duplicate values which violate the principle of being a tidy
data."
```

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

   (1) How many unique male names, unique female names, and total unique names are in the data?

```
    #Type your code in the code chunk; and write a paragraph answering the questions
    unique_male_names <- nrow(unique(babynames[babynames$Sex == "Male", "Name"]))
    cat("There are total: ",unique_male_names,"unique male names in the dataset. \n"
```

```
There are total:  43653 unique male names in the dataset.
```

```
    unique_female_names <- nrow(unique(babynames[babynames$Sex == "Female", "Name"])
    cat("There are total: ",unique_female_names,"unique female names in the dataset.
```

```
There are total:  70225 unique female names in the dataset.
```

```
    total_unique_names <- nrow(unique(babynames[, "Name"]))
    cat("There are total: ",total_unique_names,"unique names in the dataset. \n")
```

```
There are total:  102447 unique names in the dataset.
```

```
\(2\) How many years of names does this data record?
```

```
    years <- nrow(unique(babynames[, "Year"]))
    cat("This data records ",years," years of names \n")
```

```
This data records  143  years of names
```

```
\(3\) Summarize the min, mean, median, and max of "Occurrence". (Must use
summarize())
```

```
  occurrence_summary <- babynames %>%
    summarize(
      min_occurrence = min(Occurrences),
      mean_occurrence = mean(Occurrences),
      median_occurrence = median(Occurrences),
      max_occurrence = max(Occurrences)
    )
  print("min, mean, median, and max of 'Occurrence' :")
```

```
[1] "min, mean, median, and max of 'Occurrence' :"
```

```
  occurrence_summary
```

```
# A tibble: 1 × 4
  min_occurrence mean_occurrence median_occurrence max_occurrence
           <dbl>           <dbl>             <dbl>          <dbl>
1              5            175.                12          99693
```

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

```
    babynames_decade <- babynames %>%
      mutate(Decade = 10 * (Year %/% 10))
    occurrence_by_decade_summary <- babynames_decade %>%
      group_by(Decade) %>%
      summarize(
        min_occurrence_decade = min(Occurrences),
        mean_occurrence_decade = mean(Occurrences),
        median_occurrence_decade = median(Occurrences),
        max_occurrence_decade = max(Occurrences)
      )
      print("min, mean, median, and max of 'Occurrence by Decade' :")
```

```
[1] "min, mean, median, and max of 'Occurrence by Decade' :"
```

```
      occurrence_by_decade_summary
```

```
# A tibble: 15 × 5
   Decade min_occurrence_decade mean_occurrence_decade median_occurrence_decade
    <dbl>                 <dbl>                  <dbl>                     <dbl>
 1   1880                     5                   106.                        13
 2   1890                     5                   114.                        13
 3   1900                     5                   117.                        12
 4   1910                     5                   184.                        12
 5   1920                     5                   218.                        12
 6   1930                     5                   232.                        12
 7   1940                     5                   307.                        13
 8   1950                     5                   357.                        13
 9   1960                     5                   302.                        13
10   1970                     5                   191.                        11
11   1980                     5                   173.                        11
12   1990                     5                   142.                        11
13   2000                     5                   118.                        11
14   2010                     5                   109.                        11
15   2020                     5                   106.                        12
# ℹ 1 more variable: max_occurrence_decade <dbl>
```

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".**

```
library(haven)
ESS_5 <- read_dta("~/Desktop/DACSS 601/DACSS_601_datasets/ESS_5.dta")
head(ESS_5)#Type your code here
```

```
# A tibble: 6 × 696
   idno essround  male   age   edu income_10 eth_major  media  obey trust_court
  <dbl>    <dbl> <dbl> <dbl> <dbl>     <dbl>     <dbl>  <dbl> <dbl>       <dbl>
1 15906        5     0    14     1         2         1 0.312   1           1
2 21168        5     0    14     1         2         1 0.438   1           0.75
3    40        5     0    14     1         8        NA 0.375   0.5         0.5
4  2108        5     0    14     1        NA         1 0.0625  0.75        0.75
5   519        5     0    14     1        NA         1 0.125   1           1
6  2304        5     0    14     1        NA         1 0.25    0.5         0.25
# i 686 more variables: cntry <chr>, commonlaw <dbl>, PostComm <dbl>, tv <dbl>,
#   radio <dbl>, papers <dbl>, Internet <dbl>, name <chr>, edition <chr>,
#   proddate <chr>, tvtot <dbl+lbl>, tvpol <dbl+lbl>, rdtot <dbl+lbl>,
#   rdpol <dbl+lbl>, nwsptot <dbl+lbl>, nwsppol <dbl+lbl>, netuse <dbl+lbl>,
#   ppltrst <dbl+lbl>, pplfair <dbl+lbl>, pplhlp <dbl+lbl>, polintr <dbl+lbl>,
#   trstprl <dbl+lbl>, trstlgl <dbl+lbl>, trstplc <dbl+lbl>, trstplt <dbl+lbl>,
#   trstprt <dbl+lbl>, trstep <dbl+lbl>, trstun <dbl+lbl>, vote <dbl+lbl>, …
```

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)?

```
#Type your code here; and write a paragraph answering the questions.
# \(1\) What is the dimension of the data (# of rows and columns)?
data_dimension <- dim(ESS_5)
cat("Data Dimensions of the dataset ESS_5 are: ",data_dimension[1]," rows and ",
```

```
Data Dimensions of the dataset ESS_5 are:  52458  rows and  696  columns
```

```
#
```

```
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".

```
    #Type your code here; and write a paragraph answering the questions.
    selected_columns <- c("idno", "essround", "male", "age", "edu", "income_10", "et
    selected_data <- ESS_5[, selected_columns]
    print(selected_data)
```

```
# A tibble: 52,458 × 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
 7   290        5     0    14     1        NA         1 0.312  PT
 8  3977        5     0    14     1        NA         1 0.375  BG
 9 23244        5     0    14     1        NA         1 0.375  IE
10 19417        5     0    14     1        NA         1 0.438  IE
# i 52,448 more rows
```

\(2\) For the reloaded/smaller data, what do the rows and columns mean in this data?

```
    #Type your code here; and write a paragraph answering the questions.
  #Data description
    print("The dataset comprises 52458 observations and 696 columns, with each row r
```

```
[1] "The dataset comprises 52458 observations and 696 columns, with each row
representing a unique respondent. Some of its columns are 'idno' (identification
number), 'essround' (survey round), 'male' (gender), 'age' (age of the respondent),
'edu' (educational level), 'income_10' (income category), 'eth_major' (major
ethnicity), 'media' (media consumption habits), and 'cntry' (country of residence)."
```

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

```
    print("Each case, or each row in the dataset, corresponds to a single survey res
```

```
[1] "Each case, or each row in the dataset, corresponds to a single survey
respondent. Each row in the dataset represents the responses and characteristics of
a specific person who participated in the survey."
```

\(4\) According to the lecture, is this a "tidy" data?

```
    print("Yes this is a tidy data. Each variable forms a column and each row forms
```

```
[1] "Yes this is a tidy data. Each variable forms a column and each row forms an
observation. Also there is no duplicate data in the dataset which proves that it is
a tidy data."
```

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

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

   (2) What are the range and average of the following variables: “age”, “edu”, and “media”? Must use summarize().

   (3) How many missing data (NA) are in the following variables: “eth_major” and “income_10”? (tips: use is.na())

```
#Type your code here; and write a paragraph answering the questions.
#\(1\) How many unique countries are in the data?
unique_countries <- n_distinct(selected_data$cntry)
cat("1. Number of unique countries in the data:", unique_countries, "\n")
```

```
1. Number of unique countries in the data: 27
```

```
#\(2\) What are the range and average of the following variables: "age", "edu",
summary_stats_data <- selected_data %>%
summarize(
  age_range = range(age),
  age_avg = mean(age, na.rm = TRUE),
  edu_range = range(edu),
  edu_avg = mean(edu, na.rm = TRUE),
  media_range = range(media),
  media_avg = mean(media, na.rm = TRUE)
)
```

```
Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated
in
dplyr 1.1.0.
ℹ Please use `reframe()` instead.
ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
  always returns an ungrouped data frame and adjust accordingly.
```

```
cat("2. Summary Statistics for 'age', 'edu', and 'media':\n")
```

```
2. Summary Statistics for 'age', 'edu', and 'media':
```

```
print(summary_stats_data)
```

```
# A tibble: 2 × 6
  age_range age_avg edu_range edu_avg media_range media_avg
      <dbl>   <dbl>     <dbl>   <dbl>       <dbl>     <dbl>
1        NA    47.9        NA    2.77          NA     0.479
2        NA    47.9        NA    2.77          NA     0.479
```

```
#\(3\) How many missing data (NA) are in the following variables: "eth_major" ar
```

```
    eth_major_missingdata <- sum(is.na(selected_data$eth_major))
    income_10_missingdata <- sum(is.na(selected_data$income_10))
    cat("\n\n 3. Number of missing data (NA) for 'eth_major':", eth_major_missingdat
```

```
3. Number of missing data (NA) for 'eth_major': 1310
```

```
    cat("   Number of missing data (NA) for 'income_10':", income_10_missingdata, "\
```

```
   Number of missing data (NA) for 'income_10': 12620
```

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.

```
railroad_data <- read_excel("~/Desktop/DACSS 601/DACSS_601_datasets/railroads.xl
```

```
New names:
• `` -> `...2`
• `` -> `...3`
• `` -> `...4`
• `` -> `...5`
• `` -> `...6`
```

```
head(railroad_data)
```

```
# A tibble: 6 × 6
  `TOTAL RAILROAD EMPLOYMENT BY STATE AND COUNTY` ...2   ...3  ...4  ...5  ...6
  <chr>                                           <chr>  <lgl> <chr> <lgl> <chr>
1 CALENDAR YEAR 2012                              <NA>   NA    <NA>  NA    <NA>
2 <NA>                                            <NA>   NA    <NA>  NA    <NA>
3 <NA>                                            STATE  NA    COUN… NA    TOTAL
4 <NA>                                            AE     NA    APO   NA    2.0
5 <NA>                                            AE To… NA    <NA>  NA    2
6 <NA>                                            AK     NA    ANCH… NA    7.0
```

\(1\) Read the first sheet of the Excel file;
\(2\) Skipping the title rows;

```
railroad_data <- read_excel("~/Desktop/DACSS 601/DACSS_601_datasets/railroads.xl
```

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

```
    railroad_data
```

```
# A tibble: 2,990 × 5
   STATE     ...2  COUNTY               ...4  TOTAL
   <chr>     <lgl> <chr>                <lgl> <dbl>
 1 AE        NA    APO                  NA        2
 2 AE Total1 NA    <NA>                 NA        2
 3 AK        NA    ANCHORAGE            NA        7
 4 AK        NA    FAIRBANKS NORTH STAR NA        2
 5 AK        NA    JUNEAU               NA        3
 6 AK        NA    MATANUSKA-SUSITNA    NA        2
 7 AK        NA    SITKA                NA        1
 8 AK        NA    SKAGWAY MUNICIPALITY NA       88
 9 AK Total  NA    <NA>                 NA      103
10 AL        NA    AUTAUGA              NA      102
# ℹ 2,980 more rows
```

\(3\) Removing empty columns

```
    railroad_data <- select(railroad_data, -where(~all(is.na(.))))
    railroad_data
```

```
# A tibble: 2,990 × 3
   STATE     COUNTY               TOTAL
   <chr>     <chr>                <dbl>
 1 AE        APO                      2
 2 AE Total1 <NA>                     2
 3 AK        ANCHORAGE                7
 4 AK        FAIRBANKS NORTH STAR     2
 5 AK        JUNEAU                   3
 6 AK        MATANUSKA-SUSITNA        2
 7 AK        SITKA                    1
 8 AK        SKAGWAY MUNICIPALITY    88
 9 AK Total  <NA>                   103
10 AL        AUTAUGA                102
# ℹ 2,980 more rows
```

\(4\) Deleting rows that contain the name "total", e.g. "WI total"

```
#railroad_data_updated <- railroad_data_updated[!grepl("total", railroad_data_update
    railroad_data <- railroad_data %>%
      filter(!grepl("Total", STATE, ignore.case = TRUE))
    railroad_data
```

```
# A tibble: 2,936 × 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
 7 AK    SKAGWAY MUNICIPALITY    88
 8 AL    AUTAUGA                102
 9 AL    BALDWIN                143
10 AL    BARBOUR                  1
# i 2,926 more rows
```

\(5\) Deleting the row for State "CANADA"

```
    railroad_data <- filter(railroad_data, STATE != "CANADA")
    railroad_data
```

```
# A tibble: 2,932 × 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
 7 AK    SKAGWAY MUNICIPALITY    88
 8 AL    AUTAUGA                102
 9 AL    BALDWIN                143
10 AL    BARBOUR                  1
# i 2,922 more rows
```

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

```
    #Type your code here
    railroad_data <- head(railroad_data, n = nrow(railroad_data) - 2)
    railroad_data
```

```
# A tibble: 2,930 × 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
 7 AK    SKAGWAY MUNICIPALITY    88
 8 AL    AUTAUGA                102
 9 AL    BALDWIN                143
10 AL    BARBOUR                  1
# i 2,920 more rows
```

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)?

(2) What do the rows and columns mean?

(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?

```
#Type your code here; and write a paragraph answering the questions.
#\(1\) What is the dimension of the data (# of rows and columns)?
railroad_data_dimensions <- dim(railroad_data)
cat("Ans 1: The dimension of the data are: ", railroad_data_dimensions[1]," rows
```

```
Ans 1: The dimension of the data are:  2930  rows and  3  columns.
```

```
#\(2\) What do the rows and columns mean?

cat("Ans 2: The dataset consists of", railroad_data_dimensions[1], "rows and", r
```

```
Ans 2: The dataset consists of 2930 rows and 3 columns. Each row represents a unique
observation or entry, and each column corresponds to a different variable or
attribute. Understanding the dimensions of the data is crucial for further analysis
and exploration, helping to comprehend the scale and structure of the dataset.
```

```
#\(3\) What is the unit of observation? In other words, what does each case mear
cat("Ans 3: In this each row represents a unit of observation. It shows that in
```

```
Ans 3: In this each row represents a unit of observation. It shows that in a state,
a particular county has these many number of railroads employees and hence, giving
the number of employees in each county of any state.
```

```
#\(4\) According to the lecture, is this a "tidy" data?
cat("Ans 4: No, the railroads data is not a tidy data. This is because while the
```

```
Ans 4: No, the railroads data is not a tidy data. This is because while the data is
organized into rows and columns, there are a lot of missing values and repeated data
which we cleaned in the 1st part of the question.
```

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

   (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)

   (2) What is the total number of employees (total_employees) in this data?

   (3) What are the min, max, mean, and median of “total_employees”

   (4) Which states have the most employees? And which countries have the most employees? (tips: use group_by() and arrange())

```
#Type your code here; and write a paragraph answering the questions.
library(dplyr)
```

```r
#\(1\) How many unique counties and states are in the data? (tips: you can try u
unique_count <- railroad_data %>%
distinct(across(c(COUNTY, STATE))) %>%
summarize(
  unique_counties = n_distinct(COUNTY),
  unique_states = n_distinct(STATE))
print(unique_count)
```

```
# A tibble: 1 × 2
  unique_counties unique_states
            <int>         <int>
1            1709            53
```

```r
#\(2\) What is the total number of employees (total_employees) in this data?
total_employees <- sum(railroad_data$TOTAL, na.rm = TRUE)
cat("\n\n The total number of employees in this data are:", total_employees,"\n\n"
```

```
 The total number of employees in this data are: 255432
```

```r
#\(3\) What are the min, max, mean, and median of "total_employees"

min_total_employees <- min(railroad_data$TOTAL, na.rm = TRUE)
cat("Min Total Employees: ",min_total_employees,"\n")
```

```
Min Total Employees:  1
```

```r
max_total_employees <- max(railroad_data$TOTAL, na.rm = TRUE)
cat("Max Total Employees: ",max_total_employees,"\n")
```

```
Max Total Employees:  8207
```

```r
mean_total_employees <- mean(railroad_data$TOTAL, na.rm = TRUE)
cat("Mean Total Employees: ",mean_total_employees,"\n")
```

```
Mean Total Employees:  87.17816
```

```r
median_total_employees <- median(railroad_data$TOTAL, na.rm = TRUE)
cat("Mean Total Employees: ",median_total_employees,"\n")
```

```
Mean Total Employees:  21
```

```r
#\(4\) Which states have the most employees? And which countries have the most emp
total_state_employees <- railroad_data %>%
group_by(STATE) %>%
summarize(total_employees = sum(TOTAL, na.rm = TRUE)) %>%
```

```
  arrange(desc(total_employees))
  print("The states which have the most employeees are: ")
```

```
[1] "The states which have the most employeees are: "
```

```
  print(total_state_employees)
```

```
# A tibble: 53 × 2
   STATE total_employees
   <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
```

```
  total_county_employees <- railroad_data %>%
  group_by(COUNTY) %>%
  summarize(total_employees = sum(TOTAL, na.rm = TRUE)) %>%
  arrange(desc(total_employees))
  print("The Counties that have the most employeees are: ")
```

```
[1] "The Counties that have the most employeees are: "
```

```
  print(total_county_employees)
```

```
# A tibble: 1,709 × 2
   COUNTY           total_employees
   <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
```