

CODE FOR ASTHMA PROJECT

Eve Rodler

2024-11-05

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Clean Up Data

County level summary data: average asthma_pctl per county

```
# By averaging the asthma percent measures for each county,  
# this does not pick up areas within certain counties with higher  
# distributions of asthma.  
  
asthma_measures_df3 <- asthma_measures_df2 %>%  
  group_by(county) %>%  
  summarize(asthma_mean_percent_by_county=  
    mean(asthma_pctl, na.rm = TRUE), .groups = "drop") %>% ungroup()  
  
asthma_measures_df3 <- asthma_measures_df3 %>%  
  arrange(desc(asthma_mean_percent_by_county))
```

It is important to consider impact of census tracts within counties

```
asthma_measures_df4 <- asthma_measures_df2 %>% arrange(county, desc(asthma_pctl))  
  
# Reduce observations from 8035 to 803  
# by subsetting highest percentile (asthma >= 90th percentile)  
  
asthma_measures_df5 <- asthma_measures_df2 %>%  
  filter(asthma_pctl >= 90) %>%  
  arrange(county, desc(asthma_pctl))  
  
# 25 out of 58 counties are represented in asthma top 10 percentile (90 - 100)  
# which means certain census tracts within multiple counties have the highest  
# distribution of asthma cases. This needs to be considered for  
# county level interventions.  
  
unique(asthma_measures_df5$county)
```

```
## [1] "Alameda" "Amador" "Contra Costa" "Fresno"
```

```
## [5] "Humboldt"      "Imperial"      "Kern"          "Kings"
## [9] "Lake"          "Los Angeles"   "Madera"        "Mendocino"
## [13] "Merced"        "Modoc"         "Monterey"      "Plumas"
## [17] "Riverside"     "Sacramento"    "San Bernardino" "San Diego"
## [21] "San Francisco" "San Joaquin"   "Santa Clara"   "Solano"
## [25] "Stanislaus"

# [1] "Solano"          "San Joaquin"    "Alameda"        "San Bernardino"
# "Santa Clara"    "Sacramento"
# [7] "Contra Costa"   "Kern"           "Los Angeles"    "Imperial"
# "Fresno"         "Merced"
# [13] "Lake"           "San Francisco" "Modoc"          "Kings"
# "Stanislaus"     "San Diego"
# [19] "Amador"         "Riverside"     "Humboldt"       "Mendocino"
# "Monterey"       "Plumas"
# [25] "Madera"
```

If restrict dataset to top 5% of asthma then 16 counties are represented

with 402 observations.

```
asthma_measures_df6 <- asthma_measures_df2 %>%
  filter(asthma_pctl >= 95) %>%
  arrange(county, desc(asthma_pctl))

unique(asthma_measures_df6$county)
```

```
## [1] "Alameda"      "Contra Costa" "Fresno"        "Imperial"
## [5] "Kern"         "Kings"        "Lake"          "Los Angeles"
## [9] "Merced"       "Modoc"        "Sacramento"    "San Bernardino"
## [13] "San Francisco" "San Joaquin"  "Santa Clara"   "Solano"
```

```
# [1] "Alameda"      "Contra Costa" "Fresno"        "Imperial"    "Kern"        "Kings"
# [7] "Lake"         "Los Angeles"  "Merced"        "Modoc"       "Sacramento"  "San Bernard
# [13] "San Francisco" "San Joaquin"  "Santa Clara"   "Solano"
```

```
# If restrict dataset tp top 1% of asthma, then 10 counties are
# represented with 81 observations.
```

```
asthma_measures_df9 <- asthma_measures_df2 %>%
  filter(asthma_pctl >= 99) %>%
  arrange(county, desc(asthma_pctl))

unique(asthma_measures_df9$county)
```

```
## [1] "Alameda"      "Contra Costa" "Imperial"      "Kern"
## [5] "Los Angeles"  "Sacramento"   "San Bernardino" "San Joaquin"
## [9] "Santa Clara"  "Solano"
```

```
# [1] "Alameda"      "Contra Costa" "Imperial"      "Kern"
# "Los Angeles"  "Sacramento"
# [7] "San Bernardino" "San Joaquin"  "Santa Clara"   "Solano"
```

```
# Reorder the columns to move asthma_pctl to 3rd position.
```

```
asthma_measures_df8 <- asthma_measures_df6 %>%
```

```
select(1:2, asthma_pctl, 3:6, 8:ncol(asthma_measures_df6))
```

Will keep dataset with top 5% of asthma percentile and compare to other variables.

For comparison to other variables, will categorize the Pollution and Population Characteristic variables into Very High, High, Medium, Low categories.

```
asthma_measures_compare1 <- asthma_measures_df8 %>%
  mutate(pm2.5_pctl_cat =
    case_when(pm2.5_pctl < 50 ~ "Low", pm2.5_pctl < 90 ~ "Medium",
              pm2.5_pctl < 95 ~ "High", TRUE ~ "Very High"))

asthma_measures_compare2 <- asthma_measures_compare1 %>%
  mutate(diesel_pm_pctl_cat =
    case_when(diesel_pm_pctl < 50 ~ "Low", diesel_pm_pctl < 90 ~ "Medium",
              diesel_pm_pctl < 95 ~ "High", TRUE ~ "Very High"))

asthma_measures_compare3 <- asthma_measures_compare2 %>%
  mutate(traffic_pctl_cat =
    case_when(traffic_pctl < 50 ~ "Low", traffic_pctl < 90 ~ "Medium",
              traffic_pctl < 95 ~ "High", TRUE ~ "Very High"))

asthma_measures_compare4 <- asthma_measures_compare3 %>%
  mutate(tox._release_pctl_cat =
    case_when(tox._release_pctl < 50 ~ "Low", tox._release_pctl < 90 ~ "Medium",
              tox._release_pctl < 95 ~ "High", TRUE ~ "Very High"))

asthma_measures_compare5 <- asthma_measures_compare4 %>%
  mutate(housing_burden_pctl_cat =
    case_when(housing_burden_pctl < 50 ~ "Low", housing_burden_pctl < 90 ~ "Medium",
              housing_burden_pctl < 95 ~ "High",
              housing_burden_pctl <=100 ~ "Very High", TRUE ~ NA))

asthma_measures_compare6 <- asthma_measures_compare5 %>%
  mutate(education_pctl_cat =
    case_when(education_pctl < 50 ~ "Low", education_pctl < 90 ~ "Medium",
              education_pctl < 95 ~ "High",
              education_pctl <=100 ~ "Very High", TRUE ~ NA))

asthma_measures_compare7 <- asthma_measures_compare6 %>%
  mutate(unemployment_pctl_cat =
    case_when(unemployment_pctl < 50 ~ "Low",
              unemployment_pctl < 90 ~ "Medium",
              unemployment_pctl < 95 ~ "High",
              unemployment_pctl <=100 ~ "Very High", TRUE ~ NA))

asthma_measures_compare8 <- asthma_measures_compare7 %>%
  mutate(poverty_pctl_cat =
    case_when(poverty_pctl < 50 ~ "Low", poverty_pctl < 90 ~ "Medium",
              poverty_pctl < 95 ~ "High",
              poverty_pctl <=100 ~ "Very High", TRUE ~ NA))
```

Exploring correlation between asthma and pollution variables in asthma 90 percentile.

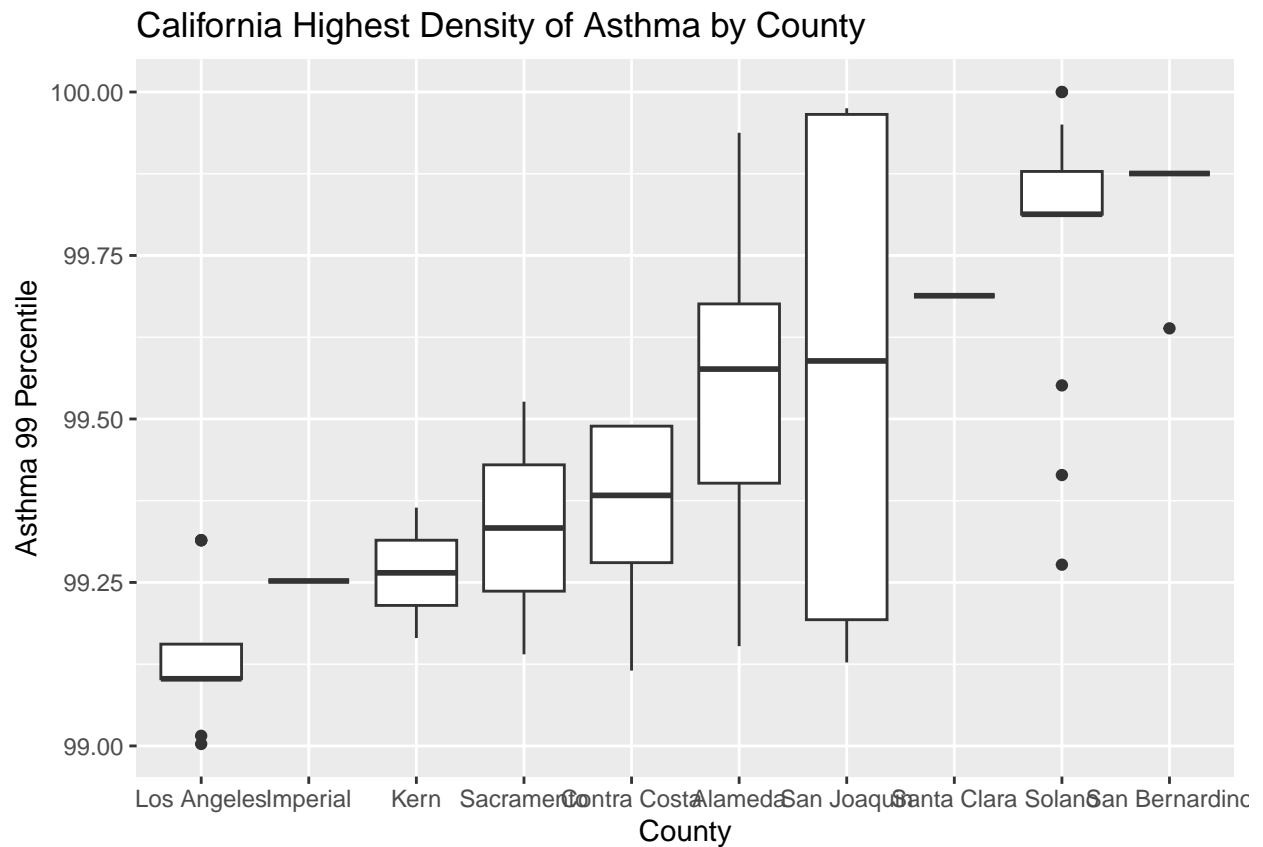
Strength of correlations, for the absolute value of correlation coefficients (based on Pearson's correlation):

0.00 - 0.19: Very weak or negligible correlation

0.20 - 0.39: Weak correlation

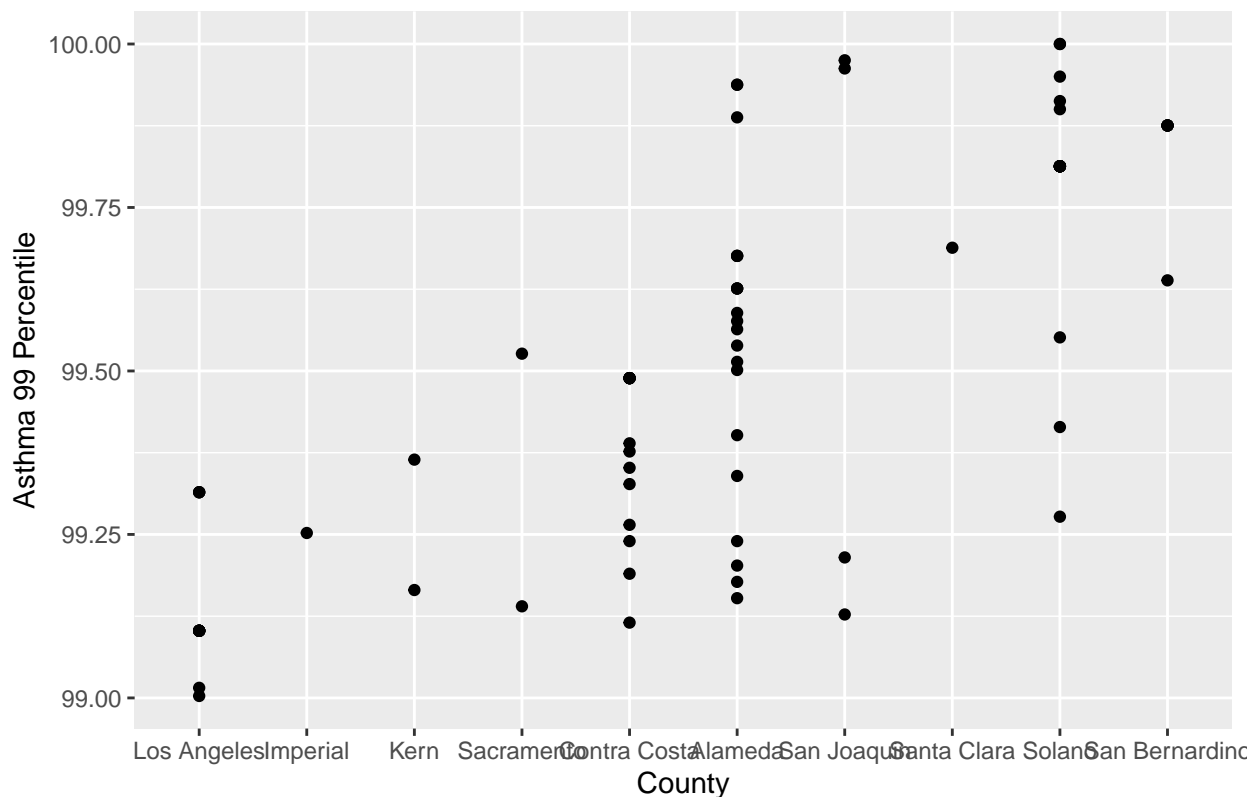
0.40 - 0.59: Moderate correlation 0.60 - 0.79: Strong correlation 0.80 - 1.00: Very strong correlation

```
cor_matrix <-  
  cor(asthma_measures_df2[, c("asthma_pctl", "pm2.5_pctl", "diesel_pm_pctl",  
                             "traffic_pctl", "tox._release_pctl")],  
      use = "complete.obs")  
  
      #asthma_pctl pm2_5_pctl diesel_pm_pctl traffic_pctl tox_release_pctl  
# asthma_pctl      1.000000000 -0.1606919      0.0213816  -0.07081408      -0.002961313  
  
# Running the code assuming data set of 95 asthma_pctl,  
# also no correlations other than a weak correlation with PM.2.5 percentile:  
  
cor_matrix_95_pctl <-  
  cor(asthma_measures_df6[, c("asthma_pctl", "pm2.5_pctl", "diesel_pm_pctl",  
                             "traffic_pctl", "tox._release_pctl")],  
      use = "complete.obs")  
  
      #asthma_pctl pm2_5_pctl diesel_pm_pctl traffic_pctl tox_release_pctl  
# asthma_pctl      1.000000000 -0.3451474  -0.0001463002  -0.1452523      -0.16277202  
  
# ggplot to explore relationship between county and asthma percentile  
  
ggplot(asthma_measures_df9,  
  aes(x = fct_reorder(county, asthma_pctl), y = asthma_pctl)) +  
  geom_boxplot() +  
  labs(title = "California Highest Density of Asthma by County", x = "County",  
       y = "Asthma 99 Percentile") + scale_y_continuous()
```



```
ggplot(asthma_measures_df9,
  aes(x = fct_reorder(county, asthma_pctl), y = asthma_pctl)) +
  geom_point() + labs(title = "California Highest Density of Asthma by County",
    x = "County", y = "Asthma 99 Percentile") +
  scale_y_continuous()
```

California Highest Density of Asthma by County



Summary county-level asthma correlation with pollution variable pm2.5_pctl

```
asthma_measures_compare10 <- asthma_measures_compare8 %>%
  filter(pm2.5_pctl_cat %in% c("Very High", "High"))

unique(asthma_measures_compare10$county)

## [1] "Fresno" "Kern" "Kings"
# [1] "Kern" "Fresno" "Kings"

asthma_measures_compare10_sum <- asthma_measures_compare10 %>%
  select(census_tract, county, asthma_pctl, pm2.5_pctl_cat) %>%
  group_by(county) %>% ungroup()

print(asthma_measures_compare10_sum)

## # A tibble: 56 x 4
##   census_tract county asthma_pctl pm2.5_pctl_cat
##   <dbl> <chr> <dbl> <chr>
## 1 6019003302 Fresno 98.6 Very High
## 2 6019008502 Fresno 98.6 Very High
## 3 6019003702 Fresno 98.5 Very High
## 4 6019004704 Fresno 98.5 Very High
## 5 6019000700 Fresno 98.2 Very High
## 6 6019001000 Fresno 98.2 Very High
## 7 6019000300 Fresno 98.2 Very High
```

```
## 8 6019000902 Fresno 98.2 Very High
## 9 6019000901 Fresno 98.2 Very High
## 10 6019000800 Fresno 98.2 High
## # i 46 more rows
```

Summary county-level asthma correlation with pollution variable

traffic_pctl

```
asthma_measures_compare11<- asthma_measures_compare8 %>%
  filter(traffic_pctl_cat %in% c("Very High", "High"))

unique(asthma_measures_compare11$county)
```

```
## [1] "Alameda"      "Contra Costa" "Fresno"      "Los Angeles"
## [5] "Sacramento"   "San Francisco" "Solano"
```

```
# [1] "Alameda"      "Solano"      "Los Angeles" "Sacramento" "Fresno"
# "San Francisco" "Contra Costa"
```

```
asthma_measures_compare11_sum <- asthma_measures_compare11 %>%
  select(census_tract, county, asthma_pctl, traffic_pctl_cat) %>%
  group_by(county) %>% ungroup()

print(asthma_measures_compare11_sum)
```

```
## # A tibble: 29 x 4
##   census_tract county      asthma_pctl traffic_pctl_cat
##   <dbl> <chr>      <dbl> <chr>
## 1 6001409100 Alameda      99.9 Very High
## 2 6001408300 Alameda      99.6 High
## 3 6001401700 Alameda      97.2 High
## 4 6001409200 Alameda      96.5 Very High
## 5 6001425101 Alameda      95.8 Very High
## 6 6013315000 Contra Costa 95.7 High
## 7 6019002400 Fresno      97.0 High
## 8 6019002502 Fresno      95.9 High
## 9 6037900605 Los Angeles 99.1 Very High
## 10 6037900803 Los Angeles 99.0 Very High
## # i 19 more rows
```

Summary county-level asthma correlation with pollution variable

diesel_pm_pctl

```
asthma_measures_compare12 <- asthma_measures_compare8 %>%
  filter(diesel_pm_pctl_cat %in% c("Very High", "High"))

unique(asthma_measures_compare12$county)
```

```
## [1] "Alameda"      "Contra Costa" "Fresno"      "Kern"
## [5] "Los Angeles"   "Merced"      "San Bernardino" "San Francisco"
## [9] "San Joaquin"   "Santa Clara"
```

```
# [1] "San Joaquin"   "Alameda"      "Santa Clara"   "Contra Costa" "Kern"
# "San Bernardino"
# [7] "Los Angeles"   "Merced"      "Fresno"      "San Francisco"
```

```
asthma_measures_compare12_sum <- asthma_measures_compare12 %>%
  select(census_tract, county, asthma_pctl, diesel_pm_pctl_cat) %>%
  group_by(county) %>% ungroup()

print(asthma_measures_compare12_sum)
```

```
## # A tibble: 74 x 4
##   census_tract county  asthma_pctl diesel_pm_pctl_cat
##   <dbl> <chr>      <dbl> <chr>
## 1  6001409100 Alameda      99.9 High
## 2  6001409500 Alameda      99.7 High
## 3  6001409000 Alameda      99.5 High
## 4  6001402200 Alameda      99.0 Very High
## 5  6001410500 Alameda      99.0 High
## 6  6001402500 Alameda      99.0 Very High
## 7  6001401800 Alameda      99.0 Very High
## 8  6001402400 Alameda      99.0 High
## 9  6001983200 Alameda      99.0 Very High
## 10 6001981900 Alameda      99.0 Very High
## # i 64 more rows
```

Summary county-level asthma correlation with pollution variable

tox._release_pctl

```
asthma_measures_compare13 <- asthma_measures_compare8 %>%
  filter(tox._release_pctl_cat %in% c("Very High", "High"))

unique(asthma_measures_compare13$county)
```

```
## [1] "Contra Costa" "Fresno" "Los Angeles"
```

```
# [1] "Los Angeles" "Contra Costa" "Fresno"
```

```
asthma_measures_compare13_sum <- asthma_measures_compare13 %>%
  select(census_tract, county, asthma_pctl, tox._release_pctl_cat) %>%
  group_by(county) %>% ungroup()

print(asthma_measures_compare13_sum)
```

```
## # A tibble: 83 x 4
##   census_tract county  asthma_pctl tox._release_pctl_cat
##   <dbl> <chr>      <dbl> <chr>
## 1  6013314103 Contra Costa      97.9 Very High
## 2  6013378000 Contra Costa      97.2 High
## 3  6019001100 Fresno          97.2 High
## 4  6019001304 Fresno          95.2 High
## 5  6037900806 Los Angeles      99.3 High
## 6  6037900703 Los Angeles      99.3 High
## 7  6037900505 Los Angeles      99.1 Very High
## 8  6037900605 Los Angeles      99.1 High
## 9  6037900608 Los Angeles      99.1 High
## 10 6037900609 Los Angeles      99.1 High
## # i 73 more rows
```


Summary county-level asthma correlation with population characteristic variable

housing_burden_pctl

```
asthma_measures_compare14 <- asthma_measures_compare8 %>%
  filter(housing_burden_pctl_cat %in% c("Very High", "High"))

unique(asthma_measures_compare14$county)

## [1] "Alameda"      "Contra Costa"  "Fresno"        "Imperial"
## [5] "Kern"         "Los Angeles"   "Merced"        "Sacramento"
## [9] "San Bernardino" "San Francisco" "San Joaquin"    "Solano"

# [1] "San Joaquin"    "Alameda"        "Solano"          "Sacramento"
# "Contra Costa"  "Kern"
# [7] "Los Angeles"    "Fresno"          "Merced"          "Imperial"
# "San Francisco" "San Bernardino"

asthma_measures_compare14_sum <- asthma_measures_compare14 %>%
  select(census_tract, county, asthma_pctl, housing_burden_pctl_cat) %>%
  group_by(county) %>% ungroup()

print(asthma_measures_compare14_sum)
```

```
## # A tibble: 110 x 4
##   census_tract county  asthma_pctl housing_burden_pctl_cat
##   <dbl> <chr>      <dbl> <chr>
## 1  6001409100 Alameda    99.9 Very High
## 2  6001409500 Alameda    99.7 High
## 3  6001408900 Alameda    99.7 High
## 4  6001408400 Alameda    99.6 Very High
## 5  6001409400 Alameda    99.6 High
## 6  6001408600 Alameda    99.2 Very High
## 7  6001408700 Alameda    99.2 High
## 8  6001408500 Alameda    99.2 High
## 9  6001406202 Alameda    97.2 High
## 10 6001403000 Alameda    96.5 High
## # i 100 more rows
```

Summary county-level asthma correlation with population characteristic

variable education_pctl

```
asthma_measures_compare15 <- asthma_measures_compare8 %>%
  filter(education_pctl_cat %in% c("Very High", "High"))

unique(asthma_measures_compare15$county)

## [1] "Alameda"      "Contra Costa"  "Fresno"        "Kern"
## [5] "Los Angeles"   "Merced"        "San Bernardino" "San Joaquin"

# [1] "San Joaquin"    "Alameda"        "San Bernardino" "Los Angeles"
# "Fresno"        "Contra Costa"
# [7] "Merced"         "Kern"

asthma_measures_compare15_sum <- asthma_measures_compare15 %>%
  select(census_tract, county, asthma_pctl, education_pctl_cat) %>%
```

```

group_by(county) %>% ungroup()

print(asthma_measures_compare15_sum)

## # A tibble: 72 x 4
##   census_tract county      asthma_pctl education_pctl_cat
##   <dbl> <chr>          <dbl> <chr>
## 1  6001408800 Alameda      99.7 High
## 2  6001409500 Alameda      99.7 High
## 3  6001409400 Alameda      99.6 High
## 4  6001407200 Alameda      97.2 High
## 5  6013365002 Contra Costa  98.6 High
## 6  6013314200 Contra Costa  97.9 High
## 7  6019008502 Fresno       98.6 Very High
## 8  6019000700 Fresno       98.2 High
## 9  6019001000 Fresno       98.2 High
## 10 6019000902 Fresno       98.2 Very High
## # i 62 more rows

```

Summary county-level asthma correlation with population characteristic

variable unemployment_pctl

```

asthma_measures_compare16 <- asthma_measures_compare8 %>%
  filter(unemployment_pctl_cat %in% c("Very High", "High"))

unique(asthma_measures_compare16$county)

## [1] "Alameda"      "Contra Costa" "Fresno"      "Imperial"
## [5] "Kern"         "Lake"         "Los Angeles" "Merced"
## [9] "Modoc"        "Sacramento"   "San Bernardino" "San Francisco"
## [13] "San Joaquin"  "Solano"

# [1] "Solano"      "San Joaquin" "Alameda"     "San Bernardino"
# "Sacramento" "Contra Costa"
# [7] "Kern"        "Los Angeles" "Imperial"    "Merced"
# "Fresno"     "Lake"
# [13] "San Francisco" "Modoc"

asthma_measures_compare16_sum <- asthma_measures_compare16 %>%
  select(census_tract, county, asthma_pctl, unemployment_pctl_cat) %>%
  group_by(county) %>% ungroup()

print(asthma_measures_compare16_sum)

## # A tibble: 120 x 4
##   census_tract county      asthma_pctl unemployment_pctl_cat
##   <dbl> <chr>          <dbl> <chr>
## 1  6001410300 Alameda      99.9 High
## 2  6001408300 Alameda      99.6 High
## 3  6001410500 Alameda      99.0 Very High
## 4  6001402500 Alameda      99.0 Very High
## 5  6001407200 Alameda      97.2 High
## 6  6001407400 Alameda      97.2 High
## 7  6001407500 Alameda      95.6 High

```

```
## 8 6013307205 Contra Costa 99.5 High
## 9 6013305000 Contra Costa 99.4 Very High
## 10 6013307202 Contra Costa 99.4 Very High
## # i 110 more rows
```

Summary county-level asthma correlation with population characteristic

variable poverty_pctl

```
asthma_measures_compare17 <- asthma_measures_compare8 %>%
  filter(poverty_pctl_cat %in% c("Very High", "High"))

unique(asthma_measures_compare17$county)

## [1] "Alameda" "Contra Costa" "Fresno" "Imperial"
## [5] "Kern" "Lake" "Los Angeles" "Merced"
## [9] "Sacramento" "San Bernardino" "San Francisco" "San Joaquin"
## [13] "Solano"

# [1] "San Joaquin" "San Bernardino" "Solano" "Alameda"
# "Sacramento" "Kern"
# [7] "Contra Costa" "Los Angeles" "Fresno" "Merced"
# "Imperial" "Lake"
# [13] "San Francisco"

asthma_measures_compare17_sum <- asthma_measures_compare17 %>%
  select(census_tract, county, asthma_pctl, poverty_pctl_cat) %>%
  group_by(county) %>% ungroup()

print(asthma_measures_compare17_sum)
```

```
## # A tibble: 122 x 4
##   census_tract county asthma_pctl poverty_pctl_cat
##   <dbl> <chr> <dbl> <chr>
## 1 6001408800 Alameda 99.7 Very High
## 2 6001409500 Alameda 99.7 High
## 3 6001410500 Alameda 99.0 High
## 4 6001402500 Alameda 99.0 Very High
## 5 6001402600 Alameda 97.6 High
## 6 6001407500 Alameda 95.6 Very High
## 7 6013307202 Contra Costa 99.4 High
## 8 6019003302 Fresno 98.6 High
## 9 6019008502 Fresno 98.6 Very High
## 10 6019004704 Fresno 98.5 Very High
## # i 112 more rows
```

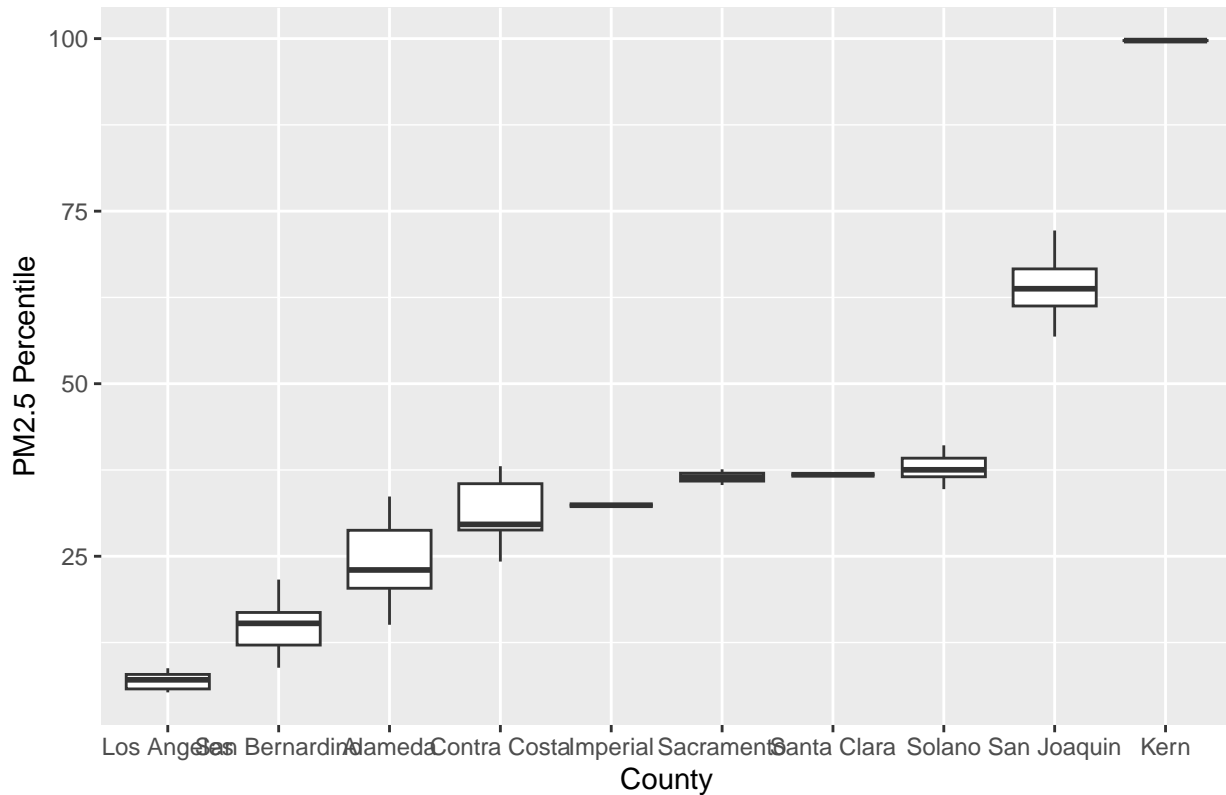
Top ten counties having highest density of asthma (99th percentile) include: Los Angeles, Imperial, Kern, Sacramento, Contra Costa, Alameda, San Joaquin, Santa Clara, Solano, San Bernadino.

Explore association between these counties and other potential Pollution and Population risk factors for asthma

```
# ggplot to explore relationship between counties with highest distribution of
# asthma and pm2.5_pctl
```

```
ggplot(asthma_measures_df9 %>%
  filter(county %in% c("Los Angeles", "Imperial", "Kern", "Sacramento", "Contra Costa", "Alameda",
    "San Joaquin", "Santa Clara", "Solano", "San Bernardino")),
  aes(x = fct_reorder(county, pm2.5_pctl), y = pm2.5_pctl)) +
geom_boxplot() +
labs(title = "California Counties with Highest Density of Asthma Association with PM2.5",
  x = "County",
  y = "PM2.5 Percentile") +
scale_y_continuous()
```

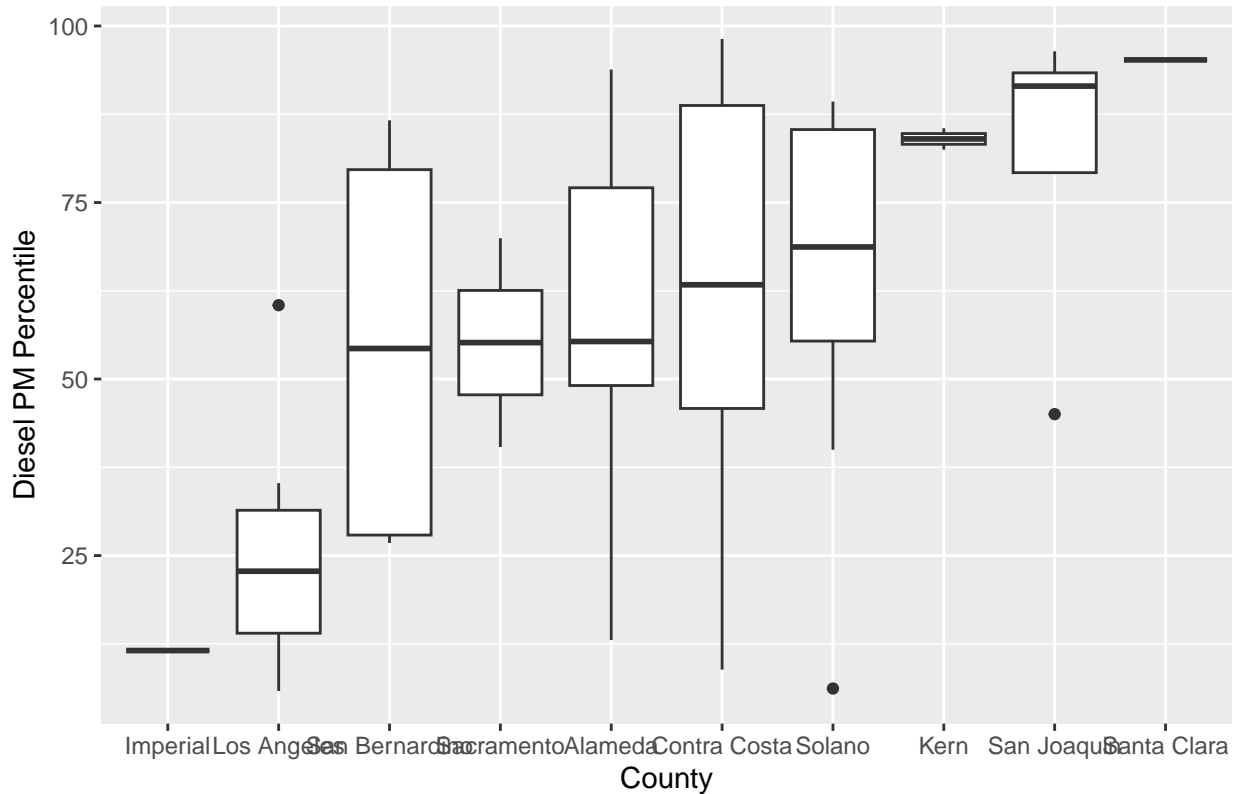
California Counties with Highest Density of Asthma Association with PM2.5



```
# ggplot to explore relationship between counties with highest distribution of
# and asthma and diesel_pm_pctl

ggplot(asthma_measures_df9 %>%
  filter(county %in% c("Los Angeles", "Imperial", "Kern", "Sacramento", "Contra Costa", "Alameda",
    "San Joaquin", "Santa Clara", "Solano", "San Bernardino")),
  aes(x = fct_reorder(county, diesel_pm_pctl), y = diesel_pm_pctl)) +
geom_boxplot() +
labs(title = "California Counties with Highest Density of Asthma Association with Diesel PM",
  x = "County",
  y = "Diesel PM Percentile") +
scale_y_continuous()
```

California Counties with Highest Density of Asthma Association with Diesel



```
# selecting variables of interest, (11 variables, 8035 obs.)
```

```
asthma_measures_df2a <- asthma_measures_df1 %>%
  select(census_tract, county, pm2.5, diesel_pm,
         traffic, tox_release, asthma, housing_burden,
         education, unemployment, poverty)
```

```
# Analysis of NA missing data from variables of interest before subsetting.
# Total NA = 1408 all occurring in population characteristics;
# none in pollution burden.
```

```
sum(is.na(asthma_measures_df2a$census_tract))
```

```
## [1] 0
```

```
# [1] 0
```

```
sum(is.na(asthma_measures_df2a$county))
```

```
## [1] 0
```

```
# [1] 0
```

```
sum(is.na(asthma_measures_df2a$pm2.5))
```

```
## [1] 0
```

```
# [1] 0
```

```

sum(is.na(asthma_measures_df2a$diesel_pm))

## [1] 0
# [1] 0

sum(is.na(asthma_measures_df2a$tox._release))

## [1] 0
# [1] 0

sum(is.na(asthma_measures_df2a$traffic))

## [1] 35
# [1] 35

sum(is.na(asthma_measures_df2a$asthma))

## [1] 11
# [1] 11

sum(is.na(asthma_measures_df2a$education))

## [1] 103
# [1] 103

sum(is.na(asthma_measures_df2a$unemployment))

## [1] 335
# [1] 335

sum(is.na(asthma_measures_df2a$housing_burden))

## [1] 145
# [1] 145

sum(is.na(asthma_measures_df2a$poverty))

## [1] 75
# [1] 75

```

County level summary data: median level of asthma per county

```

# Remove rows with NA values in columns
asthma_measures_df2a_clean <- asthma_measures_df2a[!is.na(asthma_measures_df2a$asthma) & !is.na(asthma_

asthma_measures_df3a <- asthma_measures_df2a_clean %>%
  group_by(county) %>%
  summarize(
    asthma_median_by_county = median(asthma, na.rm = TRUE),
    pm2.5_median_by_county = median(pm2.5, na.rm = TRUE),
    diesel_pm_median_by_county = median(diesel_pm, na.rm = TRUE),

```

```

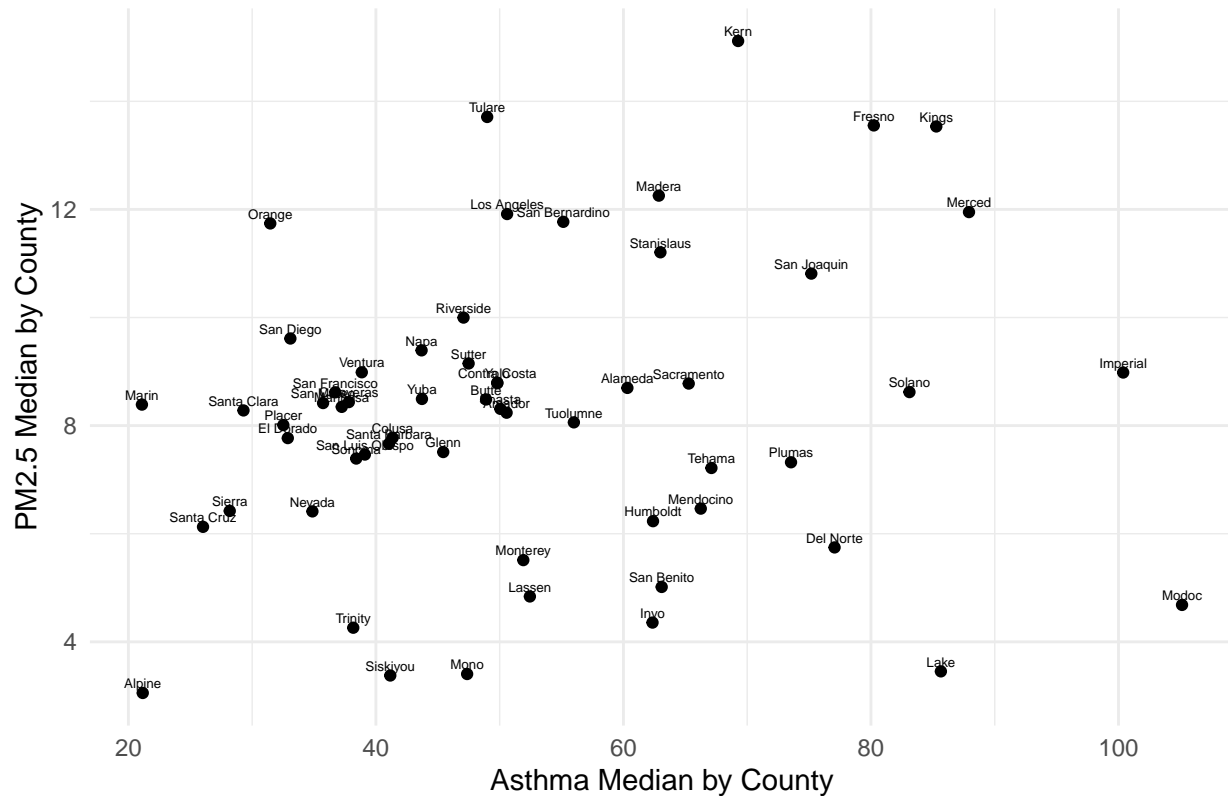
    traffic_median_by_county = median(traffic, na.rm = TRUE),
    tox._release_median_by_county = median(tox._release, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  ungroup()
asthma_measures_df3a <- asthma_measures_df3a %>%
  arrange(desc(asthma_median_by_county))

asthma_measures_df3b <- asthma_measures_df2a_clean %>%
  group_by(county) %>%
  summarize(
    asthma_median_by_county = median(asthma, na.rm = TRUE),
    housing_burden_median_by_county = median(housing_burden, na.rm = TRUE),
    education_median_by_county = median(education, na.rm = TRUE),
    unemployment_median_by_county = median(unemployment, na.rm = TRUE),
    poverty_median_by_county = median(poverty, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  ungroup()
asthma_measures_df3b <- asthma_measures_df3b %>%
  arrange(desc(asthma_median_by_county))

ggplot(asthma_measures_df3a, aes(x = asthma_median_by_county, y = pm2.5_median_by_county)) +
  geom_point() + # Scatter plot points
  geom_text(aes(label = county), hjust = 0.5, vjust = -0.5, size = 1.85) + # Labels each point with th
  labs(
    x = "Asthma Median by County",
    y = "PM2.5 Median by County",
    title = "Scatter Plot of Asthma vs PM2.5 by County"
  ) +
  theme_minimal() # Minimal theme for a clean look

```

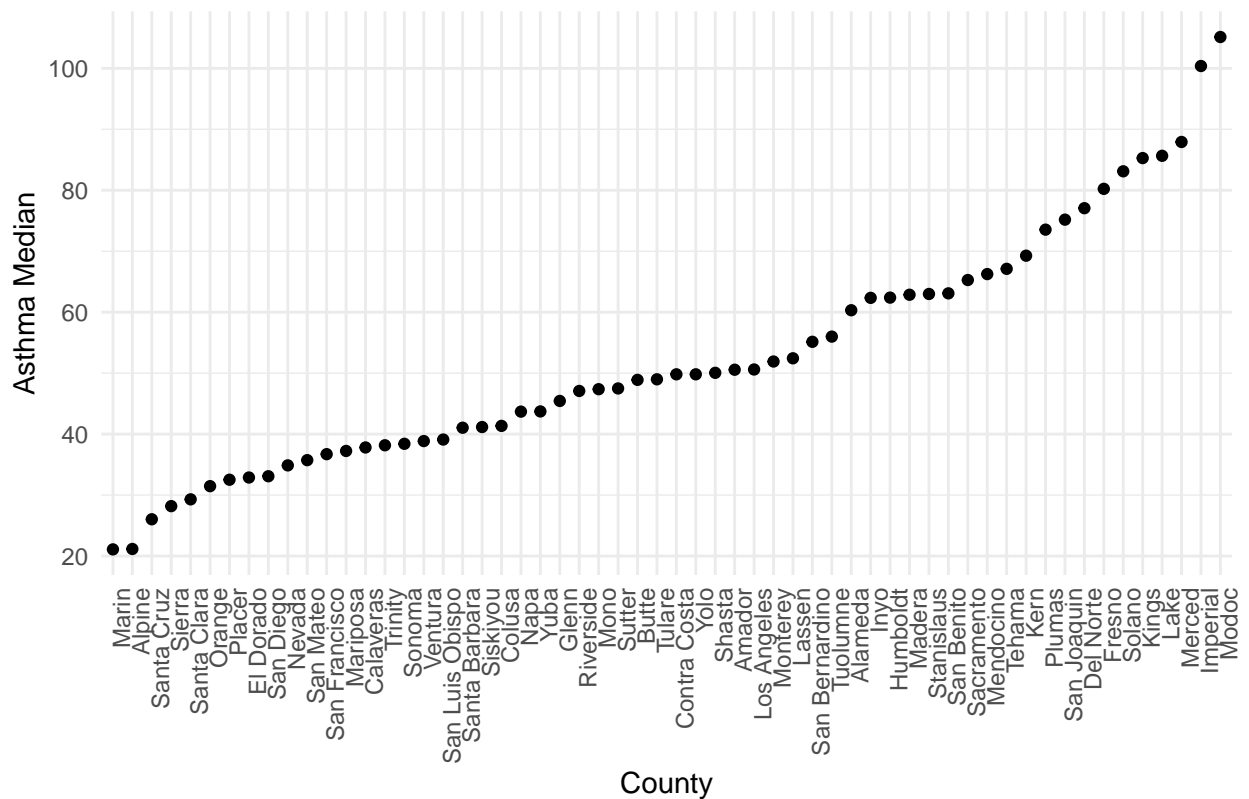
Scatter Plot of Asthma vs PM2.5 by County



```
asthma_measures_df13a <- asthma_measures_df3a %>%
  mutate(county = factor(county, levels = county[order(asthma_median_by_county)]))

# Plotting
ggplot(asthma_measures_df13a, aes(x = county, y = asthma_median_by_county)) +
  geom_point() + # Scatter plot points
  labs(
    x = "County",
    y = "Asthma Median",
    title = "Scatter Plot of Asthma Median per County"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1) # Rotate x-axis labels
  )
```


Scatter Plot of Asthma Median per County



It is important to consider impact of census tracts within counties

```
# Reduce observations from 8035 to 803
# by subsetting highest percentile (asthma >= 90th percentile)

asthma_measures_df5 <- asthma_measures_df2 %>%
  filter(asthma_pctl >= 90) %>%
  arrange(county, desc(asthma_pctl))

# 25 out of 58 counties are represented in asthma top 10 percentile (90 - 100)
# which means certain census tracts within multiple counties have the highest
# distribution of asthma cases. This needs to be considered for
# county level interventions.
```

```
unique(asthma_measures_df5$county)
```

```
## [1] "Alameda"      "Amador"      "Contra Costa" "Fresno"
## [5] "Humboldt"     "Imperial"    "Kern"         "Kings"
## [9] "Lake"         "Los Angeles" "Madera"       "Mendocino"
## [13] "Merced"       "Modoc"       "Monterey"     "Plumas"
## [17] "Riverside"    "Sacramento"  "San Bernardino" "San Diego"
## [21] "San Francisco" "San Joaquin" "Santa Clara"   "Solano"
## [25] "Stanislaus"
```

```
# [1] "Solano"      "San Joaquin" "Alameda"     "San Bernardino"
# "Santa Clara" "Sacramento"
# [7] "Contra Costa" "Kern"        "Los Angeles" "Imperial"
```

```

# "Fresno"          "Merced"
# [13] "Lake"            "San Francisco"  "Modoc"          "Kings"
# "Stanislaus"      "San Diego"
# [19] "Amador"          "Riverside"      "Humboldt"        "Mendocino"
# "Monterey"        "Plumas"
# [25] "Madera"

library(ggplot2)
library(dplyr)
library(ggrepel) # Optional for better label positioning

# Ensure counties are ordered by asthma median
asthma_measures_df3a <- asthma_measures_df3a %>%
  arrange(asthma_median_by_county) %>% # Order by asthma median
  mutate(county = factor(county, levels = county)) # Set factor levels by this order

theme(axis.text.x = element_blank(), # Remove x-axis text (counties)
      axis.ticks.x = element_blank()) # Remove x-axis ticks

## List of 2
## $ axis.text.x : list()
## .. attr(*, "class")= chr [1:2] "element_blank" "element"
## $ axis.ticks.x: list()
## .. attr(*, "class")= chr [1:2] "element_blank" "element"
## - attr(*, "class")= chr [1:2] "theme" "gg"
## - attr(*, "complete")= logi FALSE
## - attr(*, "validate")= logi TRUE

library(kableExtra)

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
## group_rows

library(dplyr)

# Selecting only numeric columns
correlations <- asthma_measures_df3a %>%
  select(asthma_median_by_county, pm2.5_median_by_county, diesel_pm_median_by_county,
         traffic_median_by_county, tox._release_median_by_county) %>%
  cor(use = "complete.obs") # Use complete cases to avoid NA issues

# Display correlations only involving `asthma_median_by_county`
correlations["asthma_median_by_county", ]

##      asthma_median_by_county      pm2.5_median_by_county
##      1.00000000      0.16820601
##      diesel_pm_median_by_county      traffic_median_by_county
##      -0.08258447      -0.19643727
##      tox._release_median_by_county
##      -0.07230597

```

Table 1: Correlation Matrix of Asthma and Pollution Factors by County

	Asthma	pm 2.5	Diesel	Traffic	Tox. Release
Asthma	1.000	0.168	-0.083	-0.196	-0.072
pm 2.5	0.168	1.000	0.423	0.438	0.435
Diesel	-0.083	0.423	1.000	0.682	0.349
Traffic	-0.196	0.438	0.682	1.000	0.434
Tox. Release	-0.072	0.435	0.349	0.434	1.000

Table 2: Correlation Matrix of Asthma and Socioeconomic Factors by County

	Asthma	Housing Burden	Education	Unemployment	Poverty
Asthma	1.000	0.168	-0.083	-0.196	-0.072
Housing Burden	0.168	1.000	0.423	0.438	0.435
Education	-0.083	0.423	1.000	0.682	0.349
Unemployment	-0.196	0.438	0.682	1.000	0.434
Poverty	-0.072	0.435	0.349	0.434	1.000

```

correlations <- round(correlations, 3)

# Rename columns and row names for readability
colnames(correlations) <- c("Asthma", "pm 2.5", "Diesel", "Traffic", "Tox. Release")
rownames(correlations) <- c("Asthma", "pm 2.5", "Diesel", "Traffic", "Tox. Release")

kable(correlations, caption = "Correlation Matrix of Asthma and Pollution Factors by County")

# Compute the correlation matrix
correlations1 <- asthma_measures_df3b %>%
  select(asthma_median_by_county, housing_burden_median_by_county, education_median_by_county,
         unemployment_median_by_county, poverty_median_by_county) %>%
  cor(use = "complete.obs") # Use complete cases to avoid NA issues

# Display correlations only involving `asthma_median_by_county`
correlations1["asthma_median_by_county", ]

##          asthma_median_by_county housing_burden_median_by_county
##                1.000000000                -0.04073784
##          education_median_by_county unemployment_median_by_county
##                0.53216114                0.43151174
##          poverty_median_by_county
##                0.55769216

correlations1 <- round(correlations, 3)

# Rename columns and row names for readability
colnames(correlations1) <- c("Asthma", "Housing Burden", "Education", "Unemployment", "Poverty")
rownames(correlations1) <- c("Asthma", "Housing Burden", "Education", "Unemployment", "Poverty")

# Display the correlation matrix as a kable table
kable(correlations1, caption = "Correlation Matrix of Asthma and Socioeconomic Factors by County")

```