data_cleaning

Madison Thompson

2025-12-08

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE)
# imports
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
```

```
library(tidyr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.1     v readr     2.1.5
## v ggplot2   4.0.0     v stringr   1.5.2
## v lubridate 1.9.4     v tibble    3.3.0
## v purrr     1.1.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# get data
data <- read.csv("nc_county_health_data_2021.csv")

# remove irrelevant columns
data <- data %>%
  select(-c(Classification, Definition, Source.Label, Source.Link.s., X, Data.Notes)) %>%
  drop_na() %>%
  rename("NC" = X.1) %>%
  mutate(Metric = stringr::str_trim(Metric))
```

```r
# from metric, remove unwanted rows
data <- data %>%
  filter(Metric %in% c(
      "Metropolitan or Nonmetropolitan",
      "Population",
      "College Graduation",
      "Uninsured Adults",
      "Health Care Workforce - Primary Care Physicians",
      "Adult Smoking",
      "Poverty",
      "Food Insecurity",
      "Transportation",
      "Air Pollution"
    )
  )

# peek at data
data
```

```
##                                        Metric         NC Alamance
## 1                  Metropolitan or Nonmetropolitan              Metro
## 2                                  Population 10,488,084  169,509
## 3                             College Graduation      21.3%    22.9%
## 4                               Uninsured Adults      15.5%    16.8%
## 5  Health Care Workforce - Primary Care Physicians          8        6
## 6                                Adult Smoking      17.0%    17.0%
## 7                                      Poverty     13.60%    14.6%
## 8                               Food Insecurity     14.00%    14.0%
## 9                                Transportation       6.5%     5.5%
## 10                               Air Pollution        9.8     10.5
##    Alexander Alleghany    Anson     Ashe    Avery Beaufort   Bertie   Bladen
## 1      Metro  Nonmetro Nonmetro Nonmetro Nonmetro Nonmetro Nonmetro Nonmetro
## 2     37,497    11,137   24,446   27,203   17,557   46,994   18,947   32,722
## 3      13.9%     18.7%     9.2%    19.5%    20.2%    19.6%    12.3%    14.5%
## 4      15.2%     20.7%    16.6%    18.2%    21.1%    15.7%    15.2%    21.7%
## 5        2.3       4.4      1.2      4.7        5      3.8      3.6      3.5
## 6      17.0%     17.0%    20.0%    17.0%    17.0%    18.0%    20.0%    19.0%
## 7      11.7%     16.9%    21.4%    14.6%    16.4%    17.6%    24.2%    21.2%
## 8      14.0%     18.0%    16.6%    14.4%    14.5%    16.2%    18.4%    19.1%
## 9       3.9%      4.3%    10.3%     6.0%     4.8%     8.0%     8.6%     8.9%
## 10      10.1       8.7     10.2      8.5      8.5      9.2      9.1       10
##    Brunswick Buncombe  Burke Cabarrus Caldwell   Camden Carteret  Caswell
## 1      Metro    Metro  Metro    Metro    Metro Nonmetro Nonmetro Nonmetro
## 2    142,820  261,191 90,485  216,453   82,178   10,867   69,473   22,604
## 3      28.0%    38.5%  16.7%    30.0%    14.8%    18.1%    26.7%    13.7%
## 4      15.7%    14.8%  16.3%    13.4%    17.8%    14.1%    15.6%    15.6%
## 5        4.2       13    8.1      7.9      4.9      0.9      5.1        3
## 6      15.0%    17.0%  17.0%    16.0%    17.0%    16.0%    16.0%    18.0%
## 7      10.2%    12.2%  18.4%     7.9%    12.0%     7.6%    10.4%    16.2%
## 8      13.7%    12.4%  15.5%    11.1%    14.9%    11.7%    13.8%    15.9%
## 9       4.0%     5.2%   5.8%     4.5%     4.6%     2.2%     4.8%       10
## 10       9.3        9   10.1       11      9.9      8.4      8.7     10.2
##    Catawba Chatham Cherokee   Chowan     Clay Cleveland Columbus  Craven
## 1    Metro   Metro Nonmetro Nonmetro Nonmetro  Nonmetro Nonmetro   Metro
```

```
## 2  159,551  74,470   28,612   13,943   11,231    97,947   55,508 102,139
## 3    21.5%   40.1%    19.2%    19.0%    23.1%     16.5%    12.5%   24.3%
## 4    16.1%   15.9%    17.9%    14.3%    19.2%     16.1%    19.4%   14.0%
## 5      6.8       3      2.4      7.8      3.4       6.2      4.4     6.8
## 6    16.0%   14.0%    17.0%    18.0%    16.0%     18.0%    19.0%   16.0%
## 7    13.3%    8.7%    17.7%    18.5%    14.0%     19.0%    22.3%   13.8%
## 8    12.8%   11.9%    15.3%    15.9%    14.9%     16.2%    18.1%   15.0%
## 9     5.0%    4.7%     5.3%     9.4%     5.7%      6.9%     7.5%    7.0%
## 10    10.5    10.3        9      8.6      8.5      10.6      9.7     9.4
##    Cumberland Currituck     Dare Davidson  Davie  Duplin  Durham Edgecombe
## 1       Metro     Metro Nonmetro    Metro  Metro Nometro   Metro      Metro
## 2     335,509    27,763   37,009  167,609 42,846  58,741 321,488     51,472
## 3       24.3%     22.9%    29.8%    18.1%  23.0%   10.8%   47.3%      11.6%
## 4       13.8%     14.6%    17.4%    17.2%  15.5%   25.7%   15.1%      16.0%
## 5         5.8       2.2      6.7      3.6    3.5     2.2    12.6        5.1
## 6       18.0%     15.0%    16.0%    18.0%  15.0%   21.0%   15.0%      21.0%
## 7       18.0%      8.8%     8.9%    15.2%  10.9%   17.7%   14.0%      21.0%
## 8       16.9%     11.4%    11.8%    14.1%  12.6%   17.1%   13.5%      19.6%
## 9        6.5%      3.8%     3.6%     5.7%   3.9%    7.6%    8.2%      11.6%
## 10       10.6       8.5      8.2     10.8   10.4     9.9    10.6       10.2
##    Forsyth Franklin  Gaston  Gates   Graham Granville   Greene Guilford
## 1    Metro     Metro   Metro  Metro Nonmetro  Nonmetro Nonmetro    Metro
## 2  382,295   69,685 224,529 11,562    8,441    60,443   21,069  537,174
## 3    33.8%    21.0%   20.5%  15.7%    14.2%     21.0%    9.8%    34.9%
## 4    16.1%    16.8%   15.2%  12.7%    22.2%     15.0%   23.8%    14.0%
## 5     11.4      0.7     5.8      0      4.6         6      4.8      7.8
## 6    18.0%    18.0%   17.0%  16.0%    18.0%     17.0%   20.0%    16.0%
## 7    15.2%    11.6%   11.6%  14.7%    16.8%     14.6%   20.2%    16.0%
## 8    13.8%    13.7%   14.4%  14.2%    16.5%     13.4%   18.5%    13.9%
## 9     7.7%     6.0%    6.2%   4.7%     6.2%      5.2%    7.7%     6.6%
## 10     9.4     10.6    10.9    8.8      8.4      10.5    10.4     10.5
##     Halifax  Harnett Haywood Henderson Hertford   Hoke     Hyde Iredell
## 1  Nonmetro Nonmetro   Metro      Metro Nonmetro  Metro Nonmetro   Metro
## 2    50,010  135,976  62,317   117,417   23,677 55,234    4,937 181,806
## 3     14.0%    20.3%   24.3%     31.2%    15.3%  18.3%     7.2%   26.9%
## 4     16.4%    17.9%   14.6%     17.5%    15.1%  19.9%    20.4%   14.5%
## 5       4.5      4.3     7.1       7.5      9.6    2.9      1.9     6.9
## 6     20.0%    19.0%   18.0%     14.0%    19.0%  20.0%    17.0%   15.0%
## 7     23.8%    15.6%   10.6%     10.6%    23.0%  16.9%    19.2%    8.2%
## 8     19.9%    14.5%   13.6%     11.8%    18.9%  16.7%    18.1%   12.1%
## 9     12.6%     5.4%    4.8%      4.8%     8.8%   5.2%     3.2%    3.7%
## 10      9.8     10.9     8.7       9.2      9.1   10.4      7.8     10.9
##     Jackson Johnston Jones      Lee   Lenoir Lincoln    Macon Madison   Martin
## 1  Nonmetro    Metro Metro Nonmetro Nonmetro   Metro Nonmetro   Metro Nonmetro
## 2    43,938  209,339 9,419   61,779   55,949  86,111   35,858  21,755   22,440
## 3     30.5%    21.7% 14.2%    21.2%    13.5%   20.2%    22.2%   25.9%    15.8%
## 4     20.7%    16.8% 18.6%    20.1%    17.7%   14.9%    21.4%   14.7%    14.3%
## 5       6.3        3   4.9      6.2      5.5       5      7.1     4.9      5.6
## 6     18.0%    18.0% 19.0%    17.0%    19.0%   16.0%    16.0%   17.0%    19.0%
## 7     19.3%    12.5% 18.8%    14.2%    23.1%    9.0%    14.3%   14.6%    20.6%
## 8     14.4%    12.6% 19.5%    14.2%    18.9%   12.9%    14.5%   14.7%    16.7%
## 9      4.7%     4.4%  6.6%     6.2%    11.5%    3.0%     5.9%    3.8%     6.4%
## 10      8.4     10.4   9.5     10.5     10.3    10.8      8.6     8.8      8.9
##    McDowell Mecklenburg Mitchell Montgomery   Moore   Nash New.Hanover
```

```
## 1  Nonmetro       Metro Nonmetro   Nonmetro Nonmetro  Metro       Metro
## 2    45,756   1,110,356   14,964     27,173  100,880 94,298     234,473
## 3     15.9%       44.1%    18.1%      14.0%    36.0%  20.3%       38.9%
## 4     17.0%       15.8%    17.1%      18.9%    14.2%  14.7%       13.8%
## 5       5.1         9.5      8.5        2.5        7    5.9         8.1
## 6     19.0%       14.0%    16.0%      18.0%    14.0%  18.0%       16.0%
## 7     13.6%       10.3%    14.8%      16.1%    11.3%  16.4%       13.0%
## 8     15.5%       12.0%    14.5%      14.5%    12.8%  15.4%       14.1%
## 9      5.8%        6.2%     6.0%       8.8%     4.6%   8.2%        6.3%
## 10      9.2        11.3      8.5        9.7     10.4   10.5         8.4
##    Northampton  Onslow  Orange Pamlico Pasquotank Pender Perquimans Person
## 1     Nonmetro   Metro   Metro   Metro   Nonmetro  Metro   Nonmetro  Metro
## 2       19,483 197,938 148,476  12,726     39,824 63,060     13,463 39,490
## 3        12.8%   20.2%   57.6%   19.4%      20.4%  25.6%      19.7%  15.3%
## 4        14.8%   12.7%   12.2%   15.2%      14.2%  15.7%      15.8%  14.1%
## 5          1.5       3    22.1     3.8        7.6    2.7        3.7    3.2
## 6        19.0%   18.0%   14.0%   16.0%      18.0%  16.0%      17.0%  18.0%
## 7        21.6%   12.5%   13.4%   15.9%      14.3%  11.5%      15.0%  15.4%
## 8        18.5%   15.2%   10.8%   15.1%      15.3%  13.6%      14.6%  15.5%
## 9        11.1%    5.0%    4.8%    6.0%      10.9%   5.0%       8.4%   8.8%
## 10         9.6     9.5    10.7     8.6        8.5    9.4        8.5   10.1
##       Pitt     Polk Randolph Richmond  Robeson Rockingham   Rowan Rutherford
## 1    Metro Nonmetro    Metro Nonmetro Nonmetro      Metro   Metro   Nonmetro
## 2  180,742   20,724  143,667   44,829  130,625     91,010 142,088     67,029
## 3    30.9%    31.9%    15.3%    14.3%    12.8%      14.7%  18.4%      17.3%
## 4    14.4%    16.4%    19.7%    17.1%    23.8%      16.3%  16.8%      17.1%
## 5     11.1      3.7      3.5      3.8      4.4        4.9    4.4        4.2
## 6    17.0%    14.0%    18.0%    21.0%    24.0%      18.0%  17.0%      19.0%
## 7    19.2%    12.1%    14.1%    25.8%    31.5%      18.4%  13.9%      18.5%
## 8    16.1%    12.6%    13.8%    18.4%    19.4%      15.2%  14.3%      16.4%
## 9     7.8%     6.1%     5.5%     9.5%    10.3%       7.5%   6.2%       6.5%
## 10     9.8      9.5     10.9     10.2     10.1       10.3   10.6       10.1
##     Sampson Scotland   Stanly Stokes    Surry    Swain Transylvania  Tyrrell
## 1  Nonmetro Nonmetro Nonmetro  Metro Nonmetro Nonmetro     Nonmetro Nonmetro
## 2    63,531   34,823   62,806 45,591   71,783   14,271       34,385    4,016
## 3     12.5%    15.9%    16.5%  13.9%    16.5%    15.0%        29.9%     7.6%
## 4     24.0%    18.7%    15.8%  14.6%    18.4%    23.5%        19.8%    21.0%
## 5       4.7      6.2      4.1    3.7      5.5     10.5          7.6        0
## 6     19.0%    23.0%    17.0%  16.0%    17.0%    22.0%        16.0%    21.0%
## 7     16.8%    28.5%    10.7%  13.0%    16.0%    16.1%        13.1%    25.4%
## 8     16.9%    20.9%    13.4%  13.0%    15.0%    15.5%        13.6%    20.3%
## 9      8.0%    11.7%     5.8%   4.1%     6.9%     4.3%         4.4%     8.5%
## 10     10.7     10.2     10.5    9.8      9.6      8.4          8.7      7.9
##      Union    Vance      Wake   Warren Washington  Watauga   Wayne   Wilkes
## 1    Metro Nonmetro     Metro Nonmetro   Nonmetro Nonmetro   Metro Nonmetro
## 2  239,859   44,535 1,111,761   19,731     11,580   56,177 123,131   68,412
## 3    34.0%    12.1%     51.0%    15.8%       9.1%    41.7%   19.7%    15.4%
## 4    15.3%    16.5%     11.8%    19.7%      15.5%    14.4%   18.9%    18.8%
## 5      4.5      5.9       8.5      0.5        2.5      5.9     5.6        5
## 6    15.0%    20.0%     12.0%    19.0%      19.0%    18.0%   18.0%    18.0%
## 7     7.3%    18.5%      8.0%    21.7%      21.3%    21.4%   18.6%    15.2%
## 8     9.5%    19.0%     10.4%    17.6%      19.5%    13.8%   16.6%    15.8%
## 9     2.3%    12.7%      4.1%     8.7%      12.0%     5.6%    7.8%     6.2%
## 10    10.9       10        11      9.7        8.6      8.3    10.7      9.6
```

```
##       Wilson Yadkin   Yancey Data.year
## 1  Nonmetro  Metro Nonmetro      2013
## 2    81,801 37,667   18,069      2019
## 3     18.8%  12.0%    19.7% 2013-2017
## 4     18.3%  17.4%    17.9%      2018
## 5       6.1    2.9      4.3      2019
## 6     18.0%  18.0%    17.0%      2017
## 7     21.5%  13.9%    14.2%      2019
## 8     17.8%  13.4%    15.4%      2018
## 9      8.7%   6.2%     6.4%      2017
## 10     10.7    9.9      8.6      2014
```

```r
# print(colnames(data))

# save data to cleaned file
write.csv(data, "nc_county_health_data_2021_cleaned.csv", row.names = FALSE)
```

Using the updated data values, convert to z-scores

```r
# read in cleaned data
data <- read.csv("nc_county_health_data_2021_cleaned.csv")

# function to convert to z-scores
convert_to_zscore <- function(x) {
  (x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)
}

# pivot data to long format
data_long <- data %>%
  pivot_longer(
    cols = Alamance:Yancey,
    names_to = "County",
    values_to = "Value_Raw"
  ) %>%
  select(!Data.year)

# remove categorical data to avoid errors when scoring
df_categorical <- data_long %>%
  filter(Metric == "Metropolitan or Nonmetropolitan") %>%
  mutate(Z_Score = NA) %>%
  rename(Value = Value_Raw) %>%
  select(Metric, County, Value)

# numeric data cleaning
df_numeric_raw <- data_long %>%
  filter(Metric != "Metropolitan or Nonmetropolitan") %>%
  mutate(
    Value_Clean = gsub(",", "", Value_Raw),
    Value_Clean = gsub("%", "", Value_Clean),
    Value_Clean = as.numeric(Value_Clean)
  ) %>%
  filter(!is.na(Value_Clean))
```

```r
# calculate z-scores for numeric data
df_numeric_zscore <- df_numeric_raw %>%
  group_by(Metric) %>%
  mutate(Z_Score = convert_to_zscore(Value_Clean)) %>%
  ungroup() %>%
  select(Metric, County, Value = Value_Clean, Z_Score) %>%
  rename(Value_Clean = Value)

zscore_data <- bind_rows(df_numeric_zscore, df_categorical)

# view data
data
```

```
##                                        Metric         NC Alamance
## 1                  Metropolitan or Nonmetropolitan              Metro
## 2                                  Population 10,488,084  169,509
## 3                               College Graduation      21.3%    22.9%
## 4                                 Uninsured Adults      15.5%    16.8%
## 5  Health Care Workforce - Primary Care Physicians          8        6
## 6                                  Adult Smoking      17.0%    17.0%
## 7                                        Poverty     13.60%    14.6%
## 8                                 Food Insecurity     14.00%    14.0%
## 9                                  Transportation       6.5%     5.5%
## 10                                 Air Pollution        9.8     10.5
##    Alexander Alleghany    Anson     Ashe    Avery Beaufort   Bertie   Bladen
## 1      Metro  Nonmetro Nonmetro Nonmetro Nonmetro Nonmetro Nonmetro Nonmetro
## 2     37,497    11,137   24,446   27,203   17,557   46,994   18,947   32,722
## 3      13.9%     18.7%     9.2%    19.5%    20.2%    19.6%    12.3%    14.5%
## 4      15.2%     20.7%    16.6%    18.2%    21.1%    15.7%    15.2%    21.7%
## 5        2.3       4.4      1.2      4.7        5      3.8      3.6      3.5
## 6      17.0%     17.0%    20.0%    17.0%    17.0%    18.0%    20.0%    19.0%
## 7      11.7%     16.9%    21.4%    14.6%    16.4%    17.6%    24.2%    21.2%
## 8      14.0%     18.0%    16.6%    14.4%    14.5%    16.2%    18.4%    19.1%
## 9       3.9%      4.3%    10.3%     6.0%     4.8%     8.0%     8.6%     8.9%
## 10      10.1       8.7     10.2      8.5      8.5      9.2      9.1       10
##    Brunswick Buncombe  Burke Cabarrus Caldwell   Camden Carteret  Caswell
## 1      Metro    Metro  Metro    Metro    Metro Nonmetro Nonmetro Nonmetro
## 2    142,820  261,191 90,485  216,453   82,178   10,867   69,473   22,604
## 3      28.0%    38.5%  16.7%    30.0%    14.8%    18.1%    26.7%    13.7%
## 4      15.7%    14.8%  16.3%    13.4%    17.8%    14.1%    15.6%    15.6%
## 5        4.2       13    8.1      7.9      4.9      0.9      5.1        3
## 6      15.0%    17.0%  17.0%    16.0%    17.0%    16.0%    16.0%    18.0%
## 7      10.2%    12.2%  18.4%     7.9%    12.0%     7.6%    10.4%    16.2%
## 8      13.7%    12.4%  15.5%    11.1%    14.9%    11.7%    13.8%    15.9%
## 9       4.0%     5.2%   5.8%     4.5%     4.6%     2.2%     4.8%       10
## 10       9.3        9   10.1       11      9.9      8.4      8.7     10.2
##    Catawba Chatham Cherokee   Chowan     Clay Cleveland Columbus  Craven
## 1    Metro   Metro Nonmetro Nonmetro Nonmetro  Nonmetro Nonmetro   Metro
## 2  159,551  74,470   28,612   13,943   11,231    97,947   55,508 102,139
## 3    21.5%   40.1%    19.2%    19.0%    23.1%     16.5%    12.5%   24.3%
## 4    16.1%   15.9%    17.9%    14.3%    19.2%     16.1%    19.4%   14.0%
## 5      6.8       3      2.4      7.8      3.4       6.2      4.4     6.8
## 6    16.0%   14.0%    17.0%    18.0%    16.0%     18.0%    19.0%   16.0%
## 7    13.3%    8.7%    17.7%    18.5%    14.0%     19.0%    22.3%   13.8%
```

```
## 8    12.8%   11.9%    15.3%    15.9%    14.9%     16.2%    18.1%   15.0%
## 9     5.0%    4.7%     5.3%     9.4%     5.7%      6.9%     7.5%    7.0%
## 10    10.5    10.3        9      8.6      8.5      10.6      9.7     9.4
##    Cumberland Currituck     Dare Davidson  Davie  Duplin  Durham Edgecombe
## 1       Metro     Metro Nonmetro    Metro  Metro Nometro   Metro      Metro
## 2     335,509    27,763   37,009  167,609 42,846  58,741 321,488     51,472
## 3       24.3%     22.9%    29.8%    18.1%  23.0%   10.8%   47.3%      11.6%
## 4       13.8%     14.6%    17.4%    17.2%  15.5%   25.7%   15.1%      16.0%
## 5         5.8       2.2      6.7      3.6    3.5     2.2    12.6        5.1
## 6       18.0%     15.0%    16.0%    18.0%  15.0%   21.0%   15.0%      21.0%
## 7       18.0%      8.8%     8.9%    15.2%  10.9%   17.7%   14.0%      21.0%
## 8       16.9%     11.4%    11.8%    14.1%  12.6%   17.1%   13.5%      19.6%
## 9        6.5%      3.8%     3.6%     5.7%   3.9%    7.6%    8.2%      11.6%
## 10       10.6       8.5      8.2     10.8   10.4     9.9    10.6       10.2
##    Forsyth Franklin  Gaston  Gates   Graham Granville   Greene Guilford
## 1    Metro    Metro   Metro  Metro Nonmetro  Nonmetro Nonmetro    Metro
## 2  382,295   69,685 224,529 11,562    8,441    60,443   21,069  537,174
## 3    33.8%    21.0%   20.5%  15.7%    14.2%     21.0%     9.8%    34.9%
## 4    16.1%    16.8%   15.2%  12.7%    22.2%     15.0%    23.8%    14.0%
## 5     11.4      0.7     5.8      0      4.6         6      4.8      7.8
## 6    18.0%    18.0%   17.0%  16.0%    18.0%     17.0%    20.0%    16.0%
## 7    15.2%    11.6%   11.6%  14.7%    16.8%     14.6%    20.2%    16.0%
## 8    13.8%    13.7%   14.4%  14.2%    16.5%     13.4%    18.5%    13.9%
## 9     7.7%     6.0%    6.2%   4.7%     6.2%      5.2%     7.7%     6.6%
## 10     9.4     10.6    10.9    8.8      8.4      10.5     10.4     10.5
##     Halifax  Harnett Haywood Henderson Hertford   Hoke     Hyde Iredell
## 1  Nonmetro Nonmetro   Metro     Metro Nonmetro  Metro Nonmetro   Metro
## 2    50,010  135,976  62,317   117,417   23,677 55,234     4,937 181,806
## 3     14.0%    20.3%   24.3%     31.2%    15.3%  18.3%     7.2%   26.9%
## 4     16.4%    17.9%   14.6%     17.5%    15.1%  19.9%    20.4%   14.5%
## 5       4.5      4.3     7.1       7.5      9.6    2.9      1.9     6.9
## 6     20.0%    19.0%   18.0%     14.0%    19.0%  20.0%    17.0%   15.0%
## 7     23.8%    15.6%   10.6%     10.6%    23.0%  16.9%    19.2%    8.2%
## 8     19.9%    14.5%   13.6%     11.8%    18.9%  16.7%    18.1%   12.1%
## 9     12.6%     5.4%    4.8%      4.8%     8.8%   5.2%     3.2%    3.7%
## 10      9.8     10.9     8.7       9.2      9.1   10.4      7.8     10.9
##     Jackson Johnston Jones      Lee   Lenoir Lincoln    Macon Madison   Martin
## 1  Nonmetro    Metro Metro Nonmetro Nonmetro   Metro Nonmetro   Metro Nonmetro
## 2    43,938  209,339 9,419   61,779   55,949  86,111   35,858  21,755   22,440
## 3     30.5%    21.7% 14.2%    21.2%    13.5%   20.2%    22.2%   25.9%    15.8%
## 4     20.7%    16.8% 18.6%    20.1%    17.7%   14.9%    21.4%   14.7%    14.3%
## 5       6.3        3   4.9      6.2      5.5       5      7.1     4.9      5.6
## 6     18.0%    18.0% 19.0%    17.0%    19.0%   16.0%    16.0%   17.0%    19.0%
## 7     19.3%    12.5% 18.8%    14.2%    23.1%    9.0%    14.3%   14.6%    20.6%
## 8     14.4%    12.6% 19.5%    14.2%    18.9%   12.9%    14.5%   14.7%    16.7%
## 9      4.7%     4.4%  6.6%     6.2%    11.5%    3.0%     5.9%    3.8%     6.4%
## 10      8.4     10.4   9.5     10.5     10.3     10.8      8.6     8.8      8.9
##    McDowell Mecklenburg Mitchell Montgomery    Moore   Nash New.Hanover
## 1  Nonmetro       Metro Nonmetro   Nonmetro Nonmetro  Metro       Metro
## 2    45,756   1,110,356   14,964     27,173  100,880 94,298     234,473
## 3     15.9%       44.1%    18.1%      14.0%    36.0%  20.3%       38.9%
## 4     17.0%       15.8%    17.1%      18.9%    14.2%  14.7%       13.8%
## 5       5.1         9.5      8.5        2.5        7    5.9         8.1
## 6     19.0%       14.0%    16.0%      18.0%    14.0%  18.0%       16.0%
```

```
## 7     13.6%       10.3%    14.8%      16.1%    11.3%  16.4%       13.0%
## 8     15.5%       12.0%    14.5%      14.5%    12.8%  15.4%       14.1%
## 9      5.8%        6.2%     6.0%       8.8%     4.6%   8.2%        6.3%
## 10      9.2        11.3      8.5        9.7     10.4   10.5         8.4
##    Northampton  Onslow  Orange Pamlico Pasquotank Pender Perquimans Person
## 1     Nonmetro   Metro   Metro   Metro   Nonmetro  Metro   Nonmetro  Metro
## 2       19,483 197,938 148,476  12,726     39,824 63,060     13,463 39,490
## 3        12.8%   20.2%   57.6%   19.4%      20.4%  25.6%      19.7%  15.3%
## 4        14.8%   12.7%   12.2%   15.2%      14.2%  15.7%      15.8%  14.1%
## 5          1.5       3    22.1     3.8        7.6    2.7        3.7    3.2
## 6        19.0%   18.0%   14.0%   16.0%      18.0%  16.0%      17.0%  18.0%
## 7        21.6%   12.5%   13.4%   15.9%      14.3%  11.5%      15.0%  15.4%
## 8        18.5%   15.2%   10.8%   15.1%      15.3%  13.6%      14.6%  15.5%
## 9        11.1%    5.0%    4.8%    6.0%      10.9%   5.0%       8.4%   8.8%
## 10         9.6     9.5    10.7     8.6        8.5    9.4        8.5   10.1
##       Pitt     Polk Randolph Richmond  Robeson Rockingham   Rowan Rutherford
## 1    Metro Nonmetro    Metro Nonmetro Nonmetro      Metro   Metro   Nonmetro
## 2  180,742   20,724  143,667   44,829  130,625     91,010 142,088     67,029
## 3    30.9%    31.9%    15.3%    14.3%    12.8%      14.7%  18.4%      17.3%
## 4    14.4%    16.4%    19.7%    17.1%    23.8%      16.3%  16.8%      17.1%
## 5     11.1      3.7      3.5      3.8      4.4        4.9    4.4        4.2
## 6    17.0%    14.0%    18.0%    21.0%    24.0%      18.0%  17.0%      19.0%
## 7    19.2%    12.1%    14.1%    25.8%    31.5%      18.4%  13.9%      18.5%
## 8    16.1%    12.6%    13.8%    18.4%    19.4%      15.2%  14.3%      16.4%
## 9     7.8%     6.1%     5.5%     9.5%    10.3%       7.5%   6.2%       6.5%
## 10     9.8      9.5     10.9     10.2     10.1       10.3   10.6       10.1
##     Sampson Scotland   Stanly Stokes    Surry    Swain Transylvania  Tyrrell
## 1  Nonmetro Nonmetro Nonmetro  Metro Nonmetro Nonmetro     Nonmetro Nonmetro
## 2    63,531   34,823   62,806 45,591   71,783   14,271       34,385    4,016
## 3     12.5%    15.9%    16.5%  13.9%    16.5%    15.0%        29.9%     7.6%
## 4     24.0%    18.7%    15.8%  14.6%    18.4%    23.5%        19.8%    21.0%
## 5       4.7      6.2      4.1    3.7      5.5     10.5          7.6        0
## 6     19.0%    23.0%    17.0%  16.0%    17.0%    22.0%        16.0%    21.0%
## 7     16.8%    28.5%    10.7%  13.0%    16.0%    16.1%        13.1%    25.4%
## 8     16.9%    20.9%    13.4%  13.0%    15.0%    15.5%        13.6%    20.3%
## 9      8.0%    11.7%     5.8%   4.1%     6.9%     4.3%         4.4%     8.5%
## 10     10.7     10.2     10.5    9.8      9.6      8.4          8.7      7.9
##      Union    Vance      Wake   Warren Washington  Watauga   Wayne   Wilkes
## 1    Metro Nonmetro     Metro Nonmetro   Nonmetro Nonmetro   Metro Nonmetro
## 2  239,859   44,535 1,111,761   19,731     11,580   56,177 123,131   68,412
## 3    34.0%    12.1%     51.0%    15.8%      9.1%    41.7%   19.7%    15.4%
## 4    15.3%    16.5%     11.8%    19.7%     15.5%    14.4%   18.9%    18.8%
## 5      4.5      5.9       8.5      0.5       2.5      5.9     5.6        5
## 6    15.0%    20.0%     12.0%    19.0%     19.0%    18.0%   18.0%    18.0%
## 7     7.3%    18.5%      8.0%    21.7%     21.3%    21.4%   18.6%    15.2%
## 8     9.5%    19.0%     10.4%    17.6%     19.5%    13.8%   16.6%    15.8%
## 9     2.3%    12.7%      4.1%     8.7%     12.0%     5.6%    7.8%     6.2%
## 10    10.9       10        11      9.7       8.6      8.3    10.7      9.6
##      Wilson Yadkin   Yancey Data.year
## 1  Nonmetro  Metro Nonmetro      2013
## 2    81,801 37,667   18,069      2019
## 3     18.8%  12.0%    19.7% 2013-2017
## 4     18.3%  17.4%    17.9%      2018
## 5       6.1    2.9      4.3      2019
```

```
## 6     18.0%  18.0%    17.0%      2017
## 7     21.5%  13.9%    14.2%      2019
## 8     17.8%  13.4%    15.4%      2018
## 9      8.7%   6.2%     6.4%      2017
## 10     10.7    9.9      8.6      2014
```

```
zscore_data
```

```
## # A tibble: 1,000 x 5
##    Metric     County    Value_Clean Z_Score Value
##    <chr>      <chr>           <dbl>   <dbl> <chr>
##  1 Population Alamance       169509   0.380 <NA>
##  2 Population Alexander       37497  -0.396 <NA>
##  3 Population Alleghany       11137  -0.551 <NA>
##  4 Population Anson           24446  -0.473 <NA>
##  5 Population Ashe            27203  -0.457 <NA>
##  6 Population Avery           17557  -0.513 <NA>
##  7 Population Beaufort        46994  -0.340 <NA>
##  8 Population Bertie          18947  -0.505 <NA>
##  9 Population Bladen          32722  -0.424 <NA>
## 10 Population Brunswick      142820   0.223 <NA>
## # i 990 more rows
```

```
# saving z-score data
write.csv(zscore_data, "nc_county_health_data_2021_zscores.csv", row.names = FALSE)

summary(zscore_data)
```

```
##     Metric             County           Value_Clean           Z_Score
##  Length:1000        Length:1000        Min.   :      0.0   Min.   :-2.7396
##  Class :character   Class :character   1st Qu.:      9.2   1st Qu.:-0.6249
##  Mode  :character   Mode  :character   Median :     14.7   Median :-0.1929
##                                        Mean   :  11665.4   Mean   : 0.0000
##                                        3rd Qu.:     19.0   3rd Qu.: 0.5663
##                                        Max.   :1111761.0   Max.   : 5.9209
##                                        NA's   :100         NA's   :100
##     Value
##  Length:1000
##  Class :character
##  Mode  :character
##
##
##
##
```

Getting summary statistics for each metric, and showing any outliers

```
# imports
library(ggplot2)
library(ggrepel)

# getting summary stats for quantitative variables
```

```r
summary_stats <- zscore_data %>%
  filter(!Metric == "Metropolitan or Nonmetropolitan") %>%
  group_by(Metric) %>%
  summarise(
    Mean = mean(Z_Score, na.rm = TRUE),
    Median = median(Z_Score, na.rm = TRUE),
    SD = sd(Z_Score, na.rm = TRUE),
    Min = min(Z_Score, na.rm = TRUE),
    Max = max(Z_Score, na.rm = TRUE)
  )

# list all low outliers
outliers_low <- zscore_data %>%
  filter(!Metric == "Metropolitan or Nonmetropolitan") %>%
  group_by(Metric) %>%
  filter(Z_Score < (quantile(Z_Score, 0.25) - 1.5 * IQR(Z_Score))) %>%
  select(Metric, County, Z_Score)

# list all high outliers
outliers_high <- zscore_data %>%
  filter(!Metric == "Metropolitan or Nonmetropolitan") %>%
  group_by(Metric) %>%
  filter(Z_Score > (quantile(Z_Score, 0.75) + 1.5 * IQR(Z_Score))) %>%
  select(Metric, County, Z_Score)

# join summary stats and outliers
summary_stats <- summary_stats %>%
  left_join(
    outliers_low %>%
      group_by(Metric) %>%
      summarise(Low_Outliers = paste(County, collapse = ", ")),
    by = "Metric"
  ) %>%
  left_join(
    outliers_high %>%
      group_by(Metric) %>%
      summarise(High_Outliers = paste(County, collapse = ", ")),
    by = "Metric"
  )

summary_stats
```

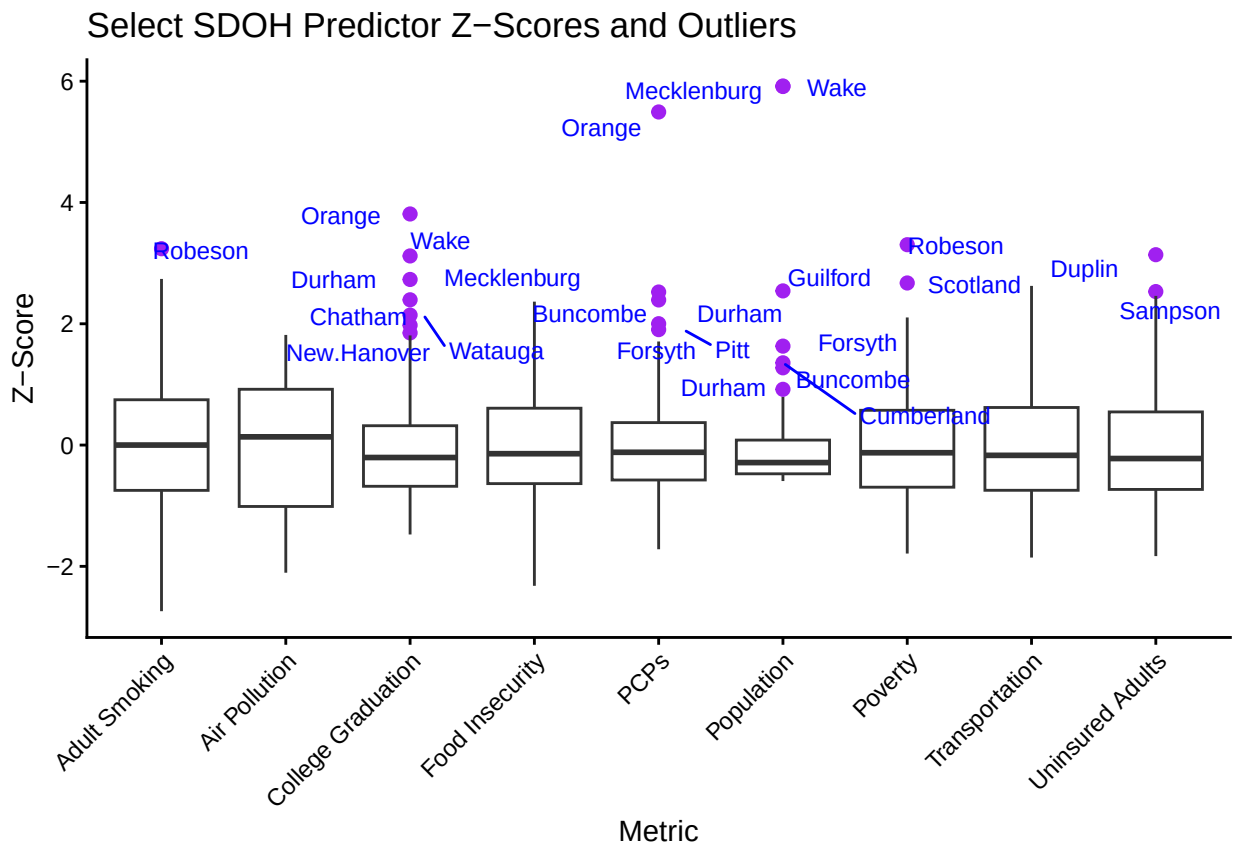
```
## # A tibble: 9 x 8
##   Metric               Mean Median    SD    Min   Max Low_Outliers High_Outliers
##   <chr>               <dbl>  <dbl> <dbl>  <dbl> <dbl> <chr>        <chr>
## 1 Adult Smoking   -4.72e-17  0      1    -2.74   3.24 <NA>         Robeson
## 2 Air Pollution   -3.31e-16  0.136  1    -2.10   1.82 <NA>         <NA>
## 3 College Gradua~ -6.52e-17 -0.205  1    -1.47   3.81 <NA>         Chatham, Dur~
## 4 Food Insecurity -1.05e-16 -0.143  1.00 -2.32   2.36 <NA>         <NA>
## 5 Health Care Wo~ -1.55e-17 -0.119  1    -1.72   5.49 <NA>         Buncombe, Du~
## 6 Population      -8.88e-17 -0.289  1    -0.593  5.92 <NA>         Buncombe, Cu~
## 7 Poverty         -1.55e-16 -0.126  1    -1.79   3.30 <NA>         Robeson, Sco~
## 8 Transportation   3.84e-17 -0.170  1    -1.85   2.62 <NA>         <NA>
```

```
## 9 Uninsured Adul~  1.88e-16 -0.222  1    -1.83   3.14 <NA>         Duplin, Samp~
```

```r
# readability for plot (shorten healthcare name)
zscore_data <- zscore_data %>% mutate(Metric = if_else(
    Metric == "Health Care Workforce - Primary Care Physicians",
    "PCPs",
    Metric
  ))

# putting summary stats in a csv to transfer to report
write.csv(summary_stats, "sdoh_zscore_summary_stats.csv", row.names = FALSE)

# boxplot to visualize outliers, with labels on each outlier
ggplot(zscore_data %>% filter(!Metric == "Metropolitan or Nonmetropolitan"), aes(x = Metric, y = Z_Score
  geom_boxplot(outlier.colour = "purple", outlier.size = 2) +
  #geom_jitter(width = 0.2, alpha = 0.5) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Select SDOH Predictor Z-Scores and Outliers", x = "Metric", y = "Z-Score") +
  geom_text_repel(
    data = zscore_data %>%
      filter(!Metric == "Metropolitan or Nonmetropolitan") %>%
      group_by(Metric) %>%
      filter(Z_Score > (quantile(Z_Score, 0.75) + 1.5 * IQR(Z_Score)) |
               Z_Score < (quantile(Z_Score, 0.25) - 1.5 * IQR(Z_Score))),
    aes(label = County),
    position = position_jitter(width = 0.2),
    vjust = -0.5,
    size = 3,
    color = "blue"
  ) +
  theme_classic() +
  theme(
      axis.text.x = element_text(
          angle = 45,
          hjust = 1
      )
  )
```




```
# correlation matrix
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
correlation_data <- zscore_data %>%
  filter(!Metric == "Metropolitan or Nonmetropolitan") %>%
  select(Metric, County, Z_Score) %>%
  pivot_wider(names_from = Metric, values_from = Z_Score)

correlation_matrix <- cor(correlation_data %>% select(-County), use = "pairwise.complete.obs")

# peek data
# correlation_matrix

# visualize correlation matrix
library(ggcorrplot)
ggcorrplot(correlation_matrix,
           method = "square",
           type = "lower",
```

```
          lab = TRUE,
          title = "Correlation Matrix of Select SDOH Predictor Z-Scores",
          colors = c("red", "white", "blue")) +
theme_classic() +
labs(x = "", y = "") +
theme(
    axis.text.x = element_text(
        angle = 45,
        hjust = 1,
    )
)
```