Jadr Health Insights: Capstone Project

Introductions

Justin Bartell

Bachelor's degree in Chemical-Biomedical Engineering (May 2019)

Abbey Guilliat

Bachelor's degree in Astronomy & Astrophysics (May 2021)
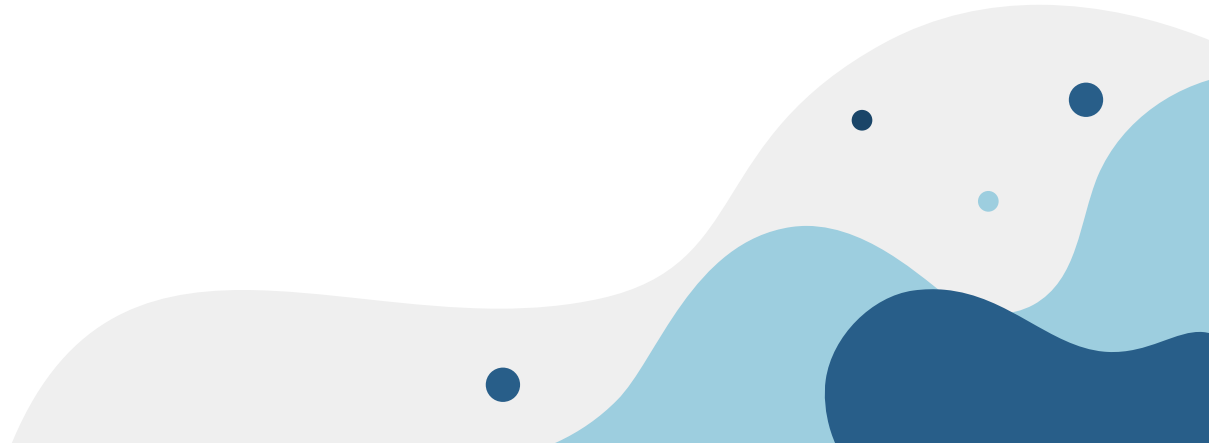
Darrell Gerber

Bachelor's and Master's degrees in Aerospace Engineering (June 1993)

Regina Huber

Bachelor's degree in Mathematical Sciences (May 2020)

Project Purpose

Provide local asthma risk assessments for health professionals and individuals with asthma via a Power BI dashboard

Initial Questions

1. What are the relationships between:

   - **Environmental air quality factors**
   - **Local industry activity**
   - **Asthma-induced emergency department visits**

2. What machine learning model can **predict asthma incidence rates** for a given county based on local air quality metrics?
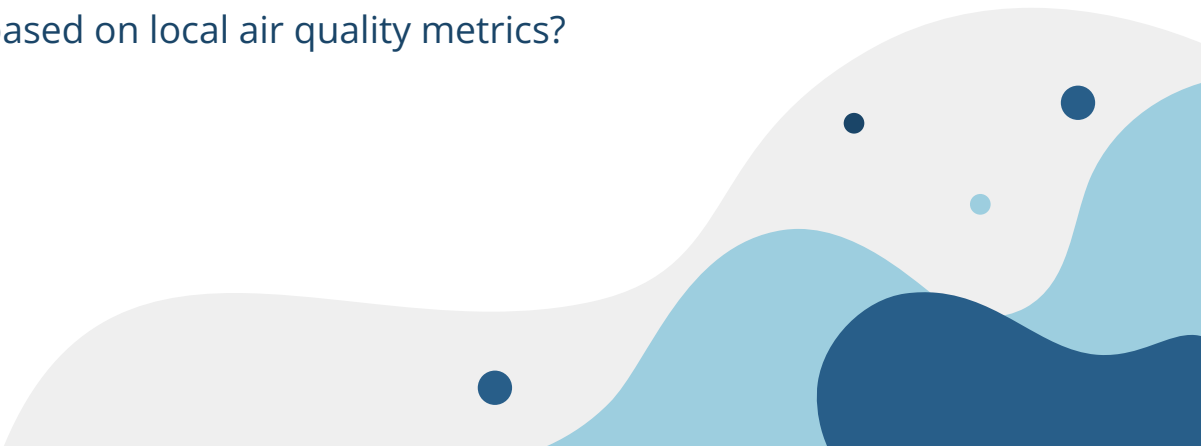
Research Process

Datasets Chosen

1

California Department of Public Health's Asthma Emergency Department Visit Rates

2015-2019 Annual Summary

2

US Environmental Protection Agency's National Air Quality Data API
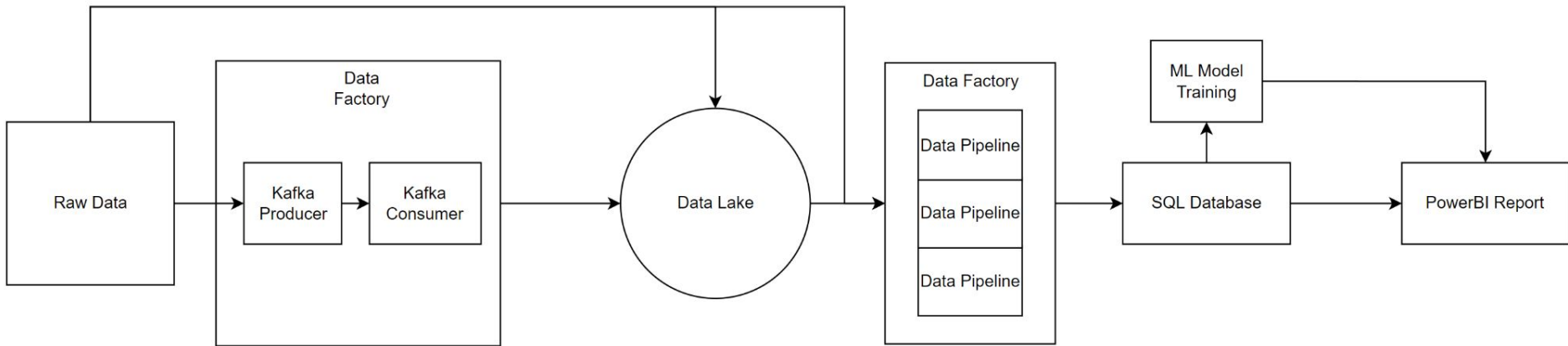
2015-2019 and 2021 Annual Summary

3
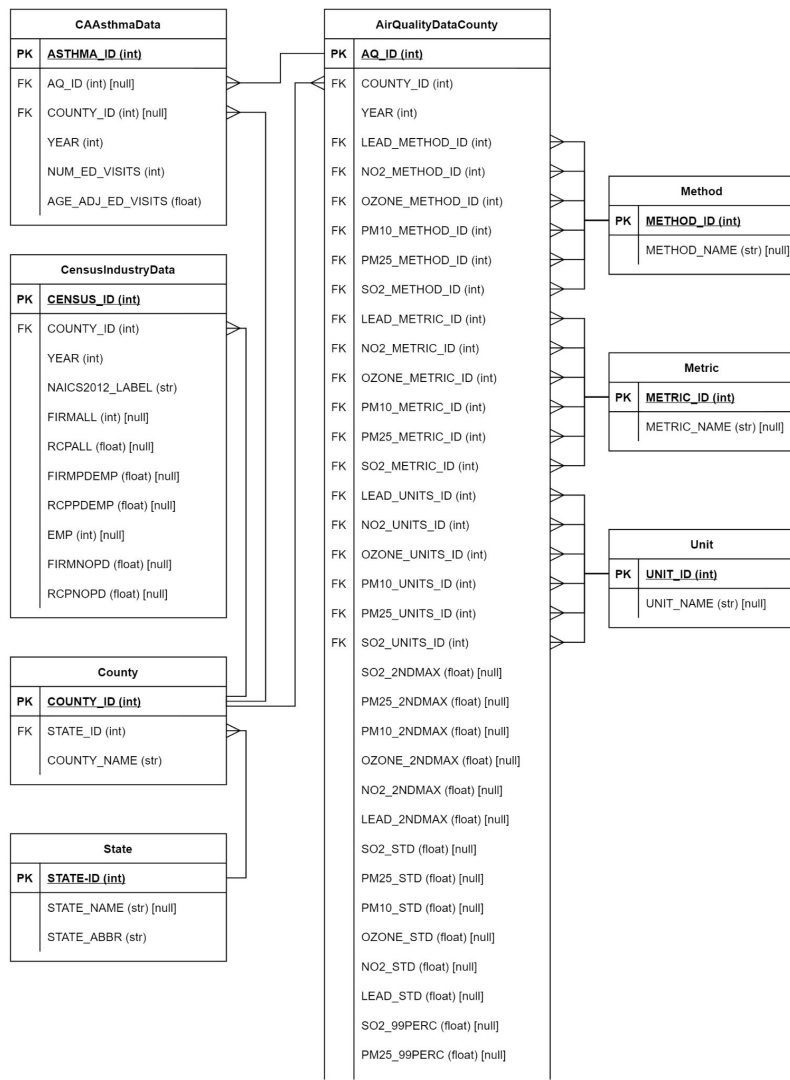
US Census Survey of Business Owners Industry Dataset
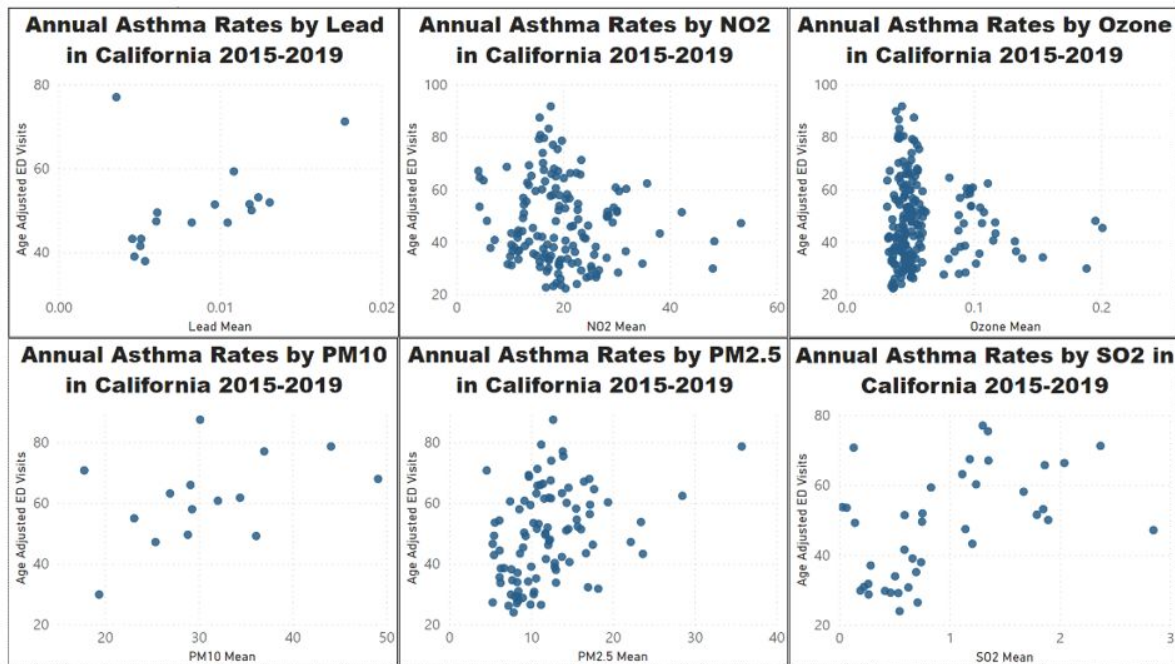
2012 Annual Summary

Data Platform

SQL Entity Relationship Diagram

CAAsthmaData

PK	ASTHMA_ID (int)
FK	AQ_ID (int) [null]
FK	COUNTY_ID (int) [null]
	YEAR (int)
	NUM_ED_VISITS (int)
	AGE_ADJ_ED_VISITS (float)

CensusIndustryData

PK	CENSUS_ID (int)
FK	COUNTY_ID (int)
	YEAR (int)
	NAICS2012_LABEL (str)
	FIRMALL (int) [null]
	RCPALL (float) [null]
	FIRMPDEMP (float) [null]
	RCPPDEMP (float) [null]
	EMP (int) [null]
	FIRMNOPD (float) [null]
	RCPNOPD (float) [null]

County

PK	COUNTY_ID (int)
FK	STATE_ID (int)
	COUNTY_NAME (str)

State

PK	STATE-ID (int)
	STATE_NAME (str) [null]
	STATE_ABBR (str)

AirQualityDataCounty

PK	AQ_ID (int)
FK	COUNTY_ID (int)
	YEAR (int)
FK	LEAD_METHOD_ID (int)
FK	NO2_METHOD_ID (int)
FK	OZONE_METHOD_ID (int)
FK	PM10_METHOD_ID (int)
FK	PM25_METHOD_ID (int)
FK	SO2_METHOD_ID (int)
FK	LEAD_METRIC_ID (int)
FK	NO2_METRIC_ID (int)
FK	OZONE_METRIC_ID (int)
FK	PM10_METRIC_ID (int)
FK	PM25_METRIC_ID (int)
FK	SO2_METRIC_ID (int)
FK	LEAD_UNITS_ID (int)
FK	NO2_UNITS_ID (int)
FK	OZONE_UNITS_ID (int)
FK	PM10_UNITS_ID (int)
FK	PM25_UNITS_ID (int)
FK	SO2_UNITS_ID (int)
	SO2_2NDMAX (float) [null]
	PM25_2NDMAX (float) [null]
	PM10_2NDMAX (float) [null]
	OZONE_2NDMAX (float) [null]
	NO2_2NDMAX (float) [null]
	LEAD_2NDMAX (float) [null]
	SO2_STD (float) [null]
	PM25_STD (float) [null]
	PM10_STD (float) [null]
	OZONE_STD (float) [null]
	NO2_STD (float) [null]
	LEAD_STD (float) [null]
	SO2_99PERC (float) [null]
	PM25_99PERC (float) [null]

Method

PK	METHOD_ID (int)
	METHOD_NAME (str) [null]

Metric

PK	METRIC_ID (int)
	METRIC_NAME (str) [null]

Unit

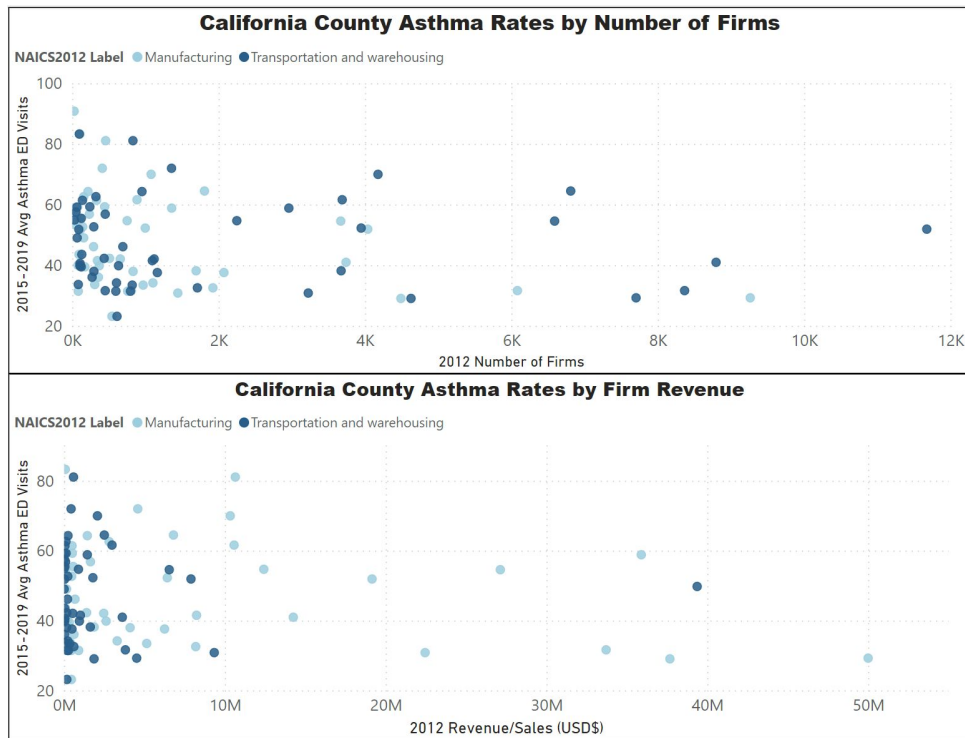|	
PK	UNIT_ID (int)
	UNIT_NAME (str) [null]

Exploratory Analysis:
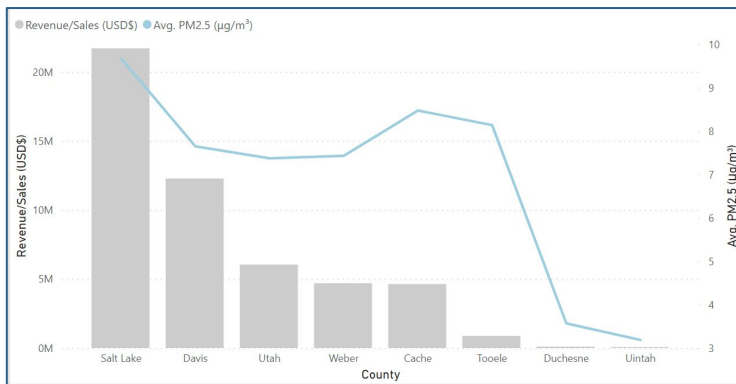# Asthma Rates and Air Quality

Exploratory Analysis:
Asthma Rates and Census Industries

Exploratory Analysis:
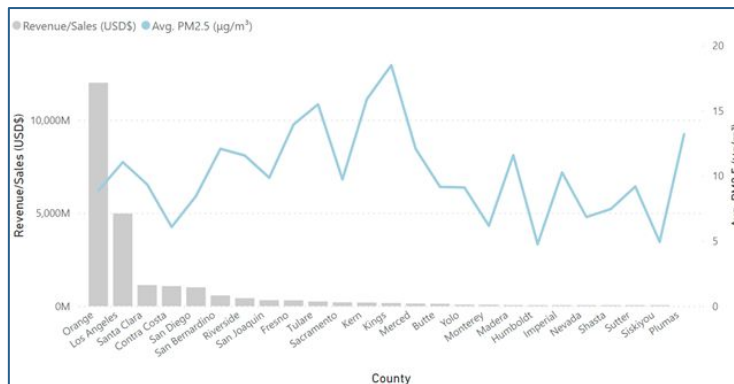# Air Quality and Census Industries

Manufacturing Industry Revenue vs. Average PM2.5 Measure
for US Counties

Expectation

Utah

Reality

California

Machine Learning

Models Developed

Linear Regression

Lasso

Random Forest

AdaBoost

Support Vector Regression

Voting Regression

Final Machine Learning Model

Predictors - EPA air quality measurements averaged annually
California counties 2015 to 2019
(mean, 1st max, 2nd max, 99 percentile, standard deviation)

- Ozone
- Very small particulate matter (PM2.5)
- SO_2

Outcome - Age-adjust asthma emergency department visits averaged annually
California counties 2015 to 2019

Imputing- Drop missing values

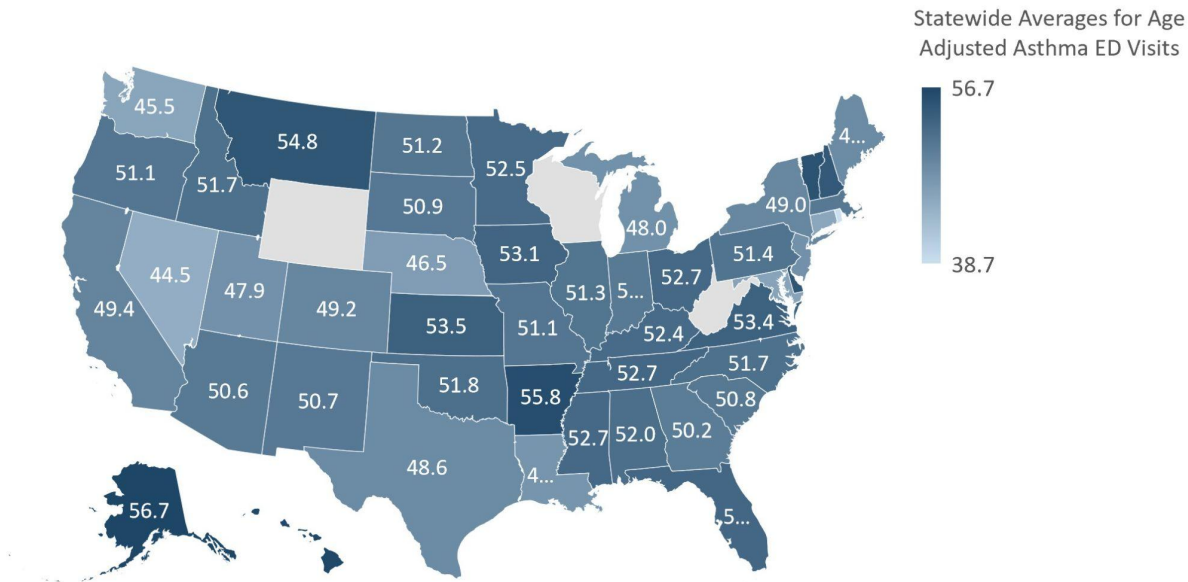**Model** - Voting Regression using AdaBoost and Support Vector Regression

Model Accuracy Results
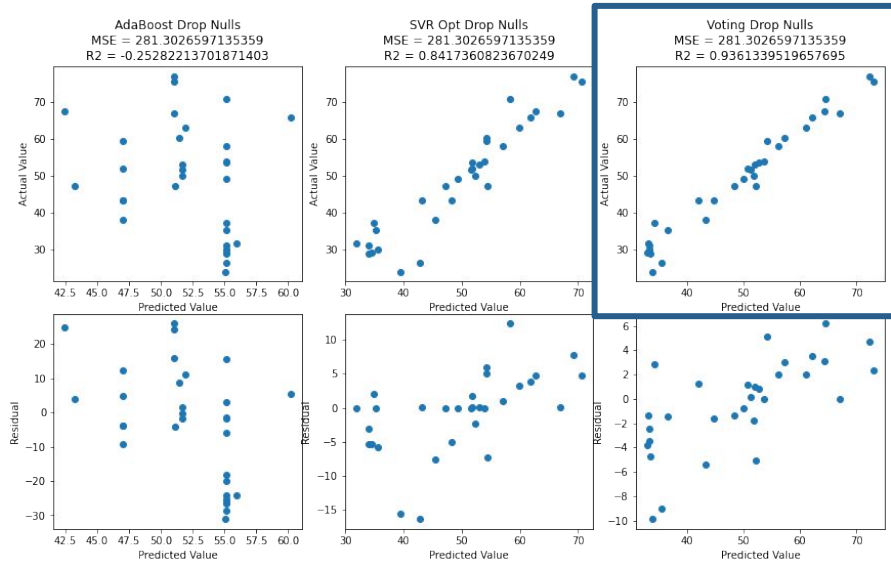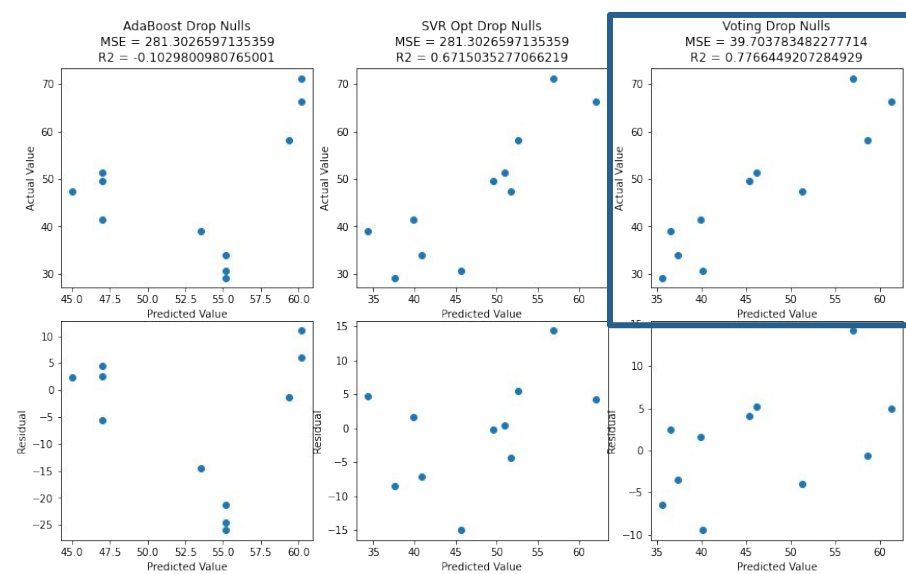
Model Predictions

ML Model Predicts Age-Adjusted Emergency Department Visits for US Counties

Highlights and Predictions

→

State, County ⌄

- ⌄ ○ Alabama
- ⌄ ○ Alaska
- ⌄ ○ Arizona
- ⌄ ○ Arkansas
- ⌃ ◉ California
  - ○ Alameda
  - ○ Alpine
  - ○ Amador
  - ○ Butte
  - ○ Calaveras
  - ○ Colusa
  - ○ Contra Costa
  - ○ Del Norte
  - ○ El Dorado
  - ◉ Fresno
  - ○ Glenn
  - ○ Humboldt
  - ○ Imperial
  - ○ Inyo
  - ○ Kern
  - ○ Kings
  - ○ Lake
  - ○ Lassen
  - ○ Los Angeles
  - ○ Madera
  - ○ Marin
  - ○ Mariposa
  - ○ Mendocino
  - ○ Merced
  - ○ Modoc

US Average Annual Prediction of Age-Adjusted Asthma Emergency Department Visits

63.473

% Age-Adjusted * 10,000

Relative Risk Assessment:

High

County Map

Top Industries by County (2012)

Number of Firms	Industry
9208	Other services (except public administration)
7915	Health care and social assistance
6546	Professional, scientific, and technical services
6315	Retail trade
5339	Administrative and support and waste management and remediation services
4924	Real estate and rental and leasing
4609	Construction
4177	Transportation and warehousing
2045	Finance and insurance
1905	Accommodation and food services
1745	Arts, entertainment, and recreation
1571	Wholesale trade
1077	Manufacturing
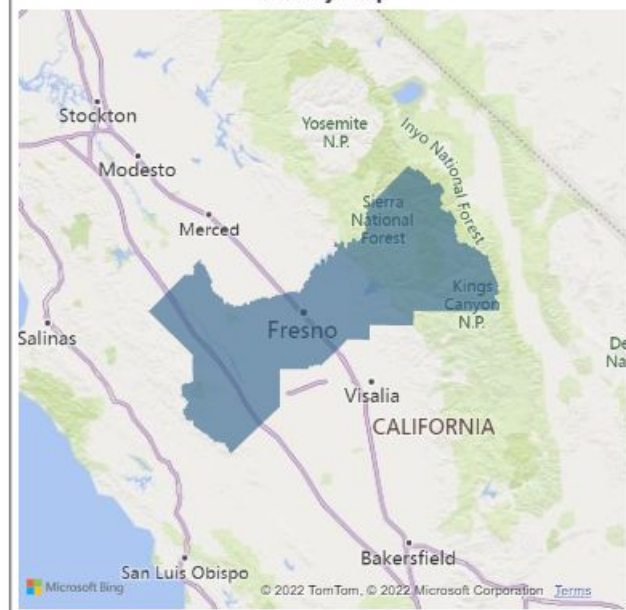
Air Quality Annual Averages in 2021

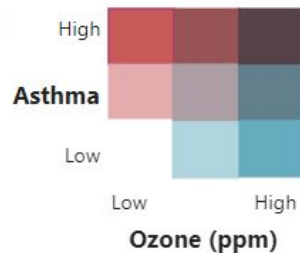| Year	Lead	NO2	Ozone	PM10	PM25	SO2
2021		1,583.34	4.72		1,628.03	29.44
Total		**1,583.34**	**4.72**		**1,628.03**	**29.44**

Asthma Rates and Air Quality in California

On this page, "asthma rate" refers to the age-adjusted number of asthma-related emergency department visits per 10,000 people.
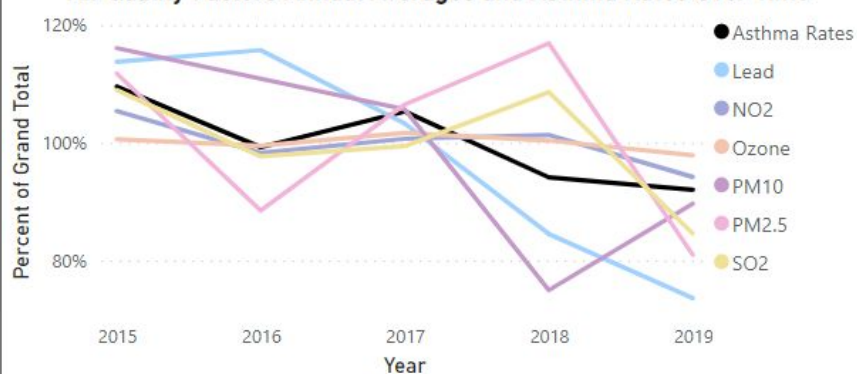

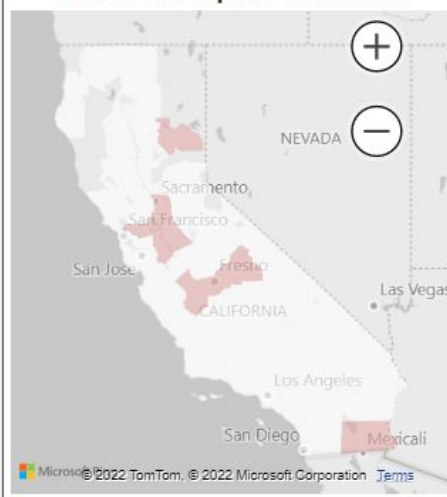
PM2.5 Annual Averages and Asthma Rates in California

Year
2015
2019

Air Quality Factors Annual Averages and Asthma Rates Over Time

- Asthma Rates
- Lead
- NO2
- Ozone
- PM10
- PM2.5
- SO2

Asthma (High / Low)
PM2.5 (ppm) (Low / High)
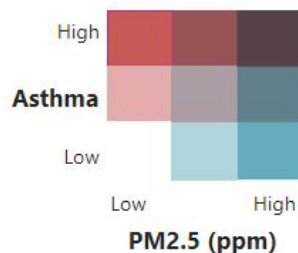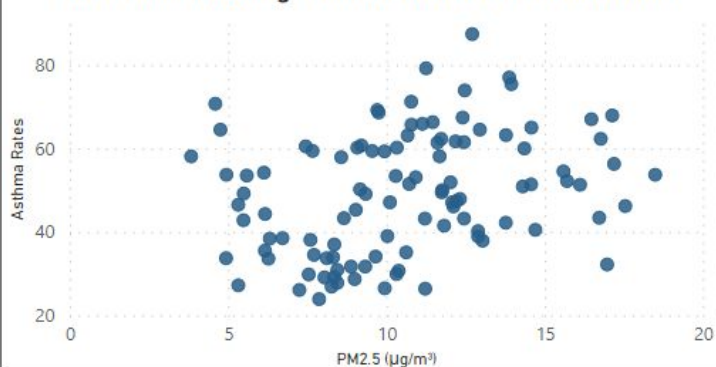
Select Year:
- 2015
- 2016
- 2017
- 2018
- ● 2019

Asthma (High / Low)
Ozone (ppm) (Low / High)

Asthma Incidence vs PM2.5 Bivariate Map for California

Asthma Incidence vs Ozone Bivariate Map for California

© 2022 TomTom, © 2022 Microsoft Corporation Terms
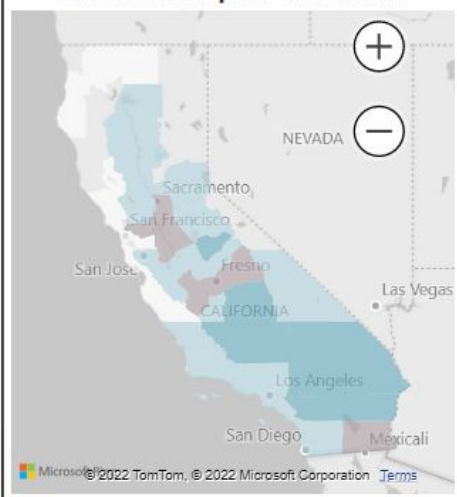
Total Revenue for All Manufacturing Firms:

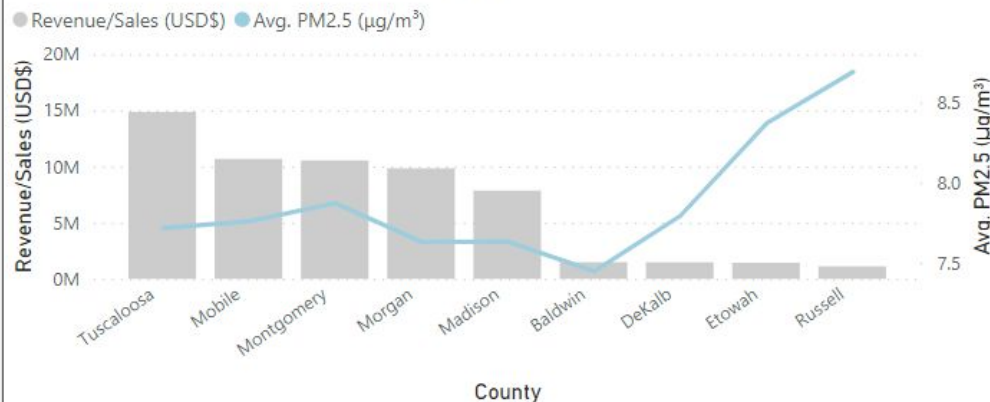# 118.4M

(USD$)

Filter by State:

- ☑ Alabama
- ☐ Arizona
- ☐ California
- ☐ Colorado
- ☐ Connecticut
- ☐ Florida
- ☐ Georgia
- ☐ Idaho
- ☐ Illinois
- ☐ Indiana
- ☐ Iowa
- ☐ Kentucky
- ☐ Louisiana
- ☐ Massachusetts
- ☐ Michigan
- ☐ Minnesota
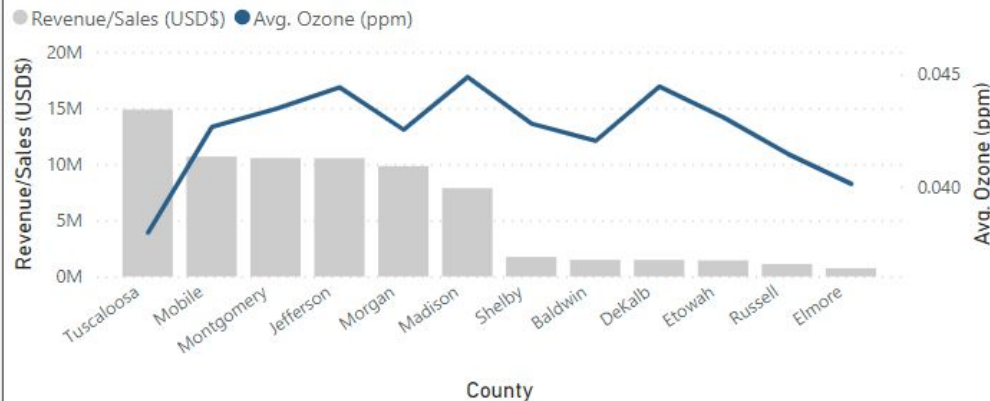- ☐ Nebraska
- ☐ New Jersey
- ☐ New York

Manufacturing Industry Revenue vs. Average PM2.5 Measure for US Counties

● Revenue/Sales (USD$) ● Avg. PM2.5 ($\mu g/m^3$)

Y-axis left: Revenue/Sales (USD$) — 0M, 5M, 10M, 15M, 20M
Y-axis right: Avg. PM2.5 ($\mu g/m^3$) — 7.5, 8.0, 8.5

X-axis (County): Tuscaloosa, Mobile, Montgomery, Morgan, Madison, Baldwin, DeKalb, Etowah, Russell

Manufacturing Industry Revenue vs. Average Ozone Measure for US Counties

● Revenue/Sales (USD$) ● Avg. Ozone (ppm)

Y-axis left: Revenue/Sales (USD$) — 0M, 5M, 10M, 15M, 20M
Y-axis right: Avg. Ozone (ppm) — 0.040, 0.045

X-axis (County): Tuscaloosa, Mobile, Montgomery, Jefferson, Morgan, Madison, Shelby, Baldwin, DeKalb, Etowah, Russell, Elmore
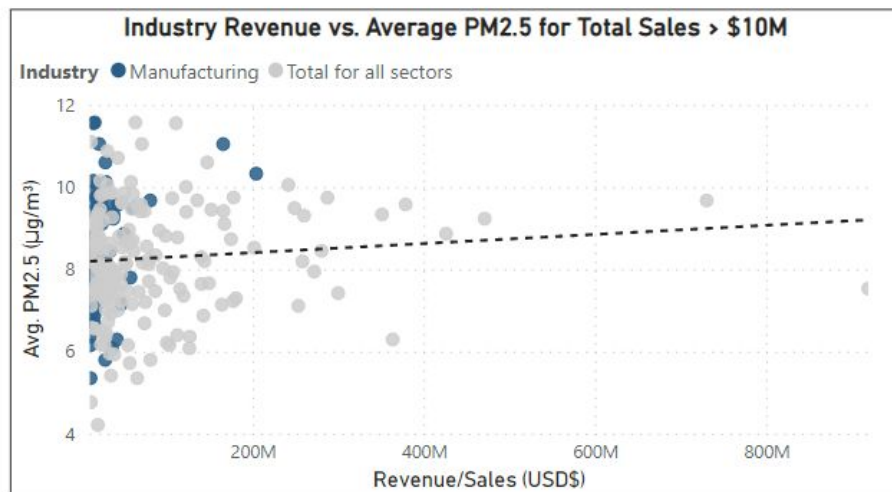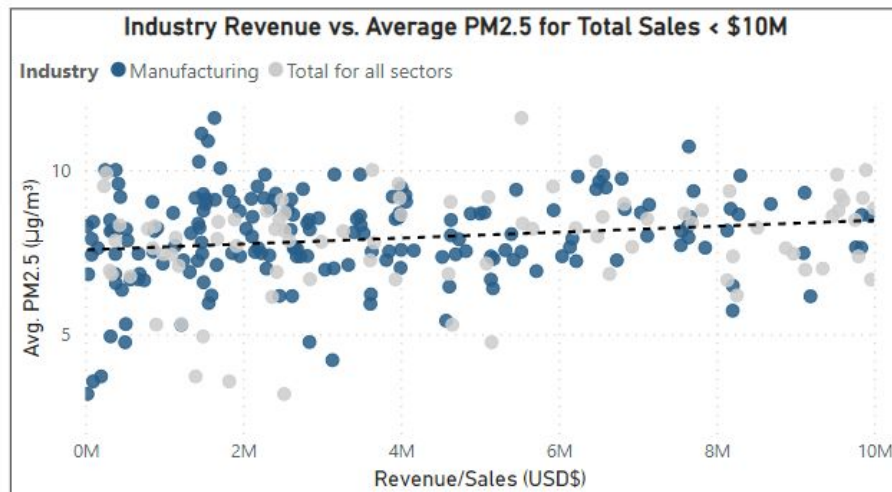
7.80

Average PM25 ($\mu g/m^3$)

0.042

Average Ozone (ppm)

Comparing Industries

Select Industry: ⌄
- ☐ Construction
- ☐ Finance and insurance
- ☐ Health care and social assistance
- ☐ Information
- ☑ Manufacturing
- ☐ Professional, scientific, and technical services
- ☐ Retail trade
- ☑ Total for all sectors
- ☐ Transportation and warehousing
- ☐ Wholesale trade

Industry Revenue vs. Average PM2.5 for Total Sales < $10M

Industry ● Manufacturing ● Total for all sectors

Avg. PM2.5 (µg/m³) vs. Revenue/Sales (USD$)

Industry Revenue vs. Average PM2.5 for Total Sales > $10M

Industry ● Manufacturing ● Total for all sectors

Avg. PM2.5 (µg/m³) vs. Revenue/Sales (USD$)

Machine Learning Model

For our machine learning model, we have developed an ensemble voting regressor which incorporates both the Adaboost and Support Vector Regression algorithms. Our model fits each of these algorithms to the data and averages the individual predictions to make a final prediction. For this model, we have focused on air quality measures of Ozone, PM25, and SO2.

ML Model Results for California Age-Adjusted Asthma Emergency Department Visits
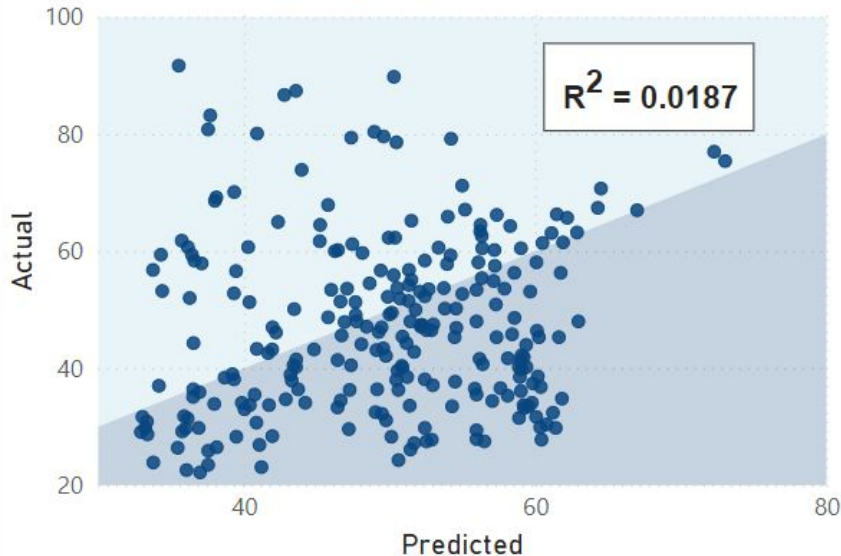
$R^2 = 0.0187$

(Scatter plot: Actual vs. Predicted)

ML Model Predicts Age-Adjusted Emergency Department Visits for US Counties

(Choropleth map of the United States)
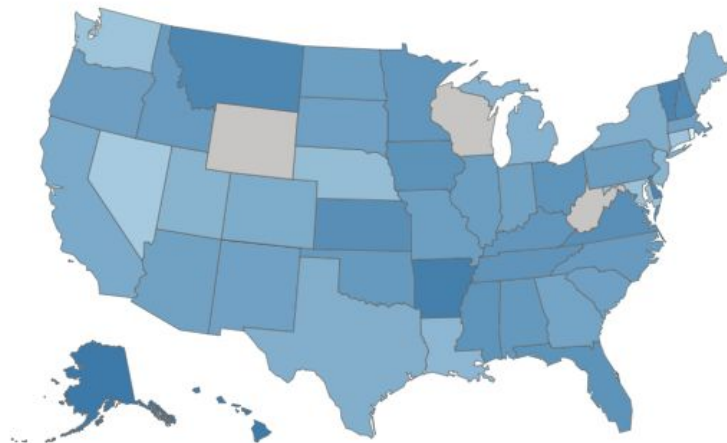
Final Thoughts

Recommendations

Daily Asthma and
AQ Data

More Asthma
Datasets

Additional
Predictor Variables

Conclusion

- Extracted and loaded data to a SQL database in order to explore relationships between **1)** air quality **2)** industry activity **3)** asthma ED visits

- Developed an ensemble voting regressor, incorporating both Adaboost and Support Vector Regression machine learning algorithms

- Designed Power BI dashboard to provide countywide asthma risk assessments based on local air quality measurements

- Access to more data resources would improve the accuracy and significance of our ML results

Any Questions?

Learn more about our project @
https://github.com/jlb14n/Capstone !

References

1. California Department of Public Health. (2019). Asthma ED Visit Rates by County (November 10, 2021). Retrieved from https://data.chhs.ca.gov/dataset/asthma-emergency-department-visit-rates. Accessed February 3, 2022.

2. US Environmental Protection Agency. Air Quality System Data Mart [internet database] available via https://www.epa.gov/outdoor-air-quality-data. Accessed February 10, 2022.

3. SB1200CSA05 - Statistics for All US Firms by Industry, Gender, and Receipts Size of Firm for the US and States: 2012. (2015, December 15). Retrieved February 4, 2022, from https://data.census.gov/cedsci/table?q=SB1200CSA05&tid=SBOCS2012.SB1200CSA05.