HW 3, STAT 650

Due: Friday, October 11

Directions: This assignment should be completed using Quarto and submitted to Canvas as a self-contained HTML or PDF file.

Reading: Chapter 6 from *Modern Data Science with R*

```r
library(tidyverse)
```

Exercise 1

The data set `tech_stock.csv` contains daily stock prices in the year 2021 for three different tech companies: Apple (AAPL), Amazon (AMZN), and Alphabet (GOOGL). Variable descriptions:

- `company`: Company name, abbreviated with stock symbol (AAPL, AMZN, GOOGL)
- `date`: date
- `high`: the highest price for that day
- `low`: the lowest price for that day

a

Use the `read_csv()` function to read `tech_stock.csv` into R. You can download the CSV file from Canvas. After reading in the file, answer the following questions: What are the dimensions of the data frame (i.e., number of rows and columns)? What are the data types for the columns?

b

Make side-by-side box plots of the `high` price for the three tech companies.

c

Add a new column to the `tech_stock` data frame called `diff`, which is the difference between the `high` and `low` price for each day. Then use faceting to make a histogram of `diff` for each tech company. (Your visualization should contain three panels of density plots.)

d

Group the rows of the `tech_stock` data frame by `company`. Then for each company compute the following summary statistics: mean of `high`, standard deviation of `high`, mean of `low`, and standard deviation of `low`.

Exercise 2

Consider the following data from a Pew religion and income survey.

```
relig_income
```

```
## # A tibble: 18 x 11
##    religion `<$10k` `$10-20k` `$20-30k` `$30-40k` `$40-50k` `$50-75k` `$75-100k`
##    <chr>      <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>      <dbl>
##  1 Agnostic      27        34        60        81        76       137        122
##  2 Atheist       12        27        37        52        35        70         73
##  3 Buddhist      27        21        30        34        33        58         62
##  4 Catholic     418       617       732       670       638      1116        949
##  5 Don’t k~      15        14        15        11        10        35         21
##  6 Evangel~     575       869      1064       982       881      1486        949
##  7 Hindu          1         9         7         9        11        34         47
##  8 Histori~     228       244       236       238       197       223        131
##  9 Jehovah~      20        27        24        24        21        30         15
## 10 Jewish        19        19        25        25        30        95         69
## 11 Mainlin~     289       495       619       655       651      1107        939
## 12 Mormon        29        40        48        51        56       112         85
## 13 Muslim         6         7         9        10         9        23         16
## 14 Orthodox      13        17        23        32        32        47         38
## 15 Other C~       9         7        11        13        13        14         18
## 16 Other F~      20        33        40        46        49        63         46
## 17 Other W~       5         2         3         4         2         7          3
## 18 Unaffil~     217       299       374       365       341       528        407
## # i 3 more variables: `$100-150k` <dbl>, `>150k` <dbl>,
## #   `Don't know/refused` <dbl>
```

Use the `pivot_longer()` function to reshape `relig_income` into a tidy data set, with the variables along the columns and observations along the rows. Your code should produce the following output:

```
## # A tibble: 180 x 3
##    religion income             count
##    <chr>    <chr>              <dbl>
##  1 Agnostic <$10k                 27
##  2 Agnostic $10-20k               34
##  3 Agnostic $20-30k               60
##  4 Agnostic $30-40k               81
##  5 Agnostic $40-50k               76
##  6 Agnostic $50-75k              137
##  7 Agnostic $75-100k             122
##  8 Agnostic $100-150k            109
##  9 Agnostic >150k                 84
## 10 Agnostic Don't know/refused    96
## # i 170 more rows
```

Exercise 3

An analyst wants to calculate the pairwise differences between the treatment and control values for a small data set from a crossover trial (all subjects received both treatments) that consists of the following observations:

```
tb1 <- tibble(
  id = c(1:4, 1:4),
  group = c("t", "t", "t", "t", "c", "c", "c", "c"),
  vals = c(4, 6, 8, 11, 5, 6, 10, 16)
)
```

```
tb1
```

```
## # A tibble: 8 x 3
##      id group  vals
##   <int> <chr> <dbl>
## 1     1 t         4
## 2     2 t         6
## 3     3 t         8
## 4     4 t        11
## 5     1 c         5
## 6     2 c         6
## 7     3 c        10
## 8     4 c        16
```

Use the `pivot_wider()` and `mutate()` functions to transform this data table into the following format, which has a column with the differences between the control and treatment group values.

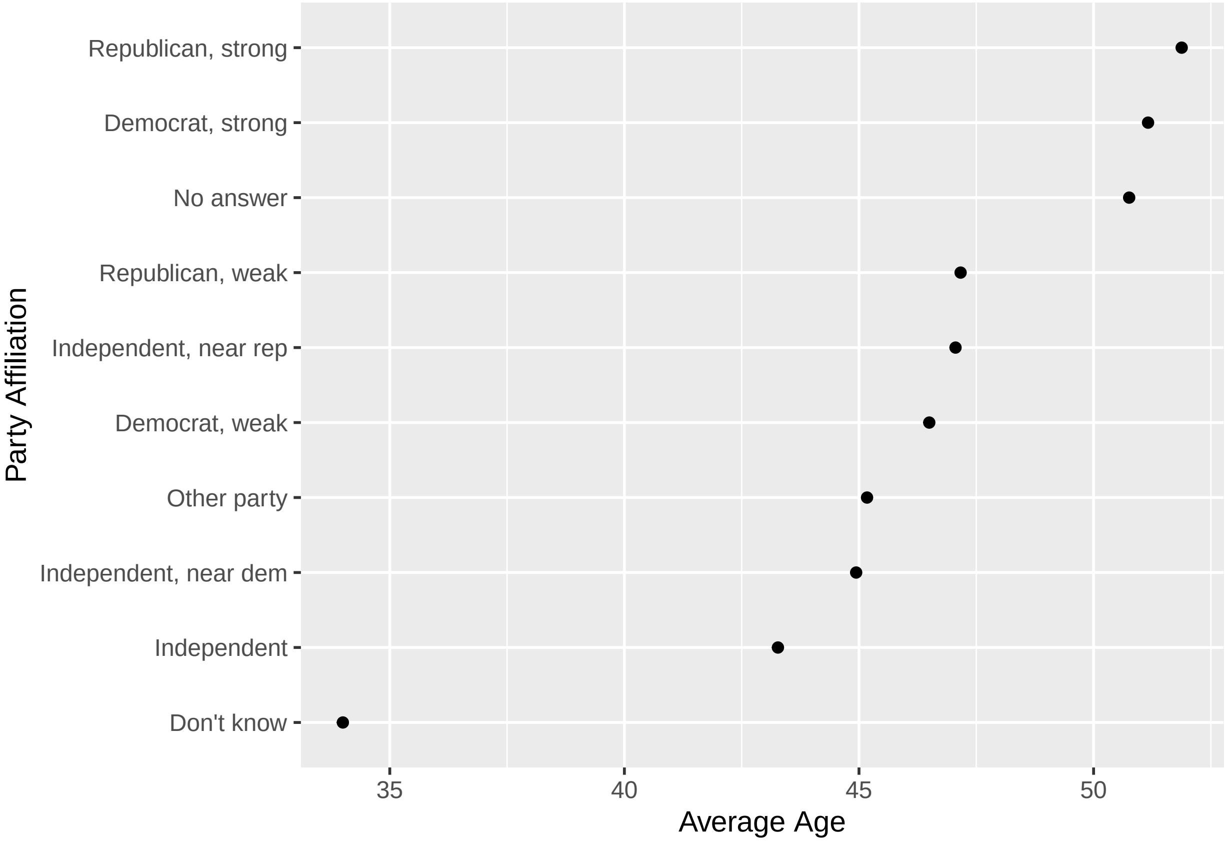
```
## # A tibble: 4 x 4
##      id     t     c  diff
##   <int> <dbl> <dbl> <dbl>
## 1     1     4     5    -1
## 2     2     6     6     0
## 3     3     8    10    -2
## 4     4    11    16    -5
```

Exercise 4

Run the following code to update the factor `partyid` with better names for the levels:

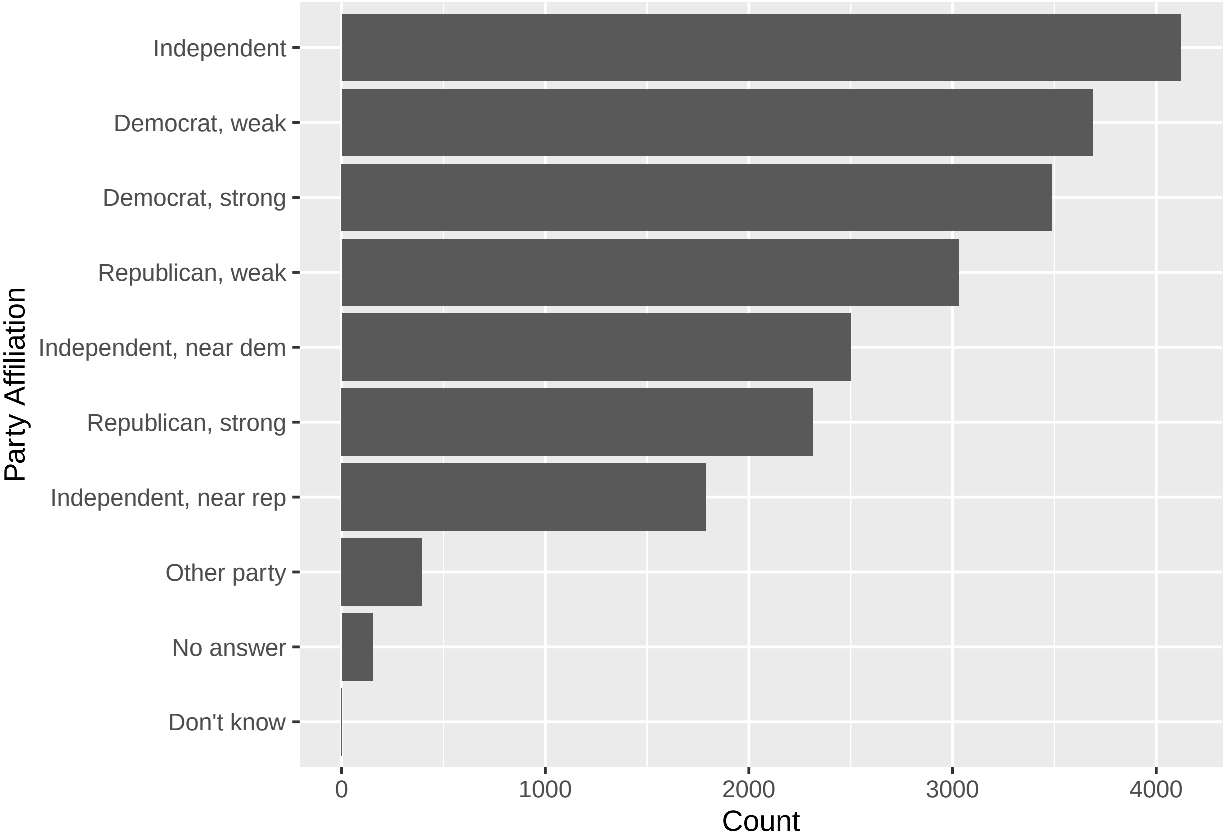
```
gss_cat2 <- gss_cat |>
  mutate(partyid = fct_recode(partyid,
    "Republican, strong"    = "Strong republican",
    "Republican, weak"      = "Not str republican",
    "Independent, near rep" = "Ind,near rep",
    "Independent, near dem" = "Ind,near dem",
    "Democrat, weak"        = "Not str democrat",
    "Democrat, strong"      = "Strong democrat"
  ))
```

Next use `group_by()` and `summarize()` to compute the average age for each category of `partyid`. Then recreate the R code that makes the graph below.



Exercise 5

Recreate the R code that makes the graph below. When creating this graph use the data frame `gss_cat2` which has the updated names for the levels of `partyid`.


Independent
Democrat, weak
Democrat, strong
Republican, weak
Independent, near dem
Republican, strong
Independent, near rep
Other party
No answer
Don't know
Party Affiliation
Count