BA_Assignment2

Thanasit C.

2024-03-11

1. Summary

Dataset provided by IPUMS USA for the year of 2021 - 2022, are observed from 51 States across the US contain more than 6.6 million observations and 21 variables. Dataset is already treated, there are no missing values nor N/A value. However, according to the Code book, in some particular variables there are some unusable values that need to be filter out before analyzing. See the full details of the findings in Section 5.

For 2021, the highest cost of electricity is USD 9,990 share between 19 Stats. The highest cost of gas is USD 9,990 share across 4 states. lastly, the cost of water, the highest value is USD 6,200 from The State of California. For 2022, all of the 51 states shares the same highest cost of electricity at USD 9,900. The cost of gas share the highest value of USD 9,900 across 49 States EXCEPT Florida and Hawaii. Lastly, Hawaii has the highest cost of water at USD 7,100 in 2022. The detail of state shows in Section 5.

Next, I explored the imbalance of gender (SEX) across the country. An imbalance of gender exists in every States, The highest imbalance state is The District of Columbia, which has female almost 7% more than male. In the other hand, the closet proportion between male and female, 0.04%, is the State of Utah. I also performed hypothesis testing to ensure the imbalance, the result rejects H0 which mean there is a different between the proportion of male and female. The mean different is -1.53%, I can say that the proportion of female is 1.53% higher than male.

For 2021 and 2022, the state that have the highest total cost of electric, gas, and water combine is The State of California. Even through, I calculate separately for each category, The State of California is still has the highest total cost. I dig a bit deeper by find ding the number of observations of each state. I found that The State of California has the highest number of observations, around 20% - 50% more than the second place, Texas. This finding back up the reason why California have the highest total cost.

The State of Maine on average has the oldest residents with the age of 46.8 years old in 2021 and 47.2 years old in 2022. I also found that the average age of USA residents grows from 42.7 years old in 2021 to 42.9 years old in 2022.

Lastly, I found some insights related to The State of Ohio for 2022. The residence lives in Ohio has an average age at 43.2 years old, a bit older compares to the nation average, 42.9 years old. I combined 'SEX' and 'AGE' in my analysis, and found that an average age of male

is 2.5 years, almost 6%, lower than female, 41.9 compared to 44.4 years old. In Ohio, the proportion of female to male is 51.15% to 48.85%. The different is -2.3% which is higher than nation wide different, -1.5%. There are all 9 races live in Ohio. However, White people is the dominant race with almost 82%. Last but not least, there are 97 languages used in the US, however, in Ohio, there are only 57 languages reported. As expected, the most use languages at home is English with the proportion of 88.7%. Unsurprisingly, there are only 51 persons who speak 'Thai/Laos at their home' which is around 0.04%.

2. Library

```
library(dplyr)
library(ExcelFunctionsR)
library(ggplot2)
library(ggpubr)
library(forcats)
library(tidyr)
```

3. Import data

```
setwd("/Users/sieng/Documents/Study/MS.Business Analytics/SPRING 
2024/Business Anaytics/BA - Assignment/BA - Assignment2")
maindf <- read.csv("usa_00006.csv")
```

4. Data preparation

```
# 4.1. Data summary
str(maindf)

## 'data.frame':    6625977 obs. of  21 variables:
##  $ YEAR     : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 ...
##  $ SAMPLE   : int  202101 202101 202101 202101 202101 202101 202101 202101 
202101 202101 ...
##  $ SERIAL   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CBSERIAL : num  2.02e+12 2.02e+12 2.02e+12 2.02e+12 2.02e+12 ...
##  $ HHWT     : num  13 51 17 61 15 46 55 31 71 48 ...
##  $ CLUSTER  : num  2.02e+12 2.02e+12 2.02e+12 2.02e+12 2.02e+12 ...
##  $ STATEFIP : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ STRATA   : int  80001 80001 120001 170001 50001 160001 130201 210001 
120001 30201 ...
##  $ GQ       : int  3 3 3 3 3 4 3 3 3 4 ...
##  $ COSTELEC : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ COSTGAS  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ COSTWATR : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ COSTFUEL : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PERNUM   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PERWT    : num  13 51 17 61 15 46 55 31 71 48 ...
##  $ SEX      : int  1 2 1 1 1 2 1 1 1 1 ...
##  $ AGE      : int  85 67 74 16 83 19 36 35 45 20 ...
```

```
##  $ RACE     : int  1 2 1 1 1 1 2 2 1 2 ...
##  $ RACED    : int  100 200 100 100 100 100 200 200 100 200 ...
##  $ LANGUAGE : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ LANGUAGED: int  100 100 100 100 100 100 100 100 100 100 ...
```

```r
head(maindf)
```

```
##   YEAR SAMPLE SERIAL    CBSERIAL HHWT   CLUSTER STATEFIP STRATA GQ
COSTELEC
## 1 2021 202101      1 2.02101e+12   13 2.021e+12        1  80001  3
0
## 2 2021 202101      2 2.02101e+12   51 2.021e+12        1  80001  3
0
## 3 2021 202101      3 2.02101e+12   17 2.021e+12        1 120001  3
0
## 4 2021 202101      4 2.02101e+12   61 2.021e+12        1 170001  3
0
## 5 2021 202101      5 2.02101e+12   15 2.021e+12        1  50001  3
0
## 6 2021 202101      6 2.02101e+12   46 2.021e+12        1 160001  4
0
##   COSTGAS COSTWATR COSTFUEL PERNUM PERWT SEX AGE RACE RACED LANGUAGE
LANGUAGED
## 1       0        0        0      1    13   1  85    1   100        1
100
## 2       0        0        0      1    51   2  67    2   200        1
100
## 3       0        0        0      1    17   1  74    1   100        1
100
## 4       0        0        0      1    61   1  16    1   100        1
100
## 5       0        0        0      1    15   1  83    1   100        1
100
## 6       0        0        0      1    46   2  19    1   100        1
100
```

```r
tail(maindf)
```

```
##         YEAR SAMPLE  SERIAL     CBSERIAL HHWT      CLUSTER STATEFIP STRATA
GQ
## 6625972 2022 202201 1505106 2.022001e+12   72 2.022015e+12       56  30056
1
## 6625973 2022 202201 1505107 2.022001e+12  119 2.022015e+12       56  40056
1
## 6625974 2022 202201 1505107 2.022001e+12  119 2.022015e+12       56  40056
1
## 6625975 2022 202201 1505107 2.022001e+12  119 2.022015e+12       56  40056
1
## 6625976 2022 202201 1505108 2.022001e+12  126 2.022015e+12       56  20056
1
## 6625977 2022 202201 1505108 2.022001e+12  126 2.022015e+12       56  20056
```

```
1
##         COSTELEC COSTGAS COSTWATR COSTFUEL PERNUM PERWT SEX AGE RACE RACED
## 6625972      840     840      410     9993      1    72   1  55    1   100
## 6625973     2400     960      300      250      1   119   1  33    1   100
## 6625974     2400     960      300      250      2    89   2  27    1   100
## 6625975     2400     960      300      250      3   177   1   1    1   100
## 6625976     3000    1320       70     9993      1   126   1  66    1   100
## 6625977     3000    1320       70     9993      2   187   2  58    1   100
##         LANGUAGE LANGUAGED
## 6625972        1       100
## 6625973        1       100
## 6625974        1       100
## 6625975        0         0
## 6625976        1       100
## 6625977        1       100
```

```
summary(maindf)
```

```
##       YEAR          SAMPLE           SERIAL           CBSERIAL
##  Min.   :2021   Min.   :202101   Min.   :      1   Min.   :2.021e+12
##  1st Qu.:2021   1st Qu.:202101   1st Qu.: 359081   1st Qu.:2.021e+12
##  Median :2022   Median :202201   Median : 732416   Median :2.022e+12
##  Mean   :2022   Mean   :202152   Mean   : 734687   Mean   :2.022e+12
##  3rd Qu.:2022   3rd Qu.:202201   3rd Qu.:1107874   3rd Qu.:2.022e+12
##  Max.   :2022   Max.   :202201   Max.   :1505108   Max.   :2.022e+12
##       HHWT            CLUSTER            STATEFIP         STRATA
##  Min.   :   1.00   Min.   :2.021e+12   Min.   : 1.00   Min.   :  10001
##  1st Qu.:  48.00   1st Qu.:2.021e+12   1st Qu.:12.00   1st Qu.:  90131
##  Median :  73.00   Median :2.022e+12   Median :27.00   Median : 230026
##  Mean   :  98.29   Mean   :2.022e+12   Mean   :27.73   Mean   : 478438
##  3rd Qu.: 118.00   3rd Qu.:2.022e+12   3rd Qu.:42.00   3rd Qu.: 460037
##  Max.   :3118.00   Max.   :2.022e+12   Max.   :56.00   Max.   :8100351
##        GQ            COSTELEC       COSTGAS         COSTWATR       COSTFUEL
##  Min.   :1.000   Min.   :   0   Min.   :   0   Min.   :   0   Min.   :   0
##  1st Qu.:1.000   1st Qu.:1200   1st Qu.: 600   1st Qu.: 200   1st Qu.:9993
##  Median :1.000   Median :1800   Median :2160   Median : 840   Median :9993
##  Mean   :1.133   Mean   :2357   Mean   :4876   Mean   :3028   Mean   :8784
##  3rd Qu.:1.000   3rd Qu.:3000   3rd Qu.:9993   3rd Qu.:9993   3rd Qu.:9993
##  Max.   :5.000   Max.   :9997   Max.   :9997   Max.   :9997   Max.   :9997
##      PERNUM           PERWT             SEX             AGE
##  Min.   : 1.000   Min.   :   1.0   Min.   :1.000   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.:  49.0   1st Qu.:1.000   1st Qu.:22.00
##  Median : 2.000   Median :  75.0   Median :2.000   Median :43.00
##  Mean   : 2.061   Mean   : 100.4   Mean   :1.509   Mean   :42.69
##  3rd Qu.: 3.000   3rd Qu.: 121.0   3rd Qu.:2.000   3rd Qu.:63.00
##  Max.   :20.000   Max.   :3223.0   Max.   :2.000   Max.   :97.00
##       RACE           RACED          LANGUAGE        LANGUAGED
##  Min.   :1.000   Min.   :100.0   Min.   : 0.000   Min.   :   0.0
##  1st Qu.:1.000   1st Qu.:100.0   1st Qu.: 1.000   1st Qu.: 100.0
##  Median :1.000   Median :100.0   Median : 1.000   Median : 100.0
```

```
##   Mean   :2.529   Mean   :257.1   Mean   : 4.859   Mean   : 486.3
##   3rd Qu.:2.000   3rd Qu.:200.0   3rd Qu.: 1.000   3rd Qu.: 100.0
##   Max.   :9.000   Max.   :990.0   Max.   :96.000   Max.   :9601.0
```

```r
# 4.2. Convert Data Attributes
maindf$LANGUAGE <- factor(maindf$LANGUAGE)
maindf$RACE <- factor(maindf$RACE)
maindf$SEX <- factor(maindf$SEX)
maindf$STATEFIP <- factor(maindf$STATEFIP)
maindf$YEAR <- factor(maindf$YEAR)
str(maindf)
```

```
## 'data.frame':    6625977 obs. of  21 variables:
##  $ YEAR     : Factor w/ 2 levels "2021","2022": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SAMPLE   : int  202101 202101 202101 202101 202101 202101 202101 202101
202101 202101 ...
##  $ SERIAL   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CBSERIAL : num  2.02e+12 2.02e+12 2.02e+12 2.02e+12 2.02e+12 ...
##  $ HHWT     : num  13 51 17 61 15 46 55 31 71 48 ...
##  $ CLUSTER  : num  2.02e+12 2.02e+12 2.02e+12 2.02e+12 2.02e+12 ...
##  $ STATEFIP : Factor w/ 51 levels "1","2","4","5",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ STRATA   : int  80001 80001 120001 170001 50001 160001 130201 210001
120001 30201 ...
##  $ GQ       : int  3 3 3 3 3 4 3 3 3 4 ...
##  $ COSTELEC : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ COSTGAS  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ COSTWATR : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ COSTFUEL : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PERNUM   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PERWT    : num  13 51 17 61 15 46 55 31 71 48 ...
##  $ SEX      : Factor w/ 2 levels "1","2": 1 2 1 1 1 2 1 1 1 1 ...
##  $ AGE      : int  85 67 74 16 83 19 36 35 45 20 ...
##  $ RACE     : Factor w/ 9 levels "1","2","3","4",..: 1 2 1 1 1 1 2 2 1 2
...
##  $ RACED    : int  100 200 100 100 100 100 200 200 100 200 ...
##  $ LANGUAGE : Factor w/ 64 levels "0","1","2","3",..: 2 2 2 2 2 2 2 2 2 2
...
##  $ LANGUAGED: int  100 100 100 100 100 100 100 100 100 100 ...
```

5. Data Analysis and Question Answering.

5.1. Question_1; Are there any missing values?

Answer_Q1; There is no missing value. Actually, all N/A values are treated for example, code 0 in column ‘LANGUAGE’ means ‘N/A or blank’.

```r
# Check number of N/A in data set
sumna <- sum(is.na(maindf))
```

```
print(paste("Number of N/A dataset is ", sumna))

colsumna <- colSums(is.na(maindf))
print(paste("Below shows the number of N/A in each column"))
colsumna

## [1] "Number of N/A dataset is  0"
## [1] "Below shows the number of N/A in each column"
##      YEAR    SAMPLE    SERIAL  CBSERIAL      HHWT   CLUSTER  STATEFIP
STRATA
##         0         0         0         0         0         0         0
0
##        GQ  COSTELEC   COSTGAS  COSTWATR  COSTFUEL    PERNUM     PERWT
SEX
##         0         0         0         0         0         0         0
0
##       AGE      RACE     RACED  LANGUAGE LANGUAGED
##         0         0         0         0         0
```

5.2 Question_2; Identify the states that have the highest cost of electricity, gas, and water.

NOTED: According to the Code book, there are few rows that unusable. So I started with filtering it out.
Answer_Q2;
For 2021, there are 19 states share the highest cost of electricity is $9,990 which are California (6), Colorado (8), Connecticut (9), District of Columbia (11), Florida (12), Hawaii (15), Indiana (18), Massachusetts (25), Michigan (26), Missouri (29), New Jersey (34), New York (36), Oregon (41), Rhode Island (44), Tennessee (47), Texas (48), Vermont (50), Virginia (51), Washington (53).
For 2021, there are 4 states share the highest cost of gas is $9,990 which are California (6), Massachusetts (25), Missouri (29), Rhode Island (44).
For 2021, the highest cost of water is $6,200 which is The State of California (6).

For 2022, there are all 51 states that share the highest cost of electricity is $9,990.
For 2022, there are 49 states share the highest cost of gas is $9,990 EXCLUDE The State of Florida (12) and The State of Hawaii (15).
For 2022, the highest cost of water is $7,100 which is The State of Hawaii (15).

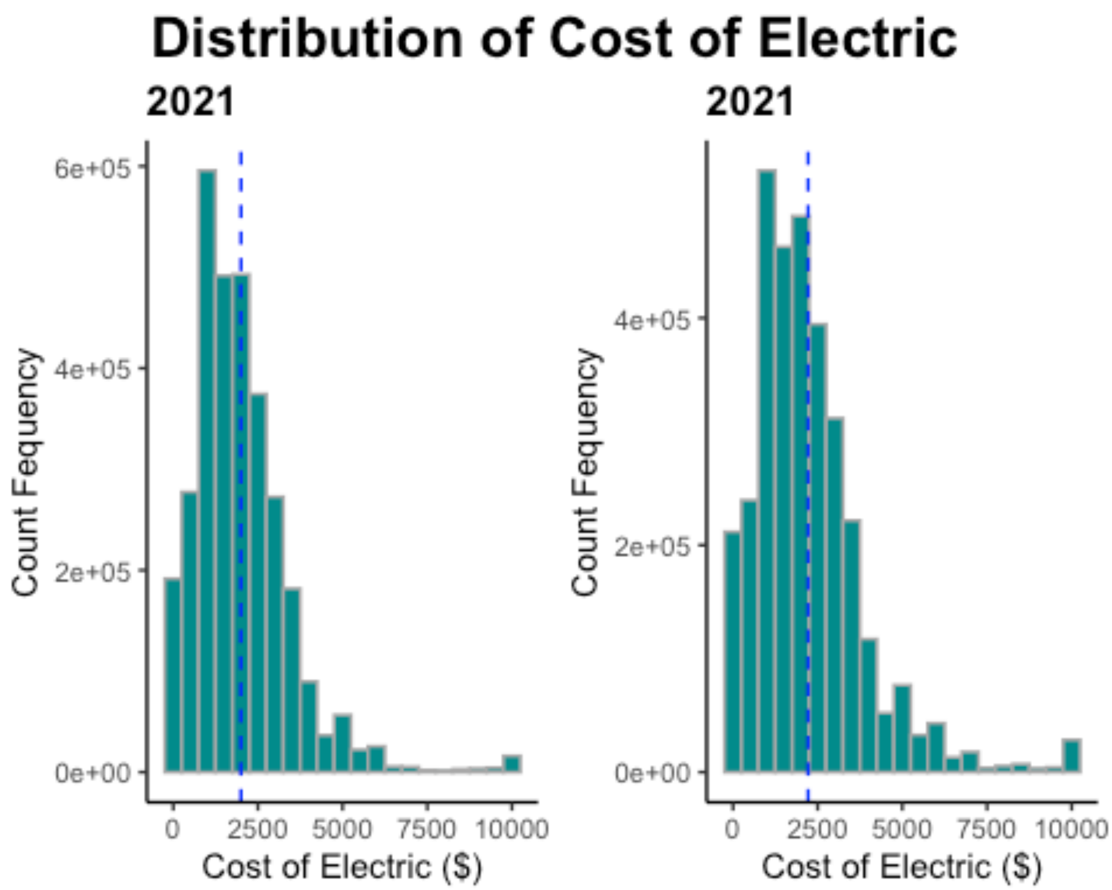
```
## Data manipulation
q2_data_electric <- maindf %>%
                      select(STATEFIP, YEAR, COSTELEC) %>%
                      filter(COSTELEC < 9993)

q2_data_gas <- maindf %>%
                 select(STATEFIP, YEAR, COSTGAS) %>%
                 filter(COSTGAS < 9992)
```

```r
q2_data_water <- maindf %>%
                   select(STATEFIP, YEAR, COSTWATR) %>%
                   filter(COSTWATR < 9993)

# Cost of Electric
q2_electric_2021 <- q2_data_electric %>%
                      filter(YEAR == 2021) %>%
                      group_by(STATEFIP) %>%
                      summarise(maxCOSTELEC21 = max(COSTELEC)) %>%
                      slice_max(maxCOSTELEC21, n = 1)

q2_electric_2022 <- q2_data_electric %>%
                      filter(YEAR == 2022) %>%
                      group_by(STATEFIP) %>%
                      summarise(maxCOSTELEC22 = max(COSTELEC)) %>%
                      slice_max(maxCOSTELEC22, n =1)

q2_electric <- merge(q2_electric_2021, q2_electric_2022, all = TRUE)
q2_electric

##    STATEFIP maxCOSTELEC21 maxCOSTELEC22
## 1         1            NA          9990
## 2         2            NA          9990
## 3         4            NA          9990
## 4         5            NA          9990
## 5         6          9990          9990
## 6         8          9990          9990
## 7         9          9990          9990
## 8        10            NA          9990
## 9        11          9990          9990
## 10       12          9990          9990
## 11       13            NA          9990
## 12       15          9990          9990
## 13       16            NA          9990
## 14       17            NA          9990
## 15       18          9990          9990
## 16       19            NA          9990
## 17       20            NA          9990
## 18       21            NA          9990
## 19       22            NA          9990
## 20       23            NA          9990
## 21       24            NA          9990
## 22       25          9990          9990
## 23       26          9990          9990
## 24       27            NA          9990
## 25       28            NA          9990
## 26       29          9990          9990
## 27       30            NA          9990
## 28       31            NA          9990
```

```
## 29       32         NA        9990
## 30       33         NA        9990
## 31       34       9990        9990
## 32       35         NA        9990
## 33       36       9990        9990
## 34       37         NA        9990
## 35       38         NA        9990
## 36       39         NA        9990
## 37       40         NA        9990
## 38       41       9990        9990
## 39       42         NA        9990
## 40       44       9990        9990
## 41       45         NA        9990
## 42       46         NA        9990
## 43       47       9990        9990
## 44       48       9990        9990
## 45       49         NA        9990
## 46       50       9990        9990
## 47       51       9990        9990
## 48       53       9990        9990
## 49       54         NA        9990
## 50       55         NA        9990
## 51       56         NA        9990
```

```
hist_COSTELEC21 <- q2_data_electric %>%
                      filter(YEAR == 2021) %>%
                      ggplot(aes(x = COSTELEC)) +
                        geom_histogram(binwidth = 500L, fill = "darkcyan",
color = "darkgrey") +
                        geom_vline(aes(xintercept = mean(COSTELEC)), color =
"blue", linetype = "dashed") +
                        labs(title = "2021") +
                        xlab(label = "Cost of Electric ($)") +
                        ylab(label = "Count Fequency") +
                        theme_classic() +
                        theme(plot.title = element_text(face = "bold"),
                              legend.position = "none")

hist_COSTELEC22 <- q2_data_electric %>%
                      filter(YEAR == 2022) %>%
                      ggplot(aes(x = COSTELEC)) +
                        geom_histogram(binwidth = 500L, fill = "darkcyan",
color = "darkgrey") +
                        geom_vline(aes(xintercept = mean(COSTELEC)), color =
"blue", linetype = "dashed") +
                        labs(title = "2021") +
                        xlab(label = "Cost of Electric ($)") +
                        ylab(label = "Count Fequency") +
                        theme_classic() +
                        theme(plot.title = element_text(face = "bold"),
```

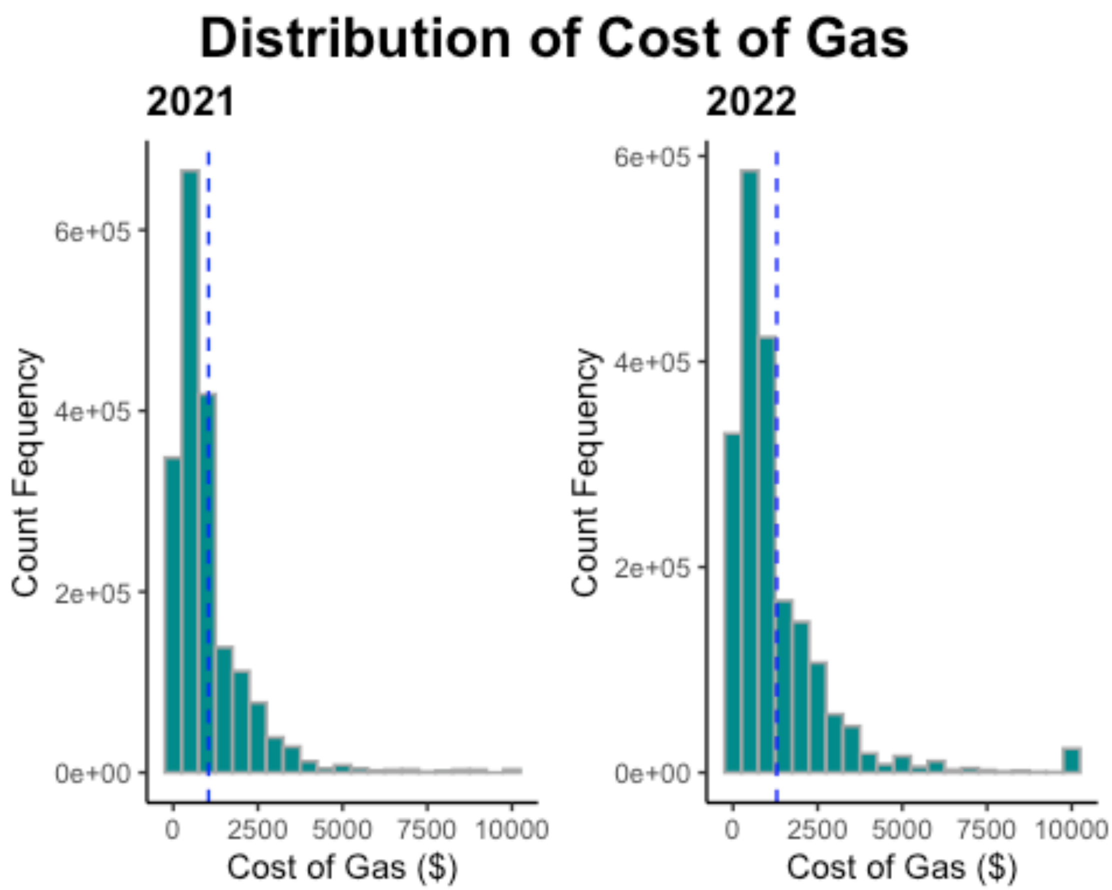
```
                                 legend.position = "none")

hist_COSTELEC <- ggarrange(hist_COSTELEC21, hist_COSTELEC22,
                           ncol = 2, nrow = 1,
                           widths = c(1,1), heights = c(1,1))
hist_COSTELEC <- annotate_figure(hist_COSTELEC,
                          top = text_grob("Distribution of Cost of Electric",
                                          color = "black",
                                          face = "bold",
                                          size = 18))
hist_COSTELEC
```




```
# Cost of Gas
q2_gas_2021 <- q2_data_gas %>%
                 filter(YEAR == 2021) %>%
                 group_by(STATEFIP) %>%
                 summarise(maxCOSTGAS21 = max(COSTGAS)) %>%
                 slice_max(maxCOSTGAS21, n = 1)

q2_gas_2022 <- q2_data_gas %>%
                 filter(YEAR == 2022) %>%
                 group_by(STATEFIP) %>%
                 summarise(maxCOSTGAS22 = max(COSTGAS)) %>%
```

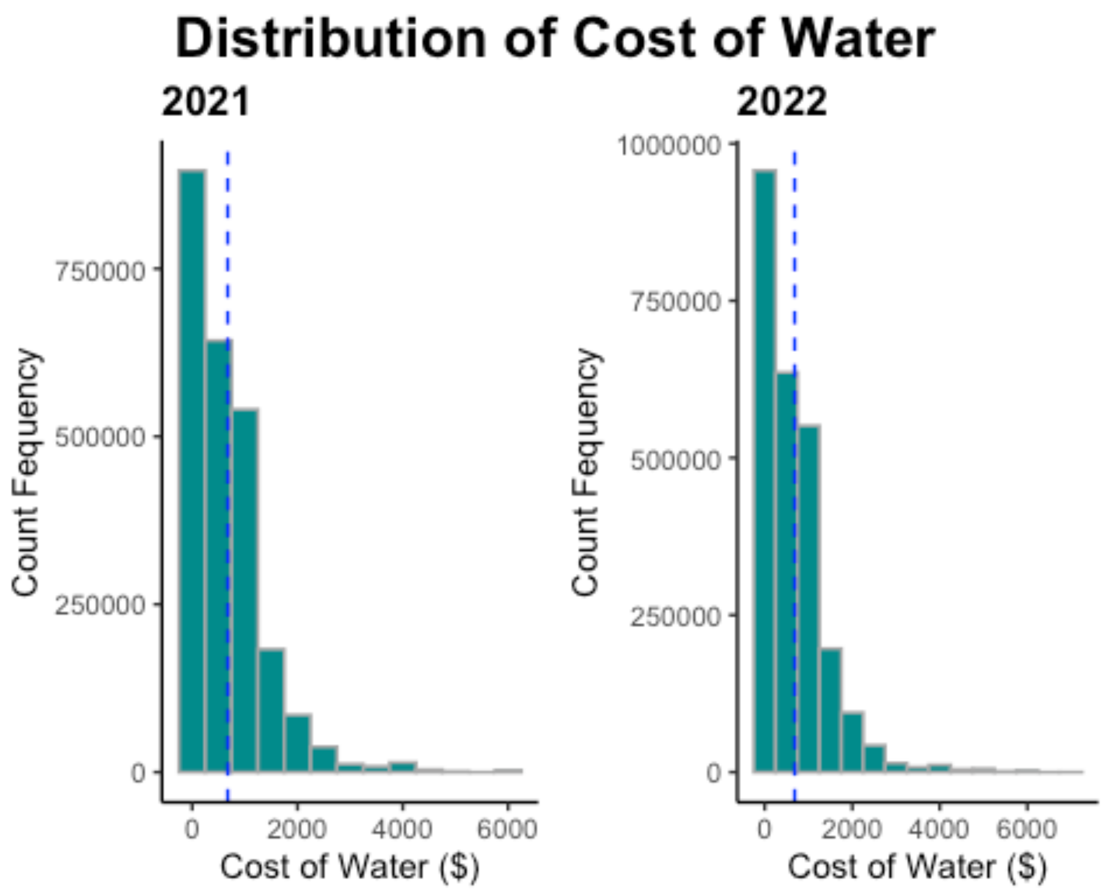
```
                  slice_max(maxCOSTGAS22, n = 1)

q2_gas <- merge(q2_gas_2021, q2_gas_2022, all = TRUE)
q2_gas

##    STATEFIP maxCOSTGAS21 maxCOSTGAS22
## 1         1           NA         9990
## 2         2           NA         9990
## 3         4           NA         9990
## 4         5           NA         9990
## 5         6         9990         9990
## 6         8           NA         9990
## 7         9           NA         9990
## 8        10           NA         9990
## 9        11           NA         9990
## 10       13           NA         9990
## 11       16           NA         9990
## 12       17           NA         9990
## 13       18           NA         9990
## 14       19           NA         9990
## 15       20           NA         9990
## 16       21           NA         9990
## 17       22           NA         9990
## 18       23           NA         9990
## 19       24           NA         9990
## 20       25         9990         9990
## 21       26           NA         9990
## 22       27           NA         9990
## 23       28           NA         9990
## 24       29         9990         9990
## 25       30           NA         9990
## 26       31           NA         9990
## 27       32           NA         9990
## 28       33           NA         9990
## 29       34           NA         9990
## 30       35           NA         9990
## 31       36           NA         9990
## 32       37           NA         9990
## 33       38           NA         9990
## 34       39           NA         9990
## 35       40           NA         9990
## 36       41           NA         9990
## 37       42           NA         9990
## 38       44         9990         9990
## 39       45           NA         9990
## 40       46           NA         9990
## 41       47           NA         9990
## 42       48           NA         9990
## 43       49           NA         9990
## 44       50           NA         9990
```

```
## 45       51           NA         9990
## 46       53           NA         9990
## 47       54           NA         9990
## 48       55           NA         9990
## 49       56           NA         9990

hist_COSTGAS21 <- q2_data_gas %>%
                    filter(YEAR == 2021) %>%
                    ggplot(aes(x = COSTGAS)) +
                      geom_histogram(binwidth = 500L, fill = "darkcyan",
color = "darkgrey") +
                      geom_vline(aes(xintercept = mean(COSTGAS)), color =
"blue", linetype = "dashed") +
                      labs(title = "2021") +
                      xlab(label = "Cost of Gas ($)") +
                      ylab(label = "Count Fequency") +
                      theme_classic() +
                      theme(plot.title = element_text(face = "bold"),
                            legend.position = "none")

hist_COSTGAS22 <- q2_data_gas %>%
                    filter(YEAR == 2022) %>%
                    ggplot(aes(x = COSTGAS)) +
                      geom_histogram(binwidth = 500L, fill = "darkcyan",
color = "darkgrey") +
                      geom_vline(aes(xintercept = mean(COSTGAS)), color =
"blue", linetype = "dashed") +
                      labs(title = "2022") +
                      xlab(label = "Cost of Gas ($)") +
                      ylab(label = "Count Fequency") +
                      theme_classic() +
                      theme(plot.title = element_text(face = "bold"),
                            legend.position = "none")

hist_COSTGAS <- ggarrange(hist_COSTGAS21, hist_COSTGAS22,
                          ncol = 2, nrow = 1,
                          widths = c(1,1), heights = c(1,1))
hist_COSTGAS <- annotate_figure(hist_COSTGAS,
                                top = text_grob("Distribution of Cost of
Gas",
                                                color = "black",
                                                face = "bold",
                                                size = 18))
hist_COSTGAS
```




```
q2_water_2021 <- q2_data_water %>%
                   filter(YEAR == 2021) %>%
                   group_by(STATEFIP) %>%
                   summarise(maxCOSTWATR21 = max(COSTWATR)) %>%
                   slice_max(maxCOSTWATR21, n = 1)

q2_water_2022 <- q2_data_water %>%
                   filter(YEAR == 2022) %>%
                   group_by(STATEFIP) %>%
                   summarise(maxCOSTWATR22 = max(COSTWATR)) %>%
                   slice_max(maxCOSTWATR22, n = 1)

q2_water <- merge(q2_water_2021, q2_water_2022, all = TRUE)
q2_water

##   STATEFIP maxCOSTWATR21 maxCOSTWATR22
## 1        6          6200            NA
## 2       15            NA          7100

hist_COSTWATR21 <- q2_data_water %>%
                     filter(YEAR == 2021) %>%
                     ggplot(aes(x = COSTWATR)) +
                       geom_histogram(binwidth = 500L, fill = "darkcyan",
```

```
color = "darkgrey") +
                       geom_vline(aes(xintercept = mean(COSTWATR)), color =
"blue", linetype = "dashed") +
                       labs(title = "2021") +
                       xlab(label = "Cost of Water ($)") +
                       ylab(label = "Count Fequency") +
                       theme_classic() +
                       theme(plot.title = element_text(face = "bold"),
                             legend.position = "none")

hist_COSTWATR22 <- q2_data_water %>%
                     filter(YEAR == 2022) %>%
                     ggplot(aes(x = COSTWATR)) +
                       geom_histogram(binwidth = 500L, fill = "darkcyan",
color = "darkgrey") +
                       geom_vline(aes(xintercept = mean(COSTWATR)), color =
"blue", linetype = "dashed") +
                       labs(title = "2022") +
                       xlab(label = "Cost of Water ($)") +
                       ylab(label = "Count Fequency") +
                       theme_classic() +
                       theme(plot.title = element_text(face = "bold"),
                             legend.position = "none")

hist_COSTWATR <- ggarrange(hist_COSTWATR21, hist_COSTWATR22,
                           ncol = 2, nrow = 1,
                           widths = c(1,1), heights = c(1,1))
hist_COSTWATR <- annotate_figure(hist_COSTWATR,
                          top = text_grob("Distribution of Cost of Water",
                                          color = "black",
                                          face = "bold",
                                          size = 18))
hist_COSTWATR
```



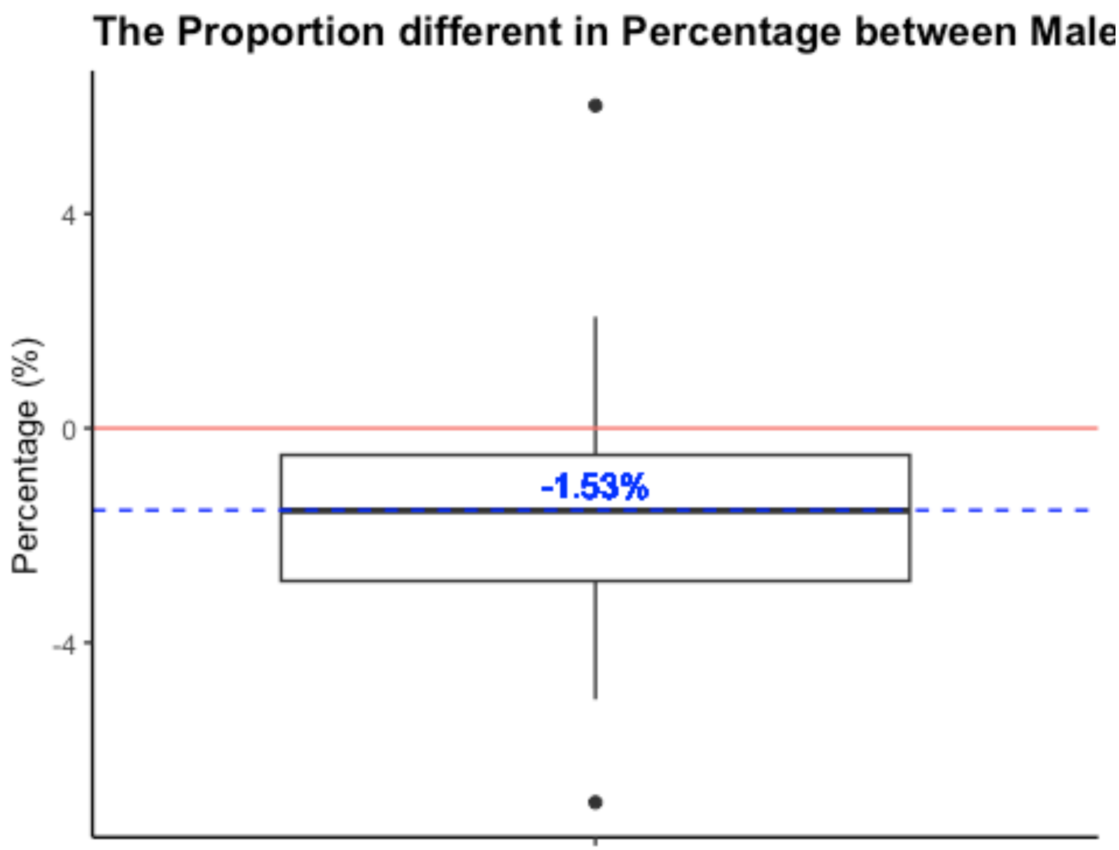
5.3 Question_3; Are there any states with an imbalance in Sex?

Answer_Q3; According to the sample, there are imbalance in Gender in every states. As the table below, the different clearly shown in every states. The highest different is 6.98% at The District of Columbia, in the other hand, the smallest different is 0.04% at The States of Utah.
For a solid conclusion, I performed hypothesis testing as 'H0; mean different between proportion of Male and Female = 0'. The p-value is 0.000001712 which is less than 0.05, so that, H0 is rejected and accept H1. There is the different between the proportion of Male and Female, the mean different is -1.53% which mean the proportion of Female is larger than Male.
I also create a box plot of the proportion different in percentage between Male and Female. The mean and median is closed to each other at -1.53%. The distribution seems to be a normal bell curve, with few potential outliers in both tails. I created the z-score, and found that there is one outlier at the right tail. Alaska is an outlier which have Male 6% more than Female populations.

```
# Data construction
q3_data <- maindf %>%
             select(STATEFIP, SEX) %>%
             filter(SEX != 9) %>%
             group_by(STATEFIP) %>%
```

```
            summarise(Male = COUNTIF(SEX, 1),
                      Female = COUNTIF(SEX, 2),
                      percMale = round(100 * (Male/(Male + Female)), digits
= 2),
                      percFemale = round(100 * (Female/(Male + Female)),
digits = 2),
                      percDiff = percMale - percFemale)

# Hypothesis testing
## H0: Male - Female = 0
## H1: Male - Female <> 0
t.test(x = q3_data$percDiff, y = NULL, alternative = c("two.side"), mu = 0)

##
##  One Sample t-test
##
## data:  q3_data$percDiff
## t = -5.4209, df = 50, p-value = 1.712e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.0966293 -0.9629786
## sample estimates:
## mean of x
## -1.529804

## [1] "p-value = 0.000001712. Then At 95% Confidence Interval, HO is
rejected as p-value < 0.05 and accept alternative hypothesis. The different
between number of Male and Female is more than 0, so there is an imbalance in
Sex"

# Create box plot the different between Male and Female proportion in
percentage
q3_box <- q3_data %>%
            ggplot() +
              geom_boxplot(aes(x = "", y = percDiff)) +
              geom_hline(aes(yintercept = mean(percDiff)), color = "blue",
linetype = "dashed") +
              geom_hline(aes(yintercept = 0, color = "red")) +
              geom_text(aes(x = "",
                            y = mean(percDiff),
                            label = paste(round(mean(percDiff), digits = 2),
"%", sep = "")),
                        color = "blue",
                        vjust = -0.5) +
              labs(title = "The Proportion different in Percentage between
Male and Female") +
              xlab(label = NULL) +
              ylab(label = "Percentage (%)") +
              theme_classic() +
              theme(plot.title = element_text(face = "bold"),
                    legend.position = "none")
```

```
q3_box
```




```
ZpercDiff <- q3_data %>%
               mutate(zScore = (percDiff-mean(percDiff))/sqrt(var(percDiff)))
ZpercDiff

## # A tibble: 51 × 7
##    STATEFIP   Male Female percMale percFemale percDiff zScore
##    <fct>     <int>  <int>    <dbl>      <dbl>    <dbl>  <dbl>
##  1 1         48892  52335     48.3       51.7   -3.40  -0.928
##  2 2          7094   6289     53.0       47.0    6.02   3.75
##  3 4         72898  74280     49.5       50.5   -0.940  0.293
##  4 5         30006  31332     48.9       51.1   -2.16  -0.313
##  5 6        384142 393090     49.4       50.6   -1.16   0.183
##  6 8         59159  58813     50.2       49.8    0.300  0.908
##  7 9         35805  38229     48.4       51.6   -3.28  -0.868
##  8 10         9050  10015     47.5       52.5   -5.06  -1.75
##  9 11         6456   7424     46.5       53.5   -6.98  -2.70
## 10 12       204971 215711     48.7       51.3   -2.56  -0.511
## # ℹ 41 more rows
```

```
ZpercDiff_outlier <- ZpercDiff %>%
                       filter(zScore < -3 | zScore > 3)
ZpercDiff_outlier
```

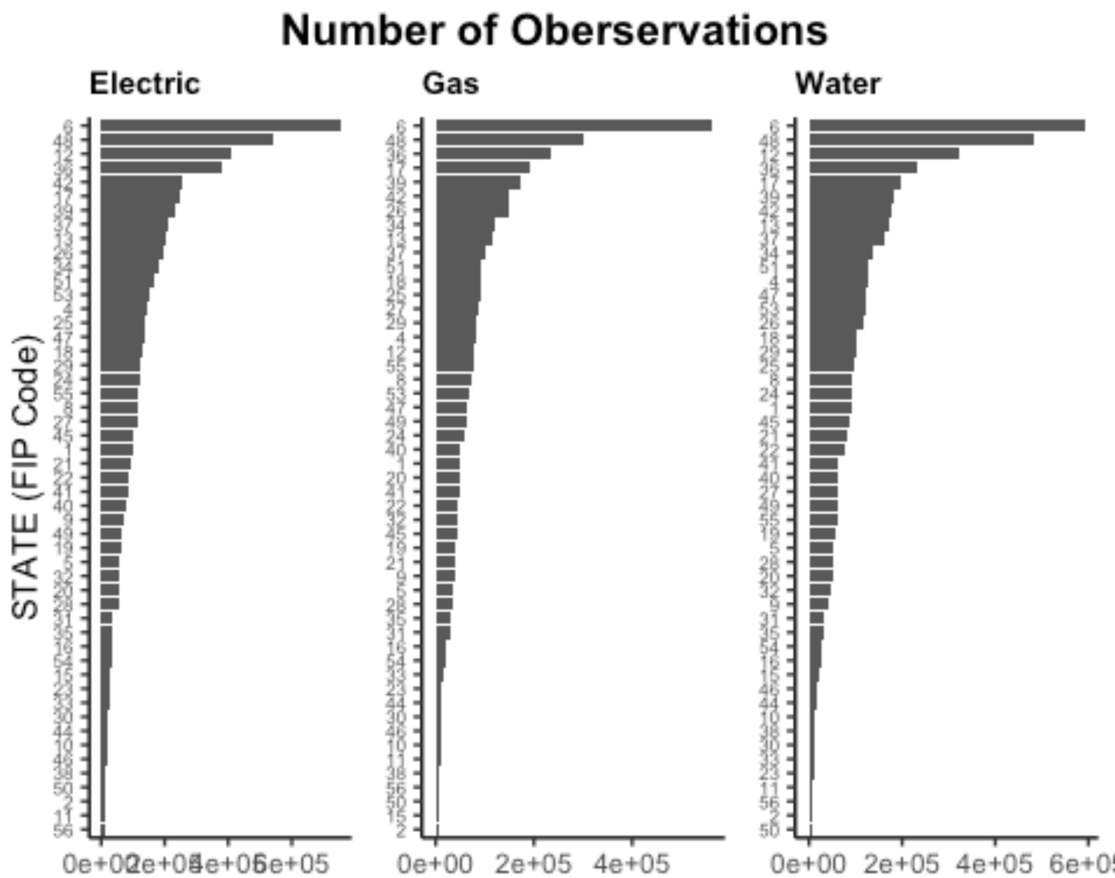
```
## # A tibble: 1 × 7
##   STATEFIP  Male Female percMale percFemale percDiff zScore
##   <fct>    <int>  <int>    <dbl>      <dbl>    <dbl>  <dbl>
## 1 2         7094   6289     53.0       47.0     6.02   3.75
```

5.4 Question_4; Create a new variable that indicates the Total Annual cost that is the sum of the cost of Electricity, Gas, and Water. Which states have the highest total cost?

Answer_Q4; The highest total cost of Electric, Gas, and Water is The State of California (FIP Code = 6), for both 2021 and 2022. I started with filter out unusable rows, according to the code book. Since I curious about the number of observations of each States which is directly effect the calculation. I found that number of observations of The State of California is the highest in every variables, that's strongly support the findings. I also found that The State of California have the highest total cost of each category in both 2021 and 2022.

```
## Subsetting data
q4_data_electric <- maindf %>%
                      select(STATEFIP, YEAR, COSTELEC) %>%
                      filter(COSTELEC < 9993)

q4_data_gas <- maindf %>%
                 select(STATEFIP, YEAR, COSTGAS) %>%
                 filter(COSTGAS < 9992)

q4_data_water <- maindf %>%
                   select(STATEFIP, YEAR, COSTWATR) %>%
                   filter(COSTWATR < 9993)

## Want to know how many observations of each STATES
q4_obs_elec <- q4_data_electric %>%
                 select(STATEFIP) %>%
                 group_by(STATEFIP) %>%
                 summarise(Obs_elec = n()) %>%
                 arrange(desc(Obs_elec))

q4_obs_elec_bar <- q4_obs_elec %>%
                     ggplot() +
                       geom_col(aes(y = fct_reorder(STATEFIP, Obs_elec), x = 
Obs_elec)) +
                       labs(title = "Electric") +
                       ylab(label = "STATE (FIP Code)") +
                       xlab(label = NULL) +
                       theme_classic() +
                       theme(axis.text.y = element_text(size = 6),
```

```
                             plot.title = element_text(size = 10, face =
"bold"))

q4_obs_gas <- q4_data_gas %>%
                select(STATEFIP) %>%
                group_by(STATEFIP) %>%
                summarise(Obs_gas = n()) %>%
                arrange(desc(Obs_gas))

q4_obs_gas_bar <- q4_obs_gas %>%
                    ggplot() +
                      geom_col(aes(y = fct_reorder(STATEFIP, Obs_gas), x =
Obs_gas)) +
                      labs(title = "Gas") +
                      ylab(label = NULL) +
                      xlab(label = NULL) +
                      theme_classic() +
                      theme(axis.text.y = element_text(size = 6),
                            plot.title = element_text(size = 10, face =
"bold"))

q4_obs_water <- q4_data_water %>%
                  select(STATEFIP) %>%
                  group_by(STATEFIP) %>%
                  summarise(Obs_water = n()) %>%
                  arrange(desc(Obs_water))

q4_obs_water_bar <- q4_obs_water %>%
                      ggplot() +
                        geom_col(aes(y = fct_reorder(STATEFIP, Obs_water), x
= Obs_water)) +
                        labs(title = "Water") +
                        ylab(label = NULL) +
                        xlab(label = NULL) +
                        theme_classic() +
                        theme(axis.text.y = element_text(size = 6),
                              plot.title = element_text(size = 10, face =
"bold"))

q4_obs_bar <- ggarrange(q4_obs_elec_bar, q4_obs_gas_bar, q4_obs_water_bar,
                        ncol = 3, nrow = 1)
q4_obs_bar <- annotate_figure(q4_obs_bar, top = text_grob("Number of
Oberservations", size = 14, face = "bold"))
q4_obs_bar
```

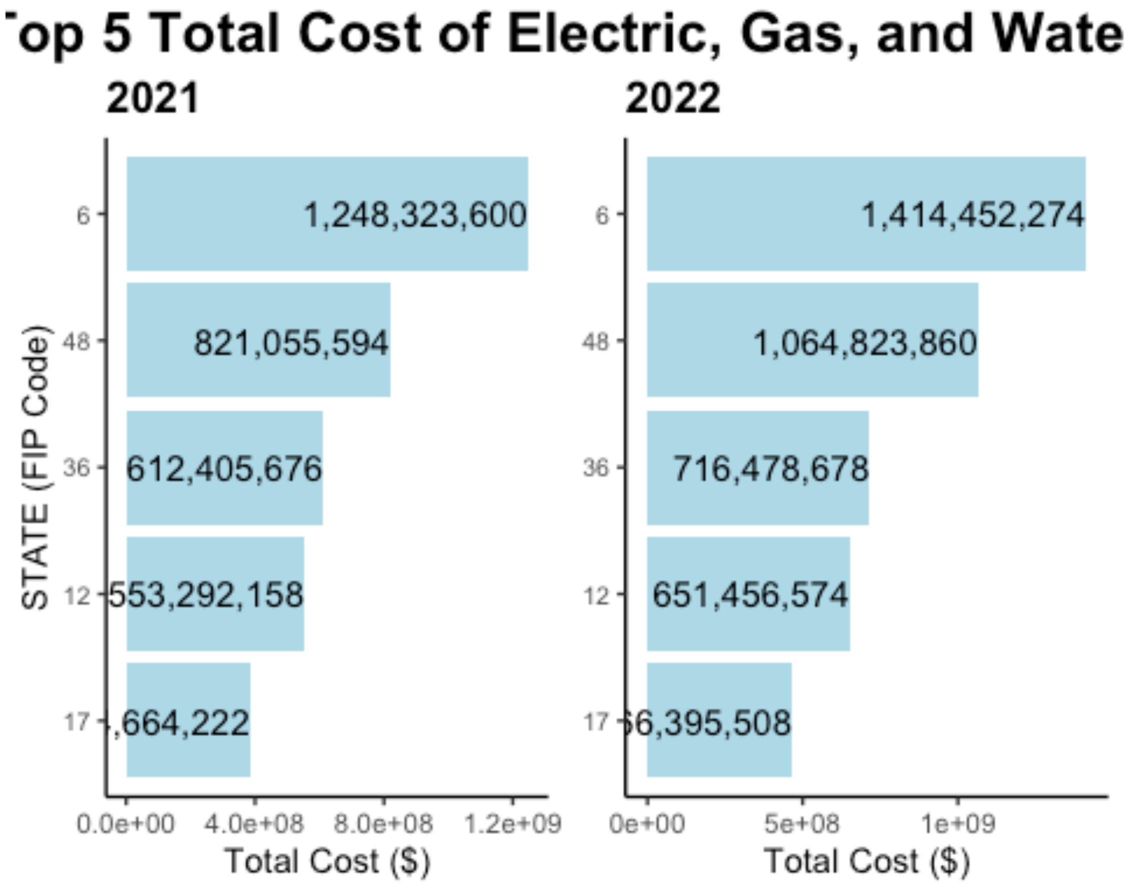



```r
# Total Cost of 2021
q4_electric_2021 <- q4_data_electric %>%
                      filter(YEAR == 2021) %>%
                      group_by(STATEFIP) %>%
                      summarise(sumCOSTELEC21 = sum(COSTELEC))

q4_gas_2021 <- q4_data_gas %>%
                 filter(YEAR == 2021) %>%
                 group_by(STATEFIP) %>%
                 summarise(sumCOSTGAS21 = sum(COSTGAS))

q4_water_2021 <- q4_data_water %>%
                   filter(YEAR == 2021) %>%
                   group_by(STATEFIP) %>%
                   summarise(sumCOSTWATR21 = sum(COSTWATR))

q4_totalcost21 <- merge(q4_electric_2021, q4_gas_2021) %>%
                    merge(q4_water_2021) %>%
                    mutate(TotalCost21 = sumCOSTELEC21 + sumCOSTGAS21 +
sumCOSTWATR21) %>%
                    arrange(desc(TotalCost21))
```

```r
# Total Cost of 2022
q4_electric_2022 <- q4_data_electric %>%
                      filter(YEAR == 2022) %>%
                      group_by(STATEFIP) %>%
                      summarise(sumCOSTELEC22 = sum(COSTELEC))

q4_gas_2022 <- q4_data_gas %>%
                 filter(YEAR == 2022) %>%
                 group_by(STATEFIP) %>%
                 summarise(sumCOSTGAS22 = sum(COSTGAS))

q4_water_2022 <- q4_data_water %>%
                   filter(YEAR == 2022) %>%
                   group_by(STATEFIP) %>%
                   summarise(sumCOSTWATR22 = sum(COSTWATR))

q4_totalcost22 <- merge(q4_electric_2022, q4_gas_2022) %>%
                    merge(q4_water_2022) %>%
                    mutate(TotalCost22 = sumCOSTELEC22 + sumCOSTGAS22 +
sumCOSTWATR22) %>%
                    arrange(desc(TotalCost22))

# Top 5 States in total cost of Electric, Gas, Water
q4_totalcost21_t5 <- slice_max(q4_totalcost21, TotalCost21, n = 5)
q4_totalcost21_t5

##   STATEFIP sumCOSTELEC21 sumCOSTGAS21 sumCOSTWATR21 TotalCost21
## 1        6     744658068    235149366     268516166  1248323600
## 2       48     555726072    108111336     157218186   821055594
## 3       36     378052986    165990360      68362330   612405676
## 4       12     431143950     22311816      99836392   553292158
## 5       17     211432296    109022112      64209814   384664222

q4_totalcost22_t5 <- slice_max(q4_totalcost22, TotalCost22, n = 5)
q4_totalcost22_t5

##   STATEFIP sumCOSTELEC22 sumCOSTGAS22 sumCOSTWATR22 TotalCost22
## 1        6     858151356    283834272     272466646  1414452274
## 2       48     726873864    155045226     182904770  1064823860
## 3       36     437374146    208562148      70542384   716478678
## 4       12     509279982     32385480     109791112   651456574
## 5       17     243387516    156151182      66856810   466395508

q4_totalcost_t5 <- merge(q4_totalcost21_t5, q4_totalcost22_t5) %>%
                     select(STATEFIP, TotalCost21, TotalCost22) %>%
                     arrange(desc(TotalCost22))
q4_totalcost_t5

##   STATEFIP TotalCost21 TotalCost22
## 1        6  1248323600  1414452274
## 2       48   821055594  1064823860
```

```
## 3       36   612405676   716478678
## 4       12   553292158   651456574
## 5       17   384664222   466395508
```

```
q4_totalcost21_bar <- q4_totalcost_t5 %>%
                        ggplot() +
                          geom_col(aes(y = fct_reorder(STATEFIP,
TotalCost21), x = TotalCost21), fill = "lightblue") +
                          geom_text(aes(x = TotalCost21, y = STATEFIP, label
= scales::comma(TotalCost21)), hjust = 1, size = 4) +
                          labs(title = "2021") +
                          ylab(label = "STATE (FIP Code)") +
                          xlab(label = "Total Cost ($)") +
                          theme_classic() +
                          theme(axis.text.y = element_text(size = 8),
                                plot.title = element_text(size = 14, face =
"bold"))

q4_totalcost22_bar <- q4_totalcost_t5 %>%
                        ggplot() +
                          geom_col(aes(y = fct_reorder(STATEFIP,
TotalCost22), x = TotalCost22), fill = "lightblue") +
                          geom_text(aes(x = TotalCost22, y = STATEFIP, label
= scales::comma(TotalCost22)), hjust = 1, size = 4) +
                          labs(title = "2022") +
                          ylab(label = NULL) +
                          xlab(label = "Total Cost ($)") +
                          theme_classic() +
                          theme(axis.text.y = element_text(size = 8),
                                plot.title = element_text(size = 14, face =
"bold"))

q4_totalcost_bar <- ggarrange(q4_totalcost21_bar, q4_totalcost22_bar, ncol =
2, nrow = 1)
q4_totalcost_bar <- annotate_figure(q4_totalcost_bar, top = text_grob("Top 5
Total Cost of Electric, Gas, and Water", size = 18, face = "bold"))
q4_totalcost_bar
```



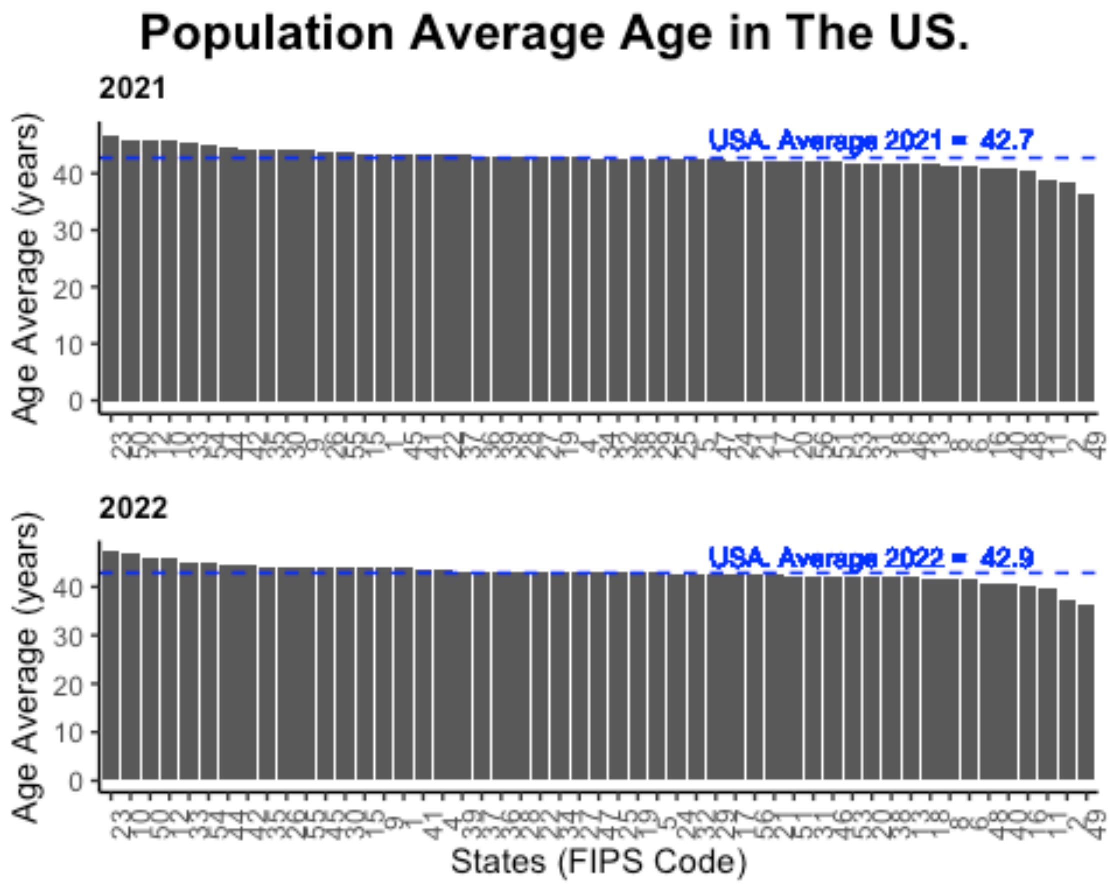
5.5 Question_5; Which state has the oldest, on average, residents?

Answer_Q5; The State of Maine (FIP Code = 23) on average has the oldest residents with the age of 46.8 years old in 2021 and 47.2 years old in 2022. The average age of USA residents grows from 42.7 years old in 2021 to 42.9 years old in 2022.

```
q5_age2021 <- maindf %>%
                select(STATEFIP, YEAR, AGE) %>%
                filter(YEAR == 2021) %>%
                group_by(STATEFIP) %>%
                summarise(avgAge21 = round(mean(AGE), digits = 1))

q5_age2022 <- maindf %>%
                select(STATEFIP, YEAR, AGE) %>%
                filter(YEAR == 2022) %>%
                group_by(STATEFIP) %>%
                summarise(avgAge22 = round(mean(AGE), digits = 1))

q5_age2021_max <- slice_max(q5_age2021, avgAge21, n = 1)
q5_age2022_max <- slice_max(q5_age2022, avgAge22, n =1)

q5_age <- merge(q5_age2021_max, q5_age2022_max, all = TRUE)
q5_age
```

```
##   STATEFIP avgAge21 avgAge22
## 1       23     46.8     47.2
```

```
q5_age21_hist <- q5_age2021 %>%
                   ggplot() +
                     geom_col(aes(x = fct_rev(fct_reorder(STATEFIP,
avgAge21)), y = avgAge21)) +
                     geom_hline(aes(yintercept = mean(avgAge21)), color =
"blue", linetype = "dashed") +
                     geom_text(aes(y = 46, x = 40, label = paste("USA.
Average 2021 = ", round(mean(avgAge21),digits = 1))),
                               color = "blue",
                               size = 3,
                               face = "bold") +
                     labs(title = "2021") +
                     ylab(label = "Age Average (years)") +
                     xlab(label = NULL) +
                     theme_classic() +
                     theme(plot.title = element_text(face = "bold", size =
10),
                           axis.text.x = element_text(angle = 90))

q5_age22_hist <- q5_age2022 %>%
                   ggplot() +
                     geom_col(aes(x = fct_rev(fct_reorder(STATEFIP,
avgAge22)), y = avgAge22)) +
                     geom_hline(aes(yintercept = mean(avgAge22)), color =
"blue", linetype = "dashed") +
                     geom_text(aes(y = 46, x = 40, label = paste("USA.
Average 2022 = ", round(mean(avgAge22),digits = 1))),
                               color = "blue",
                               size = 3,
                               face = "bold") +
                     labs(title = "2022") +
                     ylab(label = "Age Average (years)") +
                     xlab(label = "States (FIPS Code)") +
                     theme_classic() +
                     theme(plot.title = element_text(face = "bold", size =
10),
                           axis.text.x = element_text(angle = 90))

qq_age_hist <- ggarrange(q5_age21_hist, q5_age22_hist,
                         ncol = 1, nrow = 2)
qq_age_hist <- annotate_figure(qq_age_hist,
                               top = text_grob("Population Average Age in The
US.",
                                                color = "black",
                                                face = "bold",
                                                size = 16))
qq_age_hist
```

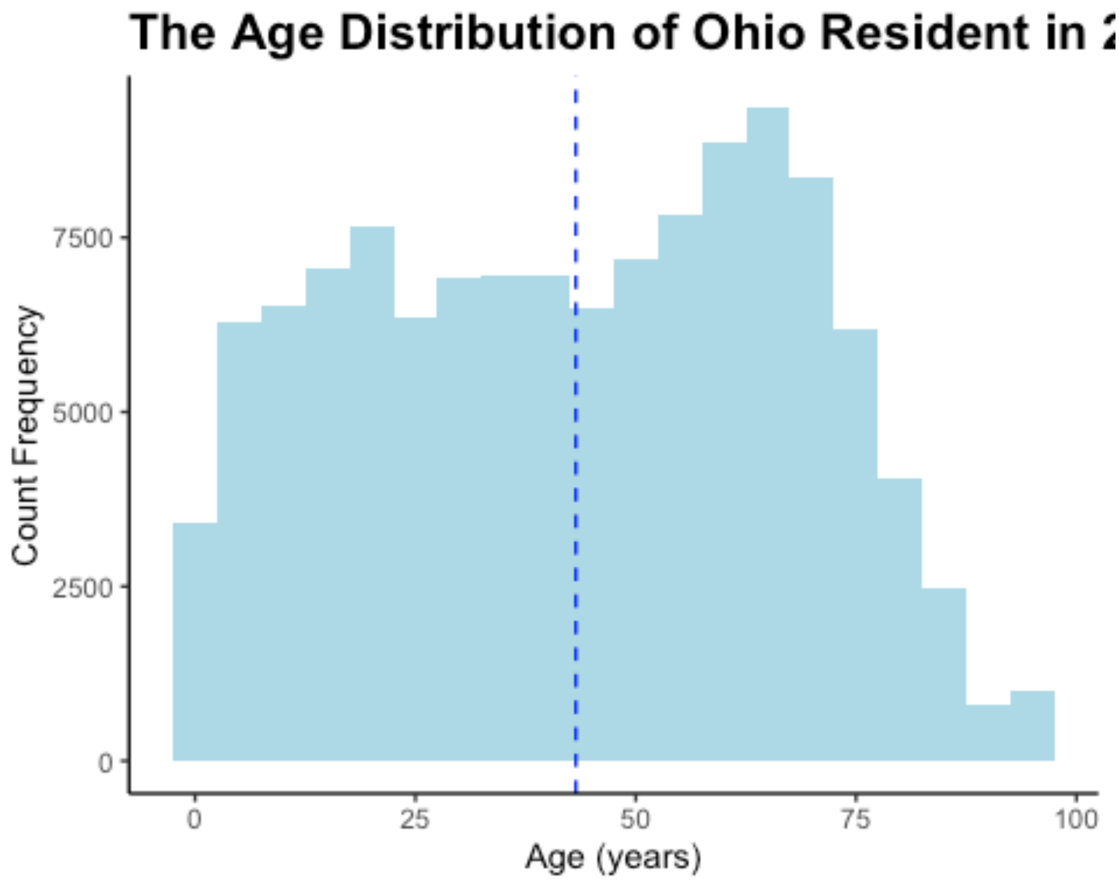


5.6 Question_6; What can you say about the residents of Ohio based on their age, sex, race, and language. Use only the most recent data.

Answer_Q6; In this question, I discovered 5 interesting things. FIP Code for The State of Ohio is 39.
NOTED: Number of resident that live in Ohio according to the data is 120,666 observations for the year of 2022.
1) The average age of resident in The State of Ohio is 43.2 which is a bit higher than US average. The distribution of Age shows a little bit of left skew since the mean is a bit less than the median, however, it also looks like normal bell curve with no skewness.
2) I analysed 'SEX' and 'AGE' together. For 2022, I found that an average age of male is 2.5 years, almost 6%, lower than female, 41.9 compared to 44.4 years. The distribution of female age skew to the left more than male, however, both of it seems to be a normal bell shape. 3) There are more Female live in Ohio than Male, with the proportion of 51.15 to 48.85. The different is -2.3% which is higher than nation wide different, -1.5%.
4) According to this data, there are all 9 races live in Ohio. However, White people is the dominant race with almost 82%. There are few Asians live in Ohio, since I'm from Thailand, my race live here only 1.7%.
5) There are 97 languages use in The US, however, in Ohio, there are only 57 languages reported. Expected, the most use languages at home is English with the proportion of

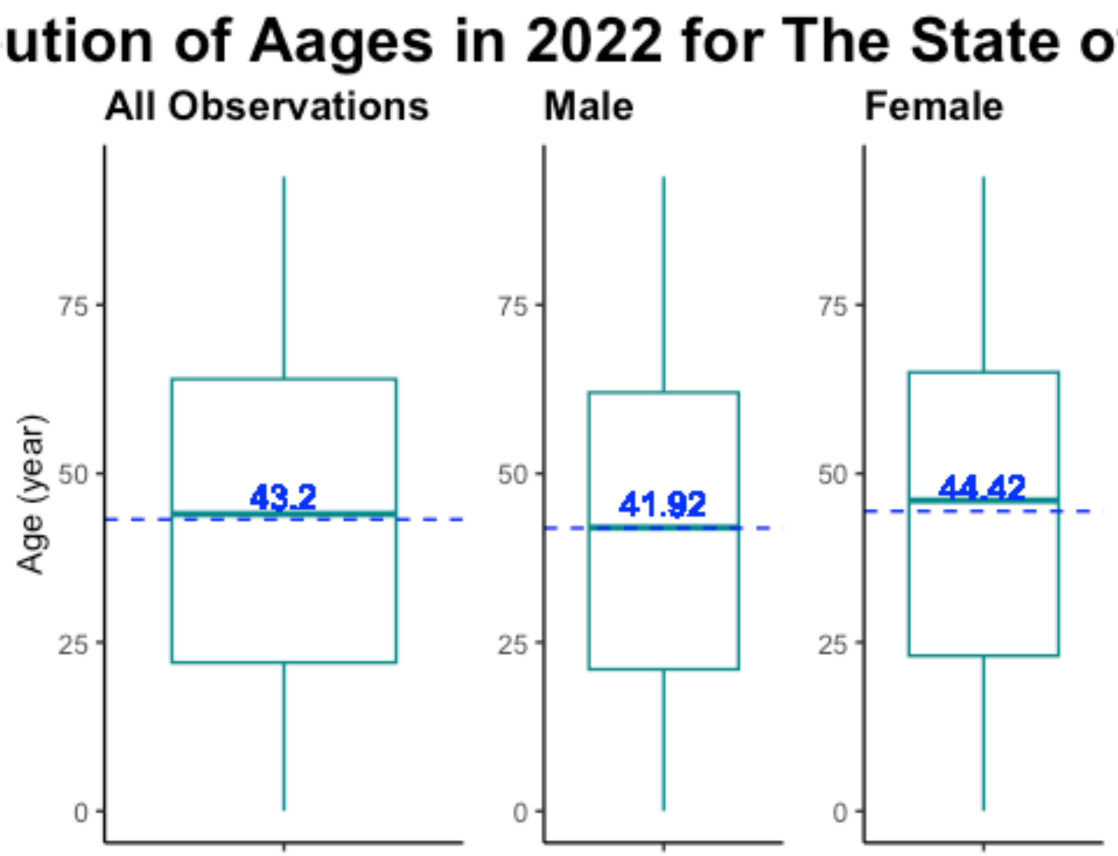
88.7%. The second place is 'N/A or blank' which means almost 5% don't answer this question. Unsurprisingly, there are only 51 persons who speak 'Thai/Laos at their home'.

```
## Subsetting data
q6_data <- maindf %>%
            filter(STATEFIP == 39, YEAR == 2022) %>%
            select(AGE, SEX, RACE, LANGUAGE)

## Q_6.1. Age
q6_age_box <- q6_data %>%
                ggplot() +
                  geom_boxplot(aes(x = "", y = AGE), color = "darkcyan") +
                  geom_hline(aes(yintercept = mean(AGE)), color = "blue",
linetype = "dashed") +
                  geom_text(aes(x = "",
                                y = mean(AGE),
                                label = round(mean(AGE), digits = 2)),
                            vjust = -0.5,
                            color = "blue") +
                  labs(title = "All Observations") +
                  ylab(label = "Age (year)") +
                  xlab(label = NULL) +
                  theme_classic() +
                  theme(plot.title = element_text(face = 'bold'))

q6_age_hist <- q6_data %>%
                 ggplot() +
                   geom_histogram(aes(x = AGE), binwidth = 5L, fill =
"lightblue") +
                   geom_vline(aes(xintercept = mean(AGE)), color = "blue",
linetype = "dashed") +
                   labs(title = "The Age Distribution of Ohio Resident in
2022") +
                   xlab(label = "Age (years)") +
                   ylab(label = "Count Frequency") +
                   theme_classic() +
                   theme(plot.title = element_text(size = 16, face = "bold"))
q6_age_hist
```



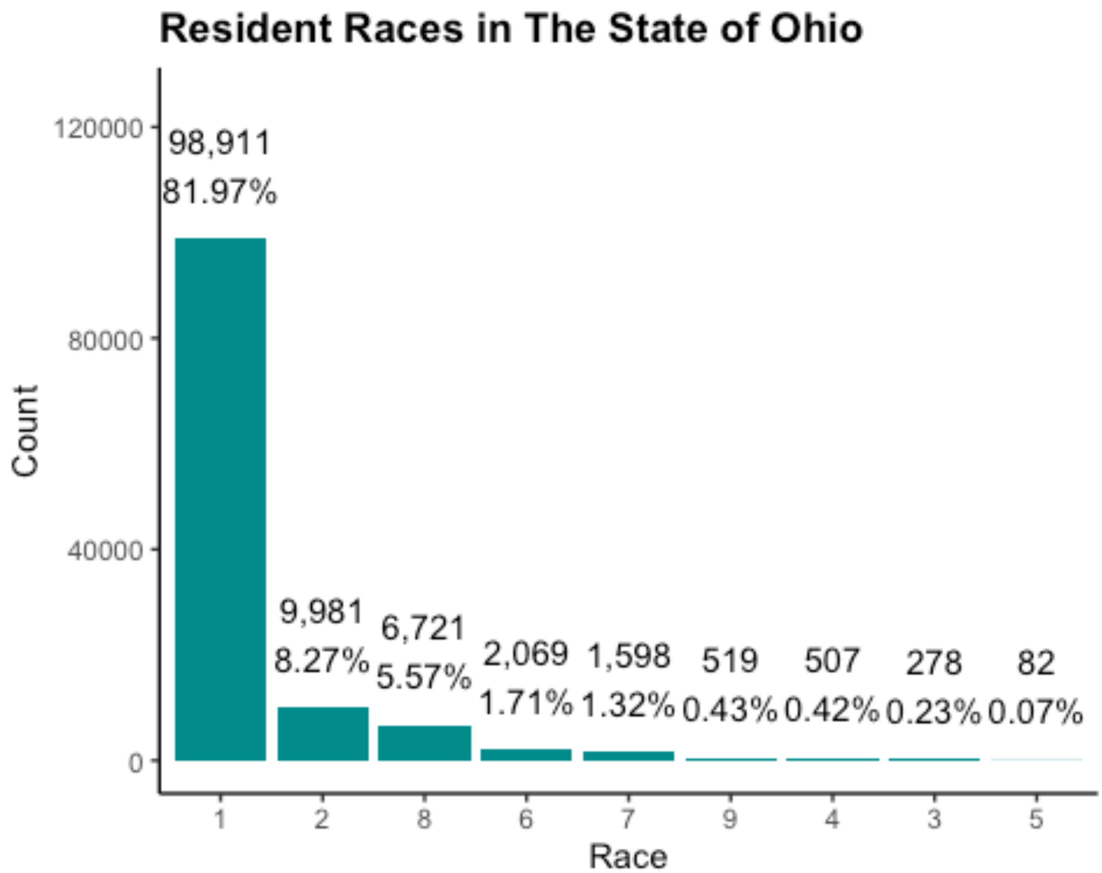

```
## Q_6.2. Age + SEX
q6_agemale_box <- q6_data %>%
                    filter(SEX == 1) %>%
                    ggplot() +
                      geom_boxplot(aes(x = "", y = AGE), color = "darkcyan")
+
                      geom_hline(aes(yintercept = mean(AGE)), color = "blue",
linetype = "dashed") +
                      geom_text(aes(x = "",
                                    y = mean(AGE),
                                    label = round(mean(AGE), digits = 2)),
                                vjust = -0.5,
                                color = "blue") +
                      labs(title = "Male") +
                      ylab(label = NULL) +
                      xlab(label = NULL) +
                      theme_classic() +
                      theme(plot.title = element_text(face = 'bold'))

q6_agefemale_box <- q6_data %>%
                      filter(SEX == 2) %>%
                      ggplot() +
                        geom_boxplot(aes(x = "", y = AGE), color =
```

```
"darkcyan") +
                        geom_hline(aes(yintercept = mean(AGE)), color =
"blue", linetype = "dashed") +
                        geom_text(aes(x = "",
                                      y = mean(AGE),
                                      label = round(mean(AGE), digits = 2)),
                                  vjust = -0.5,
                                  color = "blue") +
                        labs(title = "Female") +
                        ylab(label = NULL) +
                        xlab(label = NULL) +
                        theme_classic() +
                        theme(plot.title = element_text(face = 'bold'))

q6_box <- ggarrange(q6_age_box, q6_agemale_box, q6_agefemale_box,
                    ncol = 3, nrow = 1,
                    widths = c(1.5,1,1), heights = c(1,1,1),
                    common.legend = TRUE,
                    align = 'h')
q6_box <- annotate_figure(q6_box,
                          top = text_grob("Distribution of Aages in 2022 for
The State of Ohio ",
                                          color = "black",
                                          face = "bold",
                                          size = 20))
q6_box
```



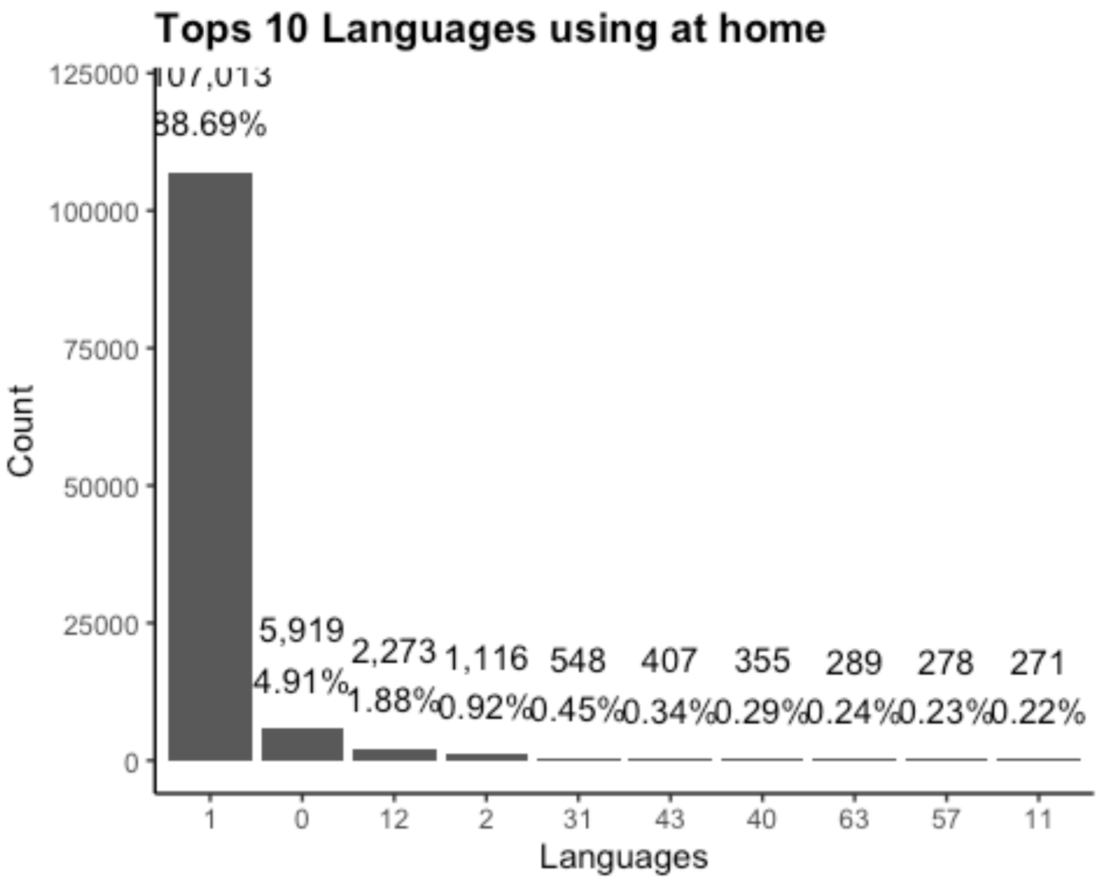

```
## Q_6.3. SEX
q6_sex <- q6_data %>%
            filter(SEX != 9) %>%
            select(SEX) %>%
            summarise(Male = COUNTIF(SEX, 1),
                      Female = COUNTIF(SEX, 2),
                      percMale = round(100 * (Male/(Male + Female)), digits =
2),
                      percFemale = round(100 * (Female/(Male + Female)),
digits = 2),
                      percDiff = percMale - percFemale)
q6_sex

##    Male Female percMale percFemale percDiff
## 1 58942  61724    48.85      51.15     -2.3

## Q_6.4. Race
q6_race <- q6_data %>%
             select(RACE) %>%
             group_by(RACE) %>%
             summarise(RaceCount = n(),
                       percRaceCount = round(100 * RaceCount/nrow(q6_data),
digits = 2)) %>%
```

```
                arrange(desc(RaceCount))
q6_race

## # A tibble: 9 × 3
##   RACE  RaceCount percRaceCount
##   <fct>     <int>         <dbl>
## 1 1         98911         82.0
## 2 2          9981          8.27
## 3 8          6721          5.57
## 4 6          2069          1.71
## 5 7          1598          1.32
## 6 9           519          0.43
## 7 4           507          0.42
## 8 3           278          0.23
## 9 5            82          0.07

q6_race_chart <- q6_race %>%
                      ggplot() +
                        geom_col(aes(x = fct_rev(fct_reorder(RACE, RaceCount)),
y = RaceCount), fill = "darkcyan") +
                        geom_text(aes(x = fct_rev(fct_reorder(RACE, RaceCount)),
                                      y = RaceCount,
                                      label = paste(scales::comma(RaceCount),
paste(percRaceCount, "%", sep = ""), sep = "\n")),
                                  vjust = -0.5) +
                        ylim(0, 125000) +
                        labs(title = "Resident Races in The State of Ohio") +
                        xlab("Race") +
                        ylab("Count") +
                        theme_classic() +
                        theme(plot.title = element_text(face = "bold"))

q6_race_chart
```




```
## Q_6.5. Language
q6_language <- q6_data %>%
                 select(LANGUAGE) %>%
                 group_by(LANGUAGE) %>%
                 summarise(LangCount = n(),
                           percLangCount = round(100 *
LangCount/nrow(q6_data), digits = 2)) %>%
                 arrange(desc(LangCount))
q6_language

## # A tibble: 57 × 3
##    LANGUAGE LangCount percLangCount
##    <fct>        <int>         <dbl>
##  1 1           107013         88.7
##  2 0             5919          4.91
##  3 12            2273          1.88
##  4 2             1116          0.92
##  5 31             548          0.45
##  6 43             407          0.34
##  7 40             355          0.29
##  8 63             289          0.24
##  9 57             278          0.23
```

```
## 10 11               271          0.22
## # ℹ 47 more rows

q6_language_thai <- q6_language %>%
                      filter(LANGUAGE == 47)
q6_language_thai

## # A tibble: 1 × 3
##   LANGUAGE LangCount percLangCount
##   <fct>        <int>         <dbl>
## 1 47              51          0.04

q6_language_chart <- q6_language %>%
                       slice_max(LangCount, n = 10) %>%
                       ggplot() +
                         geom_col(aes(x = fct_rev(fct_reorder(LANGUAGE,
LangCount)), y = LangCount)) +
                         geom_text(aes(x = fct_rev(fct_reorder(LANGUAGE,
LangCount)),
                                   y = LangCount,
                                   label = paste(scales::comma(LangCount),
paste(percLangCount, "%", sep = ""), sep = "\n")),
                               vjust = -0.5) +
                         ylim(0, 120000) +
                         labs(title = "Tops 10 Languages using at home") +
                         xlab("Languages") +
                         ylab("Count") +
                         theme_classic() +
                         theme(plot.title = element_text(face = "bold"))
q6_language_chart
```



Tops 10 Languages using at home
107,013
88.69%
5,919
4.91%
2,273
1.88%
1,116
0.92%
548
0.45%
407
0.34%
355
0.29%
289
0.24%
278
0.23%
271
0.22%
125000
100000
75000
50000
25000
0
Count
1
0
12
2
31
43
40
63
57
11
Languages