Lab Assignment 8

Summer Heschong

2025-03-07

(1) Load, Tidy, and Combine Datasets

Load packages and data

```
library(here)
```

```
## here() starts at /Users/summerheschong/stats_spring25
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- 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 error
```

```
library(dplyr)

#load data
Census_data <- read.csv(here('Data/Raw/NC_Census.csv'))
Recreation_data <- read.csv(here('Data/Raw/NC_Recreation_Acreage.csv'))
```

Tidy data

```
#filter Recreation data for relevant counties and sum local, state,
#and federal recreation acreage to form total recreation acreage by county
Recreation_data <- Recreation_data %>%
  filter(Area.Name %in% c('Chatham County', 'Wake County',
                          'Alamance County','Orange County',
                          'Durham County', 'Caswell County',
                          'Person County', 'Granville County',
                          'Franklin County', 'Vance County',
                          'Lee County', 'Moore County',
                          'Randolph County', 'Guilford County',
                          'Rockingham County', 'Halifax County',
                          'Nash County', 'Wilson County',
                          'Johnston County', 'Harnett County')) %>%
  pivot_wider(names_from = Variable, values_from = Value) %>%
  group_by(`Total Outdoor Recreation Acreage` =
             `Local Outdoor Recreation Acreage` +
             `Federal Outdoor Recreation Acreage` +
             `State Outdoor Recreation Acreage`)

#Filter census data
Census_data <- Census_data %>%
  filter(Fact %in% c("Population estimates, July 1, 2023, (V2023)",
                     "Persons under 18 years, percent",
                     "Persons 65 years and over, percent",
                     "Female persons, percent",
                     "White alone, percent",
                     "Black alone, percent (a)",
                     "Hispanic or Latino, percent (b)",
                     "Median value of owner-occupied housing units, 2019-2023",
                     "Persons per household, 2019-2023",
                     "Median households income (in 2023 dollars), 2019-2023"))
```

Combine data

```
#ready census data for combining
Census_data <- Census_data %>%
  select(!Fact.Note)%>% #remove Fact.Note column
  pivot_longer(!c(Fact), #combine county columns into one column
               names_to = 'Counties',
               values_to = 'Values') %>%
  pivot_wider(names_from = Fact, values_from = Values) %>%#make each row in Fact its own column
  mutate(Area.Name = c('Chatham County', 'Wake County',     # make a column that's exactly
                          'Alamance County','Orange County',#the same as one in the other
                          'Durham County', 'Caswell County',#dataset
                          'Person County', 'Granville County',
                          'Franklin County', 'Vance County',
                          'Lee County', 'Moore County',
                          'Randolph County', 'Guilford County',
                          'Rockingham County', 'Halifax County',
                          'Nash County', 'Wilson County',
                          'Johnston County', 'Harnett County'))

#convert characters to numbers
Census_data$`Median value of owner-occupied housing units, 2019-2023` <-
  as.numeric(gsub("[$,]", "",
  Census_data$`Median value of owner-occupied housing units, 2019-2023`))
```

```r
Census_data$`Median households income (in 2023 dollars), 2019-2023` <-
  as.numeric(gsub("[$,]", "",
  Census_data$`Median households income (in 2023 dollars), 2019-2023`))

Census_data$`Persons under 18 years, percent` <-
  as.numeric(gsub("[%]", "", Census_data$`Persons under 18 years, percent`))

Census_data$`Persons 65 years and over, percent` <-
  as.numeric(gsub("[%]", "", Census_data$`Persons 65 years and over, percent`))

Census_data$`Female persons, percent` <-
  as.numeric(gsub("[%]", "", Census_data$`Female persons, percent`))

Census_data$`White alone, percent` <-
  as.numeric(gsub("[%]", "", Census_data$`White alone, percent`))

Census_data$`Black alone, percent (a)` <-
  as.numeric(gsub("[%]", "", Census_data$`Black alone, percent (a)`))

Census_data$`Hispanic or Latino, percent (b)` <-
  as.numeric(gsub("[%]", "", Census_data$`Hispanic or Latino, percent (b)`))

Census_data$`Population estimates, July 1, 2023, (V2023)` <-
  as.numeric(gsub("[,]", "",
                  Census_data$`Population estimates, July 1, 2023, (V2023)`))

Census_data$`Persons per household, 2019-2023` <-
  as.numeric(Census_data$`Persons per household, 2019-2023`)


#combine datasets
combined_data <- left_join(Census_data, Recreation_data)
```

```
## Joining with ‘by = join_by(Area.Name)‘
```

(2) Perform a Multiple Linear Regression

Step 1 - Define Research Question

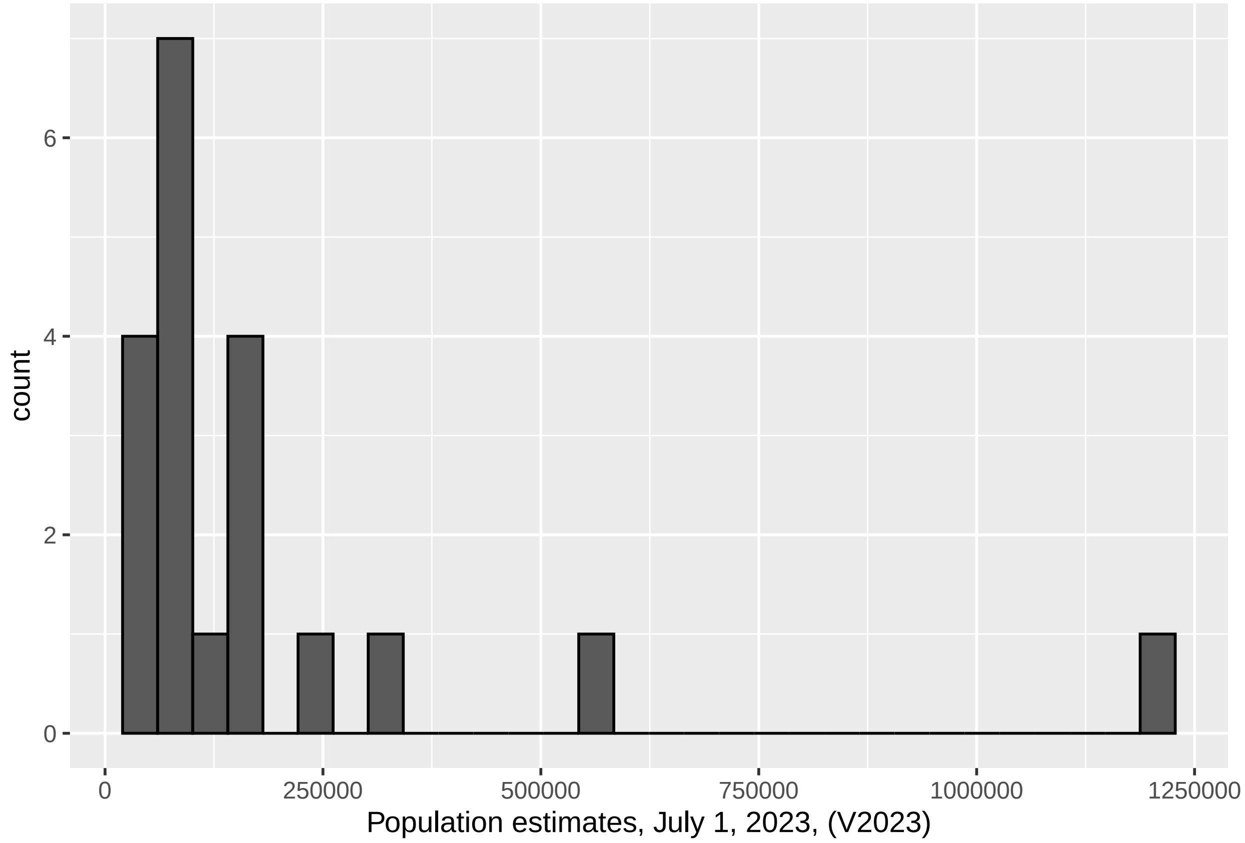
Research Question: What demographic factors significantly predict outdoor recreation acreage at the county level across 20 North Carolina counties?

Step 2 - Examine Data

Create histograms of raw values for each variable

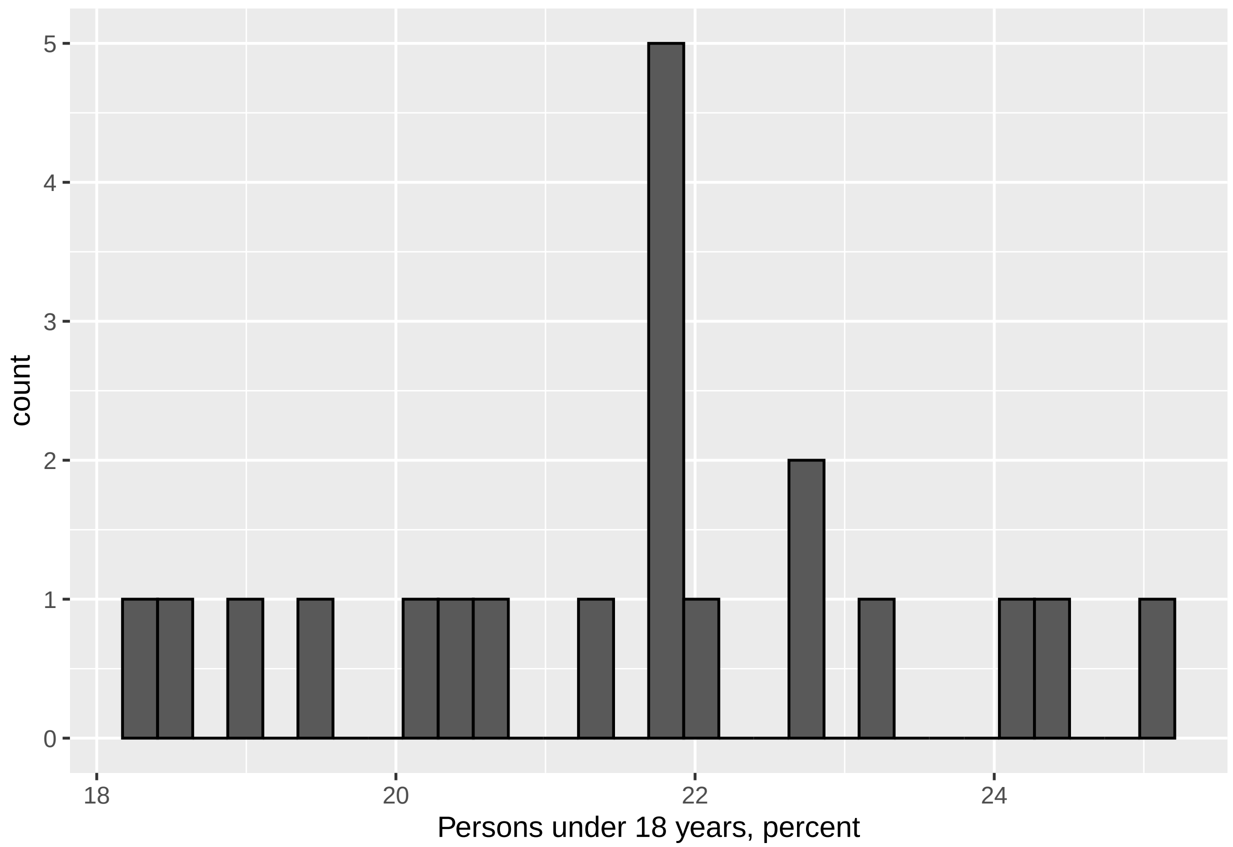
```r
#Pop estimates
ggplot(combined_data, aes(x = `Population estimates, July 1, 2023, (V2023)`)) +
geom_histogram(color = 'black')
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



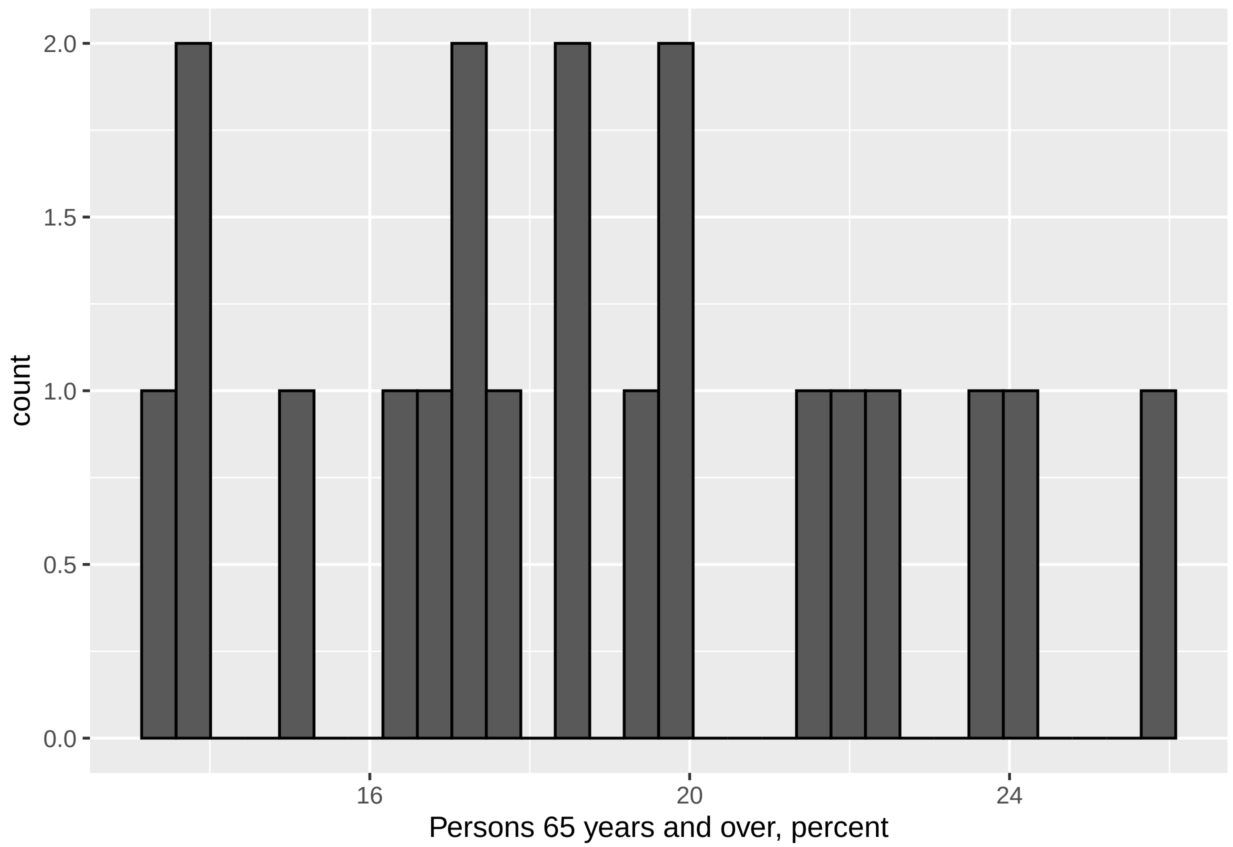
```r
#Ppl under 18
ggplot(combined_data, aes(x = `Persons under 18 years, percent`)) +
geom_histogram(color = 'black')
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



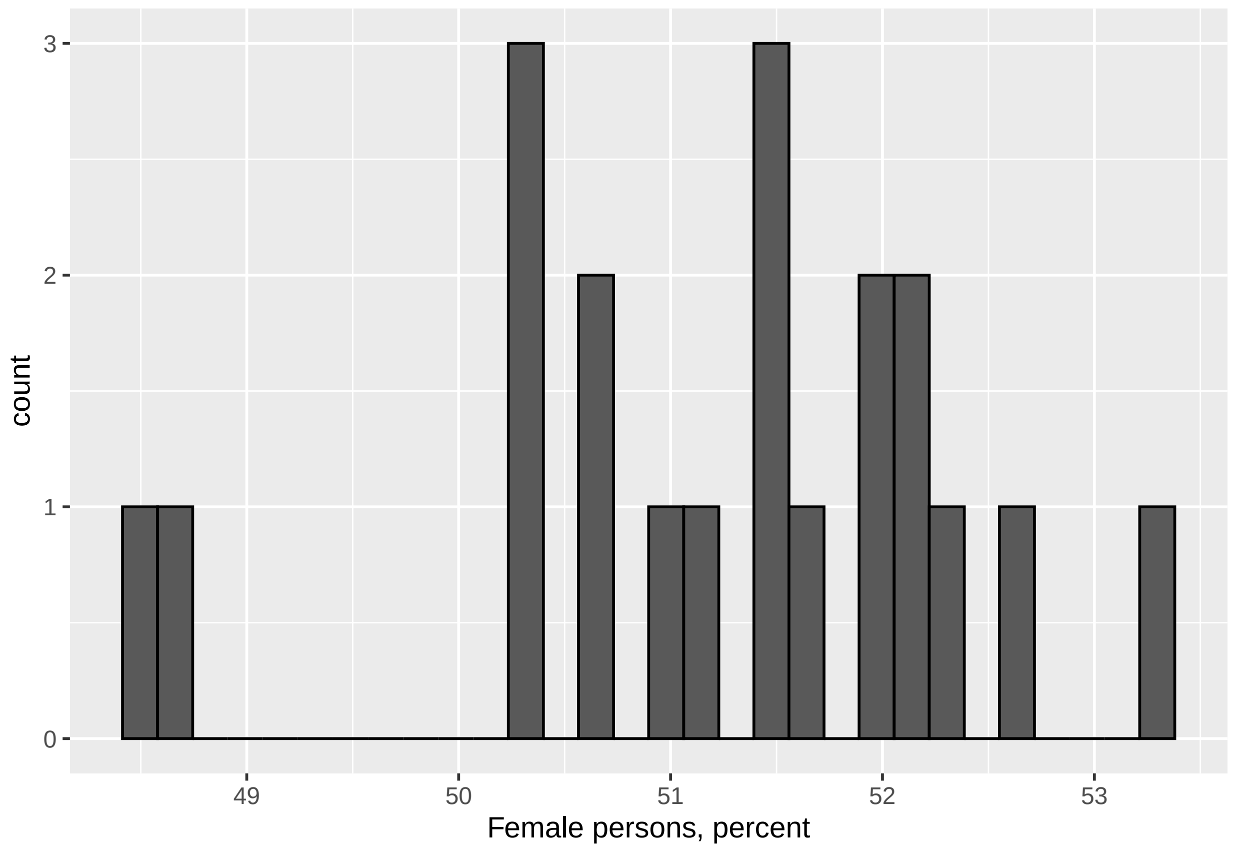
```
#Ppl 65 and over
ggplot(combined_data, aes(x = `Persons 65 years and over, percent`)) +
geom_histogram(color = 'black')
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



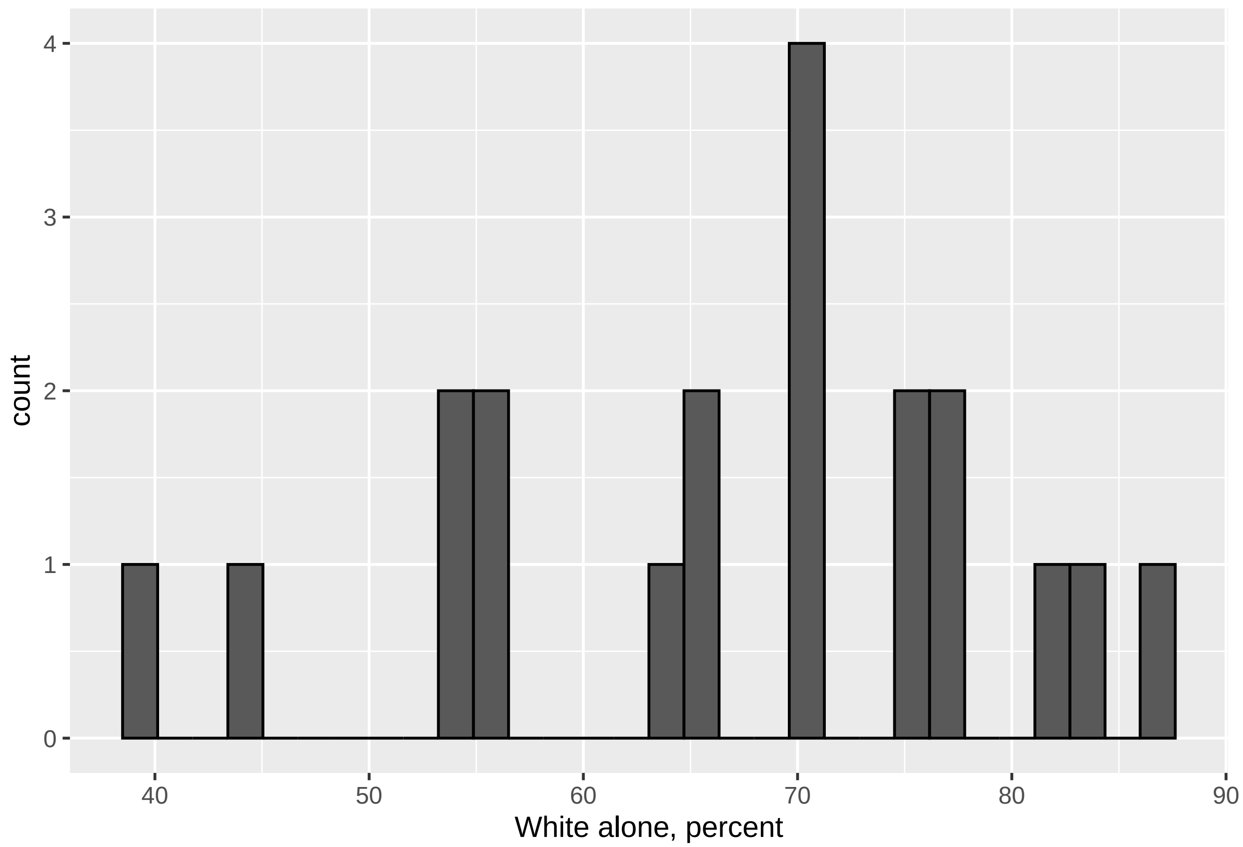
```
#XX ppl
ggplot(combined_data, aes(x = `Female persons, percent`)) +
geom_histogram(color = 'black')
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



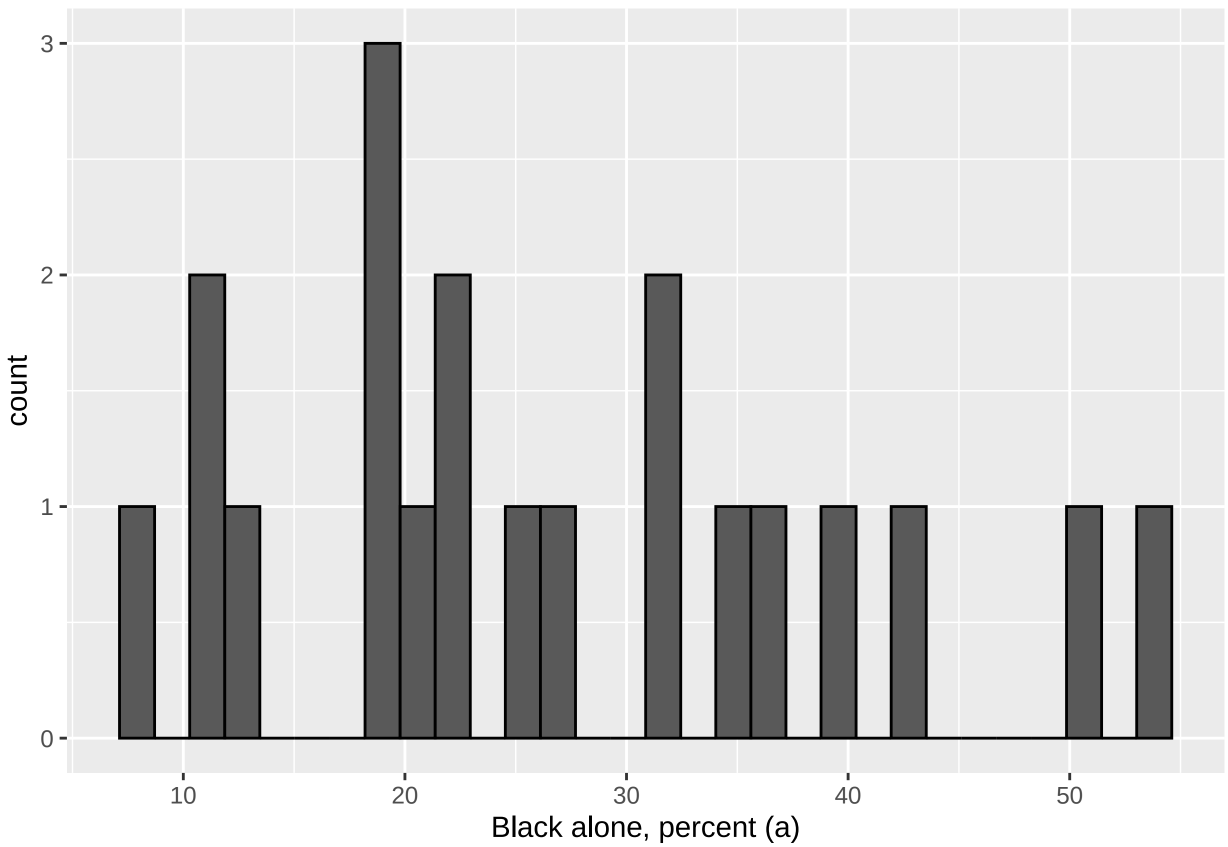
```
#White ppl
ggplot(combined_data, aes(x = `White alone, percent`)) +
geom_histogram(color = 'black')
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



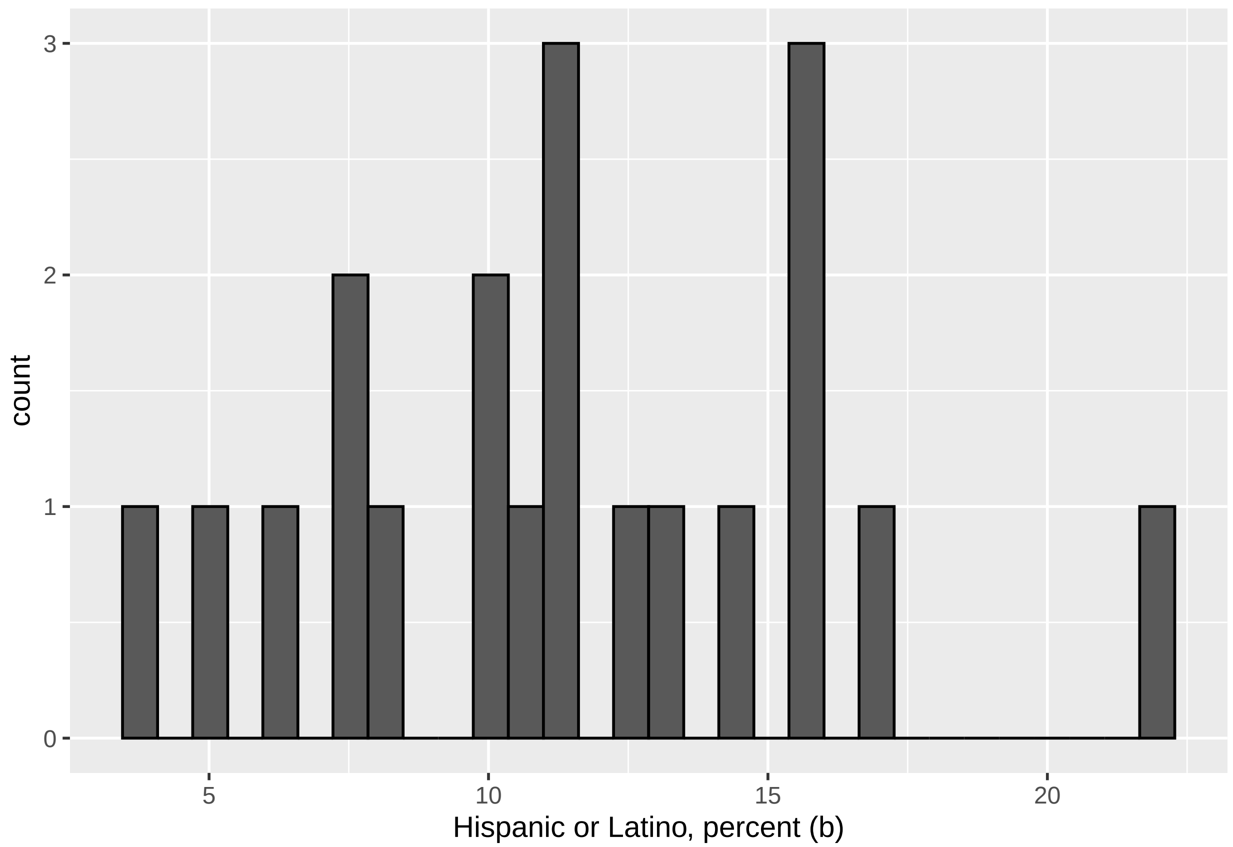
```r
#Black ppl
ggplot(combined_data, aes(x = `Black alone, percent (a)`)) +
geom_histogram(color = 'black')
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



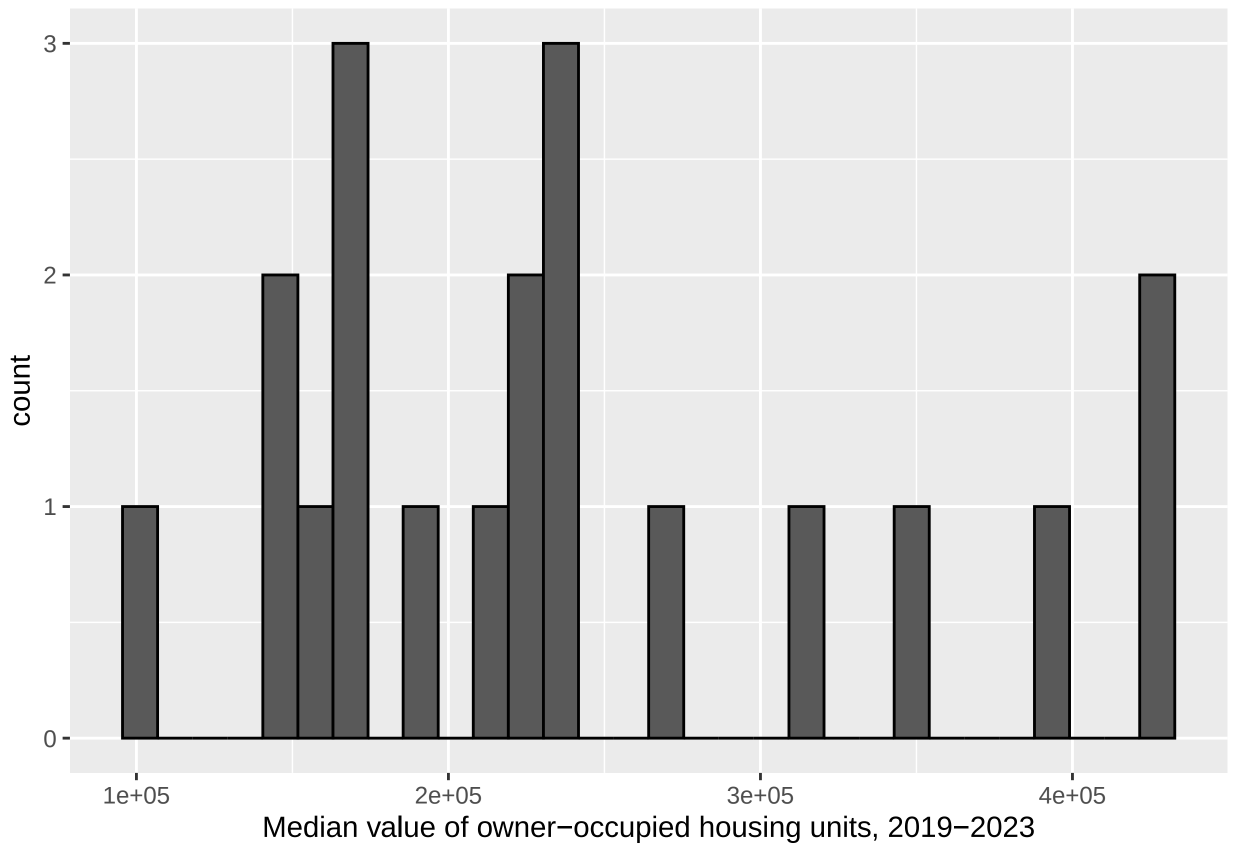
```
#Hispanic/Latine ppl
ggplot(combined_data, aes(x = `Hispanic or Latino, percent (b)`)) +
geom_histogram(color = 'black')
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



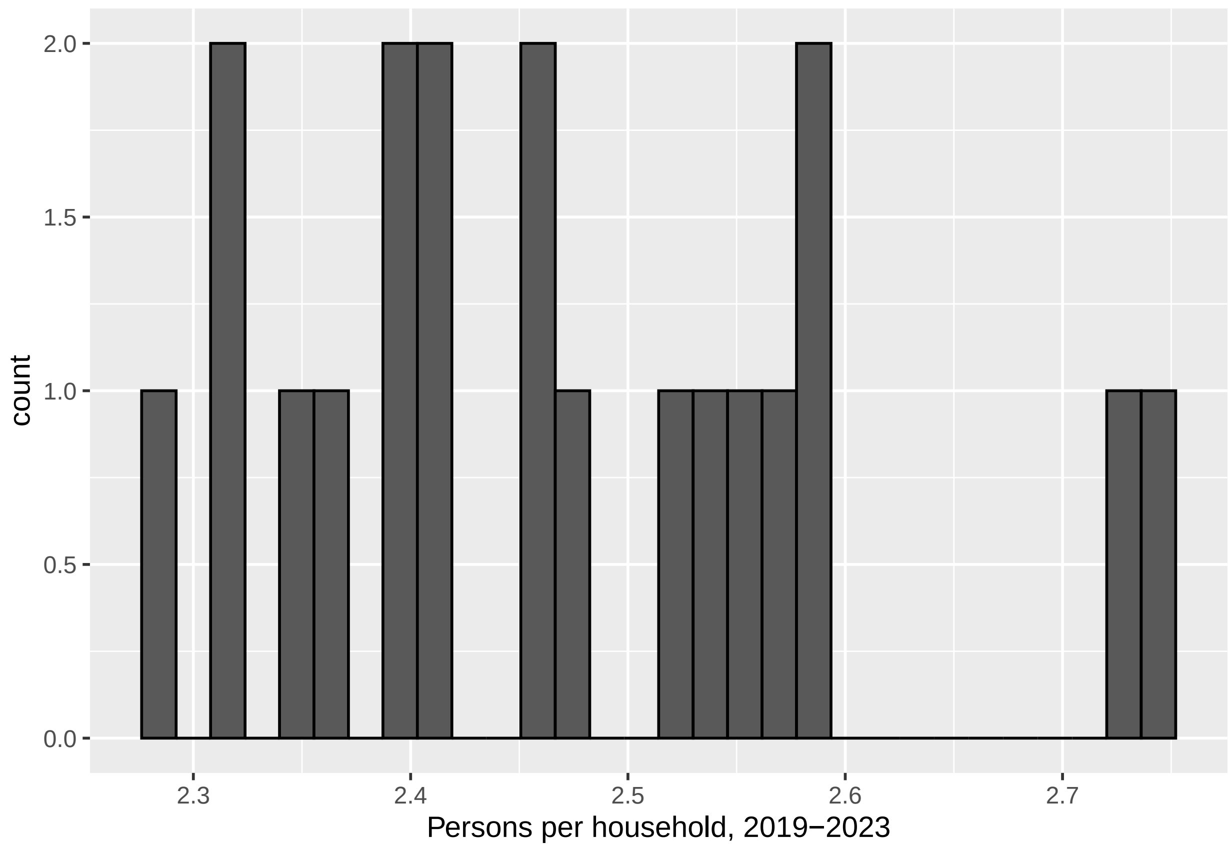
```
#Median value of owned homes
ggplot(combined_data, aes(x = `Median value of owner-occupied housing units, 2019-2023`)) +
geom_histogram(color = 'black')
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



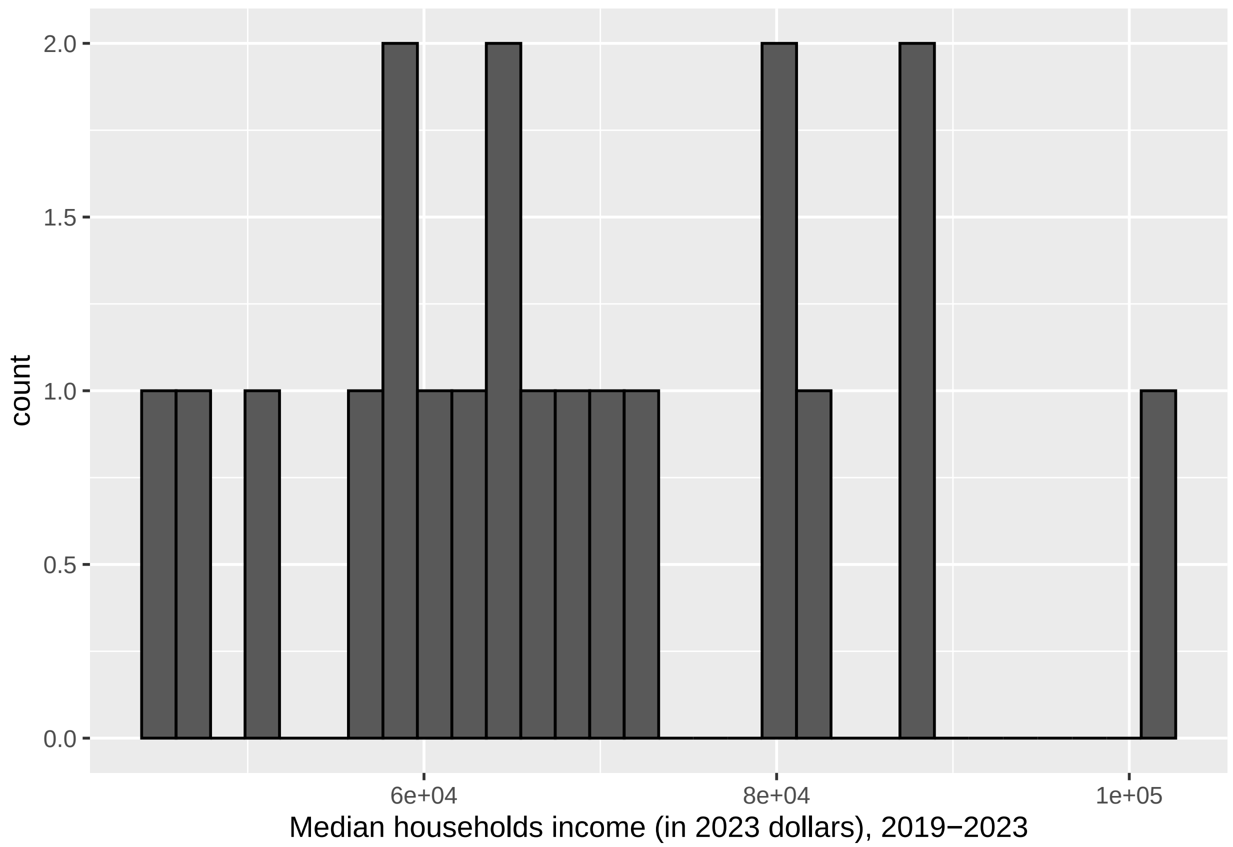
```
#Ppl/household
ggplot(combined_data, aes(x = `Persons per household, 2019-2023`)) +
geom_histogram(color = 'black')
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



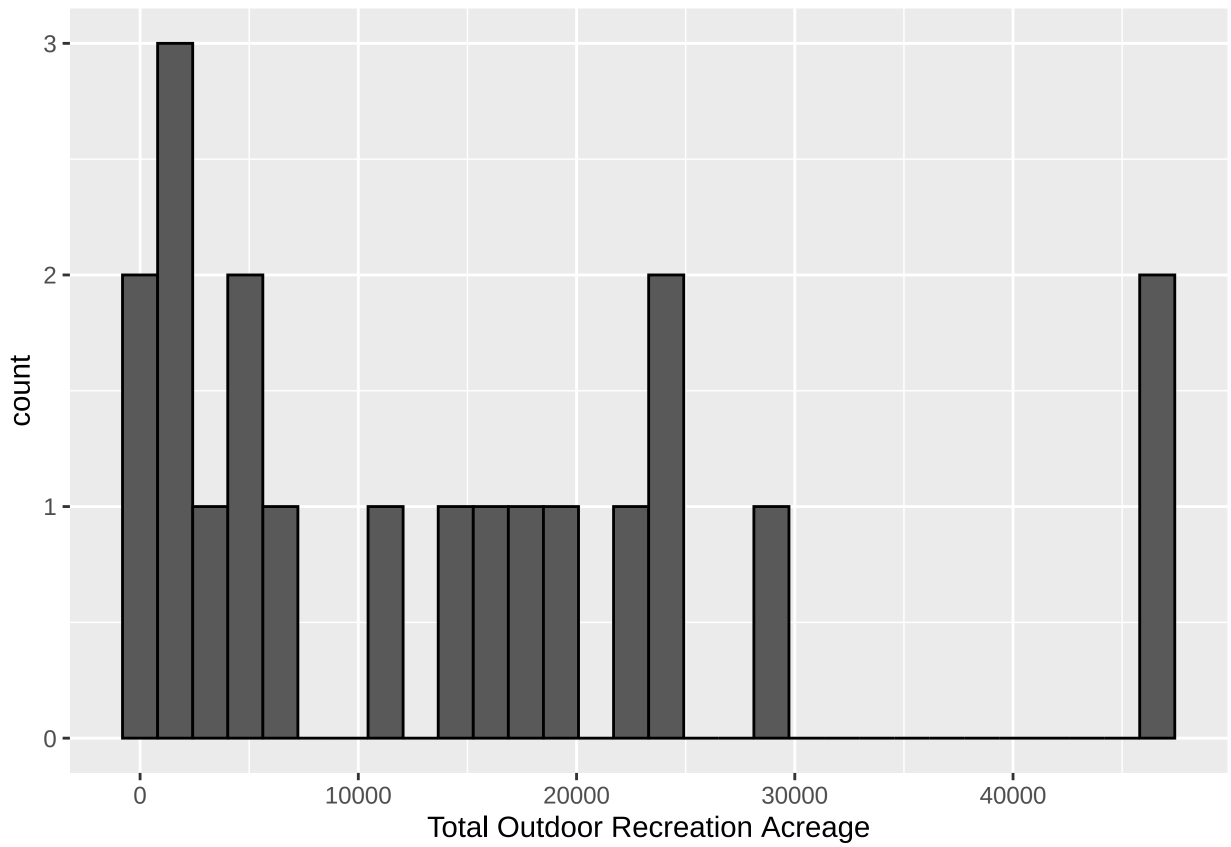
```
#Median household income
ggplot(combined_data, aes(x = `Median households income (in 2023 dollars), 2019-2023`)) +
geom_histogram(color = 'black')
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



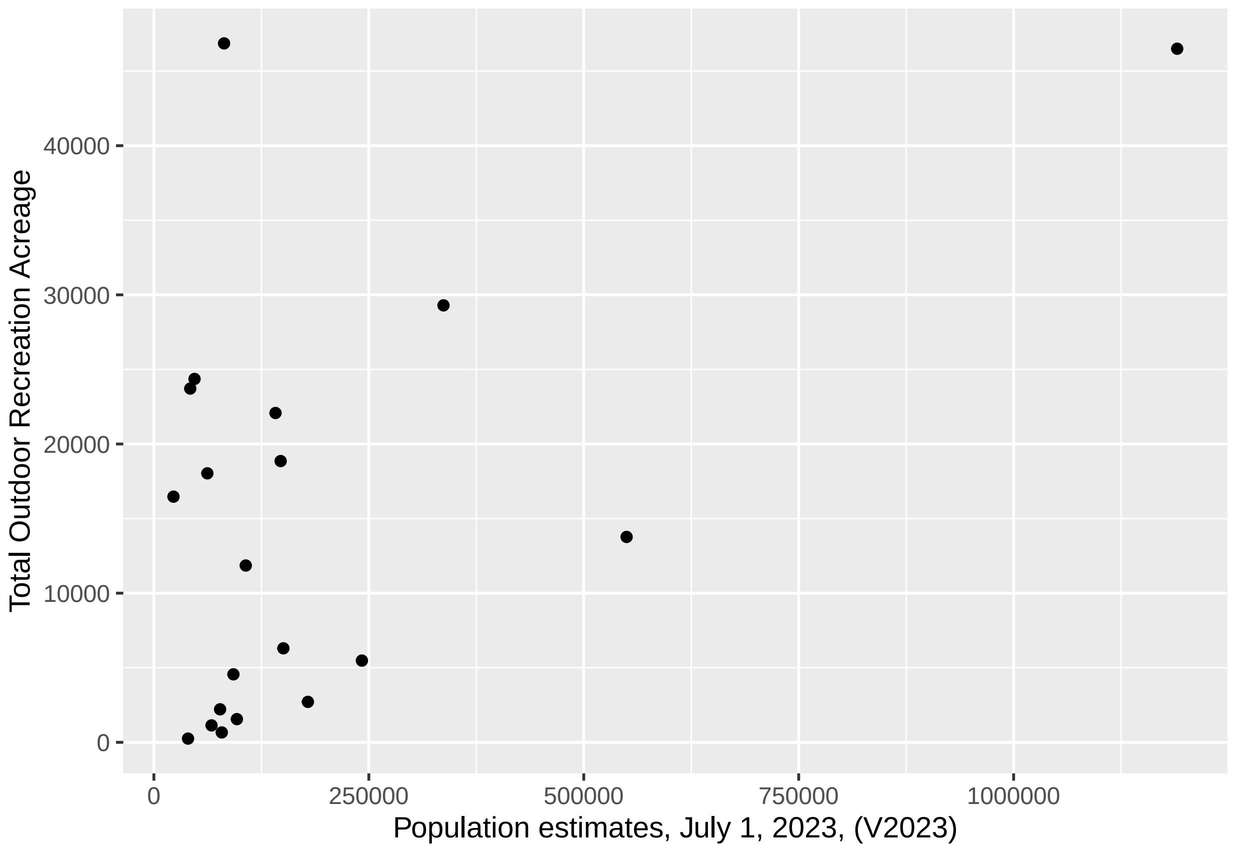
```
#Total outdoor recreation acreage
ggplot(combined_data, aes(x = `Total Outdoor Recreation Acreage`)) +
geom_histogram(color = 'black')
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

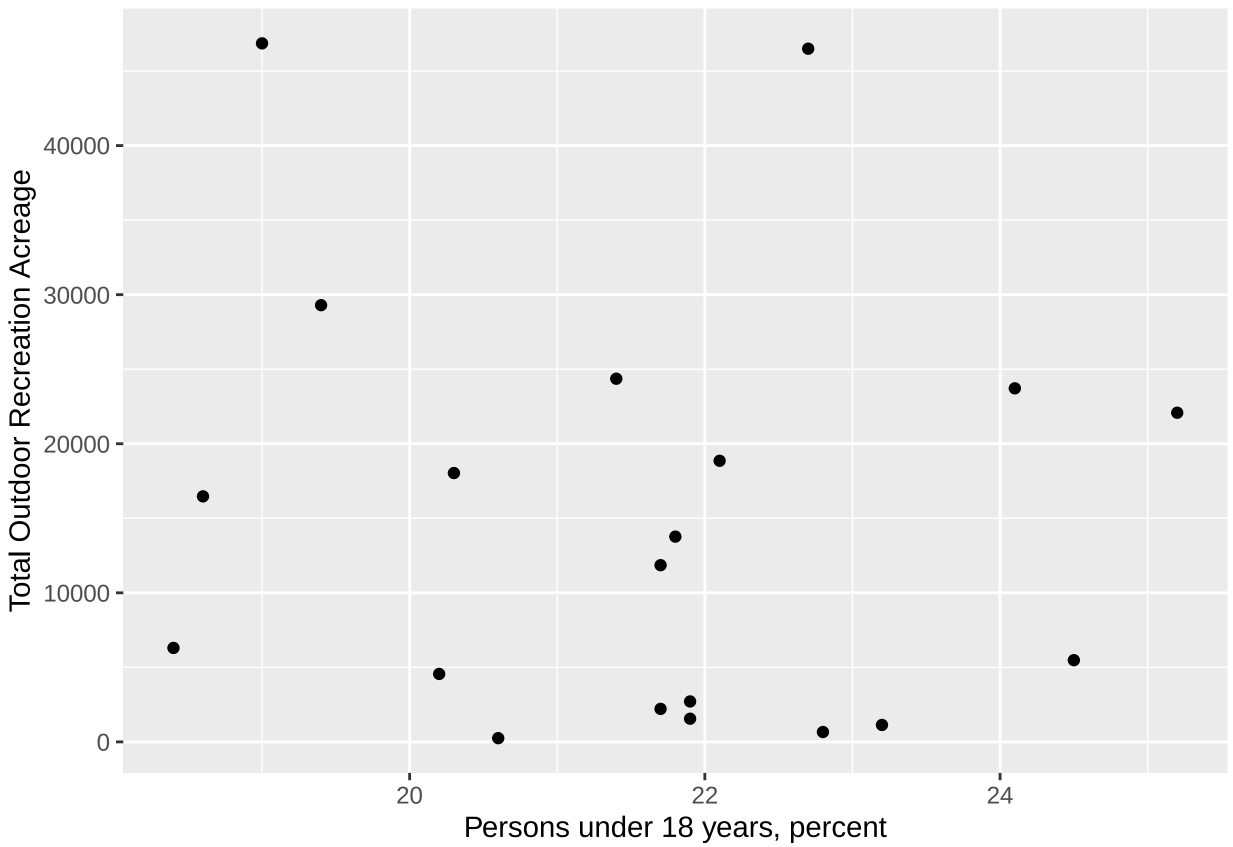


Look at relationships between independent variables and Total Outdoor Recreation Acreage

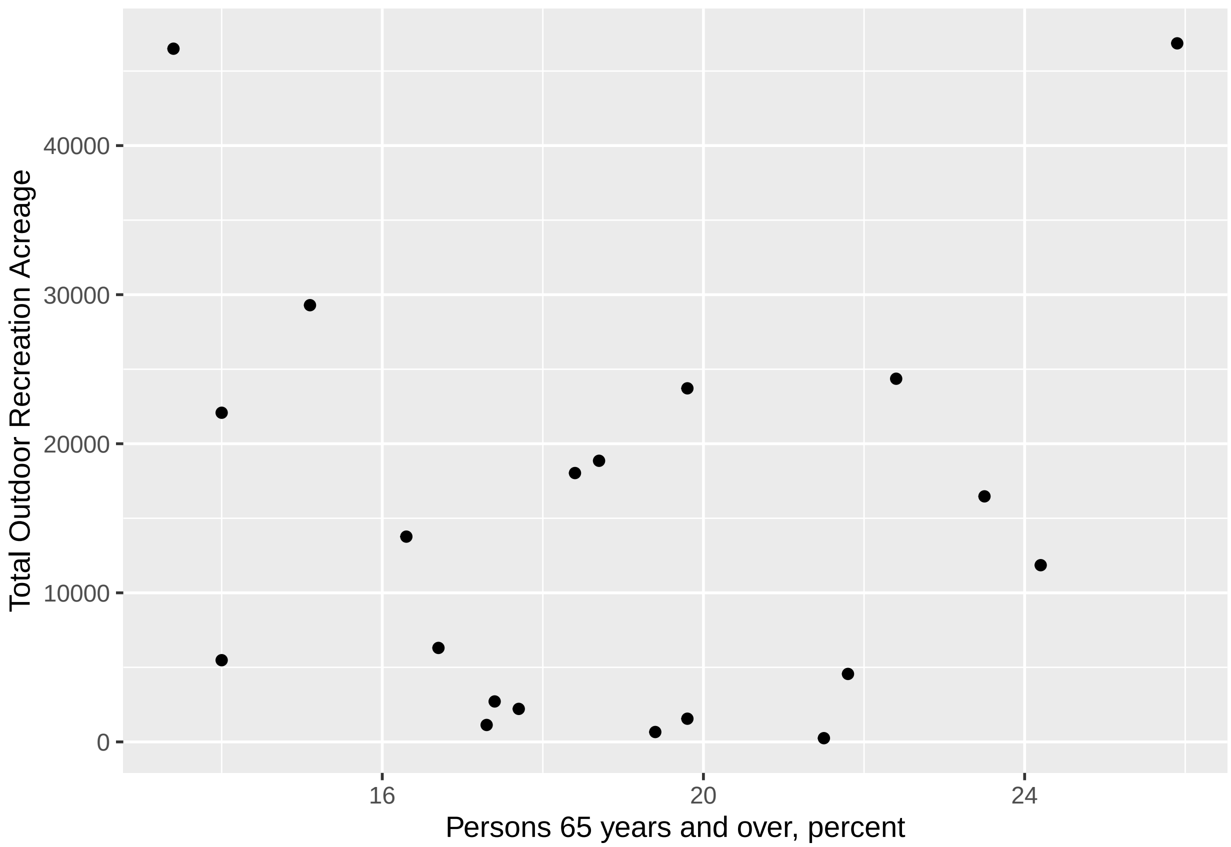
```
#Create scatterplots of recreation acreage by:

#pop estimates
ggplot(combined_data, aes(x = `Population estimates, July 1, 2023, (V2023)`,
                          y =`Total Outdoor Recreation Acreage`)) +
  geom_point()
```



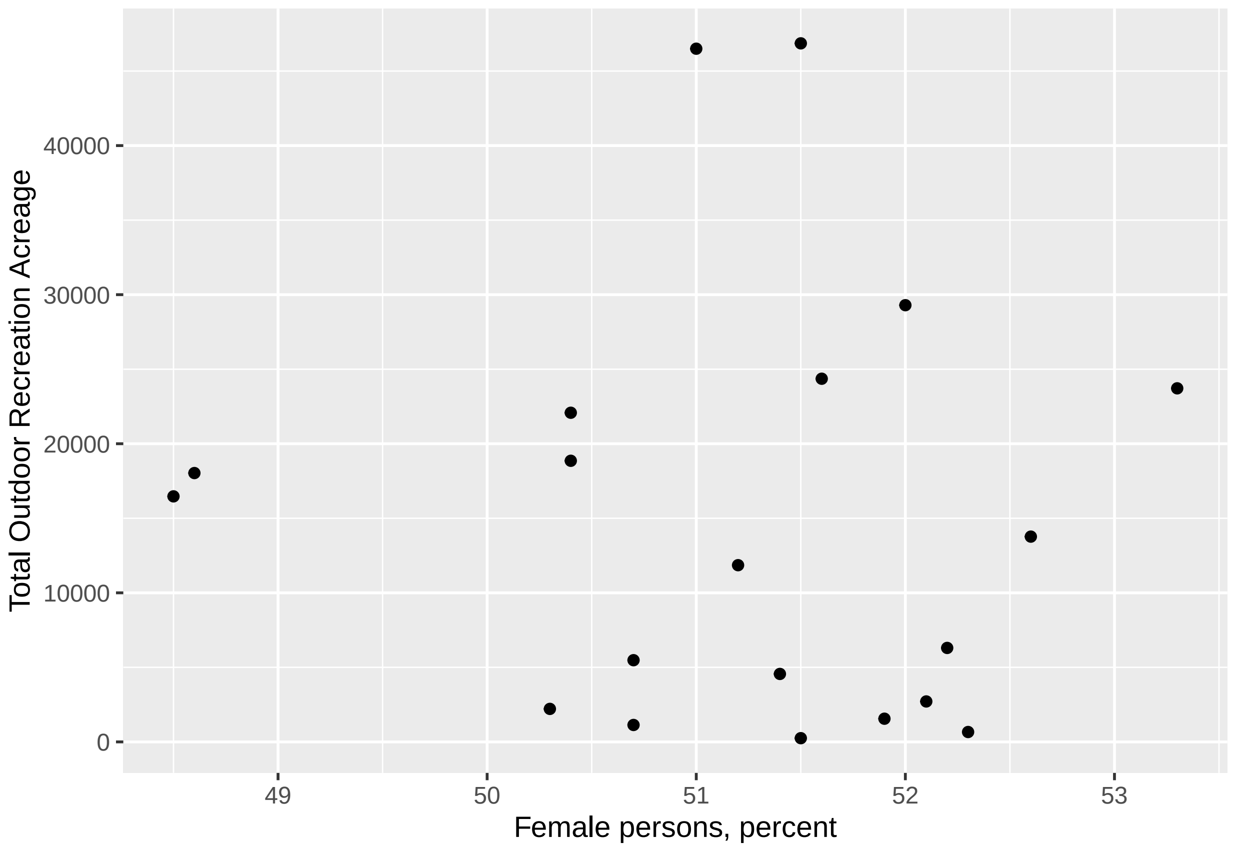
```r
#Ppl under 18
ggplot(combined_data, aes(x = `Persons under 18 years, percent`,
                          y =`Total Outdoor Recreation Acreage`)) +
  geom_point()
```



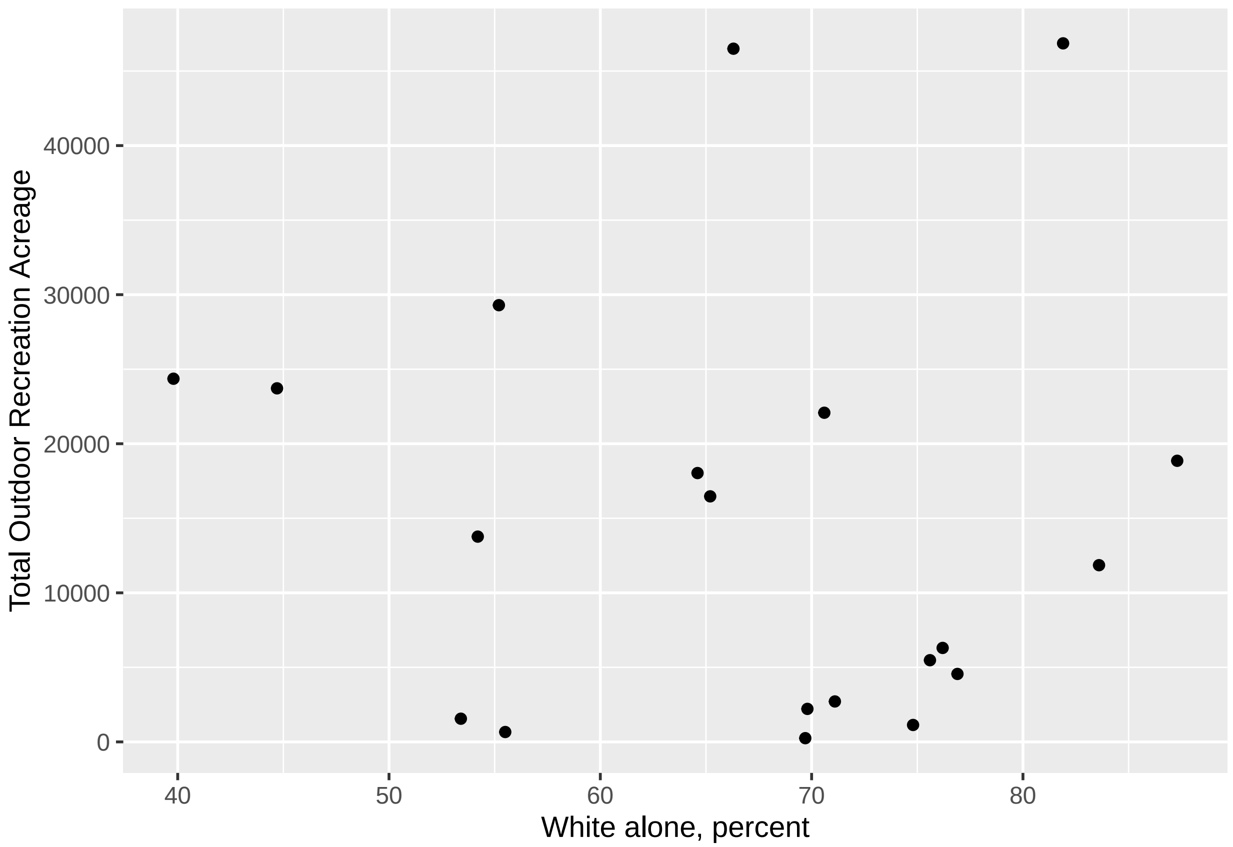
```
#Ppl 65 and older
ggplot(combined_data, aes(x = `Persons 65 years and over, percent`,
                          y =`Total Outdoor Recreation Acreage`)) +
  geom_point()
```



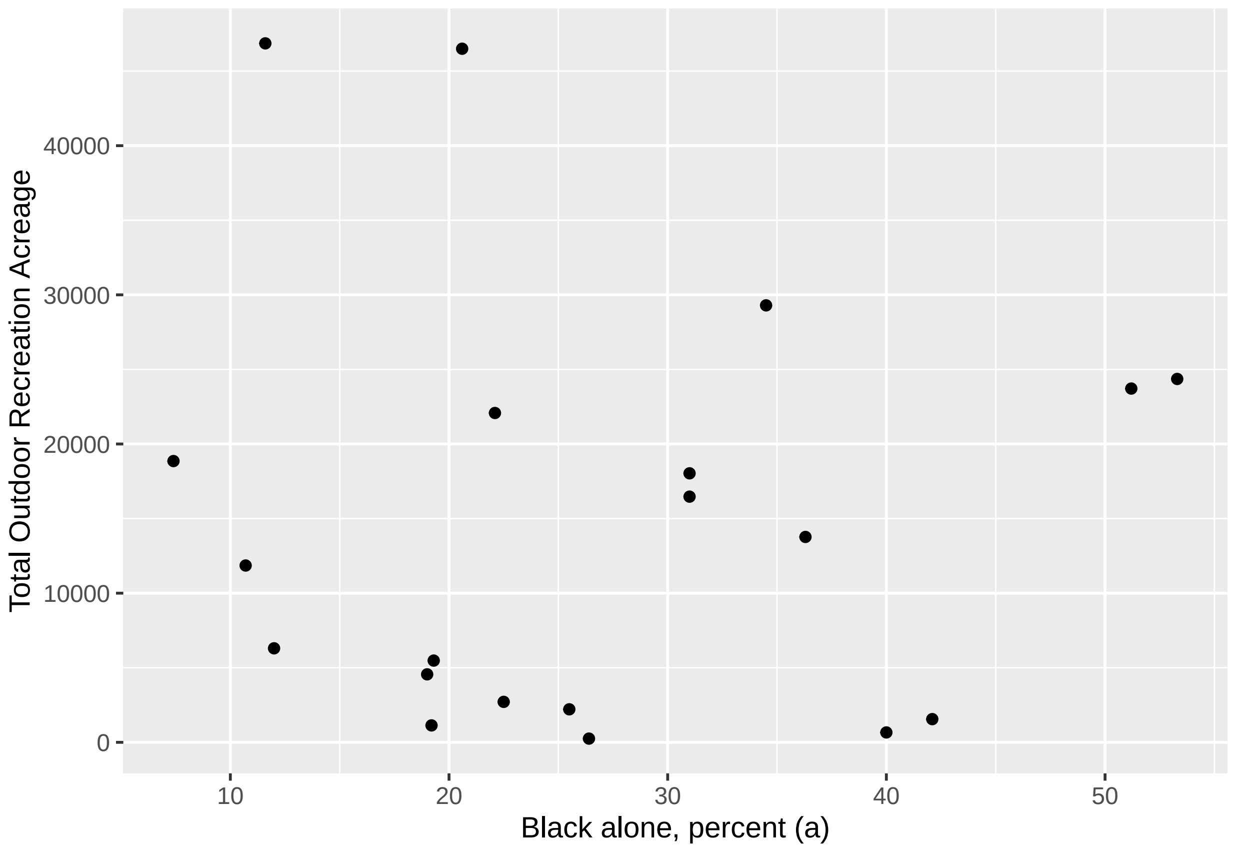
```
#XX ppl

ggplot(combined_data, aes(x = `Female persons, percent`,
                          y =`Total Outdoor Recreation Acreage`)) +
  geom_point()
```



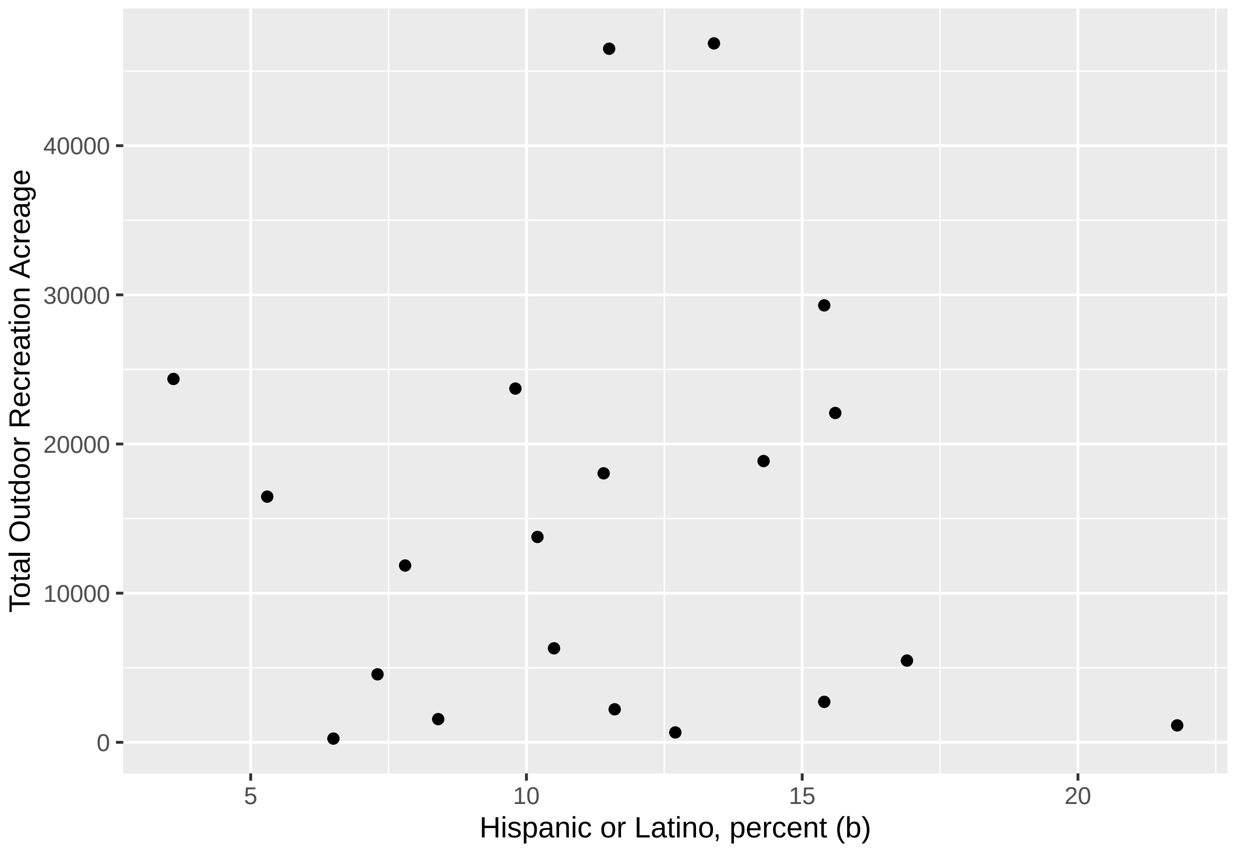
```
#White ppl
ggplot(combined_data, aes(x = `White alone, percent`,
                          y =`Total Outdoor Recreation Acreage`)) +
  geom_point()
```



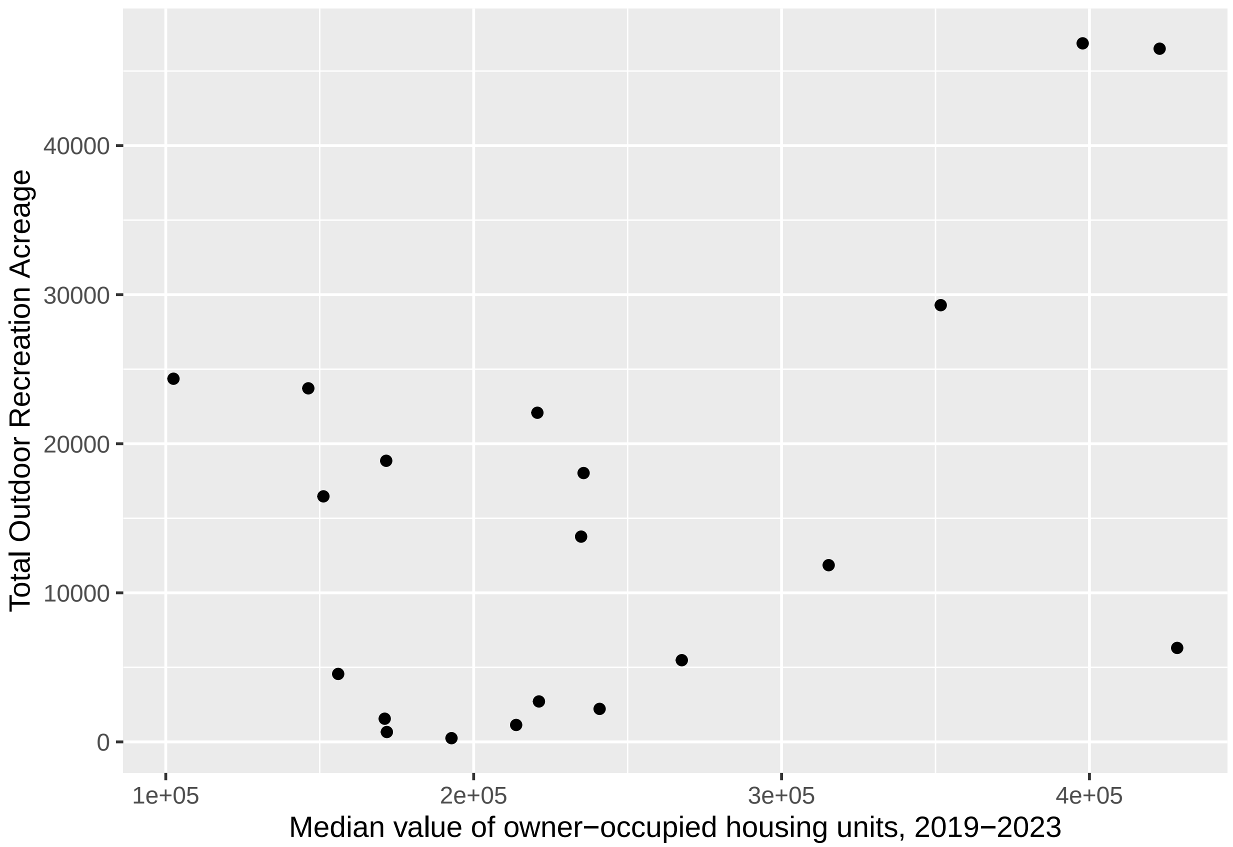
```
#Black ppl
ggplot(combined_data, aes(x = `Black alone, percent (a)`,
                          y =`Total Outdoor Recreation Acreage`)) +
  geom_point()
```



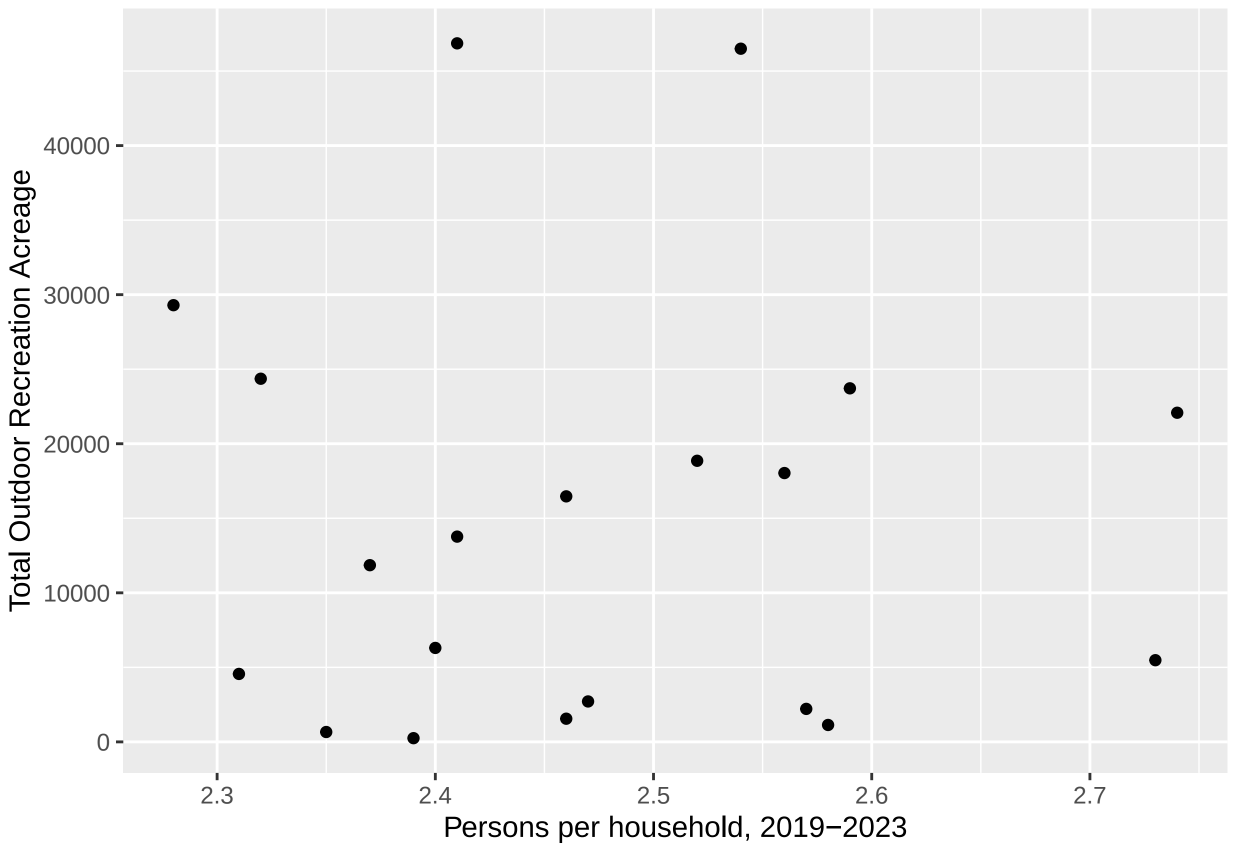
```r
#Hispanic or Latine ppl
ggplot(combined_data, aes(x = `Hispanic or Latino, percent (b)`,
                           y =`Total Outdoor Recreation Acreage`)) +
  geom_point()
```



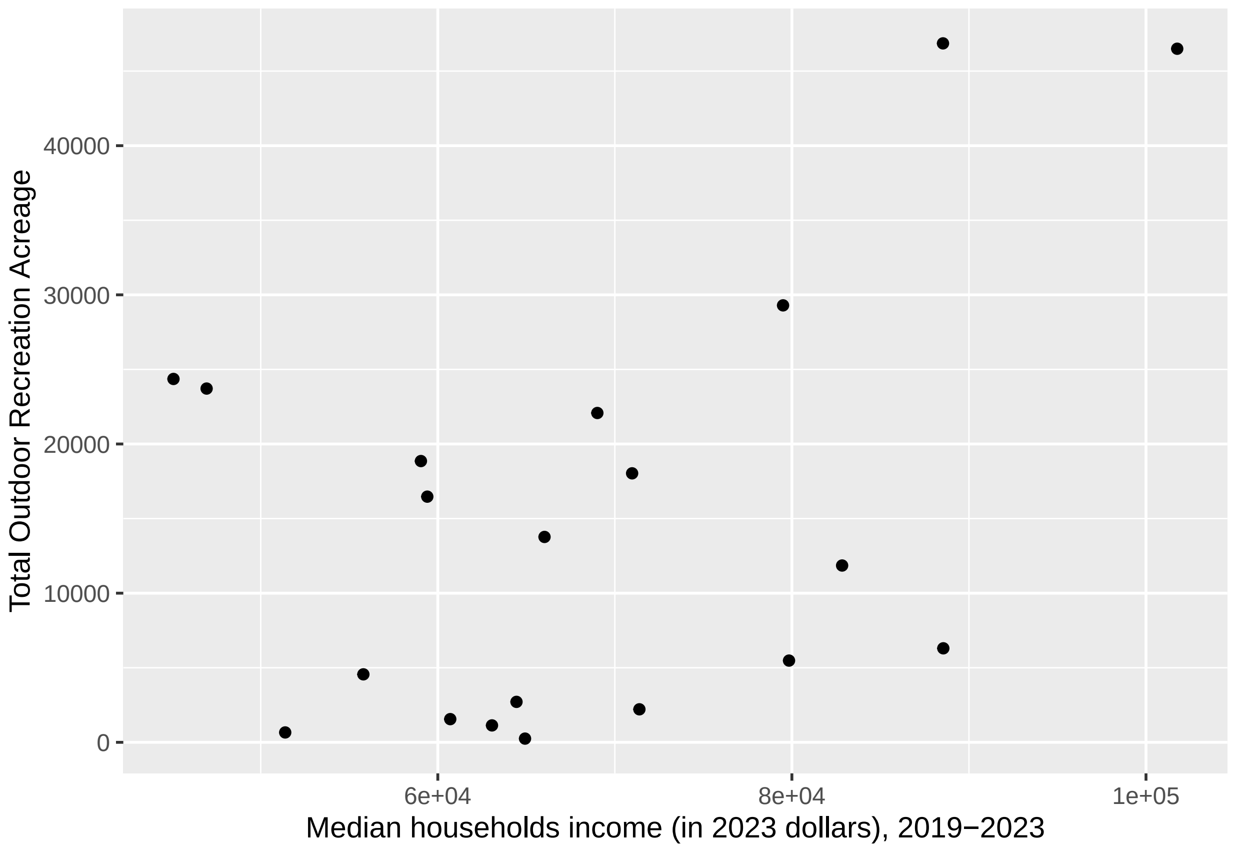
```
#Median value of owned homes
ggplot(combined_data, aes(x = `Median value of owner-occupied housing units, 2019-2023`,
                        y =`Total Outdoor Recreation Acreage`)) +
  geom_point()
```



```
#Ppl/household
ggplot(combined_data, aes(x = `Persons per household, 2019-2023`,
                          y =`Total Outdoor Recreation Acreage`)) +
  geom_point()
```



```
#Median household income
ggplot(combined_data, aes(x = `Median households income (in 2023 dollars), 2019-2023`,
                          y =`Total Outdoor Recreation Acreage`)) +
  geom_point()
```



Investigate multi-collinearity

```
#possible correlations between pop estimates and:

#Ppl under 18
cor.test(combined_data$`Population estimates, July 1, 2023, (V2023)`,
         combined_data$`Persons under 18 years, percent`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$‘Population estimates, July 1, 2023, (V2023)‘ and combined_data$‘Persons under
## t = 0.64339, df = 18, p-value = 0.5281
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3133781  0.5555926
## sample estimates:
##       cor
## 0.1499351
```

```
#Ppl 65 or older
cor.test(combined_data$`Population estimates, July 1, 2023, (V2023)`,
         combined_data$`Persons 65 years and over, percent`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Population estimates, July 1, 2023, (V2023)‘ and combined_data$‘Persons 65 yea
## t = -2.8216, df = 18, p-value = 0.0113
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8001965 -0.1473556
## sample estimates:
##        cor
## -0.5537664
```

```
#XX ppl
cor.test(combined_data$`Population estimates, July 1, 2023, (V2023)`,
         combined_data$`Female persons, percent`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Population estimates, July 1, 2023, (V2023)‘ and combined_data$‘Female persons
## t = 0.50097, df = 18, p-value = 0.6225
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3430580  0.5321701
## sample estimates:
##       cor
## 0.1172647
```

```
#White ppl
cor.test(combined_data$`Population estimates, July 1, 2023, (V2023)`,
         combined_data$`White alone, percent`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Population estimates, July 1, 2023, (V2023)‘ and combined_data$‘White alone, pe
## t = -0.21937, df = 18, p-value = 0.8288
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4831178  0.4000251
## sample estimates:
##         cor
## -0.05163636
```

```
#Black ppl
cor.test(combined_data$`Population estimates, July 1, 2023, (V2023)`,
         combined_data$`Black alone, percent (a)`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Population estimates, July 1, 2023, (V2023)‘ and combined_data$‘Black alone, pe
## t = -0.45229, df = 18, p-value = 0.6565
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5239485  0.3530775
## sample estimates:
##        cor
## -0.1060061
```

```
#Hispanic or Latine ppl
cor.test(combined_data$`Population estimates, July 1, 2023, (V2023)`,
         combined_data$`Hispanic or Latino, percent (b)`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Population estimates, July 1, 2023, (V2023)‘ and combined_data$‘Hispanic or La
## t = 0.50547, df = 18, p-value = 0.6194
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3421284  0.5329247
## sample estimates:
##       cor
## 0.1183034
```

```
#Median value of owned homes
cor.test(combined_data$`Population estimates, July 1, 2023, (V2023)`,
         combined_data$`Median value of owner-occupied housing units, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Population estimates, July 1, 2023, (V2023)‘ and combined_data$‘Median value o
## t = 2.6166, df = 18, p-value = 0.01748
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1073493 0.7850840
## sample estimates:
##       cor
## 0.5249335
```

```
#Ppl/household
cor.test(combined_data$`Population estimates, July 1, 2023, (V2023)`,
         combined_data$`Persons per household, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Population estimates, July 1, 2023, (V2023)‘ and combined_data$‘Persons per hou
## t = 0.30746, df = 18, p-value = 0.762
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3824742  0.4988452
## sample estimates:
##        cor
## 0.07228019
```

```
#Median household income
cor.test(combined_data$`Population estimates, July 1, 2023, (V2023)`,
         combined_data$`Median households income (in 2023 dollars), 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Population estimates, July 1, 2023, (V2023)‘ and combined_data$‘Median househol
## t = 3.0212, df = 18, p-value = 0.007339
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1850375 0.8137111
## sample estimates:
##       cor
## 0.5800612
```

```
#possible correlations between ppl under 18 and:

##Ppl 65 or older
cor.test(combined_data$`Persons under 18 years, percent`,
         combined_data$`Persons 65 years and over, percent`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons under 18 years, percent‘ and combined_data$‘Persons 65 years and over,
## t = -2.3136, df = 18, p-value = 0.03271
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.76022160 -0.04598527
## sample estimates:
##        cor
## -0.4787635
```

```
#XX ppl
cor.test(combined_data$`Persons under 18 years, percent`,
         combined_data$`Female persons, percent`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons under 18 years, percent‘ and combined_data$‘Female persons, percent‘
## t = 0.64765, df = 18, p-value = 0.5254
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3124835  0.5562779
## sample estimates:
##       cor
## 0.1509044
```

```r
#White ppl
cor.test(combined_data$`Persons under 18 years, percent`,
         combined_data$`White alone, percent`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons under 18 years, percent‘ and combined_data$‘White alone, percent‘
## t = -0.53519, df = 18, p-value = 0.5991
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5378842  0.3359750
## sample estimates:
##       cor
## -0.125153
```

```r
#Black ppl
cor.test(combined_data$`Persons under 18 years, percent`,
         combined_data$`Black alone, percent (a)`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons under 18 years, percent‘ and combined_data$‘Black alone, percent (a)‘
## t = 0.64451, df = 18, p-value = 0.5274
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3131442  0.5557720
## sample estimates:
##       cor
## 0.1501887
```

```r
#Hispanic or Latine ppl
cor.test(combined_data$`Persons under 18 years, percent`,
         combined_data$`Hispanic or Latino, percent (b)`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons under 18 years, percent‘ and combined_data$‘Hispanic or Latino, percen
## t = 1.933, df = 18, p-value = 0.06913
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03419427  0.72424438
## sample estimates:
##       cor
## 0.4146001
```

```r
#Median value of owned homes
cor.test(combined_data$`Persons under 18 years, percent`,
         combined_data$`Median value of owner-occupied housing units, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons under 18 years, percent‘ and combined_data$‘Median value of owner-occup
## t = -1.2768, df = 18, p-value = 0.2179
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6480518  0.1769086
## sample estimates:
##        cor
## -0.2881719
```

```
#Ppl/household
cor.test(combined_data$`Persons under 18 years, percent`,
         combined_data$`Persons per household, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons under 18 years, percent‘ and combined_data$‘Persons per household, 2019
## t = 3.9391, df = 18, p-value = 0.0009619
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3403606 0.8630616
## sample estimates:
##       cor
## 0.6804016
```

```
#Median household income
cor.test(combined_data$`Persons under 18 years, percent`,
         combined_data$`Median households income (in 2023 dollars), 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons under 18 years, percent‘ and combined_data$‘Median households income (i
## t = -0.88645, df = 18, p-value = 0.3871
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5933416  0.2616833
## sample estimates:
##        cor
## -0.2045211
```

```
#possible correlations between ppl 65 and older and:
#XX ppl
cor.test(combined_data$`Persons 65 years and over, percent`,
         combined_data$`Female persons, percent`)
```

```
##
##  Pearson’s product-moment correlation
##
```

```
## data:  combined_data$‘Persons 65 years and over, percent‘ and combined_data$‘Female persons, percent
## t = -0.20447, df = 18, p-value = 0.8403
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4804254  0.4029661
## sample estimates:
##         cor
## -0.04813882
```

```
#White ppl
cor.test(combined_data$`Persons 65 years and over, percent`,
         combined_data$`White alone, percent`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons 65 years and over, percent‘ and combined_data$‘White alone, percent‘
## t = 0.31206, df = 18, p-value = 0.7586
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3815523  0.4996556
## sample estimates:
##        cor
## 0.07335396
```

```
#Black ppl
cor.test(combined_data$`Persons 65 years and over, percent`,
         combined_data$`Black alone, percent (a)`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons 65 years and over, percent‘ and combined_data$‘Black alone, percent (a
## t = 0.1617, df = 18, p-value = 0.8733
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4113680  0.4726409
## sample estimates:
##        cor
## 0.03808589
```

```
#Hispanic or Latine ppl
cor.test(combined_data$`Persons 65 years and over, percent`,
         combined_data$`Hispanic or Latino, percent (b)`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons 65 years and over, percent‘ and combined_data$‘Hispanic or Latino, perc
## t = -2.9794, df = 18, p-value = 0.008037
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.8109760 -0.1772592
## sample estimates:
##     cor
## -0.5747
```

```
#Median value of owned homes
cor.test(combined_data$`Persons 65 years and over, percent`,
         combined_data$`Median value of owner-occupied housing units, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons 65 years and over, percent‘ and combined_data$‘Median value of owner-oc
## t = -1.1785, df = 18, p-value = 0.2539
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6349624  0.1983697
## sample estimates:
##        cor
## -0.2676458
```

```
#Ppl/household
cor.test(combined_data$`Persons 65 years and over, percent`,
         combined_data$`Persons per household, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons 65 years and over, percent‘ and combined_data$‘Persons per household,
## t = -2.5165, df = 18, p-value = 0.02155
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.77721735 -0.08735831
## sample estimates:
##        cor
## -0.5101575
```

```
#Median household income
cor.test(combined_data$`Persons 65 years and over, percent`,
         combined_data$`Median households income (in 2023 dollars), 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons 65 years and over, percent‘ and combined_data$‘Median households income
## t = -1.3296, df = 18, p-value = 0.2003
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6548999  0.1653591
## sample estimates:
##        cor
## -0.2990442
```

```r
#possible correlations between XX ppl and :
#White ppl
cor.test(combined_data$`Female persons, percent`,
         combined_data$`White alone, percent`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$‘Female persons, percent‘ and combined_data$‘White alone, percent‘
## t = -1.6326, df = 18, p-value = 0.1199
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.69172455  0.09914043
## sample estimates:
##        cor
## -0.3591363
```

```r
#Black ppl
cor.test(combined_data$`Female persons, percent`,
         combined_data$`Black alone, percent (a)`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$‘Female persons, percent‘ and combined_data$‘Black alone, percent (a)‘
## t = 1.3185, df = 18, p-value = 0.2039
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1677834  0.6534736
## sample estimates:
##      cor
## 0.296772
```

```r
#Hispanic or Latine ppl
cor.test(combined_data$`Female persons, percent`,
         combined_data$`Hispanic or Latino, percent (b)`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$‘Female persons, percent‘ and combined_data$‘Hispanic or Latino, percent (b)‘
## t = -0.084022, df = 18, p-value = 0.934
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4583054  0.4264571
## sample estimates:
##         cor
## -0.01980032
```

```r
#Median value of owned homes
cor.test(combined_data$`Female persons, percent`,
         combined_data$`Median value of owner-occupied housing units, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Female persons, percent‘ and combined_data$‘Median value of owner-occupied hous
## t = 0.34685, df = 18, p-value = 0.7327
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3745441  0.5057659
## sample estimates:
##        cor
## 0.08148171
```

```
#Ppl/household
cor.test(combined_data$`Female persons, percent`,
         combined_data$`Persons per household, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Female persons, percent‘ and combined_data$‘Persons per household, 2019-2023‘
## t = -1.6561, df = 18, p-value = 0.115
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.69441293  0.09401528
## sample estimates:
##       cor
## -0.363634
```

```
#Median household income
cor.test(combined_data$`Female persons, percent`,
         combined_data$`Median households income (in 2023 dollars), 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Female persons, percent‘ and combined_data$‘Median households income (in 2023
## t = -0.52742, df = 18, p-value = 0.6043
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5365918  0.3375858
## sample estimates:
##        cor
## -0.1233642
```

```
#possible correlations between white ppl and:
#Black ppl
cor.test(combined_data$`White alone, percent`,
         combined_data$`Black alone, percent (a)`)
```

```
##
##  Pearson’s product-moment correlation
##
```

```
## data:  combined_data$‘White alone, percent‘ and combined_data$‘Black alone, percent (a)‘
## t = -19.864, df = 18, p-value = 1.084e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9914176 -0.9439073
## sample estimates:
##        cor
## -0.9779428
```

```
#Hispanic or Latine ppl
cor.test(combined_data$`White alone, percent`,
         combined_data$`Hispanic or Latino, percent (b)`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘White alone, percent‘ and combined_data$‘Hispanic or Latino, percent (b)‘
## t = 1.606, df = 18, p-value = 0.1257
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1049347  0.6886585
## sample estimates:
##       cor
## 0.3540255
```

```
#Median value of owned homes
cor.test(combined_data$`White alone, percent`,
         combined_data$`Median value of owner-occupied housing units, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘White alone, percent‘ and combined_data$‘Median value of owner-occupied housin
## t = 1.8271, df = 18, p-value = 0.08432
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.05696813  0.71321264
## sample estimates:
##       cor
## 0.3955236
```

```
#Ppl/household
cor.test(combined_data$`White alone, percent`,
         combined_data$`Persons per household, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘White alone, percent‘ and combined_data$‘Persons per household, 2019-2023‘
## t = 0.82662, df = 18, p-value = 0.4193
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.2745112  0.5843125
## sample estimates:
##      cor
## 0.191241
```

```r
#Median household income
cor.test(combined_data$`White alone, percent`,
         combined_data$`Median households income (in 2023 dollars), 2019-2023`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$‘White alone, percent‘ and combined_data$‘Median households income (in 2023 doll
## t = 2.3572, df = 18, p-value = 0.02994
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05496442 0.76399515
## sample estimates:
##       cor
## 0.4856722
```

```r
#Possible correlations between black ppl and:
#Hispanic or Latine ppl
cor.test(combined_data$`Black alone, percent (a)`,
         combined_data$`Hispanic or Latino, percent (b)`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$‘Black alone, percent (a)‘ and combined_data$‘Hispanic or Latino, percent (b)‘
## t = -1.7723, df = 18, p-value = 0.09327
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.70732413  0.06880487
## sample estimates:
##       cor
## -0.385452
```

```r
#Median value of owned homes
cor.test(combined_data$`Black alone, percent (a)`,
         combined_data$`Median value of owner-occupied housing units, 2019-2023`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$‘Black alone, percent (a)‘ and combined_data$‘Median value of owner-occupied hou
## t = -2.7401, df = 18, p-value = 0.01345
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7943457 -0.1316095
## sample estimates:
##        cor
## -0.5425332
```

```r
#Ppl/household
cor.test(combined_data$`Black alone, percent (a)`,
         combined_data$`Persons per household, 2019-2023`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$‘Black alone, percent (a)‘ and combined_data$‘Persons per household, 2019-2023‘
## t = -0.69984, df = 18, p-value = 0.493
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5646105  0.3014799
## sample estimates:
##        cor
## -0.1627541
```

```r
#Median household income
cor.test(combined_data$`Black alone, percent (a)`,
         combined_data$`Median households income (in 2023 dollars), 2019-2023`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$‘Black alone, percent (a)‘ and combined_data$‘Median households income (in 2023
## t = -3.2957, df = 18, p-value = 0.004019
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8305239 -0.2346461
## sample estimates:
##       cor
## -0.613467
```

```r
#possible correlations between Hispanic or Latine ppl and:
#Median value of owned homes
cor.test(combined_data$`Hispanic or Latino, percent (b)`,
         combined_data$`Median value of owner-occupied housing units, 2019-2023`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$‘Hispanic or Latino, percent (b)‘ and combined_data$‘Median value of owner-occu
## t = 1.294, df = 18, p-value = 0.212
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1731480  0.6502963
## sample estimates:
##       cor
## 0.2917252
```

```r
#Ppl/household
cor.test(combined_data$`Hispanic or Latino, percent (b)`,
         combined_data$`Persons per household, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Hispanic or Latino, percent (b)‘ and combined_data$‘Persons per household, 201
## t = 2.4683, df = 18, p-value = 0.02383
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07761583 0.77330380
## sample estimates:
##       cor
## 0.5028649
```

```
#Median household income
cor.test(combined_data$`Hispanic or Latino, percent (b)`,
         combined_data$`Median households income (in 2023 dollars), 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Hispanic or Latino, percent (b)‘ and combined_data$‘Median households income (
## t = 1.1732, df = 18, p-value = 0.256
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1995364  0.6342368
## sample estimates:
##       cor
## 0.2665176
```

```
#Possible correlations between median owned homes value and:
#Ppl/household
cor.test(combined_data$`Median value of owner-occupied housing units, 2019-2023`,
         combined_data$`Persons per household, 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Median value of owner-occupied housing units, 2019-2023‘ and combined_data$‘Pe
## t = -0.11967, df = 18, p-value = 0.9061
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4649151  0.4195606
## sample estimates:
##         cor
## -0.02819495
```

```
#Median household income
cor.test(combined_data$`Median value of owner-occupied housing units, 2019-2023`,
         combined_data$`Median households income (in 2023 dollars), 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
```

```
## data:  combined_data$‘Median value of owner-occupied housing units, 2019-2023‘ and combined_data$‘Me
## t = 13.13, df = 18, p-value = 1.169e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8792964 0.9809970
## sample estimates:
##      cor
## 0.951559
```

```
#possible correlations between ppl/household and:
#median household income
cor.test(combined_data$`Persons per household, 2019-2023`,
         combined_data$`Median households income (in 2023 dollars), 2019-2023`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  combined_data$‘Persons per household, 2019-2023‘ and combined_data$‘Median households income
## t = 0.51009, df = 18, p-value = 0.6162
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3411732  0.5336985
## sample estimates:
##       cor
## 0.1193696
```

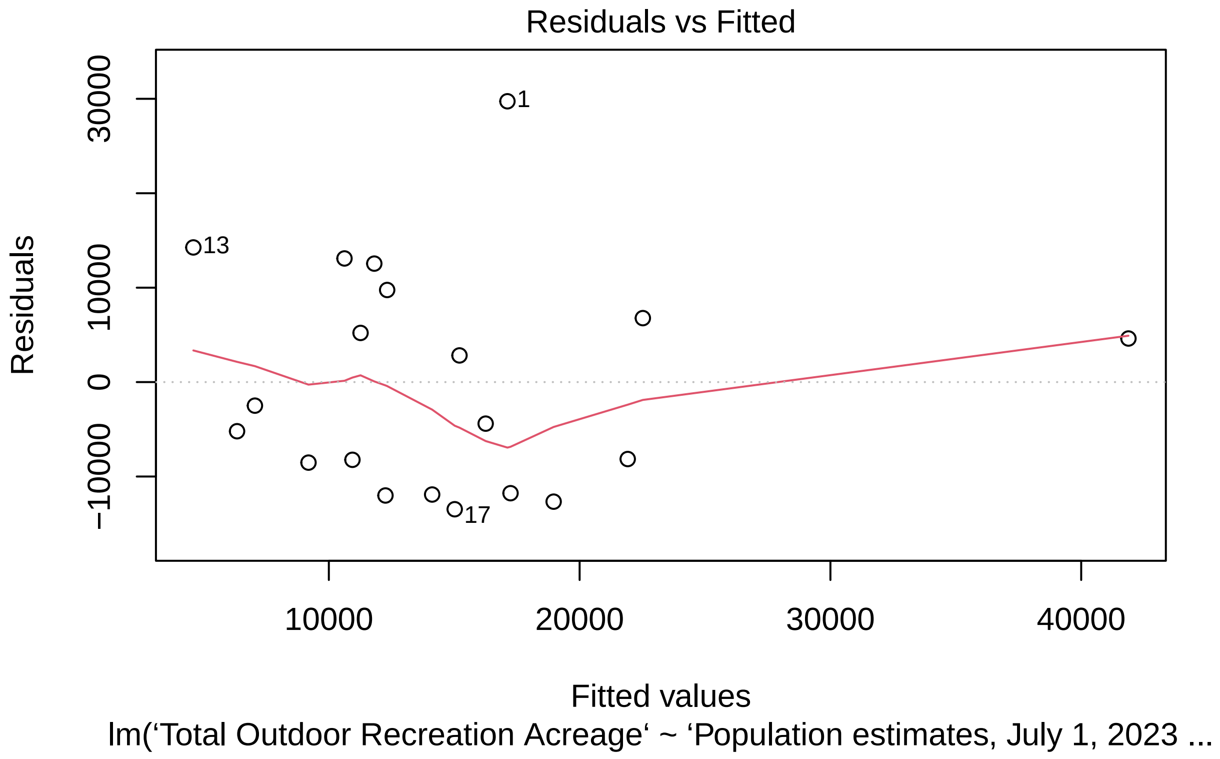
Step 3 - Fit regression model

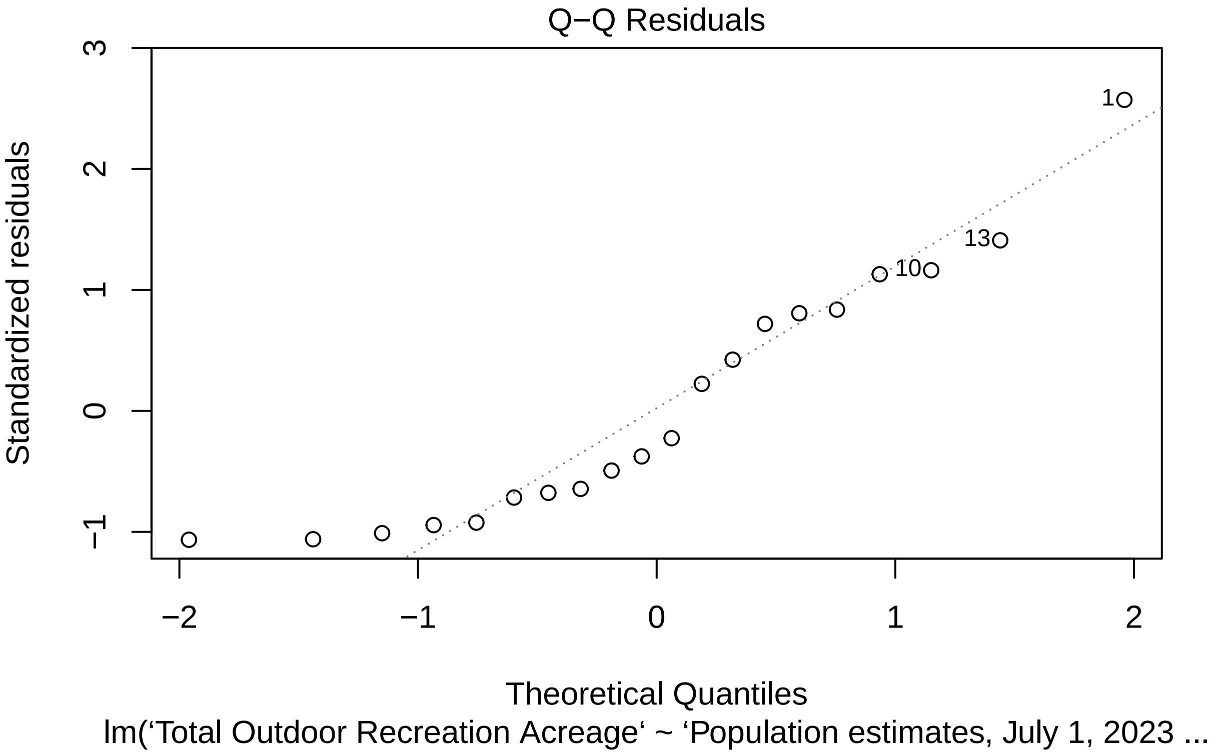
```
#fit first regression model
Model1 <- lm(`Total Outdoor Recreation Acreage` ~
     `Population estimates, July 1, 2023, (V2023)`+
     `Persons per household, 2019-2023` +
     `Black alone, percent (a)` +
     `Hispanic or Latino, percent (b)` +
     `Median households income (in 2023 dollars), 2019-2023`,
      data = combined_data)

#examine model
summary(Model1)
```

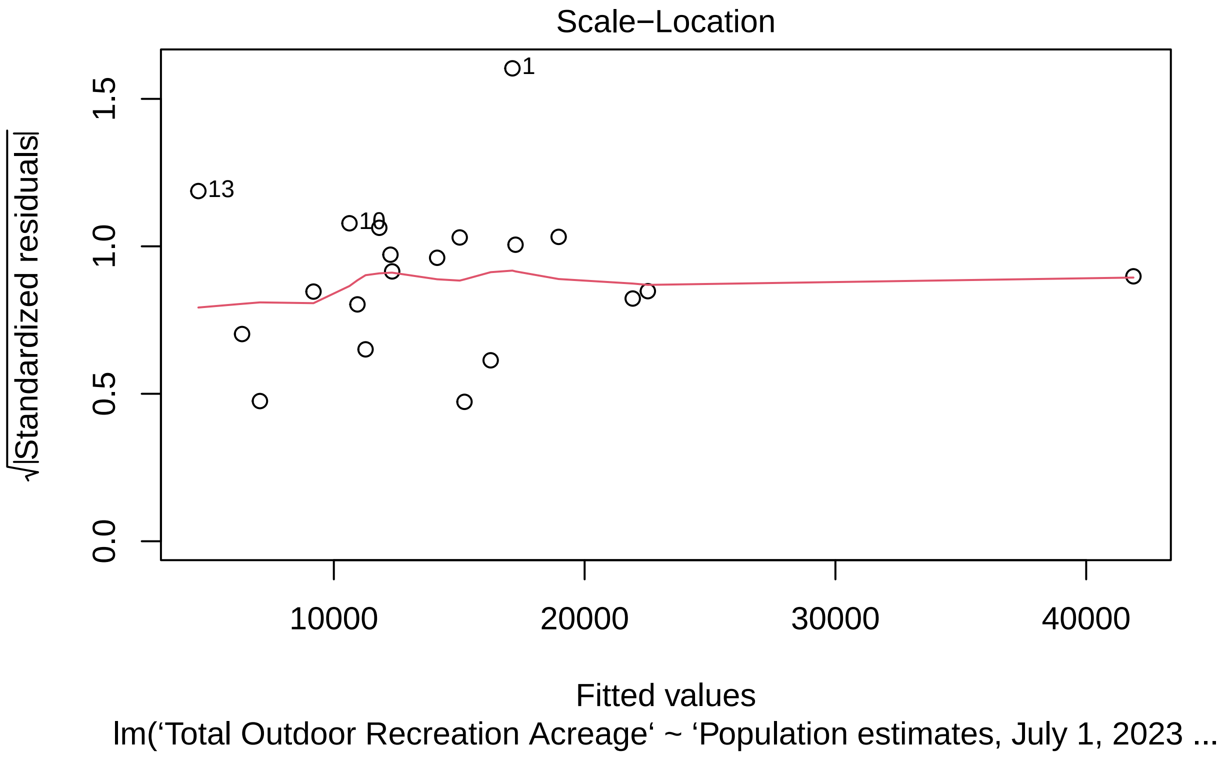
```
##
## Call:
## lm(formula = ‘Total Outdoor Recreation Acreage‘ ~ ‘Population estimates, July 1, 2023, (V2023)‘ +
##     ‘Persons per household, 2019-2023‘ + ‘Black alone, percent (a)‘ +
##     ‘Hispanic or Latino, percent (b)‘ + ‘Median households income (in 2023 dollars), 2019-2023‘,
##     data = combined_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13463  -9336  -3447   7518  29737
##
```

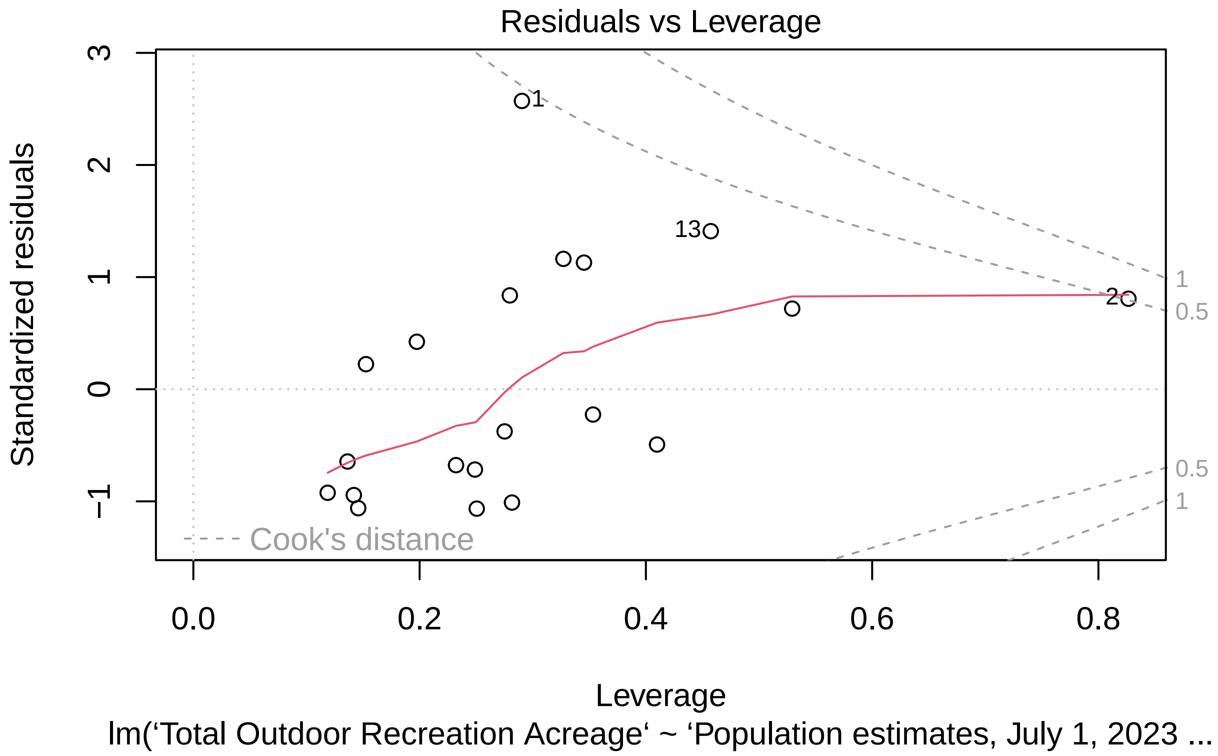
```
## Coefficients:
##                                                             Estimate Std. Error
## (Intercept)                                             -2.232e+04  7.202e+04
## ‘Population estimates, July 1, 2023, (V2023)‘            1.475e-02  1.581e-02
## ‘Persons per household, 2019-2023‘                       7.244e+02  2.812e+04
## ‘Black alone, percent (a)‘                               2.611e+02  3.477e+02
## ‘Hispanic or Latino, percent (b)‘                       -2.485e+02  8.884e+02
## ‘Median households income (in 2023 dollars), 2019-2023‘  4.156e-01  3.578e-01
##                                                         t value Pr(>|t|)
## (Intercept)                                              -0.310    0.761
## ‘Population estimates, July 1, 2023, (V2023)‘             0.933    0.367
## ‘Persons per household, 2019-2023‘                        0.026    0.980
## ‘Black alone, percent (a)‘                                0.751    0.465
## ‘Hispanic or Latino, percent (b)‘                        -0.280    0.784
## ‘Median households income (in 2023 dollars), 2019-2023‘   1.162    0.265
##
## Residual standard error: 13730 on 14 degrees of freedom
## Multiple R-squared:  0.3141, Adjusted R-squared:  0.06918
## F-statistic: 1.282 on 5 and 14 DF,  p-value: 0.3256
```

```r
#examine model residuals
plot(Model1)
```




Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
−1
0
1
2
1
13
10
Theoretical Quantiles
lm(‘Total Outdoor Recreation Acreage‘ ~ ‘Population estimates, July 1, 2023 ...

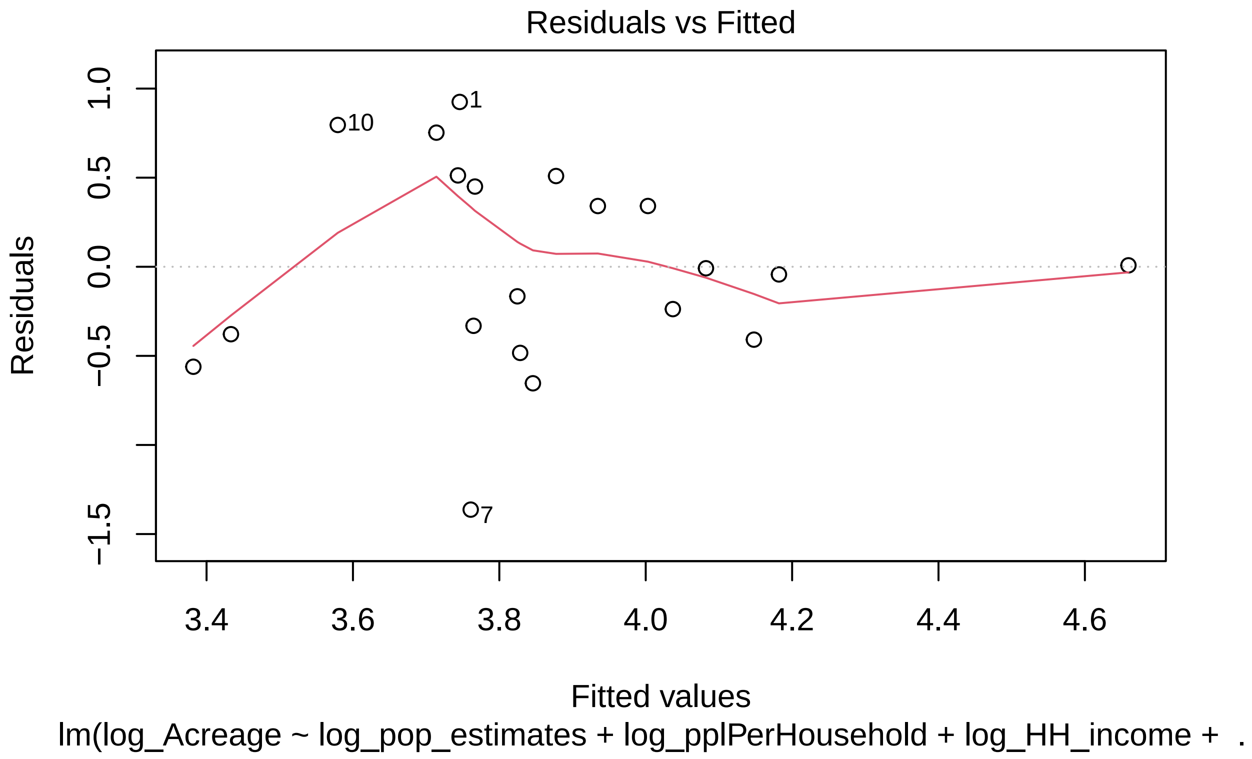

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10000
20000
30000
40000
1
13
10
Fitted values
lm(‘Total Outdoor Recreation Acreage‘ ~ ‘Population estimates, July 1, 2023 ...

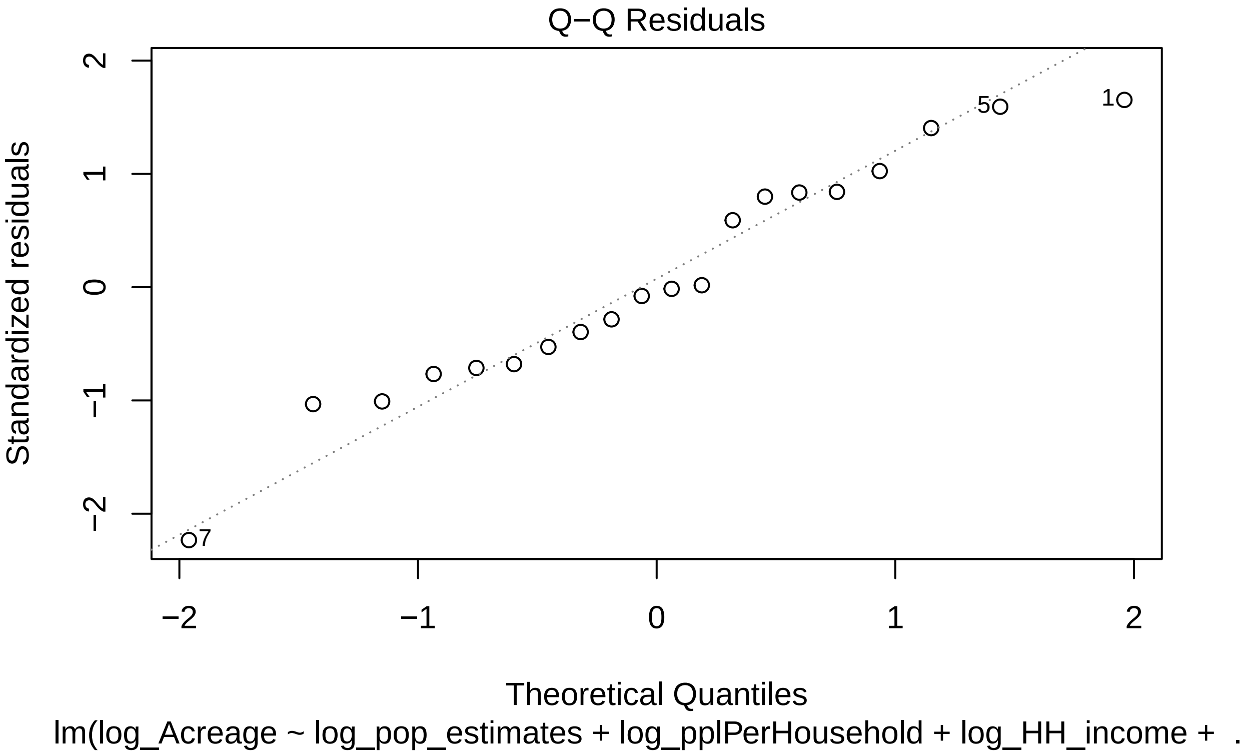


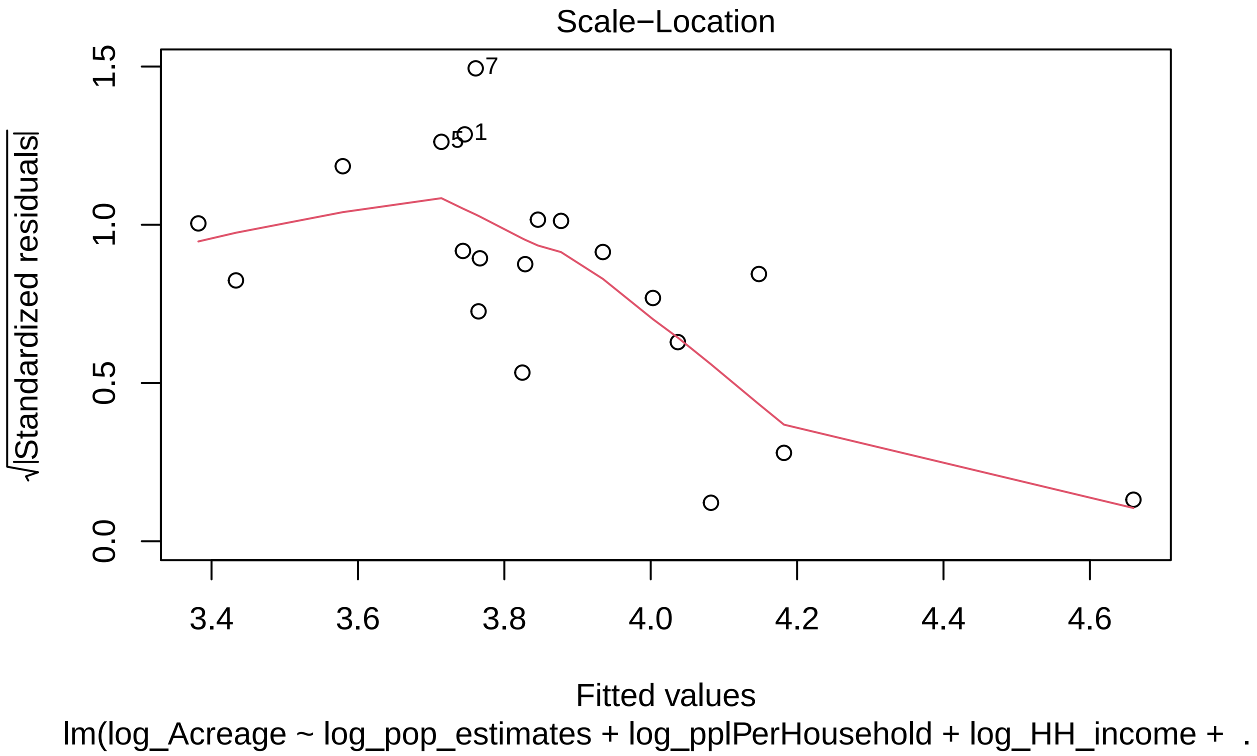
```
#log transform data
combined_data$log_pop_estimates <- log10(combined_data$`Population estimates, July 1, 2023, (V2023)`)
combined_data$log_pplPerHousehold <- log10(combined_data$`Persons per household, 2019-2023`)
combined_data$log_Black_ppl <- log10(combined_data$`Black alone, percent (a)`)
combined_data$log_Hispanic_ppl <- log10(combined_data$`Hispanic or Latino, percent (b)`)
combined_data$log_HH_income <- log10(combined_data$`Median households income (in 2023 dollars), 2019-20
combined_data$log_Acreage <- log10(combined_data$`Total Outdoor Recreation Acreage`)

#re-run model
Model2 <- lm(log_Acreage ~ log_pop_estimates + log_pplPerHousehold +
                log_HH_income + log_Black_ppl +log_Hispanic_ppl,
              data = combined_data)

summary(Model2)
```

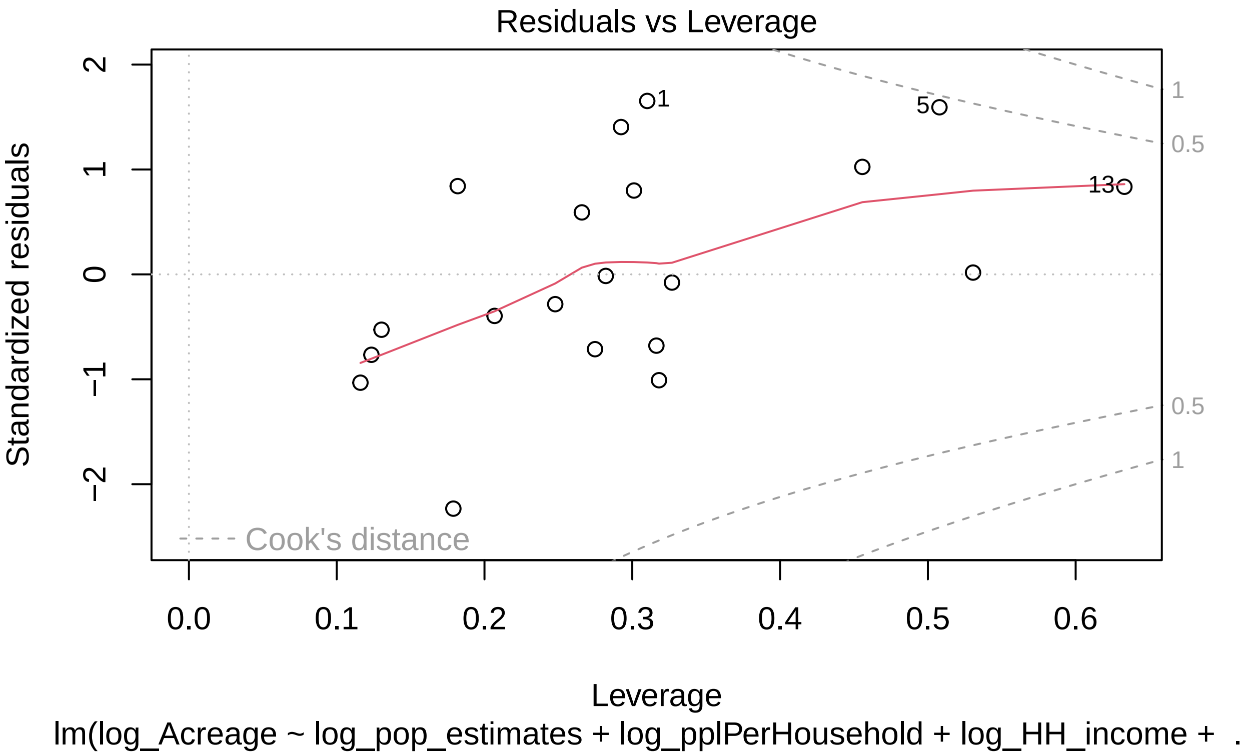
```
##
## Call:
## lm(formula = log_Acreage ~ log_pop_estimates + log_pplPerHousehold +
##     log_HH_income + log_Black_ppl + log_Hispanic_ppl, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36294 -0.38602 -0.02576  0.46485  0.92490
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -3.0567    11.3613  -0.269    0.792
## log_pop_estimates     0.6143     0.5132   1.197    0.251
## log_pplPerHousehold   6.3757     8.0861   0.788    0.444
## log_HH_income         0.6475     2.4413   0.265    0.795
## log_Black_ppl        -0.3049     0.8337  -0.366    0.720
## log_Hispanic_ppl     -1.3565     1.1227  -1.208    0.247
##
## Residual standard error: 0.6736 on 14 degrees of freedom
## Multiple R-squared:  0.191,  Adjusted R-squared:  -0.09796
## F-statistic: 0.661 on 5 and 14 DF,  p-value: 0.6589
```

```
plot(Model2)
```




Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(log_Acreage ~ log_pop_estimates + log_pplPerHousehold + log_HH_income + .
7
5
1


Scale−Location
1.5
1.0
0.5
0.0
√|Standardized residuals|
7
5
1
3.4
3.6
3.8
4.0
4.2
4.4
4.6
Fitted values
lm(log_Acreage ~ log_pop_estimates + log_pplPerHousehold + log_HH_income + .



(3) Narrate Decision-making

I decided to choose the following variables for my model: population estimates, persons per household, percent black alone, percent Hispanic or Latino, and median households income. I first looked at variables which had very low correlations with each other. Then out of those I chose variables that seemed to have a relationship with total outdoor recreation acreage based on their scatterplots. I fit these variables into a multiple linear regression model. The R2 and adjusted R2 were very low suggesting the model explains very little of the variability of the dependent variable. Also the first plot of the residuals shows some heteroscedasticity so I decided to log transform the data and re-run the model. The results of the new model are worse than the first. They show that these variables do not predict variability in total outdoor recreation acreage.

(4) Describe Final Model Output

The results of a multiple linear regression (F = 0.66, p = 0.66, Adjusted R2 = -0.98) suggest that total outdoor recreation acreage is not significantly predicted by population, people per household, percent of black people, percent of Hispanic or Latino people, or median household income. None of the independent variables were statistically significant (all p-values $>$ 0.2). Although a log-transformation was applied to improve model fit it did not meaningfully enhance the results.This suggests that different variables ought to be used to predict recreation acreage.