Project Exploration

Joyce Yan (qy249), Lydia Lin (dl2253), Victoria Xiao (sx287)

2024-11-14

Objective(s)

The objective of this project is to develop a model that predicts the count of adult arrests (ages 18 and older) in New York State by offense type, based on county-level social, economic, and housing characteristics from 2015 to 2022. This model aims to identify the key socioeconomic factors that influence arrest rates, and to provide insights that could help policymakers understand the relationship between community conditions and criminal activity.

Data collection and cleaning

Data Collection

This project utilizes four datasets obtained from two government organizations, Data.NY.gov and the United States Census Bureau.

Sources:

https://data.ny.gov/Public-Safety/Adult-Arrests-18-and-Older-by-County-Beginning-197/rikd-mt35/about_data https://data.census.gov/profile/New_York?g=040XX00US36

The first dataset, "New_Arrests.csv," includes adult arrest records (ages 18 and older) by county in New York State, beginning from 1970. We downloaded this dataset from Data.NY.gov. To focus on recent trends, we filtered it to cover only the years 2015 - 2022.

The other three datasets, "social_data," "economic_data," and "housing_data," were created from raw data provided by the United States Census Bureau, specifically selecting information for New York State to align with the arrest dataset. We chose three key indicators—social, economic, and housing characteristics. Since the data for each indicator was originally separated by year, we consolidated the datasets for each characteristic into a single dataset covering 2015 - 2022. Finally, we merged all three characteristic datasets with the original arrest dataset by joining them on county name and year.

Below is the code we used for data collection and cleaning.

Data cleaning

Set up:

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(readr)
library(ggplot2)
library(reshape2)
```

Social Characteristics:

```
# Add variables from Social Characteristics in the United States census dataset.
# Source: https://data.census.gov/table?g=040XX00US36,36$0500000_050XX00US36047&d=ACS%205-Yea

# Load each CSV file from the social_data folder
social_data <- list.files(path = "data/social_data", full.names = TRUE, pattern = "*.csv") |>
  lapply(function(file) {
    year <- as.numeric(gsub(".*(\\d{4}).*", "\\1", file))
    data <- read_csv(file) |>
      select(County = NAME,                          # Rename NAME to County
             Less_than_9th_grade = DP02_0059E,       # Educational Attainment
             With_disability = DP02_0071E,           # Disability Status
             Civilian_veterans = DP02_0069E,         # Veteran Status
             Limited_English = DP02_0113E) |>        # Language spoken at home, limited Englis
      mutate(
        County = sub(" County.*", "", County), # Keep only the part before " County"
        Year = year # Add the extracted Year as a new column
```

```
    ) |>
    slice(-1) # Drop the first row, as it contains metadata information from the original c
}) |>
bind_rows() # Combine all years into one dataset
```

```
New names:
Rows: 64 Columns: 611
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(610): GEO_ID, NAME, DP02_0001E, DP02_0001M, DP02_0001PE, DP02_0001PM, D... lgl
(1): ...611
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 611
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(610): GEO_ID, NAME, DP02_0001E, DP02_0001M, DP02_0001PE, DP02_0001PM, D... lgl
(1): ...611
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 611
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(610): GEO_ID, NAME, DP02_0001E, DP02_0001M, DP02_0002E, DP02_0002M, DP0... lgl
(1): ...611
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 611
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(610): GEO_ID, NAME, DP02_0001E, DP02_0001M, DP02_0002E, DP02_0002M, DP0... lgl
(1): ...611
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 615
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(614): GEO_ID, NAME, DP02_0001E, DP02_0001M, DP02_0002E, DP02_0002M, DP0... lgl
```

```
(1): ...615
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 619
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(618): GEO_ID, NAME, DP02_0001E, DP02_0001M, DP02_0002E, DP02_0002M, DP0... lgl
(1): ...619
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 619
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(618): GEO_ID, NAME, DP02_0001E, DP02_0001M, DP02_0002E, DP02_0002M, DP0... lgl
(1): ...619
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 619
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(618): GEO_ID, NAME, DP02_0001E, DP02_0001M, DP02_0002E, DP02_0002M, DP0... lgl
(1): ...619
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...611`
```

```r
print(names(social_data)) # Check for 'County' and 'Year'
```

```
[1] "County"              "Less_than_9th_grade" "With_disability"
[4] "Civilian_veterans"   "Limited_English"     "Year"
```

```r
# Load Adult Arrests data(from 2015 to 2022)
adult_arrests <- read_csv("data/New_Arrests.csv") |>
  filter(Year >= 2015 & Year <= 2022) # Filter to years between 2015 and 2022
```

```
Rows: 496 Columns: 13
-- Column specification --------------------------------------------------------
Delimiter: ","
```

```
chr  (1): County
dbl (12): Year, Total, Felony Total, Drug Felony, Violent Felony, DWI Felony...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Perform the left join with adult_arrests and social_data on County and Year
merged_data <- adult_arrests |>
  left_join(social_data, by = c("County", "Year"))

# Display the merged data to check the result
head(merged_data)
```

```
# A tibble: 6 x 17
  County   Year Total `Felony Total` `Drug Felony` `Violent Felony` `DWI Felony`
  <chr>   <dbl> <dbl>          <dbl>         <dbl>             <dbl>        <dbl>
1 Albany   2015  7412           2590           509               533          161
2 Allega~  2015   792            209            33                39           38
3 Bronx    2015 59576          16121          4333              5623          156
4 Broome   2015  4968           1532           216               389           91
5 Cattar~  2015  1516            439           103                86           47
6 Cayuga   2015  1327            431            68                57           27
# i 10 more variables: `Other Felony` <dbl>, `Misdemeanor Total` <dbl>,
#   `Drug Misdemeanor` <dbl>, `DWI Misdemeanor` <dbl>,
#   `Property Misdemeanor` <dbl>, `Other Misdemeanor` <dbl>,
#   Less_than_9th_grade <chr>, With_disability <chr>, Civilian_veterans <chr>,
#   Limited_English <chr>
```

```r
# Check the total rows
nrow(merged_data)
```

```
[1] 504
```

Housing Characteristics:

```r
# Add variables from Housing Characteristics in the United States census dataset.
# Source: https://data.census.gov/table/ACSDP5Y2021.DP04?g=040XX00US36,36$0500000_050XX00US3

housing_data <- list.files(path = "data/housing_data", full.names = TRUE, pattern = "*.csv")
  lapply(function(file) {
    year <- as.numeric(gsub(".*(\\d{4}).*", "\\1", file))
```

```
    data <- read_csv(file) |>
      select(County = NAME,                          # Rename NAME to County
             Total_housing_units = DP04_0001E,       # Total housing units of occupancy
             With_mortgage = DP04_0091E,             # Housing with mortgage
             Without_mortgage = DP04_0092E) |>       # Housing without mortgage
      mutate(
        County = sub(" County.*", "", County), # Keep only the part before " County"
        Year = year # Add the extracted Year as a new column
      ) |>
      slice(-1) # Drop the first row, as it contains metadata information from the original
  }) |>
  bind_rows() # Combine all years into one dataset
```

```
New names:
Rows: 64 Columns: 575
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(574): GEO_ID, NAME, DP04_0001E, DP04_0001M, DP04_0001PE, DP04_0001PM, D... lgl
(1): ...575
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 575
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(574): GEO_ID, NAME, DP04_0001E, DP04_0001M, DP04_0001PE, DP04_0001PM, D... lgl
(1): ...575
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 575
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(574): GEO_ID, NAME, DP04_0001E, DP04_0001M, DP04_0002E, DP04_0002M, DP0... lgl
(1): ...575
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 575
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(574): GEO_ID, NAME, DP04_0001E, DP04_0001M, DP04_0002E, DP04_0002M, DP0... lgl
```

```
(1): ...575
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 575
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(574): GEO_ID, NAME, DP04_0001E, DP04_0001M, DP04_0002E, DP04_0002M, DP0... lgl
(1): ...575
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 575
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(574): GEO_ID, NAME, DP04_0001E, DP04_0001M, DP04_0002E, DP04_0002M, DP0... lgl
(1): ...575
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 575
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(574): GEO_ID, NAME, DP04_0001E, DP04_0001M, DP04_0002E, DP04_0002M, DP0... lgl
(1): ...575
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 575
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(574): GEO_ID, NAME, DP04_0001E, DP04_0001M, DP04_0002E, DP04_0002M, DP0... lgl
(1): ...575
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...575`
```

```r
print(names(housing_data)) # Check for 'County' and 'Year'
```

```
[1] "County"              "Total_housing_units" "With_mortgage"
[4] "Without_mortgage"    "Year"
```

```r
# Perform the left join between adult_arrests and housing_data on County and Year
merged_data <- merged_data |>
  left_join(housing_data, by = c("County", "Year"))
```

```
Warning in left_join(merged_data, housing_data, by = c("County", "Year")): Detected an unexpe
i Row 31 of `x` matches multiple rows in `y`.
i Row 1 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

```r
# Display the merged data to check the result
head(merged_data)
```

```
# A tibble: 6 x 20
  County   Year Total `Felony Total` `Drug Felony` `Violent Felony` `DWI Felony`
  <chr>   <dbl> <dbl>          <dbl>         <dbl>            <dbl>        <dbl>
1 Albany   2015  7412           2590           509              533          161
2 Allega~  2015   792            209            33               39           38
3 Bronx    2015 59576          16121          4333             5623          156
4 Broome   2015  4968           1532           216              389           91
5 Cattar~  2015  1516            439           103               86           47
6 Cayuga   2015  1327            431            68               57           27
# i 13 more variables: `Other Felony` <dbl>, `Misdemeanor Total` <dbl>,
#   `Drug Misdemeanor` <dbl>, `DWI Misdemeanor` <dbl>,
#   `Property Misdemeanor` <dbl>, `Other Misdemeanor` <dbl>,
#   Less_than_9th_grade <chr>, With_disability <chr>, Civilian_veterans <chr>,
#   Limited_English <chr>, Total_housing_units <chr>, With_mortgage <chr>,
#   Without_mortgage <chr>
```

```r
# Check the total rows
nrow(merged_data)
```

```
[1] 520
```

Economic Characteristics:

```r
# Add variables from Economic Characteristics in the United States census dataset.
# Source: https://data.census.gov/table?g=040XX00US36,36$0500000_050XX00US36047&d=ACS%205-Yea
```

```r
economic_data <- list.files(path = "data/economic_data", full.names = TRUE, pattern = "*.csv
  lapply(function(file) {
    year <- as.numeric(gsub(".*(\\d{4}).*", "\\1", file))
    data <- read_csv(file) |>
      select(County = NAME,
             Unemployment_rate = DP03_0009PE,
             Median_household_income = DP03_0062E,
             Below_poverty_level = DP03_0133PE,      # Percentage of people whose income is l
             No_health_insurance = DP03_0099PE) |>   # Percentage of non-institutionalized po
      mutate(
        County = sub(" County.*", "", County), # Keep only the part before " County"
        Year = year # Add the extracted Year as a new column
      ) |>
      slice(-1) # Drop the first row, as it contains metadata information from the original d
  }) |>
  bind_rows() # Combine all years into one dataset
```

```
New names:
Rows: 64 Columns: 551
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(550): GEO_ID, NAME, DP03_0001E, DP03_0001M, DP03_0001PE, DP03_0001PM, D... lgl
(1): ...551
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 551
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(550): GEO_ID, NAME, DP03_0001E, DP03_0001M, DP03_0001PE, DP03_0001PM, D... lgl
(1): ...551
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 551
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(550): GEO_ID, NAME, DP03_0001E, DP03_0001M, DP03_0002E, DP03_0002M, DP0... lgl
(1): ...551
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
```

```
Rows: 64 Columns: 551
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(550): GEO_ID, NAME, DP03_0001E, DP03_0001M, DP03_0002E, DP03_0002M, DP0... lgl
(1): ...551
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 551
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(550): GEO_ID, NAME, DP03_0001E, DP03_0001M, DP03_0002E, DP03_0002M, DP0... lgl
(1): ...551
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 551
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(550): GEO_ID, NAME, DP03_0001E, DP03_0001M, DP03_0002E, DP03_0002M, DP0... lgl
(1): ...551
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 551
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(550): GEO_ID, NAME, DP03_0001E, DP03_0001M, DP03_0002E, DP03_0002M, DP0... lgl
(1): ...551
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
Rows: 64 Columns: 551
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(550): GEO_ID, NAME, DP03_0001E, DP03_0001M, DP03_0002E, DP03_0002M, DP0... lgl
(1): ...551
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...551`
```

```r
print(names(economic_data)) # Check for 'County' and 'Year'
```

```
[1] "County"                  "Unemployment_rate"
[3] "Median_household_income" "Below_poverty_level"
[5] "No_health_insurance"     "Year"
```

```r
# Perform the left join with adult_arrests and economic_data on County and Year
merged_data <- merged_data |>
  left_join(economic_data, by = c("County", "Year"))
```

```
Warning in left_join(merged_data, economic_data, by = c("County", "Year")): Detected an unexp
i Row 31 of `x` matches multiple rows in `y`.
i Row 1 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

```r
# Display the merged data to check the result
head(merged_data)
```

```
# A tibble: 6 x 24
  County   Year Total `Felony Total` `Drug Felony` `Violent Felony` `DWI Felony`
  <chr>   <dbl> <dbl>          <dbl>         <dbl>            <dbl>        <dbl>
1 Albany   2015  7412           2590           509              533          161
2 Allega~  2015   792            209            33               39           38
3 Bronx    2015 59576          16121          4333             5623          156
4 Broome   2015  4968           1532           216              389           91
5 Cattar~  2015  1516            439           103               86           47
6 Cayuga   2015  1327            431            68               57           27
# i 17 more variables: `Other Felony` <dbl>, `Misdemeanor Total` <dbl>,
#   `Drug Misdemeanor` <dbl>, `DWI Misdemeanor` <dbl>,
#   `Property Misdemeanor` <dbl>, `Other Misdemeanor` <dbl>,
#   Less_than_9th_grade <chr>, With_disability <chr>, Civilian_veterans <chr>,
#   Limited_English <chr>, Total_housing_units <chr>, With_mortgage <chr>,
#   Without_mortgage <chr>, Unemployment_rate <chr>,
#   Median_household_income <chr>, Below_poverty_level <chr>, ...
```

```r
# Check the total rows
nrow(merged_data)
```

```
[1] 552
```

Convert the non-numeric columns to numeric:

```
cols_to_convert <- c("Less_than_9th_grade", "With_disability", "Civilian_veterans",
                     "Limited_English", "Total_housing_units", "With_mortgage",
                     "Without_mortgage", "Unemployment_rate", "Median_household_income",
                     "Below_poverty_level", "No_health_insurance")

merged_data[cols_to_convert] <- lapply(
  merged_data[cols_to_convert], function(x) as.numeric(as.character(x))
  )
```

Save the cleaned dataset:

```
# Save merged_data to the data folder
write_csv(merged_data, "data/merged_data.csv")
```

Data Appendix

The goal of data cleaning is to prepare an analysis-ready dataset that consolidates social, economic, and housing characteristics data with adult arrest data.

Loading and Organizing Files: We loaded each of the seven yearly CSV files (one for each year from 2015 to 2022) from the social_data, economic_data, and housing_data folders. Since each file name includes the year, we extracted the year from the file name to use as a new Year column for each dataset. Each dataset was read into R, and only relevant columns from the characteristic datasets were selected:

- Social: DP02_0059E (Less than 9th-grade education), DP02_0071E (Population with a disability), DP02_0069E (Veteran status), DP02_0113E (Limited English proficiency)
- Economic: DP03_0009PE (Unemployment rate), DP03_0062E (Median household income), DP03_0133PE (Percentage of people whose income is below the poverty level), DP03_0133PE (Percentage of non-institutionalized population without health insurance)
- Housing: DP04_0001E (Total housing units), DP04_0091E (Housing with mortgage), DP04_0092E (Housing without mortgage)

These columns represent key social, economic, and housing characteristics that we intend to analyze.

Standardizing County Names: Each file contained a NAME column representing the county and state in the format "XXX County, New York." We cleaned this column to retain only the county name and removed any extraneous text. Additionally, "County" was removed from each county name to standardize the format and match it with county names in the adult arrests data (e.g., "Kings County, New York" was transformed to "Kings").

Combining Datasets: After standardizing county names and adding the Year column, each dataset was appended to form a single dataframe (social_data, economic_data, and housing_data) containing characteristics data from 2015 to 2022.

Merging with Adult Arrest Data: We loaded the adult arrest data and extracted specific year ranges from Adult_Arrests.csv to New_Arrests.csv, which includes county-level adult arrest data for New York from 2015 to 2022. Using the left_join function in R, we merged each characteristics dataset with adult_arrests based on County and Year to create a comprehensive dataset (merged_data). This dataset includes both social characteristics and arrest data for each county-year pair.

Saving the Final Dataset: The merged dataset was saved as merged_data.csv in the data folder for further analysis.

Data description

The merged_data.csv file is an analysis-ready dataset that combines social, economic, and housing characteristics with adult arrest data for New York State counties from 2015 to 2022. It includes the following columns:

- County: Name of the county in New York State.
- Year: The year of the data, from 2015 to 2022.
- Total (from Arrests data): Total adult arrests in each county for the respective year.
- Felony Total: Total felony arrests.
- Drug Felony: Number of drug-related felony arrests.
- Violent Felony: Number of violent felony arrests.
- DWI Felony: Number of felony arrests for driving while intoxicated (DWI).
- Other Felony: Other types of felony arrests.
- Misdemeanor Total: Total misdemeanor arrests.
- Drug Misdemeanor: Number of drug-related misdemeanor arrests.
- DWI Misdemeanor: Number of misdemeanor arrests for DWI.
- Property Misdemeanor: Number of property-related misdemeanor arrests.
- Other Misdemeanor: Other types of misdemeanor arrests.
- Less_than_9th_grade: Estimate of population aged 25 years and over with less than a 9th-grade education.
- With_disability: Estimate of the total civilian noninstitutionalized population with a disability.
- Civilian_veterans: Estimate of the civilian population aged 18 years and over who are veterans.
- Limited_English: Estimate of the population aged 5 years and over who speak a language other than English at home and speak English less than “very well.”
- Total_housing_units: Estimate of the total housing units in terms of occupancy.

- With_mortgage: Estimate of the number of owner-occupied housing units that have a mortgage. Essentially, it is the count of homes occupied by their owners where there is still a mortgage loan on the property.
- Without_mortgage: Estimate of the number of owner-occupied housing units that do not have a mortgage.
- Unemployment_rate: The percentage of the civilian labor force that is unemployed.
- Median_household_income: The median income of all households in the county, adjusted for 2022 inflation.
- Below_poverty_level: The percentage of individuals aged 18 and over whose income is below the poverty threshold.
- No_health_insurance: The percentage of the civilian non-institutionalized population without health insurance coverage.

Data limitations

1. The dataset includes only counties within New York State, so findings may not be generalizable to other states or regions.
2. To focus on recent trends, we included data from 2015 to 2022 for social, economic, housing characteristics, and adult arrests. It may not capture the long-term trends and cyclical patterns, which limits historical comparisons.
3. Many values are estimates from the American Community Survey (ACS) and may contain sampling errors. Margins of error are not included, so precision of the estimates cannot be directly evaluated.
4. Characteristics data are aggregated annually, which prevents tracking changes at the individual level over time.
5. Economic, social, and housing indicators may correlate with arrest rates, but they don't necessarily indicate causation. Many external factors (e.g., policing policies, state-level legislative changes) could influence arrest trends independently of socioeconomic factors.

Exploratory data analysis

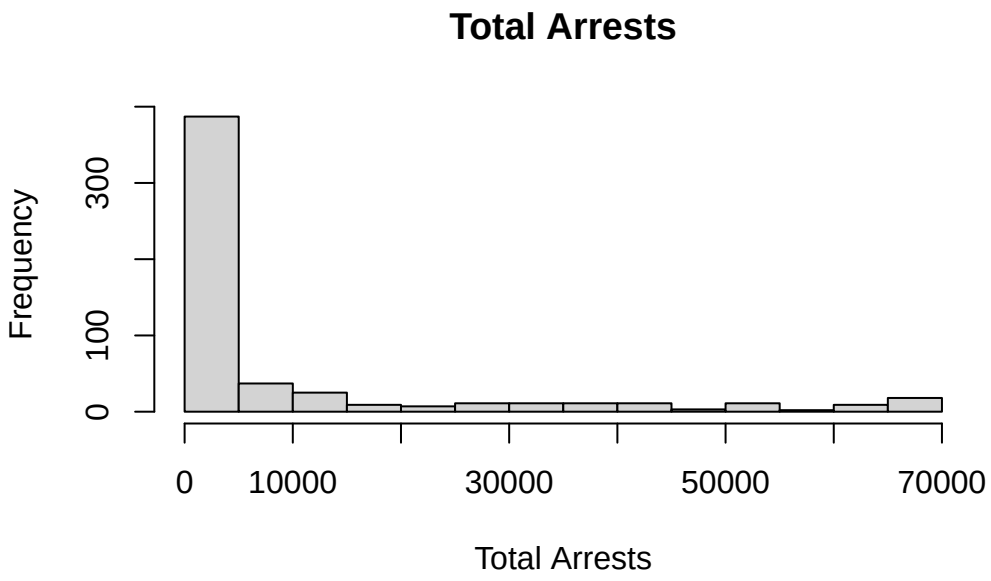
```
summary(merged_data)
```

```
   County               Year          Total         Felony Total
 Length:552         Min.   :2015   Min.   :   34   Min.   :    8.0
 Class :character   1st Qu.:2017   1st Qu.: 1022   1st Qu.:  295.8
 Mode  :character   Median :2018   Median : 1878   Median :  577.5
                    Mean   :2018   Mean   :10249   Mean   : 3447.4
```

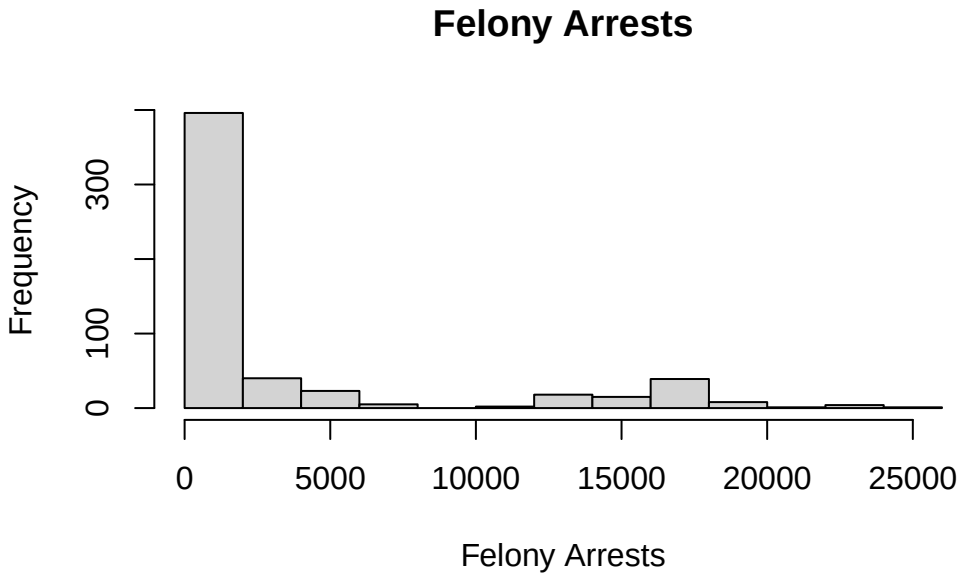
```
                  3rd Qu.:2020   3rd Qu.: 8743   3rd Qu.: 3070.5
                  Max.   :2022   Max.   :68475   Max.   :25031.0
 Drug Felony     Violent Felony     DWI Felony      Other Felony
Min.   :   0.0   Min.   :   1.0   Min.   :  0.00   Min.   :    3.0
1st Qu.:  44.0   1st Qu.:  51.0   1st Qu.: 29.00   1st Qu.:  162.8
Median :  96.0   Median : 102.5   Median : 46.00   Median :  320.5
Mean   : 515.6   Mean   : 959.6   Mean   : 77.79   Mean   : 1894.4
3rd Qu.: 483.5   3rd Qu.: 796.2   3rd Qu.: 99.00   3rd Qu.: 1634.8
Max.   :4472.0   Max.   :8111.0   Max.   :564.00   Max.   :13476.0
Misdemeanor Total Drug Misdemeanor   DWI Misdemeanor  Property Misdemeanor
Min.   :   26.0   Min.   :    1.00   Min.   :   5.0   Min.   :    1.0
1st Qu.:  705.2   1st Qu.:   79.75   1st Qu.: 145.5   1st Qu.:  178.0
Median : 1287.0   Median :  185.00   Median : 273.0   Median :  409.5
Mean   : 6801.7   Mean   : 1270.01   Mean   : 496.5   Mean   : 2535.3
3rd Qu.: 5768.2   3rd Qu.:  770.50   3rd Qu.: 648.0   3rd Qu.: 1708.8
Max.   :50228.0   Max.   :11861.00   Max.   :3186.0   Max.   :24088.0
Other Misdemeanor Less_than_9th_grade With_disability    Civilian_veterans
Min.   :    8.0   Min.   :      88    Min.   :     809   Min.   :     313
1st Qu.:  250.8   1st Qu.:    2306    1st Qu.:   12885   1st Qu.:    6714
Median :  421.0   Median :   34307    Median :   50758   Median :   37160
Mean   : 2499.8   Mean   :  562865    Mean   :  827845   Mean   :  626234
3rd Qu.: 1996.2   3rd Qu.:  112346    3rd Qu.:  167157   3rd Qu.:   92971
Max.   :17157.0   Max.   :14081080    Max.   :19878007   Max.   :15872052
Limited_English    Total_housing_units With_mortgage     Without_mortgage
Min.   :      63   Min.   :   7920     Min.   :    426   Min.   :    498
1st Qu.:    1739   1st Qu.:  28510     1st Qu.:   8268   1st Qu.:   7435
Median :   34215   Median :  49666     Median :  15521   Median :  11549
Mean   :  519644   Mean   : 641052     Mean   : 180095   Mean   : 116497
3rd Qu.:  152913   3rd Qu.: 181415     3rd Qu.:  61500   3rd Qu.:  36514
Max.   :13177639   Max.   :8494452     Max.   :2462215   Max.   :1695945
Unemployment_rate Median_household_income Below_poverty_level
Min.   : 2.100    Min.   : 34299          Min.   : 5.00
1st Qu.: 5.100    1st Qu.: 52546          1st Qu.:10.30
Median : 5.900    Median : 59304          Median :12.20
Mean   : 6.137    Mean   : 63631          Mean   :12.32
3rd Qu.: 7.000    3rd Qu.: 70431          3rd Qu.:13.90
Max.   :14.000    Max.   :137709          Max.   :26.60
No_health_insurance
Min.   : 2.30
1st Qu.: 4.30
Median : 5.40
Mean   : 5.97
3rd Qu.: 6.90
```

```
 Max.   :21.10
```

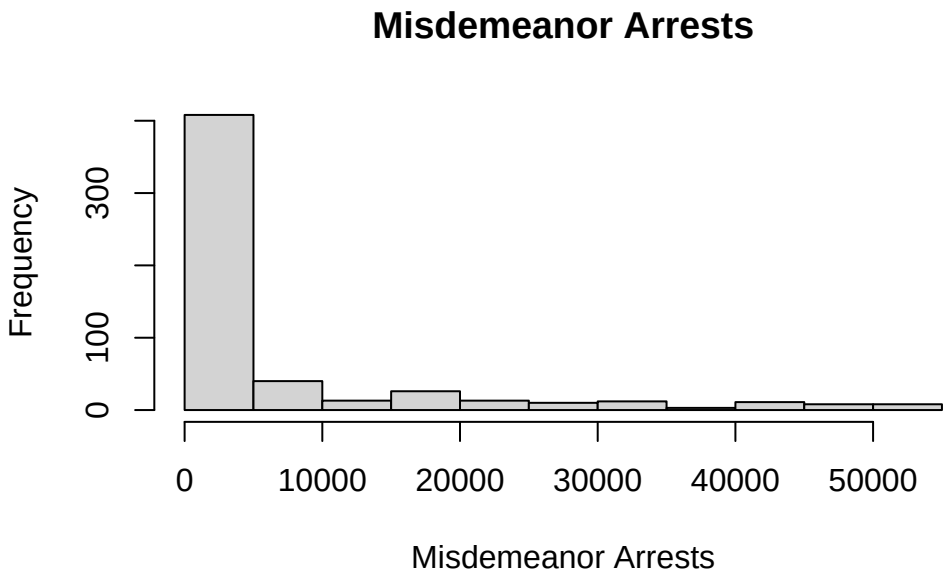
```r
# Histograms for key variables
# Display distributions for Total Arrests, Felony Arrests, Misdemeanor Arrests,
# Unemployment Rate, Median Household Income, With Disability, and Total Housing Units

hist(merged_data$Total, main = "Total Arrests", xlab = "Total Arrests")
```



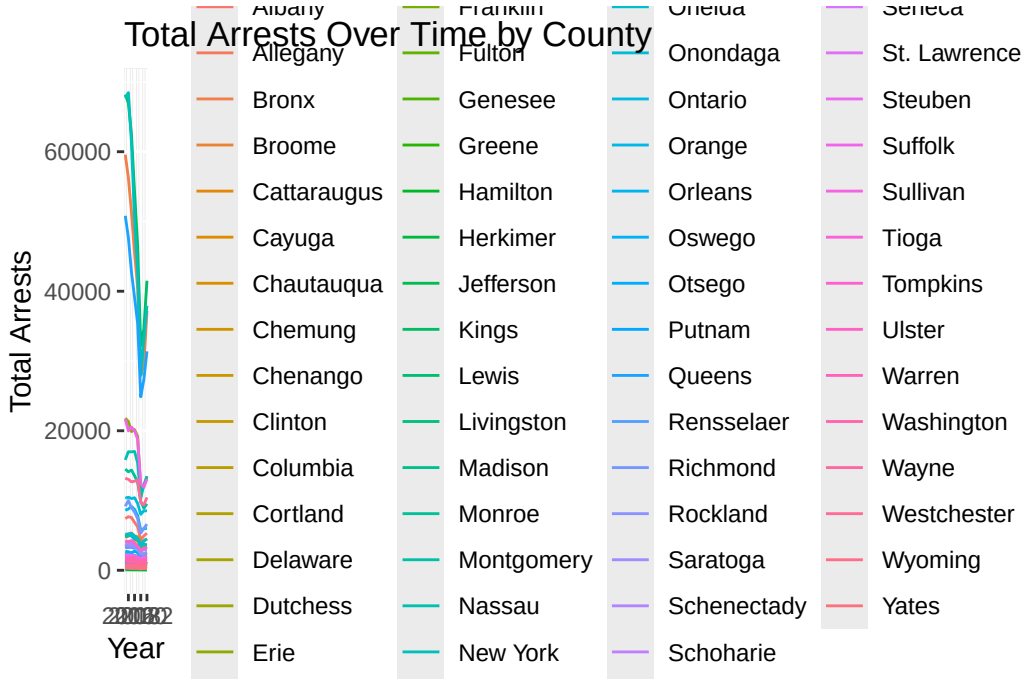

```r
hist(merged_data$`Felony Total`, main = "Felony Arrests", xlab = "Felony Arrests")
```



```r
hist(merged_data$`Misdemeanor Total`, main = "Misdemeanor Arrests", xlab = "Misdemeanor Arres
```

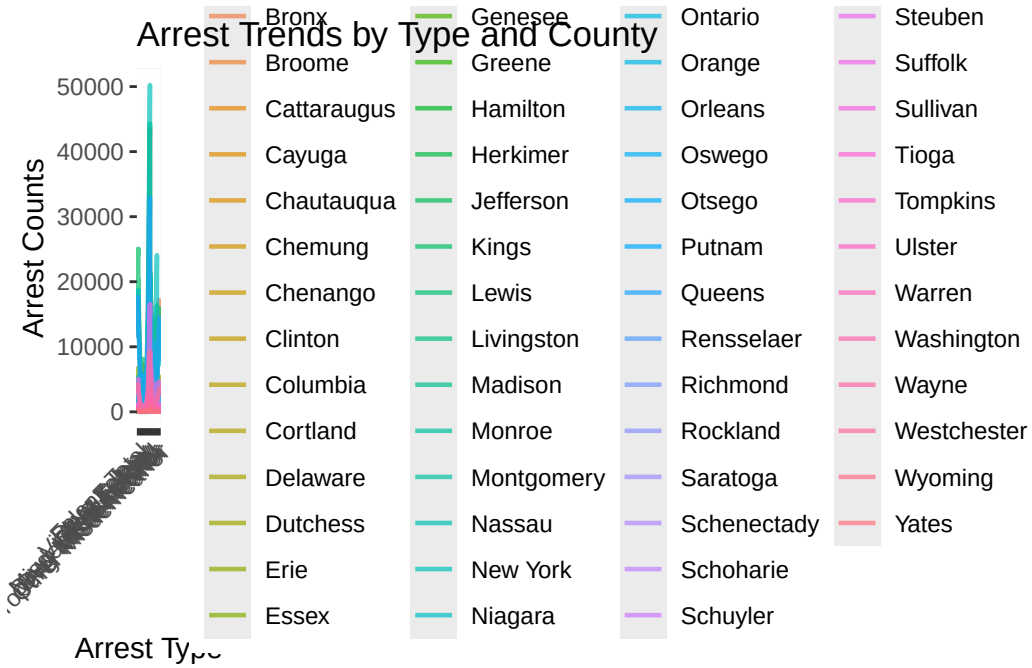



```r
# Transform data to long format
arrests_stacked <- melt(
  merged_data,
  id.vars = c("County", "Year"),
  measure.vars = c("Felony Total", "Drug Felony", "Violent Felony", "DWI Felony", "Other Felo
                   "Misdemeanor Total", "Drug Misdemeanor", "DWI Misdemeanor", "Property Misc
  variable.name = "Arrest_Type",
  value.name = "Count"
  )
# Line plot for total arrests over time
ggplot(merged_data, aes(x = Year, y = Total, color = County)) +
  geom_line() +
  labs(title = "Total Arrests Over Time by County", x = "Year", y = "Total Arrests")
```

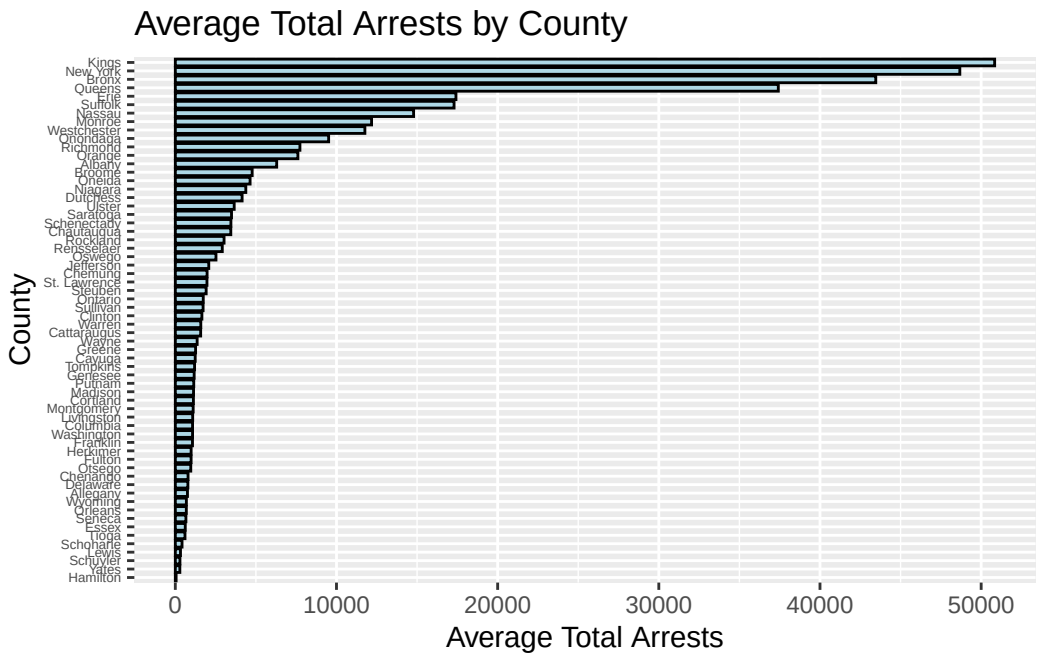



```
#Grouped Line Plot for Arrest Trends by Type and County
ggplot(arrests_stacked, aes(x = Arrest_Type, y = Count, color = County, group = County)) +
  geom_line(alpha = 0.7, size = 0.8) + # Adjust line transparency and width
  labs(title = "Arrest Trends by Type and County", x = "Arrest Type", y = "Arrest Counts") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1) # Rotate x-axis labels
  )
```

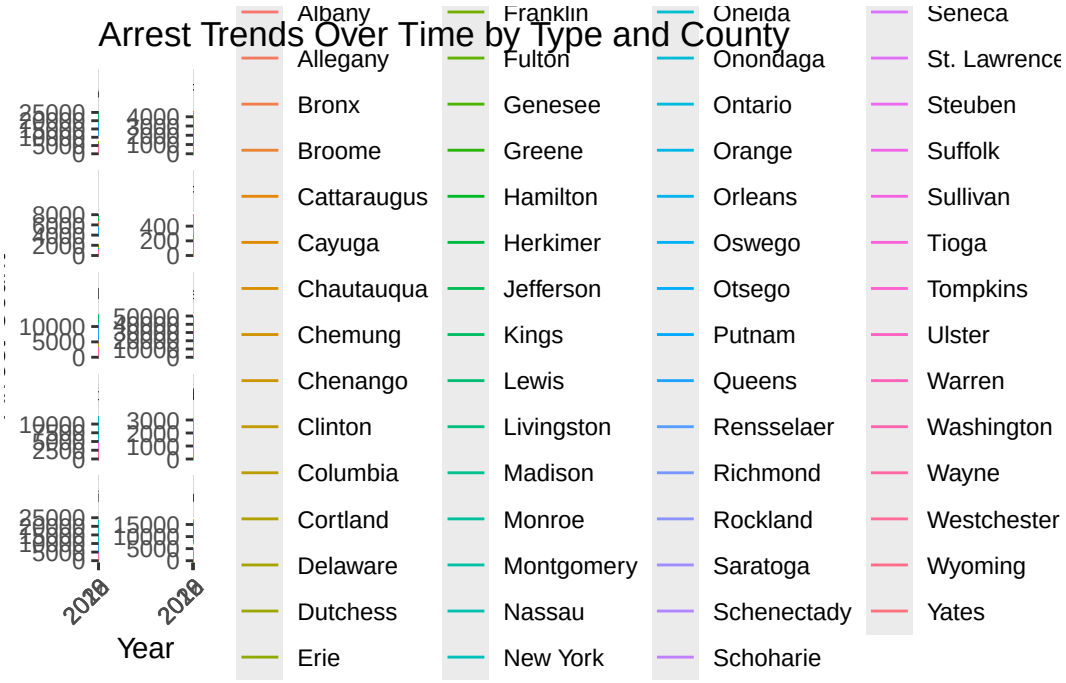
```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



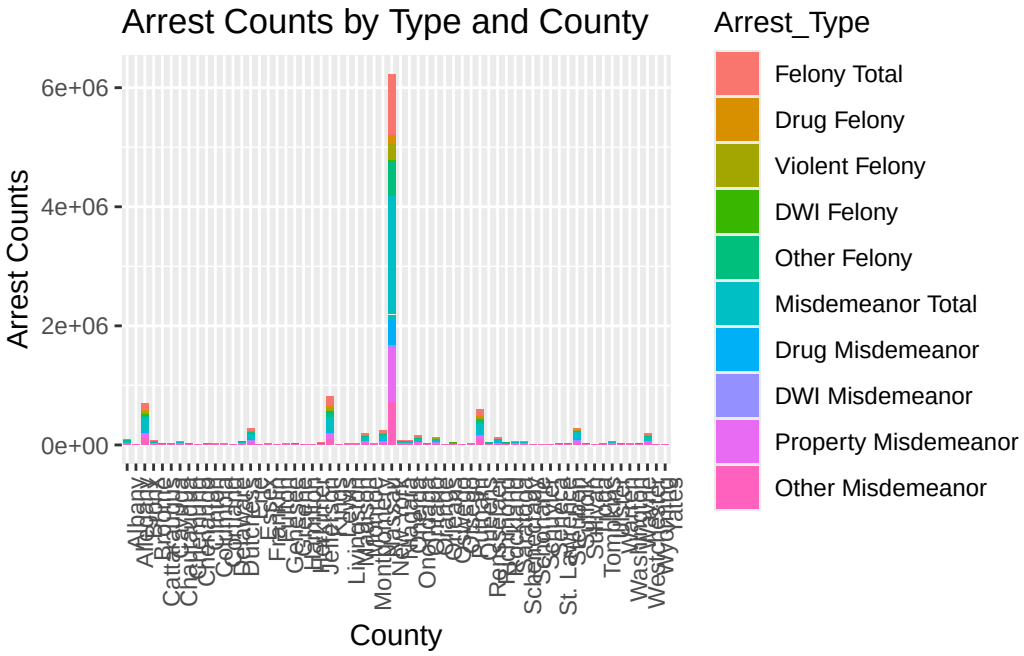

```
# Bar plot for average arrests by county
  # Average total arrests by county
county_summary <- merged_data |>
  group_by(County) |>
  summarize(Average_Total_Arrests = mean(Total, na.rm = TRUE))
ggplot(county_summary, aes(x = reorder(County, Average_Total_Arrests), y = Average_Total_Arre
  geom_bar(stat = "identity", fill = "lightblue", color = "black", width = 0.8) +
  coord_flip() +
  labs(title = "Average Total Arrests by County", x = "County", y = "Average Total Arrests")
  theme(
    axis.text.y = element_text(size = 5)
    )
```



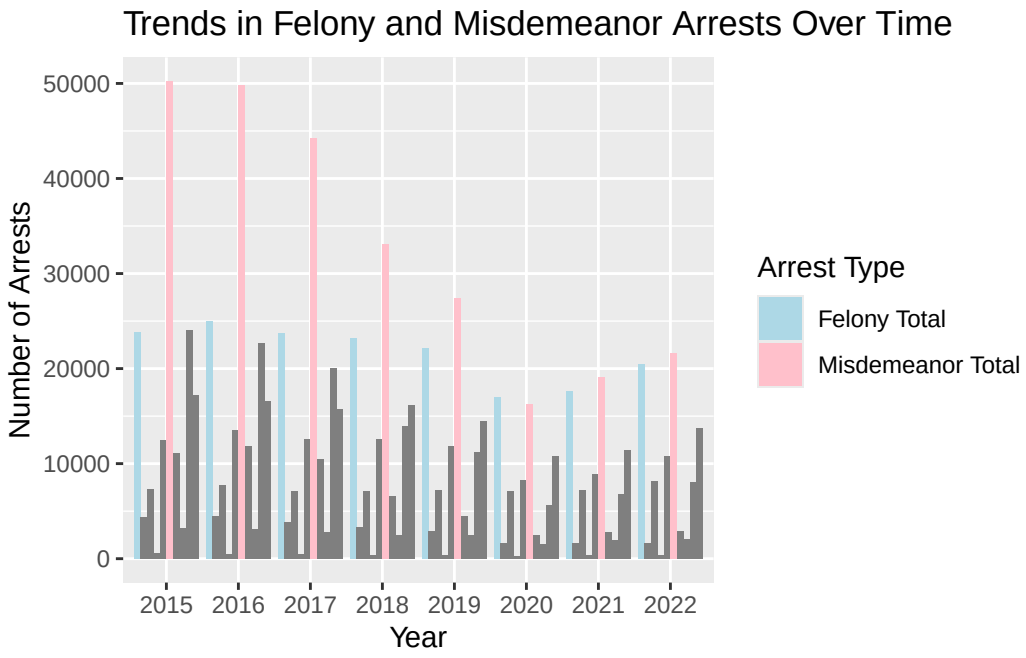

```r
#Faceted Line Plot by Arrest Type Over Time
ggplot(arrests_stacked, aes(x = Year, y = Count, color = County)) +
  geom_line() +
  facet_wrap(~ Arrest_Type, scales = "free_y", ncol = 2) + # Adjust ncol as needed
  labs(title = "Arrest Trends Over Time by Type and County", x = "Year", y = "Arrest Count")
  theme(
    strip.text = element_text(size = 6, margin = margin(t = 5, b = 5)), # Decrease label size
    axis.text.x = element_text(angle = 45, hjust = 1) # Rotate x-axis labels if needed
  )
```




```
#Stacked Bar Plot by County and Arrest Type
ggplot(arrests_stacked, aes(x = County, y = Count, fill = Arrest_Type)) +
  geom_bar(stat = "identity") +
  labs(title = "Arrest Counts by Type and County", x = "County", y = "Arrest Counts") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
# Bar plot for felony and misdemeanor arrests over time
ggplot(arrests_stacked, aes(x = as.factor(Year), y = Count, fill = Arrest_Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("Felony Total" = "lightblue", "Misdemeanor Total" = "pink")) +
  labs(title = "Trends in Felony and Misdemeanor Arrests Over Time",
       x = "Year",
       y = "Number of Arrests",
       fill = "Arrest Type")
```

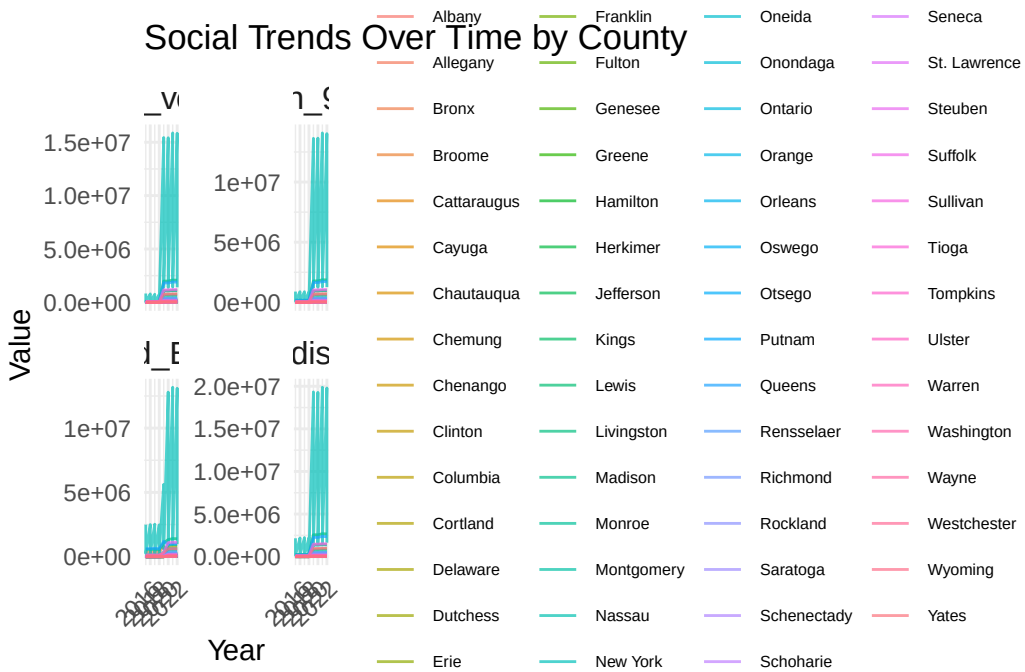



```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
-- 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 becor
```

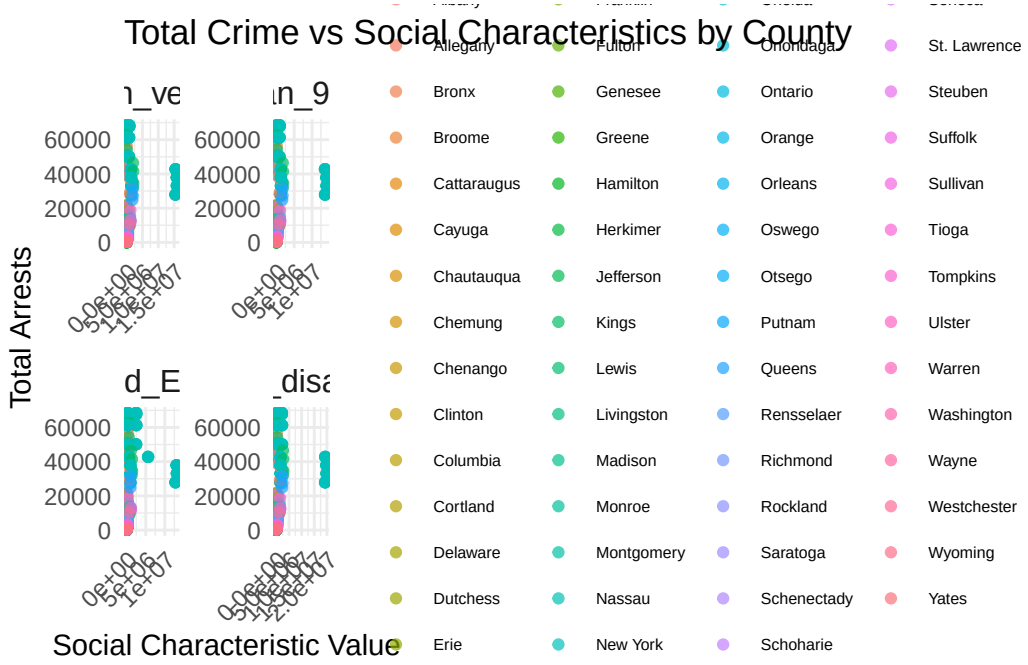
```
#Explore social data individually:

#Time Series Line Plots for Social Trends by County
# Reshape data to long format for the four social variables
social_data_long <- merged_data %>%
  select(Year, County, Less_than_9th_grade, With_disability, Civilian_veterans, Limited_Engli
  pivot_longer(cols = c(Less_than_9th_grade, With_disability, Civilian_veterans, Limited_Engl
               names_to = "Social_Variable", values_to = "Value")
ggplot(social_data_long, aes(x = Year, y = Value, group = County)) +
  geom_line(aes(color = County), alpha = 0.7) + # Slight transparency for visual clarity
  facet_wrap(~ Social_Variable, scales = "free_y", nrow = 2, ncol = 2) +
  labs(title = "Social Trends Over Time by County", x = "Year", y = "Value") +
```

```
  theme_minimal() +
  theme(
    legend.position = "right",                # Position the legend on the right
    legend.title = element_text(size = 8),    # Reduce legend title text size
    legend.text = element_text(size = 6),     # Reduce legend item text size
    strip.text = element_text(size = 12),     # Increase facet label text size
    axis.text.x = element_text(angle = 45, hjust = 1) # Rotate x-axis labels for readability
  )
```

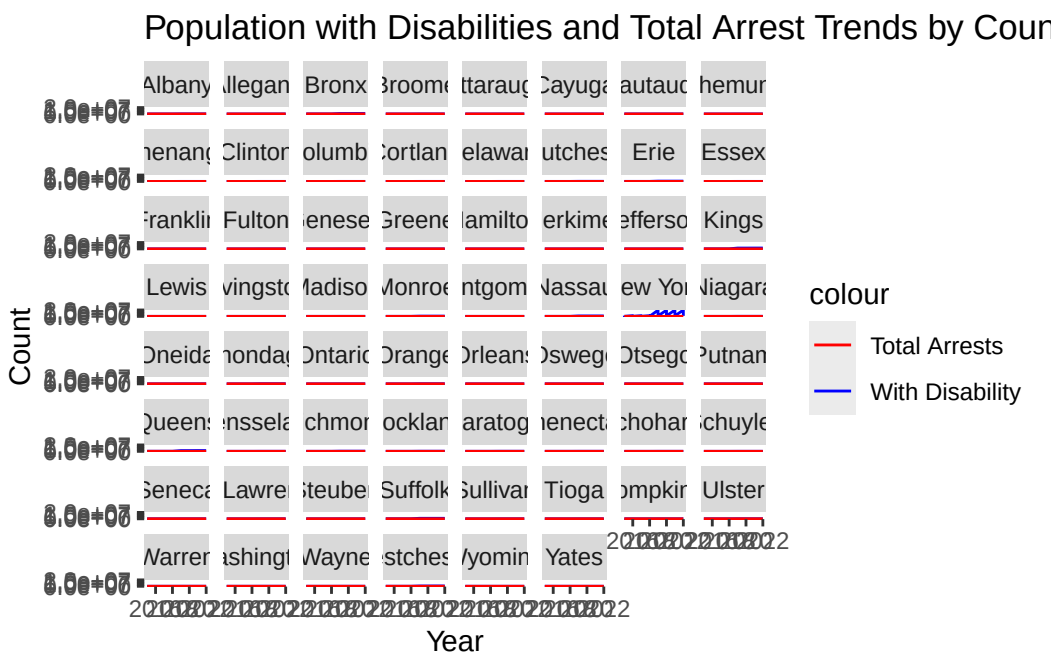



```
#Explore Relationships Between Housing and Crime Data:

# Convert data to long format for plotting
social_crime_data <- merged_data |>
  select(County, Year, Total, Less_than_9th_grade, With_disability, Civilian_veterans, Limite
  pivot_longer(cols = c(Less_than_9th_grade, With_disability, Civilian_veterans, Limited_Engl
               names_to = "Social_Variable", values_to = "Social_Value")
# Scatter plot of each social variable vs total crime with faceting and a reduced legend
ggplot(social_crime_data, aes(x = Social_Value, y = Total, color = County)) +
  geom_point(alpha = 0.7) +  # Add transparency to reduce overlap
  facet_wrap(~ Social_Variable, scales = "free", nrow = 2, ncol = 2) +
  labs(title = "Total Crime vs Social Characteristics by County", x = "Social Characteristic
  theme_minimal() +
```

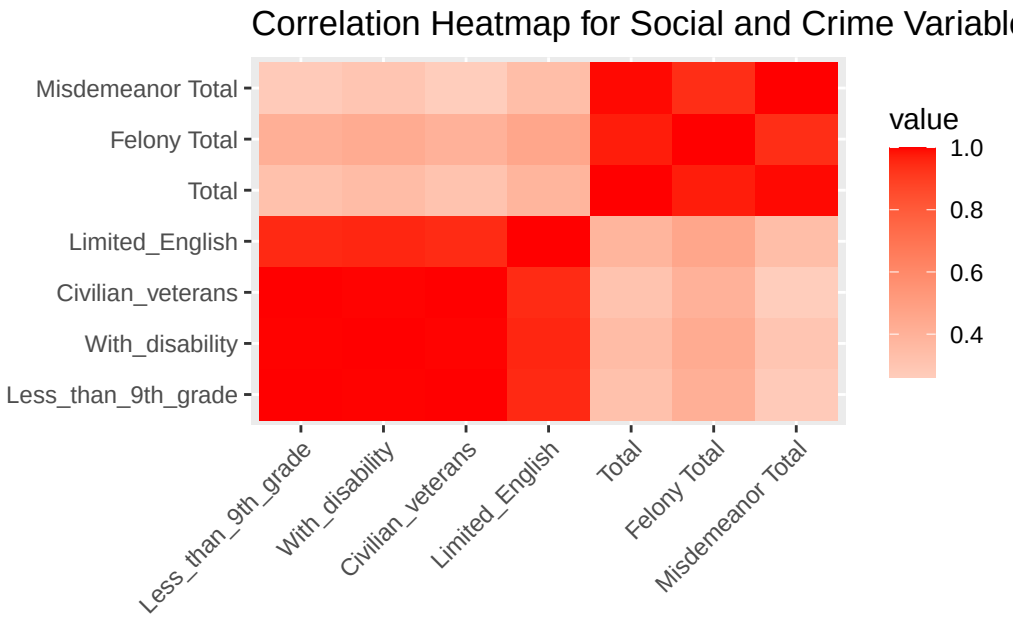
```
  theme(
    legend.position = "right",                  # Position legend on the right side
    legend.title = element_text(size = 8),      # Reduce legend title size
    legend.text = element_text(size = 6),       # Reduce legend text size
    strip.text = element_text(size = 12),       # Increase facet label text size
    axis.text.x = element_text(angle = 45, hjust = 1)  # Rotate x-axis labels for readability
  )
```



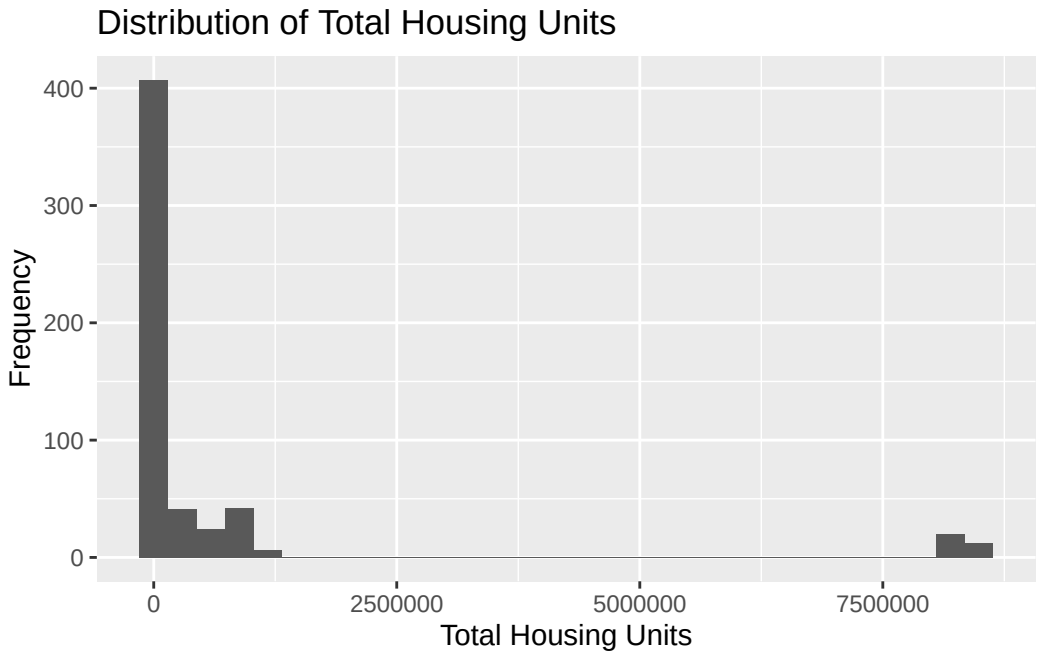

```
#Facet Grid for Social and Crime Trends Together by County
ggplot(merged_data, aes(x = Year)) +
  geom_line(aes(y = With_disability, color = "With Disability")) +
  geom_line(aes(y = Total, color = "Total Arrests")) +
  facet_wrap(~ County) +
  labs(title = "Population with Disabilities and Total Arrest Trends by County Over Time", x
  scale_color_manual(values = c("With Disability" = "blue", "Total Arrests" = "red"))
```



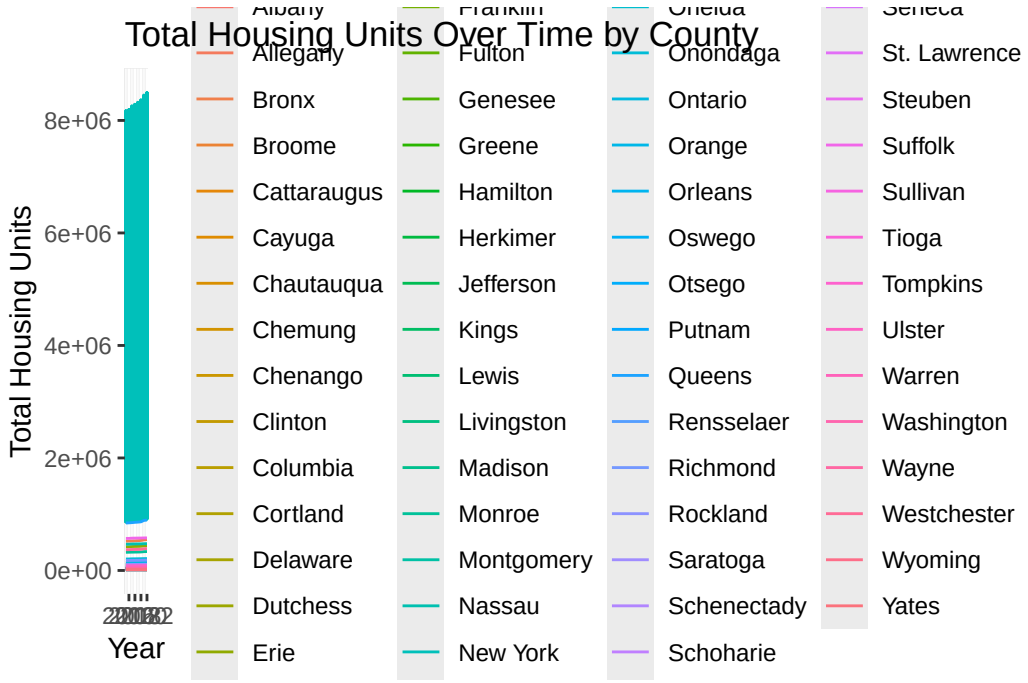

```
#Correlation Heatmap for Social and Crime Variables
correlation_matrix_social <- merged_data |>
  select(Less_than_9th_grade, With_disability, Civilian_veterans, Limited_English, Total, `Fe
  cor()
ggplot(melt(correlation_matrix_social), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  labs(title = "Correlation Heatmap for Social and Crime Variables", x = "", y = "") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        axis.text.y = element_text(angle = 0))
```



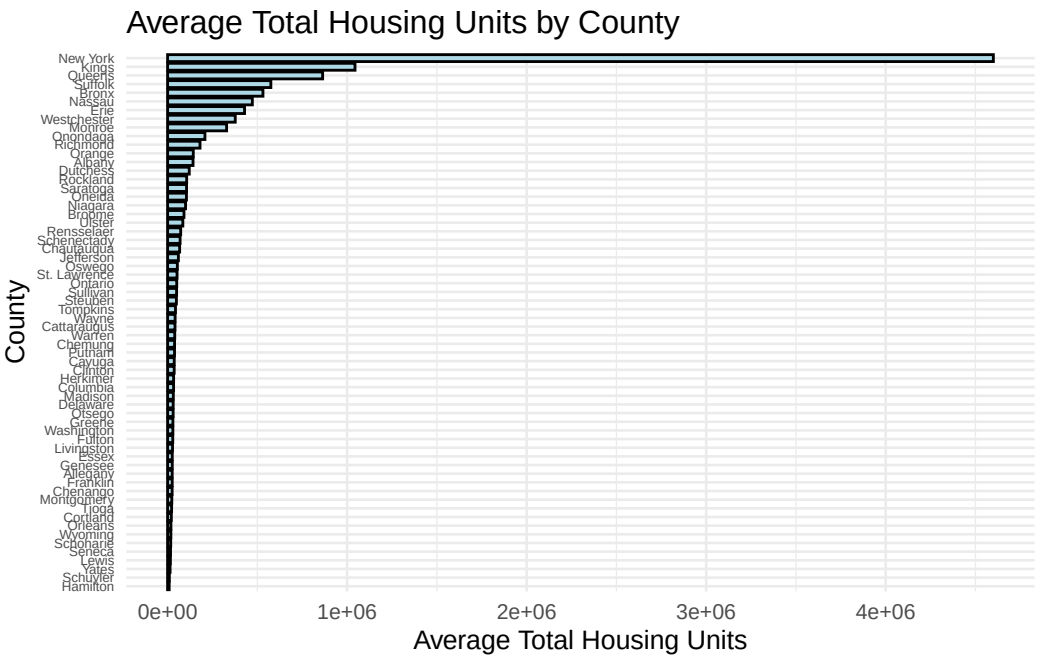

```r
#Explore housing data individually:

#Histograms and Density Plots for Housing Variables
ggplot(merged_data, aes(x = Total_housing_units)) +
  geom_histogram(bins = 30) +
  labs(title = "Distribution of Total Housing Units", x = "Total Housing Units", y = "Freque
```



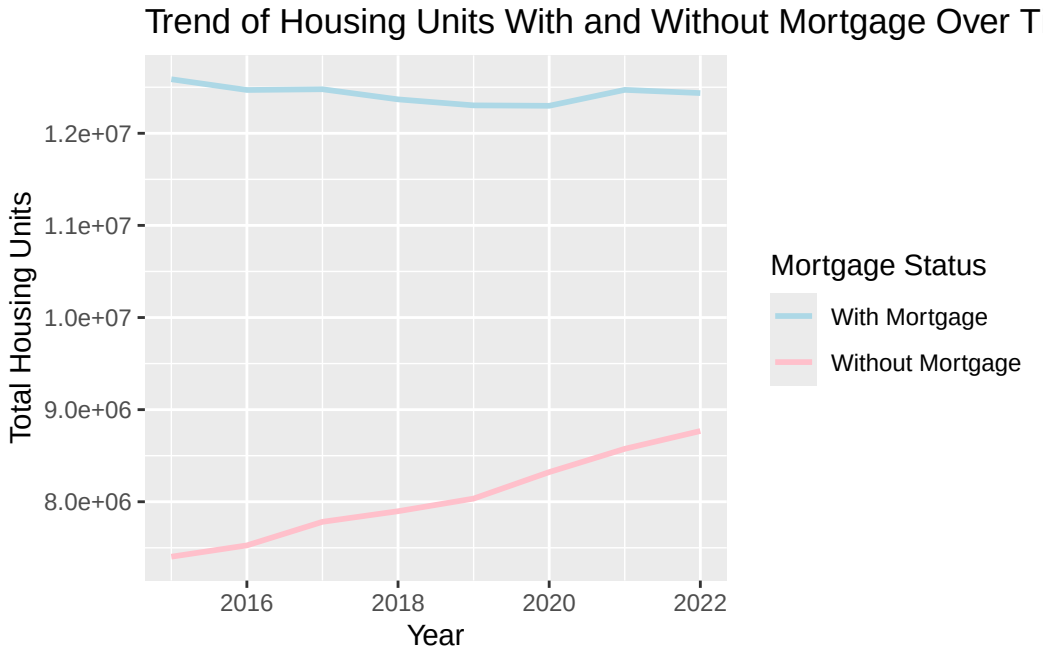
```
#Time Series Line Plots for Housing Trends by County
ggplot(merged_data, aes(x = Year, y = Total_housing_units, color = County)) +
  geom_line() +
  labs(title = "Total Housing Units Over Time by County", x = "Year", y = "Total Housing Unit
```




```
# Average total housing units by county
housing_summary <- merged_data |>
  group_by(County) |>
  summarize(Average_Total_Housing_Units = mean(Total_housing_units, na.rm = TRUE))

# Bar plot for average total housing units by county
ggplot(housing_summary, aes(x = reorder(County, Average_Total_Housing_Units), y = Average_Tot
  geom_bar(stat = "identity", fill = "lightblue", color = "black", width = 0.8) +
  coord_flip() +
  labs(title = "Average Total Housing Units by County", x = "County",
       y = "Average Total Housing Units") +
  theme_minimal(base_size = 10) +
  theme(
    axis.text.y = element_text(size = 5)
  )
```

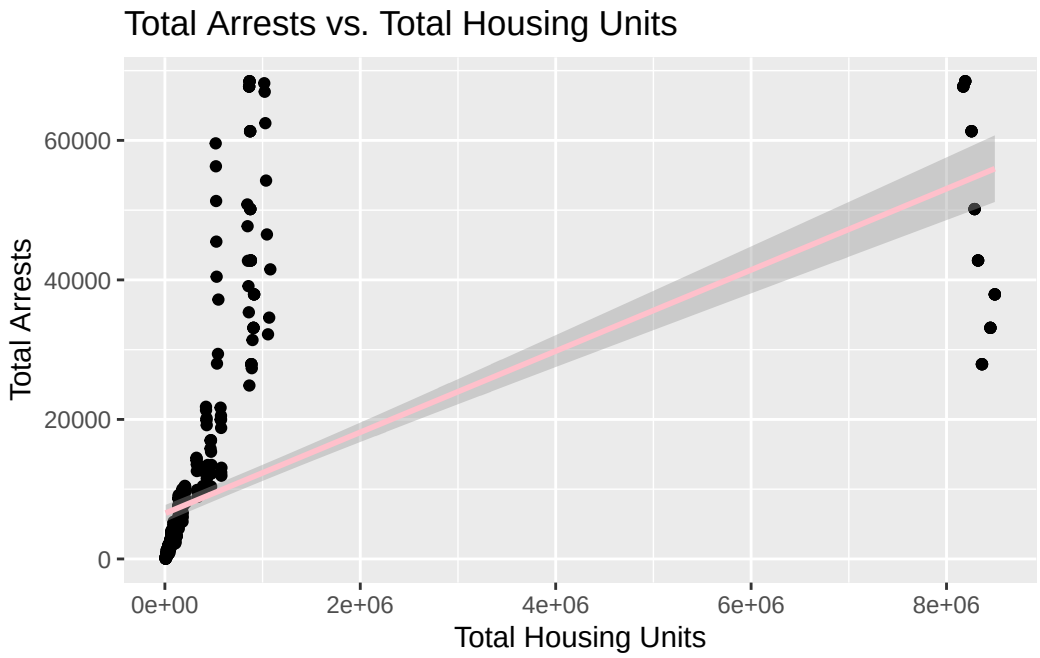



```
# Aggregate data by year for both with and without mortgage
with_mortgage_summary <- merged_data |>
  group_by(Year) |>
  summarize(Total_With_Mortgage = sum(With_mortgage, na.rm = TRUE))
without_mortgage_summary <- merged_data |>
  group_by(Year) |>
  summarize(Total_Without_Mortgage = sum(Without_mortgage, na.rm = TRUE))

# Line plot for housing units with and without mortgage over time
ggplot() +
  geom_line(data = with_mortgage_summary, aes(x = Year, y = Total_With_Mortgage, color = "Wi
  geom_line(data = without_mortgage_summary, aes(x = Year, y = Total_Without_Mortgage, color
  labs(title = "Trend of Housing Units With and Without Mortgage Over Time",
       x = "Year",
       y = "Total Housing Units") +
  scale_color_manual(values = c("With Mortgage" = "lightblue", "Without Mortgage" = "pink"),
                     name = "Mortgage Status")
```

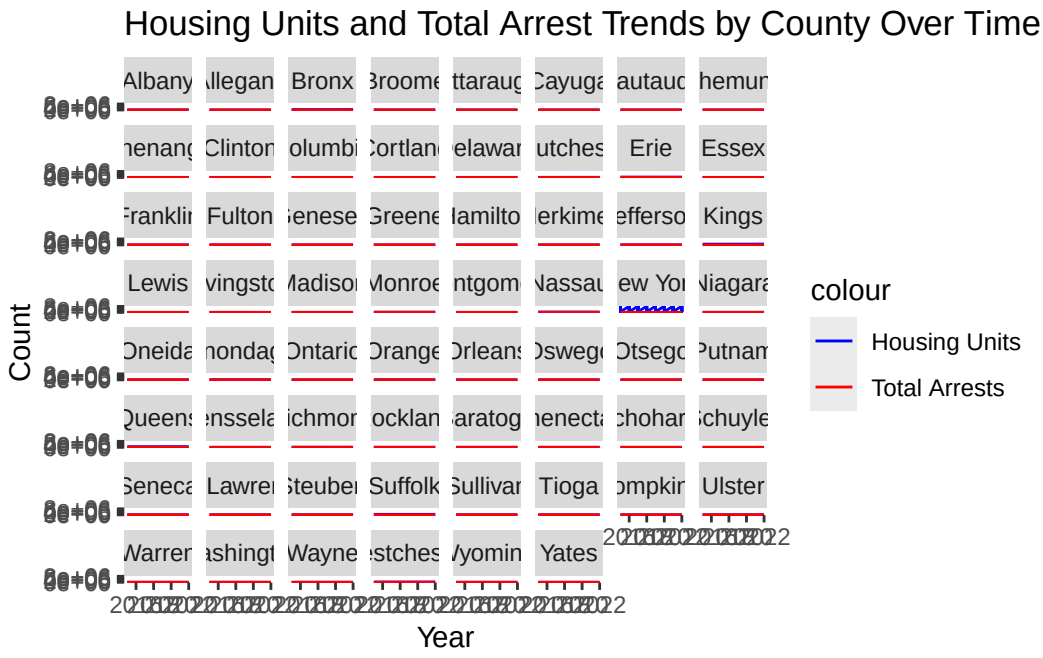



```
#Explore Relationships Between Housing and Crime Data:

# Scatter plot for total housing units vs. total arrests
ggplot(merged_data, aes(x = Total_housing_units, y = Total)) +
  geom_point() +
  geom_smooth(method = "lm", color = "pink") +
  labs(title = "Total Arrests vs. Total Housing Units", x = "Total Housing Units", y = "Total
```

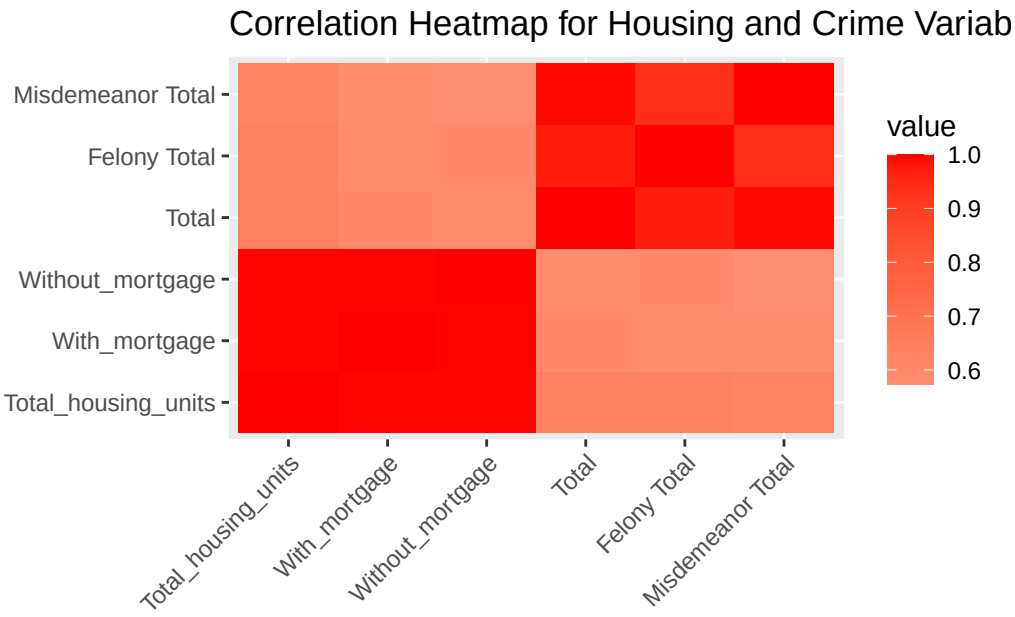
```
`geom_smooth()` using formula = 'y ~ x'
```



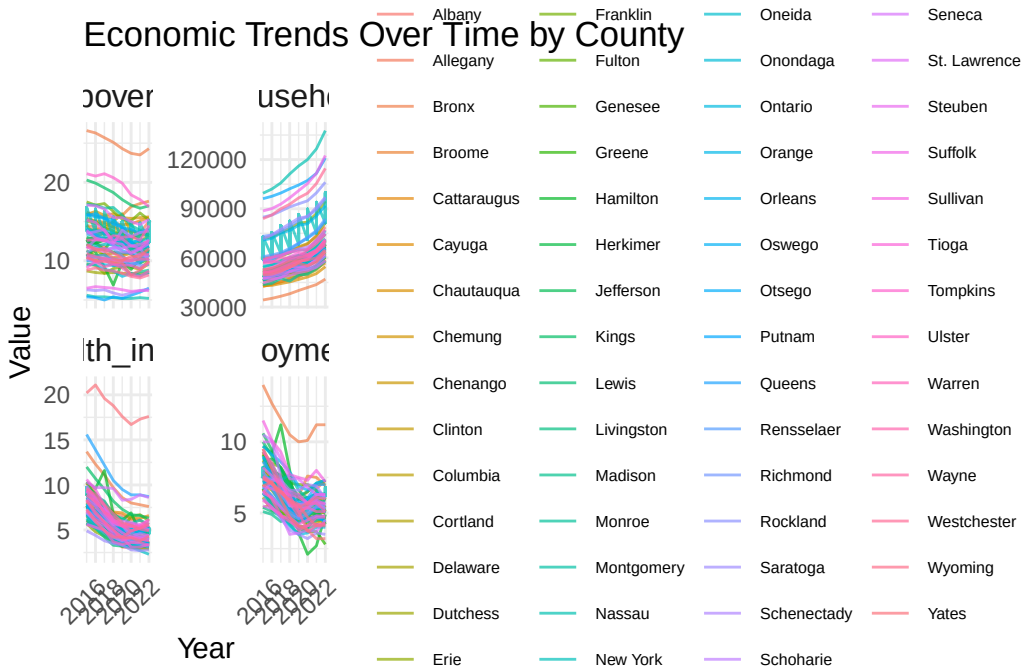

```
#Facet Grid for Housing and Crime Trends Together by County
ggplot(merged_data, aes(x = Year)) +
  geom_line(aes(y = Total_housing_units, color = "Housing Units")) +
  geom_line(aes(y = Total, color = "Total Arrests")) +
  facet_wrap(~ County) +
  labs(title = "Housing Units and Total Arrest Trends by County Over Time", x = "Year", y = "
  scale_color_manual(values = c("Housing Units" = "blue", "Total Arrests" = "red"))
```



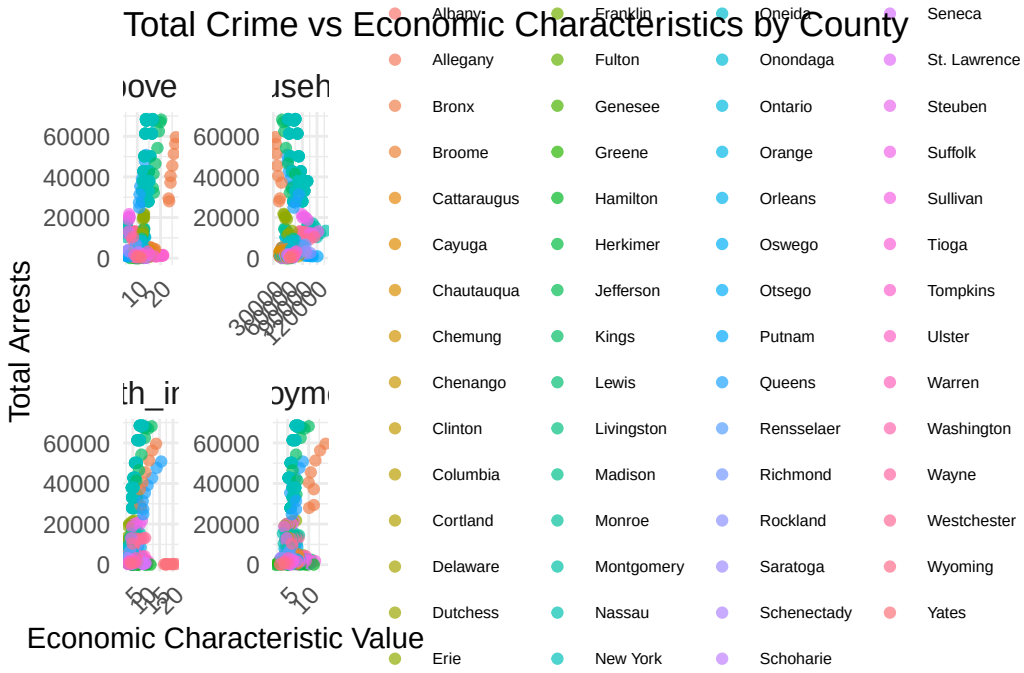

```
# Calculate correlation matrix and reshape for plotting
correlation_matrix <- merged_data |>
  select(Total_housing_units, With_mortgage, Without_mortgage, Total, `Felony Total`, `Misde
  cor()

# Heatmaps for Correlations Between Housing and Crime Variables
ggplot(melt(correlation_matrix), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  labs(title = "Correlation Heatmap for Housing and Crime Variables", x = "", y = "") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        axis.text.y = element_text(angle = 0))
```




```
# Explore economic data individually:

# Time Series Line Plots for Economic Trends by County
# Reshape data to long format for the four economic variables
economic_data_long <- merged_data |>
  select(Year, County, Unemployment_rate, Median_household_income, Below_poverty_level, No_he
  pivot_longer(cols = c(Unemployment_rate, Median_household_income, Below_poverty_level, No_l
               names_to = "Economic_Variable", values_to = "Value")

ggplot(economic_data_long, aes(x = Year, y = Value, group = County)) +
  geom_line(aes(color = County), alpha = 0.7) + # Slight transparency for visual clarity
  facet_wrap(~ Economic_Variable, scales = "free_y", nrow = 2, ncol = 2) +
  labs(title = "Economic Trends Over Time by County", x = "Year", y = "Value") +
  theme_minimal() +
  theme(
    legend.position = "right",                 # Position the legend on the right
    legend.title = element_text(size = 8),     # Reduce legend title text size
    legend.text = element_text(size = 6),      # Reduce legend item text size
    strip.text = element_text(size = 12),      # Increase facet label text size
    axis.text.x = element_text(angle = 45, hjust = 1) # Rotate x-axis labels for readability
  )
```

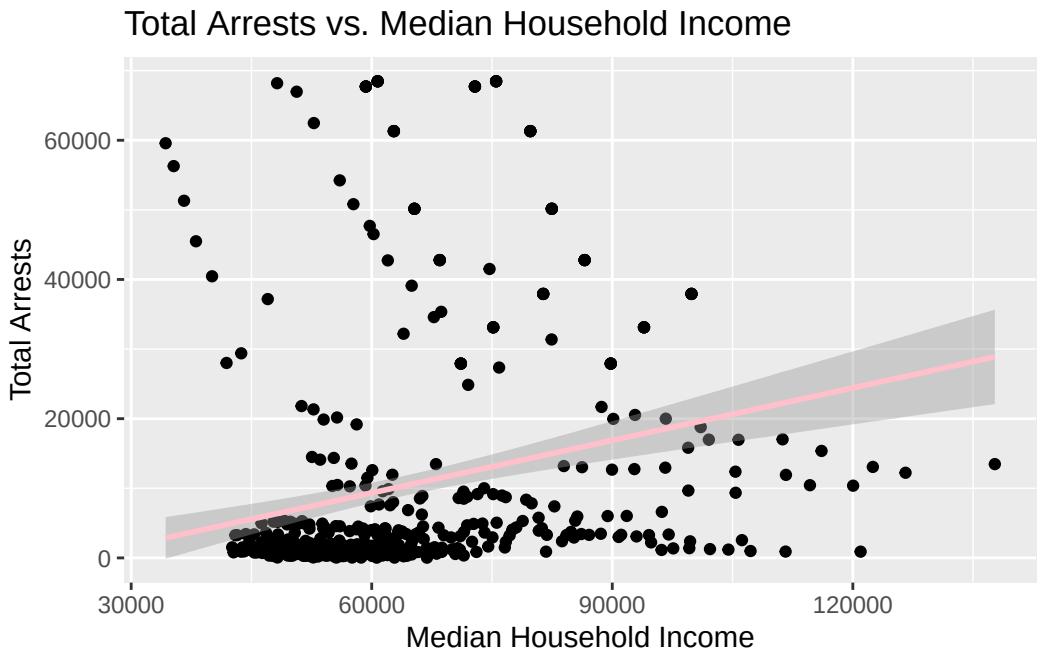



```r
# Explore Relationships Between Economic and Crime Data:

# Convert data to long format for plotting
economic_crime_data <- merged_data |>
  select(County, Year, Total, Unemployment_rate, Median_household_income, Below_poverty_level
  pivot_longer(cols = c(Unemployment_rate, Median_household_income, Below_poverty_level, No_
               names_to = "Economic_Variable", values_to = "Economic_Value")

# Scatter plot of each economic variable vs total crime with faceting and a reduced legend
ggplot(economic_crime_data, aes(x = Economic_Value, y = Total, color = County)) +
  geom_point(alpha = 0.7) +  # Add transparency to reduce overlap
  facet_wrap(~ Economic_Variable, scales = "free", nrow = 2, ncol = 2) +
  labs(title = "Total Crime vs Economic Characteristics by County", x = "Economic Characteris
  theme_minimal() +
  theme(
    legend.position = "right",                # Position legend on the right side
    legend.title = element_text(size = 8),    # Reduce legend title size
    legend.text = element_text(size = 6),     # Reduce legend text size
    strip.text = element_text(size = 12),     # Increase facet label text size
    axis.text.x = element_text(angle = 45, hjust = 1)  # Rotate x-axis labels for readability
  )
```



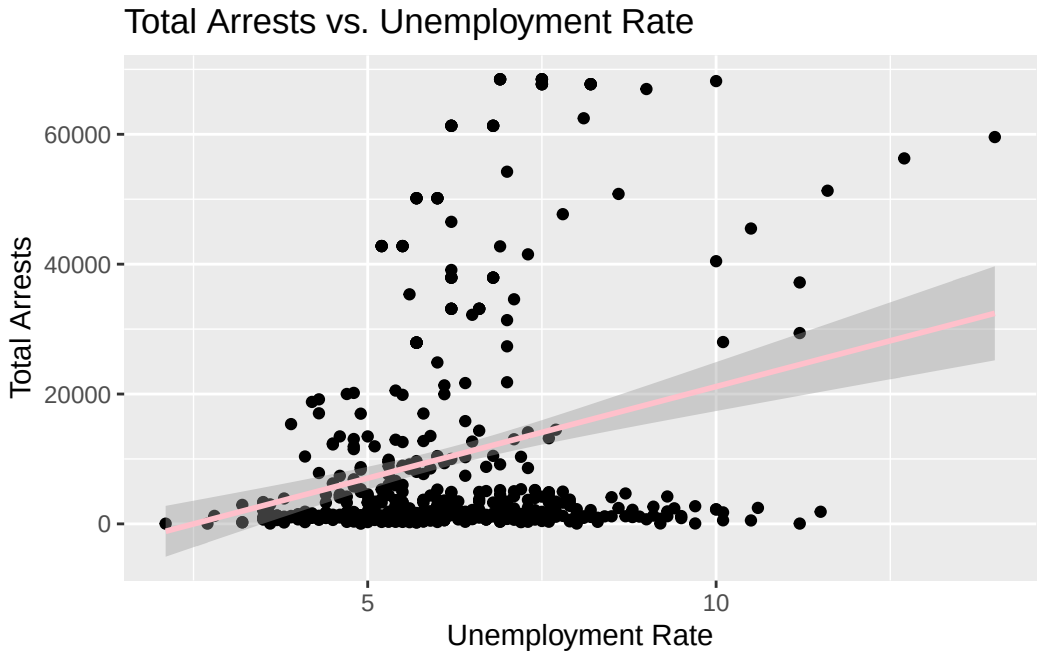

```
# Check linear relationship between variables
# Scatter plot for median household income vs. total arrests
ggplot(merged_data, aes(x = Median_household_income, y = Total)) +
  geom_point() +
  geom_smooth(method = "lm", color = "pink") +
  labs(title = "Total Arrests vs. Median Household Income",
       x = "Median Household Income", y = "Total Arrests")
```

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

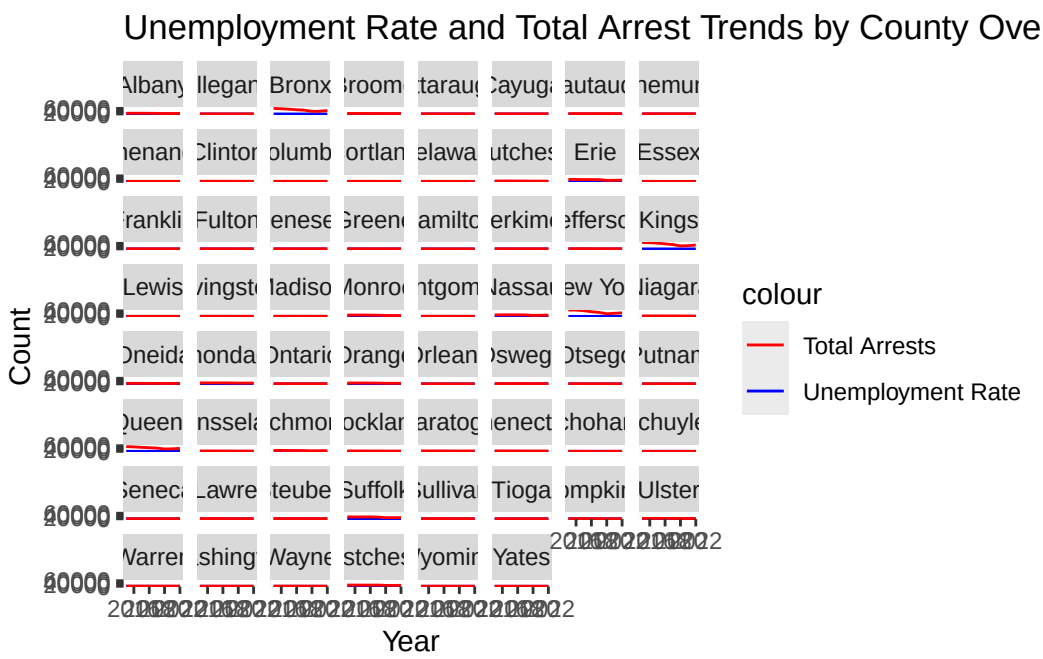



```
# Scatter plot for unemployment rate vs. total arrests
ggplot(merged_data, aes(x = Unemployment_rate, y = Total)) +
  geom_point() +
  geom_smooth(method = "lm", color = "pink") +
  labs(title = "Total Arrests vs. Unemployment Rate", x = "Unemployment Rate", y = "Total Ar
```

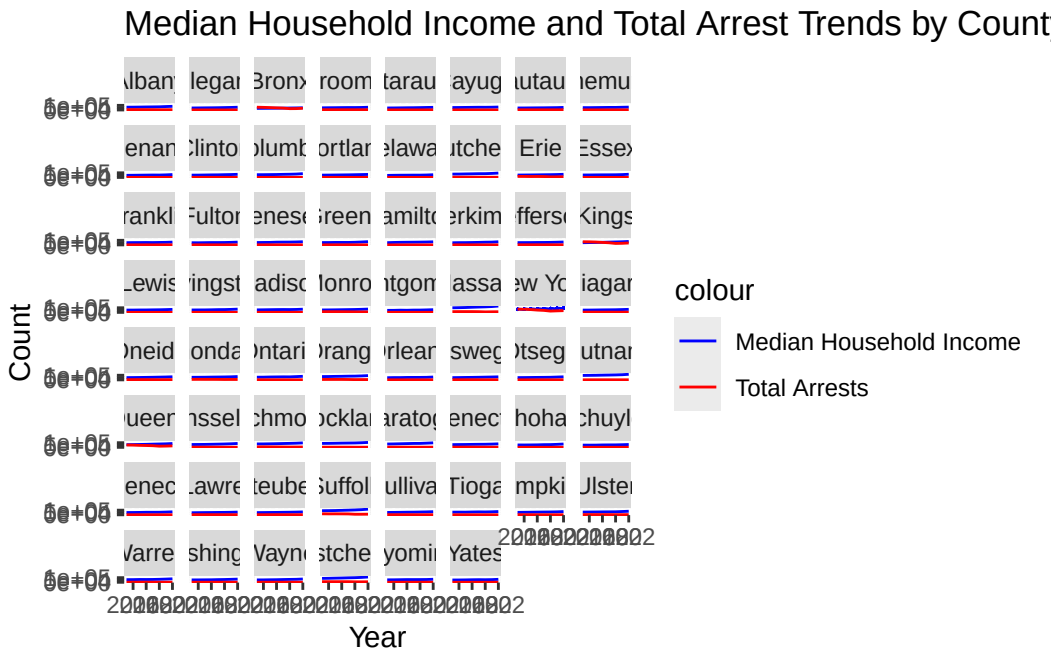
```
`geom_smooth()` using formula = 'y ~ x'
```



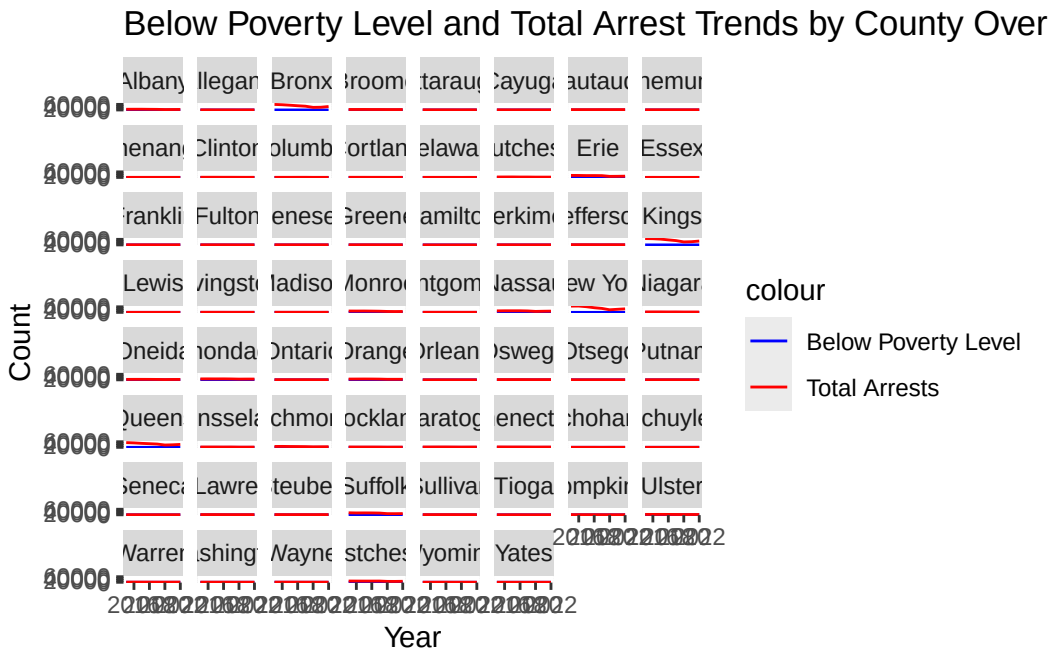

```
# Facet Grid for Economic and Crime Trends Together by County
ggplot(merged_data, aes(x = Year)) +
  geom_line(aes(y = Unemployment_rate, color = "Unemployment Rate")) +
  geom_line(aes(y = Total, color = "Total Arrests")) +
  facet_wrap(~ County) +
  labs(title = "Unemployment Rate and Total Arrest Trends by County Over Time", x = "Year", y
  scale_color_manual(values = c("Unemployment Rate" = "blue", "Total Arrests" = "red"))
```



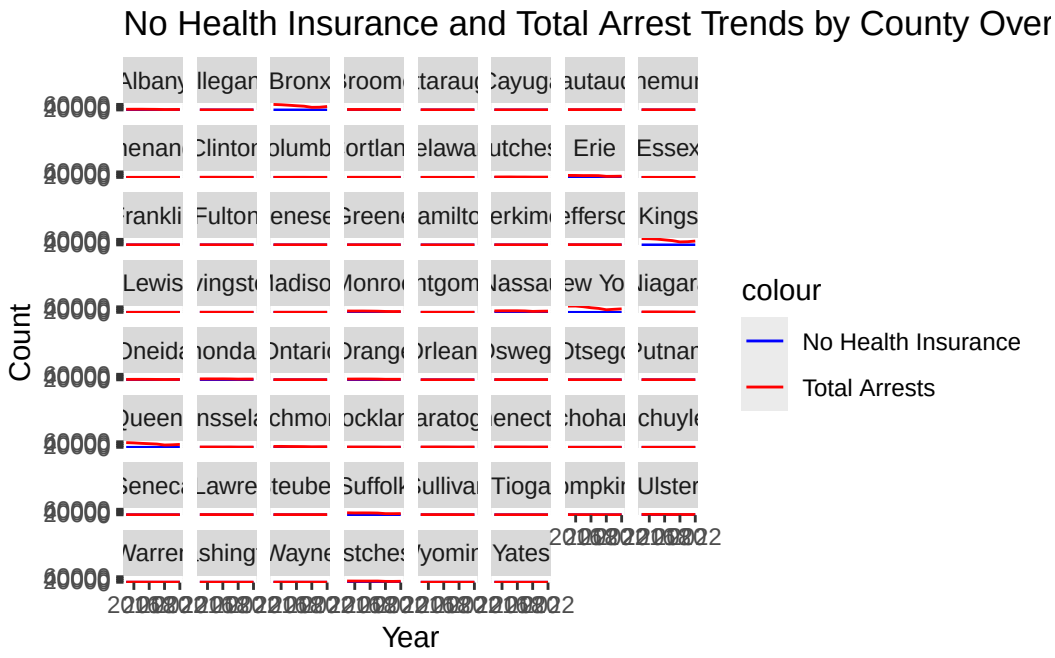

```
ggplot(merged_data, aes(x = Year)) +
  geom_line(aes(y = Median_household_income, color = "Median Household Income")) +
  geom_line(aes(y = Total, color = "Total Arrests")) +
  facet_wrap(~ County) +
  labs(title = "Median Household Income and Total Arrest Trends by County Over Time", x = "Ye
  scale_color_manual(values = c("Median Household Income" = "blue", "Total Arrests" = "red"))
```



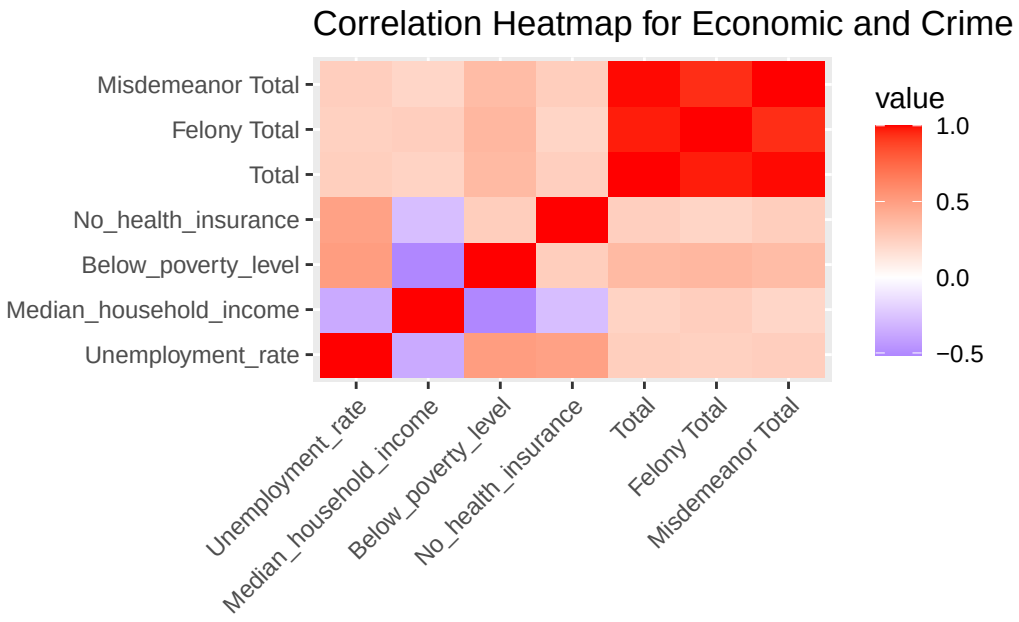

```
ggplot(merged_data, aes(x = Year)) +
  geom_line(aes(y = Below_poverty_level, color = "Below Poverty Level")) +
  geom_line(aes(y = Total, color = "Total Arrests")) +
  facet_wrap(~ County) +
  labs(title = "Below Poverty Level and Total Arrest Trends by County Over Time", x = "Year"
  scale_color_manual(values = c("Below Poverty Level" = "blue", "Total Arrests" = "red"))
```




```
ggplot(merged_data, aes(x = Year)) +
  geom_line(aes(y = No_health_insurance, color = "No Health Insurance")) +
  geom_line(aes(y = Total, color = "Total Arrests")) +
  facet_wrap(~ County) +
  labs(title = "No Health Insurance and Total Arrest Trends by County Over Time", x = "Year"
  scale_color_manual(values = c("No Health Insurance" = "blue", "Total Arrests" = "red"))
```




```
# Correlation Heatmap for Economic and Crime Variables
correlation_matrix_economic <- merged_data |>
  select(Unemployment_rate, Median_household_income, Below_poverty_level, No_health_insurance
  cor()

ggplot(melt(correlation_matrix_economic), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  labs(title = "Correlation Heatmap for Economic and Crime Variables", x = "", y = "") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        axis.text.y = element_text(angle = 0))
```



EDA Findings summary

The plots indicates that economic and social factors—such as unemployment, poverty, income, and specific social needs (like disabilities or limited English proficiency)—are closely linked to crime rates. Counties with higher poverty and unemployment rates tend to have higher arrest counts, suggesting that incorporating these variables in our model could improve its ability to predict crime trends.

While housing data, such as the number of housing units and mortgage trends, shows less direct relevance, it may still hold predictive value, especially in urban counties where housing characteristics correlate more with crime rates. Adding features to account for population density differences across counties could further enhance our model's accuracy.

Given these patterns, using models that can capture complex relationships is advisable. Poisson regression is a strong starting point for our project since we're working with count data (total arrests). we could also possibly explore alternatives like random forests for capturing non-linear patterns.

Questions for reviewers

1. Are there any additional data cleaning steps or transformations you would recommend?
2. Does the structure of the merged_data.csv dataset match with our analysis goals?

3. Our social characteristics data are based on total counts rather than rates, which doesn't account for population differences between counties. Could this affect our model's accuracy, and would you recommend normalizing these counts by population to better reflect relative prevalence?
4. Our initial exploratory data analysis shows that housing data has a weaker correlation with crime rates compared to economic and social factors. Would you suggest omitting these housing variables, substituting other factors, or applying feature engineering to enhance their relevance?
5. Should we consider including margins of error for the social characteristics estimates? Would they add valuable context to our analysis?