Learning R: Descriptives and Histograms

Michelle Chiu

Last updated on 2022-02-09

Contents

Click on subheaders under table of contents to skip to that section.

This demo is adapted from DataCamp's "Tidy Analysis of Pew Research Data Using R" and R Tutorials (Debbie Yee and Sara Weston).

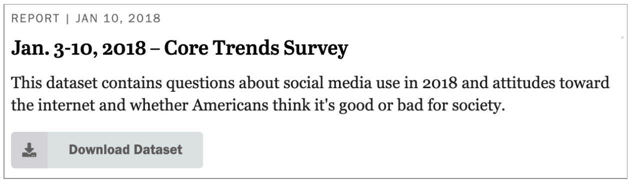
The following tutorial demonstrates how to do some basic data manipulation and plotting using survey data downloaded from the Pew Research Center.

Access the dataset

1. To access and download the dataset used in this demo, you will have to first register for a Pew Research Center account. The registration process should be fairly quick and straightforward.

2. After you've successfully signed up for an account (received an email confirmation, etc.), navigate here to find a list of datasets grouped under *Internet and Technology* on the Pew site.

3. Scroll down to the dataset labeled **Jan. 3-10, 2018 - Core Trends Survey** and click on Download Dataset.



REPORT | JAN 10, 2018

Jan. 3-10, 2018 – Core Trends Survey

This dataset contains questions about social media use in 2018 and attitudes toward the internet and whether Americans think it's good or bad for society.

4. **Once download is complete**, you should see a zipped folder named "January 3-10, 2018 - Core Trends Survey" on your local computer (likely in your Downloads folder). Depending on how your computer is set up, you may need to unzip the folder to access its contents. The unzipped folder should contain five files.

   In this demo, we will be using the following two files:

   - ***January 3-10, 2018 - Core Trends Survey - Questionnaire.docx***: A Word Doc copy of the administered survey. Akin to a *data dictionary* (See: IBM Dictionary of Computing Terminology), this doc includes the abbreviated variable names used in our dataset alongside their respective questions and answer choices, which will be crucial in data interpretation.
   - ***January 3-10, 2018 - Core Trends Survey - CSV.csv***: This CSV file is the raw dataset we will be working with.

Code

Load packages

```
library(skimr)
library(psych)
library(tidyverse)
library(ggplot2)
```

Load dataset

```
# set working directory, if needed
# working directory should match your .RProj location
setwd("~/Downloads")

# load dataset
jan_core_trends_survey <-
  read.csv("January 3-10, 2018 - Core Trends Survey - CSV.csv")
```

Pew dataset at a glance

```
# number of observations
nrow(jan_core_trends_survey)
```

```
## [1] 2002
```

```
# number of column variables
length(jan_core_trends_survey)
```

```
## [1] 70
```

```
# check out the first 6 (default) rows of the data
head(jan_core_trends_survey)
```

```
##   respid sample comp int_date lang cregion state density usr qs1 sex eminuse
## 1      1      1    1   180103    1       1    42       5   U  NA   2       1
## 2      2      1    1   180103    1       3    45       2   S  NA   2       2
## 3      3      1    1   180103    1       1    34       5   S  NA   2       1
## 4      4      1    1   180103    1       3    24       4   S  NA   2       1
## 5      5      1    1   180103    1       1    33       2   R  NA   1       1
## 6      6      1    1   180103    1       3    37       3   U  NA   1       1
##   intmob intfreq home4nw bbhome1 bbhome2 device1a smart2 snsint2 device1b
## 1      1       1       1       2      NA        1      1       1        1
## 2      2      NA      NA      NA      NA        1      2       2        2
## 3      2       3       1       2      NA        1      1       2        1
## 4      1       4       1       2      NA        1      1       1        2
## 5      1       2       1       2      NA        1      1       1        1
## 6      1       2       1       2      NA        1      1       1        1
##   device1c device1d web1a web1b web1c web1d web1e web1f web1g web1h sns2a sns2b
## 1        1        1     2     1     1     1     1     1     2     2    NA     1
## 2        2        2     2     2     2     2     2     2     2     2    NA    NA
## 3        1        2     2     2     2     2     2     2     2     2    NA    NA
## 4        1        2     2     2     1     2     2     2     2     2    NA    NA
## 5        1        1     2     2     1     2     1     2     1     1    NA    NA
## 6        1        2     1     2     1     2     1     2     1     1     2    NA
##   sns2c sns2d sns2e pial5a pial5b pial5c pial5d pial11 pial11a
## 1     1     3     3      2      1      2      3      1       1
## 2    NA    NA    NA      2      3     NA     NA      8      NA
## 3    NA    NA    NA      1      2      1     NA      1       1
## 4     3    NA    NA      2      3      3      3      2       1
## 5     3    NA     2      1      2      1      3      1       1
## 6     1    NA     3      3      5      1      1      3      NA
##                                                     pial11ao. pial11_igbm
## 1 information has become available more frequently and easier           1
## 2                                                                        9
## 3                                 it connects people together           2
## 4                               kids spend to much time on it           5
## 5                    it's just another tool for people to use           1
## 6                                                                        9
##   pial12 books1 books2a books2b books2c age marital educ2 emplnw hisp racem1
## 1      1      1       1       2       2  33       2     3      1    2      1
## 2     NA      5       1       2       2  76       1    98      3    2      1
## 3      1      0      NA      NA      NA  99       5     5      5    2      1
## 4      1      2       1       2       2  60       2     5      8    2      1
## 5      1      6       1       2       1  55       1     4      1    2      1
## 6      1     18       1       2       1  58       1     7      1    2      1
##   racem2 racem3 racem4 racecmb birth_hisp inc party partyln hh1 hh3 ql1 ql1a
## 1     NA     NA     NA       1         NA   6     2      NA   5   4   1   NA
## 2     NA     NA     NA       1         NA   4     3       8   2   2   2    2
## 3     NA     NA     NA       1         NA   4     1      NA   1  NA   1   NA
## 4     NA     NA     NA       1         NA   2     2      NA   2   2   1   NA
## 5     NA     NA     NA       1         NA   7     1      NA   3   3   1   NA
## 6     NA     NA     NA       1         NA   7     3       2   2   2   1   NA
##   qc1    weight cellweight
## 1  NA 1.7463586         NA
## 2  NA 1.6597644         NA
## 3  NA 0.4908044         NA
## 4  NA 0.9479652         NA
## 5  NA 0.9159586         NA
## 6  NA 0.4850252         NA
```

```
# skim function from skimr package prints summary stats
# outputs grouped by variable type
skim(jan_core_trends_survey)
```

Table 1: Data summary

Name	jan_core_trends_survey
Number of rows	2002
Number of columns	70

Table 1: Data summary

Column type frequency:	
character	2
numeric	68
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
usr	0	1	1	1	0	4	148
pial11ao.	0	1	1	300	0	1468	392

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
respid	0	1.00	76009.78	43691.05	1.00	100004.00	100819.00	101577.50	102430.00	
sample	0	1.00	1.75	0.43	1.00	2.00	2.00	2.00	2.00	
comp	0	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	
int_date	0	1.00	180105.64	2.28	180103.00	180104.00	180105.00	180108.00	180110.00	
lang	0	1.00	1.09	0.28	1.00	1.00	1.00	1.00	2.00	
cregion	0	1.00	2.68	1.01	1.00	2.00	3.00	3.00	4.00	
state	0	1.00	28.16	16.06	1.00	12.00	28.00	42.00	56.00	
density	0	1.00	3.05	1.42	1.00	2.00	3.00	4.00	5.00	
qs1	500	0.75	2.00	0.00	2.00	2.00	2.00	2.00	2.00	
sex	0	1.00	1.46	0.50	1.00	1.00	1.00	2.00	2.00	
eminuse	0	1.00	1.14	0.38	1.00	1.00	1.00	1.00	8.00	
intmob	0	1.00	1.19	0.49	1.00	1.00	1.00	1.00	8.00	
intfreq	217	0.89	2.15	1.13	1.00	1.00	2.00	2.00	9.00	
home4nw	217	0.89	1.21	0.61	1.00	1.00	1.00	1.00	8.00	
bbhome1	536	0.73	2.32	1.36	1.00	2.00	2.00	2.00	9.00	
bbhome2	1963	0.02	1.21	0.41	1.00	1.00	1.00	1.00	2.00	
device1a	1502	0.25	1.14	0.35	1.00	1.00	1.00	1.00	2.00	
smart2	69	0.97	1.30	0.94	1.00	1.00	1.00	1.00	9.00	
snsint2	0	1.00	1.34	0.54	1.00	1.00	1.00	2.00	8.00	
device1b	0	1.00	1.49	0.65	1.00	1.00	1.00	2.00	9.00	
device1c	0	1.00	1.25	0.46	1.00	1.00	1.00	1.00	8.00	
device1d	0	1.00	1.68	0.59	1.00	1.00	2.00	2.00	9.00	
web1a	49	0.98	1.77	0.48	1.00	2.00	2.00	2.00	9.00	
web1b	49	0.98	1.69	0.54	1.00	1.00	2.00	2.00	9.00	
web1c	49	0.98	1.34	0.61	1.00	1.00	1.00	2.00	9.00	
web1d	49	0.98	1.78	0.47	1.00	2.00	2.00	2.00	9.00	
web1e	49	0.98	1.27	0.55	1.00	1.00	1.00	2.00	9.00	
web1f	49	0.98	1.82	0.64	1.00	2.00	2.00	2.00	9.00	
web1g	49	0.98	1.76	0.64	1.00	1.00	2.00	2.00	9.00	
web1h	49	0.98	1.77	0.71	1.00	1.00	2.00	2.00	9.00	
sns2a	1544	0.23	2.76	1.44	1.00	1.00	3.00	4.00	8.00	
sns2b	1375	0.31	2.39	1.37	1.00	1.00	2.00	3.00	8.00	
sns2c	666	0.67	2.01	1.26	1.00	1.00	2.00	3.00	8.00	

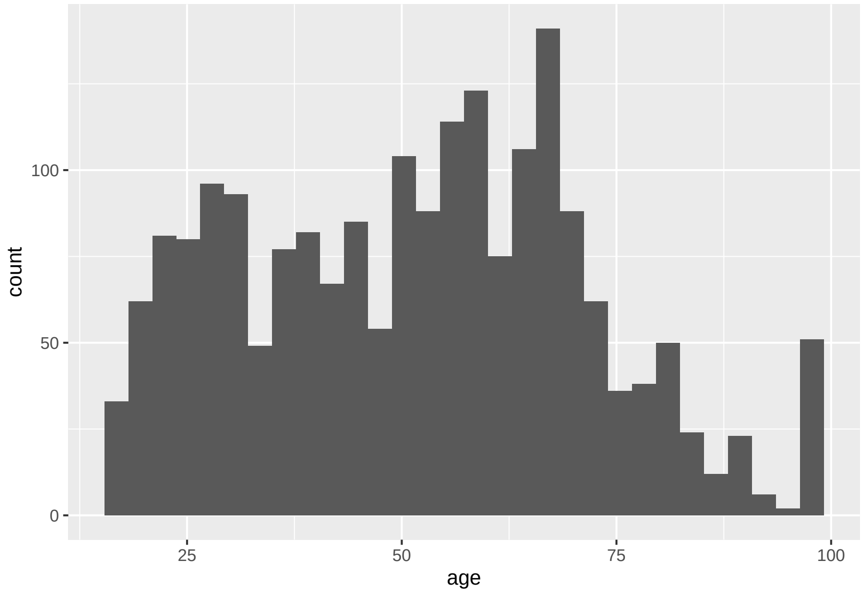
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
sns2d	1551	0.23	2.29	1.48	1.00	1.00	2.00	3.00	8.00	
sns2e	552	0.72	2.65	1.30	1.00	1.00	3.00	3.00	9.00	
pial5a	0	1.00	2.43	1.35	1.00	1.00	2.00	4.00	9.00	
pial5b	69	0.97	1.98	1.29	1.00	1.00	1.00	3.00	9.00	
pial5c	217	0.89	1.96	1.22	1.00	1.00	1.00	3.00	8.00	
pial5d	659	0.67	2.80	1.05	1.00	2.00	3.00	4.00	8.00	
pial11	0	1.00	1.73	1.51	1.00	1.00	1.00	2.00	9.00	
pial11a	370	0.82	1.10	0.84	1.00	1.00	1.00	1.00	9.00	
pial11_igbm	0	1.00	4.07	3.47	1.00	1.00	2.00	9.00	9.00	
pial12	217	0.89	1.28	1.08	1.00	1.00	1.00	1.00	9.00	
books1	0	1.00	14.70	24.89	0.00	1.00	5.00	12.75	99.00	
books2a	447	0.78	1.13	0.54	1.00	1.00	1.00	1.00	9.00	
books2b	447	0.78	1.77	0.54	1.00	2.00	2.00	2.00	9.00	
books2c	447	0.78	1.70	0.75	1.00	1.00	2.00	2.00	9.00	
age	0	1.00	51.79	19.94	18.00	35.00	53.00	66.00	99.00	
marital	0	1.00	2.98	2.26	1.00	1.00	2.00	5.00	9.00	
educ2	0	1.00	6.25	11.93	1.00	3.00	5.00	6.00	99.00	
emplnw	0	1.00	3.46	10.40	1.00	1.00	2.00	3.00	99.00	
hisp	0	1.00	1.98	1.07	1.00	2.00	2.00	2.00	9.00	
racem1	0	1.00	2.10	2.23	1.00	1.00	1.00	2.00	9.00	
racem2	1948	0.03	3.76	2.03	1.00	2.00	4.00	5.00	7.00	
racem3	1995	0.00	4.43	2.23	1.00	3.00	5.00	6.00	7.00	
racem4	1999	0.00	7.00	0.00	7.00	7.00	7.00	7.00	7.00	
racecmb	0	1.00	2.04	2.01	1.00	1.00	1.00	2.00	9.00	
birth_hisp	1679	0.16	2.26	1.15	1.00	1.00	3.00	3.00	9.00	
inc	0	1.00	21.07	35.13	1.00	3.00	6.00	9.00	99.00	
party	0	1.00	2.71	1.84	1.00	2.00	2.00	3.00	9.00	
partyln	1048	0.48	4.63	3.40	1.00	2.00	2.00	8.00	9.00	
hh1	0	1.00	3.01	1.94	1.00	2.00	2.00	4.00	9.00	
hh3	406	0.80	2.76	1.62	1.00	2.00	2.00	3.00	9.00	
ql1	1502	0.25	1.40	1.47	1.00	1.00	1.00	1.00	9.00	
ql1a	1952	0.02	4.36	3.69	1.00	1.00	2.00	9.00	9.00	
qc1	500	0.75	1.80	1.21	1.00	1.00	2.00	2.00	9.00	
weight	0	1.00	1.00	0.48	0.38	0.63	0.89	1.28	2.11	
cellweight	500	0.75	1.00	0.45	0.43	0.65	0.89	1.24	2.04	

```r
# glimpse at the dataset
# glimpse(jan_core_trends_survey)

# print first ten responses for age variable
head(jan_core_trends_survey$age, n = 10)
```

```
##  [1] 33 76 99 60 55 58 99 72 58 68
```

Histogram of Age

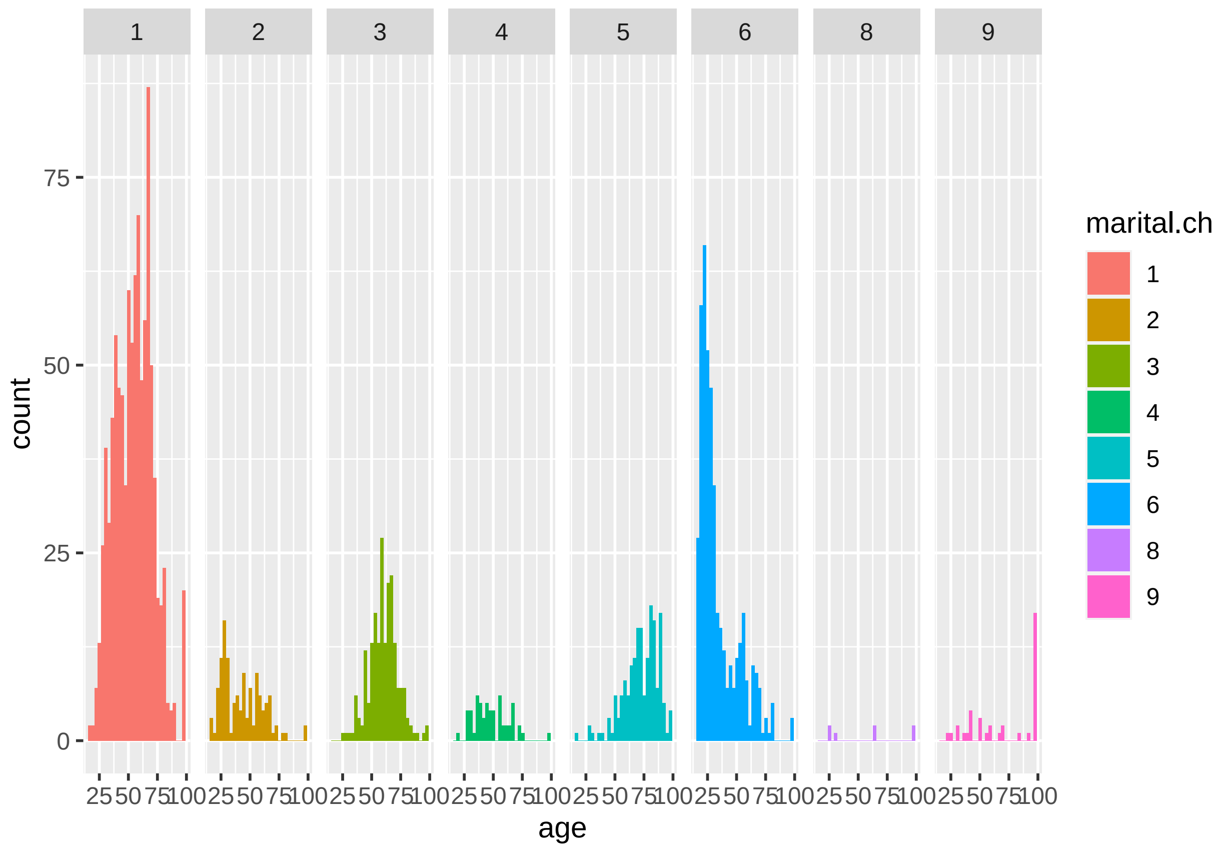
```r
# age distribution
ggplot(data = jan_core_trends_survey, aes(x = age)) +
  geom_histogram()
```



Group data by category

Example. Age distribution by marital status

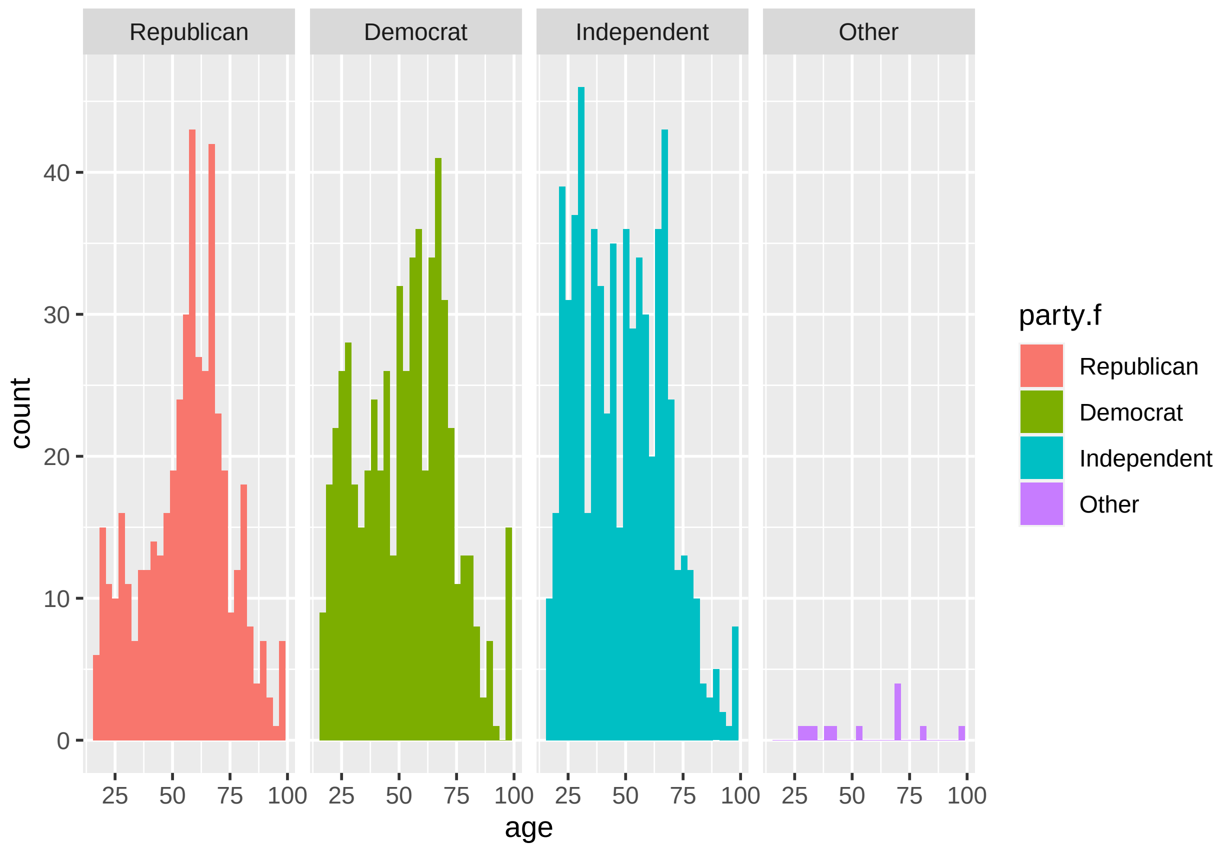
```r
# convert marital variable from numeric to character
jan_core_trends_survey$marital.ch <- as.character(jan_core_trends_survey$marital)

# plot
ggplot(data = jan_core_trends_survey, aes(x = age, fill = marital.ch)) +
  geom_histogram() +
  facet_grid(~ marital.ch)
```



Filter data by category levels

Example 1. Age distribution by political party (excluding "No preference," "Don't know" or "Refused")

```
# select relevant variables, then filter party variable
data_age_party <-
  jan_core_trends_survey %>%
  select(age, party) %>%
  filter(party==1 | party == 2 | party ==3 | party == 5)

# convert party to character
data_age_party$party.ch <- as.character(data_age_party$party)

# create factor for party variable, add meaningful labels for levels
data_age_party$party.f <- factor(data_age_party$party.ch,
                                 levels = c(1, 2, 3, 5),
                                 labels = c("Republican", "Democrat",
                                            "Independent", "Other"))

# plot
ggplot(data = data_age_party, aes(x = age, fill = party.f)) +
  geom_histogram() +
  facet_grid(~ party.f)
```

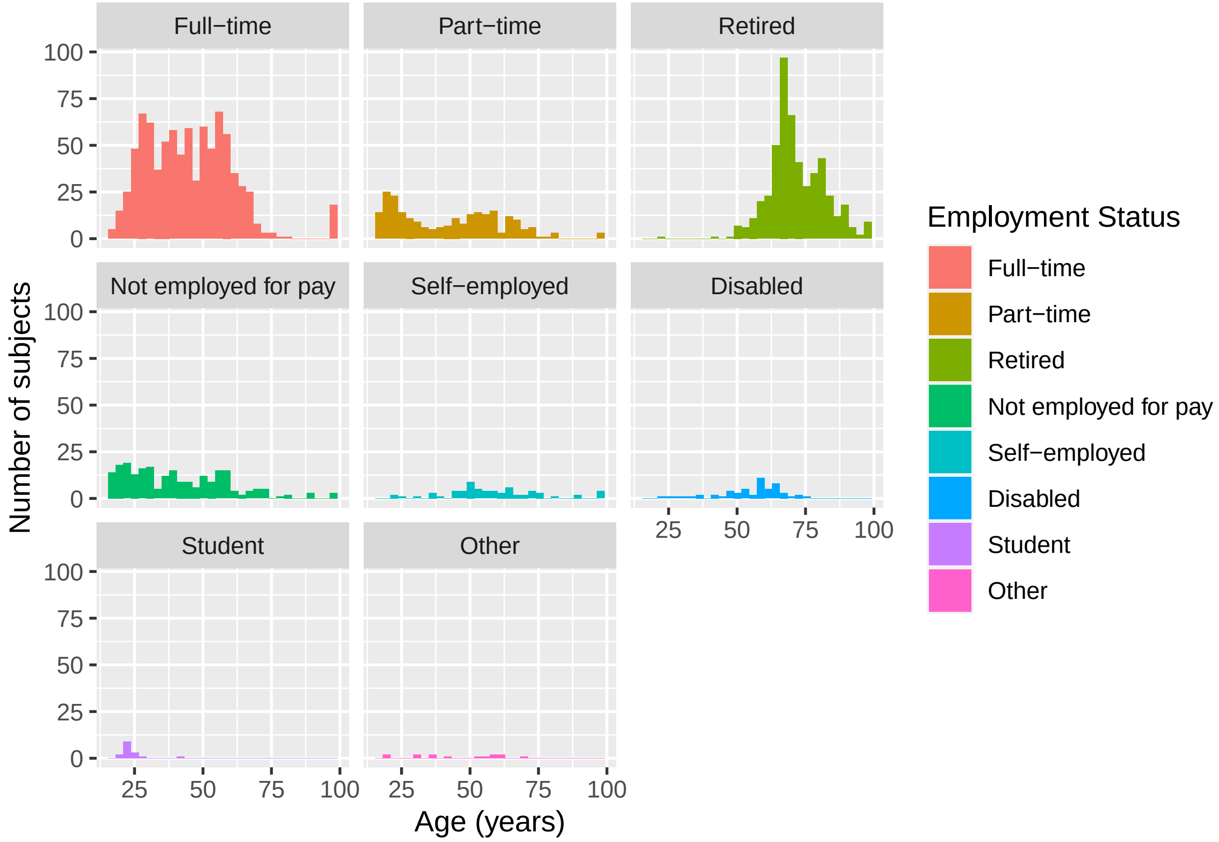


Example 2. Age distribution by employment status (excluding “Don’t know” or “Refused”)

```
# filter employment variable
data_age_employment <-
  jan_core_trends_survey %>%
  select(age, emplnw) %>%
  filter(emplnw < 9)

# check filtering using describe function
# min and max should be 1 and 8, respectively
describe(data_age_employment$emplnw)
```

```
##    vars    n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 1979 2.35 1.49      2    2.13 1.48   1   8     7 1.08     0.99 0.03
```

```
# plot histogram for filtered dataset
# convert employment to character
data_age_employment$emplnw.ch <- as.character(data_age_employment$emplnw)

# create new factor variable with new level names for party
data_age_employment$emplnw.f <- factor(data_age_employment$emplnw.ch,
                                       levels = c(1, 2, 3, 4,
                                                  5, 6, 7, 8),
                                       labels = c("Full-time", "Part-time",
                                                  "Retired", "Not employed for pay",
                                            "Self-employed", "Disabled",
                                            "Student", "Other"))

# plot
ggplot(data = data_age_employment, aes(x = age, fill = emplnw.f)) +
  geom_histogram() +
  facet_wrap(~ emplnw.f) +
  # add label names for both axes and fill legend
  labs(x = "Age (years)", y = "Number of subjects") +
  guides(fill = guide_legend(title = "Employment Status"))
```


Full−time
Part−time
Retired
Not employed for pay
Self−employed
Disabled
Student
Other
Number of subjects
Age (years)
Employment Status
Full−time
Part−time
Retired
Not employed for pay
Self−employed
Disabled
Student
Other