Class Activity 10

Arthur Viegas Eguia

April 15 2024

Your Turn 1

```
set.seed(123)
students <- tibble(
  id = 1:24,
  grade = sample(c("9th", "10th", "11th"), 24, replace = TRUE),
  region = sample(c("North America", "Europe", "Asia", "South America", "Middle East", "Africa"), 24, r
  score = round(runif(24,50, 100))
)
```

a. Create a new column `grade_fac` by converting the grade column into a factor. Reorder the levels of `grade_fac` to be "9th", "10th", and "11th". Sort the dataset based on the `grade_fac` column.

Answer:

```
students_a <- students %>%
  mutate(grade_factor = factor(grade)) %>%
  mutate(grade_factor =
           fct_relevel(grade_factor,
                       c("9th", "10th", "11th")))
levels(students_a$grade_factor)
[1] "9th"  "10th" "11th"
```

b. Create a new column `region_fac` by converting the `region` column into a factor. Collapse the `region_fac` levels into three categories: "Americas", "EMEA" and "Asia". Count the number of students in each collapsed region category.

```
students_b <- students_a %>% mutate(region_factor = factor(region)) %>%
  mutate(region_collapsed =
           fct_collapse(region_factor, Americas = c("North America", "South America"),
                        EMEA = c("Europe", "Middle East", "Africa"),
                        Asia = c("Asia"))) %>%
count(region_collapsed) #alternate for group by and summarize
students_b
# A tibble: 3 x 2
```

```
  region_collapsed     n
  <fct>            <int>
1 EMEA                12
2 Asia                 4
3 Americas             8
```

c. Create a new column `grade_infreq` that is a copy of the `grade_fac` column. Reorder the levels of `grade_infreq` based on their frequency in the dataset. Print the levels of `grade_infreq` to check the ordering.

```
students_c <- students_a %>%
  mutate(grade_infrequent = grade_factor) %>%
  mutate(grade_infrequent = fct_infreq(grade_infrequent))

students_c
# A tibble: 24 x 6
      id grade region        score grade_factor grade_infrequent
   <int> <chr> <chr>         <dbl> <fct>        <fct>
 1     1 11th  Europe           83 11th         11th
 2     2 11th  North America    55 11th         11th
 3     3 11th  Africa           69 11th         11th
 4     4 10th  Asia             64 10th         10th
 5     5 11th  South America    91 11th         11th
 6     6 10th  Africa           72 10th         10th
 7     7 10th  North America    91 10th         10th
 8     8 10th  Asia             91 10th         10th
 9     9 11th  Middle East      90 11th         11th
10    10 9th   South America    72 9th          9th
# i 14 more rows
students_a %>%  mutate(grade_infreq = grade_factor) %>%  count(grade_infreq)
# A tibble: 3 x 2
  grade_infreq     n
  <fct>        <int>
1 9th              7
2 10th             8
3 11th             9
```

d. Create a new column `grade_lumped` by lumping the least frequent level of the `grade_fac` column into an 'Others' category.

Count the number of students in each of the categories of the `grade_lumped` column.

```
students_d <- students_a %>%
  mutate(grade_lumped = fct_lump(grade_factor, n = 1, other_level = "others")) %>%  count(grade_lumped)

students_d
# A tibble: 2 x 2
  grade_lumped     n
  <fct>        <int>
1 11th             9
2 others          15
```

Your Turn 2

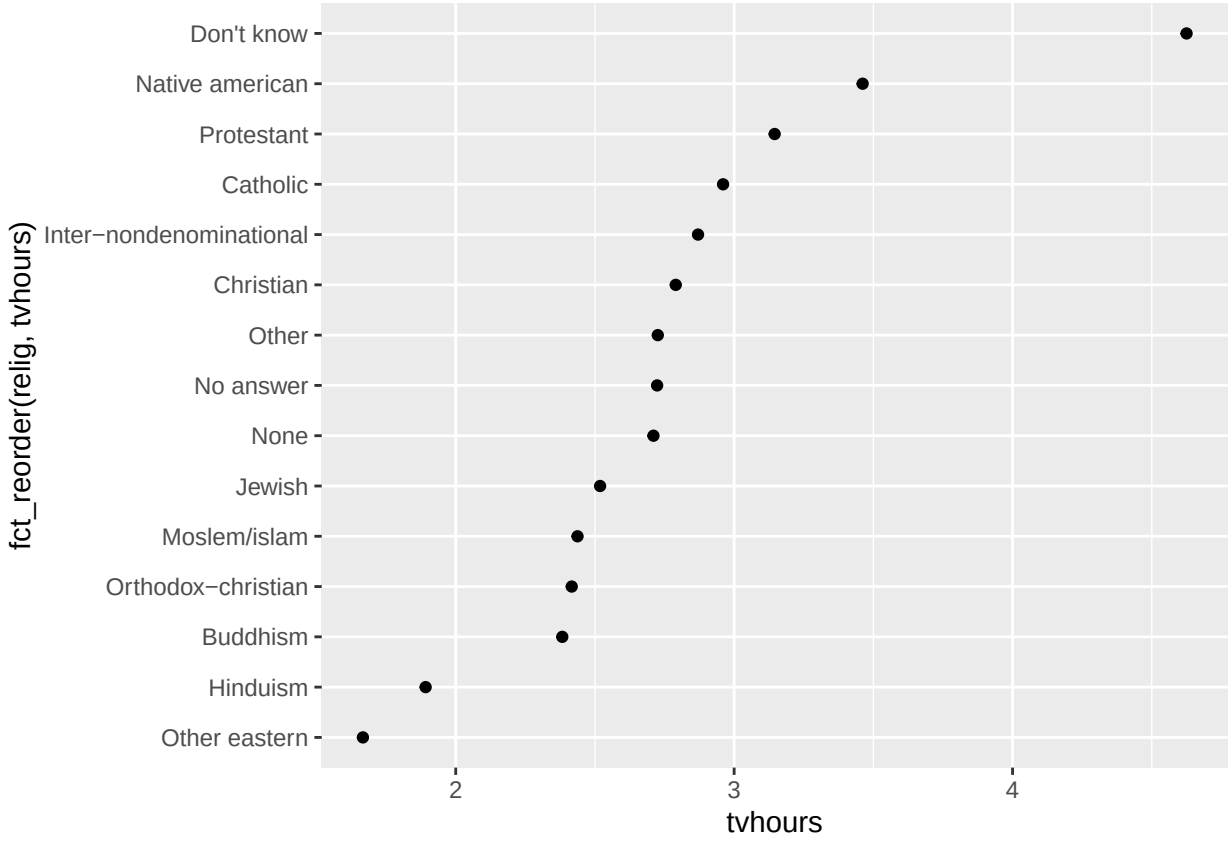
Lets import the `gss_cat` dataset from the `forcats` library. This dataset contains a sample of categorical variables from the General Social survey.

```
# import gss_cat dataset from forcats library
forcats::gss_cat
# A tibble: 21,483 x 9
    year marital         age race  rincome        partyid    relig denom tvhours
   <int> <fct>         <int> <fct> <fct>          <fct>      <fct> <fct>   <int>
 1  2000 Never married    26 White $8000 to 9999  Ind,near ~ Prot~ Sout~      12
 2  2000 Divorced         48 White $8000 to 9999  Not str r~ Prot~ Bapt~      NA
 3  2000 Widowed          67 White Not applicable Independe~ Prot~ No d~       2
 4  2000 Never married    39 White Not applicable Ind,near ~ Orth~ Not ~       4
 5  2000 Divorced         25 White Not applicable Not str d~ None  Not ~       1
 6  2000 Married          25 White $20000 - 24999 Strong de~ Prot~ Sout~      NA
 7  2000 Never married    36 White $25000 or more Not str r~ Chri~ Not ~       3
 8  2000 Divorced         44 White $7000 to 7999  Ind,near ~ Prot~ Luth~      NA
 9  2000 Married          44 White $25000 or more Not str d~ Prot~ Other       0
10  2000 Married          47 White $25000 or more Strong re~ Prot~ Sout~       3
# i 21,473 more rows
```

Use `gss_cat` to answer the following questions.

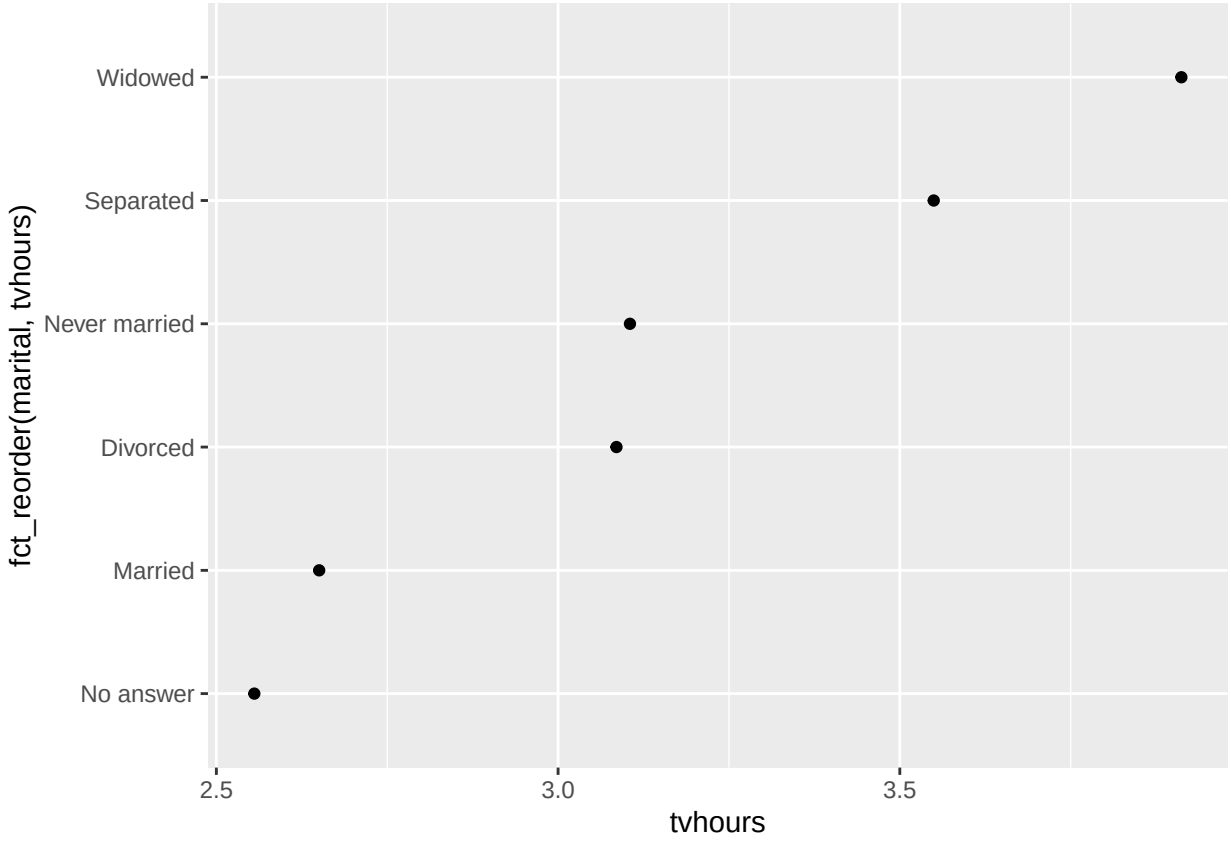
a. Which religions watch the least TV?

```
# your r-code

gss_cat %>% drop_na(tvhours) %>%
  group_by(relig) %>%
  summarise(tvhours = mean(tvhours, na.rm = TRUE)) %>%
  ggplot(aes(x = tvhours, y = fct_reorder(relig, tvhours))) + geom_point()
```



b. Do married people watch more or less TV than single people?

```
# your r-code

gss_cat %>% drop_na(tvhours) %>%
  group_by(marital) %>%
  summarise(tvhours = mean(tvhours, na.rm = TRUE)) %>%
  ggplot(aes(x = tvhours, y = fct_reorder(marital, tvhours))) + geom_point()
```



c. Collapse the `marital` variable to have levels `Married`, `Not_married`, and `No_answer` . Include `"Never married"`, `"Divorced"`, and "`Widowed`" in `Not_married`

```
# your r-code

gss_cat %>% mutate(marital_collapsed =
                     fct_collapse(marital,
                                  Married = c("Married", "Separated"),
                                  Not_married = c("Never Married", "Divorced", "Widowed"),
                                  No_answer = c("No answer")))
# A tibble: 21,483 x 10
    year marital         age race  rincome        partyid    relig denom tvhours
   <int> <fct>         <int> <fct> <fct>          <fct>      <fct> <fct>   <int>
 1  2000 Never married    26 White $8000 to 9999  Ind,near ~ Prot~ Sout~      12
 2  2000 Divorced         48 White $8000 to 9999  Not str r~ Prot~ Bapt~      NA
 3  2000 Widowed          67 White Not applicable Independe~ Prot~ No d~       2
 4  2000 Never married    39 White Not applicable Ind,near ~ Orth~ Not ~       4
 5  2000 Divorced         25 White Not applicable Not str d~ None  Not ~       1
 6  2000 Married          25 White $20000 - 24999 Strong de~ Prot~ Sout~      NA
 7  2000 Never married    36 White $25000 or more Not str r~ Chri~ Not ~       3
 8  2000 Divorced         44 White $7000 to 7999  Ind,near ~ Prot~ Luth~      NA
 9  2000 Married          44 White $25000 or more Not str d~ Prot~ Other       0
10  2000 Married          47 White $25000 or more Strong re~ Prot~ Sout~       3
# i 21,473 more rows
# i 1 more variable: marital_collapsed <fct>
```