Air Pollution and Social Vulnerability in California

ENVIRON 872 - Environmental Data Exploration

Sujay Dhanagare, Weilin Wang, Emily Guyu Yang

Contents

List of Tables

List of Figures

1 Research Questions

Research Question: Do communities facing higher pollution levels have less access to health insurance and medical care?

2 Objective

This study explores the relationships between PM2.5 levels (air pollution) and social vulnerability indicators, including poverty rates, minority percentages, and health insurance coverage, across California counties for the years 2000 and 2022.

3 Dataset Information

Source and Content of Data

The data used in this analysis was obtained from two main sources: the Environmental Protection Agency (EPA) for PM2.5 concentration data, and the Centers for Disease Control and Prevention (CDC) for the Social Vulnerability Index (SVI) data.

The PM2.5 concentration data was collected for the years 2000 and 2022. This dataset contains the daily mean PM2.5 levels measured in micrograms per cubic meter (mu * g/m3) for each county in California. The data was aggregated to calculate the yearly average PM2.5 concentration for each county.

The SVI data provides information on the social vulnerability of California counties across several socioeconomic indicators. The specific variables used in this analysis include:

- Percent of population below 150% of the poverty level
- Percent of minority (non-white) population
- Percent of population without health insurance coverage

Data Wrangling Process

To integrate the PM2.5 and SVI data, an inner join was performed on the county FIPS code to create a merged dataset for analysis. This allowed us to examine the relationships between air pollution levels and socioeconomic factors at the county level.

Dataset Structure Summary

The final dataset has the following structure:

Column Name	Data Type	Description
County FIPS Code	Character	Unique identifier for each county
County	Character	Name of the county
Percent_Below_150_Poverty	Numeric	Percentage of population below 150% of the poverty level
Percent_Uninsured	Numeric	Percentage of population without health insurance coverage
Minority_Percentage	Numeric	Percentage of minority (non-white) population
yearly_avg_PM25_2000	Numeric	Yearly average PM2.5 concentration in 2000

Column Name	Data Type	Description
yearly_avg_PM25_2022	Numeric	Yearly average PM2.5 concentration in 2022

By combining the air quality and socioeconomic data, this dataset enables the investigation of the relationships between environmental exposures and demographic factors across California counties over the 22-year period from 2000 to 2022.

4 Exploratory Analysis

```
Svi_data_2022 <- SVI_CA_County_2022 %>%
  select(
    COUNTY, FIPS, EP_POV150, E_MINRTY, E_TOTPOP, EP_UNINSUR
    ) %>% # Include uninsured percentage
  mutate(
    FIPS = sub("^06", "", FIPS),
    Minority_Percentage = (E_MINRTY / E_TOTPOP) * 100 # Calculate minority percentage
  ) %>%
  rename(
    County = COUNTY,
    `County FIPS Code` = FIPS,
    Percent_Below_150_Poverty = EP_POV150,
    Percent_Uninsured = EP_UNINSUR # Add a new column for uninsured percentage
  )

Svi_data_2000 <- SVI_CA_County_2000 %>%
  select(
    COUNTY, CNTY_FIPS, G3V1N, G1V1N, Totpop2000
    ) %>% # Include total population column
  mutate(
    Percent_Below_Poverty = (G1V1N / Totpop2000) * 100, # Calculate poverty percentage
    Minority_Percentage = (G3V1N / Totpop2000) * 100,  # Calculate minority percentage
    FIPS = sub("^06", "", CNTY_FIPS)
  ) %>%
  rename(
    County = COUNTY,
    `County FIPS Code` = CNTY_FIPS
  )
```

```
# Fixing date parsing for PM2.5 2022 dataset
PM2_5_2022 <- PM2_5_2022 %>%
  mutate(Date = parse_date(Date, format = "%m/%d/%Y")) # Corrected format

# Calculate the daily mean PM2.5 for each date within each group
# Calculate the yearly average daily mean PM2.5 concentration per county
PM2_5_2022_result <- PM2_5_2022 %>%
  mutate(year = year(Date)) %>%
  group_by(`County FIPS Code`, year) %>%
  summarise(`Daily Mean PM2.5 Concentration` = mean(`Daily Mean PM2.5 Concentration`, na.rm = TRUE)) %>%
  summarise(
```

```r
    yearly_avg_PM25 =
      mean(`Daily Mean PM2.5 Concentration`, na.rm = TRUE)
    )
print(PM2_5_2022_result)
```

```
## # A tibble: 50 x 2
##    `County FIPS Code` yearly_avg_PM25
##    <chr>                        <dbl>
##  1 001                           8.20
##  2 007                           6.19
##  3 009                           6.04
##  4 011                           7.61
##  5 013                           8.25
##  6 015                           4.97
##  7 017                           4.07
##  8 019                          10.2
##  9 021                           5.34
## 10 023                           6.76
## # i 40 more rows
```

```r
# Fixing date parsing for PM2.5 2000 dataset
PM2_5_2000 <- PM2_5_2000 %>%
  mutate(Date = parse_date(Date, format = "%m/%d/%Y")) # Corrected format

# Calculate the daily mean PM2.5 for each date within each group
# Calculate the yearly average daily mean PM2.5 concentration per county
PM2_5_2000_result <- PM2_5_2000 %>%
  mutate(year = year(Date)) %>%
  group_by(`County FIPS Code`, year) %>%
  summarise(
    `Daily Mean PM2.5 Concentration` =
      mean(`Daily Mean PM2.5 Concentration`, na.rm = TRUE)
    ) %>%
  summarise(yearly_avg_PM25 = mean(`Daily Mean PM2.5 Concentration`, na.rm = TRUE))
print(PM2_5_2000_result)
```

```
## # A tibble: 48 x 2
##    `County FIPS Code` yearly_avg_PM25
##    <chr>                        <dbl>
##  1 001                          12.0
##  2 007                          15.8
##  3 009                           8.98
##  4 011                           8.20
##  5 013                          12.9
##  6 015                           3.86
##  7 017                           4.86
##  8 019                          20.4
##  9 023                           9.27
## 10 025                          14.4
## # i 38 more rows
```

```r
# Merge SVI and PM2.5 datasets for 2000
Merged_2000 <- inner_join(Svi_data_2000, PM2_5_2000_result, by = "County FIPS Code")

# Merge SVI and PM2.5 datasets for 2022
Merged_2022 <- inner_join(Svi_data_2022, PM2_5_2022_result, by = "County FIPS Code")

# Standardize county names in 2022 dataset by removing "County"
Merged_2022 <- Merged_2022 %>%
  mutate(County = str_remove(County, " County$"))

head(Merged_2000)
```

```
## # A tibble: 6 x 9
##   County       `County FIPS Code`  G3V1N  G1V1N Totpop2000 Percent_Below_Poverty
##   <chr>        <chr>               <dbl>  <dbl>      <dbl>                 <dbl>
## 1 Alameda      001                854498 156804    1443741                 10.9
## 2 Butte        007                 41029  39148     203171                 19.3
## 3 Calaveras    009                  5026   4704      40554                 11.6
## 4 Colusa       011                  9865   2964      18804                 15.8
## 5 Contra Costa 013                400979  71575     948816                  7.54
## 6 Del Norte    015                  8235   4765      27507                 17.3
## # i 3 more variables: Minority_Percentage <dbl>, FIPS <chr>,
## #   yearly_avg_PM25 <dbl>
```

```r
head(Merged_2022)
```

```
## # A tibble: 6 x 8
##   County       `County FIPS Code` Percent_Below_150_Poverty E_MINRTY E_TOTPOP
##   <chr>        <chr>                                  <dbl>    <dbl>    <dbl>
## 1 Alameda      001                                     14.1  1176371  1663823
## 2 Butte        007                                     28.2    66604   213605
## 3 Calaveras    009                                     21.4     9927    45674
## 4 Colusa       011                                     22.6    14508    21811
## 5 Contra Costa 013                                     13.5   690897  1162648
## 6 Del Norte    015                                     25.3    10786    27462
## # i 3 more variables: Percent_Uninsured <dbl>, Minority_Percentage <dbl>,
## #   yearly_avg_PM25 <dbl>
```

```
## Reading layer `CA_Counties' from data source
##   `/home/guest/Project_main/DATA/CA_Counties.shp' using driver `ESRI Shapefile'
## Simple feature collection with 58 features and 19 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -13857270 ymin: 3832931 xmax: -12705030 ymax: 5162404
## Projected CRS: WGS 84 / Pseudo-Mercator
```

```r
# Adding 100% poverty level data to the SVI_CA_County_2022 dataset
# to make it comparable with 2000 data

# Keep only relevant columns and rename them
poverty_data_2022 <- poverty_data_2022 %>%
```

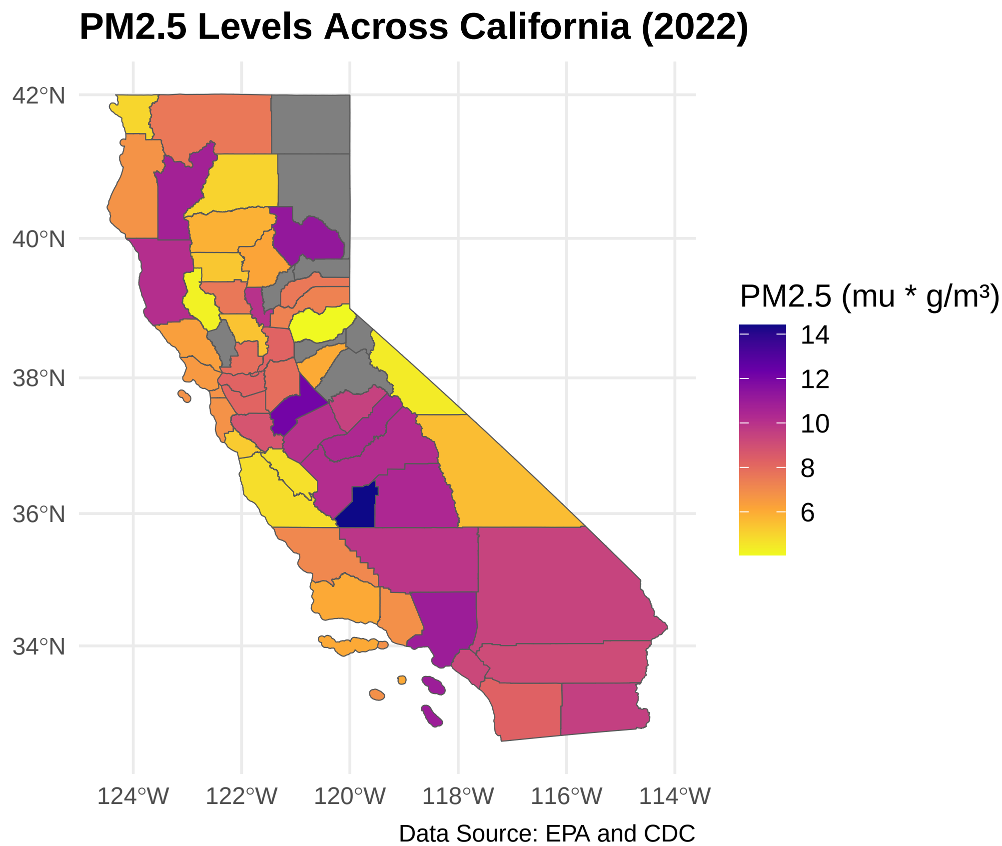

Figure 1: PM2.5 Levels Across California (2022)

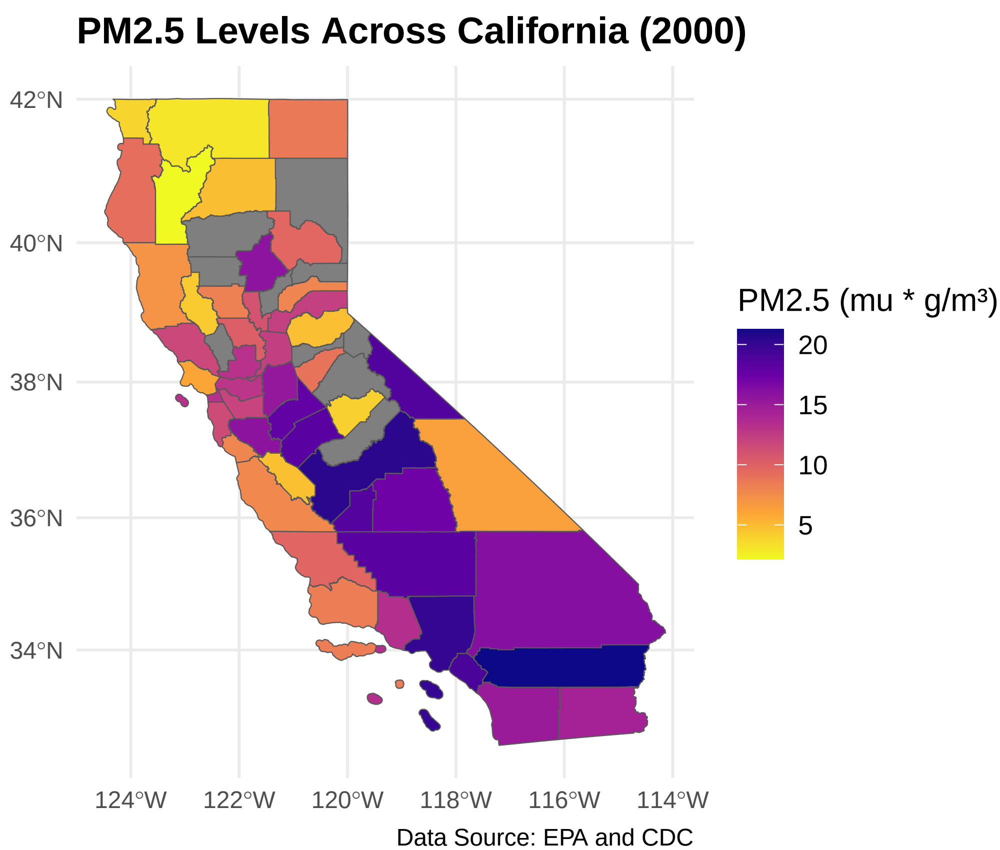

Figure 2: PM2.5 Levels Across California (2022)

```r
  select(
    `State FIPS Code`,                   # State FIPS Code
    `County FIPS Code`,                  # County FIPS Code
    `Poverty Percent, All Ages`          # Percent below 100% poverty
  ) %>%
  rename(
    `Percent_Below_Poverty` = `Poverty Percent, All Ages`,  # Rename for consistency
  )

# Filter for California data (State FIPS = 06)
california_poverty <- poverty_data_2022 %>%
  filter(`State FIPS Code` == "06") %>%
  select(`County FIPS Code`, `Percent_Below_Poverty`) # Keep only required columns

# Merge the poverty data into the SVI 2022 dataset
Merged_2022 <- Merged_2022 %>%
  left_join(california_poverty, by = "County FIPS Code")

# Inspect the updated dataset
head(Merged_2022)
```

```
## # A tibble: 6 x 9
##   County       `County FIPS Code` Percent_Below_150_Poverty E_MINRTY E_TOTPOP
##   <chr>        <chr>                                  <dbl>    <dbl>    <dbl>
## 1 Alameda      001                                     14.1  1176371  1663823
## 2 Butte        007                                     28.2    66604   213605
## 3 Calaveras    009                                     21.4     9927    45674
## 4 Colusa       011                                     22.6    14508    21811
## 5 Contra Costa 013                                     13.5   690897  1162648
## 6 Del Norte    015                                     25.3    10786    27462
## # i 4 more variables: Percent_Uninsured <dbl>, Minority_Percentage <dbl>,
## #   yearly_avg_PM25 <dbl>, Percent_Below_Poverty <dbl>
```

4.1 Correlation Analysis

```r
# This examines associations between PM2.5 and other variables
  #(Percent Uninsured, Percent Below Poverty, Minority Percentage)

# CORRELATION MATRIX - 2022
# Test associations between PM2.5 and socioeconomic variables for 2022 data
correlation_matrix_2022 <- cor(Merged_2022[, c("yearly_avg_PM25",
                                               "Percent_Uninsured",
                                               "Percent_Below_Poverty",
                                               "Percent_Below_150_Poverty",
                                               "Minority_Percentage")],
                               use = "complete.obs", method = "pearson")

# Print the correlation matrix for 2022
print(correlation_matrix_2022)
```

```
##                             yearly_avg_PM25 Percent_Uninsured
```

```
## yearly_avg_PM25                1.0000000        0.10152052
## Percent_Uninsured              0.1015205        1.00000000
## Percent_Below_Poverty          0.3833281        0.36294705
## Percent_Below_150_Poverty      0.3447585        0.46532629
## Minority_Percentage            0.3670272        0.06639129
##                          Percent_Below_Poverty Percent_Below_150_Poverty
## yearly_avg_PM25                      0.3833281                 0.3447585
## Percent_Uninsured                    0.3629470                 0.4653263
## Percent_Below_Poverty                1.0000000                 0.9493793
## Percent_Below_150_Poverty            0.9493793                 1.0000000
## Minority_Percentage                  0.1158359                 0.1187493
##                          Minority_Percentage
## yearly_avg_PM25                   0.36702715
## Percent_Uninsured                 0.06639129
## Percent_Below_Poverty             0.11583589
## Percent_Below_150_Poverty         0.11874926
## Minority_Percentage               1.00000000
```

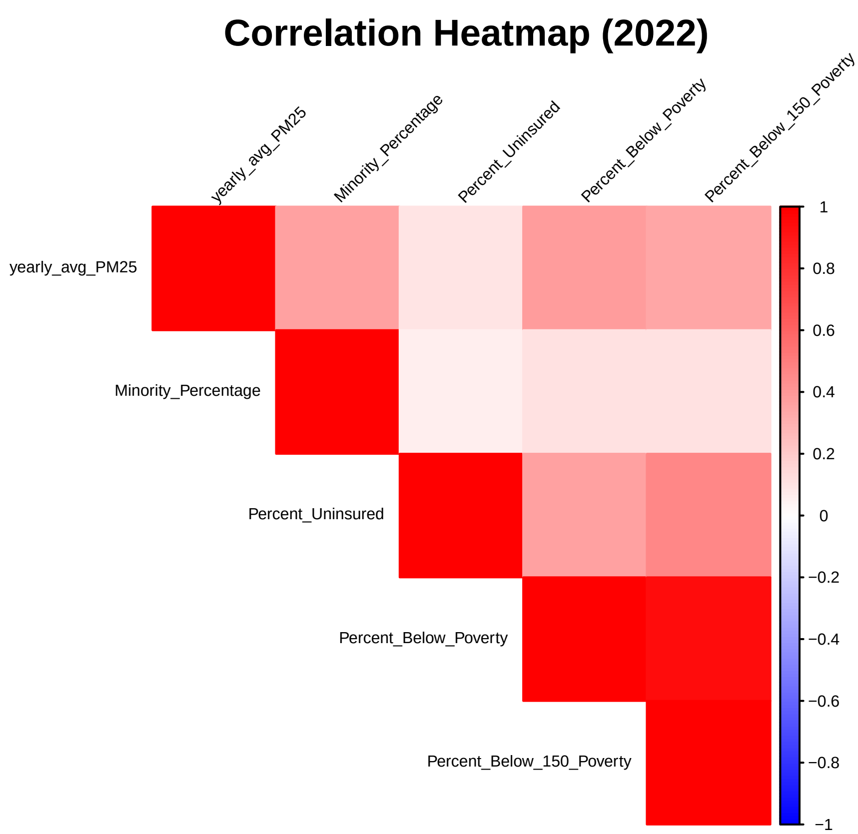

Figure 3: Correlation Heatmap (2022)

```r
# CORRELATION MATRIX - 2000
# Test associations between PM2.5 and socioeconomic variables for 2000 data
correlation_matrix_2000 <- cor(Merged_2000[, c("yearly_avg_PM25",
                                                "Percent_Below_Poverty",
                                                "Minority_Percentage")],
```

```r
                        use = "complete.obs", method = "pearson")

# Print the correlation matrix for 2000
print(correlation_matrix_2000)
```

```
##                       yearly_avg_PM25 Percent_Below_Poverty Minority_Percentage
## yearly_avg_PM25             1.0000000             0.1685785           0.6180766
## Percent_Below_Poverty       0.1685785             1.0000000           0.1946454
## Minority_Percentage         0.6180766             0.1946454           1.0000000
```

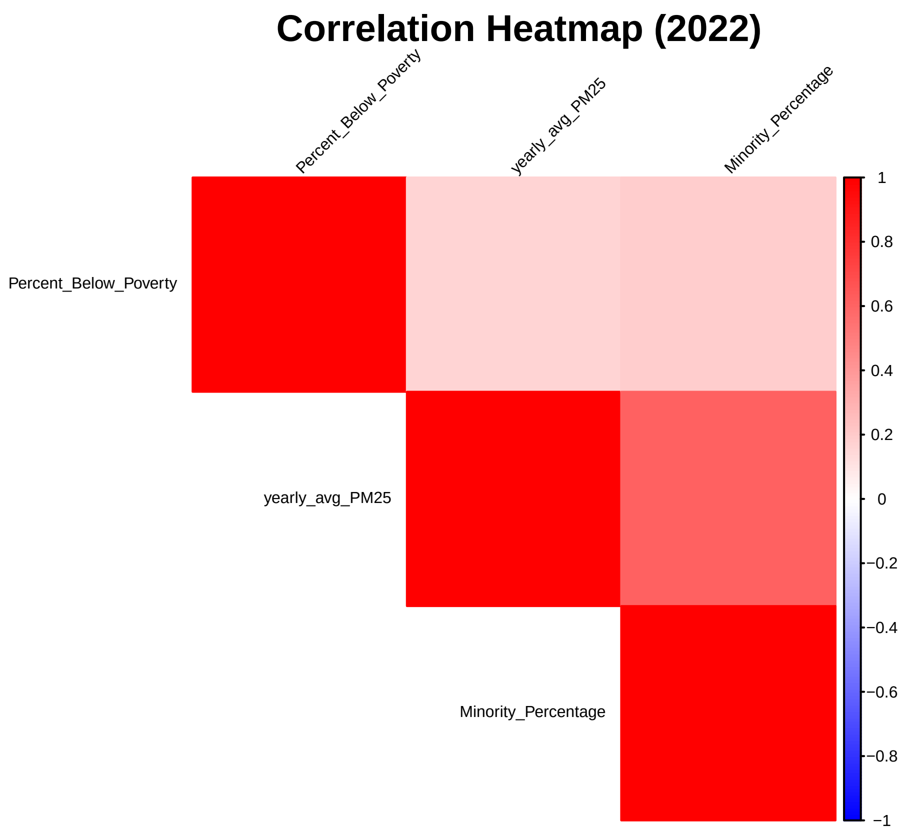

Figure 4: Correlation Heatmap (2000)

4.2 Multivariate Regression

Note: This is for exploring associations, not causation.

```r
# This models PM2.5 as a function of Percent Uninsured,
  # Percent Below Poverty, and Minority Percentage.
# This regression is for associational analysis, not to infer causation.

# REGRESSION MODELS - 2022
# MODEL 1: Using Percent Below Poverty
pm25_model_2022_poverty <-
```

```r
  lm(yearly_avg_PM25 ~
       Percent_Uninsured + Percent_Below_Poverty + Minority_Percentage,
     data = Merged_2022)
summary(pm25_model_2022_poverty)
```

```
##
## Call:
## lm(formula = yearly_avg_PM25 ~ Percent_Uninsured + Percent_Below_Poverty +
##     Minority_Percentage, data = Merged_2022)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7228 -1.6085  0.1525  1.1774  4.8895
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.35826    1.33486   2.516   0.0154 *
## Percent_Uninsured     -0.05278    0.13730  -0.384   0.7025
## Percent_Below_Poverty  0.21472    0.08089   2.655   0.0109 *
## Minority_Percentage    0.03864    0.01509   2.561   0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.062 on 46 degrees of freedom
## Multiple R-squared:  0.2548, Adjusted R-squared:  0.2062
## F-statistic: 5.244 on 3 and 46 DF,  p-value: 0.003386
```

```r
# MODEL 2: Using Percent Below 150% Poverty
pm25_model_2022_150_poverty <-
  lm(yearly_avg_PM25 ~
       Percent_Uninsured + Percent_Below_150_Poverty + Minority_Percentage,
     data = Merged_2022)
summary(pm25_model_2022_150_poverty)
```

```
##
## Call:
## lm(formula = yearly_avg_PM25 ~ Percent_Uninsured + Percent_Below_150_Poverty +
##     Minority_Percentage, data = Merged_2022)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7002 -1.5559  0.1983  0.9924  5.1776
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.77577    1.30961   2.883  0.00597 **
## Percent_Uninsured         -0.08029    0.14673  -0.547  0.58690
## Percent_Below_150_Poverty  0.11920    0.05106   2.335  0.02399 *
## Minority_Percentage        0.03904    0.01532   2.548  0.01424 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.094 on 46 degrees of freedom
## Multiple R-squared:  0.2317, Adjusted R-squared:  0.1816
## F-statistic: 4.624 on 3 and 46 DF,  p-value: 0.006565
```

```
# REGRESSION MODEL - 2000
# Association between PM2.5 and socioeconomic variables
pm25_model_2000 <-
  lm(yearly_avg_PM25 ~
       Percent_Below_Poverty + Minority_Percentage,
     data = Merged_2000)
summary(pm25_model_2000)
```

```
##
## Call:
## lm(formula = yearly_avg_PM25 ~ Percent_Below_Poverty + Minority_Percentage,
##     data = Merged_2000)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2839 -2.8851  0.4562  2.6248  9.7625
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.12689    2.12746   1.940   0.0587 .
## Percent_Below_Poverty  0.05574    0.13247   0.421   0.6759
## Minority_Percentage    0.17360    0.03403   5.102 6.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.217 on 45 degrees of freedom
## Multiple R-squared:  0.3844, Adjusted R-squared:  0.3571
## F-statistic: 14.05 on 2 and 45 DF,  p-value: 1.814e-05
```

Interpretation:

Both models for 2022 show that Minority Percentage and poverty measures (whether below poverty or 150% poverty) are significantly associated with higher PM2.5 levels, supporting the hypothesis that vulnerable populations are more exposed to pollution. However, Percent Uninsured does not show a significant relationship with PM2.5 in either model. The models explain about 23–25% of the variance, indicating moderate explanatory power.The 2000 regression model reveals a significant positive association between Minority Percentage and PM2.5 levels but Percent Below Poverty shows no significant relationship with PM2.5. This differs from the 2022 results, highlighting a possible shift over time in how demographic and socioeconomic factors relate to pollution exposure.

5 Analysis

5.1 Comparative Analysis

```
# Merge the 2000 and 2022 datasets for paired analysis
differences <- Merged_2000 %>%
  inner_join(Merged_2022, by = "County FIPS Code", suffix = c("_2000", "_2022")) %>%
```

```r
  mutate(
    PM2.5_Difference = yearly_avg_PM25_2022 - yearly_avg_PM25_2000,
    Poverty_Difference = Percent_Below_Poverty_2022 - Percent_Below_Poverty_2000,
    Minority_Difference = Minority_Percentage_2022 - Minority_Percentage_2000,
    PM2.5_Change = ifelse(PM2.5_Difference > 0.01, "Increase",
                          ifelse(PM2.5_Difference < -0.01, "Decrease", "No Change")),
    Poverty_Change = ifelse(Poverty_Difference > 0.01, "Increase",
                            ifelse(Poverty_Difference < -0.01, "Decrease", "No Change")),
    Minority_Change = ifelse(Minority_Difference > 0.01, "Increase",
                             ifelse(Minority_Difference < -0.01, "Decrease", "No Change"))
  )


# County-Level Change Summary
# Paired Comparison Results
paired_comparison_results <- data.frame(
  Metric = c("PM2.5", "Percent Below Poverty", "Minority Percentage"),
  Mean_Change = c(
    mean(differences$PM2.5_Difference, na.rm = TRUE),
    mean(differences$Poverty_Difference, na.rm = TRUE),
    mean(differences$Minority_Difference, na.rm = TRUE)
  ),
  T_Statistic = c(
    t.test(differences$PM2.5_Difference, mu = 0)$statistic,
    t.test(differences$Poverty_Difference, mu = 0)$statistic,
    t.test(differences$Minority_Difference, mu = 0)$statistic
  ),
  P_Value = c(
    t.test(differences$PM2.5_Difference, mu = 0)$p.value,
    t.test(differences$Poverty_Difference, mu = 0)$p.value,
    t.test(differences$Minority_Difference, mu = 0)$p.value
  )
)

county_level_summary <- data.frame(
  Metric = c("PM2.5", "Percent Below Poverty", "Minority Percentage"),
  Increase = sapply(c("PM2.5_Change", "Poverty_Change", "Minority_Change"),
                    function(col) sum(differences[[col]] == "Increase")),
  Decrease = sapply(c("PM2.5_Change", "Poverty_Change", "Minority_Change"),
                    function(col) sum(differences[[col]] == "Decrease")),
  No_Change = sapply(c("PM2.5_Change", "Poverty_Change", "Minority_Change"),
                     function(col) sum(differences[[col]] == "No Change"))
)

# PM2.5 Changes By Poverty Trends
poverty_pm25_summary <- differences %>%
  group_by(Poverty_Change) %>%
  summarise(
    Average_PM2.5_Change = mean(PM2.5_Difference, na.rm = TRUE),
    Number_of_Counties = n()
  ) %>%
  filter(Poverty_Change %in% c("Increase", "Decrease"))
```

```r
# Print the results
# Print the Paired Comparison Table
print("Paired Comparison Results for 2000 vs. 2022")
```

```
## [1] "Paired Comparison Results for 2000 vs. 2022"
```

```r
print(paired_comparison_results)
```

```
##                  Metric Mean_Change T_Statistic      P_Value
## 1                 PM2.5  -3.7641671   -5.645080 9.823930e-07
## 2 Percent Below Poverty  -0.7963269   -2.944006 5.065187e-03
## 3   Minority Percentage  12.3181269   24.042011 1.091449e-27
```

```r
print("County-Level Change Summary")
```

```
## [1] "County-Level Change Summary"
```

```r
print(county_level_summary)
```

```
##                                 Metric Increase Decrease No_Change
## PM2.5_Change                     PM2.5        8       39         0
## Poverty_Change  Percent Below Poverty       17       30         0
## Minority_Change   Minority Percentage       47        0         0
```

```r
print("PM2.5 Changes By Poverty Trends")
```

```
## [1] "PM2.5 Changes By Poverty Trends"
```

```r
print(poverty_pm25_summary)
```

```
## # A tibble: 2 x 3
##   Poverty_Change Average_PM2.5_Change Number_of_Counties
##   <chr>                         <dbl>              <int>
## 1 Decrease                      -4.31                 30
## 2 Increase                      -2.81                 17
```

Interpretation: Over the past 22 years, there has been a significant reduction in PM2.5 levels and poverty rates, indicating improvements in air quality and socioeconomic conditions. However, the minority population percentage has significantly increased, reflecting notable demographic shifts. These trends suggest progress in environmental and economic factors, alongside evolving population dynamics, which may have implications for policy and resource allocation in addressing environmental justice and equity. In counties where poverty decreased, PM2.5 levels also decreased significantly, with an average reduction of 4.31 units. Conversely, in counties where poverty increased, PM2.5 levels still decreased on average, but by a smaller margin of 2.81 units. This suggests that PM2.5 has generally declined across counties, regardless of poverty trends, with a greater reduction observed in counties experiencing poverty decreases.

5.2 Linear Regression

```r
model <-
  lm(yearly_avg_PM25 ~ Percent_Below_150_Poverty + Minority_Percentage,
     data = Merged_2022)
summary(model)
```

```
##
## Call:
## lm(formula = yearly_avg_PM25 ~ Percent_Below_150_Poverty + Minority_Percentage,
##     data = Merged_2022)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8246 -1.5909  0.2719  0.9759  5.2138
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.51153    1.20821   2.906  0.00556 **
## Percent_Below_150_Poverty  0.10630    0.04495   2.365  0.02222 *
## Minority_Percentage        0.03893    0.01521   2.560  0.01373 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.078 on 47 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.1938
## F-statistic:  6.89 on 2 and 47 DF,  p-value: 0.002377
```

```r
model <-
  lm(yearly_avg_PM25 ~ Percent_Below_Poverty + Minority_Percentage, data = Merged_2000)
summary(model)
```

```
##
## Call:
## lm(formula = yearly_avg_PM25 ~ Percent_Below_Poverty + Minority_Percentage,
##     data = Merged_2000)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2839 -2.8851  0.4562  2.6248  9.7625
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.12689    2.12746   1.940   0.0587 .
## Percent_Below_Poverty  0.05574    0.13247   0.421   0.6759
## Minority_Percentage    0.17360    0.03403   5.102 6.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.217 on 45 degrees of freedom
## Multiple R-squared:  0.3844, Adjusted R-squared:  0.3571
## F-statistic: 14.05 on 2 and 45 DF,  p-value: 1.814e-05
```

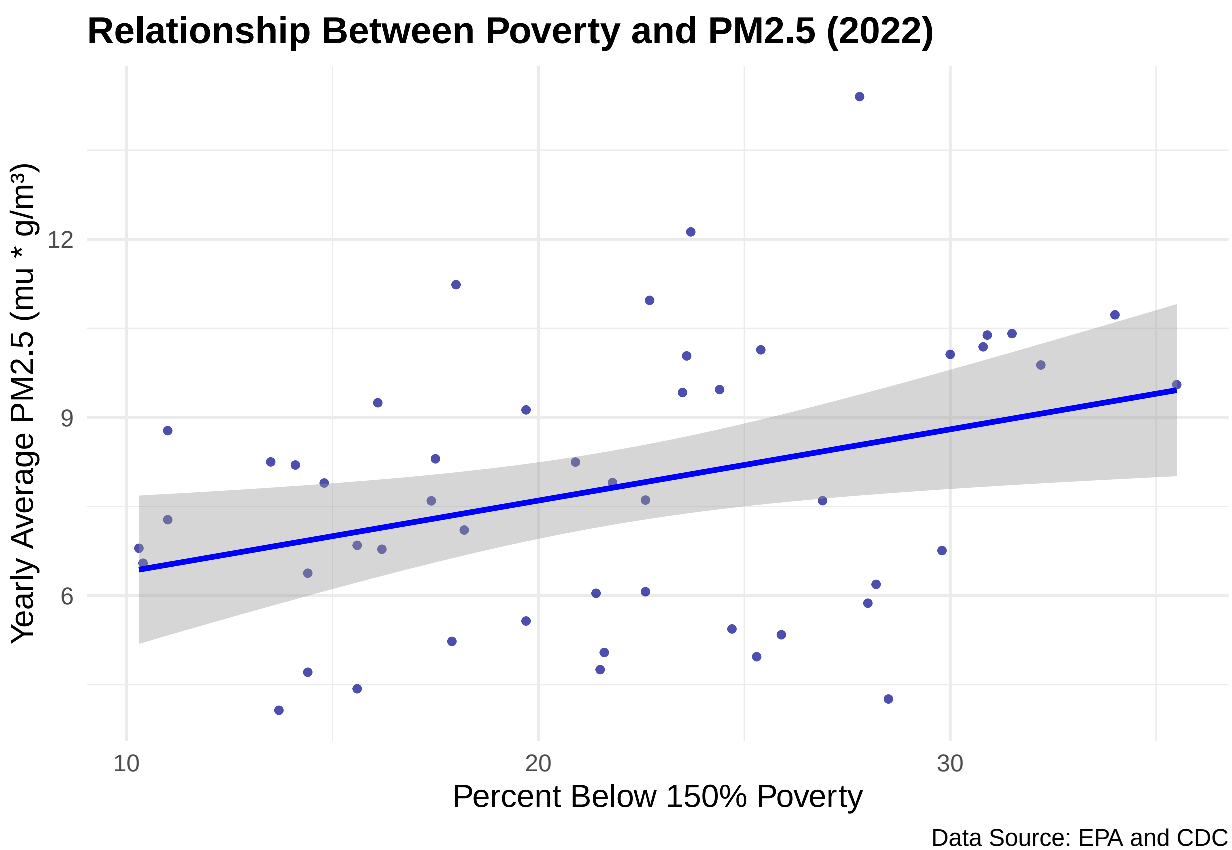

Figure 5: Relationship Between Poverty and PM2.5 (2022)

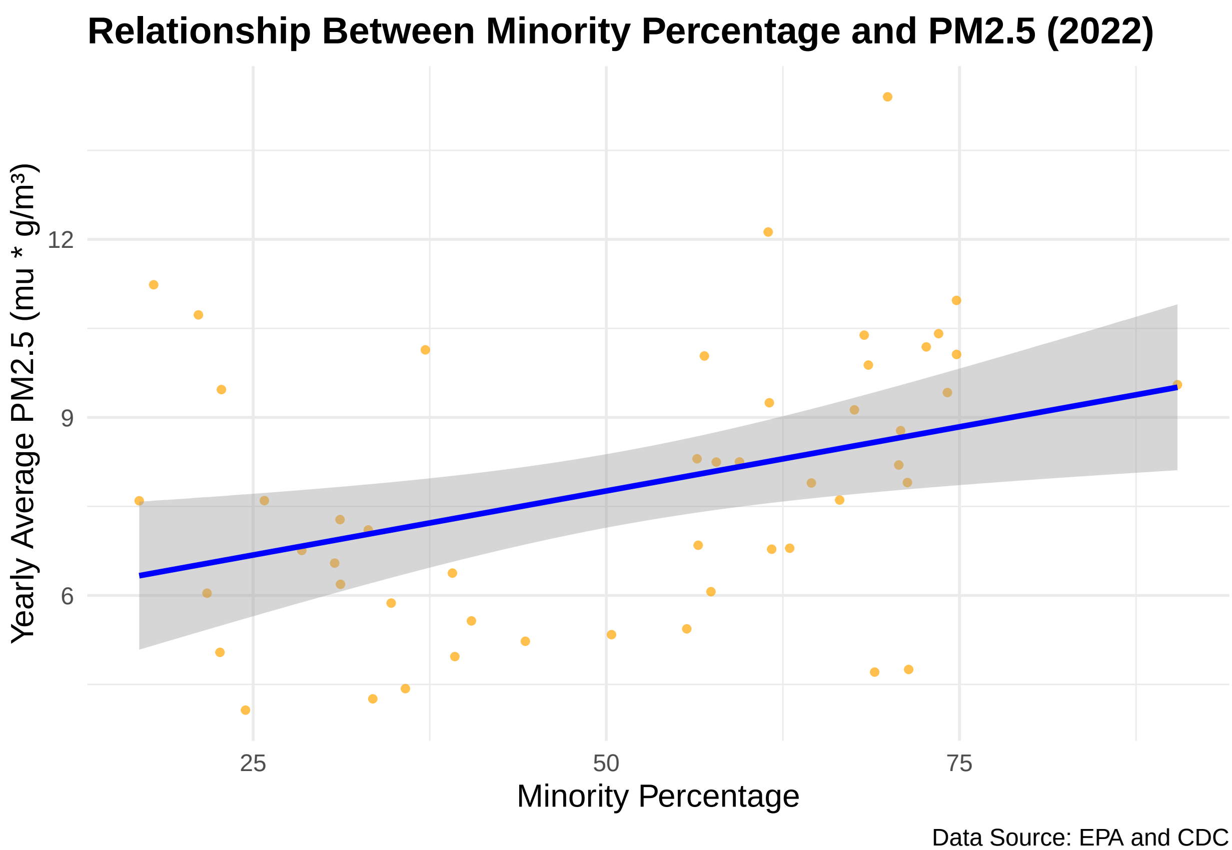

Figure 6: Relationship Between Minority Percentage and PM2.5 (2022)

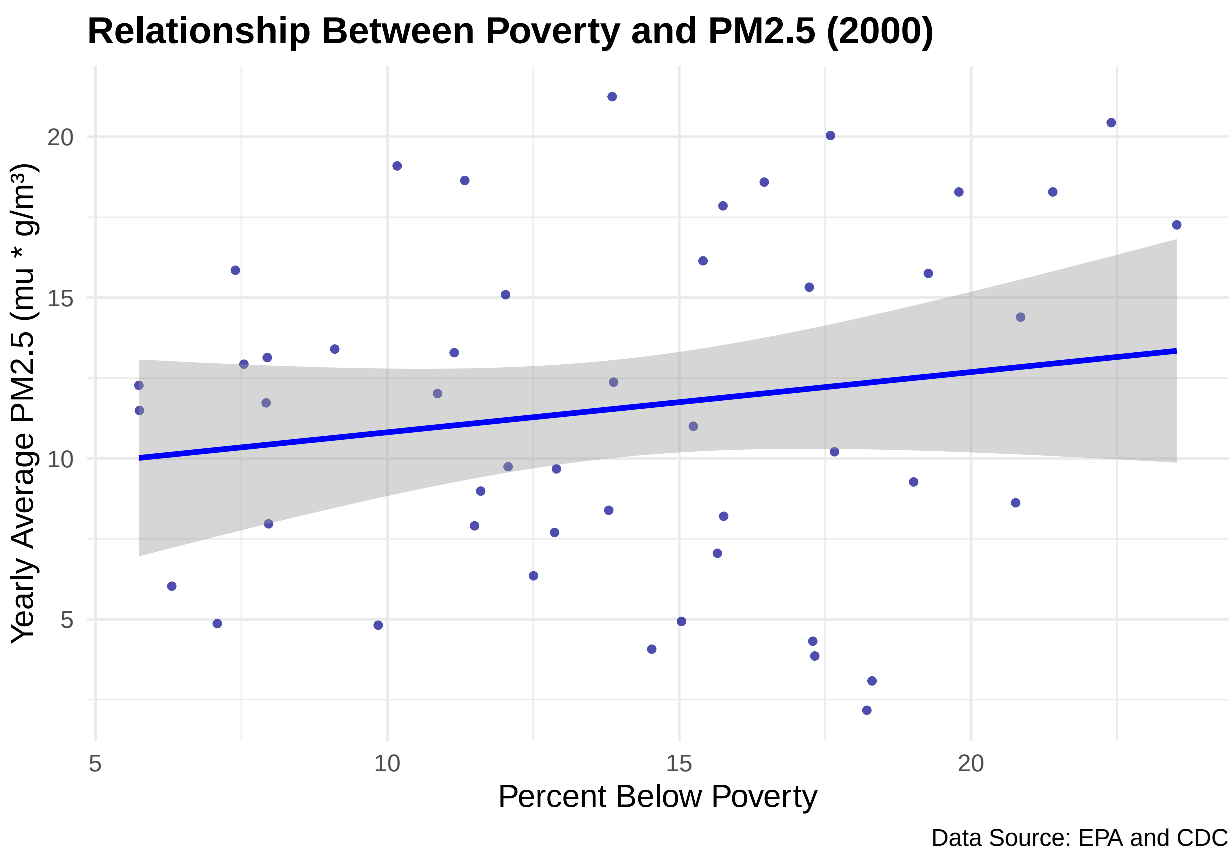

Figure 7: Relationship Between Poverty and PM2.5 (2000)

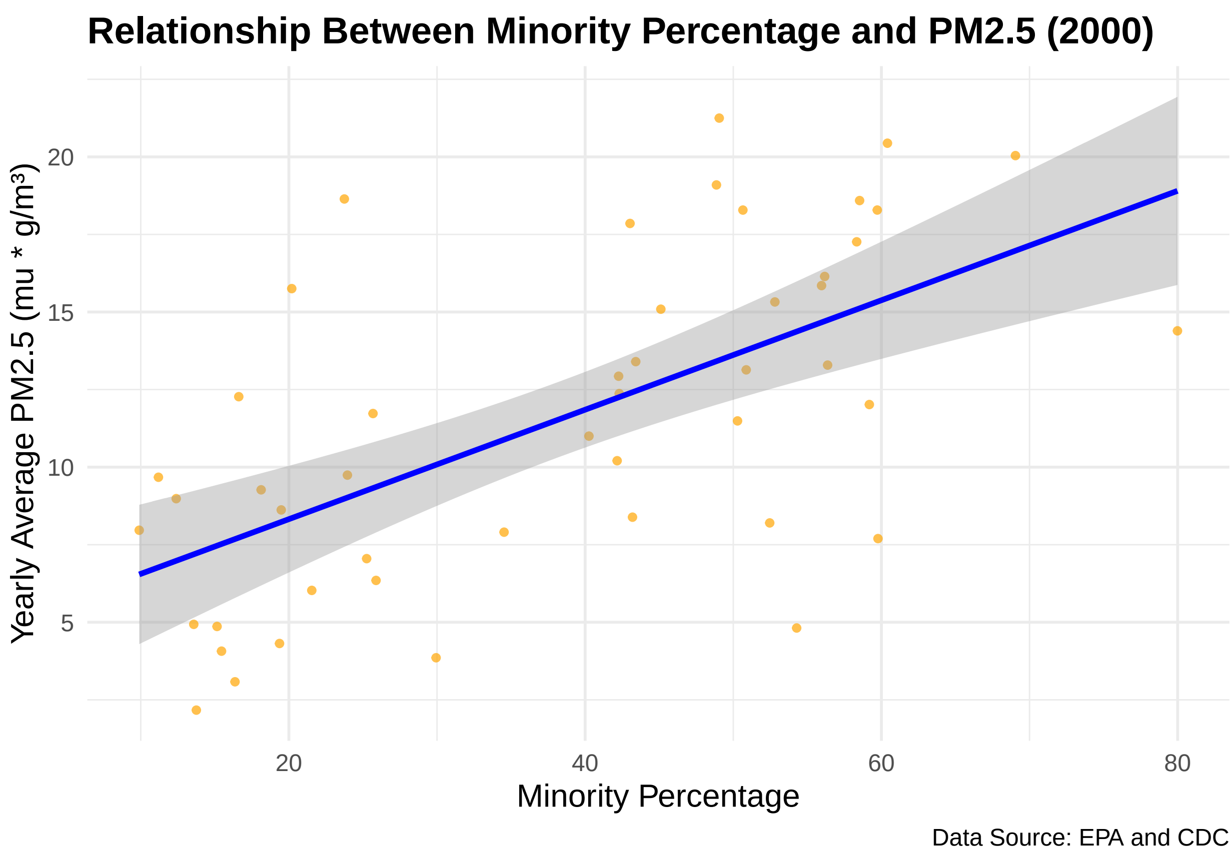

Figure 8: Relationship Between Minority Percentage and PM2.5 (2000)

6 Summary and Conclusions

This project investigates the relationship between air pollution levels, measured by PM2.5 concentrations, and social vulnerability indicators, such as poverty, minority population percentage, and uninsured rates, across California counties for the years 2000 and 2022. The analysis integrates data from the U.S. Environmental Protection Agency (EPA) and the Centers for Disease Control and Prevention (CDC) Social Vulnerability Index (SVI). Key steps in this study included:

1. Cleaning and merging PM2.5 and SVI datasets by county FIPS codes to create unified datasets for both years.
2. Exploratory data analysis using summary statistics, maps, and visualizations to understand spatial distributions of PM2.5 and socioeconomic factors.
3. Correlation analysis to identify associations between PM2.5 levels and social vulnerability indicators.
4. Multivariate regression modeling to evaluate the combined effects of socioeconomic factors on PM2.5 levels.

The findings revealed significant relationships between PM2.5 levels and social vulnerability indicators. In 2022, counties with higher minority percentages and poverty rates exhibited higher PM2.5 levels. Similarly, the analysis for 2000 identified a positive association between PM2.5 levels and minority percentages, though the association with poverty was weaker. Comparisons between 2000 and 2022 highlighted a notable decline in PM2.5 levels over time, accompanied by reductions in poverty rates, while minority percentages increased.

This study highlights the intersection between environmental quality and social equity, emphasizing that vulnerable communities face disproportionate exposure to air pollution. The result indicates 1) minority populations and economically disadvantaged communities are more exposed to higher PM2.5 levels, reinforcing concerns about environmental justice. 2) From 2000 to 2022, there was a significant reduction in PM2.5 levels across California counties, coinciding with poverty reductions. However, despite these improvements, disparities persist, particularly in counties with high minority populations.3) Policies aimed at improving air quality must address the unequal burden on vulnerable populations. Intersectional approaches are needed to target areas where social vulnerabilities overlap with high pollution levels. 5) Future Research: Additional studies should explore causal mechanisms, focusing on factors such as proximity to pollution sources (e.g., industrial facilities or highways) and changes in county-level demographics.

This analysis highlights the importance of integrating environmental and social data to inform policies promoting equity and sustainability. By reducing pollution exposure in vulnerable communities, policymakers can advance both public health and environmental justice objectives.