Week 4

Kevin Chen

2024-09-20

```r
# Set a CRAN mirror (choose any mirror that works for you)
options(repos = c(CRAN = "https://cran.rstudio.com/"))

options(repos = c(CRAN = "https://cran.rstudio.com/"))

getwd()
```

```
## [1] "/Users/kevinsmac"
```

```r
# Load the CSV file
my_data <- read.csv("Dataset_123.csv")

# View the first few rows of the dataset
head(my_data)
```

```
##        Year StateAbbr StateDesc      CountyName CountyFIPS LocationName
## 1 2017-2021        AL   Alabama  Baldwin County       1003   1003010300
## 2 2017-2021        AL   Alabama  Baldwin County       1003   1003011201
## 3 2017-2021        AL   Alabama  Barbour County       1005   1005950400
## 4 2017-2021        AL   Alabama  Calhoun County       1015   1015002503
## 5 2017-2021        AL   Alabama Chambers County       1017   1017954700
## 6 2017-2021        AL   Alabama  Chilton County       1021   1021060104
##   DataSource Category                                      Measure
## 1 5-year ACS     SDOH           Housing cost burden among households
## 2 5-year ACS     SDOH Persons living below 150% of the poverty level
## 3 5-year ACS     SDOH           Housing cost burden among households
## 4 5-year ACS     SDOH                       Single-parent households
## 5 5-year ACS     SDOH Persons living below 150% of the poverty level
## 6 5-year ACS     SDOH           Housing cost burden among households
##   Data_Value_Unit Data_Value_Type Data_Value  MOE TotalPopulation
LocationID
## 1               %      Percentage       12.8  6.6            8863
1003010300
## 2               %      Percentage       13.9  6.4            4345
1003011201
## 3               %      Percentage       21.3  7.4            3859
1005950400
## 4               %      Percentage        0.8  1.7            3689
1015002503
## 5               %      Percentage       24.1  8.5            4586
1017954700
## 6               %      Percentage       23.6 11.6            3077
1021060104
```

```
##   CategoryID MeasureID DataValueTypeID      Short_Question_Text
## 1       SDOH     HCOST         Percent      Housing cost burden
## 2       SDOH    POV150         Percent                  Poverty
## 3       SDOH     HCOST         Percent      Housing cost burden
## 4       SDOH    SNGPNT         Percent Single-parent households
## 5       SDOH    POV150         Percent                  Poverty
## 6       SDOH     HCOST         Percent      Housing cost burden
##                      Geolocation
## 1 POINT (-87.8786378 30.8412001)
## 2  POINT (-87.891618 30.5408065)
## 3 POINT (-85.5577196 31.6794432)
## 4  POINT (-85.870404 33.7823094)
## 5 POINT (-85.1744682 32.7702191)
## 6 POINT (-86.6087644 32.8511448)
```

```
# Check the structure of the dataset
str(my_data)
```

```
## 'data.frame':    751509 obs. of  20 variables:
##  $ Year               : chr  "2017-2021" "2017-2021" "2017-2021" "2017-
2021" ...
##  $ StateAbbr          : chr  "AL" "AL" "AL" "AL" ...
##  $ StateDesc          : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ CountyName         : chr  "Baldwin County" "Baldwin County" "Barbour
County" "Calhoun County" ...
##  $ CountyFIPS         : int  1003 1003 1005 1015 1017 1021 1029 1039 1039
1039 ...
##  $ LocationName       : num  1.00e+09 1.00e+09 1.01e+09 1.02e+09 1.02e+09
...
##  $ DataSource         : chr  "5-year ACS" "5-year ACS" "5-year ACS" "5-
year ACS" ...
##  $ Category           : chr  "SDOH" "SDOH" "SDOH" "SDOH" ...
##  $ Measure            : chr  "Housing cost burden among households"
"Persons living below 150% of the poverty level" "Housing cost burden among
households" "Single-parent households" ...
##  $ Data_Value_Unit    : chr  "%" "%" "%" "%" ...
##  $ Data_Value_Type    : chr  "Percentage" "Percentage" "Percentage"
"Percentage" ...
##  $ Data_Value         : num  12.8 13.9 21.3 0.8 24.1 23.6 16.9 3.9 0.8
17.5 ...
##  $ MOE                : num  6.6 6.4 7.4 1.7 8.5 11.6 6.5 5.4 0.8 7 ...
##  $ TotalPopulation    : int  8863 4345 3859 3689 4586 3077 3018 1828 3338
1870 ...
##  $ LocationID         : num  1.00e+09 1.00e+09 1.01e+09 1.02e+09 1.02e+09
...
##  $ CategoryID         : chr  "SDOH" "SDOH" "SDOH" "SDOH" ...
##  $ MeasureID          : chr  "HCOST" "POV150" "HCOST" "SNGPNT" ...
##  $ DataValueTypeID    : chr  "Percent" "Percent" "Percent" "Percent" ...
##  $ Short_Question_Text: chr  "Housing cost burden" "Poverty" "Housing cost
burden" "Single-parent households" ...
```

```
##  $ Geolocation        : chr  "POINT (-87.8786378 30.8412001)" "POINT (-
87.891618 30.5408065)" "POINT (-85.5577196 31.6794432)" "POINT (-85.870404
33.7823094)" ...

# View summary statistics
summary(my_data)

##      Year            StateAbbr          StateDesc          CountyName
##  Length:751509      Length:751509      Length:751509      Length:751509
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    CountyFIPS     LocationName        DataSource          Category
##  Min.   : 1001   Min.   :1.001e+09   Length:751509      Length:751509
##  1st Qu.:12119   1st Qu.:1.212e+10   Class :character   Class :character
##  Median :27145   Median :2.715e+10   Mode  :character   Mode  :character
##  Mean   :27865   Mean   :2.786e+10
##  3rd Qu.:41067   3rd Qu.:4.107e+10
##  Max.   :56045   Max.   :5.605e+10
##
##    Measure          Data_Value_Unit    Data_Value_Type      Data_Value
##  Length:751509      Length:751509      Length:751509      Min.   :  0.00
##  Class :character   Class :character   Class :character   1st Qu.:  4.50
##  Mode  :character   Mode  :character   Mode  :character   Median : 11.10
##                                                           Mean   : 16.37
##                                                           3rd Qu.: 21.80
##                                                           Max.   :100.00
##                                                           NA's   :1237
##       MOE           TotalPopulation   LocationID         CategoryID
##  Min.   :   0.000   Min.   :   52   Min.   :1.001e+09   Length:751509
##  1st Qu.:   3.200   1st Qu.: 2737   1st Qu.:1.212e+10   Class :character
##  Median :   5.600   Median : 3768   Median :2.715e+10   Mode  :character
##  Mean   :   6.994   Mean   : 3949   Mean   :2.786e+10
##  3rd Qu.:   9.500   3rd Qu.: 4949   3rd Qu.:4.107e+10
##  Max.   :1200.000   Max.   :38223   Max.   :5.605e+10
##  NA's   :1477
##   MeasureID          DataValueTypeID    Short_Question_Text Geolocation
##  Length:751509      Length:751509      Length:751509       Length:751509
##  Class :character   Class :character   Class :character    Class
:character
##  Mode  :character   Mode  :character   Mode  :character    Mode
:character
##
##
##
##
```

```r
# Load necessary library
install.packages("dplyr")
```

```
##
## The downloaded binary packages are in
##
/var/folders/cr/d_vl212s1d9d6cqyv8m8ty0h0000gn/T//Rtmp6IkfA4/downloaded_packa
ges
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Filter the dataset for specific categories in Short_Question_Text and
Measure_ID
filtered_data <- my_data %>%
  filter(
    Short_Question_Text %in% c("Poverty", "Crowding") &
    MeasureID %in% c("POV150", "CROWD")
  )

# View the filtered data
head(filtered_data)
```

```
##        Year StateAbbr StateDesc       CountyName CountyFIPS LocationName
## 1 2017-2021        AL   Alabama   Baldwin County       1003   1003011201
## 2 2017-2021        AL   Alabama  Chambers County       1017   1017954700
## 3 2017-2021        AL   Alabama Covington County       1039   1039962100
## 4 2017-2021        AL   Alabama Covington County       1039   1039962300
## 5 2017-2021        AL   Alabama   Cullman County       1043   1043965501
## 6 2017-2021        AL   Alabama   Houston County       1069   1069040802
##   DataSource Category                                      Measure
## 1 5-year ACS     SDOH Persons living below 150% of the poverty level
## 2 5-year ACS     SDOH Persons living below 150% of the poverty level
## 3 5-year ACS     SDOH                   Crowding among housing units
## 4 5-year ACS     SDOH                   Crowding among housing units
## 5 5-year ACS     SDOH                   Crowding among housing units
## 6 5-year ACS     SDOH                   Crowding among housing units
##   Data_Value_Unit Data_Value_Type Data_Value MOE TotalPopulation
LocationID
## 1               %      Percentage       13.9 6.4            4345
1003011201
```

```
## 2               %      Percentage       24.1 8.5            4586
1017954700
## 3               %      Percentage        3.9 5.4            1828
1039962100
## 4               %      Percentage        0.8 0.8            3338
1039962300
## 5               %      Percentage        0.0 2.4            1693
1043965501
## 6               %      Percentage        1.2 1.5            3885
1069040802
##   CategoryID MeasureID DataValueTypeID Short_Question_Text
## 1       SDOH    POV150         Percent             Poverty
## 2       SDOH    POV150         Percent             Poverty
## 3       SDOH     CROWD         Percent            Crowding
## 4       SDOH     CROWD         Percent            Crowding
## 5       SDOH     CROWD         Percent            Crowding
## 6       SDOH     CROWD         Percent            Crowding
##                      Geolocation
## 1  POINT (-87.891618 30.5408065)
## 2 POINT (-85.1744682 32.7702191)
## 3   POINT (-86.507274 31.326445)
## 4 POINT (-86.3833989 31.2543283)
## 5 POINT (-86.9158949 33.9712693)
## 6 POINT (-85.4186788 31.1494473)
```

```r
# View the first few rows
head(filtered_data)
```

```
##        Year StateAbbr StateDesc       CountyName CountyFIPS LocationName
## 1 2017-2021        AL   Alabama   Baldwin County       1003   1003011201
## 2 2017-2021        AL   Alabama  Chambers County       1017   1017954700
## 3 2017-2021        AL   Alabama Covington County       1039   1039962100
## 4 2017-2021        AL   Alabama Covington County       1039   1039962300
## 5 2017-2021        AL   Alabama   Cullman County       1043   1043965501
## 6 2017-2021        AL   Alabama   Houston County       1069   1069040802
##   DataSource Category                                        Measure
## 1 5-year ACS     SDOH Persons living below 150% of the poverty level
## 2 5-year ACS     SDOH Persons living below 150% of the poverty level
## 3 5-year ACS     SDOH                   Crowding among housing units
## 4 5-year ACS     SDOH                   Crowding among housing units
## 5 5-year ACS     SDOH                   Crowding among housing units
## 6 5-year ACS     SDOH                   Crowding among housing units
##   Data_Value_Unit Data_Value_Type Data_Value MOE TotalPopulation
LocationID
## 1               %      Percentage       13.9 6.4            4345
1003011201
## 2               %      Percentage       24.1 8.5            4586
1017954700
## 3               %      Percentage        3.9 5.4            1828
1039962100
```

```
## 4               %      Percentage        0.8 0.8            3338
1039962300
## 5               %      Percentage        0.0 2.4            1693
1043965501
## 6               %      Percentage        1.2 1.5            3885
1069040802
##   CategoryID MeasureID DataValueTypeID Short_Question_Text
## 1       SDOH    POV150         Percent              Poverty
## 2       SDOH    POV150         Percent              Poverty
## 3       SDOH     CROWD         Percent             Crowding
## 4       SDOH     CROWD         Percent             Crowding
## 5       SDOH     CROWD         Percent             Crowding
## 6       SDOH     CROWD         Percent             Crowding
##                       Geolocation
## 1  POINT (-87.891618 30.5408065)
## 2 POINT (-85.1744682 32.7702191)
## 3   POINT (-86.507274 31.326445)
## 4 POINT (-86.3833989 31.2543283)
## 5 POINT (-86.9158949 33.9712693)
## 6 POINT (-85.4186788 31.1494473)

# Check unique entries in Short_Question_Text
unique(filtered_data$Short_Question_Text)

## [1] "Poverty"  "Crowding"

# Count occurrences of each category in Short_Question_Text
table(filtered_data$Short_Question_Text)

##
## Crowding  Poverty
##    83501    83501

table(filtered_data$MeasureID)

##
##  CROWD POV150
##  83501  83501

# Create a contingency table

contingency_table <- table(filtered_data$MeasureID,
filtered_data$Short_Question_Text)


# View the table
print(contingency_table)

##
##          Crowding Poverty
```

```
##   CROWD     83501       0
##   POV150        0   83501

# Total counts
total <- sum(contingency_table)

# Prevalence: Proportion of the outcome variable
prevalence <- sum(contingency_table[2, ]) / total

# Sensitivity
true_positives <- contingency_table[2, 2]  # TP
false_negatives <- contingency_table[2, 1]  # FN
sensitivity <- true_positives / (true_positives + false_negatives)

# Specificity
true_negatives <- contingency_table[1, 1]  # TN
false_positives <- contingency_table[1, 2]  # FP
specificity <- true_negatives / (true_negatives + false_positives)

# Print results
print(paste("Prevalence:", prevalence))

## [1] "Prevalence: 0.5"

print(paste("Sensitivity:", sensitivity))

## [1] "Sensitivity: 1"

print(paste("Specificity:", specificity))

## [1] "Specificity: 1"

# Calculate PPV and NPV
ppv <- true_positives / (true_positives + false_positives)
npv <- true_negatives / (true_negatives + false_negatives)

# Print PPV and NPV
print(paste("PPV (Positive Predictive Value):", ppv))

## [1] "PPV (Positive Predictive Value): 1"

print(paste("NPV (Negative Predictive Value):", npv))

## [1] "NPV (Negative Predictive Value): 1"
```