Week 8 Assignment

Jessalyn Chuang

Your research question for this assignment is - What demographic factors significantly predict outdoor recreation acreage at the county level across 20 North Carolina counties?

#(1) Load, tidy, and combine the datasets in RStudio. Filter the Office of State Budget data to include only data for the 20 relevant counties (see census data) and sum together local, state, and federal recreation acreage to form your outcome or dependent variable value of total recreation acreage by county. Filter the census data down to include only the following ten items:

Population estimates, July 1, 2023 (V2023) Persons under 18 years, percent Persons 65 years and over, percent Female persons, percent White alone, percent Black or African American alone, percent (a) Hispanic or Latino, percent (b) Median value of owner-occupied housing units, 2019-2023 Persons per household, 2019-2023 Median household income (in 2023 dollars), 2019-2023

```
library(here)
```

```
## here() starts at /home/guest/Statistical_Modeling_Sp25
```

```
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 1.9.3     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 errors
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(grid)

NC_Census <- read.csv(here("./Assignments/week8/NC_Census.csv"))
NC_Recreation_Acreage <- read.csv(("./Assignments/week8/NC_Recreation_Acreage.csv"))
```

Cleaning up data:

```
#renaming columns in NC_Census so they match up with NC_Recreation_Acreage county names
NC_Census <- NC_Census %>% rename(`Chatham County` = Chatham.County..North.Carolina,
                      `Wake County` = Wake.County..North.Carolina,
                      `Alamance County` = Alamance.County..North.Carolina,
                      `Orange County` = Orange.County..North.Carolina,
                      `Durham County` = Durham.County..North.Carolina,
                      `Caswell County` = Caswell.County..North.Carolina,
                      `Person County` = Person.County..North.Carolina,
                      `Granville County` = Granville.County..North.Carolina,
                      `Franklin County` = Franklin.County..North.Carolina,
                      `Vance County` = Vance.County..North.Carolina,
                      `Lee County` = Lee.County..North.Carolina,
                      `Moore County` = Moore.County..North.Carolina,
                      `Randolph County` = Randolph.County..North.Carolina,
                      `Guilford County` = Guilford.County..North.Carolina,
                      `Rockingham County` = Rockingham.County..North.Carolina,
                      `Halifax County` = Halifax.County..North.Carolina,
                      `Nash County` = Nash.County..North.Carolina,
                      `Wilson County` = Wilson.County..North.Carolina,
                      `Johnston County` = Johnston.County..North.Carolina,
                      `Harnett County` = Harnett.County..North.Carolina)

#filter to only include the aforementioned items from census data
NC_Census_filtered <- NC_Census %>% slice(c(2, 10, 11, 12, 13, 14, 19, 25, 31, 48))

#filter the NC_Recreation_Acreage data to only have the counties in the
#NC_Census data file (remove counties that don't appear in NC_Census)
NC_Recreation_Acreage_filtered <- NC_Recreation_Acreage %>%
  filter(Area.Name %in% colnames(NC_Census_filtered))

NC_Recreation_Acreage_Sum <- NC_Recreation_Acreage_filtered %>%
  group_by(Area.Name) %>%
  summarize(Total_Recreation_Acreage = sum(Value)) %>%
  ungroup()

#transpose NC_Recreation_Acreage_Sum to make it easier to merge with NC_Census_filtered
NC_Recreation_Acreage_Sum <- NC_Recreation_Acreage_Sum %>%
  pivot_wider(names_from = Area.Name, values_from = Total_Recreation_Acreage)

#merge the two data frames together
NC_data_merged <- merge(NC_Census_filtered, NC_Recreation_Acreage_Sum, all = TRUE)

NC_data_merged$Fact[8] <- "Total Recreation Acreage"

#Cleaning up merged data frame
NC_data_merged <- NC_data_merged %>%
  filter(Fact != "Total Recreation Acreage") %>%
  bind_rows(filter(NC_data_merged, Fact == "Total Recreation Acreage")) %>%
  subset(select = -ncol(NC_data_merged))

#transpose so rows are counties and columns are the facts
NC_data_transposed <- NC_data_merged %>%
```

```
  pivot_longer(cols = -Fact, names_to = "County", values_to = "Value") %>%
  pivot_wider(names_from = Fact, values_from = Value)

#removing dollar signs and turning into num values
NC_data_cleaned <- NC_data_transposed %>%
  mutate(across(-1, ~ str_remove_all(., "[$,()%]"))) %>%
  mutate(across(-1, as.numeric))
```

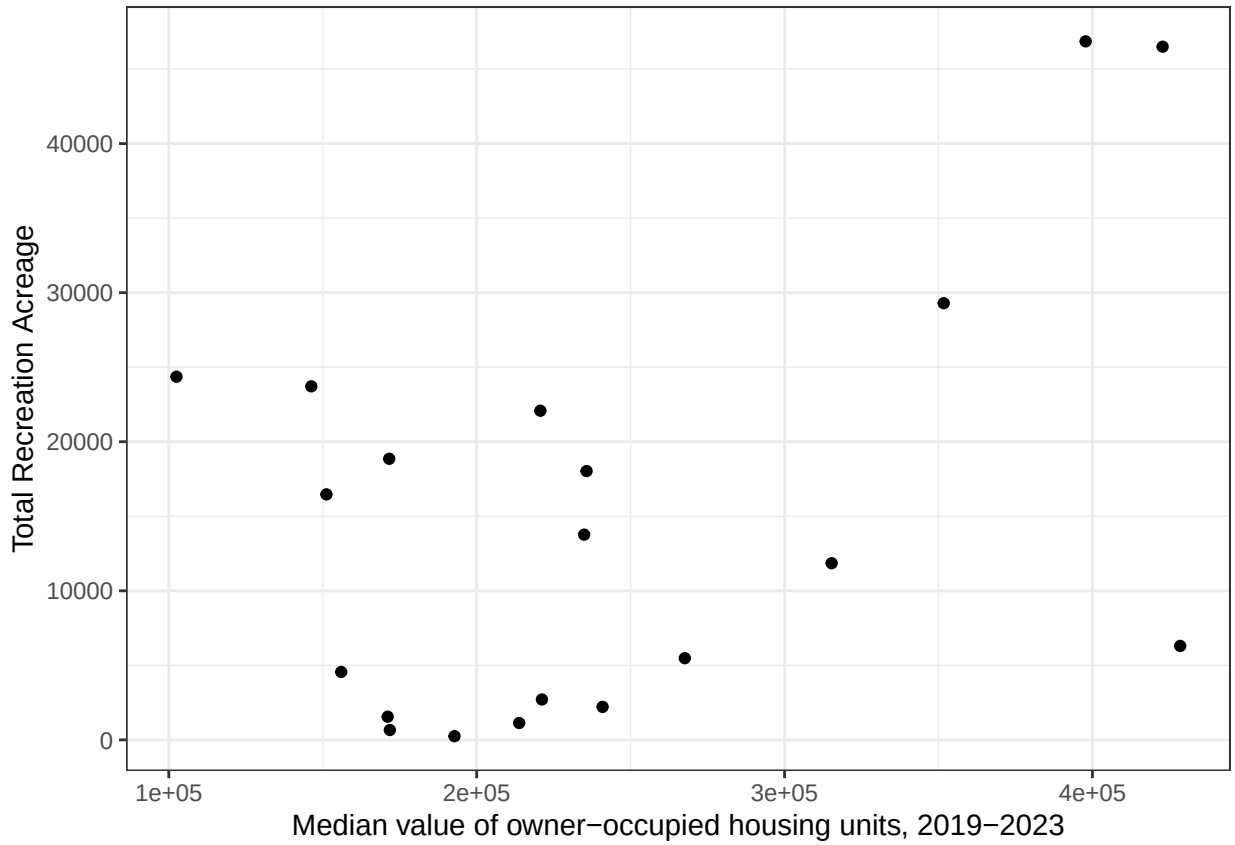
#(2) Using the workflow presented in lab, perform a multiple linear regression to investigate potential predictors of total outdoor recreation acreage. You must decide which of the initial 10 variables make sense to include based on your conceptual understanding of relevant factors, model outputs, diagnostics, and comparative model fits (AIC values). There is no "right answer" for this analysis—make an informed decision based on the data and understanding that you have.

##Step 1 - Defining the research question

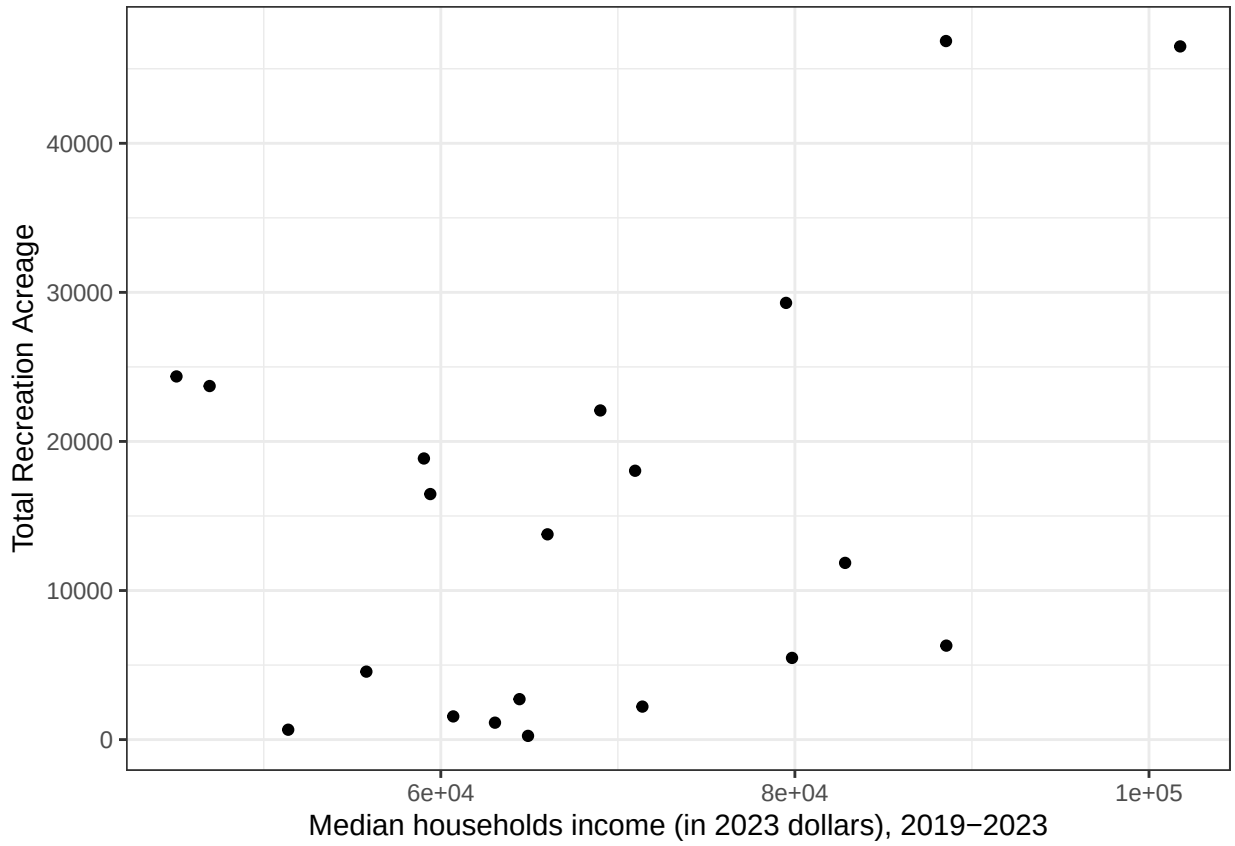
Research question: What are the potential census predictors of total outdoor recreation acreage?

##Step 2 - Examining the data and possible correlations

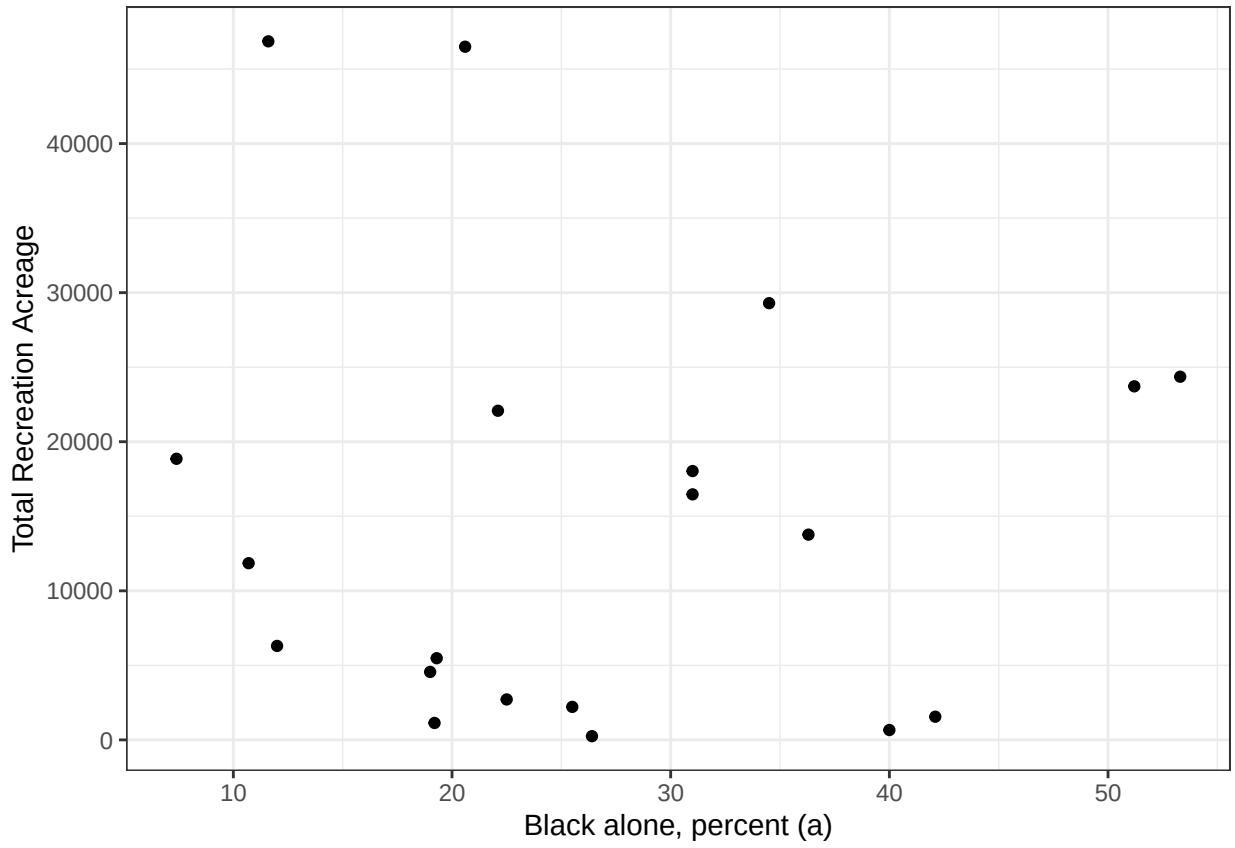
```
# Create and examine initial scatterplots of data.
#Median value of owner-occupied housing units, 2019-2023
scatter_1 <- ggplot(NC_data_cleaned, aes(x = `Median value of owner-occupied housing units, 2019-2023`,
                                         y = `Total Recreation Acreage`)) +
    geom_point() +
    theme_bw()
scatter_1
```



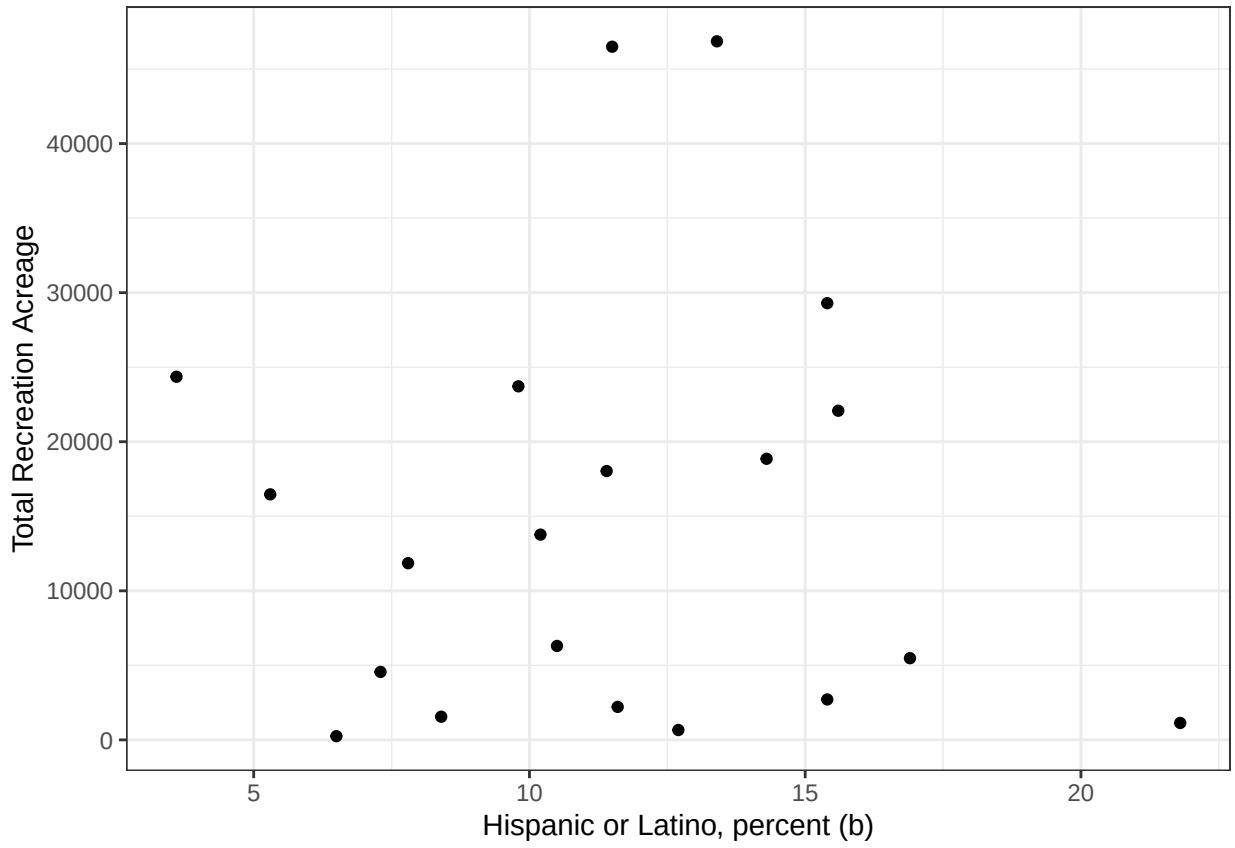
```r
#Median households income (in 2023 dollars), 2019-2023
scatter_2 <- ggplot(NC_data_cleaned, aes(x = `Median households income (in 2023 dollars), 2019-2023`,
                                         y = `Total Recreation Acreage`)) +
    geom_point() +
    theme_bw()
scatter_2
```



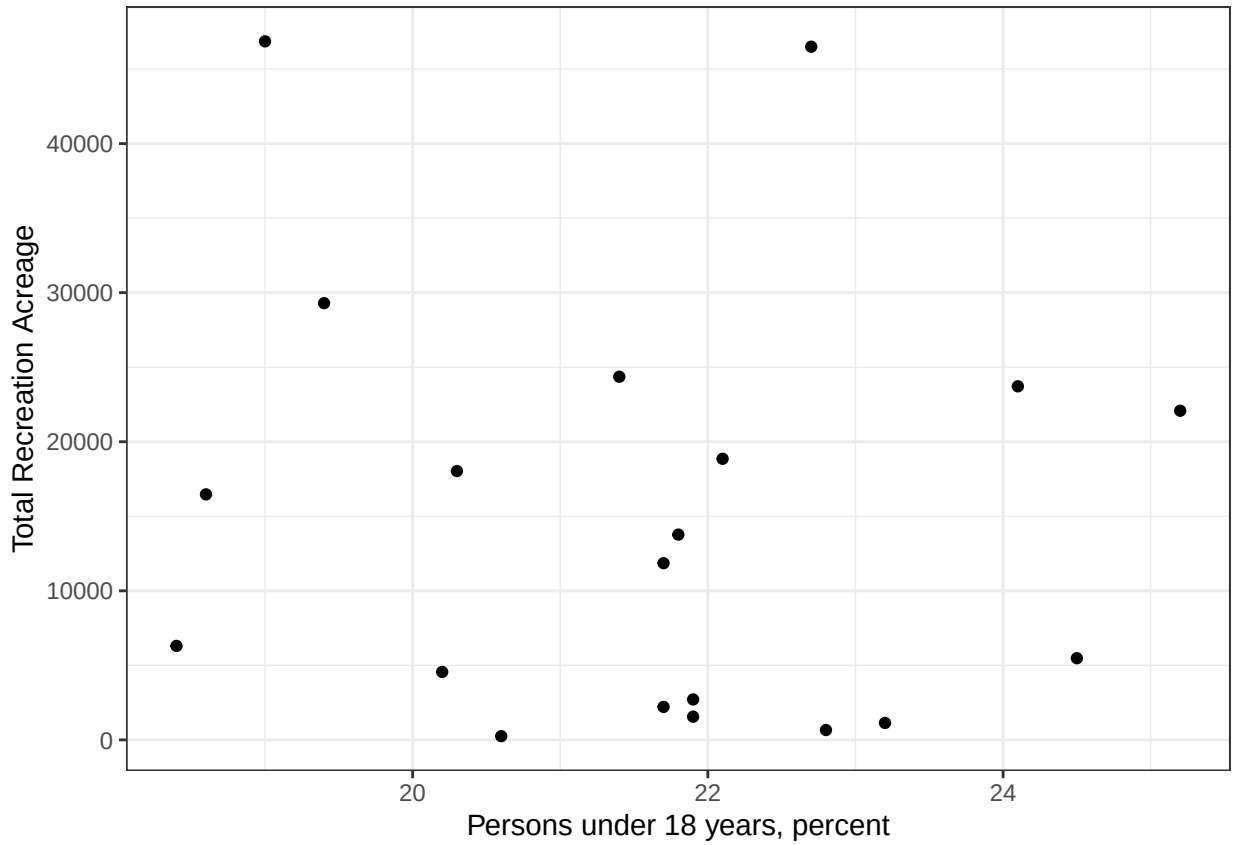

```r
#Black alone, percent (a)
scatter_3 <- ggplot(NC_data_cleaned, aes(x = `Black alone, percent (a)`,
                                         y = `Total Recreation Acreage`)) +
    geom_point() +
    theme_bw()
scatter_3
```



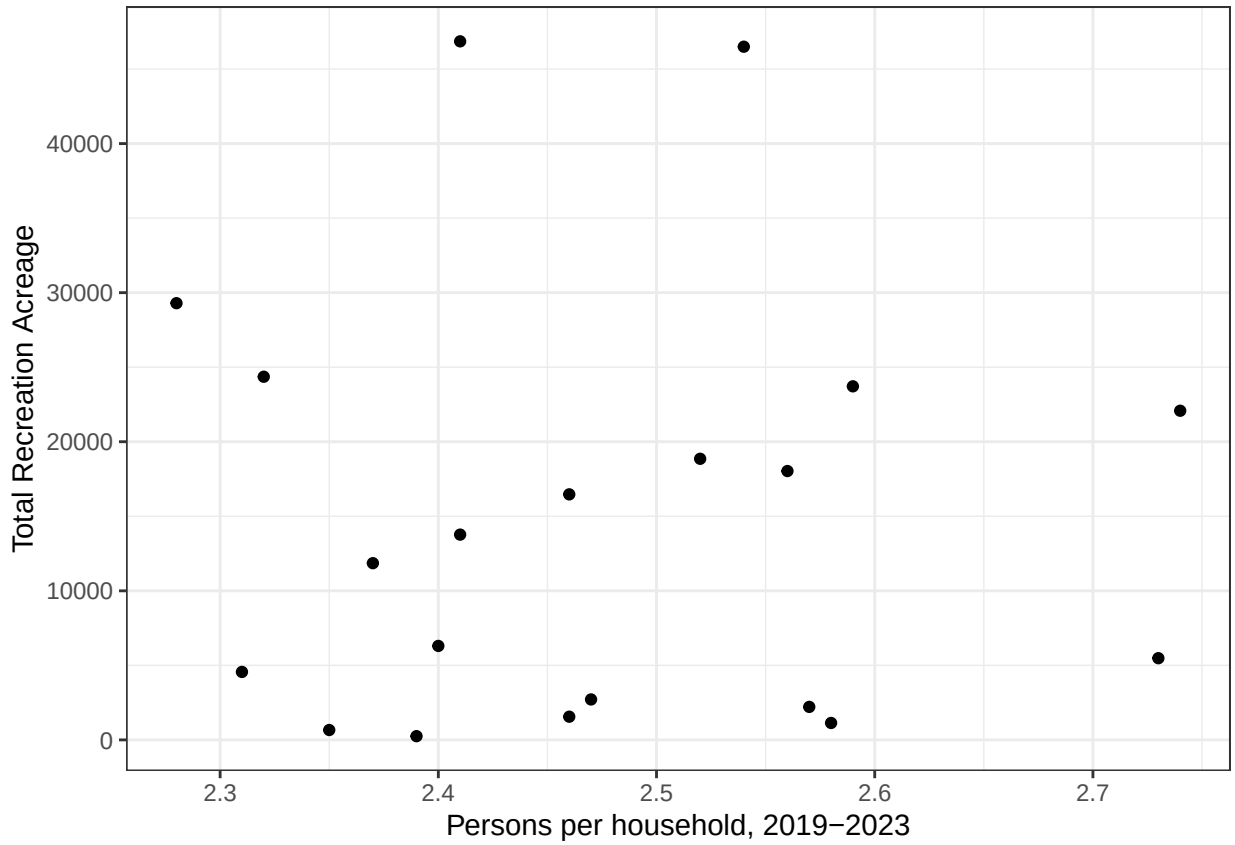

```
#Hispanic or Latino, percent (b)
scatter_4 <- ggplot(NC_data_cleaned, aes(x = `Hispanic or Latino, percent (b)`,
                                         y = `Total Recreation Acreage`)) +
    geom_point() +
    theme_bw()
scatter_4
```



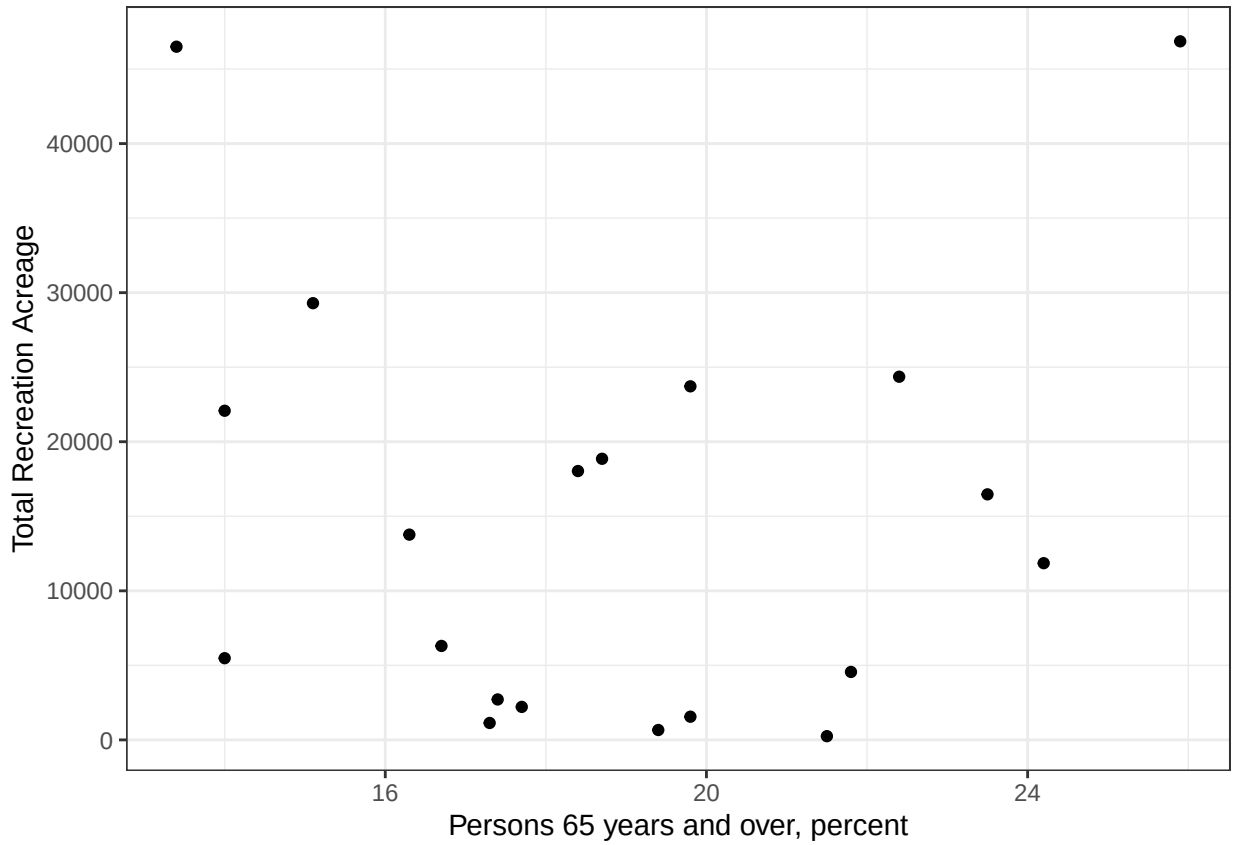

```
#Persons under 18 years, percent
scatter_5 <- ggplot(NC_data_cleaned, aes(x = `Persons under 18 years, percent`,
                                         y = `Total Recreation Acreage`)) +
    geom_point() +
    theme_bw()
scatter_5
```



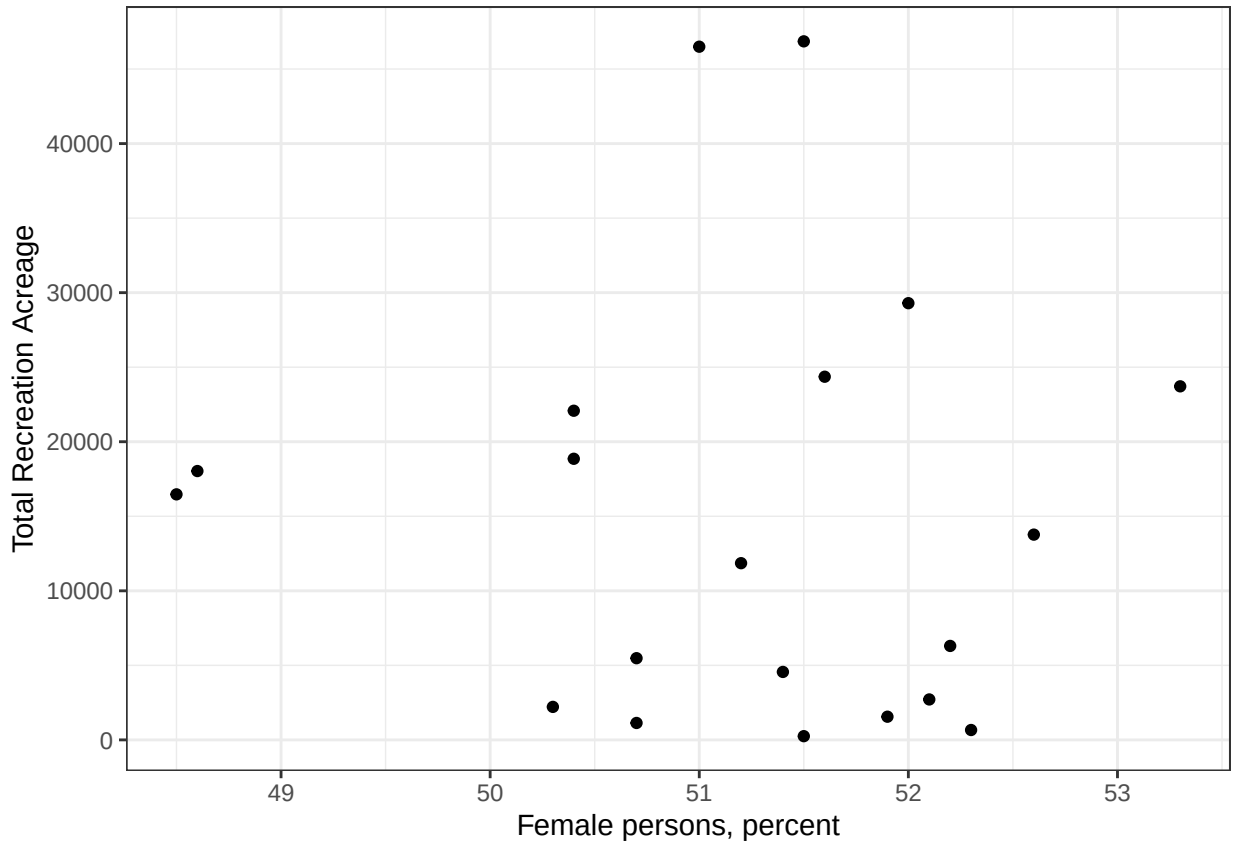

```
#Persons per household, 2019-2023
scatter_6 <- ggplot(NC_data_cleaned, aes(x = `Persons per household, 2019-2023`,
                                         y = `Total Recreation Acreage`)) +
    geom_point() +
    theme_bw()
scatter_6
```



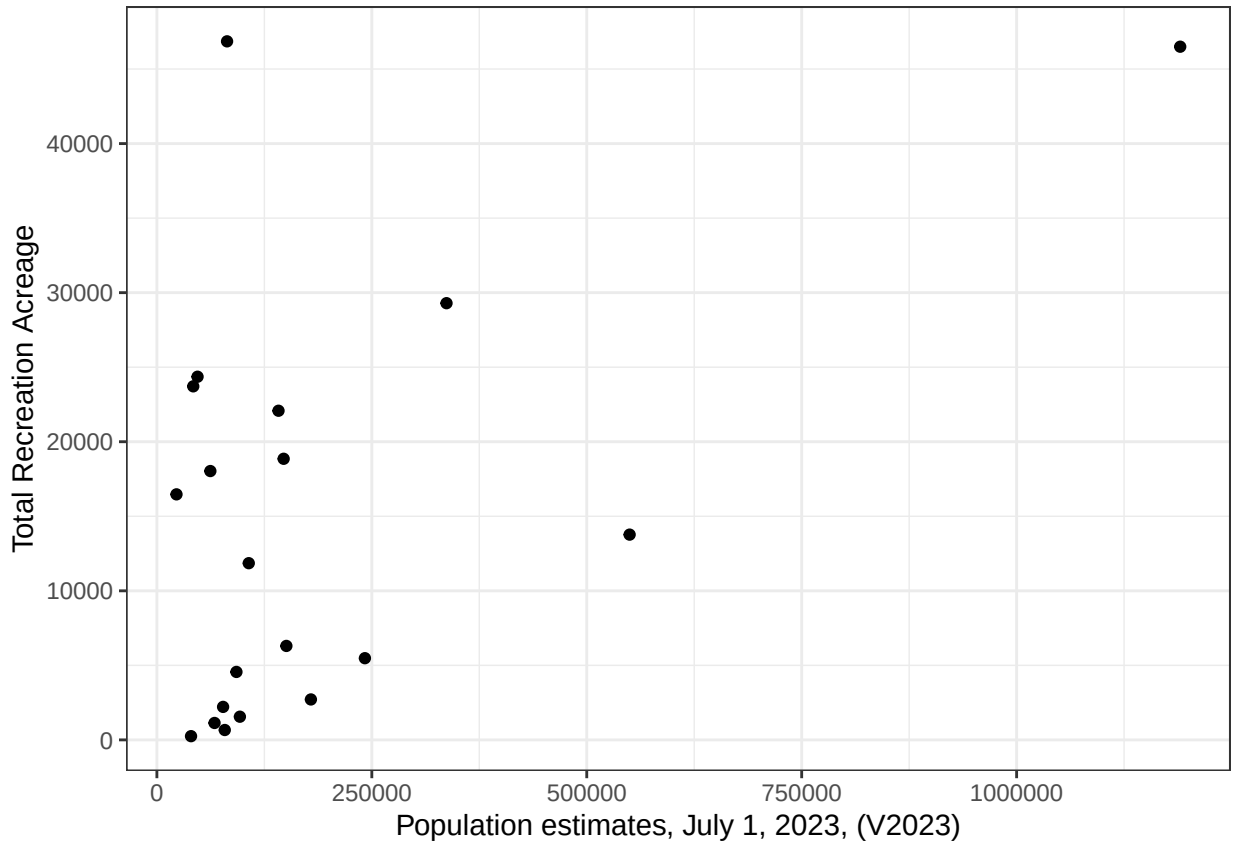

```
#Persons 65 years and over, percent
scatter_7 <- ggplot(NC_data_cleaned, aes(x = `Persons 65 years and over, percent`,
                                         y = `Total Recreation Acreage`)) +
    geom_point() +
    theme_bw()
scatter_7
```



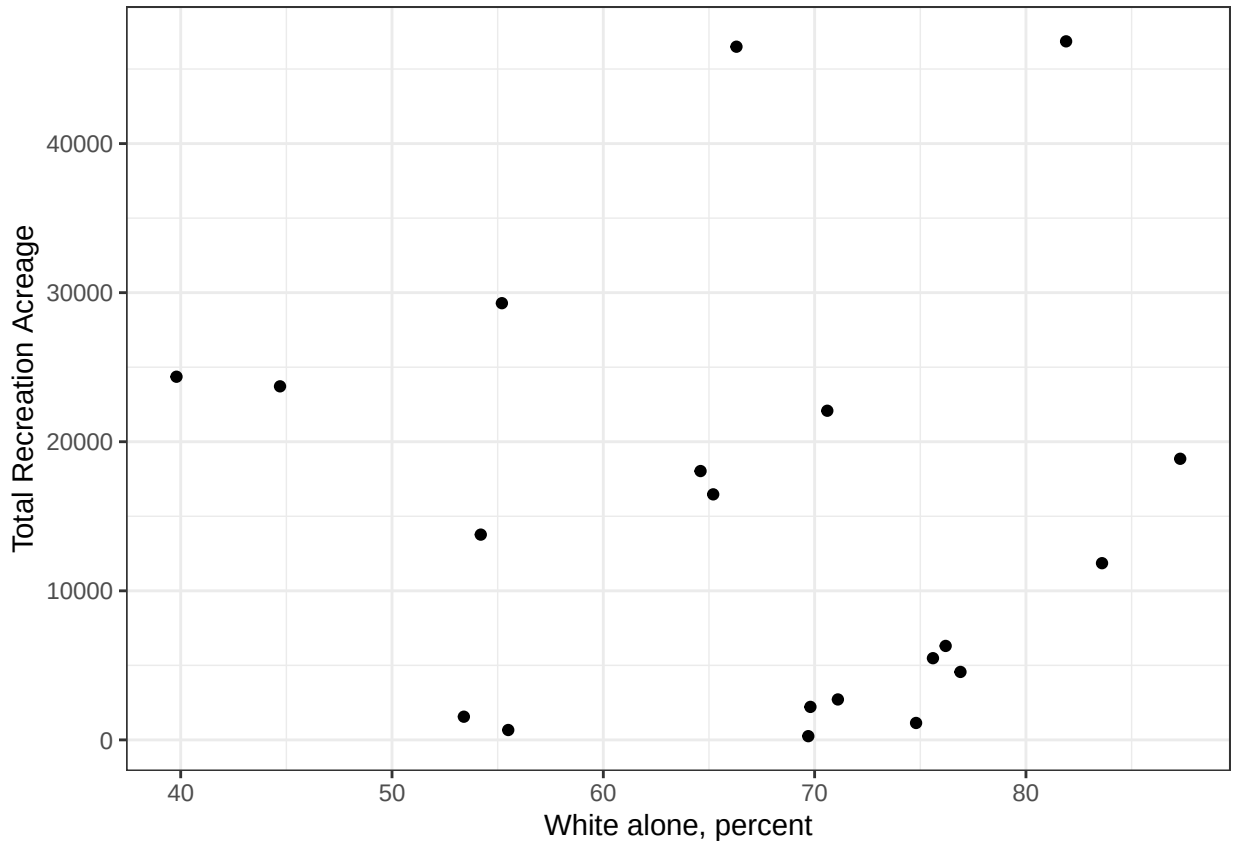

```
#Female persons, percent
scatter_8 <- ggplot(NC_data_cleaned, aes(x = `Female persons, percent`,
                                         y = `Total Recreation Acreage`)) +
    geom_point() +
    theme_bw()
scatter_8
```




```
#Population estimates, July 1, 2023, (V2023)
scatter_9 <- ggplot(NC_data_cleaned, aes(x = `Population estimates, July 1, 2023, (V2023)`,
                                         y = `Total Recreation Acreage`)) +
    geom_point() +
    theme_bw()
scatter_9
```




```
#White alone, percent
scatter_10 <- ggplot(NC_data_cleaned, aes(x = `White alone, percent`,
                                          y = `Total Recreation Acreage`)) +
    geom_point() +
    theme_bw()
scatter_10
```



Investigating multi-collinearity:

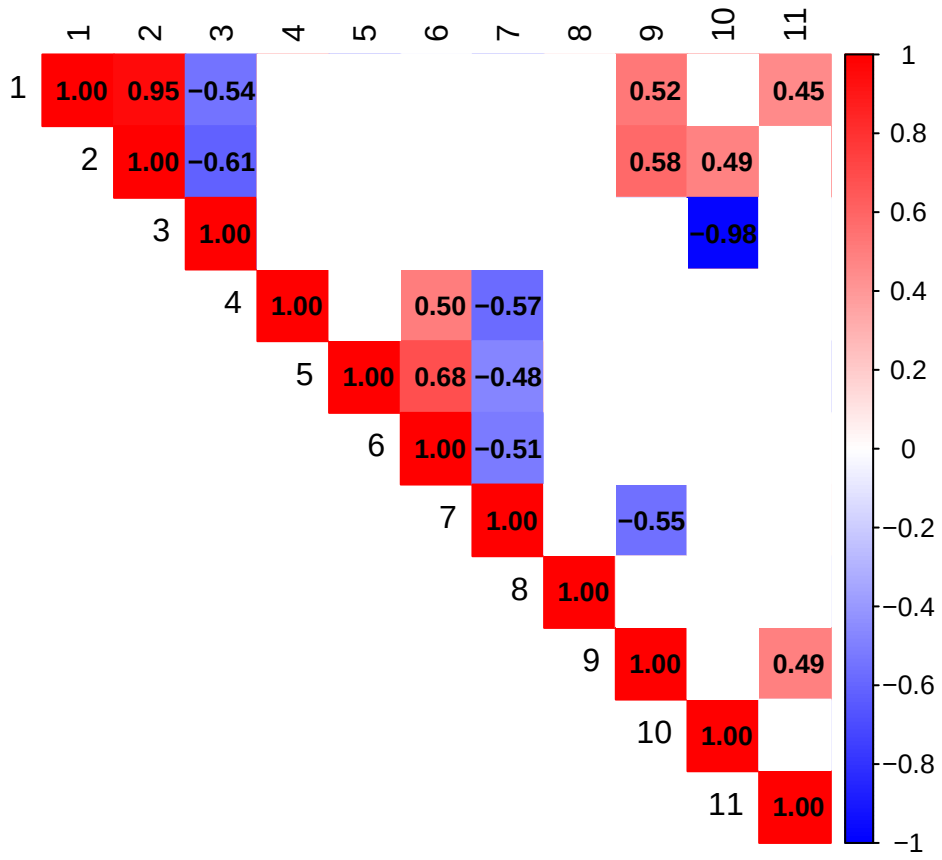
```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
numeric_data <- NC_data_cleaned %>% select(where(is.numeric))
#Learned about pairwise.complete.obs from:
#https://corrr.tidymodels.org/reference/correlate.html
cor_matrix <- cor(numeric_data, use = "pairwise.complete.obs")

# Compute p-values, cor.mtest() found at the following page:
#https://www.rdocumentation.org/packages/corrplot/versions/0.95/topics/cor.mtest
p_matrix <- cor.mtest(numeric_data)$p

# Create a numbered key
variable_names <- colnames(cor_matrix)  # Extract variable names
numbered_names <- paste0(1:length(variable_names))  # Create numbers

# Rename matrix columns and rows with numbers
colnames(cor_matrix) <- numbered_names
rownames(cor_matrix) <- numbered_names

#Learned corrplot in ENV 872, documentation at this page:
#https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
```

```
corrplot(cor_matrix, method = "color", type = "upper",
         col = colorRampPalette(c("blue", "white", "red"))(200),
         tl.col = "black", tl.cex = 1,
         addCoef.col = "black", number.cex = 0.8,
         p.mat = p_matrix, sig.level = 0.05, insig = "blank") #
```

```
## Warning in corrplot(cor_matrix, method = "color", type = "upper", col =
## colorRampPalette(c("blue", : p.mat and corr may be not paired, their rownames
## and colnames are not totally same!
```




```
key <- data.frame(
  Variable = variable_names
)

key <- head(key, -1)

#for identifying what I mean by X1 - X10
key
```

```
##                                                Variable
## 1  Median value of owner-occupied housing units, 2019-2023
## 2    Median households income (in 2023 dollars), 2019-2023
## 3                                 Black alone, percent (a)
## 4                         Hispanic or Latino, percent (b)
```

```
## 5                          Persons under 18 years, percent
## 6                        Persons per household, 2019-2023
## 7                     Persons 65 years and over, percent
## 8                                Female persons, percent
## 9           Population estimates, July 1, 2023, (V2023)
## 10                                White alone, percent
```

If there are pairs with correlations that are greater than 0.6 with p value < 0.05, only one of the variables can be kept in a linear model. Pairs that appear to have higher correlations based on the colors on the correlation plot (follow key for combinations):
X1 and X2
X2 and X3
X3 and X10
X5 and X6

Based on these findings, I will fit 6 different models using the following combinations that avoids multicollinearity:
X1, X3, X4, X5, X7, X8, X9
X1, X4, X5, X7, X8, X9, X10
X1, X3, X4, X6, X7, X8, X9
X1, X4, X6, X7, X8, X9, X10
X2, X4, X5, X7, X8, X9, X10
X2, X4, X6, X7, X8, X9, X10

##Step 3 - Fit regression model(s)

```
model_1 <- lm(data = NC_data_cleaned, `Total Recreation Acreage` ~
                `Median value of owner-occupied housing units, 2019-2023` +
                `Black alone, percent (a)` +
                `Hispanic or Latino, percent (b)` +
                `Persons under 18 years, percent` +
                `Persons 65 years and over, percent` +
                `Female persons, percent` +
                `Population estimates, July 1, 2023, (V2023)`)

model_2 <- lm(data = NC_data_cleaned, `Total Recreation Acreage` ~
                `Median value of owner-occupied housing units, 2019-2023` +
                `Hispanic or Latino, percent (b)` +
                `Persons under 18 years, percent` +
                `Persons 65 years and over, percent` +
                `Female persons, percent` +
                `Population estimates, July 1, 2023, (V2023)` +
                `White alone, percent`)

model_3 <- lm(data = NC_data_cleaned, `Total Recreation Acreage` ~
                `Median value of owner-occupied housing units, 2019-2023` +
                `Black alone, percent (a)` +
                `Hispanic or Latino, percent (b)` +
                `Persons per household, 2019-2023` +
                `Persons 65 years and over, percent` +
                `Female persons, percent` +
                `Population estimates, July 1, 2023, (V2023)`)

model_4 <- lm(data = NC_data_cleaned, `Total Recreation Acreage` ~
                `Median value of owner-occupied housing units, 2019-2023` +
```

```
                `Hispanic or Latino, percent (b)` +
                `Persons per household, 2019-2023` +
                `Persons 65 years and over, percent` +
                `Female persons, percent` +
                `Population estimates, July 1, 2023, (V2023)` +
                `White alone, percent`)

model_5 <- lm(data = NC_data_cleaned, `Total Recreation Acreage` ~
                `Median households income (in 2023 dollars), 2019-2023` +
                `Hispanic or Latino, percent (b)` +
                `Persons under 18 years, percent` +
                `Persons 65 years and over, percent` +
                `Female persons, percent` +
                `Population estimates, July 1, 2023, (V2023)` +
                `White alone, percent`)

model_6 <- lm(data = NC_data_cleaned, `Total Recreation Acreage` ~
                `Median households income (in 2023 dollars), 2019-2023` +
                `Hispanic or Latino, percent (b)` +
                `Persons per household, 2019-2023` +
                `Persons 65 years and over, percent` +
                `Female persons, percent` +
                `Population estimates, July 1, 2023, (V2023)` +
                `White alone, percent`)
```

##Step 4 - Evaluate model diagnostics

```
model_1_AIC <- AIC(model_1)
model_2_AIC <- AIC(model_2)
model_3_AIC <- AIC(model_3)
model_4_AIC <- AIC(model_4)
model_5_AIC <- AIC(model_5)
model_6_AIC <- AIC(model_6)
```

model 4 does the best as it received the lowest AIC of 437.42, and this includes the following: `Median value of owner-occupied housing units, 2019-2023 + Hispanic or Latino, percent (b) + Persons per household, 2019-2023 + Persons 65 years and over, percent + Female persons, percent + Population estimates, July 1, 2023, (V2023) + White alone, percent`

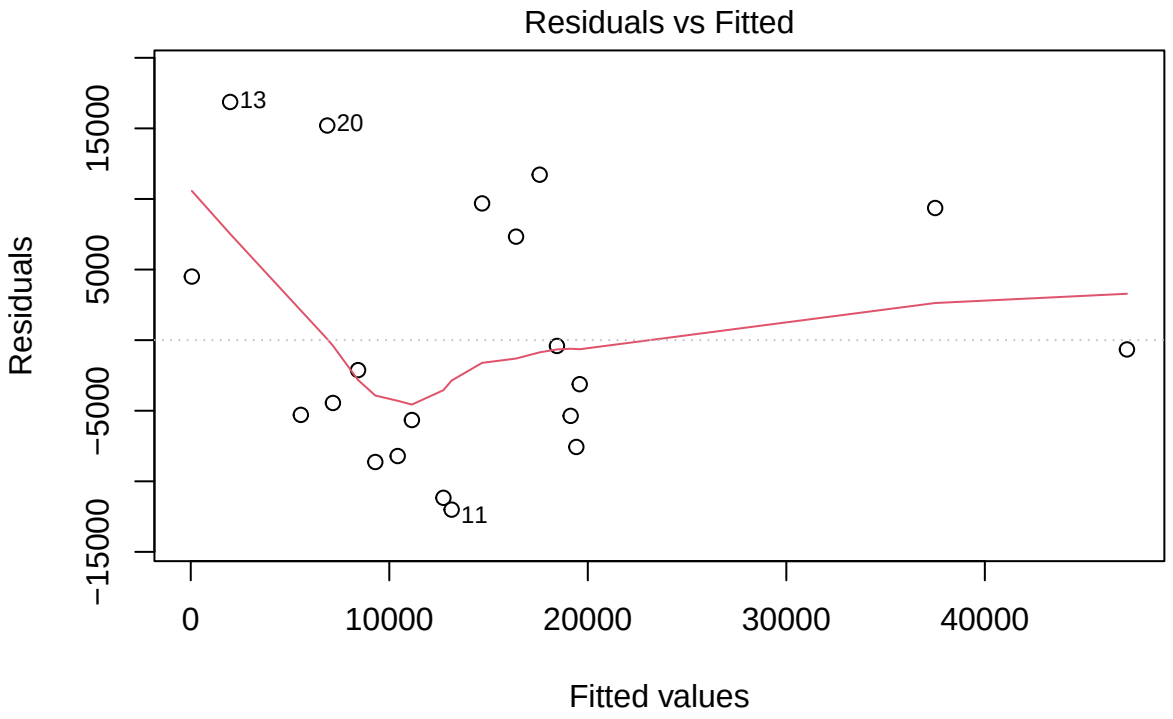
Next is to examine the results and the residuals of this model:

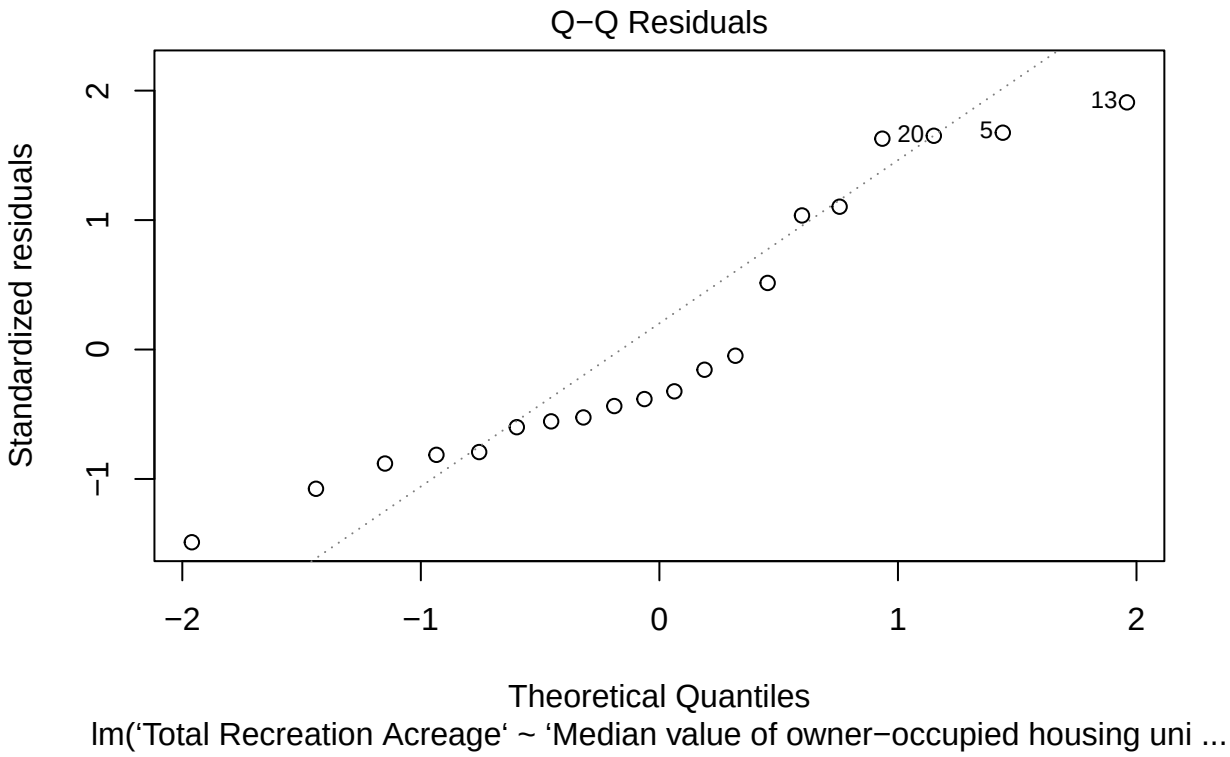
```
# Examine the results and the residuals.
summary(model_4)
```

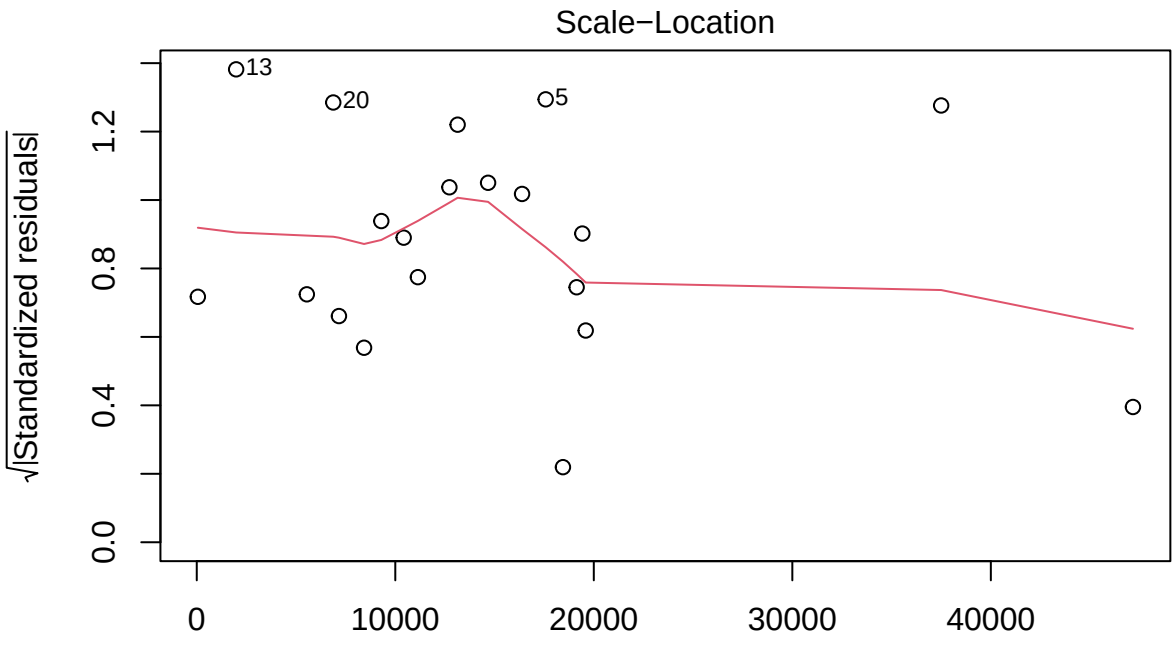
```
##
## Call:
## lm(formula = `Total Recreation Acreage` ~ `Median value of owner-occupied housing units, 2019-2023` +
##     `Hispanic or Latino, percent (b)` + `Persons per household, 2019-2023` +
##     `Persons 65 years and over, percent` + `Female persons, percent` +
##     `Population estimates, July 1, 2023, (V2023)` + `White alone, percent`,
##     data = NC_data_cleaned)
##
## Residuals:
```

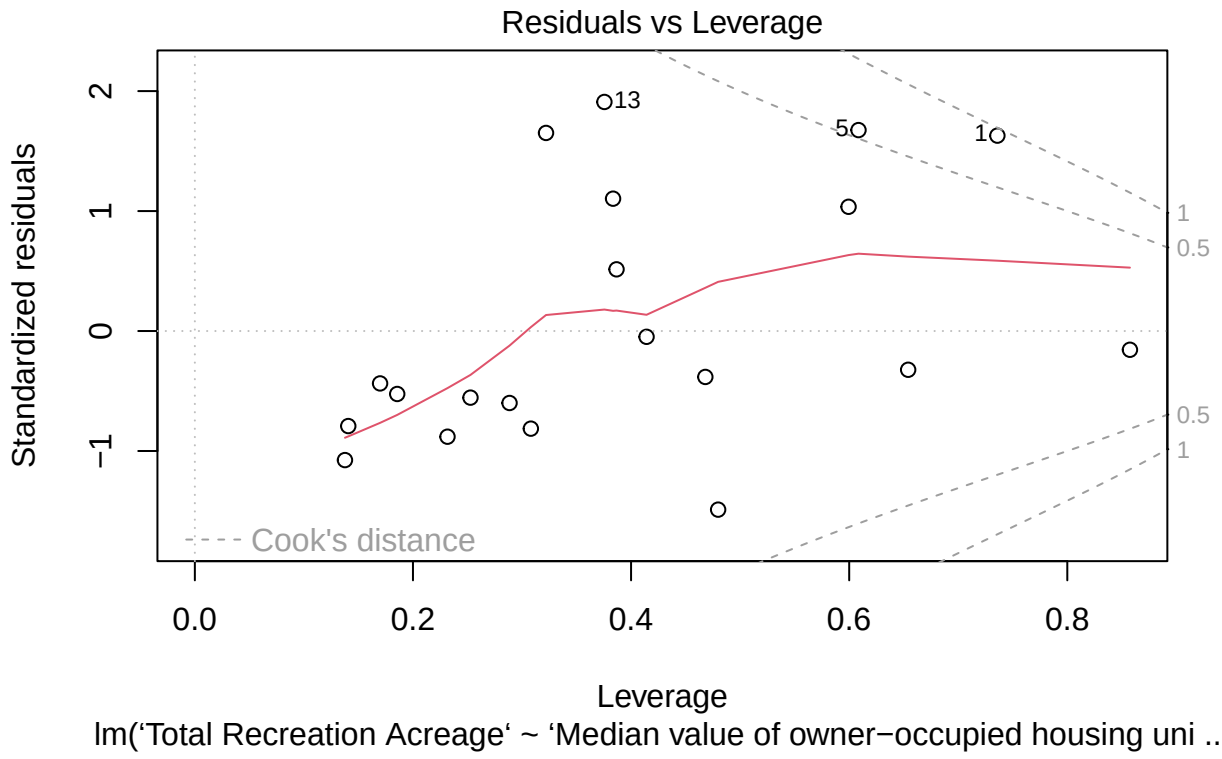
```
##    Min     1Q Median     3Q    Max
## -12006  -6136  -2623   7838  16872
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                      -7.790e+03  1.801e+05
## `Median value of owner-occupied housing units, 2019-2023`  7.218e-02  3.864e-02
## `Hispanic or Latino, percent (b)`                 9.769e+02  9.093e+02
## `Persons per household, 2019-2023`                3.184e+04  2.825e+04
## `Persons 65 years and over, percent`              3.510e+03  1.312e+03
## `Female persons, percent`                        -2.312e+03  2.590e+03
## `Population estimates, July 1, 2023, (V2023)`     3.532e-02  1.464e-02
## `White alone, percent`                           -5.858e+02  2.822e+02
##                                                  t value Pr(>|t|)
## (Intercept)                                       -0.043   0.9662
## `Median value of owner-occupied housing units, 2019-2023`   1.868   0.0863 .
## `Hispanic or Latino, percent (b)`                 1.074   0.3038
## `Persons per household, 2019-2023`                1.127   0.2818
## `Persons 65 years and over, percent`              2.675   0.0202 *
## `Female persons, percent`                        -0.893   0.3896
## `Population estimates, July 1, 2023, (V2023)`     2.413   0.0327 *
## `White alone, percent`                           -2.076   0.0601 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11180 on 12 degrees of freedom
## Multiple R-squared:  0.6101, Adjusted R-squared:  0.3827
## F-statistic: 2.683 on 7 and 12 DF,  p-value: 0.06396
```

```
plot(model_4)
```


Residuals vs Fitted
Residuals
Fitted values
lm(‘Total Recreation Acreage‘ ~ ‘Median value of owner−occupied housing uni ...
13
20
11
15000
5000
−5000
−15000
0
10000
20000
30000
40000


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(‘Total Recreation Acreage‘ ~ ‘Median value of owner−occupied housing uni ...
13
20
5
−2
−1
0
1
2


Scale−Location
13
20
5
√|Standardized residuals|
0.0
0.4
0.8
1.2
0
10000
20000
30000
40000
Fitted values
lm(‘Total Recreation Acreage‘ ~ ‘Median value of owner−occupied housing uni ...



It looks like only Persons 65 years and over, percent and Population estimates, July 1, 2023, (V2023) were the significant predictors, so I am revising the model to only include these two predictors. Additionally, based on plotting the residuals, it looks like row 1 could be an outlier, so I am also going to remove this and rerun the linear model.

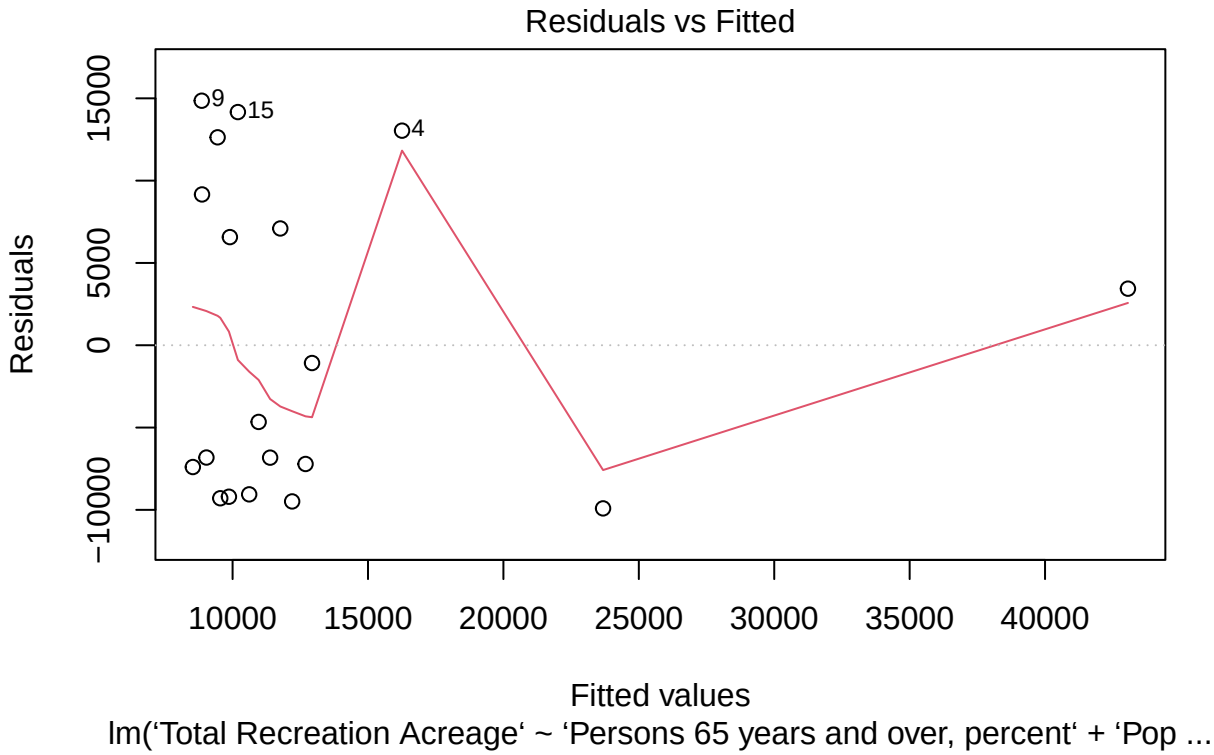
```
NC_data_cleaned_updated <- NC_data_cleaned[-1, ]
model_revised <- lm(data = NC_data_cleaned_updated,
`Total Recreation Acreage` ~ `Persons 65 years and over, percent` +
  `Population estimates, July 1, 2023, (V2023)`)
model_revised_AIC <- AIC(model_revised)
```

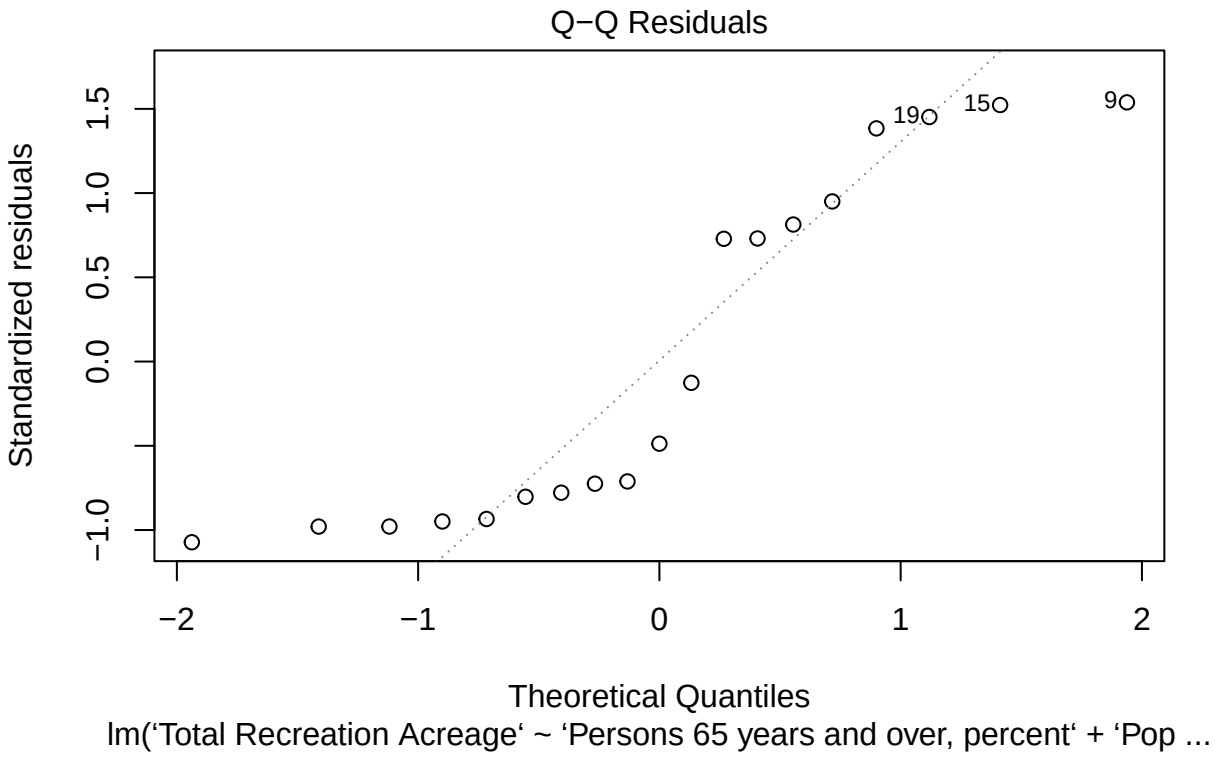
```
# Examine the results and the residuals.
summary(model_revised)
```

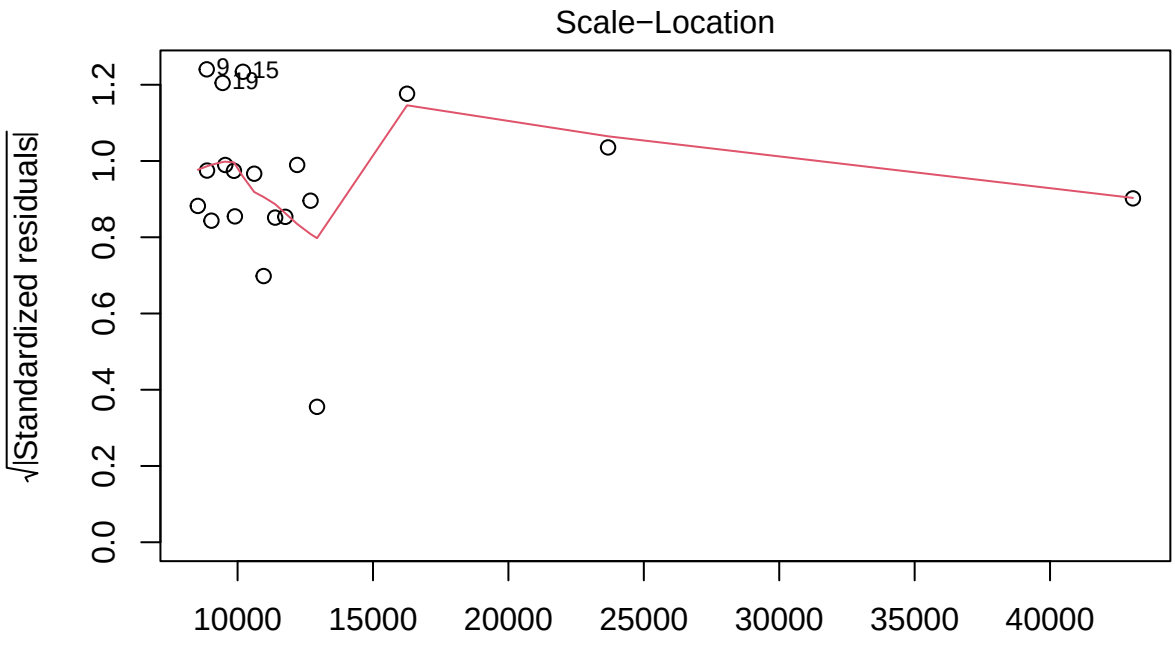
```
##
## Call:
## lm(formula = ‘Total Recreation Acreage‘ ~ ‘Persons 65 years and over, percent‘ +
##     ‘Population estimates, July 1, 2023, (V2023)‘, data = NC_data_cleaned_updated)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9910  -8228  -4659   8128  14855
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                            -1.442e+03  1.806e+04  -0.080
```

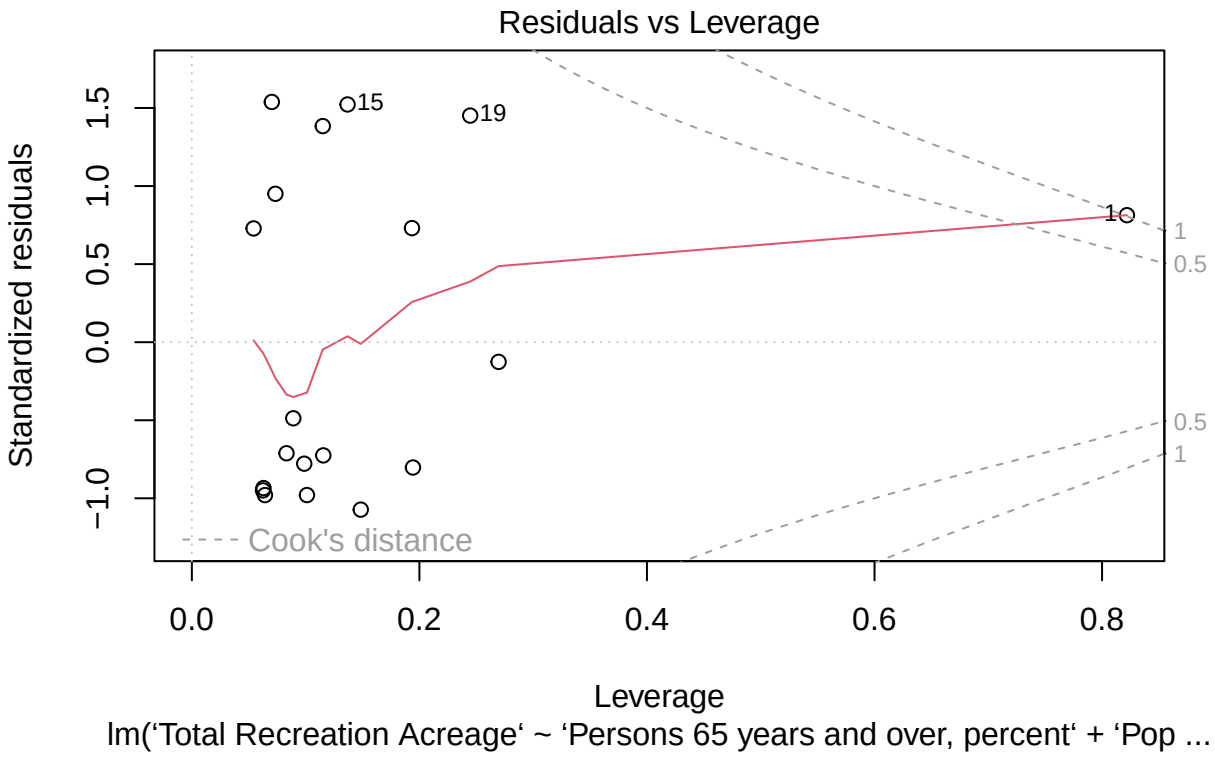
```
## `Persons 65 years and over, percent`           4.513e+02  9.002e+02   0.501
## `Population estimates, July 1, 2023, (V2023)`  3.231e-02  1.064e-02   3.038
##                                                Pr(>|t|)
## (Intercept)                                     0.93739
## `Persons 65 years and over, percent`            0.62295
## `Population estimates, July 1, 2023, (V2023)`   0.00784 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10010 on 16 degrees of freedom
## Multiple R-squared:  0.4206, Adjusted R-squared:  0.3481
## F-statistic: 5.807 on 2 and 16 DF,  p-value: 0.01271
```

```
plot(model_revised)
```




Q−Q Residuals
Standardized residuals
−1.0
0.0
0.5
1.0
1.5
19
15
9
−2
−1
0
1
2
Theoretical Quantiles
lm(‘Total Recreation Acreage‘ ~ ‘Persons 65 years and over, percent‘ + ‘Pop ...


Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
9
15
19
10000
15000
20000
25000
30000
35000
40000
Fitted values
lm(‘Total Recreation Acreage‘ ~ ‘Persons 65 years and over, percent‘ + ‘Pop ...



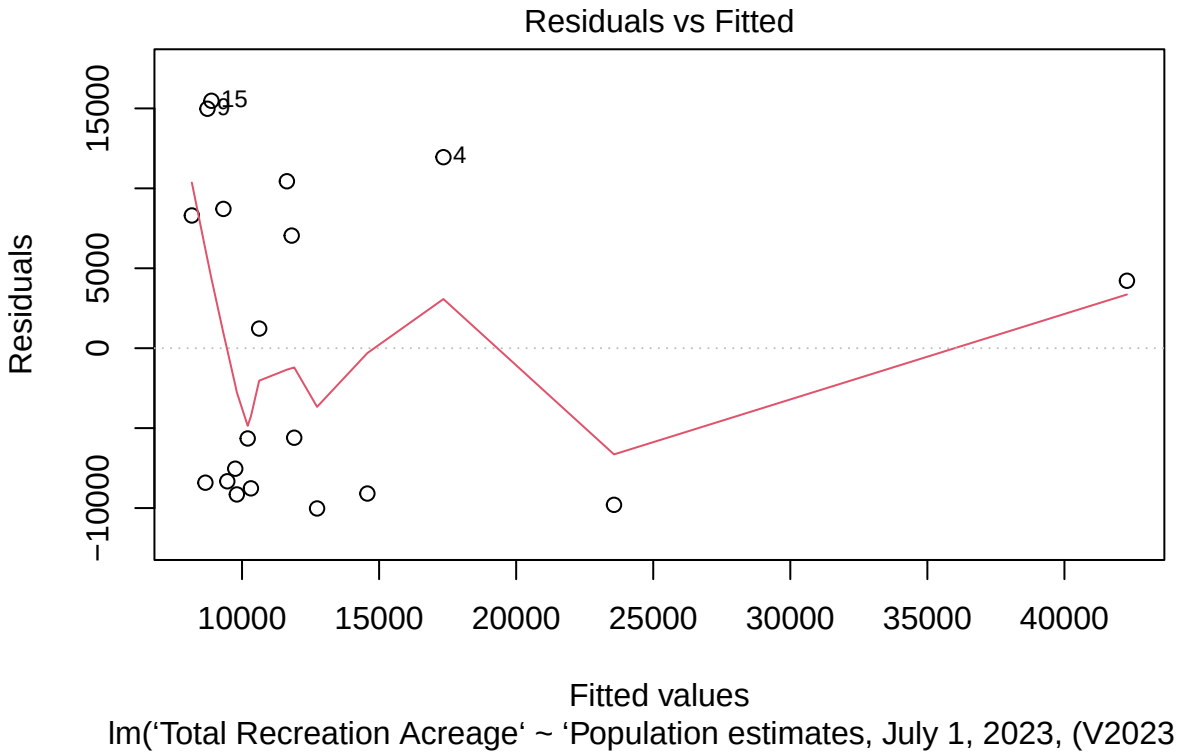
Now it looks like Population estimates, July 1, 2023, (V2023) is the only significant indicator.

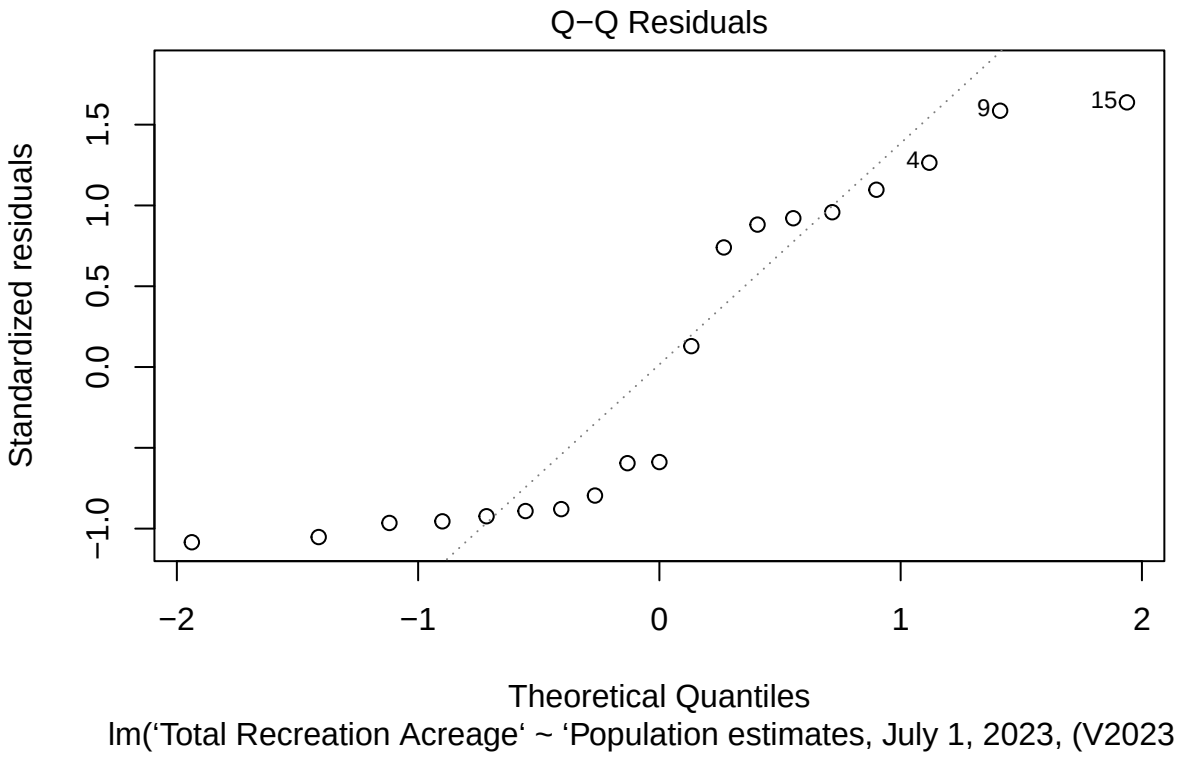
```
model_revised_2 <- lm(data = NC_data_cleaned_updated,
    `Total Recreation Acreage` ~ `Population estimates, July 1, 2023, (V2023)`)
model_revised_2_AIC <- AIC(model_revised_2)
summary(model_revised_2)
```

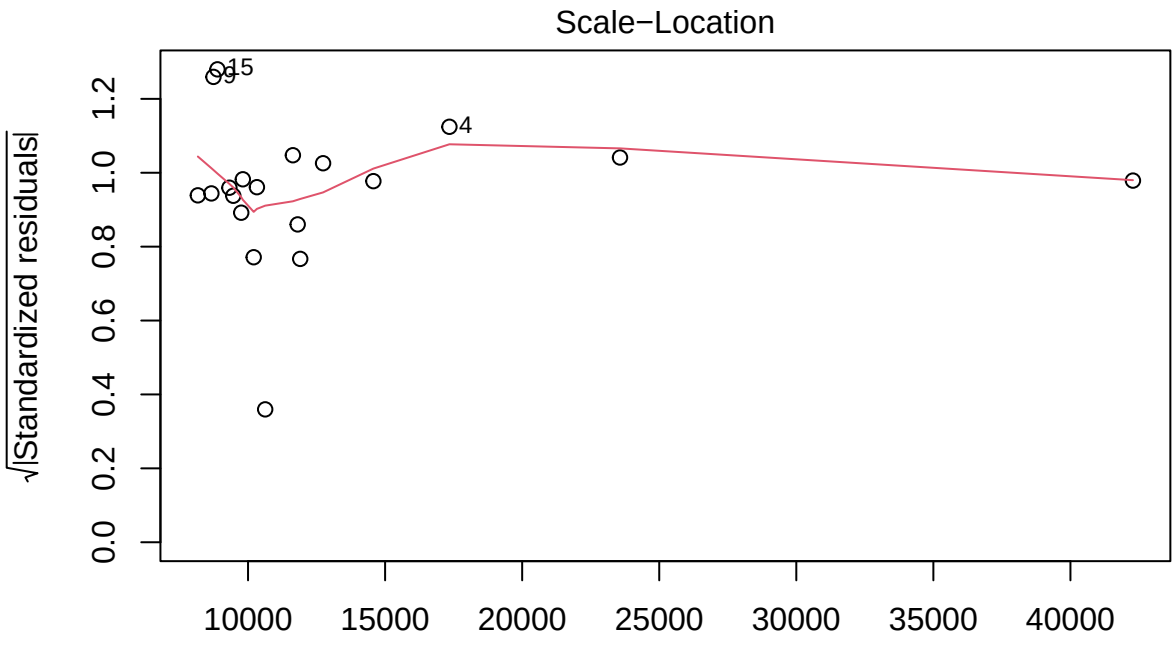
```
##
## Call:
## lm(formula = ‘Total Recreation Acreage‘ ~ ‘Population estimates, July 1, 2023, (V2023)‘,
##     data = NC_data_cleaned_updated)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10025  -8590  -5600   8508  15477
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 7.502e+03  2.780e+03   2.699
## ‘Population estimates, July 1, 2023, (V2023)‘ 2.922e-02  8.475e-03   3.447
##                                             Pr(>|t|)
## (Intercept)                                  0.01521 *
## ‘Population estimates, July 1, 2023, (V2023)‘  0.00307 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

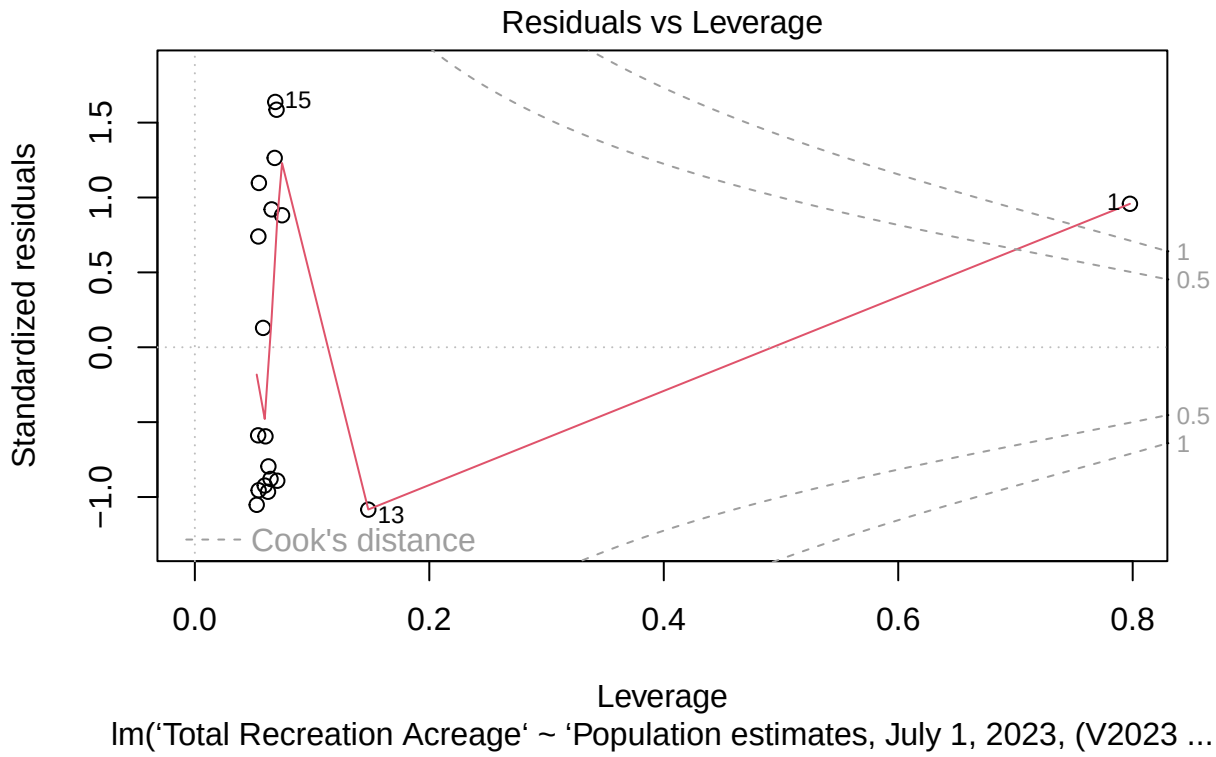
```
## Residual standard error: 9790 on 17 degrees of freedom
## Multiple R-squared:  0.4115, Adjusted R-squared:  0.3768
## F-statistic: 11.89 on 1 and 17 DF,  p-value: 0.003075
```

```
plot(model_revised_2)
```




Q−Q Residuals
Standardized residuals
−1.0
0.0
0.5
1.0
1.5
−2
−1
0
1
2
Theoretical Quantiles
lm(‘Total Recreation Acreage‘ ~ ‘Population estimates, July 1, 2023, (V2023 ...
9
15
4


Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
15
9
4
10000
15000
20000
25000
30000
35000
40000
Fitted values
lm(‘Total Recreation Acreage‘ ~ ‘Population estimates, July 1, 2023, (V2023 ...



It turns out, it appears that Population estimates, July 1, 2023, (V2023) is the only significant indicator for Total Recreation Acreage. Granted, there is some variance in the residuals, but I was originally weary of log-transforming data considering how I don't have many data points to begin with! It looks like the AIC has improved even further with this second revision.

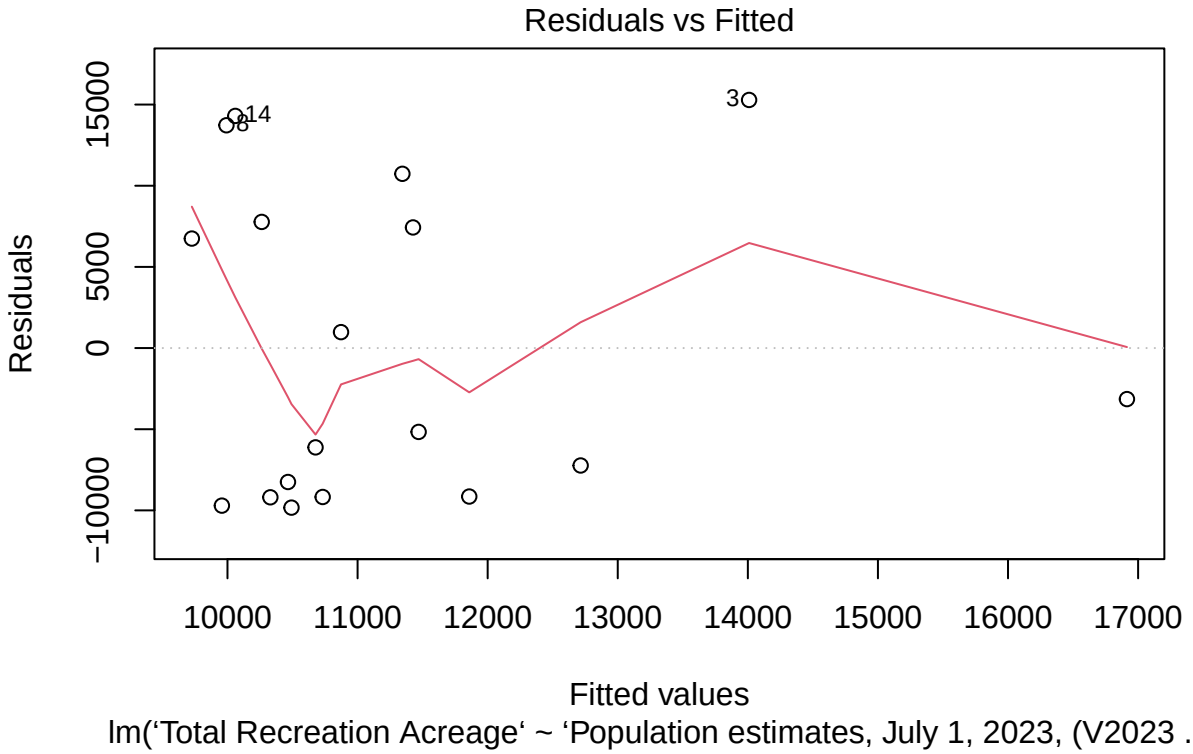
There is one more outlier to remove it seems like based on Cook's Distance on the Residuals vs Leverage chart. This point is removed and the model is revised again:

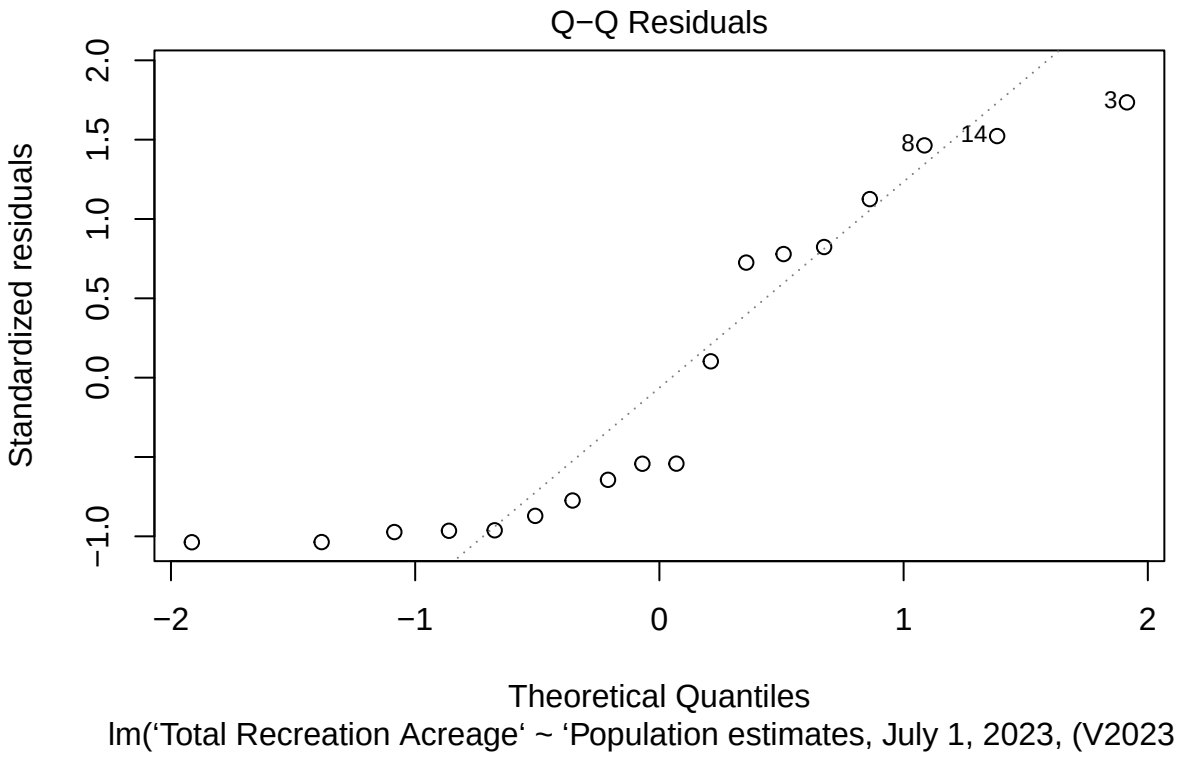
```
NC_data_cleaned_updated_2 <- NC_data_cleaned_updated[-1, ]
model_revised_3 <- lm(data = NC_data_cleaned_updated_2,
    `Total Recreation Acreage` ~ `Population estimates, July 1, 2023, (V2023)`)
model_revised_3_AIC <- AIC(model_revised_3)
summary(model_revised_3)
```

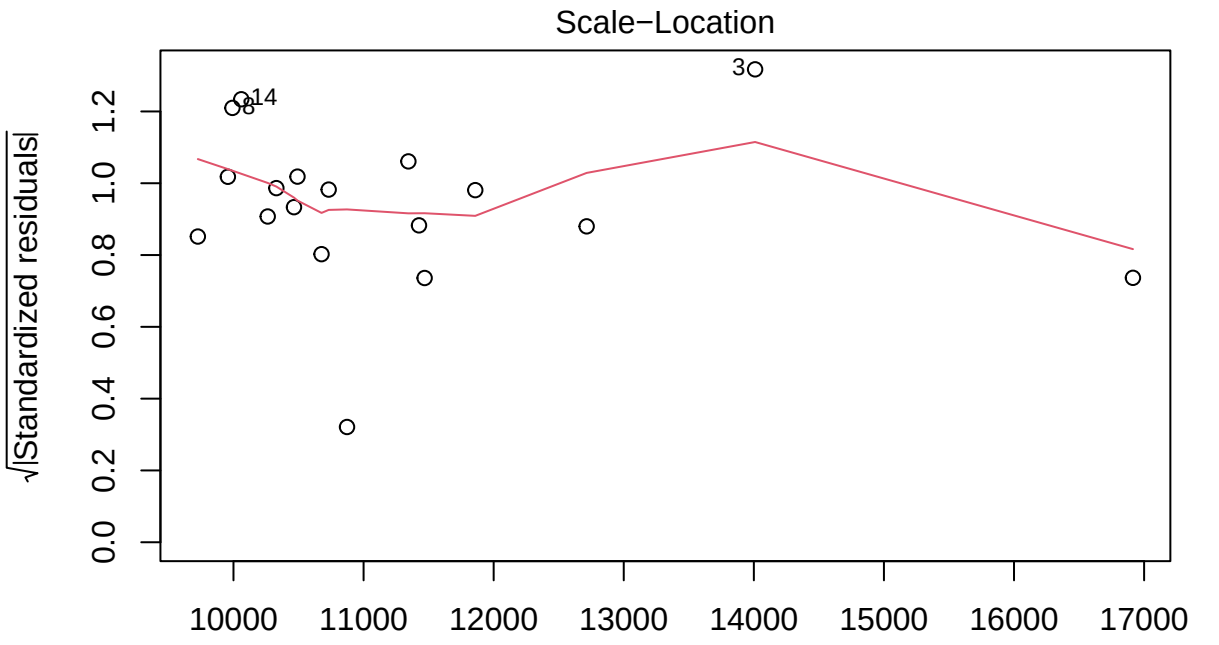
```
##
## Call:
## lm(formula = ‘Total Recreation Acreage‘ ~ ‘Population estimates, July 1, 2023, (V2023)‘,
##     data = NC_data_cleaned_updated_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9830  -8923  -4155   7685  15284
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             9.415e+03  3.431e+03   2.744
```

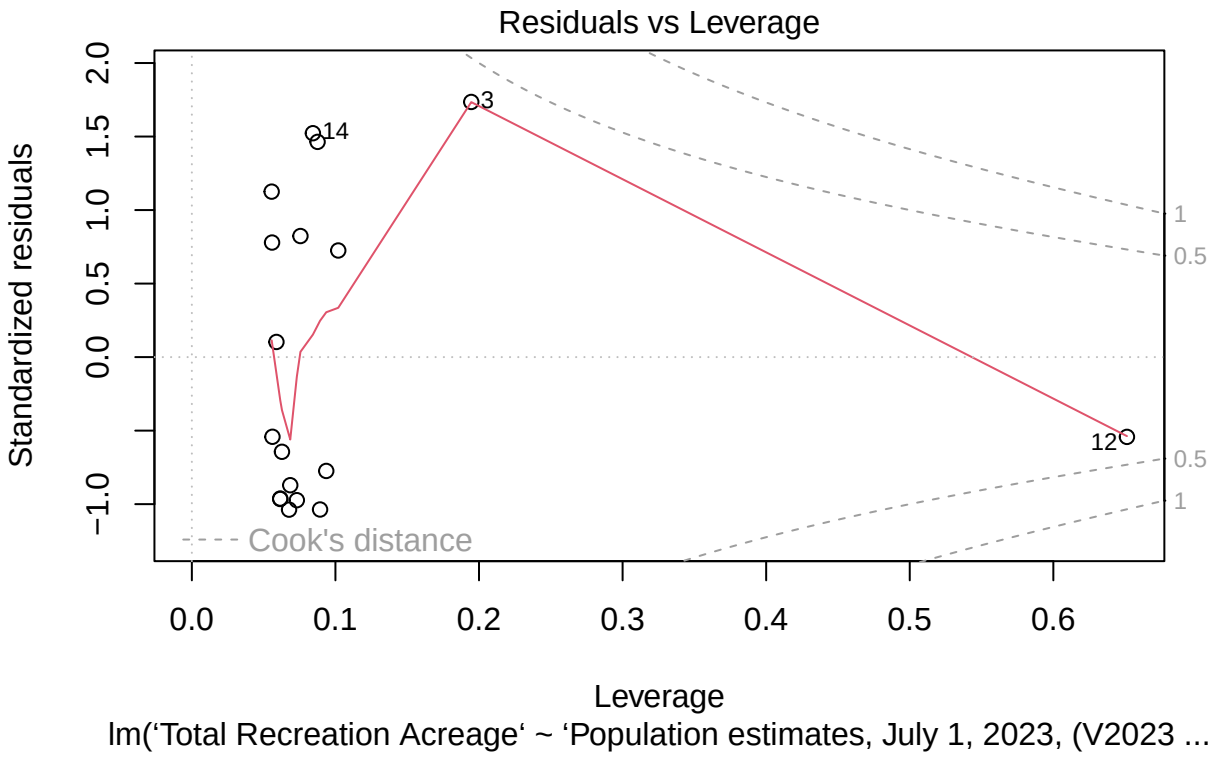
```
## `Population estimates, July 1, 2023, (V2023)` 1.364e-02  1.839e-02   0.742
##                                                Pr(>|t|)
## (Intercept)                                      0.0144 *
## `Population estimates, July 1, 2023, (V2023)`    0.4690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9815 on 16 degrees of freedom
## Multiple R-squared:  0.03325,    Adjusted R-squared:  -0.02717
## F-statistic: 0.5503 on 1 and 16 DF,  p-value: 0.469
```

```
plot(model_revised_3)
```




Q−Q Residuals
Standardized residuals
−1.0
0.0
0.5
1.0
1.5
2.0
3
8
14
−2
−1
0
1
2
Theoretical Quantiles
lm(‘Total Recreation Acreage‘ ~ ‘Population estimates, July 1, 2023, (V2023 ...


Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
3
14
8
10000
11000
12000
13000
14000
15000
16000
17000
Fitted values
lm(‘Total Recreation Acreage‘ ~ ‘Population estimates, July 1, 2023, (V2023 ...



Now it appears that not even Population estimates is a good predictor once the outlier is removed. Thus, I will be concluding that none of the variables are a good predictor of Total Recreation Acreage!

#(3) In 6-8 sentences, narrate your decision-making. Justify your inclusion of certain variables, investigation of any collinearity of variables, and model diagnostics. Did you choose to log-transform any data? Did you have concerns after exploring any of the model diagnostics? If so, how did that change your model selection? Please be sure to include information that helped you determine whether multi-collinearity, heteroscedasticity, or poor model fit was a concern.

When looking at the scatter plots of each independent variable, the total recreation acreage as a function of each variable did not see a wide range. Additionally, there are so few data points for each, that I decided to not log-transform any data. From a visual inspection, it does not look like heteroscedasticity is a major issue because there looks to generally be a relatively even spread of data points across different independent variables. Overall, the values for each didn't show a very obvious correlation, though interestingly in the scatter plot representing Total Recreation Acreage vs Population estimates, July 1, 2023, (V2023), most of that data was clustered to the left, indicating that most of the population estimates were smaller, but they encompassed a wide range of acreages. This would be the main variable that would make me concerned as there might be some heteroscedasticity present that might make log-transforming worth it. However, again there are so few data points, that I avoided log transforming this. From there, I investigated multi-collinearity, and this was done by creating a correlation matrix showing every pairwise correlation for those with p-values less than 0.05. Any pairs with correlations greater than 0.6 that were also significant were noted to only keep one of the independent variables in the pair for a potential linear model. I found 4 pairs that were shown to have multi-collinearity:

- Median value of owner-occupied housing units, 2019-2023 and Median households income (in 2023 dollars), 2019-2023
- Median households income (in 2023 dollars), 2019-2023 and Black alone, percent (a)

- Black alone, percent (a) and White alone, percent
- Persons under 18 years, percent and Persons per household, 2019-2023

With this in mind, I came up with 6 different models that did not pair two of these variables together. I then evaluated the models I created by using AIC values to describe them relative to one another and stating that the best model is the one with the lowest AIC value. Additionally, I also interpreted the actual model results and studied the residuals to further decide whether I should remove any variables I initially included (keeping only those that are considered significant in the model) or if I need to clean the data at all (i.e. log-transform or remove outliers).

#(4)

In 3-4 sentences, describe the final model output as you would for a final report.

A multiple linear regression analysis was conducted to examine whether demographic factors predict total recreation acreage across 20 North Carolina counties. Initial models included multiple predictors, and outliers were identified and removed based on diagnostic plots of residuals to improve model fit. After testing multiple model specifications and addressing potential data concerns, the final model included only Population Estimates, July 1, 2023, (V2023) as the predictor but was not statistically significant ($F(1,16) = 0.5503$, $p = 0.469$). The predictor, Population Estimates, was not significant ($B = 0.01364$, $p = 0.469$), and the model had a very low R^2 value (0.033); given these results, it's likely that other factors not analyzed may contribute to differences in recreation acreage in North Carolina.

Works Cited: Juice, S. and T. Fahey (2019). Health and mycorrhizal colonization response of sugar maple (Acer saccharum) seedlings to calcium addition in Watershed 1 at the Hubbard Brook Experimental Forest ver 3. Environmental Data Initiative. https://doi.org/10.6073/pasta/0ade53ede9a916a36962799b2407097e