Lab 02

CS3172-1, Spring 2023, Effat University

Joud AlFarra

Packages

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

Data

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

Tasks

Task 1: Data cleaning

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 <- mutate(cas, marital_status = factor(marital_status))
```

Task 2: Variable Recoding

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 <- mutate(cas, heat_equip = case_when(heat_equip == 1 ~ 'steam',
                                          heat_equip == 2 ~ 'forced air',
                                          heat_equip == 3 ~ 'stove',
                                          heat_equip == 4 ~ 'electric
heating'))
```

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

Task 3: Group_by and Summarize

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

```
cas1 <- cas %>% group_by(heat_equip,heat_fuel) %>%
  summarise(mean_ener_exp = mean(energy_expense),
            median_ener_exp = median(energy_expense),
            sd_ener_exp = sd(energy_expense), .groups = "drop")
```

- Provide the answer to the theoretical questions here:

o What combination of fuel type and equipment has the highest average energy expense? - “forced air” and “oil”

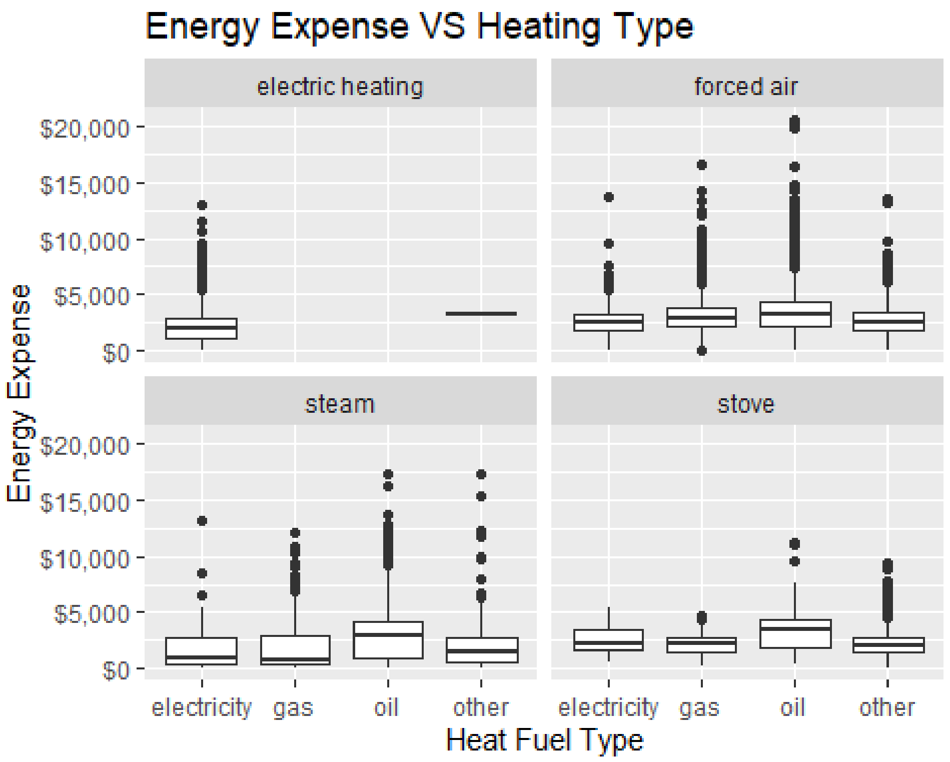
o Which combination has the most variability with regards to energy expense? - “steam” and “other”

o Which type of heating equipment doesn’t take all possible fuel types? - “electric heating”

Task 4: Data Visualization

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



Task 5: Mutate()

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.

```
cas2 <- cas %>% mutate(energy_prop = energy_expense/household_income) %>%
  arrange(desc(energy_prop)) %>% slice(1,n()) %>% glimpse()

## Rows: 2
## Columns: 25
## $ year                    <fct> 2009, 2009
## $ province                <fct> Saskatchewan, Ontario
## $ dwelling_type           <fct> "Single detached", "Apartment"
## $ year_built              <fct> 1971-1980, 1971-1980
## $ rooms                   <dbl> 7, 6
## $ beds                    <dbl> 3, 2
## $ baths                   <dbl> 1, 1
## $ heat_equip              <chr> "forced air", "forced air"
## $ heat_age                <fct> 2, 5
## $ heat_fuel               <chr> "gas", "gas"
## $ water_fuel              <fct> 2, 4
## $ cook_fuel               <fct> 2, 2
## $ income                  <dbl> 100, 67000
## $ marital_status      <fct> 3, 3
## $ age                 <fct> 08, 14
## $ sex                 <fct> 2, 2
## $ education           <fct> 6, 1
## $ household_income    <dbl> 100, 67000
## $ energy_expense      <dbl> 3780, 1
## $ water_expense       <dbl> 540, 1
## $ electricity_expense <dbl> 1716, 0
## $ nat_gas_expense     <dbl> 1524, 0
## $ other_fuel_expense  <dbl> 0, 0
## $ consumption         <dbl> 19908, 16423
## $ energy_prop         <dbl> 3.780000e+01, 1.492537e-05
```

Task 6: Pipeline

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`.

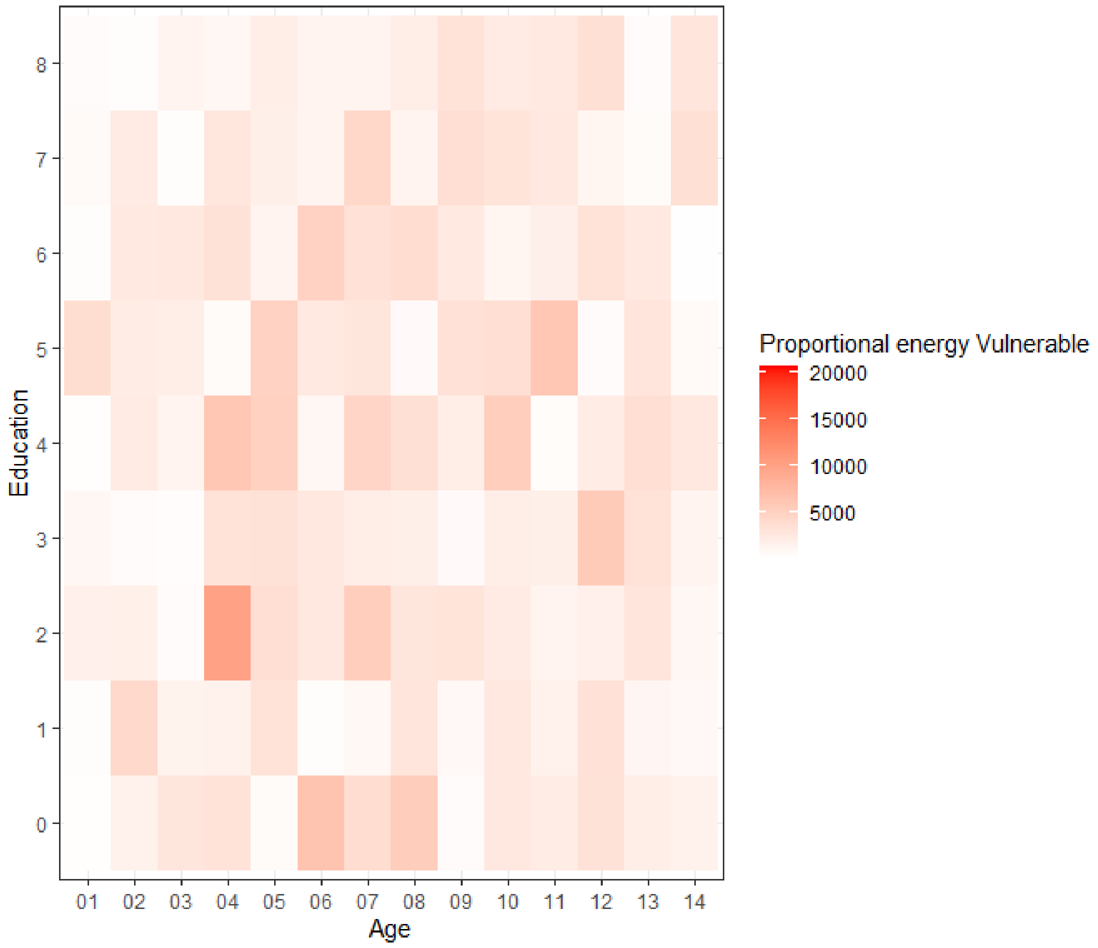
```
cas3 <- cas %>% group_by(year, province) %>%
  summarize(median_energy_expense = median(energy_expense)) %>%
  filter(median_energy_expense == min(median_energy_expense)) %>%
  select(year, province, median_energy_expense)
```

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, which shows the proportion of respondents who are energy vulnerable for each combination of age and education.

```
cas <- cas %>% mutate(energy_vulnerable = (energy_expense /
household_income))

ggplot(cas, aes(x = age, y = education, fill = energy_expense)) +
  geom_raster() +
  scale_fill_gradient(low = "white", high = "red") +
  labs(x = "Age", y = "Education", fill = "Proportional energy Vulnerable") +
  theme_bw()
```



In 2 - 3 sentences, describe what you observe in the plot. As shown in the plot, as age decreases and education increases, energy vulnerability decreases.