Lab 02

CS3172-1, Spring 2023, Effat University

Mehreen Junaid

Packages

```
library(tidyverse)
library(scales)
```

Data

```
cas <- read_rds("data/canada_survey.rds")
```

Tasks:

Task 1:

• Subset `cas` so that variables `energy_expense` and `household_income` only contain values greater than 0. Overwrite `cas`.

```
cas <- filter(cas,energy_expense > 0, household_income > 0)
```

• Use function `factor()` to change the variable `marital_status` to be a factor rather than double. Overwrite `cas`. Consult the data dictionary and write-out what the marital status codes mean.

```
cas$marital_status <- factor(cas$marital_status)
# 1 = married or common law
# 2 = never married
# 3 = other (separated, divorced, or widowed)
```

Task 2

Use function `case_when()` from `dplyr` to recode the two variables below. Overwrite `cas` after each recoding. Do not use function `if_else()` to complete this task.

Recode `heat_equip` so instead of having values 1, 2, 3, 4, it contains values `"steam"`, `"forced air"`, `"stove"`, and `"electric heating"` according to the data dictionary. These new values are as defined below: o `steam`: steam or water furnace o `forced air`: forced air furnace o `stove`: heating stoves, cookstove, or other o `electric heating`: electric

```
cas <- cas %>%
  mutate(heat_equip = case_when(
```

```
    heat_equip == 1 ~ "steam",
    heat_equip == 2 ~ "forced air",
    heat_equip == 3 ~ "stove",
    heat_equip == 4 ~ "electric heating",
    TRUE ~ heat_equip
  ))
```

Recode `heat_fuel` so instead of having values 1, 2, 3, 4, it contains values `"oil"`, `"gas"`, `"electricity"`, and `"other"` according to the data dictionary. These new values are as defined below: o `oil`: oil or other liquid fuel o `gas`: natural gas o `electricity`: electricity o `other`: bottled gas, wood, or other

```
cas <- cas %>%
  mutate(heat_fuel = case_when(
    heat_fuel == 1 ~ "oil",
    heat_fuel == 2 ~ "gas",
    heat_fuel == 3 ~ "electricity",
    heat_fuel == 4 ~ "other",
    TRUE ~ heat_fuel
  ))
write.csv(cas, "cas.csv", row.names = FALSE)
```

Task 3

For each combination of heating fuel type and heating equipment, find the mean, median, and standard deviation of household energy expenditures. Print your results.

```
cas %>%
  group_by(heat_fuel, heat_equip) %>%
  summarize(mean_energy_expense = mean(energy_expense, na.rm = TRUE),
            median_energy_expense = median(energy_expense, na.rm = TRUE),
            sd_energy_expense = sd(energy_expense, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'heat_fuel'. You can override using
the
## `.groups` argument.
```

```
## # A tibble: 14 × 5
## # Groups:   heat_fuel [4]
##    heat_fuel   heat_equip       mean_energy_expense median_energy_expe…¹
sd_en…²
##    <chr>       <chr>                          <dbl>                <dbl>
<dbl>
##  1 electricity electric heating               2084.                1956
1270.
##  2 electricity forced air                     2590.                2462.
1293.
##  3 electricity steam                          1708.                 915
1692.
##  4 electricity stove                          2443.                2120
1229.
```

```
##  5 gas         forced air                    3047.                2960
1395.
##  6 gas         steam                         1698.                 720
1820.
##  7 gas         stove                         2178.                2202
1024.
##  8 oil         forced air                    3499.                3200
2156.
##  9 oil         steam                         2887.                2900
2142.
## 10 oil         stove                         3396.                3395
2074.
## 11 other       electric heating              3240                 3240
NA
## 12 other       forced air                    2861.                2526
1655.
## 13 other       steam                         2047.                1555
2279.
## 14 other       stove                         2210.                2025
1140.
## # … with abbreviated variable names ¹median_energy_expense, ²
sd_energy_expense
```

- What combination of fuel type and equipment has the highest average energy expense?

The combination of fuel type "oil" and equipment (forced air) has the highest average energy expense of 3498.850.

- Which combination has the most variability with regards to energy expense?

The combination of "oil steam" has the most variability with regards to energy expense, with a standard deviation of 2887.383.

- Which type of heating equipment doesn't take all possible fuel types?

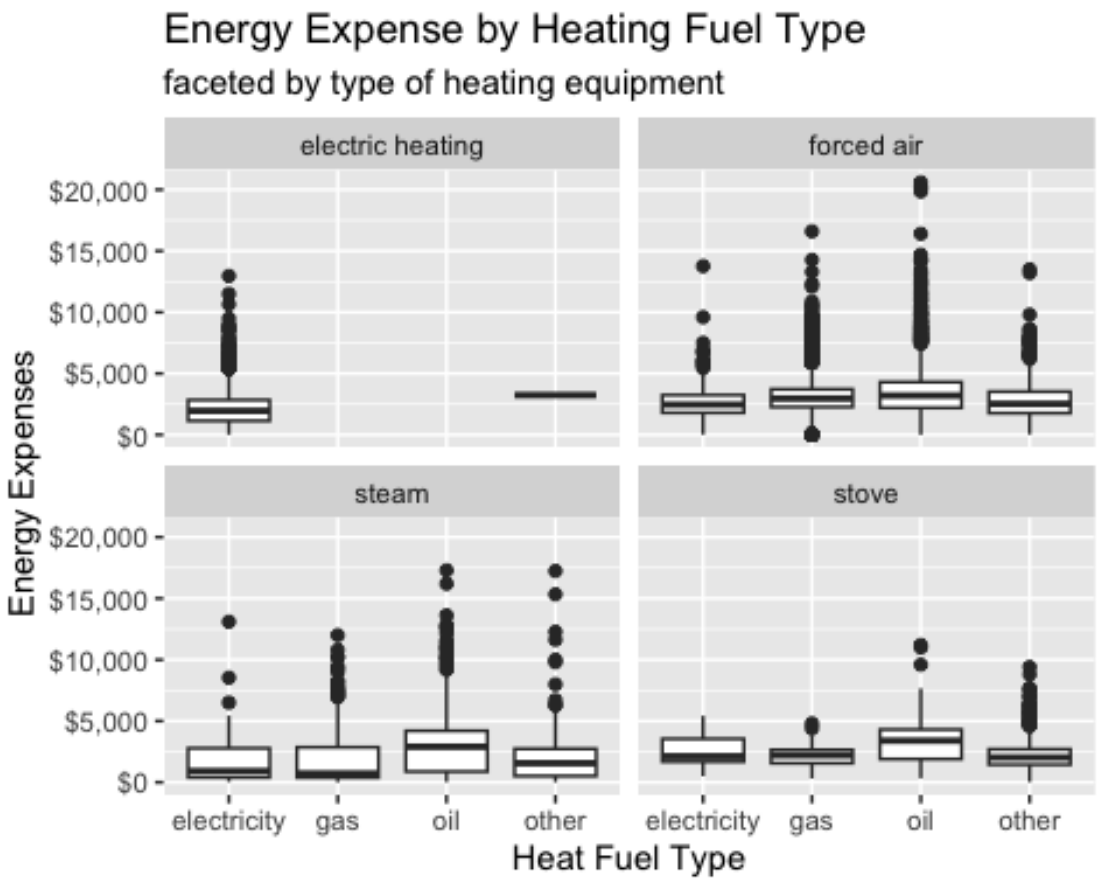
"electric heating" is only present for the "electricity" fuel type, which means it doesn't take all possible fuel types.

Task 4

Create a bar chart of energy expenses by heating fuel type and faceted by the type of heating equipment in a 2 x 2 grid. Your axis should be appropriately labeled with a dollar sign and commas. The `scales` package may be helpful here.

```
ggplot(data=cas,
       mapping=aes(x=heat_fuel, y=energy_expense)) +
  geom_boxplot() +
  facet_wrap(.~heat_equip, nrow = 2)+
  scale_y_continuous(labels = scales :: dollar_format())+
  labs(title = "Energy Expense by Heating Fuel Type",
```

```
       subtitle = "faceted by type of heating equipment",
       x="Heat Fuel Type",
       y="Energy Expenses")
```



Task 5

Create a new variable describing the proportion of household income spent on energy related expenses, and then find the respondent that spent the highest proportion of their household income on energy and the respondent that spent the lowest proportion of their household income on energy. End your pipeline with the tibble being passed into `glimpse()`. Describe these respondents based on the data they have provided.

```
cas %>%
  mutate(prop_energy_income = energy_expense / household_income) %>%
  arrange(prop_energy_income) %>%
  slice(c(1, n())) %>%
  glimpse()
```

```
## Rows: 2
## Columns: 25
## $ year                <fct> 2009, 2009
## $ province            <fct> Ontario, Saskatchewan
## $ dwelling_type       <fct> "Apartment", "Single detached"
```

```
## $ year_built          <fct> 1971-1980, 1971-1980
## $ rooms               <dbl> 6, 7
## $ beds                <dbl> 2, 3
## $ baths               <dbl> 1, 1
## $ heat_equip          <chr> "forced air", "forced air"
## $ heat_age            <fct> 5, 2
## $ heat_fuel           <chr> "gas", "gas"
## $ water_fuel          <fct> 4, 2
## $ cook_fuel           <fct> 2, 2
## $ income              <dbl> 67000, 100
## $ marital_status      <fct> 3, 3
## $ age                 <fct> 14, 08
## $ sex                 <fct> 2, 2
## $ education           <fct> 1, 6
## $ household_income    <dbl> 67000, 100
## $ energy_expense      <dbl> 1, 3780
## $ water_expense       <dbl> 1, 540
## $ electricity_expense <dbl> 0, 1716
## $ nat_gas_expense     <dbl> 0, 1524
## $ other_fuel_expense  <dbl> 0, 0
## $ consumption         <dbl> 16423, 19908
## $ prop_energy_income  <dbl> 1.492537e-05, 3.780000e+01
```

Describe these respondents based on the data they have provided:

The first respondent (row 1) lives in an apartment in Ontario built between 1971-1980 with 6 rooms and 2 beds. They use gas for heating, have a household income of 67000 CAD, and spent 1 CAD on energy-related expenses, which represents a very small proportion of their income (1.492537e-05). They are 14 years old, female, and have completed only 1 year of education.

The second respondent (row 2) lives in a single detached house in Saskatchewan built between 1971-1980 with 7 rooms and 3 beds. They also use gas for heating, but have a much lower household income of only 100 CAD, and spent 3780 CAD on energy-related expenses, which represents a very high proportion of their income (3.78e+01). They are 8 years old, female, and have completed 6 years of education.

Overall, these differences suggest that households in Ontario have a higher income and lower energy expenses compared to households in Saskatchewan.

Task 6

For each year, find the province with the cheapest median energy expense per room. Your answer should consist of a single `dplyr` pipeline that results in two rows and three columns – `year`, `province`, and `median_energy_expense_per_room`.

```
library(dplyr)
cas %>%
  group_by(year, province) %>%
  summarise(median_energy_expense_per_room = median(energy_expense/rooms))
```

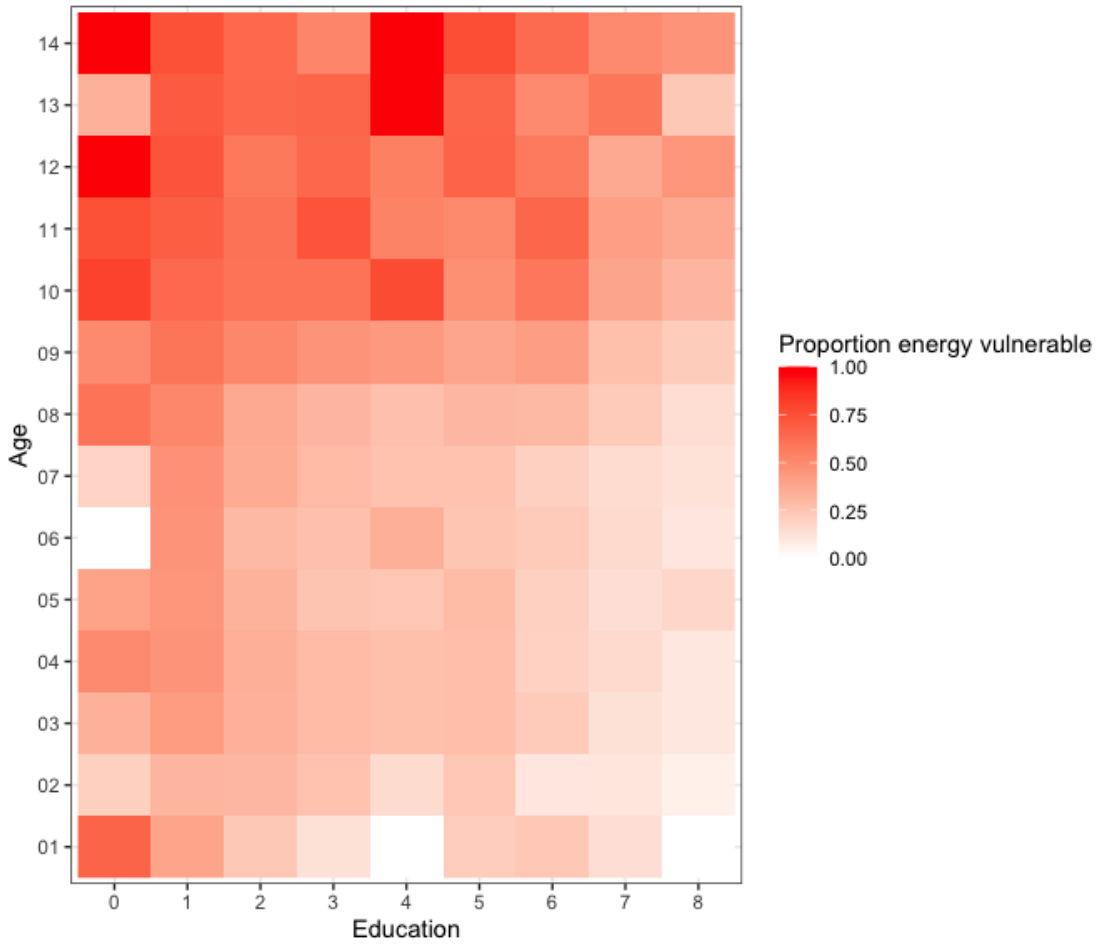
```
%>%
  arrange(year, median_energy_expense_per_room) %>%
  slice(1)
```

```
## # A tibble: 2 × 3
## # Groups:   year [2]
##   year  province median_energy_expense_per_room
##   <fct> <fct>                             <dbl>
## 1 2007  Quebec                             275
## 2 2009  Quebec                             269.
```

Task 7

A respondent is considered to be “energy vulnerable” if they spend more than 5% of their household income on energy expenses. Recreate the plot below, which shows the proportion of respondents who are energy vulnerable for each combination of age and education. In 2 - 3 sentences, describe what you observe in the plot. Hints: o You will need to use the variable created in task 5. o use `geom_raster()` o colors are from `"white"` to `"red"` in `scale_fill_gradient()` o theme is `bw` o figure width is 7, figure height is 6

```
cas %>% mutate(energy_vulnerable =
                 ifelse(energy_expense > household_income*0.05, "Yes", "No"))
%>%
  group_by(education, age) %>%
  summarise(prop_vulnerable =
              mean(energy_vulnerable == "Yes")) %>%
  ungroup() %>%
ggplot(aes(x = education,
           y = age,
           fill = prop_vulnerable)) +
  geom_raster() +
  scale_fill_gradient( low = "white",
                       high = "red") +
  theme_bw() +
  labs(x = "Education",
       y = "Age",
       fill = "Proportion energy vulnerable")
```


Age
14
13
12
11
10
09
08
07
06
05
04
03
02
01
0
1
2
3
4
5
6
7
8
Education
Proportion energy vulnerable
1.00
0.75
0.50
0.25
0.00