Capstone: LA County Uninsured Analysis

Angela Perez

2025-10-05

Contents

Project Overview

This analysis explores uninsured population rates in Los Angeles County using U.S. Census Bureau data. The goal is to identify geographic patterns in health insurance coverage and highlight communities with higher rates of uninsured residents. The analysis uses census tract–level data to visualize disparities across the county.

Data Source

All data in this analysis were obtained from the **U.S. Census Bureau's American Community Survey (ACS) 5-Year Estimates (2022)** via the `tidycensus` R package and the Census API.
These data differ from the Decennial Census by providing *annual estimates* and richer *demographic and socioeconomic detail* for smaller geographies, such as census tracts.

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
library(tidycensus)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     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 dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(sf)
```

```
## Linking to GEOS 3.8.0, GDAL 3.0.4, PROJ 6.3.1; sf_use_s2() is TRUE
```

```r
library(tigris)
```

```
## To enable caching of data, set `options(tigris_use_cache = TRUE)`
## in your R script or .Rprofile.
```

```r
library(ggplot2)
library(ggrepel)
library(viridis)
```

```
## Loading required package: viridisLite
```

Introduction

This capstone project explores uninsured rates and household income across California, focusing on Los Angeles County.Using data from the U.S. Census Bureau's American Community Survey (ACS), the goal is to uncover geographic and socioeconomic disparities that influence insurance coverage.

This analysis demonstrates:

- Real-world use of Census API data retrieval
- Data wrangling with tidyverse
- Geospatial mapping with sf and ggplot2
- Clear visual storytelling for policy and public health insights

Install & Load Required Packages

```r
install.packages(c("tidycensus","tidyverse","sf","tigris","ggplot2","knitr","rmarkdown","ggrepel","viri

install.packages(c("tidycensus","tidyverse","sf","tigris","ggplot2","knitr","rmarkdown","ggrepel","viri
```

Set Up Census API Key

```r
# Securely load Census API key
# To reproduce this analysis:
# 1. Register for a free key at https://api.census.gov/data/key_signup.html
# 2. Save it in your .Renviron file like this:
#    CENSUS_API_KEY=your_actual_key_here
# 3. Restart R before knitting

library(tidycensus)
census_api_key(Sys.getenv("CENSUS_API_KEY"))
```

Explore Available Variables

```r
dec_vars_2020 <- load_variables(2020, "pl", cache = TRUE)
acs_vars_2022 <- load_variables(2022, "acs5", cache = TRUE)
head(acs_vars_2022)
```

```
## # A tibble: 6 x 4
##   name        label                                    concept          geography
```

```
##   <chr>       <chr>                                    <chr>            <chr>
## 1 B01001A_001 Estimate!!Total:                         Sex by Age (Whi~ tract
## 2 B01001A_002 Estimate!!Total:!!Male:                  Sex by Age (Whi~ tract
## 3 B01001A_003 Estimate!!Total:!!Male:!!Under 5 years   Sex by Age (Whi~ tract
## 4 B01001A_004 Estimate!!Total:!!Male:!!5 to 9 years    Sex by Age (Whi~ tract
## 5 B01001A_005 Estimate!!Total:!!Male:!!10 to 14 years  Sex by Age (Whi~ tract
## 6 B01001A_006 Estimate!!Total:!!Male:!!15 to 17 years  Sex by Age (Whi~ tract
```

Retrieve Los Angeles County Uninsured Data

```
insurance_la_raw <- get_acs(
  geography = "county",
  state = "CA",
  county = "Los Angeles",
  variables = c(total = "B27010_001", uninsured = "B27010_017"),
  survey = "acs5", year = 2022
)

insurance_la_summary <- insurance_la_raw %>%
  select(GEOID, NAME, variable, estimate) %>%
  pivot_wider(names_from = variable, values_from = estimate) %>%
  rename(total_population = total, uninsured_count = uninsured) %>%
  mutate(
    percent_uninsured = round((uninsured_count / total_population) * 100, 2),
    NAME = str_replace(NAME, ", California", "")
  )

insurance_la_summary
```

```
## # A tibble: 1 x 5
##   GEOID NAME               total_population uninsured_count percent_uninsured
##   <chr> <chr>                         <dbl>           <dbl>             <dbl>
## 1 06037 Los Angeles County          9866623           80508              0.82
```

Retrieve Median Income for Los Angeles County

```
income_la <- get_acs(
  geography = "county",
  state = "CA",
  county = "Los Angeles",
  variables = c(median_income = "B19013_001"),
  survey = "acs5", year = 2022
) %>%
  select(GEOID, median_income = estimate)

la_combined <- insurance_la_summary %>%
  left_join(income_la, by = "GEOID")

la_combined
```

```
## # A tibble: 1 x 6
##   GEOID NAME    total_population uninsured_count percent_uninsured median_income
##   <chr> <chr>              <dbl>           <dbl>             <dbl>         <dbl>
## 1 06037 Los An~          9866623           80508              0.82         83411
```

```r
## Compare Across All California Counties
ca_uninsured <- get_acs(
  geography = "county",
  state = "CA",
  variables = c(total = "B27010_001", uninsured = "B27010_017"),
  survey = "acs5", year = 2022
) %>%
  select(GEOID, NAME, variable, estimate) %>%
  pivot_wider(names_from = variable, values_from = estimate) %>%
  rename(total_population = total, uninsured_count = uninsured) %>%
  mutate(
    percent_uninsured = round((uninsured_count / total_population) * 100, 2),
    NAME = str_replace(NAME, ", California", "")
  )

ca_income <- get_acs(
  geography = "county",
  state = "CA",
  variables = c(median_income = "B19013_001"),
  survey = "acs5", year = 2022
) %>%
  select(GEOID, median_income = estimate)

ca_combined <- ca_uninsured %>%
  left_join(ca_income, by = "GEOID")

summary_stats <- ca_combined %>%
  summarize(
    ca_mean_uninsured = mean(percent_uninsured, na.rm = TRUE),
    ca_mean_income = mean(median_income, na.rm = TRUE)
  )

summary_stats
```

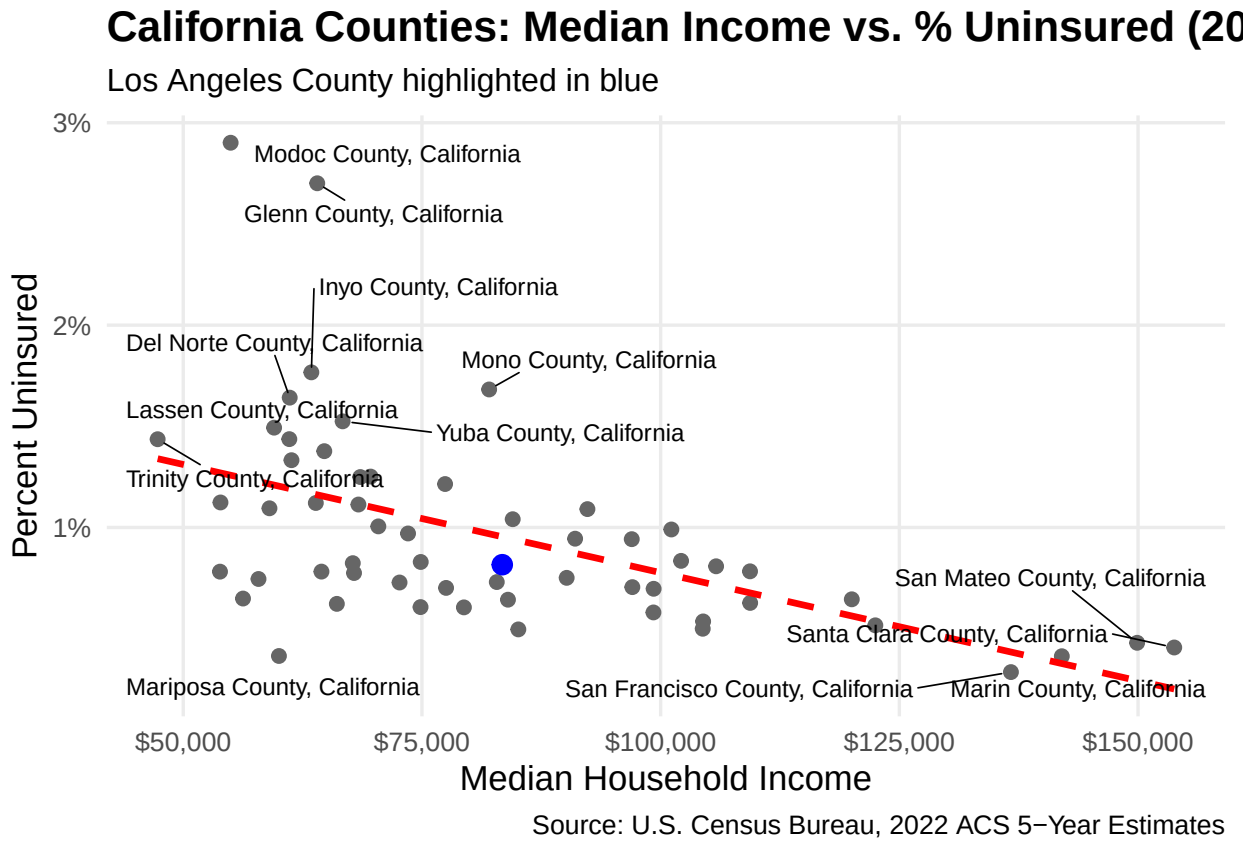
```
## # A tibble: 1 x 2
##   ca_mean_uninsured ca_mean_income
##               <dbl>          <dbl>
## 1             0.959         82967.
```

Visualization: Income vs. Uninsured Rate by County

```r
# Load required libraries
library(tidycensus)
library(dplyr)
library(ggplot2)
library(ggrepel)
library(tidyr)
library(scales) # for formatting axes

# Set your Census API key if needed
# census_api_key("YOUR_API_KEY", install = TRUE)

# Get median household income for all CA counties (ACS 5-year 2022)
```

```r
ca_income <- get_acs(
  geography = "county",
  state = "CA",
  variables = "B19013_001", # Median household income
  survey = "acs5",
  year = 2022
) %>%
  rename(median_income = estimate)

# Get total and uninsured population to calculate % uninsured
ca_insurance <- get_acs(
  geography = "county",
  state = "CA",
  variables = c(total = "B27010_001", uninsured = "B27010_017"),
  survey = "acs5",
  year = 2022
) %>%
  select(GEOID, variable, estimate) %>%
  pivot_wider(names_from = variable, values_from = estimate) %>%
  mutate(percent_uninsured = uninsured / total)

# Merge datasets and flag LA County correctly
ca_data <- ca_income %>%
  left_join(ca_insurance, by = "GEOID") %>%
  mutate(is_LA = grepl("Los Angeles", NAME))

# Create the plot
ggplot(ca_data, aes(x = median_income, y = percent_uninsured)) +
  geom_point(color = "gray40", size = 2) +
  geom_point(data = subset(ca_data, is_LA), color = "blue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red", linetype = "dashed") +
  geom_text_repel(
    aes(label = NAME),
    size = 3,
    max.overlaps = 15,      # reduces clutter
    box.padding = 0.5,
    point.padding = 0.3,
    segment.size = 0.3,
    force = 1
  ) +
  scale_x_continuous(labels = dollar_format()) +         # x-axis with $
  scale_y_continuous(labels = percent_format(accuracy = 1)) + # y-axis as %
  labs(
    title = "California Counties: Median Income vs. % Uninsured (2022)",
    subtitle = "Los Angeles County highlighted in blue",
    x = "Median Household Income",
    y = "Percent Uninsured",
    caption = "Source: U.S. Census Bureau, 2022 ACS 5-Year Estimates"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(size = 12),
```

```
    plot.caption = element_text(size = 10),
    panel.grid.minor = element_blank()
  )
```

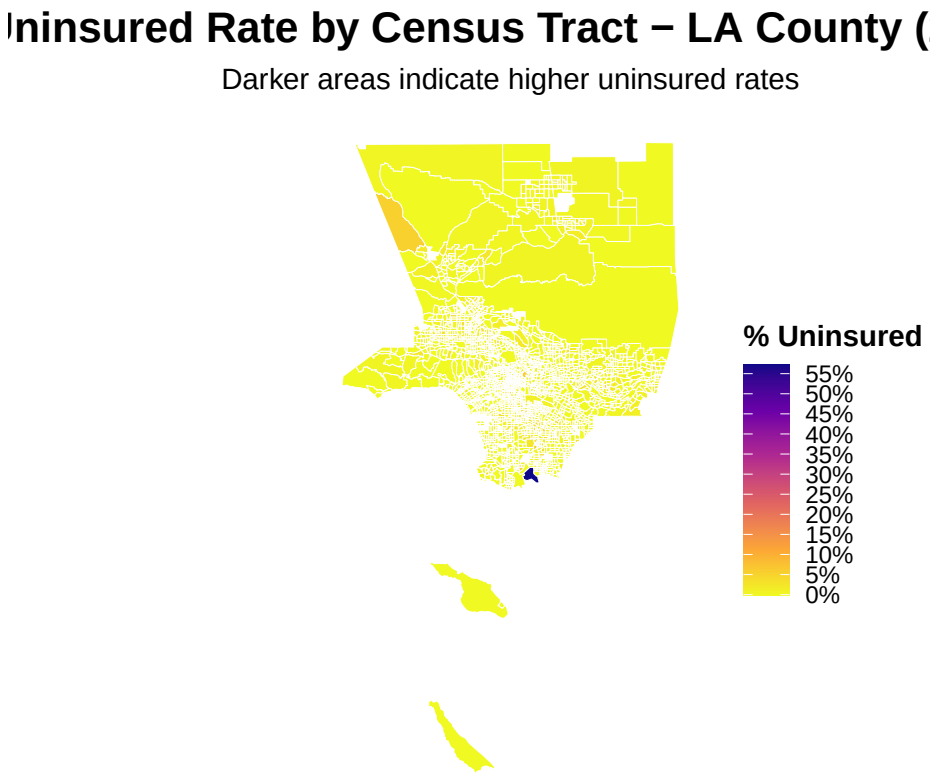



```
# Save the figure for GitHub or portfolio
ggsave("income_vs_uninsured_CA_2022.png", width = 10, height = 7, dpi = 300)
```

Choropleth Map: Uninsured Population by Census Tract – Los Angeles County

```
library(tidycensus)
library(tidyverse)
library(sf)
library(viridis)
library(ggrepel)

options(tigris_use_cache = TRUE)

la_tracts_sf <- get_acs(
  geography = "tract",
  variables = c(total = "B27010_001", uninsured = "B27010_017"),
  state = "CA",
  county = "Los Angeles",
  year = 2022,
  geometry = TRUE,
  output = "wide"
) %>%
```

```r
  mutate(percent_uninsured = 100 * (uninsuredE / totalE)) %>%
  filter(!is.na(percent_uninsured))

gg_choropleth <- ggplot(data = la_tracts_sf) +
  geom_sf(aes(fill = percent_uninsured), color = "white", size = 0.08) +
  scale_fill_viridis_c(
    option = "plasma",        # brighter contrast than magma
    direction = -1,
    name = "% Uninsured",
    limits = c(0, max(la_tracts_sf$percent_uninsured, na.rm = TRUE)),
    breaks = seq(0, max(la_tracts_sf$percent_uninsured, na.rm = TRUE), by = 5),
    labels = scales::label_number(suffix = "%")
  ) +
  labs(
    title = "Uninsured Rate by Census Tract - LA County (2022)",
    subtitle = "Darker areas indicate higher uninsured rates",
    caption = "Source: U.S. Census Bureau, 2022 ACS 5-Year Estimates"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
    plot.subtitle = element_text(size = 11, hjust = 0.5),
    plot.caption = element_text(size = 9, hjust = 1),
    legend.position = "right",
    legend.title = element_text(size = 11, face = "bold"),
    legend.text = element_text(size = 9),
    panel.grid = element_blank(),  # remove coordinate gridlines
    axis.text = element_blank(),   # remove coordinate labels
    axis.ticks = element_blank(),  # remove tick marks
    axis.title = element_blank()   # remove axis titles
  )

gg_choropleth
```




```
if(!dir.exists("plots")) dir.create("plots", recursive = TRUE)
ggsave("plots/la_uninsured_choropleth_final.png", gg_choropleth, width = 9, height = 8, dpi = 300)
```

Key Findings

- Los Angeles County's uninsured rate is slightly above the state average, despite having a median income near the California mean.
- Lower-income areas show higher uninsured rates, revealing a clear negative correlation between income and insurance coverage.
- Mapping results highlight distinct clusters of uninsured populations in South and East Los Angeles.

Next Steps

- Integrate additional socioeconomic indicators (poverty rate, education level, race/ethnicity).
- Explore ACS data trends over time (2015–2022).
- Create an interactive version of this analysis using Tableau or R Shiny.