Learning R: Descriptives and Histograms

Michelle Chiu

Last update: October 20, 2020

This demo is adapted from DataCamp's Tidy Analysis of Pew Research Data Using R.

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

1 Dataset

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.

2 Code

2.1 Load packages

```
library(tidyverse) # data manipulation
library(ggplot2) # plotting
```

2.2 Load dataset

```
# set working directory, if needed
# working directory should match your .RProj location
setwd("~/Desktop/fall-2020/R-workshop")

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

2.3 Dataset at a glance

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

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

```
length(jan_core_trends_survey)
```

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

```
# answers to the question "do you use the internet or email, at least occasionally"
# 1=yes, 2=no, 8=dont know, 9=Refused
unique(jan_core_trends_survey$eminuse)
```

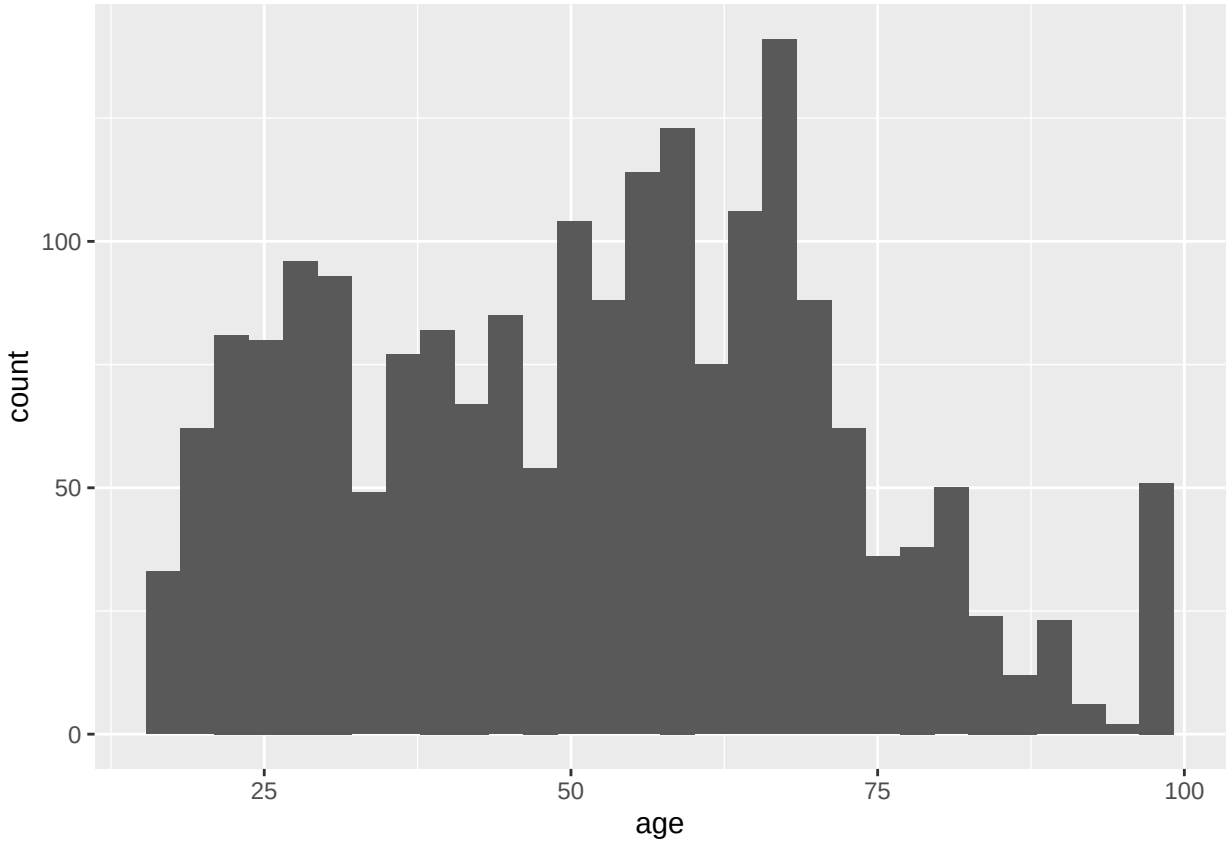
```
## [1] 1 2 8
```

```
# see 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
```

2.4 Histogram of Age

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



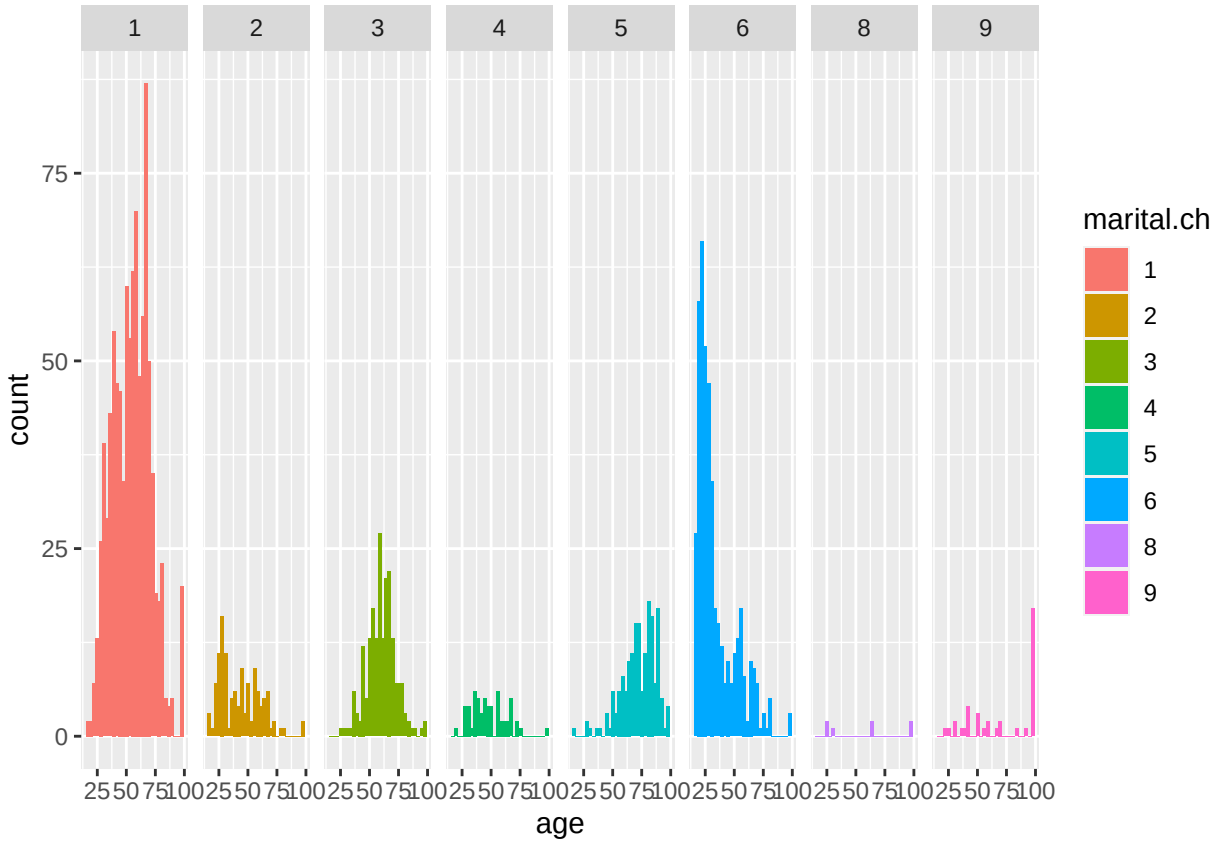
2.5 Grouping data by category

2.5.1 Age distribution by marital status

```
# 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)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2.6 Filtering data by category levels

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

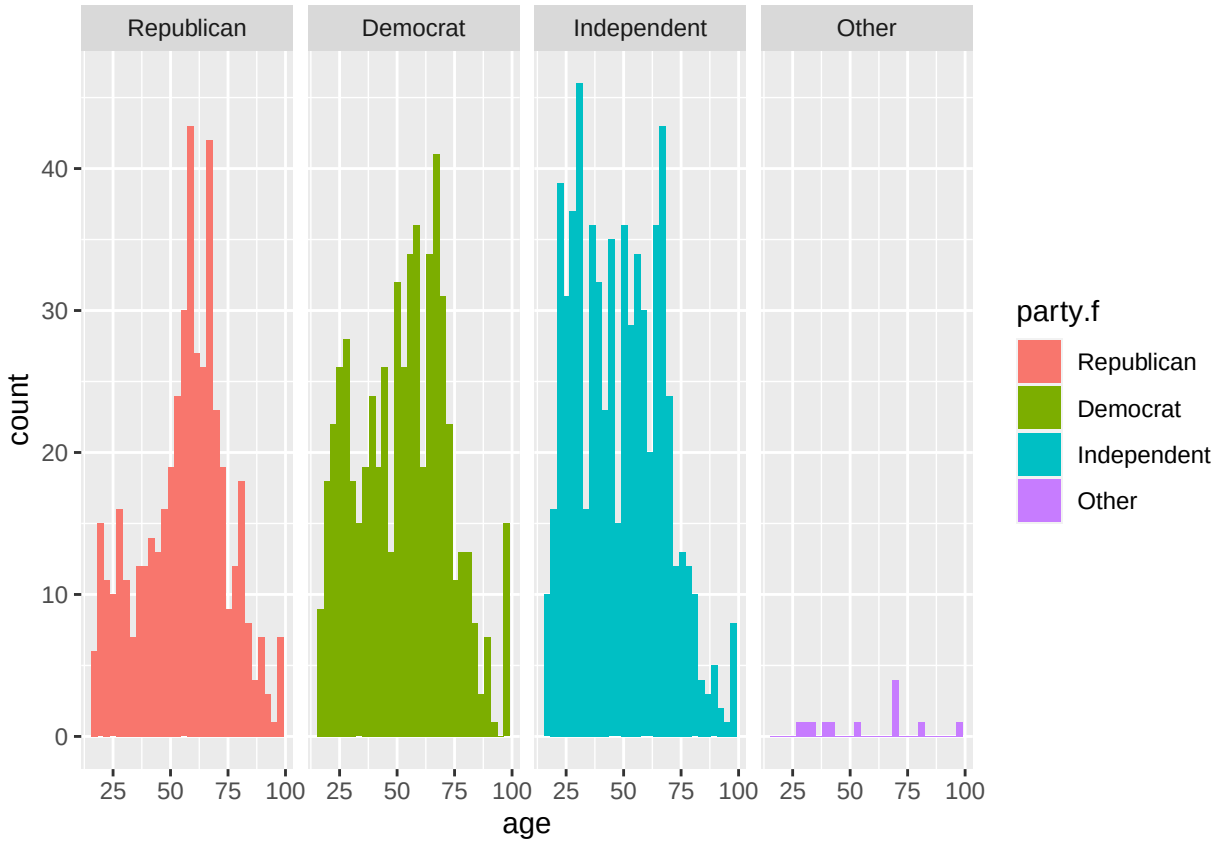
```
# filter party variable
data_age_party <-
  jan_core_trends_survey %>%
  select(age, party) %>%
  filter(party==1 | party == 2 | party ==3 | party == 5)

# 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)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2.6.2 Age distribution by employment status (excluding "Don't know" or "Refused")

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

# check
unique(data_age_employment$emplnw)
```

```
## [1] 1 3 5 8 4 2 6 7
```

```r
# 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),
```

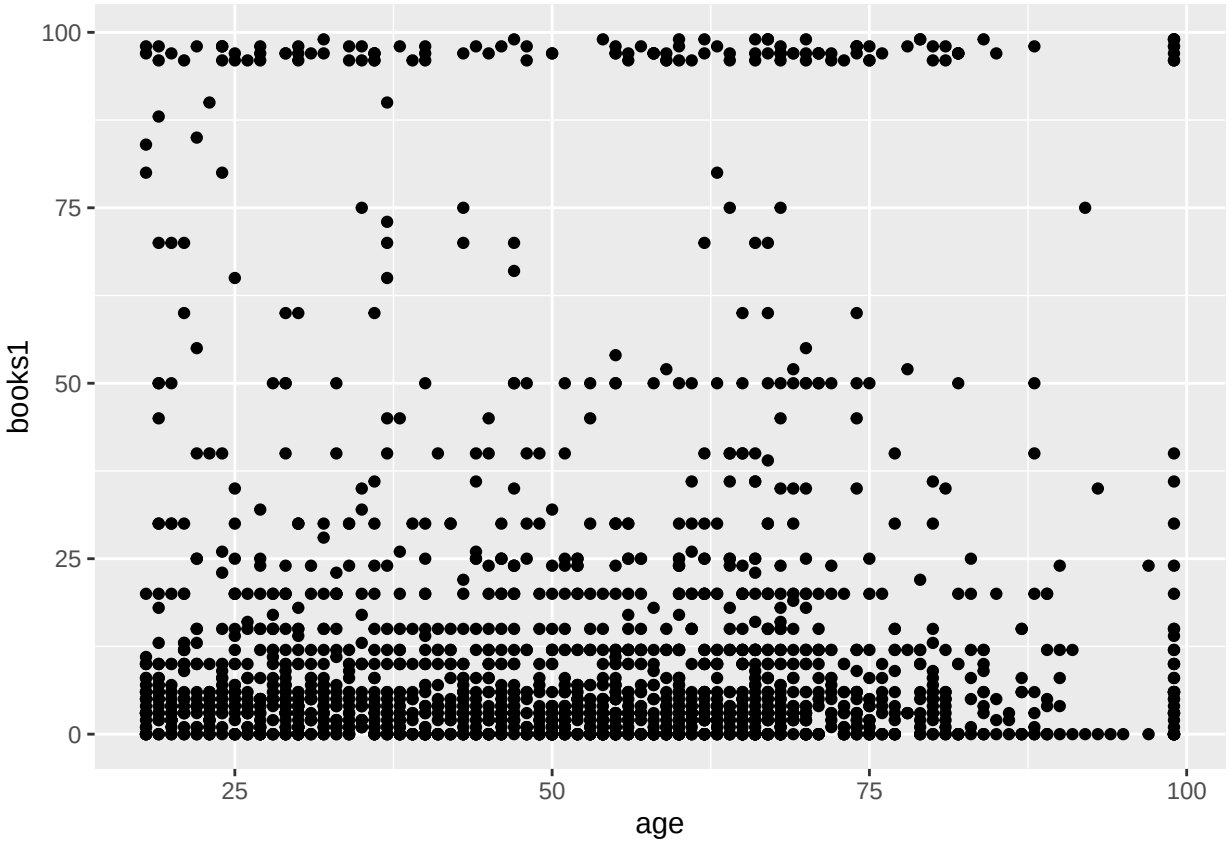
```
                                labels = c("Full-time", "Part-time",
                                           "Retired", "Not employed for pay",
                                           "Self-employed", "Disabled",
                                           "Student", "Other"))

# plot
# facet_grid vs facet_wrap
ggplot(data = data_age_employment, aes(x = age, fill = emplnw.f)) +
  geom_histogram() +
  facet_wrap(~ emplnw.f)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Scatterplot of Age X Books ### During the past 12 months, about how many BOOKS did you read either all or part of the way through? Please include any print, electronic, or audiobooks you may have read or listened to.

```
ggplot(data = jan_core_trends_survey, aes(x = age, y = books1)) +
  geom_point()
```


100
75
50
25
0
books1
25
50
75
100
age