Household Living Situations

Ronnald Le

2023-03-14

Introduction: You are provided the .csv file OR_acs_house_occ.csv which contains household level responses to the American Community Survey for households in Oregon. Technically this is a Public Use Microdata Sample (PUMS) from the 2015 1-year survey. The data were obtained from http://www2.census.gov/programs-surveys/acs/data/pums/2015/1-Year/. You are provided a subset of variables and only households that have at least one person, pay for their electricity, and are not group accommodation. You may assume this is a random sample of all such households in Oregon.

1. Explanatory Problem

Do people living in apartments pay less on electricity than those living in houses? How much?

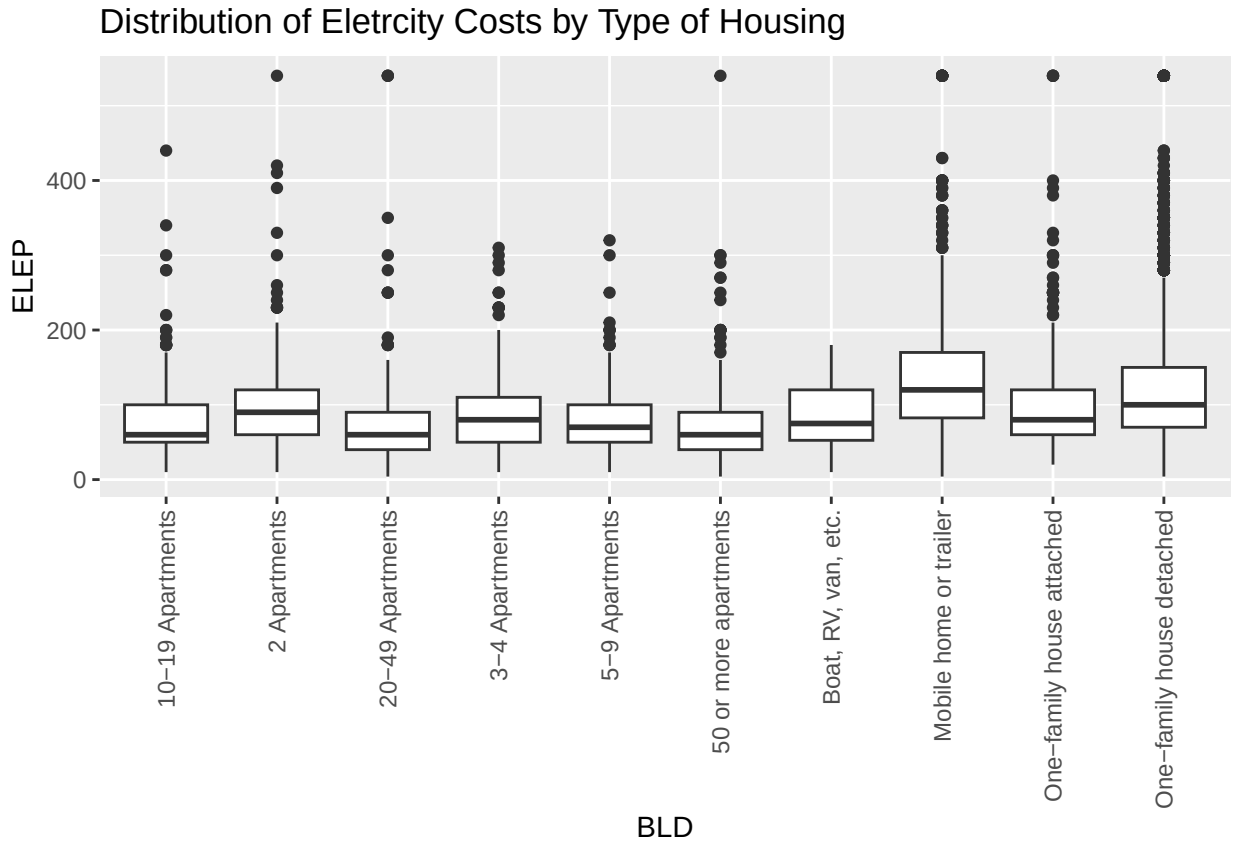
Comparing the median cost of electricity separately from houses and apartments, and factor in number of bedrooms and occupants after in those households.

```
housing |> group_by(BLD) |>
  #ELEP is monthly cost of Electricity
  summarize(median_electricity = median(ELEP)) |>
  #BLD is Units in Structures
  arrange(BLD)
```

```
## # A tibble: 10 x 2
##    BLD                       median_electricity
##    <fct>                                  <dbl>
##  1 10-19 Apartments                          60
##  2 2 Apartments                              90
##  3 20-49 Apartments                          60
##  4 3-4 Apartments                            80
##  5 5-9 Apartments                            70
##  6 50 or more apartments                     60
##  7 Boat, RV, van, etc.                       75
##  8 Mobile home or trailer                   120
##  9 One-family house attached                 80
## 10 One-family house detached                100
```

Create a boxplot to help visualize distrbution (electrictiy cost for households with factoring in outliers).

```
#To see the entire plot, make sure to view it in a separate window full screen.
ggplot(housing, aes(x = BLD, y = ELEP)) +
  geom_boxplot() +
  ggtitle("Distribution of Eletrcity Costs by Type of Housing") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



Performing a two-sample t-test. It is to help compare the mean electricity cost of the two households, which consists of number of bedrooms and number of associated occupants.For a fair comparison, I would proceed with One-family house detached vs One-family house attached. Another will be 10-19 apartments vs. 50 or more apartments.

```
# Create a new variable that consist of detached/attached
housing$detached <- ifelse(housing$BLD == 'One-family house detached', 'detached', 'attached')

# Perform t-test between
t.test(ELEP ~ detached, data = housing)
```

```
##
##  Welch Two Sample t-test
##
## data:  ELEP by detached
## t = -17.684, df = 9780.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group attached and group detached is not eq
## 95 percent confidence interval:
##  -24.61824 -19.70522
## sample estimates:
## mean in group attached mean in group detached
##               100.7344               122.8961
```

```
# Create a new variable with apartments of interest
#Other is for all the other categories of housing units not in the apartments of interest
housing$apartment_type <- ifelse(housing$BLD == '10-19 Apartments', '10-19 Apartments',
```

```
                                ifelse(housing$BLD == '50 or more apartments', '50 or more apartments'

# Filter targeted rows
housing_apartments <- subset(housing, apartment_type %in% c('10-19 Apartments', '50 or more apartments')

# Perform t-test between apartment_type and ELEP
t.test(ELEP ~ apartment_type, data = housing_apartments)
```

```
##
##  Welch Two Sample t-test
##
## data:  ELEP by apartment_type
## t = 1.4143, df = 809.12, p-value = 0.1577
## alternative hypothesis: true difference in means between group 10-19 Apartments and group 50 or more
## 95 percent confidence interval:
##  -1.90155 11.70513
## sample estimates:
##      mean in group 10-19 Apartments mean in group 50 or more apartments
##                            77.08134                            72.17955
```

2. Prediction Problem

Create a model that could be used to predict electricity costs for a household in Oregon.

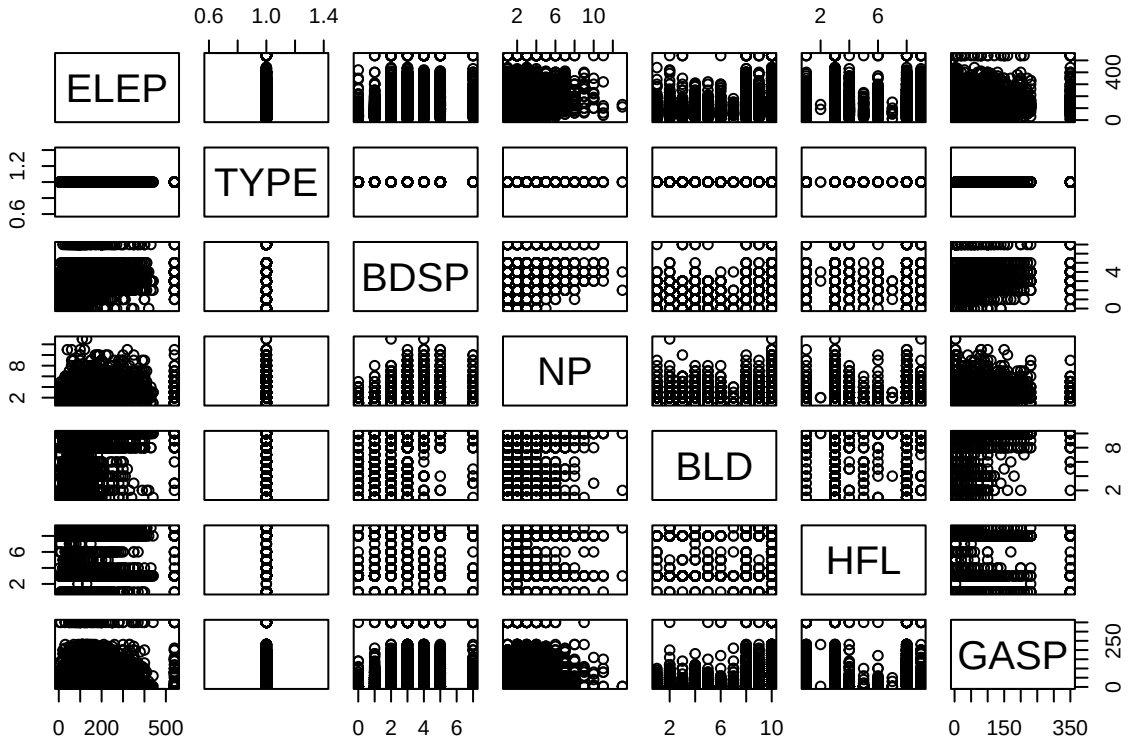
Here we are comparing the relationship between the electricity cost variable and predictor variables, such as housing type, number of bedrooms, occupants, units in structures, house heating fuel, and monthly gas cost.

```
#First is to clean up the data that includes NA
housing_cleandata <- select(housing, ELEP, TYPE, BDSP, NP, BLD, HFL, GASP)
housing_cleandata <- na.omit(housing_cleandata)
```

Creating a scatterplot here will help visualize the relationship between the electricity cost variable and predictor variables listed above.

```
# Select the variables for the scatterplot matrix
vars <- c("ELEP", "TYPE", "BDSP", "NP", "BLD", "HFL", "GASP")

# Create the scatterplot matrix
pairs(housing[, vars])
```



Fitting a linear regression model to help estimate differences towards electricity costs and the predictor values.

```
model <- lm(ELEP ~ BDSP + NP + BLD + TYPE + HFL + GASP, data = housing_cleandata)

summary(model)
```

```
##
## Call:
## lm(formula = ELEP ~ BDSP + NP + BLD + TYPE + HFL + GASP, data = housing_cleandata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -222.53  -39.61  -12.97   23.58  478.76
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.97782    5.20648   0.764  0.44487
## BDSP                           11.50383    0.68298  16.844  < 2e-16 ***
## NP                             11.76164    0.42917  27.406  < 2e-16 ***
## BLD2 Apartments                24.24521    4.82460   5.025 5.08e-07 ***
## BLD20-49 Apartments             4.02911    4.90225   0.822  0.41115
## BLD3-4 Apartments               6.24205    4.35189   1.434  0.15150
## BLD5-9 Apartments              -1.91209    4.40711  -0.434  0.66439
## BLD50 or more apartments       -1.92661    4.67185  -0.412  0.68006
## BLDBoat, RV, van, etc.         24.44747   13.55512   1.804  0.07132 .
```

```
## BLDMobile home or trailer     48.07640    3.80234  12.644  < 2e-16 ***
## BLDOne-family house attached  27.40904    4.27601   6.410 1.50e-10 ***
## BLDOne-family house detached  42.73457    3.51886  12.144  < 2e-16 ***
## TYPE                              NA         NA      NA       NA
## HFLCoal or coke              -19.72202   47.35411  -0.416  0.67706
## HFLElectricity                35.71690    3.95378   9.034  < 2e-16 ***
## HFLFuel oil, kerosene, etc.    5.05029    5.08360   0.993  0.32051
## HFLNo fuel used               24.68344   10.59285   2.330  0.01981 *
## HFLOther fuel                 33.25834    7.48626   4.443 8.95e-06 ***
## HFLSolar energy              -22.76944   23.92148  -0.952  0.34119
## HFLUtility gas               -21.62997    3.81186  -5.674 1.42e-08 ***
## HFLWood                       14.81605    4.26195   3.476  0.00051 ***
## GASP                           0.16500    0.01298  12.710  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.76 on 15145 degrees of freedom
## Multiple R-squared:  0.212,  Adjusted R-squared:  0.2109
## F-statistic: 203.7 on 20 and 15145 DF,  p-value: < 2.2e-16
```

Using the model from above to help predict electricity cost for future households based on the targeted predictors.

```
new_data <- data.frame(BDSP = 3, NP = 2, BLD = "One-family house detached", TYPE = NA, HFL = "Electrici
new_data$TYPE <- as.numeric(as.factor(new_data$TYPE))
predicted_cost <- predict(model, newdata = new_data)
cat("The predicted electricity cost for the household is $", round(predicted_cost, 2))
```

```
## The predicted electricity cost for the household is $ 148.71
```

3. Compare and Contrast

For question 1, my initial goal was to compare the median electricity cost between different types of housing. I also wanted to factor in other predictor variables to help me visualize models that can be used to help predict future electricity cost as I did in question 2. I started by filtering the data to only include households that paid for electricity and were not group accommodation, and then grouped the data by housing type. After all of that, I found the median electricity cost for each of the BLD while factoring in the number of bedrooms and occupants in each household. I created a boxplot to visualize the distribution of electricity cost for each housing type. Lastly, I performed a two-sample t-test to compare the mean electricity cost between different housing types.

For question 2, the goal was to create a model that predicts electricity costs for households in Oregon based on different predictors such as housing type, number of bedrooms, occupants, house heating fuel, and monthly gas cost. I started off by cleaning up the data to exclude rows with missing values. I created a scatterplot matrix to visualize the relationship between the target variable (electricity cost) and the predictor variables. After all of that, I fitted a linear regression model to estimate the differences in electricity costs due to the predictor values. Lastly, I used the model to predict electricity costs for new households with the targeted predictors.

Different approaches are required because both of the questions have different goals,because question 1 is an explanatory problem, and the goal is to compare the median electricity cost between different types of housing structures. Question 2 is a prediction problem, and the goal is to create a model that predicts electricity costs for households in Oregon. The approach for Question 1 involves comparing and contrasting the median and mean electricity costs between different housing types while factoring in the number of

occupants and bedrooms, while the approach for Question 2 involves creating a linear regression model that estimates the relationship between the electricity cost variable and predictor variables.

The first challenge I faced was trying to get my packages to work correctly. I had to remove some packages, which you won't be able to see now due to conflicting resources. Another challenge of grouping the data by housing type and calculating the median electricity cost for each group while factoring in the number of occupants and bedrooms.

For Question 2, the primary challenge was fitting a linear regression model that accurately estimates the relationship between the electricity cost variable and the predictor variables. I wanted to create a working predictor model that predicts cost for electricity in the future. It is my 3rd or 4th chunk of code with multiple revisions and errors. When comparing across the two task, the challenge in question 1 is obtaining a fair comparison of the median electricity costs between different housing types, while for question 2, it is to accurately estimating the relationship between the electricity cost variable and the predictor variables.