Exercise: Adult Dataset

Garridos, Charlene P.

2024-02-09

Contents

1 R library

```
library(forcats)
library(kableExtra)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v ggplot2   3.4.2     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()     masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

2 Dataset Overview

The Adult dataset contains demographic information about adults from the 1994 U.S. census. Key columns include:

- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, etc.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, etc.
- marital-status: Married-civ-spouse, Divorced, Never-married, etc.
- occupation: Tech-support, Craft-repair, Other-service, Sales, etc.
- relationship: Wife, Own-child, Husband, Not-in-family, etc.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, etc.
- income: $> 50K$, $\leq 50K$.

3 Load the Dataset

```
# Specify the file path relative to the working directory
file_path <- "/Users/User/Downloads/adult.csv"

# Read the CSV file
demographic_info <- read.csv(file_path, sep = ",", header=TRUE,
                             col.names = c("age", "workclass", "fnlwgt",
                                           "education", "educational-num",
                                           "marital-status", "occupation",
                                           "relationship", "race", "sex",
                                           "capital-gain", "capital-loss",
                                           "hours-per-week", "native-country","income"))

# View the data
head(demographic_info)
```

```
##   age  workclass fnlwgt     education educational.num      marital.status
## 1  25    Private 226802          11th              7       Never-married
## 2  38    Private  89814       HS-grad              9  Married-civ-spouse
## 3  28  Local-gov 336951    Assoc-acdm             12  Married-civ-spouse
## 4  44    Private 160323  Some-college             10  Married-civ-spouse
## 5  18          ? 103497  Some-college             10       Never-married
## 6  34    Private 198693          10th              6       Never-married
##           occupation   relationship   race     sex capital.gain capital.loss
## 1  Machine-op-inspct      Own-child  Black    Male            0            0
## 2    Farming-fishing        Husband  White    Male            0            0
## 3    Protective-serv        Husband  White    Male            0            0
## 4  Machine-op-inspct        Husband  Black    Male         7688            0
## 5                  ?      Own-child  White  Female            0            0
## 6      Other-service  Not-in-family  White    Male            0            0
##   hours.per.week native.country  income
## 1             40  United-States  <=50K.
## 2             50  United-States  <=50K.
## 3             40  United-States   >50K.
## 4             40  United-States   >50K.
## 5             30  United-States  <=50K.
## 6             30  United-States  <=50K.
```

4 Inspect the data

```
demographic_info_factor <- factor(demographic_info)
demographic_info_workclass <- factor(demographic_info$workclass)
demographic_info_education <- factor(demographic_info$education)
demographic_info_native.country <- factor(demographic_info$native.country)
```

```
# Functions like levels() and str() help inspect the structure and levels of a factor.
str(demographic_info)
```

```
## 'data.frame':    16281 obs. of  15 variables:
##  $ age            : int  25 38 28 44 18 34 29 63 24 55 ...
##  $ workclass      : chr  " Private" " Private" " Local-gov" " Private" ...
##  $ fnlwgt         : int  226802 89814 336951 160323 103497 198693 227026 104626 369667 104996 ...
##  $ education      : chr  " 11th" " HS-grad" " Assoc-acdm" " Some-college" ...
##  $ educational.num: int  7 9 12 10 10 6 9 15 10 4 ...
##  $ marital.status : chr  " Never-married" " Married-civ-spouse" " Married-civ-spouse" " Married-civ-
##  $ occupation     : chr  " Machine-op-inspct" " Farming-fishing" " Protective-serv" " Machine-op-insp
##  $ relationship   : chr  " Own-child" " Husband" " Husband" " Husband" ...
##  $ race           : chr  " Black" " White" " White" " Black" ...
##  $ sex            : chr  " Male" " Male" " Male" " Male" ...
##  $ capital.gain   : int  0 0 0 7688 0 0 0 3103 0 0 ...
##  $ capital.loss   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours.per.week : int  40 50 40 40 30 30 40 32 40 10 ...
##  $ native.country : chr  " United-States" " United-States" " United-States" " United-States" ...
##  $ income         : chr  " <=50K." " <=50K." " >50K." " >50K." ...
```

```
str(demographic_info_factor)
```

```
##  Factor w/ 15 levels "c(7, 9, 12, 10, 10, 6, 9, 15, 10, 4, 9, 13, 9, 9, 9, 14, 10, 9, 9, 16, 13, 10,
##  - attr(*, "names")= chr [1:15] "age" "workclass" "fnlwgt" "education" ...
```

```
# Functions like levels() and str() help inspect the structure and levels of a factor.
levels(demographic_info)
```

```
## NULL
```

```
levels(demographic_info_workclass)
```

```
## [1] " ?"                " Federal-gov"      " Local-gov"
## [4] " Never-worked"     " Private"          " Self-emp-inc"
## [7] " Self-emp-not-inc" " State-gov"        " Without-pay"
```

```
levels(demographic_info_education)
```

```
##  [1] " 10th"         " 11th"         " 12th"         " 1st-4th"
##  [5] " 5th-6th"      " 7th-8th"      " 9th"          " Assoc-acdm"
##  [9] " Assoc-voc"    " Bachelors"    " Doctorate"    " HS-grad"
## [13] " Masters"      " Preschool"    " Prof-school"  " Some-college"
```

```
levels(demographic_info_native.country)
```

```
##  [1] " ?"                          " Cambodia"
##  [3] " Canada"                     " China"
##  [5] " Columbia"                   " Cuba"
##  [7] " Dominican-Republic"         " Ecuador"
##  [9] " El-Salvador"                " England"
## [11] " France"                     " Germany"
## [13] " Greece"                     " Guatemala"
## [15] " Haiti"                      " Honduras"
## [17] " Hong"                       " Hungary"
## [19] " India"                      " Iran"
## [21] " Ireland"                    " Italy"
## [23] " Jamaica"                    " Japan"
## [25] " Laos"                       " Mexico"
## [27] " Nicaragua"                  " Outlying-US(Guam-USVI-etc)"
## [29] " Peru"                       " Philippines"
## [31] " Poland"                     " Portugal"
## [33] " Puerto-Rico"                " Scotland"
## [35] " South"                      " Taiwan"
## [37] " Thailand"                   " Trinadad&Tobago"
## [39] " United-States"              " Vietnam"
## [41] " Yugoslavia"
```

5 Handling the Missing Value

```
# Make NA values explicit
workclass.na_explicit <- fct_na_value_to_level(demographic_info$workclass,
                                               level = "No Response")
occupation.na_explicit <- fct_na_value_to_level(demographic_info$occupation,
                                                level = "No Response")
native.country.na_explicit <- fct_na_value_to_level(demographic_info$native.country,
                                                    level = "No Response")
```

```
# Print the resulting factor
knitr::kable(table(workclass.na_explicit))
```

workclass.na_explicit	Freq
?	963
Federal-gov	472
Local-gov	1043
Never-worked	3
Private	11210
Self-emp-inc	579
Self-emp-not-inc	1321
State-gov	683
Without-pay	7
No Response	0

```
knitr::kable(table(occupation.na_explicit))
```

occupation.na_explicit	Freq
?	966
Adm-clerical	1841
Armed-Forces	6
Craft-repair	2013
Exec-managerial	2020
Farming-fishing	496
Handlers-cleaners	702
Machine-op-inspct	1020
Other-service	1628
Priv-house-serv	93
Prof-specialty	2032
Protective-serv	334
Sales	1854
Tech-support	518
Transport-moving	758
No Response	0

```
knitr::kable(table(native.country.na_explicit))
```

native.country.na_explicit	Freq
?	274
Cambodia	9
Canada	61
China	47
Columbia	26
Cuba	43
Dominican-Republic	33
Ecuador	17
El-Salvador	49
England	37
France	9
Germany	69
Greece	20
Guatemala	24
Haiti	31
Honduras	7
Hong	10
Hungary	6
India	51
Iran	16
Ireland	13
Italy	32
Jamaica	25
Japan	30
Laos	5
Mexico	308
Nicaragua	15
Outlying-US(Guam-USVI-etc)	9
Peru	15
Philippines	97
Poland	27
Portugal	30
Puerto-Rico	70
Scotland	9
South	35
Taiwan	14
Thailand	12
Trinadad&Tobago	8
United-States	14662
Vietnam	19
Yugoslavia	7
No Response	0

6 Combining Categories

```
# fct_unique() function to provide the unique values of demographic_info$education.
fct_unique(demographic_info$education)
```

```
##  [1]  10th          11th          12th          1st-4th       5th-6th
```

```
##  [6]  7th-8th       9th           Assoc-acdm    Assoc-voc     Bachelors
## [11]  Doctorate     HS-grad       Masters       Preschool     Prof-school
## [16]  Some-college
## 16 Levels:  10th  11th  12th  1st-4th  5th-6th  7th-8th  9th ...  Some-college
```

```
# Collapse all types of primary and secondary education into a single "School" category.
education.collapsed <- fct_collapse(demographic_info$education,
                                    School =c(" 11th"," HS-grad"," 10th"," 7th-8th",
                                              " 5th-6th", " 9th",
                                              " 12th"," 1st-4th"," Preschool"))

# Print the resulting factor to see the collapse effect
knitr::kable(table(education.collapsed))
```

education.collapsed	Freq
School	7438
Assoc-acdm	534
Assoc-voc	679
Bachelors	2670
Doctorate	181
Masters	934
Prof-school	258
Some-college	3587

```
# fct_unique() function to provide the unique values of demographic_info$workclass.
fct_unique(demographic_info$workclass)
```

```
## [1]  ?                 Federal-gov       Local-gov         Never-worked
## [5]  Private           Self-emp-inc      Self-emp-not-inc  State-gov
## [9]  Without-pay
## 9 Levels:  ?  Federal-gov  Local-gov  Never-worked  Private ...  Without-pay
```

```
# Collapse  all government employees (Federal, State, Local) into a "Government" category.
workclass.collapsed <- fct_collapse(demographic_info$workclass,
                                    Government =c(" Local-gov"," State-gov",
                                                  " Federal-gov"))

# Print the resulting factor to see the collapse effect
knitr::kable(table(workclass.collapsed))
```

workclass.collapsed	Freq
?	963
Government	2198
Never-worked	3
Private	11210
Self-emp-inc	579
Self-emp-not-inc	1321
Without-pay	7

7 Reordering Levels Based on Frequency

```
# Reorder factor levels by frequency
occupation_in.freq <- fct_infreq(demographic_info$occupation)

# Display the levels of the reordered factor
levels(occupation_in.freq)
```

```
##  [1] " Prof-specialty"    " Exec-managerial"   " Craft-repair"
##  [4] " Sales"             " Adm-clerical"      " Other-service"
##  [7] " Machine-op-inspct" " ?"                 " Transport-moving"
## [10] " Handlers-cleaners" " Tech-support"      " Farming-fishing"
## [13] " Protective-serv"   " Priv-house-serv"   " Armed-Forces"
```

8 Creating an Ordinal Factor

```
education.releveled <- fct_relevel(demographic_info$education, " Preschool", " 1st-4th",
                                   " 5th-6th", " 7th-8th", " 9th", " 10th", " 11th",
                                   " 12th", " HS-grad", " Some-college", " Assoc-voc",
                                   " Assoc-acdm", " Bachelors", " Masters",
                                   " Prof-school", " Doctorate")

# Display the levels of the releveled factor
levels(education.releveled)
```

```
##  [1] " Preschool"    " 1st-4th"      " 5th-6th"      " 7th-8th"
##  [5] " 9th"          " 10th"         " 11th"         " 12th"
##  [9] " HS-grad"      " Some-college" " Assoc-voc"    " Assoc-acdm"
## [13] " Bachelors"    " Masters"      " Prof-school"  " Doctorate"
```

9 Factor Level Reduction

```
# Lump together all countries with a frequency of less than 100 into an "Other" category.
native.country.lumped <- fct_lump(demographic_info$native.country, n = 3,
                                  other_level = "Other")

# Print the resulting factor to see the lump effect
table(native.country.lumped)
```

```
## native.country.lumped
##              ?         Mexico  United-States          Other
##            274            308          14662           1037
```

10 Visualization

```
# Reorder factor levels by frequency
occupation.infreq <- fct_infreq(demographic_info$occupation)

# Print the levels of the reordered factor levels by frequency
knitr::kable(table(occupation.infreq))
```

occupation.infreq	Freq
Prof-specialty	2032
Exec-managerial	2020
Craft-repair	2013
Sales	1854
Adm-clerical	1841
Other-service	1628
Machine-op-inspct	1020
?	966
Transport-moving	758
Handlers-cleaners	702
Tech-support	518
Farming-fishing	496
Protective-serv	334
Priv-house-serv	93
Armed-Forces	6

```
# Create a bar plot with the frequency order
ggplot(data.frame(Occupation = occupation.infreq), aes(x = Occupation)) +
  geom_bar(show.legend = TRUE, fill = c("turquoise1", "turquoise2",
                                        "turquoise3", "turquoise4","skyblue3",
                                        "lightskyblue3","skyblue2","paleturquoise3",
                                        "skyblue1","skyblue","lightskyblue2",
                                        "lightskyblue1", "lightblue1","paleturquoise2",
                                        "paleturquoise1")) +
  labs(title = "Distribution of Occupation By Frequency", x = "Occupation", y = "Count") +
  scale_x_discrete(guide = guide_axis(n.dodge=4)) +
  geom_text(aes(label = ..count..), stat = "count", vjust = 0, size = 3,
            color = "black", hjust = 0.5)
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

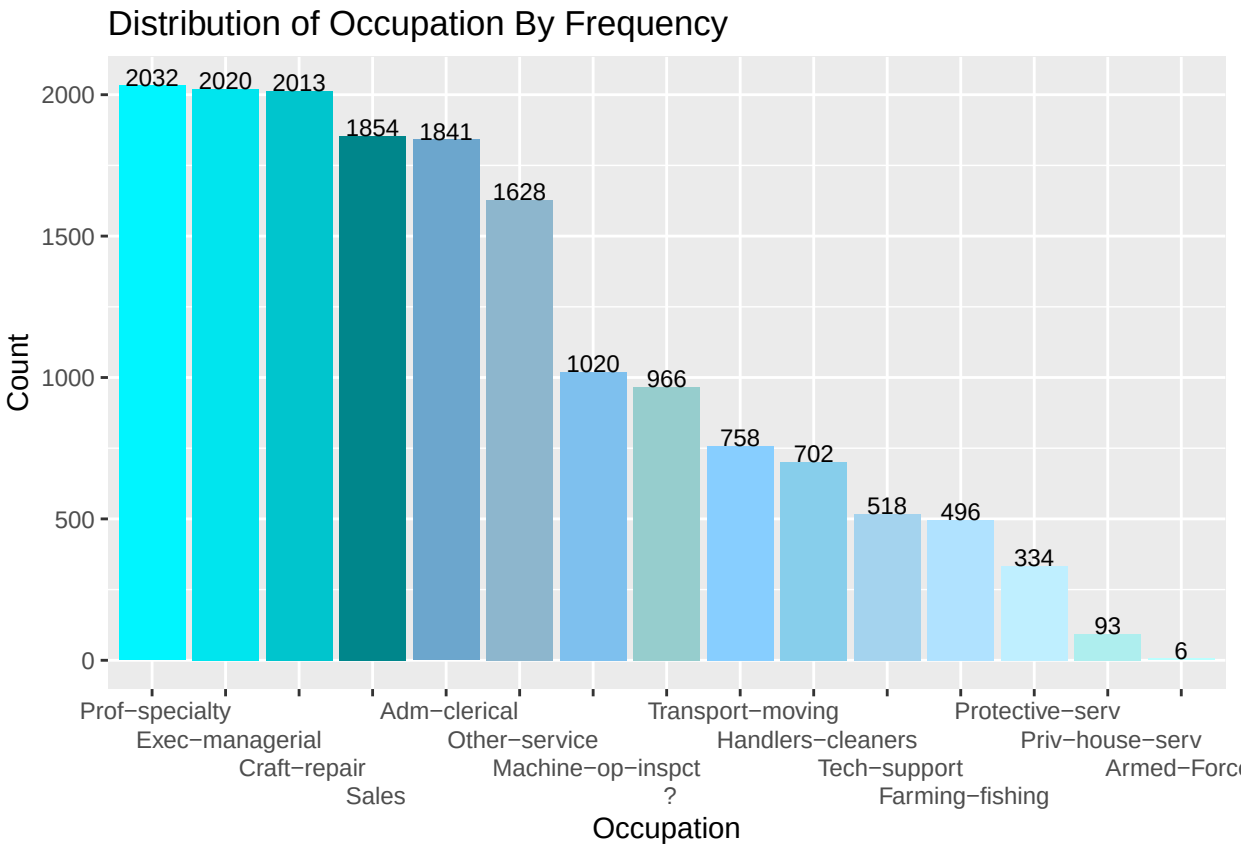

Figure 1: The Distribution of Occupation By Frequency

Interpretation

Figure 1 contains information about various occupations. The most common occupation is "Prof-specialty" with a frequency of 2032, followed closely by "Exec-managerial" with a frequency of 2020. "Craft-repair" and "Sales" are also quite common, with frequencies of 2013 and 1854, respectively. Occupations such as "Armed-Forces", "Priv-house-serv", and "Protective-serv" are among the least common, with frequencies of 6, 93, and 334, respectively. There are also 966 entries labeled as "?" which represent missing or unknown values in the dataset.

11 Summary of the Findings

Dealing with a large amount of raw data, especially like this, which is more than a thousand, can be overwhelming. We cannot simply do data inspection or data cleaning manually. It would take too much time. It involves using tools and techniques for analysis, specifically handling missing values, simplifying categories, and organizing factors. Moreover, it is beneficial to know how to deal with this large quantity of data, particularly dealing with a lot of data that is missing. values denoted as '?', ensuring accurate conclusions.

The wrangling process is so effective to use in order to improve the ways of analyzing data and also to improve data quality. There are various functions to choose from in order to achieve the way we want to analyze the data. Though it might seem confusing at first, especially since I am not too familiar with it since it was my first encounter with it when it was discussed, Nevertheless, the data wrangling procedure is efficient and improves the reliability of results.