Project Executive Summary: Jadr Health Insights

Capstone Project

Justin Bartell, Abbey Guilliat, Darrell Gerber, Regina Huber
February 16, 2022

Introduction

Our capstone project explores how asthma incidence rates are related to environmental factors, specifically focusing on regional air quality and quantifiable measures of local industrial activity. We developed a machine learning model that takes the previous year's air quality data to predict the yearly number of asthma-induced emergency room visits for any pre-selected county. Additionally, we explored the correlation between various industries and air quality by considering industry size and revenue outputs as measures of industry activity.

Our driving questions throughout our research were: What are the distributions of asthma rates, air quality, and industry by county? Do these distributions correlate with each other? Can air quality predict asthma emergency room visits by county per year?

Descriptive/Summary Data: Asthma, Air Quality, and Industries

When starting our research, we explored the descriptive statistics of our three datasets. We looked at the shape, center, outliers, spread, and geographical patterns to better understand the data that we gathered.

Asthma

> Due to the sparseness of information, the maximum granularity of asthma-related emergency room visits was within a county yearly. The only dataset available was California between 2014-2019 (1). The number of emergency room visits was age-adjusted and normalized per 10,000 residents.
>
> The normalized asthma-related emergency room visits by county in California is slightly right-skewed, with no outliers, mean of 48.33, and ranging from 23.14 to 90.82 (Figure 1). Geographically, there is not an obvious pattern to higher normalized asthma rates (Figure 2).

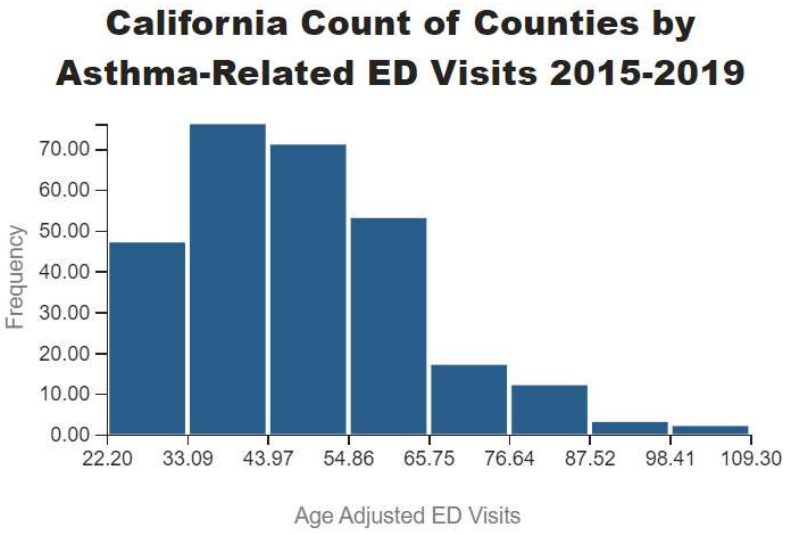

Figure 1: Histogram depicting the count of counties in California with binned numbers of asthma-related emergency department visits

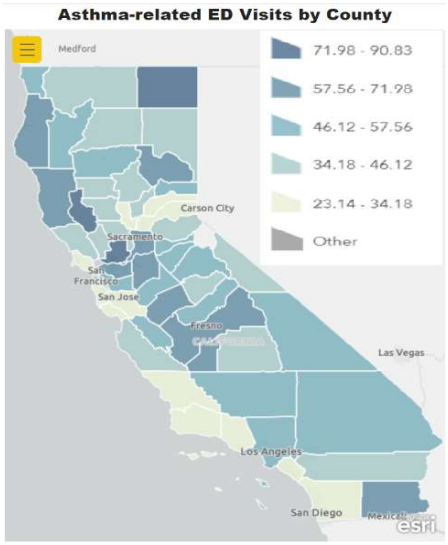

Figure 2: filled ARCGIS map depicting the number of asthma-related emergency department visits by county in California

Figure 3 shows that asthma-related emergency department visits have decreased from 2015-2019 in California, from 50 to 42. Even though there was an increase in asthma rates from 2016 to 2017, the overall trend shows a decrease in asthma rates.

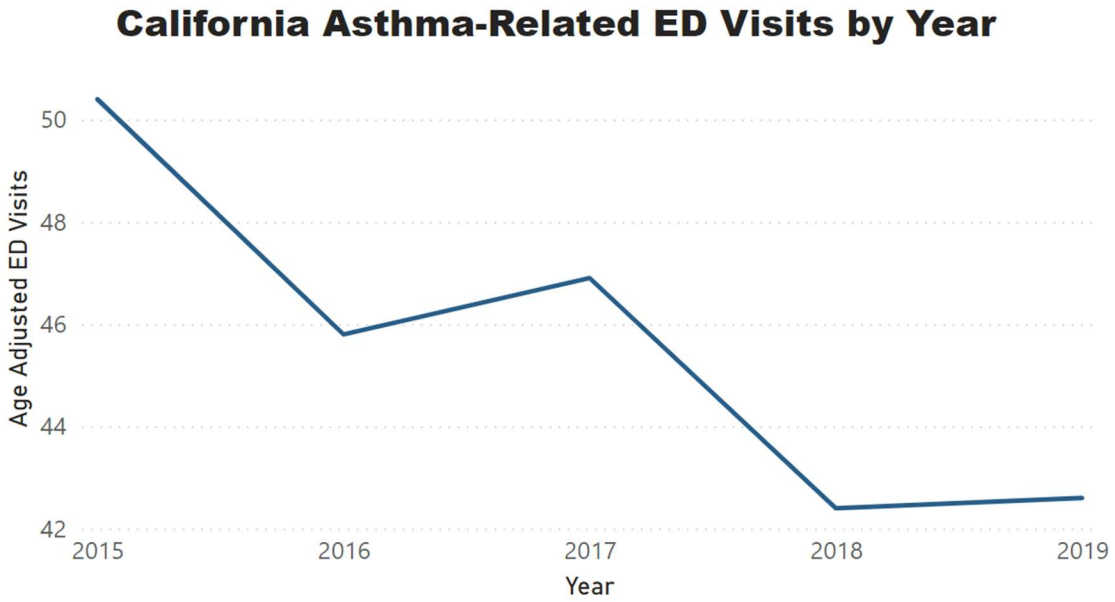

Figure 3: line chart depicting the number of asthma-related emergency department visits in California from 2015 to 2019

Air Quality

The EPA monitors various air quality metrics across the United States (2). Six of the air metrics we researched were lead, nitrogen dioxide (NO_2), ozone, particulate matter 10 micrometers and smaller (PM_{10}), particulate matter 2.5 micrometers and smaller ($PM_{2.5}$), and sulfur dioxide (SO_2). Besides ozone, the distribution of these metrics by county nationwide is right-skewed. All except lead have outliers present, as defined by greater than 1.5*IQR from Q1 or Q3. (Figure 4). Additionally, most counties tend not to track these metrics except ozone (Figure 5).

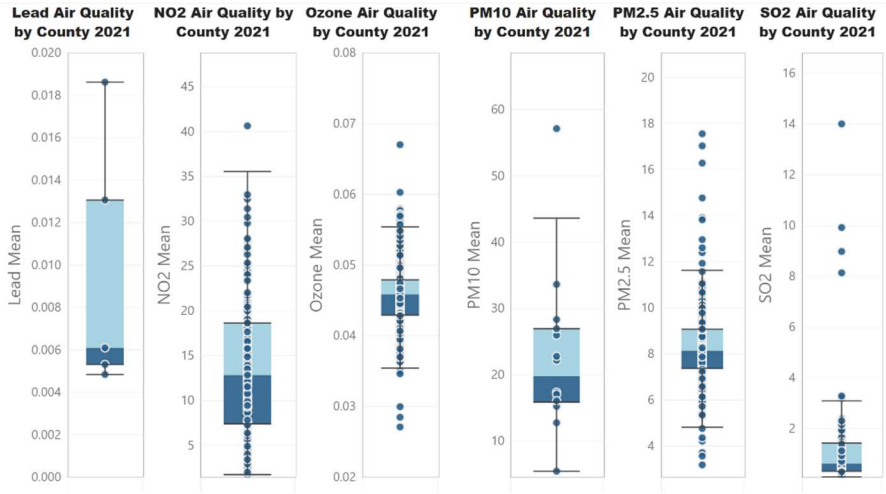

Figure 4: Box and Whisker plots of the 2021 US county averages of air quality metrics. Whisker type is 1.5 times the interquartile range or max/min if there are no outliers. Individual dots represent the counties themselves.

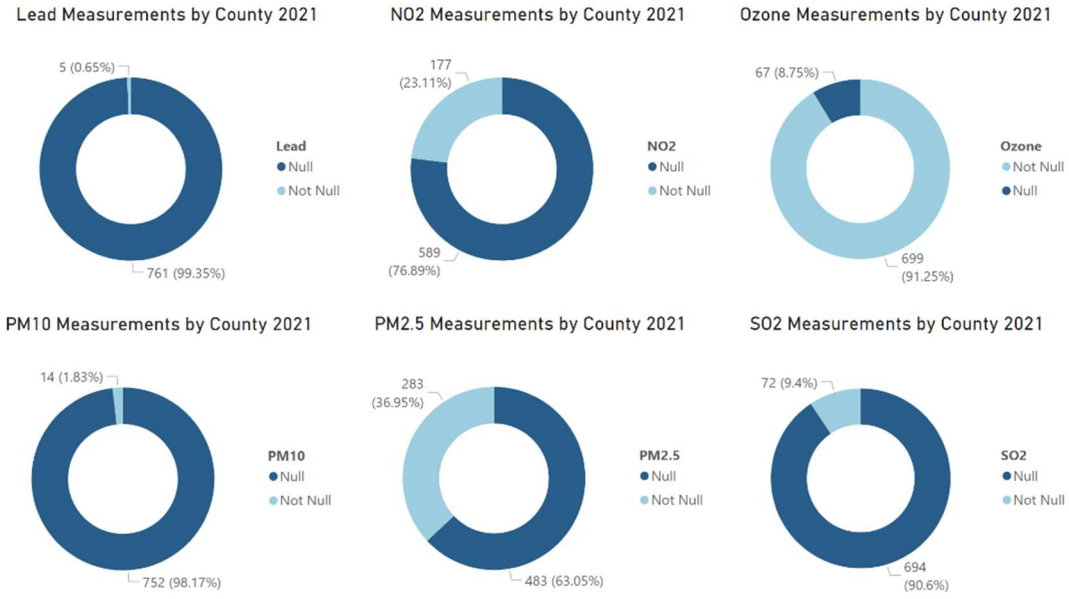

Figure 5: Donut charts depicting the percentage of null and non-null values in our dataset for various air quality factors

Industries

The US Census Bureau provides annual information about industrial sectors, yet aggregation by county is available only for 2012 (3). Nationwide, the distribution of manufacturing and transportation industries is heavily right-skewed with multiple outliers (Figure 6). For example, the county aggregated manufacturing industry revenue has a mean of $2,817,407 and ranges from $237 to $204,389,793, while the transportation industry has a mean of $415,644 and ranges from $181 to $39,365,654. The mean manufacturing industry per county is 278, ranging from 25 to 25,533, while the transportation industry has a mean of 462 per county, ranging from 25 to 51,066.

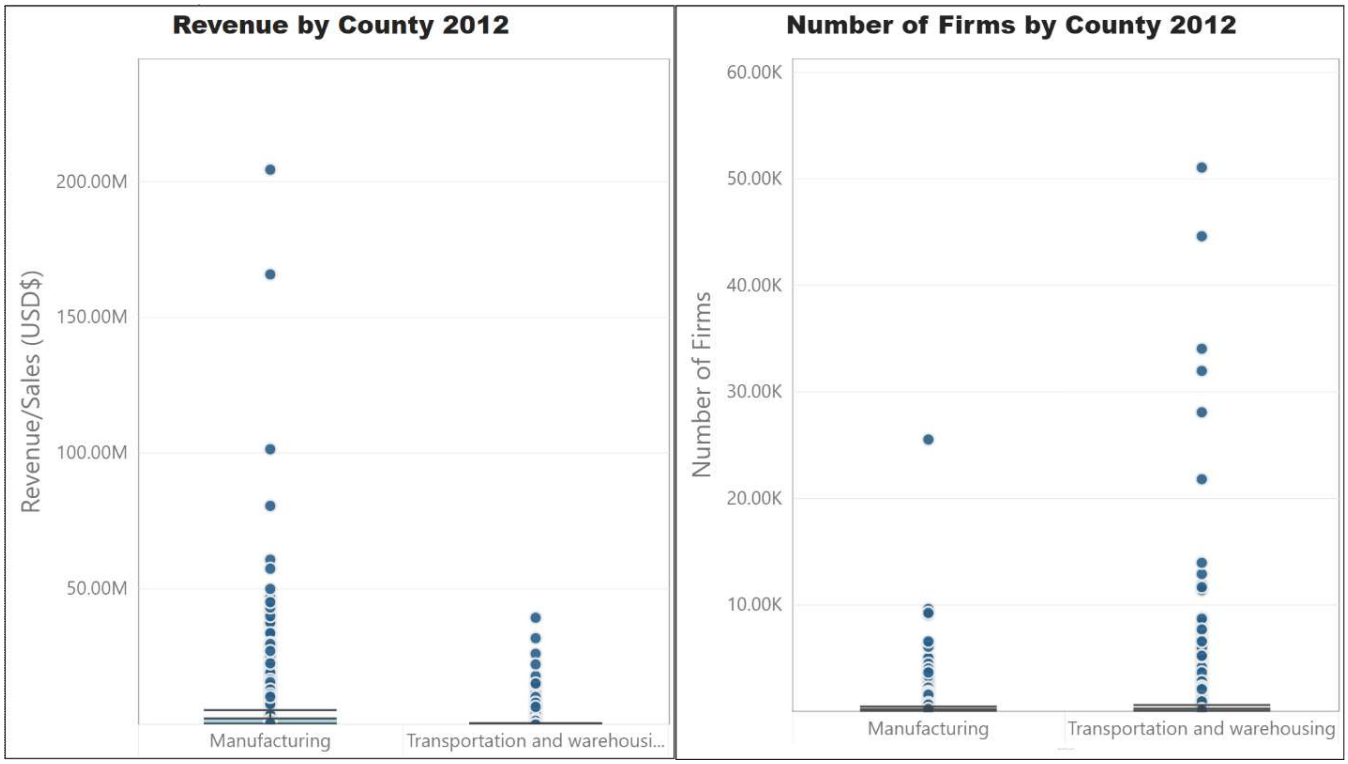

Figure 6: Box and Whisker plots of industry metrics' 2012 US county averages. Whisker type is 1.5 times the interquartile range. Individual dots represent the counties themselves.

Comparative Analysis: Asthma Rates, Air Quality, and Industries

After better understanding the datasets we chose, we looked at how the different features of each dataset compared with the others. Our focus was on finding the best parameters to predict asthma-related emergency room visits, so we looked for those features with the highest correlations with that metric.

Asthma Rates and Air Quality

Total asthma-related emergency room visits and air quality were broken down by county over an entire year. None of the outliers were removed in our analysis due to the sparsity of our dataset. In Figure 7, each point represents a given county in a given year. Each air quality metric was either weakly, positively correlated, or not linearly correlated with asthma-related emergency room visits, likely due to its highly aggregated nature: Lead R^2 is 0.168. NO_2 R^2 is 0.015. Ozone R^2 is 0.003. PM_{10} R^2 is 0.189. $PM_{2.5}$ R^2 is 0.118. SO_2 R^2 is 0.262.

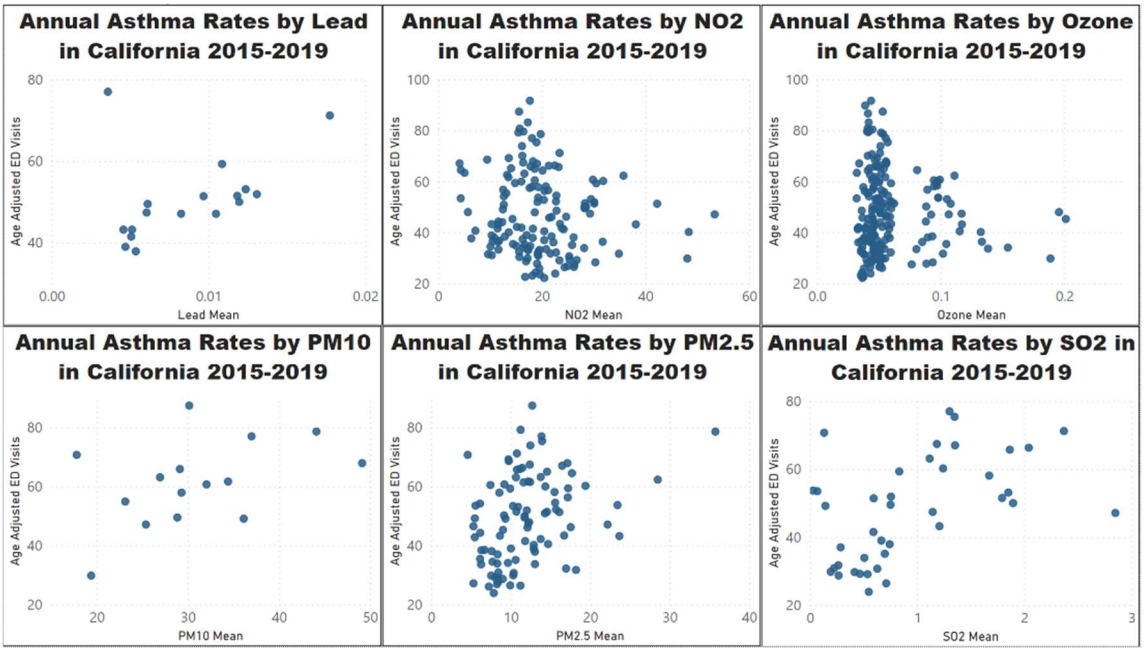

Figure 7: scatterplots depicting the correlations between annual averages of asthma-related emergency department visits and various air quality factors in California from 2015 to 2019

Figure 8 depicts the annual averages of asthma rates and various air quality factors in California from 2015 to 2019. The values represent the percent of the total, so we can easily see the correlation between air quality and asthma rates. The line graphs show that there is no significant correlation between air quality factors and asthma rates over time. However, we were only able to get data for annual averages, so we may see more of a correlation if we were able to look at seasonal or monthly data.

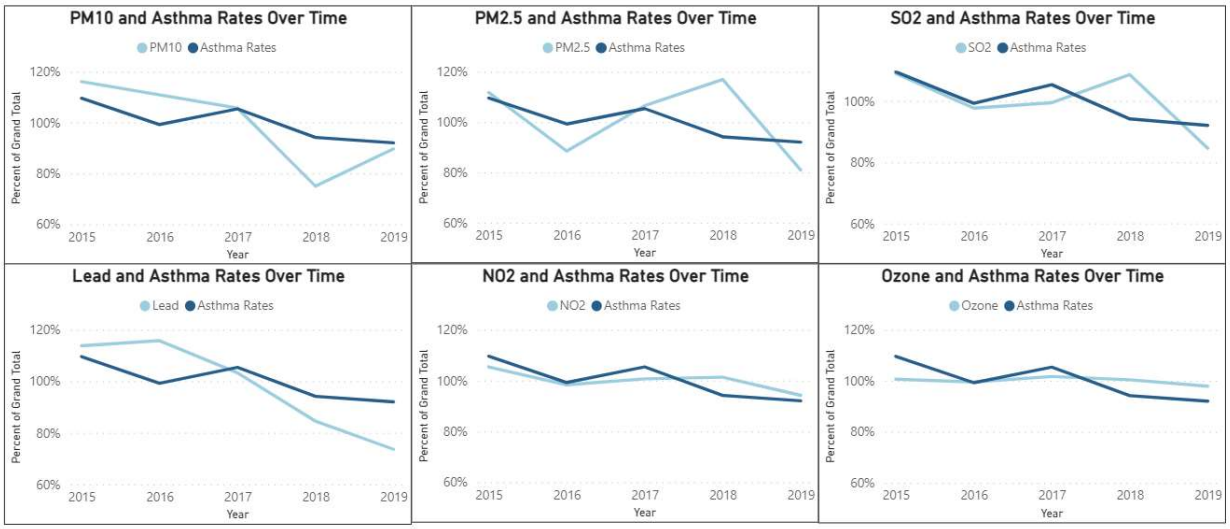

Figure 8: line charts depicting the change in the percent of the total of annual averages of asthma-related emergency department visits and various air quality factors in California from 2015 to 2019

Asthma Rates and Industries

Total asthma-related emergency room visits and industry size/revenue were broken down by county over an entire year. For the sake of our analysis, we focused only on the Manufacturing and Transportation industries. None of the outliers were removed in our study due to the sparsity of our dataset. In Figure 9, each point represents a given county in a given year. The number of firms was not correlated with the asthma-related emergency room visits: manufacturing R^2 is .02; transportation R^2 is .00001. The firm revenue was also not associated with asthma-related emergency room visits: manufacturing R^2 is .004; transportation R^2 is .00002.

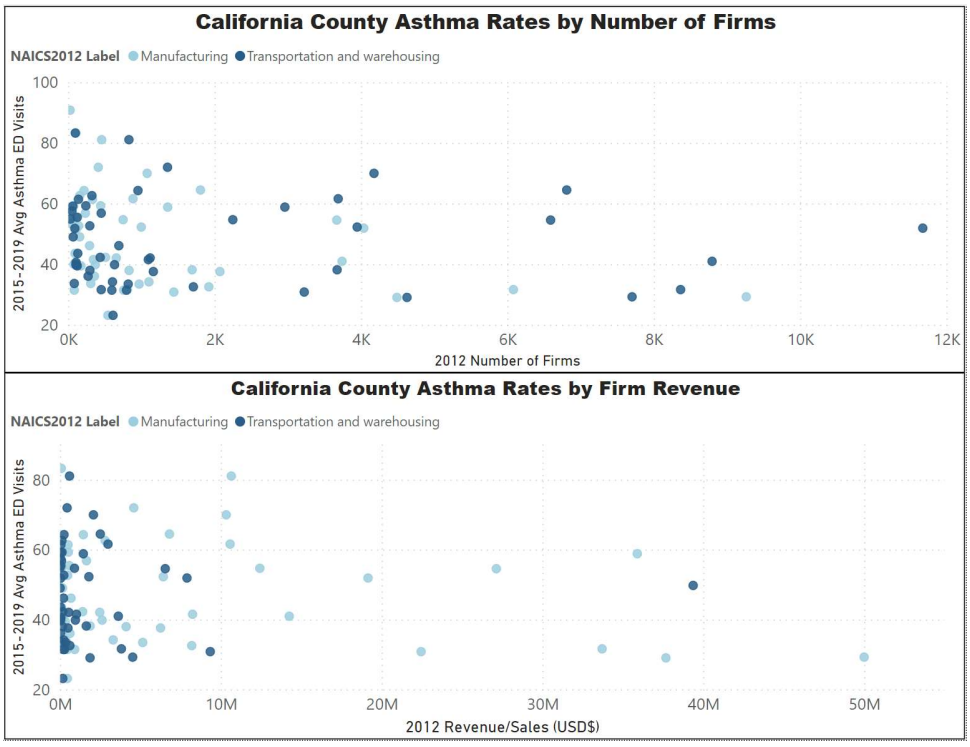

Figure 9: scatterplots depicting the correlations between asthma-related emergency department visits and number of firms (above) and firm revenue (below) in the manufacturing and transportation & warehousing industries in California

Air Quality and Industry

Focusing again on the manufacturing and transportation industries, we analyzed the correlation between industry revenue and average PM2.5 for total sales less than $10 million and more than $10 million. Figures 10 and 11 show that the manufacturing and transportation industries have a slight positive correlation between revenue and PM2.5 for total sales less than and greater than $10 million. The transportation industry for total sales greater than $10 million seems to have the highest positive correlation. However, the limited amount of data makes it difficult to determine the accuracy of this scatterplot.

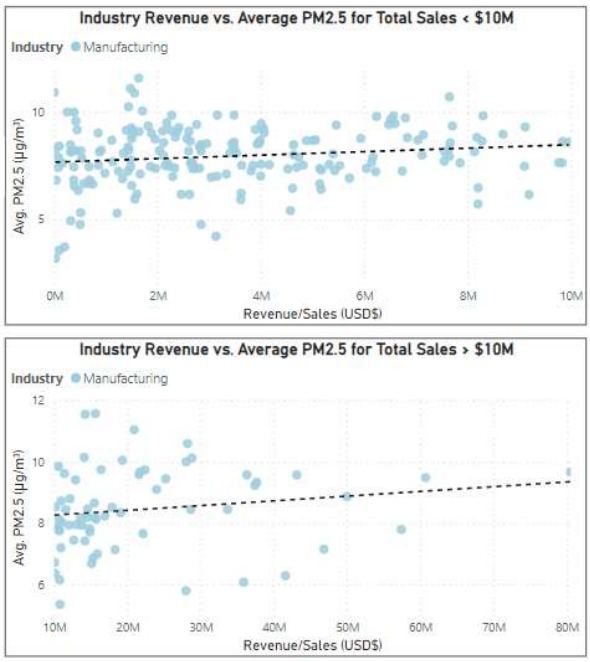

Figure 10: scatterplots depicting the correlations between industry revenue and annual PM2.5 averages for total sales less than $10 million (above) and greater than $10 million (below) in the manufacturing industry; outliers removed

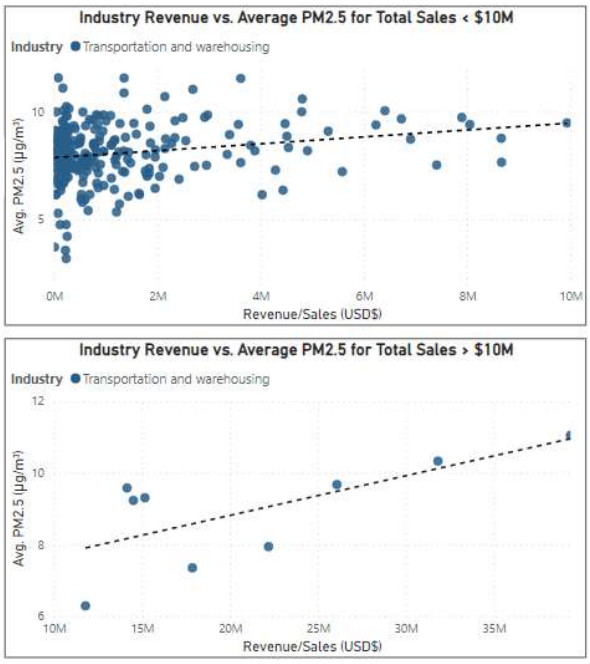

Figure 11: scatterplots depicting the correlations between industry revenue and annual PM2.5 averages for total sales less than $10 million (above) and greater than $10 million (below) in the transportation and warehousing industry

Machine Learning: Air Quality as a Predictor of Asthma Rates

Finally, we developed multiple machine learning algorithms using air quality metrics to predict asthma rates. Unfortunately, industries poorly correlate with asthma rates and won't be valuable predictors.

The machine learning algorithms tested were linear regression, LASSO, AdaBoost, support vector regression, and voting regression. The factors in the models were varied to include combinations of ozone, PM25, SO2, and NO2. Lead and PM10 were removed from consideration because very few measurements were available for the range of years and California counties in our asthma dataset (Figure 5).

Linear regression and Lasso algorithms are simple algorithms that can train quickly and give good results without overfitting for well-behaved data. The Lasso is an extension of linear regression that effectively reduces the number of dependencies. Adaboost is an ensemble regression method suitable for sparse datasets. It starts by fitting a regressor on the original dataset. It then uses boosting by making copies of the regression but with weights adjusted according to the error. The subsequent regressions end up focusing more on the problematic cases. Finally, the successive regressions are combined to produce the final prediction. Support Vector Regression (SVR) is a regression method from the more prominent family of Support

Vector Machines. They are well suited to cases where the number of dimensions is large, but there are few samples. SVR also has kernels for linear and non-linear fits. However, SVR is susceptible to overfitting. Finally, voting regression isn't a regression algorithm that combines multiple machine learning models. The method used here is to average the results from each model to make a final prediction.

The matrix in Figure 12 shows the correlation between each of the remaining air quality measurements and the asthma measurement (AGE_ADJ_ED_VISITS). Comparing correlation rates between air quality measurements and asthma show weak correlations between asthma and NO2 and moderate correlations between asthma and the ozone, PM25, and SO2. Additionally, Figure 10 shows a moderate correlation between the NO2 and ozone measurements. The NO2 measures are also removed from the data set due to low correlation with the outcome variable (age-adjusted emergency room visits due to asthma) and possible multicollinearity with ozone.

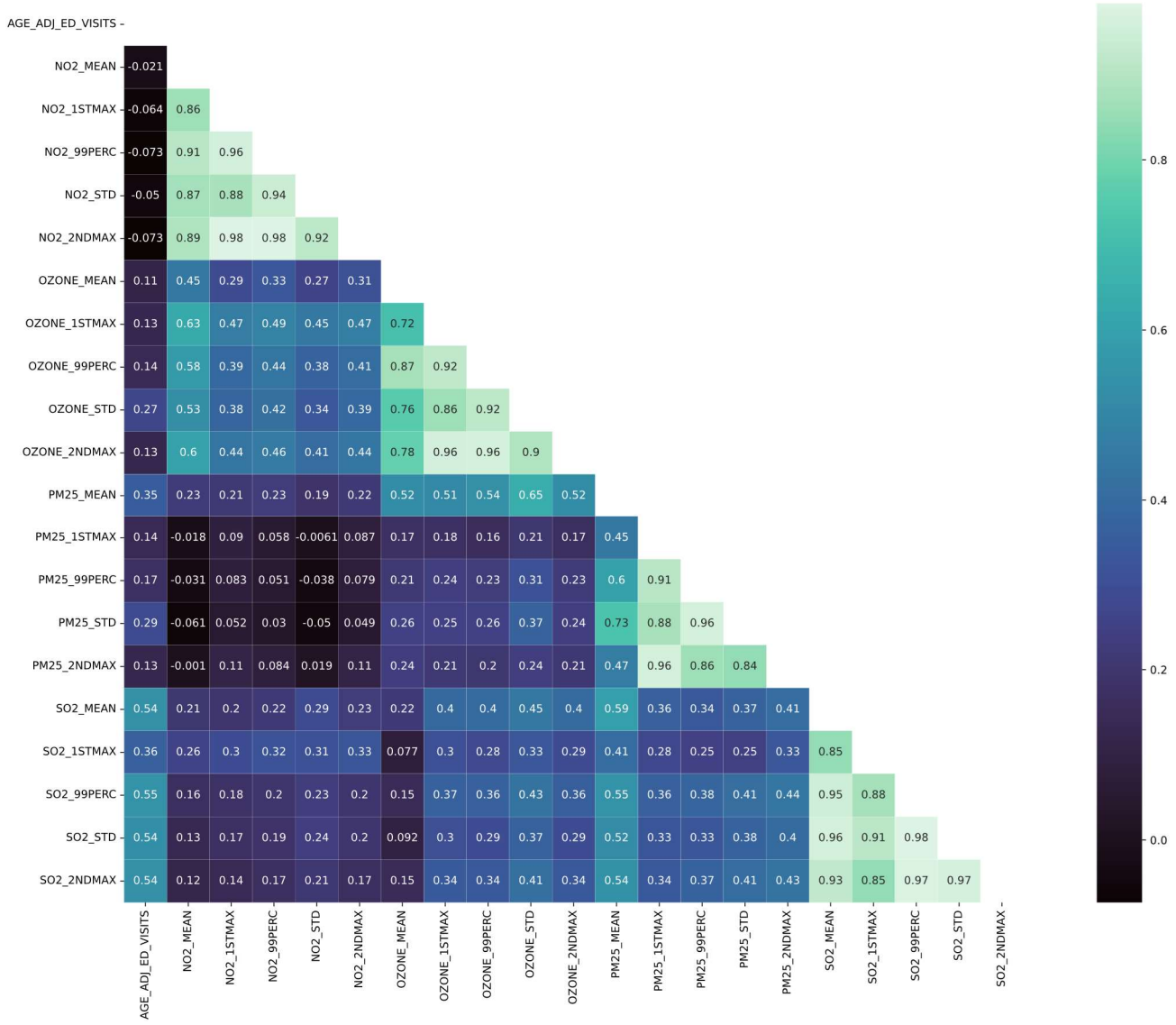

Figure 12: correlation matrix showing the correlation values between the number of asthma-related emergency department visits and the mean, first maximum, 99th-percentile, standard deviation, second maximum, and mean of NO2, ozone, PM2.5, and SO2

The SO2 measurements are missing for many counties in California. We tried two approaches to dealing with the missing data. First, remove any cases from the training without all of the measurements for ozone, PM25, and SO2. An alternative approach is to impute the missing values. Rather than fill in the missing data with the average for all available measurements, a K-Nearest Neighbor (KNN) algorithm is used. KNN was selected because each row is naturally similar to specific other rows via date and location and confounding factors leading to similar air quality conditions. KNN addresses this by filling in the missing value with the average across the K rows most similar to it in the other non-empty columns. K was set to 3 here.

The training set is 75% of the available data in all cases, with 25% reserved for testing the model. Additionally, all of the columns are standardized to have zero mean and one standard deviation.

Table 1: Training effectiveness for Linear Regression and LASSO algorithms for predicting asthma based on ozone, PM2.5, and SO2 measurements. Comparison between dropping missing measurements or imputing values with K-Nearest Neighbor (K=3).

Machine Learning Method	*Imputation*	*R^2 Training Set*	*R^2 Testing Set*
Linear Regression	Dropped Nulls	0.89	-0.555
Linear Regression	KNN	0.23	-0.196
LASSO	Dropped Nulls	0.512	0.077
LASSO	KNN	0.09	0.09

As expected, linear regression and LASSO algorithms did not perform well on our dataset. However, they both gave promising results on the dataset with dropped missing values but could not predict the test set.

Table 2: Training effectiveness for AdaBoost and Support Vector Regression (SVR) algorithms for predicting asthma based on ozone, PM2.5, and SO2 measurements. Comparison between dropping missing measurements or imputing values with K-Nearest Neighbor (K=3).

Machine Learning Method	*Imputation*	*R^2 Training Set*	*R^2 Testing Set*
AdaBoost	Dropped Nulls	-0.25	0.666
AdaBoost	KNN	0.566	-0.011
SVR	Dropped Nulls	0.842	0.67
SVR	KNN	0.748	-0.039

The AdaBoost method did not perform well in training when the missing values were dropped and gave much better results when KNN was used. However, the inverse was confirmed on the testing set. The SVR algorithm performed moderately well in training with both imputation methods. However, only the dropped nulls method could maintain any prediction ability on the testing data.

Table 3: Training effectiveness for Voting Regression (which returns the mean predicted value between the AdaBoost and SVR algorithms) for predicting asthma based on ozone, PM2.5, and SO2 measurements. Comparison between dropping missing measurements or imputing values with K-Nearest Neighbor (K=3).

Machine Learning Method	*Imputation*	*R^2 Training Set*	*R^2 Testing Set*
Voting (AdaBoost and SVR)	Dropped Nulls	0.94	0.777
Voting (AdaBoost and SVR)	KNN	0.713	-0.026

Combining the best-performing algorithms improves the results on both the training and testing data sets without imputation. However, models trained on the KNN data sets still performed poorly.

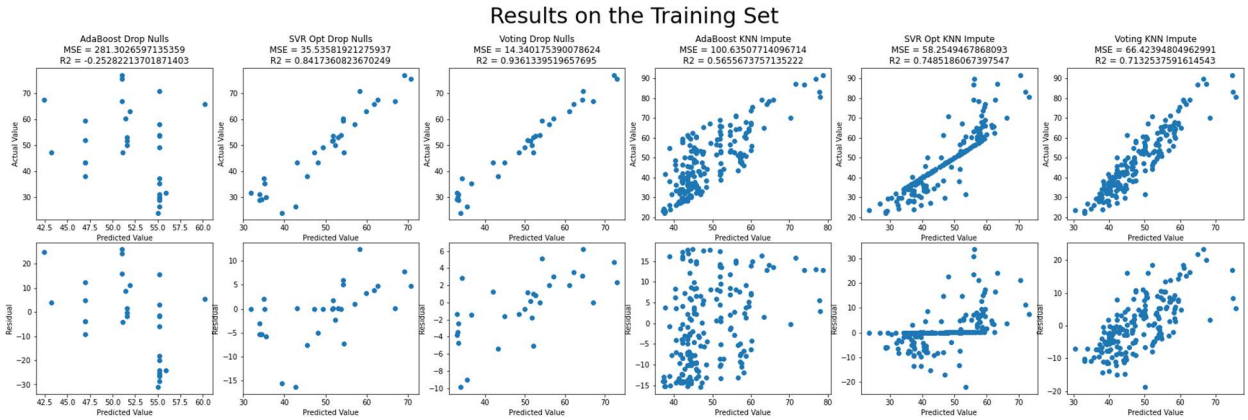

Figure 113: The actual values versus predicted values and residuals versus predicted values for an array of machine learning models evaluated to predict age-adjust asthma emergency department visits per 10,000 people. The input data is the data used to train the models.

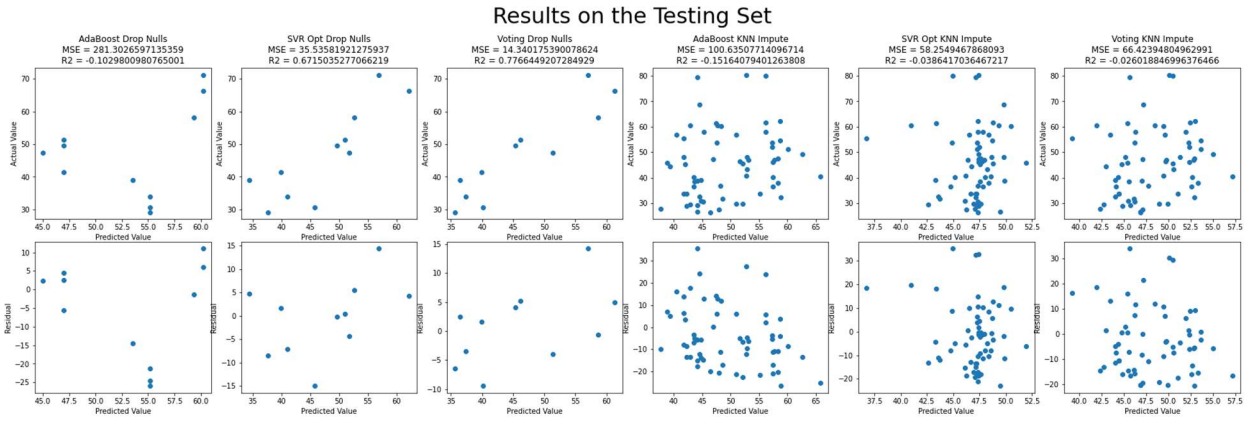

Figure 124: The actual values versus predicted values and residuals versus predicted values for an array of machine learning models evaluated to predict age-adjust asthma emergency department visits per 10,000 people. The input data is 25% of the original data reserved to test the models' response against information not used in training.

Figures 13 and 14 show the actual value versus predicted value (top row) and residual versus predicted value (bottom row) for each model explored in Tables 2 and 3. The models performing well on the training set (Figure 13) show a cluster of points along a line going upward from left to right. However, the KNN models that performed well on the training set have patterns in the residuals plot indicating that the data is not heteroskedastic (the residuals show a dependency on the predicted value). Figure 14 shows that the only model that performed well on the testing set is voting regression with dropped nulls.

The voting regression model merging SVR and AdaBoost models outperformed other models and is the selected model to predict asthma-related emergency department visits from air quality measurements. Despite the strong performance of the chosen model against the training and testing sets, we remain dubious about the model's predictive capabilities. The data shows a weak correlation between air quality measurement and asthma rates, is likely non-homoscedastic, is sparse and isolated to one geographic region (California).

To test the robustness of the selected model against a broader dataset, the model (which was trained on data where the missing values were dropped) was run on the original dataset with the missing values instead imputed using a K-Nearest Neighbors imputation. Figure 15 shows that the model performed poorly against the broader dataset yet still has weak predictive capabilities. The air quality data in Figure 15 contains a mixture of the data used to train the model (31 points), the testing data used to evaluate the model (11 points), and new data (205 points) containing values for some measurements (mostly SO2) imputed from the training and testing data values.

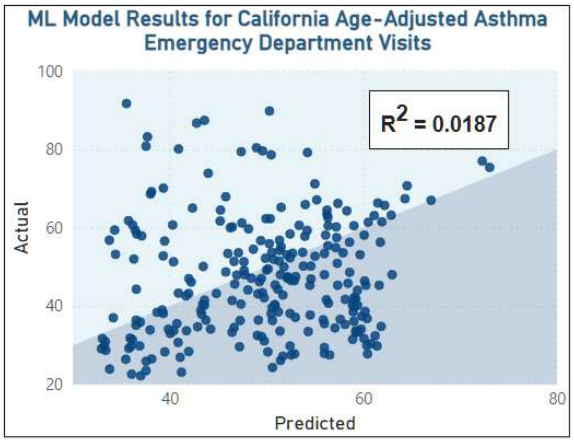

Figure 13: Actual asthma visits versus predicted visits using a model that averages predictions from Adaboost and Support Vector Regression models. The datasets used to train and test the model have no imputation. However, the data in this figure contains those values and data where missing values were imputed using K-Nearest Neighbors (K=3).

Understanding the limitations of the predictive capabilities of the selected model, Figure 16 shows predicted average county age-adjusted emergency department visits per 10,000 people for each state.

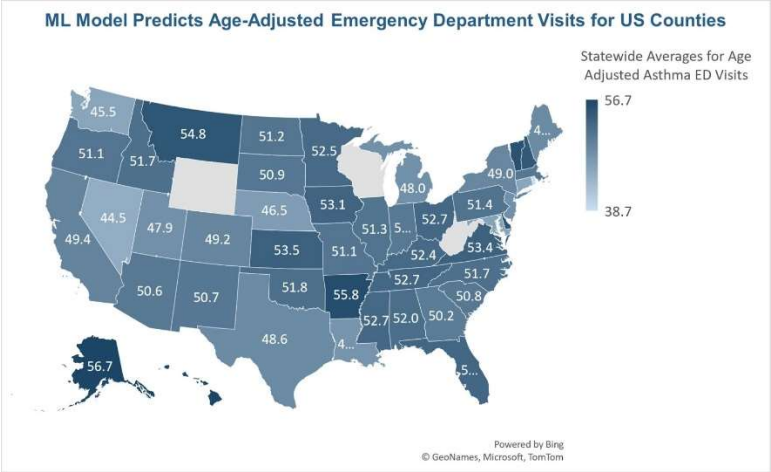

Figure 146: The predicted average county age-adjusted emergency department visits per 10,000 people for each state using a voting regression model averaging the predictions of AdaBoost and Support Vector Regression models. The feature data consists of air quality measurements for 2021 in 766 counties across the US with missing values for SO2, Ozone, or PM2.5 imputed using K-Nearest Neighbors (K=3). The county average shown is the average of counties in each state for which air quality data is available.

Conclusion

Yearly air quality data and industry revenue/number of firms are poor predictors of the county's annual asthma-related emergency room visits. While a small correlation exists between asthma-related emergency room visits and air quality data, there are significant confounding factors not included in the data leading to poor machine learning performance. The poor predictive capability is compounded by data sparsity since most of the data was dropped during training or imputed during predicting. In the future, data on a per-day basis and more complete air quality metrics over every county would likely greatly help the model. The confounding factors would remain but increased data over many days in the same geography coupled with a closer temporal tie between the air quality and asthmas emergency room visits should significantly improve the predictive ability of machine learning analysis. Further, there is a slight correlation between industry and air quality data; thus, no machine learning model was developed to predict asthma based on industry.

This project demonstrates our ability to develop stream-based processing with Kafka, cloud computing with Azure DataLake/Databricks/Data Factory, expertise with python extraction-transformation-load process, DDL to generate a normalized SQL database, and PowerBI to develop visualizations for data analysis.

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