Lab 02

CS3072-1, Spring 2023, Effat University

Leena Alam

Packages

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

Data

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

Tasks

Task 1

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

```
cas$marital_status <- factor(cas$marital_status, levels = c(1,2,3,4),
labels = c("Single", "Married", "Divorced", "Widowed"))
```

Task 2

```
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"))
```

```
cas <- cas %>% mutate(heat_fuel = case_when(
  heat_fuel == 1 ~ "oil",
  heat_fuel == 2 ~ "gas",
  heat_fuel == 3 ~ "electricity",
  heat_fuel == 4 ~ "other"))
```

Task 3

```
cas_summary <- cas %>% group_by(heat_fuel, heat_equip) %>%
summarize(mean_energy_expense = mean(energy_expense), median_energy_expense =
median(energy_expense), sd_energy_expense = sd(energy_expense))
```

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

```
print(cas_summary)
```

```
## # A tibble: 14 x 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 1: median_energy_expense,
## #   2: sd_energy_expense
```

- Provide the answer to the theoretical questions here

According to the summarized table, the oil (heat_fuel) and stove (heat_equip) combination has the highest energy expense.

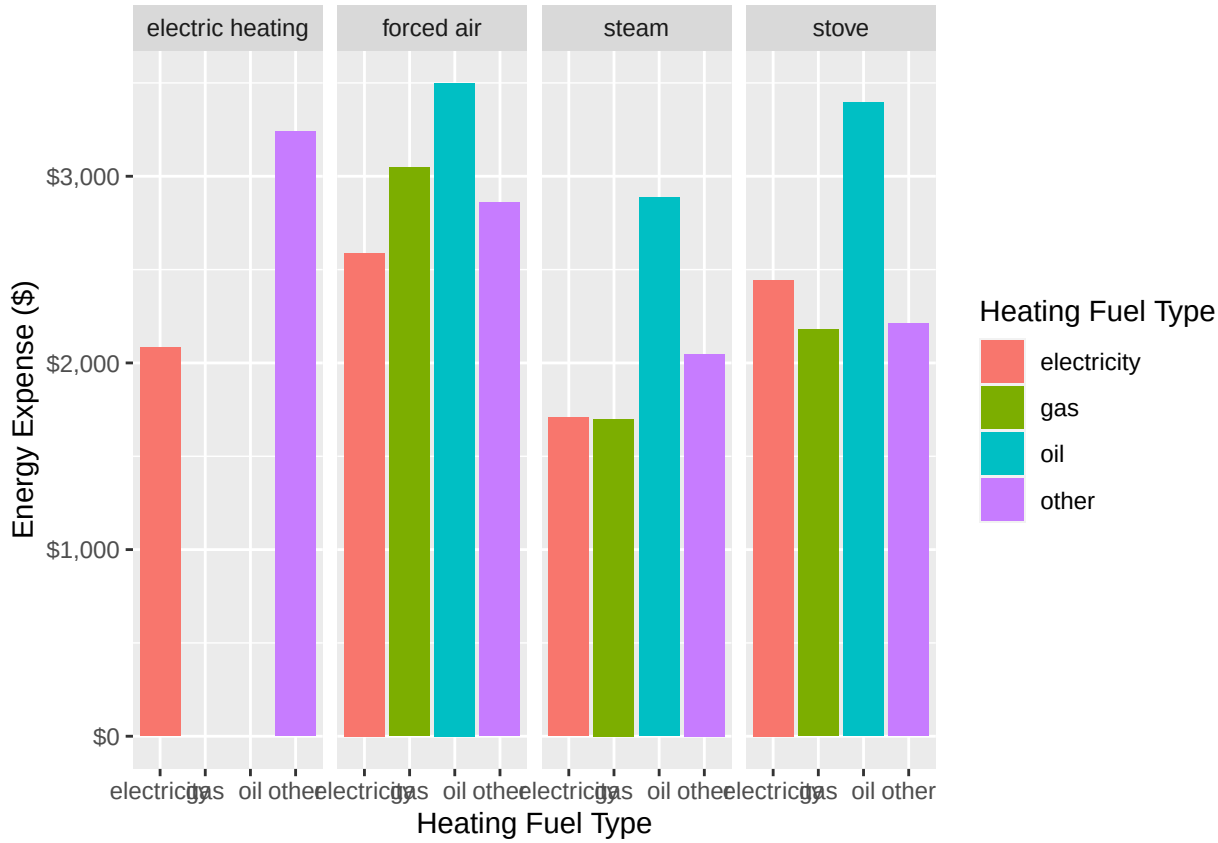
The highest variability was shown for the other (heat_fuel) and steam (heat_equip) according to the standard variation column.

Electric heating is not supported by every type of heating fuel.

Task 4

```
library(scales)

ggplot(cas, aes(x = heat_fuel, y = energy_expense, fill = heat_fuel)) +
geom_bar(stat = "summary", fun = "mean", position = "dodge") +
facet_grid(.~heat_equip) + scale_y_continuous(labels = dollar_format()) +
labs(x = "Heating Fuel Type", y = "Energy Expense ($)",
fill = "Heating Fuel Type")
```



Task 5

```
cas_summary2 <- cas %>% mutate(prop_energy_income = energy_expense /
household_income) %>% arrange(prop_energy_income) %>% slice(c(1L, 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> Divorced, Divorced
## $ 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
```

The respondent with the lowest proportion of their household income on energy is living in Ontario in an apartment and is divorced. The respondent with the highest proportion is living in Saskatchewan in a single detached house and is also divorced, however, this person's household income is very low and energy expense is very high which explains why they have a high proportion.

Task 6

```
cas_summary3 <- cas %>% group_by(year, province) %>%
summarize(median_energy_expense_per_room = median(energy_expense / rooms)) %>%
group_by(year) %>% filter(median_energy_expense_per_room ==
min(median_energy_expense_per_room)) %>% select(year, province,
median_energy_expense_per_room)
```

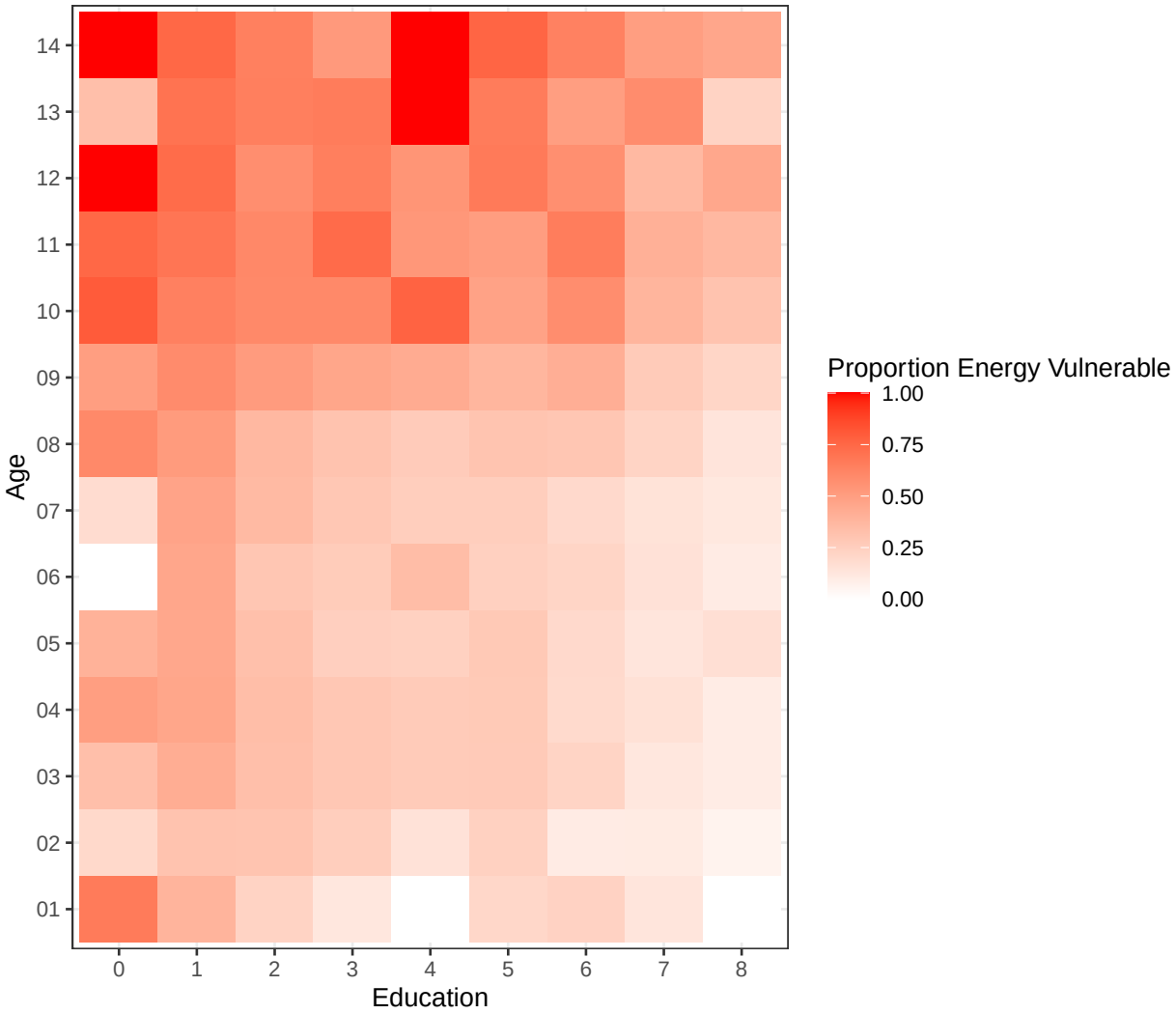
```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

Task 7

```
cas_summary4 <- cas %>% mutate(prop_energy_income = energy_expense /
household_income) %>% arrange(prop_energy_income) %>% group_by(age, education) %>%
mutate(is_energy_vulnerable = prop_energy_income > 0.05) %>%
summarize(prop_energy_vulnerable = mean(is_energy_vulnerable))
```

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

```
ggplot(cas_summary4, aes(x = education, y = age, fill = prop_energy_vulnerable)) +
geom_raster() + scale_fill_gradient(low = "white", high = "red") + labs(x =
"Education", y = "Age", fill = "Proportion Energy Vulnerable") + theme_bw()
```



As we can observe from the plot above, the most vulnerable age group is 85 years old or more. The plot shows that the least vulnerable people are those who are 25 years old or younger with a University degree that higher than a Bachelors degree. Also those individuals who are 25 years old or younger with no proper education are also somewhat vulnerable.