Assignment 2_BA

VaishnaviD

2024-03-10

Summary

My analysis is derived from “usa_00006”. This involved examining a comprehensive dataset related to demographics, utility costs, and other variables within the United States.The summary provides insights into the dataset’s characteristics and findings:

1. Missing Values: Exploration with glimpse and colSums revealed that there are no missing values ensuring the dataset’s readiness for further analysis.
2. The highest electricity costs across all 51 states are the same.
   49 states have the highest gas costs.
   Hawai has the highest water cost.
3. Yes, all the states have imbalance in sex. In every state there’s a difference in the number of men and women; none have an equal ratio of males to females.
4. California (state 6) had the highest total cost in 2021 and 2022.
5. Maine (state 23) has the oldest, on average, residents
6. In the year 2022 below points are observed,

- the average age in Ohio is 43.
- there are more female than male in Ohio.
- it is observed that white people are the largest population group in Ohio, followed by Black/African American, with Japanese being the smallest.
- English was the most widely spoken language, while Yedish, Celtic, Aleut, Eskimo, and Iroquoian were the least spoken languages.

To load required libraries and import the data

```r
library(readxl)
library(knitr)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(dplyr)
library(tidyr)
Bigdata <- read.csv("C:/Users/Vaishnavi/OneDrive - Kent State University/BA/Assignment 2/usa_00006.csv")
```

To confirm that all the data has been properly imported

```r
head(Bigdata)
```

```
##   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(Bigdata)
```

```
##         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
```

```r
nrow(Bigdata)
```

```
## [1] 6625977
```

Question 1: Are there any missing values?

Exploration with glimpse and colSums revealed that there are no missing values ensuring the dataset’s readiness for further analysis.

```r
# To check if there are any missing values
glimpse(Bigdata)
```

```
## Rows: 6,625,977
## Columns: 21
## $ YEAR      <int> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, …
## $ SAMPLE    <int> 202101, 202101, 202101, 202101, 202101, 202101, 202101, 2021…
## $ SERIAL    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1…
## $ CBSERIAL  <dbl> 2.02101e+12, 2.02101e+12, 2.02101e+12, 2.02101e+12, 2.02101e…
## $ HHWT      <dbl> 13, 51, 17, 61, 15, 46, 55, 31, 71, 48, 76, 31, 20, 106, 19,…
## $ CLUSTER   <dbl> 2.021e+12, 2.021e+12, 2.021e+12, 2.021e+12, 2.021e+12, 2.021…
## $ STATEFIP  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ STRATA    <int> 80001, 80001, 120001, 170001, 50001, 160001, 130201, 210001,…
## $ GQ        <int> 3, 3, 3, 3, 3, 4, 3, 3, 3, 4, 3, 4, 3, 4, 4, 3, 4, 3, 3, 3, …
## $ COSTELEC  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ COSTGAS   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ COSTWATR  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ COSTFUEL  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ PERNUM    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ PERWT     <dbl> 13, 51, 17, 61, 15, 46, 55, 31, 71, 48, 76, 31, 20, 106, 19,…
## $ SEX       <int> 1, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 1, 2, 1, 2, …
## $ AGE       <int> 85, 67, 74, 16, 83, 19, 36, 35, 45, 20, 27, 19, 85, 30, 34, …
## $ RACE      <int> 1, 2, 1, 1, 1, 1, 2, 2, 1, 2, 2, 1, 1, 1, 1, 1, 1, 2, 2, 1, …
## $ RACED     <int> 100, 200, 100, 100, 100, 100, 200, 200, 100, 200, 200, 100, …
## $ LANGUAGE  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ LANGUAGED <int> 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, …
```

```r
colSums(is.na(Bigdata))
```

```
##      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
```

Question 2: Identify the states with the highest cost of electricity, gas, and water.

The highest electricity costs across all 51 states are the same.
49 states have the highest gas costs.
Hawai has the highest water cost.

```r
# Highest cost of the electricity
VD<-Bigdata %>%
filter(COSTELEC < 9993)%>%
  select(STATEFIP,COSTELEC)
summary(VD)
```

```
##     STATEFIP        COSTELEC   
##  Min.   : 1.00   Min.   :   0  
##  1st Qu.:12.00   1st Qu.:1080  
##  Median :27.00   Median :1800  
##  Mean   :27.72   Mean   :2110  
##  3rd Qu.:42.00   3rd Qu.:2760  
##  Max.   :56.00   Max.   :9990
```

```r
# The maximum cost of the electricity is 9990
  High_cost_Electricity <- VD %>%
  group_by(STATEFIP)%>%
 summarise(ELECTRICITY = max(COSTELEC))%>%
  slice_max(ELECTRICITY,n=1)
```

```r
# Highest cost of the gas
VD<-Bigdata %>%
filter(COSTGAS < 9992 )%>%
  select(STATEFIP,COSTGAS)
summary(VD)
```

```
##     STATEFIP        COSTGAS    
##  Min.   : 1.00   Min.   :   0  
##  1st Qu.:13.00   1st Qu.: 360  
##  Median :27.00   Median : 720  
##  Mean   :27.55   Mean   :1171  
##  3rd Qu.:41.00   3rd Qu.:1440  
##  Max.   :56.00   Max.   :9990
```

```r
# The maximum cost of the gas is 9990
  High_cost_Gas <- VD %>%
  group_by(STATEFIP)%>%
 summarise(GAS = max(COSTGAS))%>%
  slice_max(GAS,n=1)
```

```r
# Highest cost of Water
VD<-Bigdata %>%
filter(COSTWATR < 9993 )%>%
  select(STATEFIP,COSTWATR)
summary(VD)
```

```
##     STATEFIP        COSTWATR   
##  Min.   : 1.00   Min.   :   0  
##  1st Qu.:12.00   1st Qu.: 100  
##  Median :27.00   Median : 500  
##  Mean   :27.47   Mean   : 680  
##  3rd Qu.:42.00   3rd Qu.:1000  
##  Max.   :56.00   Max.   :7100
```

```r
# The maximum cost of the Water is 7100
  High_cost_water <- VD %>%
  group_by(STATEFIP)%>%
 summarise(WATER = max(COSTWATR))%>%
  slice_max(WATER,n=1)
```

Question 3: Are there any states with an imbalance in Sex?

Yes, all the states have imbalance in sex. In every state there’s a difference in the number of men and women. None have an equal ratio of males to females.

```r
Bigdata %>%
  group_by(STATEFIP) %>%
  summarise(Female = sum(PERWT[SEX == 2], na.rm = TRUE),
            Male = sum(PERWT[SEX == 1], na.rm = TRUE)) %>%
  mutate(Imbalance = (Female - Male)) %>%
  arrange(desc(Imbalance))
```

```
## # A tibble: 51 × 4
##    STATEFIP   Female     Male Imbalance
##       <int>    <dbl>    <dbl>     <dbl>
##  1       36 20196046 19317018    879028
##  2       12 22367476 21658475    709001
##  3       13 11122256 10590186    532070
##  4       37 10866087 10384048    482039
##  5       42 13139111 12796953    342158
##  6       24  6321838  6007951    313887
##  7       25  7137816  6828881    308935
##  8       39 11922157 11613918    308239
##  9        1  5207913  4906260    301653
## 10       34  9404501  9124328    280173
## # ℹ 41 more rows
```

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?

California (state 6) had the highest total cost in 2021 and 2022.

```r
ELE21<-Bigdata %>%
  filter(COSTELEC < 9993 & YEAR==2021)%>%
  select(STATEFIP,COSTELEC)%>%
  group_by(STATEFIP)%>%
 summarise(E.21 = sum(COSTELEC))
head(ELE21)
```

```
## # A tibble: 6 × 2
##   STATEFIP      E.21
##      <int>     <int>
## 1        1 119766288
## 2        2  12607944
## 3        4 150361536
## 4        5  61909752
## 5        6 744658068
## 6        8  96558618
```

```r
ELE22<-Bigdata %>%
  filter(COSTELEC < 9993 & YEAR==2022)%>%
  select(STATEFIP,COSTELEC)%>%
  group_by(STATEFIP)%>%
 summarise(E.22 = sum(COSTELEC))
head(ELE22)
```

```
## # A tibble: 6 × 2
##   STATEFIP      E.22
##      <int>     <int>
## 1        1 133344858
## 2        2  14476350
## 3        4 158815002
## 4        5  71418264
## 5        6 858151356
## 6        8 108302376
```

```r
GAS21<-Bigdata %>%
  filter(COSTGAS < 9992 & YEAR==2021)%>%
  select(STATEFIP,COSTGAS)%>%
  group_by(STATEFIP)%>%
 summarise(G.21 = sum(COSTGAS))
head(GAS21)
```

```
## # A tibble: 6 × 2
##   STATEFIP      G.21
##      <int>     <int>
## 1        1  20684952
## 2        2   4924608
## 3        4  25782984
## 4        5  16792488
## 5        6 235149366
## 6        8  34607424
```

```r
GAS22<-Bigdata %>%
  filter(COSTGAS < 9992 & YEAR==2022)%>%
  select(STATEFIP,COSTGAS)%>%
  group_by(STATEFIP)%>%
 summarise(G.22 = sum(COSTGAS))
head(GAS22)
```

```
## # A tibble: 6 × 2
##   STATEFIP      G.22
##      <int>     <int>
## 1        1  25073178
## 2        2   5091408
## 3        4  32778570
## 4        5  21379248
## 5        6 283834272
## 6        8  46015326
```

```r
WATR21<-Bigdata %>%
  filter(COSTWATR < 9992 & YEAR==2021)%>%
  select(STATEFIP,COSTWATR)%>%
  group_by(STATEFIP)%>%
 summarise(W.21 = sum(COSTWATR))
head(WATR21)
```

```
## # A tibble: 6 × 2
##   STATEFIP      W.21
##      <int>     <int>
## 1        1  22077800
## 2        2   2725424
## 3        4  44410994
## 4        5  14052744
## 5        6 268516166
## 6        8  33007572
```

```r
WATR22<-Bigdata %>%
  filter(COSTWATR < 9992 & YEAR==2022)%>%
  select(STATEFIP,COSTWATR)%>%
  group_by(STATEFIP)%>%
 summarise(W.22 = sum(COSTWATR))
head(WATR22)
```

```
## # A tibble: 6 × 2
##   STATEFIP      W.22
##      <int>     <int>
## 1        1  22584444
## 2        2   3373802
## 3        4  44986560
## 4        5  14434326
## 5        6 272466646
## 6        8  34253386
```

```r
# Merging the datasets for each year
total_cost_2021 <- cbind(ELE21,GAS21,WATR21) %>%
  select(-3,-5) %>%
  mutate(TOTAL_21 = (E.21+G.21+W.21 ))
  head(total_cost_2021)
```

```
##   STATEFIP      E.21      G.21      W.21   TOTAL_21
## 1        1 119766288  20684952  22077800  162529040
## 2        2  12607944   4924608   2725424   20257976
## 3        4 150361536  25782984  44410994  220555514
## 4        5  61909752  16792488  14052744   92754984
## 5        6 744658068 235149366 268516166 1248323600
## 6        8  96558618  34607424  33007572  164173614
```

```r
total_cost_2022 <- cbind(ELE22,GAS22,WATR22) %>%
  select(-3,-5) %>%
  mutate(TOTAL_22 = (E.22+G.22+W.22 ))
  head(total_cost_2022)
```

```
##   STATEFIP      E.22      G.22      W.22   TOTAL_22
## 1        1 133344858  25073178  22584444  181002480
## 2        2  14476350   5091408   3373802   22941560
## 3        4 158815002  32778570  44986560  236580132
## 4        5  71418264  21379248  14434326  107231838
## 5        6 858151356 283834272 272466646 1414452274
## 6        8 108302376  46015326  34253386  188571088
```

Question 5: Which state has the oldest, on average, residents?

Maine (state 23) has the oldest, on average, residents.

```r
Bigdata %>%
  group_by(STATEFIP) %>%
  summarise(AvgAge = mean(AGE, na.rm = TRUE)) %>%
  arrange(desc(AvgAge)) %>%
  head(1)
```

```
## # A tibble: 1 × 2
##   STATEFIP AvgAge
##      <int>  <dbl>
## 1       23   47.0
```

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.

```r
ohio_data <- Bigdata %>%
  filter(STATEFIP == 39 & YEAR == 2022)
```

Analyze Age for Ohio

In the year 2022, the average age in Ohio is 43.

```r
ohio_age_summary <- ohio_data %>%
  summarise(AverageAge = mean(AGE, na.rm = TRUE))
ohio_age_summary
```

```
##   AverageAge
## 1   43.20116
```

Analyze Sex for Ohio

In the year 2022,there are more female than male in Ohio.

```r
ohio_sex_distribution <- ohio_data %>%
  group_by(SEX) %>%
  summarise(Count = n()) %>%
  mutate(SexLabel = ifelse(SEX == 1, "Male", ifelse(SEX == 2, "Female", "Missing")))
ohio_sex_distribution
```

```
## # A tibble: 2 × 3
##     SEX Count SexLabel
##   <int> <int> <chr>   
## 1     1 58942 Male    
## 2     2 61724 Female
```

Analyze Race for Ohio

In the year 2022, it is observed that white people are the largest population group in Ohio, followed by Black/African American, with Japanese being the smallest.

```r
ohio_race_distribution <- ohio_data %>%
  group_by(RACE) %>%
  summarise(Count = n()) %>%
  mutate(RaceLabel = case_when(
    RACE == 1 ~ "White",
    RACE == 2 ~ "Black/African American",
    RACE == 3 ~ "American Indian or Alaska Native",
    RACE == 4 ~ "Chinese",
    RACE == 5 ~ "Japanese",
    RACE == 6 ~ "Other Asian or Pacific Islander",
    RACE == 7 ~ "Other race, nec",
    RACE == 8 ~ "Two major races",
    RACE == 9 ~ "Three or more major races",
    TRUE ~ "Unknown"
  ))

ohio_race_distribution
```

```
## # A tibble: 9 × 3
##    RACE Count RaceLabel                       
##   <int> <int> <chr>                           
## 1     1 98911 White                           
## 2     2  9981 Black/African American          
## 3     3   278 American Indian or Alaska Native
## 4     4   507 Chinese                         
## 5     5    82 Japanese                        
## 6     6  2069 Other Asian or Pacific Islander 
## 7     7  1598 Other race, nec                 
## 8     8  6721 Two major races                 
## 9     9   519 Three or more major races
```

Analyze Language for Ohio

In the year 2022, English was the most widely spoken language, while Yedish, Celtic, Aleut, Eskimo, and Iroquoian were the least spoken languages.

```r
ohio_language_distribution <- ohio_data %>%
  group_by(LANGUAGE) %>%
  summarise(Total = n()) %>%
  mutate(.after = LANGUAGE, LABEL = case_when(
    LANGUAGE == 00 ~ "N/A or blank",
    LANGUAGE == 01 ~ "English",
    LANGUAGE == 02 ~ "German",
    LANGUAGE == 03 ~ "Yiddish, Jewish",
    LANGUAGE == 04 ~ "Dutch",
    LANGUAGE == 05 ~ "Swedish",
    LANGUAGE == 06 ~ "Danish",
    LANGUAGE == 07 ~ "Norwegian",
    LANGUAGE == 08 ~ "Icelandic",
    LANGUAGE == 09 ~ "Scandinavian",
    LANGUAGE == "10" ~ "Italian",
    LANGUAGE == "11" ~ "French",
    LANGUAGE == "12" ~ "Spanish",
    LANGUAGE == "13" ~ "Portuguese",
    LANGUAGE == "14" ~ "Rumanian",
    LANGUAGE == "15" ~ "Celtic",
    LANGUAGE == "16" ~ "Greek",
    LANGUAGE == "17" ~ "Albanian",
    LANGUAGE == "18" ~ "Russian",
    LANGUAGE == "19" ~ "Ukrainian, Ruthenian, Little Russian",
    LANGUAGE == "20" ~ "Czech",
    LANGUAGE == "21" ~ "Polish",
    LANGUAGE == "22" ~ "Slovak",
    LANGUAGE == "23" ~ "Serbo-Croatian, Yugoslavian, Slavonian",
    LANGUAGE == "24" ~ "Slovene",
    LANGUAGE == "25" ~ "Lithuanian",
    LANGUAGE == "26" ~ "Other Balto-Slavic",
    LANGUAGE == "27" ~ "Slavic unknown",
    LANGUAGE == "28" ~ "Armenian",
    LANGUAGE == "29" ~ "Persian, Iranian, Farsi",
    LANGUAGE == "30" ~ "Other Persian dialects",
    LANGUAGE == "31" ~ "Hindi and related",
    LANGUAGE == "32" ~ "Romany, Gypsy",
    LANGUAGE == "33" ~ "Finnish",
    LANGUAGE == "34" ~ "Magyar, Hungarian",
    LANGUAGE == "35" ~ "Uralic",
    LANGUAGE == "36" ~ "Turkish",
    LANGUAGE == "37" ~ "Other Altaic",
    LANGUAGE == "38" ~ "Caucasian, Georgian, Avar",
    LANGUAGE == "39" ~ "Basque",
    LANGUAGE == "40" ~ "Dravidian",
    LANGUAGE == "41" ~ "Kurukh",
    LANGUAGE == "42" ~ "Burushaski",
    LANGUAGE == "43" ~ "Chinese",
    LANGUAGE == "44" ~ "Tibetan",
    LANGUAGE == "45" ~ "Burmese, Lisu, Lolo",
    LANGUAGE == "46" ~ "Kachin",
    LANGUAGE == "47" ~ "Thai, Siamese, Lao",
    LANGUAGE == "48" ~ "Japanese",
    LANGUAGE == "49" ~ "Korean",
    LANGUAGE == "50" ~ "Vietnamese",
    LANGUAGE == "51" ~ "Other East/Southeast Asian",
    LANGUAGE == "52" ~ "Indonesian",
    LANGUAGE == "53" ~ "Other Malayan",
    LANGUAGE == "54" ~ "Filipino, Tagalog",
    LANGUAGE == "55" ~ "Micronesian, Polynesian",
    LANGUAGE == "56" ~ "Hawaiian",
    LANGUAGE == "57" ~ "Arabic",
    LANGUAGE == "58" ~ "Near East Arabic dialect",
    LANGUAGE == "59" ~ "Hebrew, Israeli",
    LANGUAGE == "60" ~ "Amharic, Ethiopian, etc.",
    LANGUAGE == "61" ~ "Hamitic",
    LANGUAGE == "62" ~ "Other Afro-Asiatic languages",
    LANGUAGE == "63" ~ "Sub-Saharan Africa",
    LANGUAGE == "64" ~ "African, n.s.",
    LANGUAGE == "70" ~ "American Indian (all)",
    LANGUAGE == "71" ~ "Aleut, Eskimo",
    LANGUAGE == "72" ~ "Algonquian",
    LANGUAGE == "73" ~ "Salish, Flathead",
    LANGUAGE == "74" ~ "Athapascan",
    LANGUAGE == "75" ~ "Navajo",
    LANGUAGE == "76" ~ "Penutian-Sahaptin",
    LANGUAGE == "77" ~ "Other Penutian",
    LANGUAGE == "78" ~ "Zuni",
    LANGUAGE == "79" ~ "Yuman",
    LANGUAGE == "80" ~ "Other Hokan languages",
    LANGUAGE == "81" ~ "Siouan languages",
    LANGUAGE == "82" ~ "Muskogean",
    LANGUAGE == "83" ~ "Keres",
    LANGUAGE == "84" ~ "Iroquoian",
    LANGUAGE == "85" ~ "Caddoan",
    LANGUAGE == "86" ~ "Shoshonean/Hopi",
    LANGUAGE == "87" ~ "Pima, Papago",
    LANGUAGE == "88" ~ "Yaqui and other Sonoran, nec",
    LANGUAGE == "89" ~ "Aztecan, Nahuatl, Uto-Aztecan",
    LANGUAGE == "90" ~ "Tanoan languages",
    LANGUAGE == "91" ~ "Other Indian languages",
    LANGUAGE == "92" ~ "Mayan languages",
    LANGUAGE == "93" ~ "American Indian, n.s.",
    LANGUAGE == "94" ~ "Native",
    LANGUAGE == "95" ~ "No language",
    LANGUAGE == "96" ~ "Other or not reported",
    LANGUAGE == "99" ~ "Not reported, blank",
    TRUE ~ "Unknown"))

ohio_language_distribution
```

```
## # A tibble: 57 × 3
##    LANGUAGE LABEL            Total
##       <int> <chr>            <int>
##  1        0 N/A or blank      5919
##  2        1 English         107013
##  3        2 German            1116
##  4        3 Yiddish, Jewish      1
##  5        4 Dutch              173
##  6        5 Swedish             10
##  7        6 Danish               2
##  8        7 Norwegian            4
##  9       10 Italian            148
## 10       11 French             271
## # ℹ 47 more rows
```