Insert title of project here

Web address for GitHub repository

Name

Contents

List of Tables

List of Figures

1 Rationale and Research Questions

In the United States, the National Ambient Air Quality Standards from the 1970 Clean Air Act help monitor air pollutant levels. While significant reductions in air pollution have been made since the standards were put in place, there are still many areas that do not meet the standards. California, in particular, has been cited as a leader in air pollution, cities such as Los Angeles-Long Beach, Bakersfield and Fresno-Madera having the highest recorded levels of ozone levels in the country.

It is well established that exposure to higher levels of air pollutants such as ozone (O3), nitrogen dioxide (NO2), and particulate matter (PM2.5) is associated with reduced lung function, asthma exacerbations, increased hospital visits, and death (Schraufnagel et al., 2019). In fact, asthma is the leading chronic condition in children, affecting 1 in 12 children in the United States (Zahran, 2018). According to a recent Center for Disease Control and Prevention (CDC) study, children have higher rates of hospital and emergency department visits associated with asthma compared to adults (Moorman et al., 2012). The health impacts of asthma are not distributed equally among children, however. Prevalence of asthma in children can differ by age, family history, racial and ethnic group, and socioeconomic status (U.S. EPA, 2013).

As such, my study seeks to answer two main questions: * What are the trends of O3, NO2 and PM2.5 in some of the most polluted cities in California over a five-year period, from 2013 to 2017 ? * How is asthma incidence in California impacted by O3 and PM2.5 levels, accounting for socioeocnomic variables? Is this relationship different in children compared to adults?

2 Dataset Information

Provide information on how the dataset for this analysis were collected, the data contained in the dataset, and any important pieces of information that are relevant to your analyses. This section should contain much of same information as the metadata file for the dataset but formatted in a way that is more narrative.

Describe how you wrangled your dataset in a format similar to a methods section of a journal article.

Add a table that summarizes your data structure (variables, units, ranges and/or central tendencies, data source if multiple are used, etc.). This table can be made in markdown text or inserted as a `kable` function in an R chunk. If the latter, do not include the code used to generate your table.

2.1 EPA air quality datasets

Air quality data were collected using EPA's Download Daily Data tool (https://www.epa.gov/outdoor-air-quality-data/download-daily-data).

The following selections were made:

Option	Selection
Pollutant	PM2.5, Ozone, and NO2
Year	2013-2017
Geographic Area	California
Download	Download CSV (spreadsheet)

The downloaded files, which were accessed on 2020-04-11, were saved in the project folder path `./Data/Raw/Air quality/` as `EPAair_[Pollutant]_CA_[Year]_raw.csv`. For example, the O3 dataset for 2017 was saved as `EPAair_O3_CA_2017_raw.csv`.

2.1.1 PM2.5 dataset content information

This datset contains daily mean PM2.5 concentrations and the corresponding air quality index (AQI) value in California over the years 2013-2017.

Year	Sites	Counties
2013	149	52
2014	152	52
2015	151	52
2016	151	52
2017	152	51

2.1.2 O3 dataset content information

This dataset contains daily maximum 8-hour O3 concentrations and the corresponding air quality index (AQI) value in California over the years 2013-2017.

Year	Sites	Counties
2013	183	49
2014	182	49
2015	182	49
2016	182	49
2017	181	49

2.1.3 NO2 dataset content information

This dataset contains the daily maximum 1-hour NO2 concentration and the corresponding air quality index (AQI) value in California over the years 2013-2017.

Year	Sites	Counties
2013	102	33
2014	105	33
2015	108	33
2016	109	33
2017	106	33

All three datasets contain 20 variables (Table 2). Variable names without descriptions are self-explanatory.

Variable	Description
Date	Month/day/year
Source	AQS (Air Quality System) or AirNow
Site ID	A unique number within the county identifying the site.
POC	"Parameter Occurrence Code" used to distinguish different instruments that measure the same parameter at the same site.
Daily Mean PM2.5 Concentration	
Daily Max 8-hour Ozone Concentration	
Daily Max 1-hour NO2 Concentration	

Variable	Description
Units	Units for concentration
Daily_AQI_VALUE	Air quality index (range 0-500)
Site Name	
DAILY_OBS_COUNT	Number of observations per day
PERCENT_COMPLETE	
AQS_PARAMETER_CODE	
AQS_PARAMETER_DESC	
CBSA_CODE	
CBSA_NAME	
STATE_CODE	
STATE	
COUNTY_CODE	
COUNTY	
SITE_LATITUDE	
SITE_LONGITUDE	

2.1.4 Data wrangling

Air quality datasets for different years were combined using `rbind` to form one dataset for each pollutant.

The following columns were selected: * Date * Site.ID * Daily.Max.1.hour.NO2.Concentration * DAILY_AQI_VALUE * Site.Name * COUNTY

Additionally, columns for `Month` and `Year` were added using the `Date` column.

2.2 Asthma datasets

Asthma data were collected using Tracking California's Asthma Data Query tool (https://trackingcalifornia.org/asthma/query).

The following selections were made:

Option	Selection
Type of Event	Emergency department visits due to asthma
Age sub-group	Age 0-17, Age 18 & over
Year	2013 - 2017
How event is measured	Age-adjusted rates per 10,000
Race/ethnicity	All Races/Ethnicities
Gender/sex	Both Sexes
Type of information	Conventional
Type of geography	Zip codes

The downloaded files, which were accessed on 2020-04-12, were saved in the project folder path `./Data/Raw/Asthma/` as `TrackingCA_Asthma_ERVisits_[Age Group]_[Year]_raw.csv`. For example, the dataset for adults in 2013 was saved as `TrackingCA_Asthma_ERVisits_Adults_2013_raw.`

2.2.1 Adult asthma dataset content information

This dataset contains the annual rates of asthma-related ER visits for adults in California over the years 2013-2017.

2.2.2 Children asthma dataset content information

This dataset contains the annual rates of asthma-related ER visits for children in California over the years 2013-2017.

Both datasets contain 2 variables. Variable names without descriptions are self-explanatory.

Variable	Description
Zip code	
Incidence	Age-adjusted rate of emergency department (ER) visits due to asthma per 10,000 California residents.

2.2.3 Data wrangling

A 'Year' column was added to all asthma datasets. Datasets for 2013-2017 were combined using `rbind` to form one dataset for each age group (Adults, Children).

Since the asthma datasets provide only zip code information but not county information, an online search was done to find zip codes and their corresponding counties in California. This information was datascraped into a data frame. This dataframe was then combined with the adult dataset and the children dataset using `left_join`. The incidence rates were then grouped by county to calculate an average incidence rate for each county.

Finally, adult and children asthma datasets were combined using `full_join`.

2.3 Demographics dataset

Demographic data were collected from County Health Rankings & Roadmaps (https://www.countyhealthrankings.org/app/california/2019/downloads).

Since demographics are assumed to remain somewhat constant over the five-year period, only one dataset was chosen. The 2019 dataset, which uses data published in 2017, was chosen.

The xls file, was accessed on 2020-04-12, was saved in the project folder path `./Data/Raw/Demographics/` as `CountyHealthRankings_CA_2019_raw.xls`. Since there multiple tabs in the xls file, relevant information from the file was taken and converted into a csv file.The csv file was saved as `CountyHealthRankings_CA_2019_filtered_raw.csv`

The following information is contained in the dataset: * FIPS * State * County * Median Household Income * Population * % African American * % Rural * % Smokers * % Uninsured

3 Exploratory Analysis

Insert exploratory visualizations of your dataset. This may include, but is not limited to, graphs illustrating the distributions of variables of interest and/or maps of the spatial context of your dataset. Format your R chunks so that graphs are displayed but code is not displayed. Accompany these graphs with text sections that describe the visualizations and provide context for further analyses.

Each figure should be accompanied by a caption, and each figure should be referenced within the text.

Scope: think about what information someone might want to know about the dataset before analyzing it statistically. How might you visualize this information?

3.1 Air quality datasets

In examining the air quality index (AQI) values for the three pollutants across all sites, there is not much temporal variation (Figure 1). However, the frequency distrbution does differ depending on the pollutant. NO2 generally has lower AQI values, which are aggregate well below 50, the cut-off number for the "Good" air quality range. O3 and PM2.5 have higher AQI values, which concentrated closer to 50, and also contain more values that are above 50.

Among the four target sites–Bakersfield, Fresno, Los Angeles and San Francisco–Bakersfield appears to have higher AQI values than the other sites across all pollutants. In contrast, San Francisco generally has lower AQI values compared to the other sites. Although AQI values for NO2 stay below the "Unhealthy" range for all sites, the AQI values for O3 and PM2.5 are generally in this range.

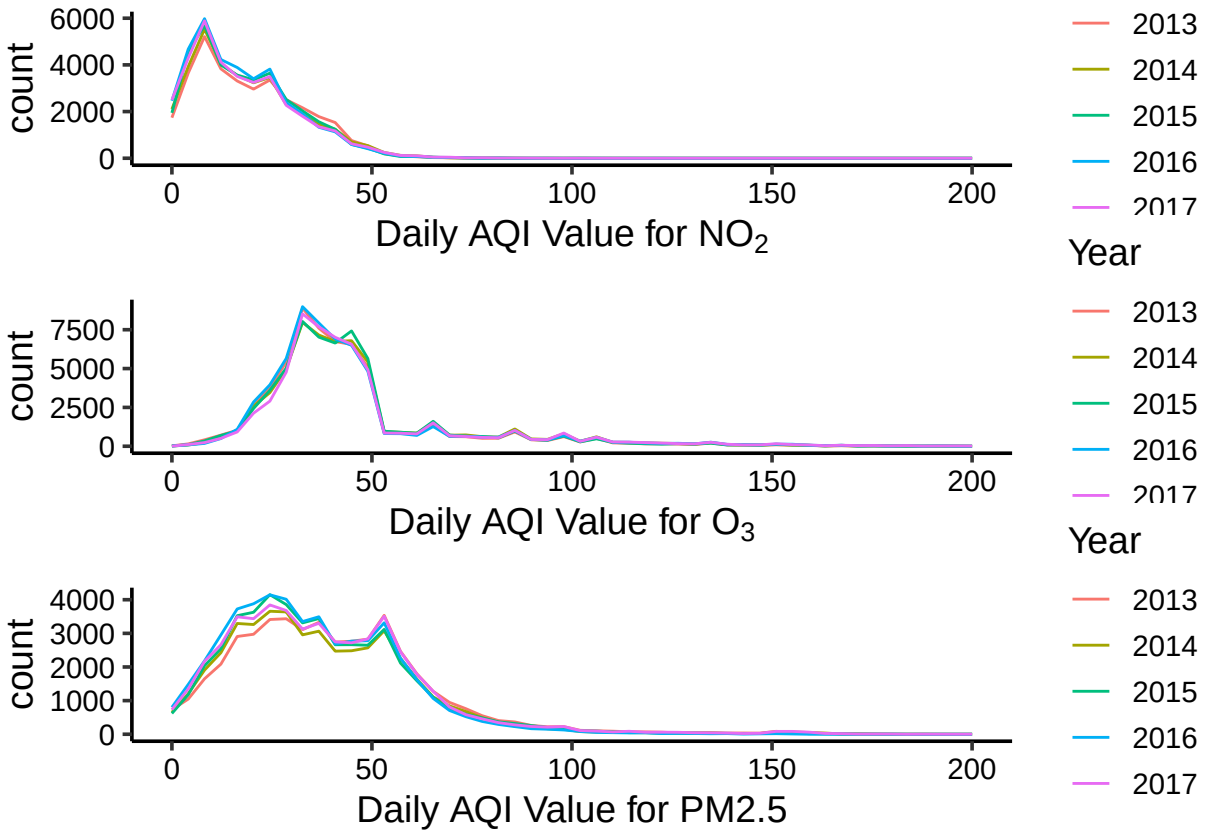

Figure 1: Frequency plots of daily air quality index (AQI) values for the three pollutants across the years 2013-2017.

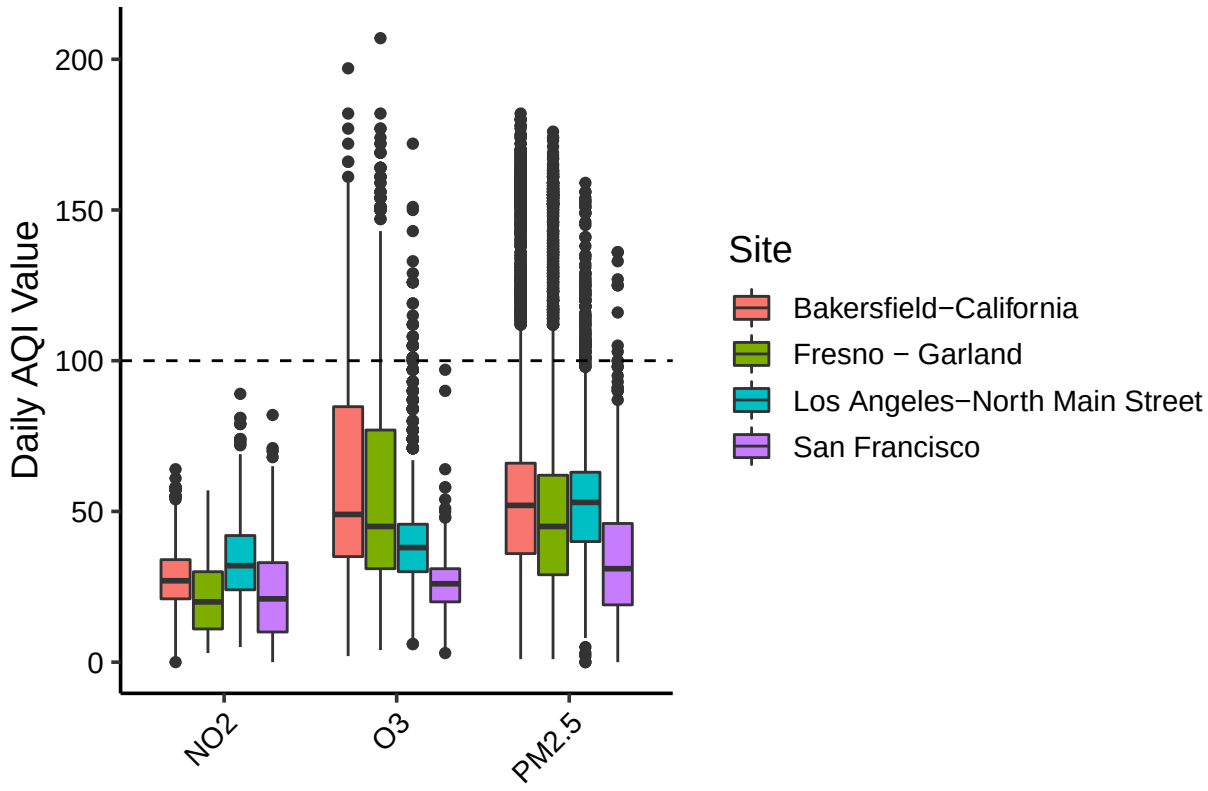

Figure 2: Boxplots of daily air quality index (AQI) values for the three pollutants in four key sites. The dashed line represents the lower bound for the `Unhealthy (For sensitive groups)` range.

4 Analysis

Insert visualizations and text describing your main analyses. Format your R chunks so that graphs are displayed but code and other output is not displayed. Instead, describe the results of any statistical tests in the main text (e.g., "Variable x was significantly different among y groups (ANOVA; df = 300, F = 5.55, p < 0.0001)"). Each paragraph, accompanied by one or more visualizations, should describe the major findings and how they relate to the question and hypotheses. Divide this section into subsections, one for each research question.

Each figure should be accompanied by a caption, and each figure should be referenced within the text

4.1 Question 1: <insert specific question here and add additional subsections for additional questions below, if needed>

4.2 Question 2:

5 Summary and Conclusions

Summarize your major findings from your analyses in a few paragraphs. What conclusions do you draw from your findings? Relate your findings back to the original research questions and rationale.

6 References

<add references here if relevant, otherwise delete this section>