Lab 02

CS3172-1, Spring 2023, Effat University

Wejdan Alshateri

# Packages

library(tidyverse)   
library(scales)  
library(ggplot2)

# 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 <- mutate(cas, marital\_status = factor(marital\_status))

1. Married - a person who is legally married and living with their spouse.
2. Widowed - a person whose spouse has died and who has not remarried.
3. Divorced - a person who has been legally divorced and has not remarried.
4. Separated - a person who is legally separated from their spouse, but not divorced.
5. Never married/single - a person who has never been married, or not currently married.
6. Unknown - a person whose marital status is unknown or not reported.

## Task 2

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

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\_equip,heat\_fuel) %>% summarise(  
 mean\_ener\_exp = mean(energy\_expense),  
 median\_ener\_exp = median(energy\_expense),  
 sd\_ener\_exp = sd(energy\_expense))

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

## # A tibble: 14 × 5  
## # Groups: heat\_equip [4]  
## heat\_equip heat\_fuel mean\_ener\_exp median\_ener\_exp sd\_ener\_exp  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 electric heating electricity 2084. 1956 1270.  
## 2 electric heating other 3240 3240 NA   
## 3 forced air electricity 2590. 2462. 1293.  
## 4 forced air gas 3047. 2960 1395.  
## 5 forced air oil 3499. 3200 2156.  
## 6 forced air other 2861. 2526 1655.  
## 7 steam electricity 1708. 915 1692.  
## 8 steam gas 1698. 720 1820.  
## 9 steam oil 2887. 2900 2142.  
## 10 steam other 2047. 1555 2279.  
## 11 stove electricity 2443. 2120 1229.  
## 12 stove gas 2178. 2202 1024.  
## 13 stove oil 3396. 3395 2074.  
## 14 stove other 2210. 2025 1140.

o What combination of fuel type and equipment has the highest average energy expense? the combination of heating equipment ( forced air) and heat fuel (oil) has the highest average energy expense equal to 3498.850.

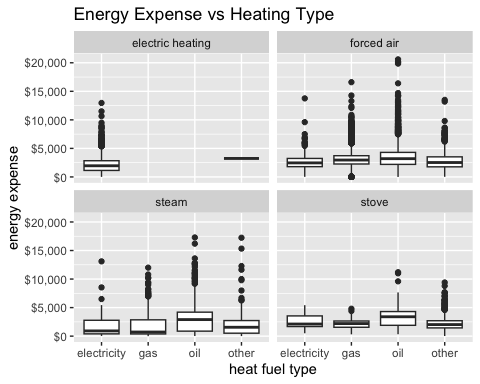
o Which combination has the most variability with regards to energy expense? the combination of heating equipment ( steam ) and heat fuel (other) will have the highest standard deviation and will have the most variability in energy expense.

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

## 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(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

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

cas1 <- 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

new column energy\_prop is created, Energy\_prop had a value of almost 3.78 for the respondent who spent the largest percentage of their household income on energy, This indicates the respondent is probably spending an excessive portion of their income to energy, we can see that this respondent is divorced, living in detached house, He has a relatively low household income, and he reported using electric heating and having a high level of energy consumption. On the other hand, the respondent who spent the lowest proportion of their household income on energy had energy\_prop value of 1.49, which is much lower. This suggests that this respondent is using energy-efficient appliances or living in a small, energy-efficient apartment. he has a high household income of $67000 and reported a low level of consumption.

## 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

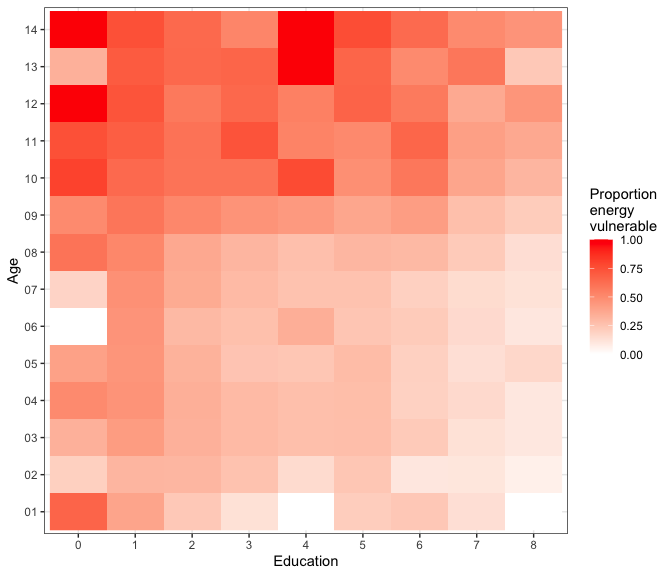
cas %>%   
 group\_by(year, province) %>%   
 summarize(median\_energy\_expense\_per\_room = median(energy\_expense/rooms)) %>%  
 group\_by(year) %>%   
 slice(which.min(median\_energy\_expense\_per\_room))

## # 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.

cas %>% mutate(energy\_prop = energy\_expense / household\_income,   
 vulnerable = if\_else(energy\_prop > 0.05, "vulnerable", "not")) %>%  
 group\_by(education, age) %>%  
 summarize(prop\_vulnerable = mean(vulnerable == "vulnerable")) %>%  
 ungroup() %>%  
   
ggplot(aes(x = education, y = age, fill = prop\_vulnerable)) +   
 geom\_raster() +  
 scale\_fill\_gradient(low = "white", high = "red") +  
 labs(x = "Education", y = "Age",  
 fill = "Proportion\nenergy\nvulnerable") +  
 theme\_bw()

 The graph shows that the proportion of energy-vulnerable susceptible people is higher for older people, older respondents are more likely to be energy vulnerable. and we observe a higher proportions of vulnerable households among those with lower levels of education.