Midterm Part2 & Part3

Ziqi Wei

a. Link to the GitHub repository containing all associated files used to create the webpage.

https://github.com/ZiqiWei217/PA-446-Midterm-Exam

b. Code and results/answers for Part II

Part II: Web scraping [20 points]

Go to the website https://www.scrapethissite.com/pages/simple/ and scrape the data to create a table with four variables: Country, Capital, Population, and Area. The table will have a total of 250 observations.

```
library(rvest)
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   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter()         masks stats::filter()
x readr::guess_encoding() masks rvest::guess_encoding()
x dplyr::lag()            masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```r
# Read the HTML
html <- read_html("https://www.scrapethissite.com/pages/simple/")

# Select the main container blocks for each country
country_blocks <- html |>
  html_elements("div.col-md-4.country")

# Extract the country name
country <- country_blocks |>
  html_element("h3.country-name") |>
  html_text2() # Use html_text2() for clean text extraction

# Extract the capital city
capital <- country_blocks |>
  html_element("span.country-capital") |>
  html_text2()

# Extract the population
population <- country_blocks |>
  html_element("span.country-population") |>
  html_text2() |>
  as.numeric() # Convert the text to a numeric type

# Extract the area
area <- country_blocks |>
  html_element("span.country-area") |>
  html_text2() |>
  as.numeric() # Convert the text to a numeric type

# Combine the extracted data into a tibble (a modern data frame)
country_data <- tibble(
  Country = country,
  Capital = capital,
  Population = population,
  "Area(km2)" = area
)

head(country_data)
```

```
# A tibble: 6 x 4
  Country              Capital          Population `Area(km2)`
  <chr>                <chr>                 <dbl>       <dbl>
```

```
1 Andorra              Andorra la Vella      84000         468
2 United Arab Emirates Abu Dhabi           4975593       82880
3 Afghanistan          Kabul              29121286      647500
4 Antigua and Barbuda  St. John's            86754         443
5 Anguilla             The Valley            13254         102
6 Albania              Tirana              2986952       28748
```

```
# Verify the count.
print(paste("Total observations scraped:", nrow(country_data)))
```

```
[1] "Total observations scraped: 250"
```

```
# Save this data to a file
write_csv(country_data, "scraped_countries.csv")
```

c. Code and results/answers for Part III

PART III: API Access [62 points]

Public administrators often use Census data to understand how civic resources - such as income, broadband, and education - vary across regions. The American Community Survey (ACS) provides detailed demographic and socioeconomic data accessible through the tidycensus package, which connects directly to the Census API.

You will use tidycensus and tidyverse tools to explore income inequality and the digital divide across Illinois counties.

Step 1: Identify Relevant Variables [6 points)

Go to the ACS 5-Year Data Dictionary and find the variables that represent the following.

```
library(tidyverse)
library(tidycensus)
# 2023 5-year American Community Survey (ACS).
v23 <- load_variables(2023, "acs5", cache = TRUE)

# Find the variable for Median Household Income
income_variable <- v23 %>%
  filter(name == "B19013_001")

print(income_variable, width = Inf)
```

```
# A tibble: 1 x 4
  name
  <chr>
1 B19013_001
  label
  <chr>
1 Estimate!!Median household income in the past 12 months (in 2023 inflation-ad~
  concept
  <chr>
1 Median Household Income in the Past 12 Months (in 2023 Inflation-Adjusted Dol~
  geography
  <chr>
1 block group
```

```
# Find variables for Broadband Internet Access
internet_variables <- v23 %>%
  filter(str_detect(name, "B28002_00"), # Look in the correct table
         str_detect(label, "Broadband|Total$")) # Find labels containing "Broadband" or endi

print(internet_variables %>% select(name, label), width = Inf)
```

```
# A tibble: 3 x 2
  name
  <chr>
1 B28002_004
2 B28002_007
3 B28002_008
  label
  <chr>
1 Estimate!!Total:!!With an Internet subscription!!Broadband of any type
2 Estimate!!Total:!!With an Internet subscription!!Broadband such as cable, fib~
3 Estimate!!Total:!!With an Internet subscription!!Broadband such as cable, fib~
```

• Median household income (in the past 12 months): B19013_001E • Households with broadband Internet: B28002_004E • Total households with any type of internet access: B28002_001E

Step 2: Retrieve Data [8 points]

Use get_acs() to retrieve the following variables for all counties in Illinois:

- Median household income (in the past 12 months)
- Households with broadband Internet
- Total households with any type of internet access

```
# Define a named vector of the variables
vars_to_get <- c(
  income = "B19013_001E",
  broadband = "B28002_004E",
  total_households = "B28002_001E"
)

# Use the get_acs() function to retrieve the data for all counties in Illinois.
il_acs_data <- get_acs(
  geography = "county",
  variables = vars_to_get,
  state = "IL",
  year = 2023,
  survey = "acs5"
)
```

```
Getting data from the 2019-2023 5-year ACS
```

```
head(il_acs_data)
```

```
# A tibble: 6 x 5
  GEOID NAME                       variable   estimate   moe
  <chr> <chr>                      <chr>         <dbl> <dbl>
1 17001 Adams County, Illinois     B19013_001    64962  2634
2 17001 Adams County, Illinois     B28002_001    27770   374
3 17001 Adams County, Illinois     B28002_004    23400   619
4 17003 Alexander County, Illinois B19013_001    43523 10035
5 17003 Alexander County, Illinois B28002_001     1826   109
6 17003 Alexander County, Illinois B28002_004     1068   117
```

Step 3: Clean and Transform Data [14 + 4 points]

a) Use pivot_wider() to create one row per county (with income and broadband side-by side). [6 points]

o BONUS: Split "County, Illinois" into County and State [4 points]

b) Calculate a new variable: broadband_rate = broadband/total_households * 100 [4 points]

c) Arrange counties from highest to lowest broadband access. [4 points]

```
il_data_clean <- il_acs_data %>%
  select(NAME, variable, estimate) %>%
  pivot_wider(
    names_from = variable,
    values_from = estimate
  ) %>%   #  a)
  glimpse() %>%
  # Rename the columns from Census codes to human-readable names
  rename(
    income = "B19013_001",
    broadband = "B28002_004",
    total_households = "B28002_001"
  ) %>%
  separate(NAME, into = c("county", "state"), sep = ", ", remove = TRUE) %>% # bonus
  mutate(broadband_rate = (broadband / total_households) * 100) %>%  # b)
  arrange(desc(broadband_rate))  # c)
```

```
Rows: 102
Columns: 4
$ NAME       <chr> "Adams County, Illinois", "Alexander County, Illinois", "Bo~
$ B19013_001 <dbl> 64962, 43523, 61603, 81638, 72288, 65894, 92095, 60871, 649~
$ B28002_001 <dbl> 27770, 1826, 6256, 19155, 2021, 14011, 1222, 6478, 5135, 83~
$ B28002_004 <dbl> 23400, 1068, 5325, 17051, 1660, 11702, 1052, 5523, 4359, 74~
```

```
head(il_data_clean,12)
```

```
# A tibble: 12 x 6
   county           state    income total_households broadband broadband_rate
   <chr>            <chr>     <dbl>            <dbl>     <dbl>          <dbl>
 1 Kendall County   Illinois 110474            44526     42382           95.2
 2 McHenry County   Illinois 102836           116329    110646           95.1
 3 DuPage County    Illinois 110502           349497    329798           94.4
 4 Lake County      Illinois 108917           256660    240027           93.5
 5 Will County      Illinois 107799           241310    224935           93.2
 6 DeKalb County    Illinois  69022            39314     36455           92.7
 7 Kane County      Illinois 100678           183196    169377           92.5
 8 Monroe County    Illinois 101635            13830     12654           91.5
```

```
 9 Grundy County    Illinois  93060           20518     18549           90.4
10 Madison County   Illinois  74800          109385     98374           89.9
11 Effingham County Illinois  75380           14030     12595           89.8
12 Clinton County   Illinois  82314           14463     12972           89.7
```

Step 4: Analyze patterns [8 points]

a) Compute the mean and median broadband rate across all Illinois counties. [4 points]

```r
summary_stats <- il_data_clean %>%
  summarise(
    mean_broadband_rate = mean(broadband_rate, na.rm = TRUE),
    median_broadband_rate = median(broadband_rate, na.rm = TRUE)
  )

summary_stats
```

```
# A tibble: 1 x 2
  mean_broadband_rate median_broadband_rate
                <dbl>                 <dbl>
1                84.8                  85.4
```

b) Identify the top 5 counties with the highest broadband access and the bottom 5 counties with the lowest. [4 points]

```r
#Since we have already arranged the sequence in step3 c)
top_5_counties <- il_data_clean %>%
  slice_head(n = 5)

top_5_counties
```

```
# A tibble: 5 x 6
  county         state    income total_households broadband broadband_rate
  <chr>          <chr>     <dbl>            <dbl>     <dbl>          <dbl>
1 Kendall County Illinois 110474            44526     42382           95.2
2 McHenry County Illinois 102836           116329    110646           95.1
3 DuPage County  Illinois 110502           349497    329798           94.4
4 Lake County    Illinois 108917           256660    240027           93.5
5 Will County    Illinois 107799           241310    224935           93.2
```

```r
bottom_5_counties <- il_data_clean %>%
  slice_tail(n = 5)

bottom_5_counties
```

```
# A tibble: 5 x 6
  county           state    income total_households broadband broadband_rate
  <chr>            <chr>     <dbl>            <dbl>     <dbl>          <dbl>
1 Saline County    Illinois  54945            10032      7614           75.9
2 Pope County      Illinois  62500             1364       996           73.0
3 Union County     Illinois  56420             6914      4911           71.0
4 Alexander County Illinois  43523             1826      1068           58.5
5 Pulaski County   Illinois  43227             1862      1057           56.8
```

Step 5: Visualize the results [14 + 8 points]

1. Scatterplot: Income (x-axis) vs Broadband Rate (y-axis). [8 points]

a. Add a regression line using geom_smooth(method = "lm").

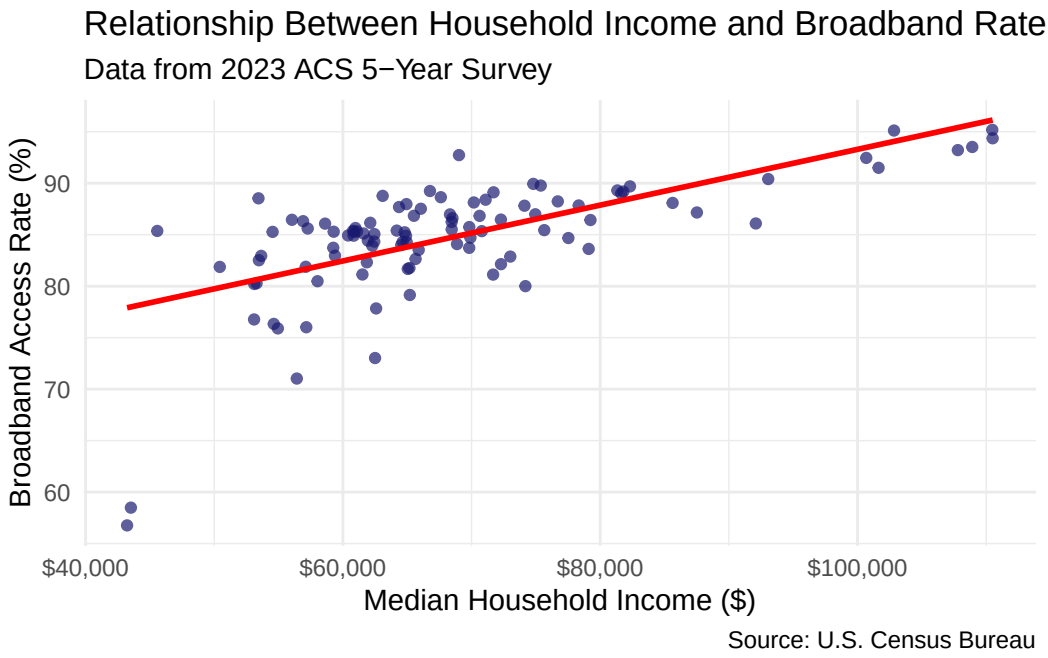
b. Label axes and add an informative title.

```r
library(scales)
```

```
Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```
ggplot(il_data_clean, aes(x = income, y = broadband_rate)) +
  geom_point(alpha = 0.7, color = "midnightblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +    # a)
  scale_x_continuous(labels = dollar_format()) +
  labs(
    title = "Relationship Between Household Income and Broadband Rate in Illinois Counties",
    subtitle = "Data from 2023 ACS 5-Year Survey",
    x = "Median Household Income ($)",
    y = "Broadband Access Rate (%)",
    caption = "Source: U.S. Census Bureau"
  ) +     # b)
  theme_minimal()
```

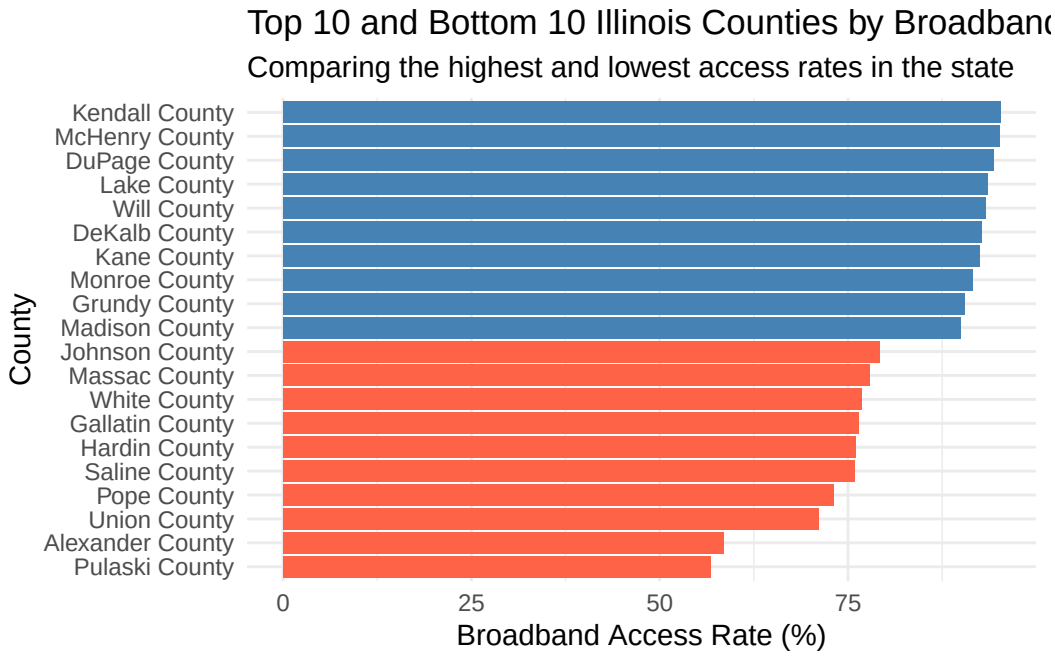
```
`geom_smooth()` using formula = 'y ~ x'
```



2. Bar chart: Show the top 10 and bottom 10 counties by broadband access. HINT: Use the reorder function to order bars. [6 points]

```
top_10 <- slice_head(il_data_clean, n = 10)
bottom_10 <- slice_tail(il_data_clean, n = 10)
top_bottom_10 <- bind_rows(top_10, bottom_10)
```

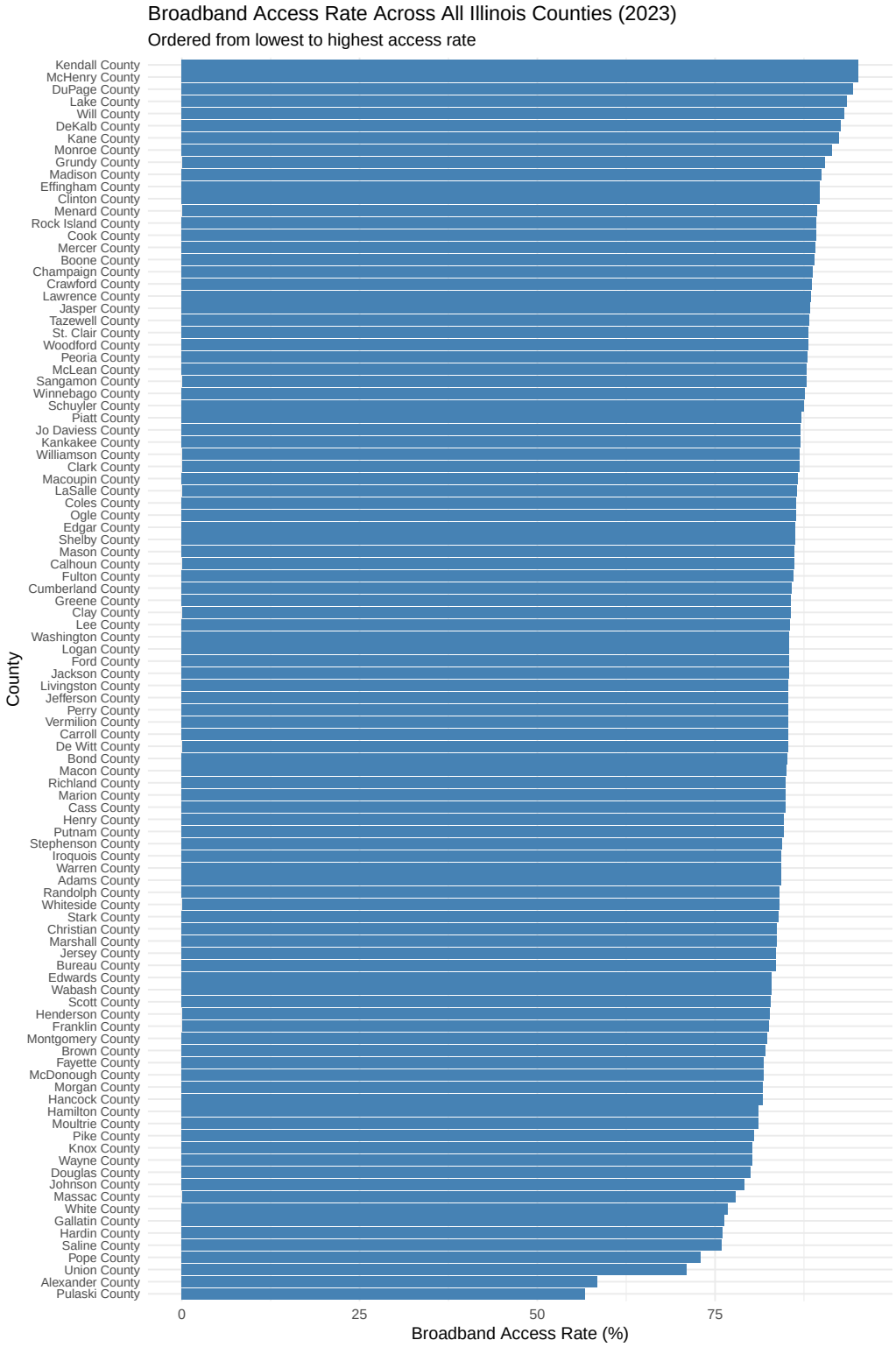
```r
ggplot(top_bottom_10, aes(y = broadband_rate, x = fct_reorder(county, broadband_rate))) +
  geom_col(aes(fill = broadband_rate > 85)) +
  coord_flip() +
  scale_fill_manual(values = c("tomato", "steelblue"), guide = "none") +
  labs(
    title = "Top 10 and Bottom 10 Illinois Counties by Broadband Access Rate",
    subtitle = "Comparing the highest and lowest access rates in the state",
    x = "County",
    y = "Broadband Access Rate (%)"
  ) +
  theme_minimal()
```



3. BONUS: Create a single ordered bar chart of ALL counties by broadband [8 points]

```r
ggplot(il_data_clean, aes(y = broadband_rate, x = fct_reorder(county, broadband_rate))) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Broadband Access Rate Across All Illinois Counties (2023)",
    subtitle = "Ordered from lowest to highest access rate",
    x = "County",
    y = "Broadband Access Rate (%)"
```

```
  ) +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 8))
```


Broadband Access Rate Across All Illinois Counties (2023)
Ordered from lowest to highest access rate
Kendall County
McHenry County
DuPage County
Lake County
Will County
DeKalb County
Kane County
Monroe County
Grundy County
Madison County
Effingham County
Clinton County
Menard County
Rock Island County
Cook County
Mercer County
Boone County
Champaign County
Crawford County
Lawrence County
Jasper County
Tazewell County
St. Clair County
Woodford County
Peoria County
McLean County
Sangamon County
Winnebago County
Schuyler County
Piatt County
Jo Daviess County
Kankakee County
Williamson County
Clark County
Macoupin County
LaSalle County
Coles County
Ogle County
Edgar County
Shelby County
Mason County
Calhoun County
Fulton County
Cumberland County
Greene County
Clay County
Lee County
Washington County
Logan County
Ford County
Jackson County
Livingston County
Jefferson County
Perry County
Vermilion County
Carroll County
De Witt County
Bond County
Macon County
Richland County
Marion County
Cass County
Henry County
Putnam County
Stephenson County
Iroquois County
Warren County
Adams County
Randolph County
Whiteside County
Stark County
Christian County
Marshall County
Jersey County
Bureau County
Edwards County
Wabash County
Scott County
Henderson County
Franklin County
Montgomery County
Brown County
Fayette County
McDonough County
Morgan County
Hancock County
Hamilton County
Moultrie County
Pike County
Knox County
Wayne County
Douglas County
Johnson County
Massac County
White County
Gallatin County
Hardin County
Saline County
Pope County
Union County
Alexander County
Pulaski County
County
0
25
50
75
Broadband Access Rate (%)

Step 6: Reflection [16 points]

Write 3-4 sentences per question:

1. What patterns do you observe between income and broadband access? [4 points]

   Based on the scatterplot, there is a clear and strong positive correlation between median household income and broadband access rates across Illinois counties. As a county's median household income increases, its broadband access rate tends to increase as well, which is confirmed by the upward-sloping regression line. The data points cluster more tightly at the higher end of the income and access spectrum, while counties with the lowest incomes, such as Alexander and Pulaski, also exhibit the lowest broadband access rates. This pattern suggests that economic prosperity is a significant factor associated with digital connectivity at the county level.

2. What might explain the variation in broadband access across counties? [4 points]

   The variation in broadband access is likely driven by a combination of economic and geographic factors. Higher-income counties, which are often more densely populated suburban areas like DuPage and Lake County, represent a more profitable market for internet service providers, encouraging investment in modern infrastructure. Conversely, lower-income counties are often more rural and less populated, leading to higher per-household costs for infrastructure deployment and lower expected returns, which deters private investment. This creates a cycle where wealthier areas receive better service, while poorer, rural areas like Pope and Saline County are left with insufficient access and limited provider choice.

3. How could public administrators use this data to inform digital inclusion policies? [4 points]

   Public administrators can use this analysis as a powerful tool for targeted intervention and policy-making. The data clearly identifies which counties are most in need, allowing administrators to direct state and federal grants for infrastructure projects to the areas with the lowest access rates, such as Pulaski and Alexander counties. Furthermore, the strong correlation between income and access provides evidence to support policies that go beyond infrastructure, such as creating subsidy programs to make internet services more affordable for low-income households. This data provides an empirical baseline for setting policy goals and measuring the effectiveness of digital equity initiatives over time.

4. What are some limitations of using ACS data for local decision-making? [4 points]

While invaluable, the ACS data has several limitations for local decision-making. First, as it is based on a survey, the data consists of estimates that come with a margin of error (moe), which can be particularly large for counties with smaller populations, making the figures less precise. Second, the 5-year estimates (2019-2023) may not reflect the most recent infrastructure changes, potentially lagging behind real-time conditions. Finally, county-level data can mask significant disparities within a county, where an affluent town and an underserved rural area might exist side-by-side, a nuance lost in the aggregated data.