Project 1: Minneapolis Demographic and Police Stop Statistics

Jena Georgopulos, jg64565

3/15/2021

Introduction

This project will analyze two datasets after combining the contents of these two datasets into a single dataset.The first dataset is titled mincop, and contains data regarding the stops made by members of the Minneapolis police department in the year 2017. The dataset contains a total of 51857 observations, each representing an individual who was pulled over by Minneapolis police in the year 2017. The variables that we will keep from this dataset are as follows: the reason for the stop ('problem'), whether the person's body was searched ('personsearch'), whether their vehical was searched ('vehiclesearch'), the race of the person pulled over ('race'), the gender of the person pulled over ('gender'), the neighborhood that they were pulled over in ('neighborhood'), and the police precinct in that area ('policePrecinct'). The second dataset is titled mindemo and contains data about the demographics of 84 Minneapolis neighborhoods in the year 2015. The data contains a total of 84 observations, each representing a neighborhood in Minneapolis. The variables that we will keep from this dataset are as follows: The neighborhood name ('neighborhood'), the population of the neighborhood ('population'), the estimated median household income of the neighborhood ('hhIncome'), and the proportion of people living below the poverty line in the neighborhood ('poverty'). Both of these datasets were downloaded as csv files from the link provided in the instructions. I expect to find an association between whether a person's car was search and whether their vehicle was searched. I also expect to find an association between an individual's race and whether their vehicle or person was searched.

Upload and Organize datasets

```
# Install necessary packages
#install.packages(magrittr)
library(magrittr)
#install.packages(tidyr)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

```
library(tibble)
#install.packages(dplyr)
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
```

```
#install.packages(ggplot2)
library(ggplot2)
```

```
# Upload the datasets
mindemo <- read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/carData/MplsDemo.csv")

mincop <- read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/carData/MplsStops.csv")

# Remove excess variables
mincop <- mincop %>%
  select(-X, -idNum, -date, -MDC, -citationIssued, -lat, -long, -preRace) %>%
  na.omit()
mindemo <- mindemo %>%
  select(-X, -white, -black, -foreignBorn, -collegeGrad)
```

Joining the Two Datasets

```
# Perfrom a left join on the datasets
combdataL <- mincop %>%
  left_join(mindemo, by="neighborhood")
#now remove all of the N/As
combdataL <- combdataL %>%
  na.omit()
# Check out first few rows of the combined dataset
head(combdataL)
```

```
##      problem personSearch vehicleSearch         race  gender policePrecinct
## 1 suspicious           NO            NO      Unknown Unknown              1
## 2 suspicious           NO            NO      Unknown    Male              1
## 3    traffic           NO            NO        White  Female              5
## 4 suspicious           NO            NO East African    Male              5
## 5    traffic           NO            NO        White  Female              1
## 6    traffic           NO            NO East African    Male              1
##      neighborhood population hhIncome poverty
## 1 Cedar Riverside       8247    18892   0.060
## 2   Downtown West       7141    67086   0.057
## 3        Whittier      14604    35855   0.038
## 4        Whittier      14604    35855   0.038
## 5   Downtown West       7141    67086   0.057
## 6   Downtown West       7141    67086   0.057
```

```
# Comare the length of mindemo with combdataL
str(mindemo)
```

```
## 'data.frame':    84 obs. of  4 variables:
##  $ neighborhood: chr  "Cedar Riverside" "Phillips West" "Downtown West" "Downtown
East" ...
##  $ population  : int  8247 5184 7141 1674 3249 6150 1676 5420 4525 5109 ...
##  $ hhIncome    : int  18892 18404 67086 70669 59414 17469 18854 43438 57148 37030
...
##  $ poverty     : num  0.06 0.042 0.057 0.071 0.11 0.048 0.074 0.089 0.066 0.053 ..
.
```

```
str(combdataL)
```

```
## 'data.frame':    41646 obs. of  10 variables:
##  $ problem       : chr  "suspicious" "suspicious" "traffic" "suspicious" ...
##  $ personSearch  : chr  "NO" "NO" "NO" "NO" ...
##  $ vehicleSearch : chr  "NO" "NO" "NO" "NO" ...
##  $ race          : chr  "Unknown" "Unknown" "White" "East African" ...
##  $ gender        : chr  "Unknown" "Male" "Female" "Male" ...
##  $ policePrecinct: int  1 1 5 5 1 1 1 2 2 4 ...
##  $ neighborhood  : chr  "Cedar Riverside" "Downtown West" "Whittier" "Whittier" ..
.
##  $ population    : int  8247 7141 14604 14604 7141 7141 7141 10496 1393 5023 ...
##  $ hhIncome      : int  18892 67086 35855 35855 67086 67086 67086 27104 83520 4044
2 ...
##  $ poverty       : num  0.06 0.057 0.038 0.038 0.057 0.057 0.057 0.042 0.076 0.103
...
##  - attr(*, "na.action")= 'omit' Named int [1:1992] 19 20 24 46 51 165 168 182 197
201 ...
##   ..- attr(*, "names")= chr [1:1992] "19" "20" "24" "46" ...
```

```
# subtract number of observations in mindemo from combdataL
41646 - 84
```

```
## [1] 41562
```

```
# Now compare the kength of mincop with combdataL
str(mincop)
```

```
## 'data.frame':    43638 obs. of  7 variables:
##  $ problem       : chr  "suspicious" "suspicious" "traffic" "suspicious" ...
##  $ personSearch  : chr  "NO" "NO" "NO" "NO" ...
##  $ vehicleSearch : chr  "NO" "NO" "NO" "NO" ...
##  $ race          : chr  "Unknown" "Unknown" "White" "East African" ...
##  $ gender        : chr  "Unknown" "Male" "Female" "Male" ...
##  $ policePrecinct: int  1 1 5 5 1 1 1 2 2 4 ...
##  $ neighborhood  : chr  "Cedar Riverside" "Downtown West" "Whittier" "Whittier" ..
.
##  - attr(*, "na.action")= 'omit' Named int [1:8282] 12 18 29 33 34 35 43 45 51 55 .
..
##   ..- attr(*, "names")= chr [1:8282] "12" "18" "29" "33" ...
```

```
str(combdataL)
```

```
## 'data.frame':    41646 obs. of  10 variables:
##  $ problem       : chr  "suspicious" "suspicious" "traffic" "suspicious" ...
##  $ personSearch  : chr  "NO" "NO" "NO" "NO" ...
##  $ vehicleSearch : chr  "NO" "NO" "NO" "NO" ...
##  $ race          : chr  "Unknown" "Unknown" "White" "East African" ...
##  $ gender        : chr  "Unknown" "Male" "Female" "Male" ...
##  $ policePrecinct: int  1 1 5 5 1 1 1 2 2 4 ...
##  $ neighborhood  : chr  "Cedar Riverside" "Downtown West" "Whittier" "Whittier" ..
.
##  $ population    : int  8247 7141 14604 14604 7141 7141 7141 10496 1393 5023 ...
##  $ hhIncome      : int  18892 67086 35855 35855 67086 67086 67086 27104 83520 4044
2 ...
##  $ poverty       : num  0.06 0.057 0.038 0.038 0.057 0.057 0.057 0.042 0.076 0.103
...
##  - attr(*, "na.action")= 'omit' Named int [1:1992] 19 20 24 46 51 165 168 182 197
201 ...
##   ..- attr(*, "names")= chr [1:1992] "19" "20" "24" "46" ...
```

```
# Subtract the number of observations in combdataL from mincop
43638 - 41646
```

```
## [1] 1992
```

A left join was perfromed in order to combine the datasets of mincop and mindemo to make the new dataset titled ‘combdataL’. This was done because we wanted to keep the rows from the first, longer, dataset mincop and add the matching columns from the second dataset, mindemo. This was necessary because we were matching the datasets based on the key variable of “neighborhood”. Therefore, we only needed to add the two remaining columns of “population” and “hhincome” from mindemo to the dataset of mincop after matching all of the neighborhoods present in both datasets. After perfroming the join, all NA values were dropped from the combined dataset. This resulted in a total of 1992 observations being dropped from the dataset mincop and zero observations being droped from the dataset mindemo. There were no observations dropped from the mindemo dataset becasue all 84 listed neighborhoods were also found to be listed in the mincop dataset.

Summary Statistics

```
## Use all six core dyplr functions (filter, select, arrange, group_by, mutate, summa
rize) to manipulate data

# Use filter to filter the dataset so that it only shows observations where the neigh
borhood is Cedar Riverside
filterdata <- combdataL %>%
  filter(neighborhood == "Cedar Riverside")
head(filterdata)
```

```
##      problem personSearch vehicleSearch            race  gender policePrecinct
## 1 suspicious           NO            NO         Unknown Unknown              1
## 2 suspicious           NO            NO           White    Male              1
## 3 suspicious           NO            NO Native American    Male              1
## 4    traffic           NO            NO    East African    Male              1
## 5 suspicious          YES            NO           Black    Male              1
## 6    traffic           NO            NO           Black  Female              1
##      neighborhood population hhIncome poverty
## 1 Cedar Riverside       8247    18892    0.06
## 2 Cedar Riverside       8247    18892    0.06
## 3 Cedar Riverside       8247    18892    0.06
## 4 Cedar Riverside       8247    18892    0.06
## 5 Cedar Riverside       8247    18892    0.06
## 6 Cedar Riverside       8247    18892    0.06
```

```
#Now use filter again to only show observations where the neighborhood is Cedar River
side, and the individual pulled over was a female
filterdata1 <- combdataL %>%
  filter(neighborhood == "Cedar Riverside" & gender == "Female")
head(filterdata1)
```

```
##      problem personSearch vehicleSearch    race gender policePrecinct
## 1    traffic           NO            NO   Black Female              1
## 2    traffic           NO            NO   Other Female              1
## 3 suspicious           NO            NO   Black Female              1
## 4    traffic           NO            NO Unknown Female              1
## 5 suspicious           NO            NO   Black Female              1
## 6    traffic           NO            NO   Black Female              1
##      neighborhood population hhIncome poverty
## 1 Cedar Riverside       8247    18892    0.06
## 2 Cedar Riverside       8247    18892    0.06
## 3 Cedar Riverside       8247    18892    0.06
## 4 Cedar Riverside       8247    18892    0.06
## 5 Cedar Riverside       8247    18892    0.06
## 6 Cedar Riverside       8247    18892    0.06
```

```
# Use select to select variables of gender, neighborhood, population and household in
come
selectdata <- combdataL %>%
  select(gender, neighborhood, population, hhIncome)
head(selectdata)
```

```
##    gender    neighborhood population hhIncome
## 1 Unknown Cedar Riverside       8247    18892
## 2    Male   Downtown West       7141    67086
## 3  Female        Whittier      14604    35855
## 4    Male        Whittier      14604    35855
## 5  Female   Downtown West       7141    67086
## 6    Male   Downtown West       7141    67086
```

```
# Use arrange to arrange the data in combdataL by household income in order of least-
to-greatest
arrangedata <- combdataL %>%
  arrange(hhIncome)
head(arrangedata)
```

```
##      problem personSearch vehicleSearch         race  gender policePrecinct
## 1    traffic           NO            NO East African    Male              3
## 2 suspicious           NO            NO        Black    Male              3
## 3 suspicious           NO            NO      Unknown    Male              3
## 4 suspicious           NO            NO      Unknown    Male              3
## 5 suspicious           NO            NO      Unknown Unknown              3
## 6    traffic          YES           YES        Black    Male              3
##      neighborhood population hhIncome poverty
## 1 Ventura Village       6150    17469   0.048
## 2 Ventura Village       6150    17469   0.048
## 3 Ventura Village       6150    17469   0.048
## 4 Ventura Village       6150    17469   0.048
## 5 Ventura Village       6150    17469   0.048
## 6 Ventura Village       6150    17469   0.048
```

```
# Use mutate to create a new variable to find the number of people living below the p
overty line in each neighborhood then arrange by the number of people in living below
the poverty line from least-to-greatest
mutatedata <- combdataL %>%
  mutate(poverty_total = population*poverty) %>%
  arrange(poverty_total)
head(mutatedata)
```

```
##   problem personSearch vehicleSearch  race gender policePrecinct
## 1 traffic           NO            NO White   Male              2
## 2 traffic           NO            NO Black   Male              2
## 3 traffic          YES           YES Black   Male              2
## 4 traffic           NO            NO White   Male              2
## 5 traffic           NO            NO White   Male              2
## 6 traffic           NO            NO White   Male              2
##            neighborhood population hhIncome poverty poverty_total
## 1 Mid - City Industrial        240    38875   0.067         16.08
## 2 Mid - City Industrial        240    38875   0.067         16.08
## 3 Mid - City Industrial        240    38875   0.067         16.08
## 4 Mid - City Industrial        240    38875   0.067         16.08
## 5 Mid - City Industrial        240    38875   0.067         16.08
## 6 Mid - City Industrial        240    38875   0.067         16.08
```

```
# Create summary statistics for each numeric variable in combdataL
combdataL %>%
  summarize(mean(population), sd(population), min(population), max(population),
            mean(hhIncome), sd(hhIncome), min(hhIncome), max(hhIncome),
            mean(poverty), sd(poverty), min(poverty), max(poverty))
```

```
##   mean(population) sd(population) min(population) max(population)
## 1         6314.545       3308.973             240           16022
##   mean(hhIncome) sd(hhIncome) min(hhIncome) max(hhIncome) mean(poverty)
## 1       49838.15     22166.61         17469        118750    0.06322814
##   sd(poverty) min(poverty) max(poverty)
## 1  0.02204858        0.031        0.135
```

```
## Create summary stats for each numeric variable in combdataL after grouping by a ca
tegorical variable

# Find the average number of people living below the poverty line by neighborhood
combdataL %>%
  mutate(poverty_total = population*poverty) %>%
  group_by(neighborhood) %>%
  summarize(mean_poverty = mean(poverty_total))
```

```
## # A tibble: 84 x 2
##    neighborhood         mean_poverty
##  * <chr>                       <dbl>
##  1 Armatage                     243.
##  2 Audubon Park                 269.
##  3 Bancroft                     188.
##  4 Beltrami                     111.
##  5 Bottineau                    134.
##  6 Bryant                       362.
##  7 Bryn - Mawr                  140.
##  8 CARAG                        281.
##  9 Cedar - Isles - Dean         266.
## 10 Cedar Riverside              495.
## # … with 74 more rows
```

```
# Find average houshold income by gender
combdataL %>%
  group_by(gender) %>%
  summarize(mean_income = mean(hhIncome))
```

```
## # A tibble: 3 x 2
##   gender  mean_income
## * <chr>         <dbl>
## 1 Female       51281.
## 2 Male         49283.
## 3 Unknown      49935.
```

```
# Find average population by neighborhood
combdataL %>%
  group_by(neighborhood) %>%
  summarize(mean_pop = mean(population))
```

```
## # A tibble: 84 x 2
##    neighborhood         mean_pop
##  * <chr>                   <dbl>
##  1 Armatage                 4864
##  2 Audubon Park             5073
##  3 Bancroft                 3542
##  4 Beltrami                 1243
##  5 Bottineau                1573
##  6 Bryant                   3178
##  7 Bryn - Mawr              2791
##  8 CARAG                    5737
##  9 Cedar - Isles - Dean     2984
## 10 Cedar Riverside          8247
## # … with 74 more rows
```

In the combdataL dataset, the average population is 6314.545 people, the standard deviation of the population is 3308.973, the minimum number of people in a neighborhood is 240 people, and the maximum number of people in a neighborhood is 16022 people.

In the dataset combdataL, the average yearly household income is 49838.15 dollars, the standard deviation of the yearly household income is 22166.61, the minimum yearly household income is 17469 dollars, and the maximum yearly household income is 118750 dollars.

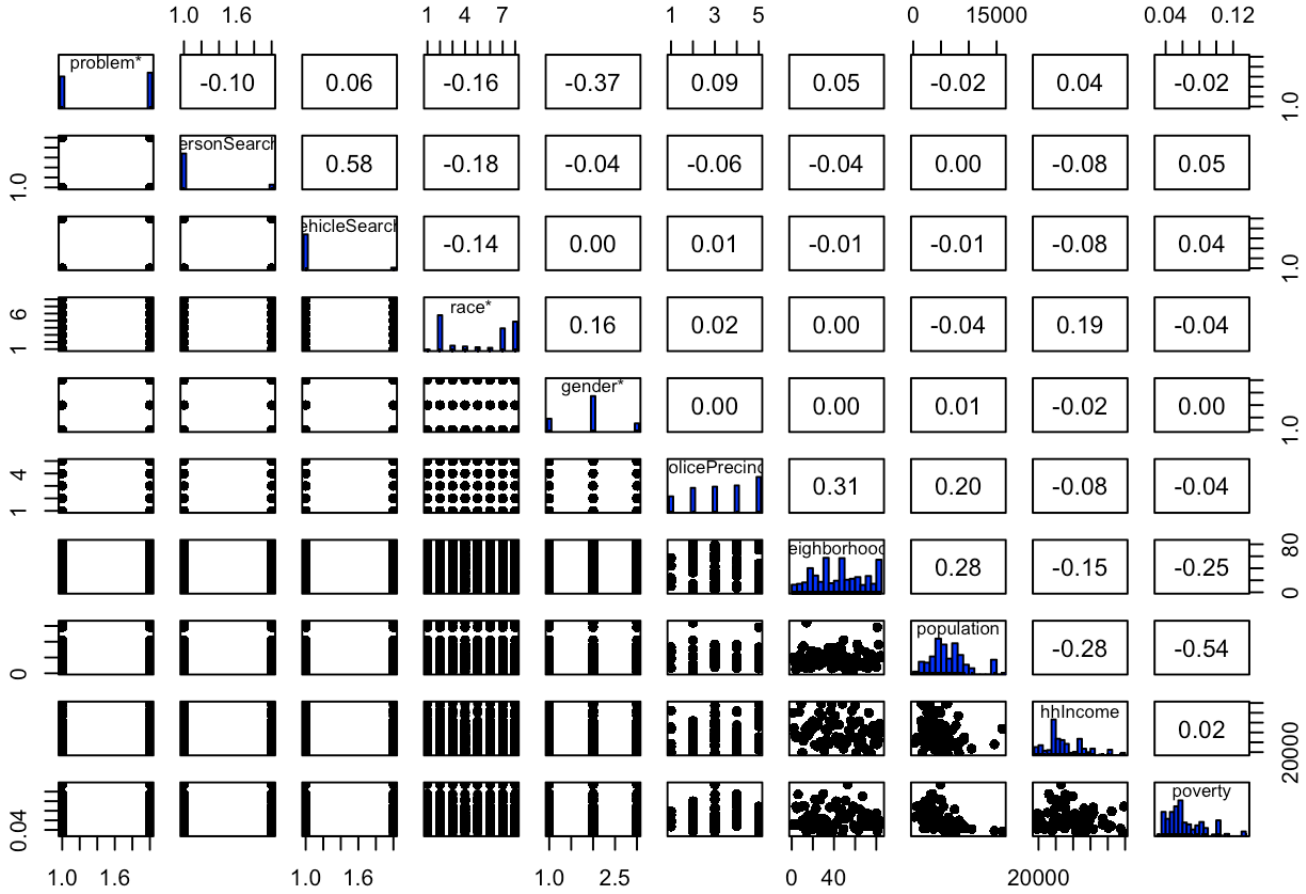
In the dataset combdataL, the average proportion of people living below the poverty line is 0.06322814 people. So on average, approximately 6.3% of the population is living below the poverty line. The standard deviation of the proportion of people living below the poverty line is 0.02204858. The minimum proportion of people living below the poverty line is 0.031 people. So at minimum, 3.1% of the population is living below the poverty line. The maximum proportion of people living below the poverty line is 0.135 people. So at maximum, 13.5% of the population is living below the poverty line.

```
# Build a correlation matrix for all variables
# install.packages(psych)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
# Correlation matrix of categorical and numeric variables
pairs.panels(combdataL,
             method = "pearson",
             hist.col = "blue",
             smooth = FALSE, density = FALSE, ellipses = FALSE)
```



The above correlation matrix shows the relationship between all numeric and categorical variables in the dataset. It can be seen that the highest correlation coefficient exists between the variables of vehicle search and person search (R = 0.58). This indicates that the decision of an officer to search people's vehicles is fairly closely associated with their decision to perform a body search on the same individual. There does not appear to be a very close relationship between the variables of neighborhood and police precinct (R = 0.31) and there also does not appear to be a close relationship between population and neighborhood (R = 0.28). There does not appear to be a large association between the race of an individual and whether their person (R = -0.18) or vehicle (R = -0.14) was searched by the officer.

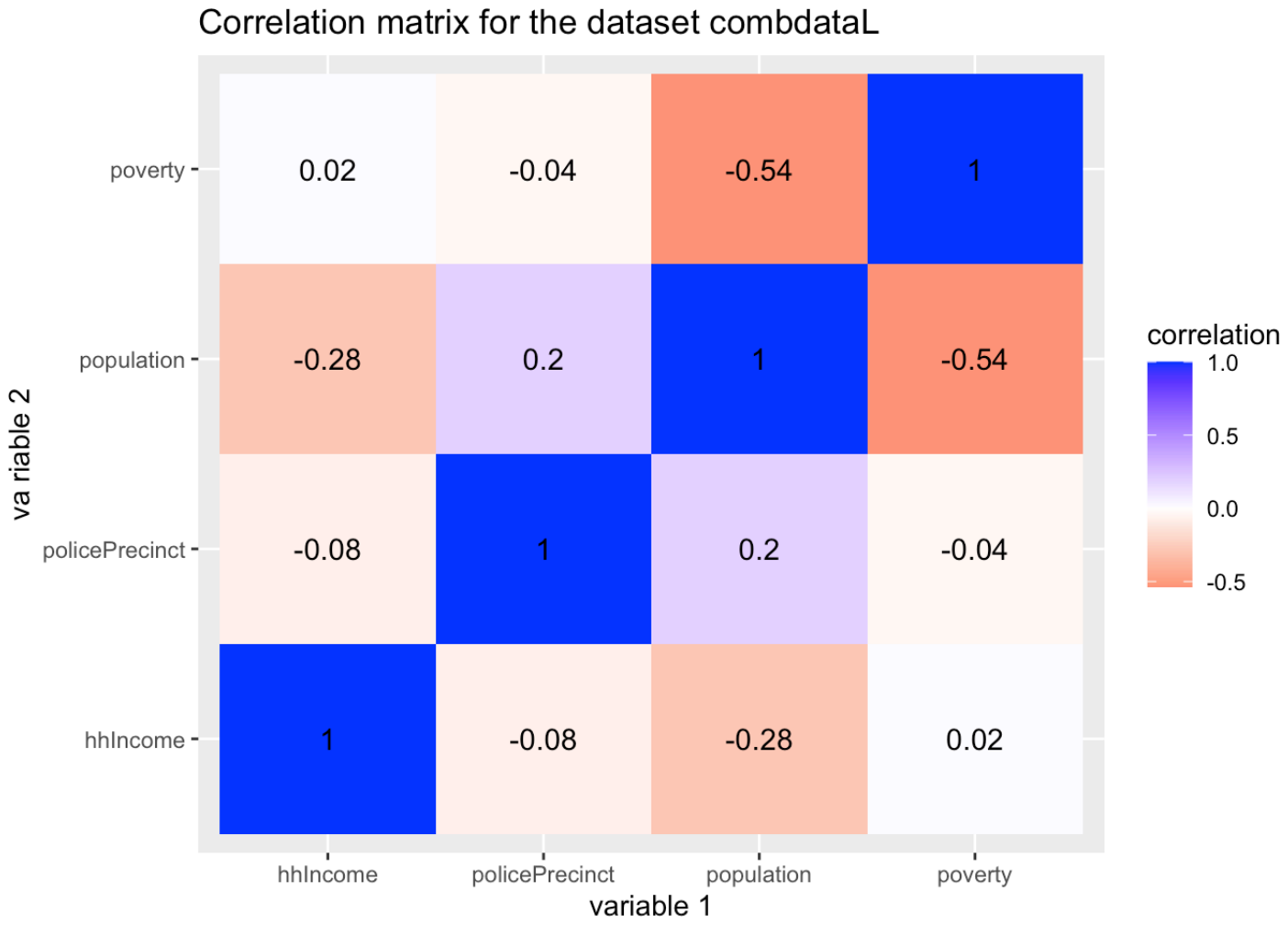
Visualizations

```
#install.packages(tidyr)
library(tidyr)
library(tibble)

# Build a correlation matrix between all numeric variables
combdataL_num <- combdataL %>% select_if(is.numeric)
cor(combdataL_num, use = "pairwise.complete.obs")
```

```
##                policePrecinct population    hhIncome     poverty
## policePrecinct     1.00000000  0.1990641 -0.07541720 -0.03649409
## population         0.19906412  1.0000000 -0.28244704 -0.53691877
## hhIncome          -0.07541720 -0.2824470  1.00000000  0.01761545
## poverty           -0.03649409 -0.5369188  0.01761545  1.00000000
```

```
#  Make it pretty using a heatmap with geom_tile!
cor(combdataL_num, use = "pairwise.complete.obs") %>%
  # Save as a data frame
as.data.frame %>%
  # Convert row names to an explicit variable
rownames_to_column %>%
  # Pivot so that all correlations appear in the same column
pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>% ggplot(aes(ro
wname, other_var, fill=correlation)) +
# Heatmap with geom_tile
geom_tile() +
  # Change the scale to make the middle appear neutral
scale_fill_gradient2(low="red",mid="white",high="blue") +
  # Overlay values
geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  # Give title and labels
labs(title = "Correlation matrix for the dataset combdataL", x = "variable 1", y = "v
a riable 2")
```



The above correlation heatmap shows that there is a somewhat high negative correlation between the population and the proportion of people living below the poverty line (R = -0.54). There appears to be a small negative correlation between population and household income (R = -0.28).