DATA 200 Graduate Project Fall21

Air Quality Index Prediction Based on Machine Learning Models

Fang Hu[1], Siqi Jiang[1], and Runda Tian[1]

[1]University of California, Berkeley

December 13, 2021

Abstract

This project looks into the climate and environment dataset (topic 2A) with a primary goal of building an effective model to predict future Air Pollution Index (AQI) using previous data including precipitation, temperature and AQI in Alameda. Our team explores several important datasets published by Global Historical Climatology Network (GHCN) and Environmental Protection Agency (EPA). Through this process, we perform data cleaning, EDA, and transformations to identify the relationship between the different variables. Besides that, our group implemented linear regression and random forest regression in order to see how these models perform in different situations. Experimental results showed that after hyperparameter tuning, the multiple linear regression model performed better in the prediction of the AQI. This work also illustrates that combining machine learning with air quality prediction is an efficient and convenient way to solve some related environmental problems.

Keywords
AQI, sliding window, linear regression, causal inference

1 Introduction

Nowadays, air pollution has become one of the most important environmental concerns around the world. Predicting air quality is a crucial topic in this area, helping people to arrange their daily activities and realize the importance of controlling air pollution. Given the advanced machine learning methods and the existence of big data, now we are able to analyze the underlying trends in data and build a model to predict future air quality.

Air Quality Index (AQI) is an important indicator that is widely used to reflect and evaluate air quality. It measures air quality based on six major pollutant including fine particulate matter (PM2.5), ozone (O3), sulfur dioxides (SO2), inhalable particles (PM10), nitrogen oxides (NO2), and carbon monoxide(CO), and is in the range between 0 and 500 [1]. Based on these numbers, we can divide air quality into six levels (good, moderate, lightly polluted, moderately polluted, heavily polluted, and severely polluted), which describes how the air pollutant can impact human health and provide a good reference for people's outdoor activities in a numerical form [2]. Predicting AQI can help us to evaluate the current air quality and solve air pollution problems in the long run.

Predicting AQI is plausible as there are several factors that correlate with AQI. Precipitation and temperature are two important factors. According to the Environmental Protection Agency (EPA), "changes in climate can result in impacts to local air quality" [3], as higher air temperatures can speed up chemical reactions in the air. On the other hand, "rain typically results in less pollution since it washes away the particulate matter and can also wash out pollutants that are dissolvable" [4]. In our model, we integrate temperature and precipitation given their potential influences on AQI forecasts.

2 Related Work

[2] studied a dataset about AQI from Beijing Municipal Environmental Monitoring Center and used support vector machines (SVMs) to predict future AQI based on information from previous days. We extend from this paper, aiming to find a simpler model that requires fewer features and less computing power.

3 Datasets

This project looked into two main datasets, both containing daily information per county. Our modeling and inference mainly focus on Alameda county, California in year 2020. The reason for choosing one county is that we would like to make temporal granularity be one day so that our model can be used in real-world daily AQI forecasting. With over $3,000$ counties nation-wide, such granularity will require a huge amount of computational resources as well as time. The model could be generalized to any other county since it only requires local weather and AQI data. The reason for using data collected in only one year is as follows. In daily AQI prediction, long-term trends such as seasonal trends and yearly trends contributes much less to prediction results when it is compared to short-term trends due to time-locality. Focusing on just short-term features will reduce model complexity at merely no expense of prediction accuracy. Following is overview on the datasets and data cleaning we perform.

3.1 Daily Weather Dataset

The first one includes data about 2020 year's daily weather that was collected by the National Centers for Environmental Information (NCEI)'s Global Historical Climatology Network (GHCN). GHCN database receives daily updates from a variety of data streams and undergoes a suite of quality checks, which ensure its accuracy and authenticity. It contains daily climate reports from over 100,000 stations in 180 countries. Using GeoReverse and the stations' latitude and longitude, we are able to get daily temperature and precipitation in Alameda, which is the area that we want to investigate.

Data Cleaning There are no missing dates there and no outlier in both temperature and precipitation. After filtering out non-U.S. counties, we find that, although some minor difference in format exists, county name and total number are consistent with county-level daily AQI which we introduce in the next paragraph.

3.2 Daily AQI Dataset

The second dataset we looked into is the dataset of county-level daily AQI from EPA's Air Data database. Combining with the previous dataset, we are able to correlate the temperature and precipitation with AQI on specific dates and areas. This dataset helps us a lot in learning time trends and the connection between whether and AQI.

Data Cleaning All data are collected from counties in the U.S., which is as expected. Then we mainly focus on three variables, namely county name, date, and AQI. County name is different in this dataset and the weather one, whereas the former has 'county' after county name while the latter does not have. Except that, county names in both dataset are the same. As for dates, there are some missing dates in the county-level daily AQI dataset. Since there are no missing dates in the aforementioned daily weather dataset, we drop missing dates in the daily county-level AQI dataset by inner-joining the two. Lastly, AQI is defined in a range from 0 to 500 (inclusive) [5]. There are no missing values (such as NaN or null) in the dataset but some outliers do present, which are higher than the upper bound of AQI, i.e., 500. Instead of removing them, we replace them by the upper bound as air quality in these days should be considered as hazardous, not none.

4 Methodology

4.1 Sliding Window

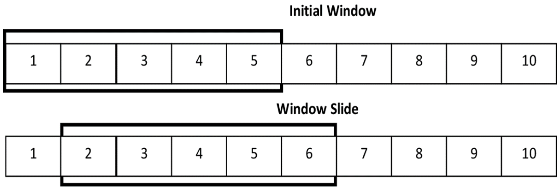


Figure 1: Sliding window

Sliding window has been applied to the final processed data frame before sending it to the model. Assumption has been made that a given day's AQI depends on the past several day's temperatures, precipitation and AQI. This assumption is backed by the EDA discoveries. The choice of window size is a hyperparameter and can be tuned from model to model. The sliding window is explained below in Fig. 1, where the past few day's (e.g. 1 to 5) features were input to predict the given day's (6) AQI, and then repeat this process by sliding the window 1 step forward.

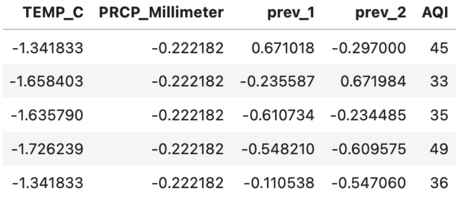
4.2 Modeling

4.2.1 Train Test Split

Since we adopt the sliding window approach, that is, data of the first n days are used to predict the AQI value of the $n + 1$ day, we can't use the regular approach of randomly shuffling the data and then split the training and testing datasets. This is because randomized splitting will cause part of testing data to be seen in the training set and it will break down the time sequencing relationship that we rely on to train the prediction model. Instead, we split the first 80% of data as a training set and leave the last 20% for testing. This splitting approach also reflects the primary research goal of this project, that is, to use the past data to predict future AQI.

4.2.2 Feature Engineering

The dataframe shown in Fig. 2 has been normalized by a standard scalar, and for a given feature, the number x at the end indicates this is the result from $x-1$ days prior. Specifically, $prev\_1$ represents AQI of the previous day, and $prev\_2$ represents AQI of the day before previous day. Each row represents a given sample to be fed into the model. For example, the first row, including today's temperature, precipitation, and AQI of the past two days, represents the features we used to predict today's AQI. The reasons for selecting temperature, precipitation as features have been explained in the former section.



TEMP_C	PRCP_Millimeter	prev_1	prev_2	AQI
-1.341833	-0.222182	0.671018	-0.297000	45
-1.658403	-0.222182	-0.235587	0.671984	33
-1.635790	-0.222182	-0.610734	-0.234485	35
-1.726239	-0.222182	-0.548210	-0.609575	49
-1.341833	-0.222182	-0.110538	-0.547060	36

Figure 2: Final dataframe

4.2.3 Loss Metric

With a goal to handle a regression problem, root mean squared loss (RMSE) was used to evaluate a given model's performance. RMSE is used both in training models and in evaluating model performance. MAE and R square were also calculated to validate judgment.

4.2.4 Cross Validation and Hyperparameter Tuning

For every model, we always tune on the training set, splitting it k folds, using every $k-1$ folds for training the model on that set of parameters and then use the remaining 1 fold for validation. We repeat this process k times for each set of parameters. Because we are handling a regression problem, we use RMSE as the evaluation metric. The lower the RMSE is, the better our model fits in with the goal. Finally, we use the best tuned parameters to construct the final model, train on the training set and evaluate on the test set.

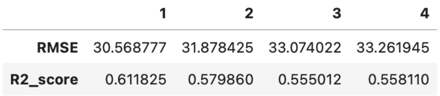
As we use sliding Window and have proposed multiple linear regression and random forest regression, it is evident that each model requires a different level of hyperparameter tuning. For every model, we will tune the sliding window size and then, depending on the model, we have different model-dependent parameters to tune: For linear regression with L1/L2 penalty, we tune the alpha parameter. For random forest regression, we tune for the number of trees in forest, the proportion of samples to be used in each tree, and the proportion of features to be used in each split.

4.2.5 Prediction Models

Multiple Linear Regression In this case, we model AQI as resulting from a linear combination of features multiplied by weights and one additional bias term. Formally:

$$\hat{y} = \sum_{i=0}^{k} \beta_i x_i + \alpha$$

where x is the feature vector, β is the weight vector and α is the bias term. With the sliding window processed features, we are using this model to see if we can find a good fit for a given day's AQI using a linear combination of features for the past few days. From the EDA section, we've seen that some features are correlated with AQI in a near-linear way. The use of linear regression helps us check the hypothesis on whether the AQI can be predicted using a linear combination of the features listed above.



	1	2	3	4
RMSE	30.568777	31.878425	33.074022	33.261945
R2_score	0.611825	0.579860	0.555012	0.558110

Figure 3: RMSE and R^2 of different sliding window size (from 1 4)

We also compared the RMSE and R square of using different sliding window sizes. Result in Fig. 3 showed that in the range of 1 to 4, the best size of sliding window is 1, which is understandable that AQI is more likely to be affected by previous day's climate.

Random Forest Regression In this case, we model AQI as the output value aggregated by a set of trees, each trained on a different subsample of the training dataset and also a subsample of features. With the use of bagging and a random sample of features at each split for each tree in the forest, we effectively generate more randomness and allow the forest to grow differently. Meanwhile, we tune each tree's depth to control the bias (deeper trees have a lower bias).

4.2.6 Modeling Process

In terms of the modeling process, we first process each individual dataset on temperature, precipitation and AQI, constraining the scope of the data to be at the Alameda County, checking outliers, and cleaning the data. Then, we aggregate the data by taking the daily summaries and then merge the datasets together, so that each row in the final data frame contains the summary statistics of temperature, precipitation, AQI, etc. for a given day in 2020. With the processed dataset, we then apply a sliding window to engineer features that we plan to use for predicting AQI, i.e., we slide the window across the dataset, creating each row that represents the past few day's temperatures, precipitation, etc. and designates the y variable to be the current day's AQI. We believe this approach is valid because of our findings in the EDA section, in which we showed that AQI has a relationship with time and previous AQI.

Then, the processed data gets split into train and test sets. Because we are dealing with time-based data, we split the first 80% of data in 2020 for training and cross validation and leave the last 20% for testing. Next, two main models are proposed, which are linear regression and random forest regression. The reasons for choosing these models are illustrated in detail in the following sections. For linear regression (ridge and lasso), we use the alpha term in hyperparameter tuning to control the degree of overfitting. For random forest regression, we use the tree depth and the number of trees in the forest to control overfitting. The modeling notebook contains cross-validation and hyperparameter tuning processes. Metric-wise, we use the root mean squared error (RMSE) to evaluate the model's performance. Visualizations on each best-tuned model's result on train and test set are also included.

5 Results and Discussion

5.1 Multiple Linear Regression

We first trained Multiple Linear Regression models, and the results shown in Fig. 4 are the best-tuned models after hyperparameter tuning.

The best Multiple Linear Regression with tuned regularization parameter generated a training RMSE of 12.46 and testing of 30.63. Fig. 4 demonstrates the performance visually, plotting the predicted and the actual y value on both the training set and test set.

5.2 Random Forest Regression

Fig. 5 demonstrates the best tuned random forest model. The best Multiple Linear Regression with tuned regularization parameter generated a training RMSE of 8.04 and testing of 39.01.

5.3 Discussion

In the MLR model, as we can see, it was able to capture the AQI increase and decrease trend. But it is not able to correctly predict the magnitude of AQI fluctuation, which is because the data potentially has high noise. This also leads to a relatively high RMSE in our test. The linear model after parameter tuning still cannot properly capture the shape of the AQI curve, as suggested by the plot for the training set's y and y_pred. This suggests that we should either engineer more useful features or try a more powerful model that can bring in non-linearity to better fit the curve.

For random forest regression, We can see that the performance is not better than the Multiple Linear Regression model. If the individual decision trees failed to capture the intrinsic trend in the AQI curve, which may because of the lack of information of whole year-round seasonal change, then the ensemble of those trees for regression will create a lot of noise, and though the averaging of these values tend to be stable, they failed to capture the periodic sharp transitions in previous days and hence failed to predict the trend in the test set. The random forest model was able to achieve a small RMSE in training, but it could not capture the spikes in the test set. Despite massively overfitting the training set, the model's feature selection was robust to the distribution of the test set, so at least it knows where the spikes would occur. However, as compared to the MLR model, this one cannot predict the magnitude of the spikes in the test set as properly as MLR. To achieve better fits, random forests require a lot of hyperparameter tuning. It is hard to adjust everything perfectly, and so we may not have found the best spot for this model. In conclusion, Multiple linear regression works better than random forest regression because of the limited dataset (we do not take data from previous years into consideration). The linear model is also good for inference. By extracting the coefficients, we find that the daily AQI is primarily influenced by the past two day's AQI and also the temperature. The precipitation is not quite correlated with

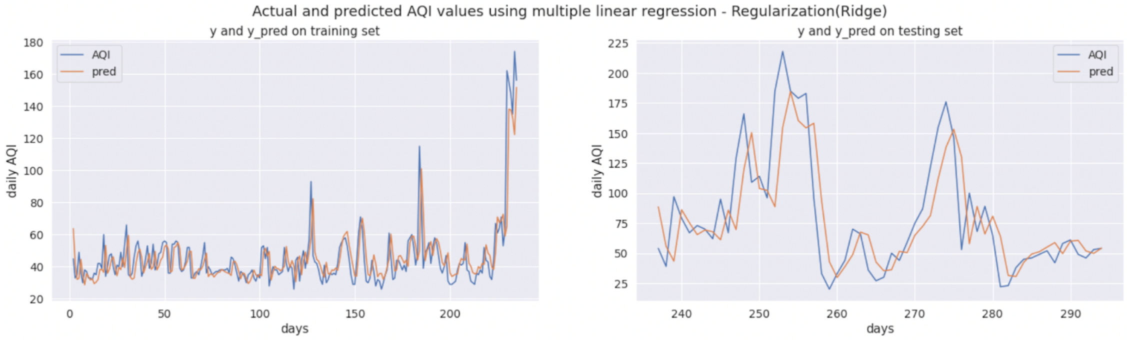


Figure 4: AQI prediction using multiple linear regression

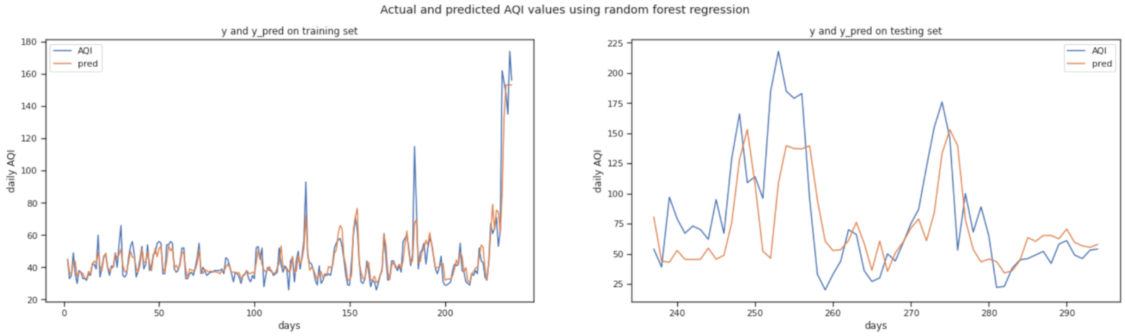


Figure 5: AQI prediction using random forest regression

the predicted AQI. This result validated our research question. At the same time, however, we see that the both two models have tried the best on fitting the curve, but it could not accurately predict the last few spikes in the test set.

5.3.1 Sampling Distributions of Parameters by Bootstrapping

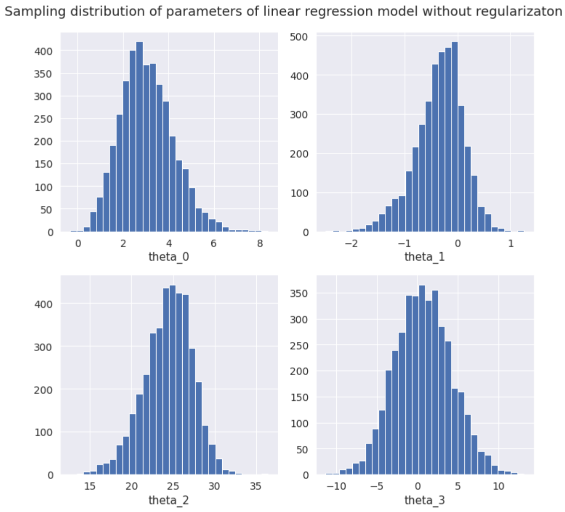


Figure 6: Bootstrap without regularization

As shown in Fig. 6, Without regularization, all four parameters of the model have a nonzero value in most estimation rounds. Notice that θ_1 and θ_3 both have mean close to 0. With input data normalization, we observe that θ_2 has the greatest absolute value, thus, it is the most useful feature when predicting AQI.

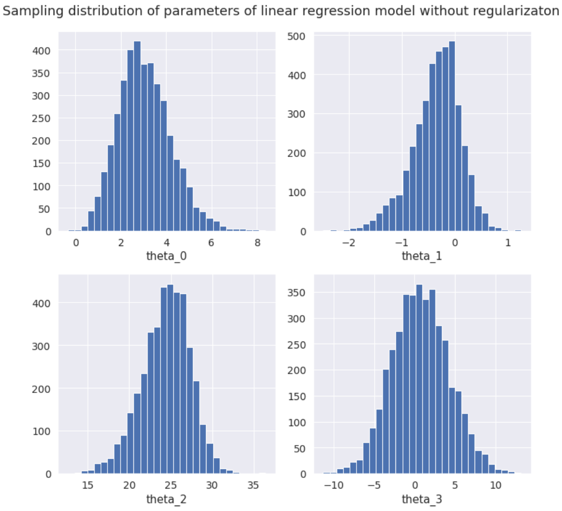


Figure 7: Bootstrap with L2 regularization (Ridge)

The effect of L2 regularization (Ridge) is not obvious in Fig. 7 since weights for 3 features and bias term does not change much, compared to the model without regularization.

We observe from Fig. 8 that L1 regularization has a significant effect on two parameters whose mean is 0 without regularization. Both θ_1 and θ_3 almost always has a 0 value, meaning that the model are forced to leave some features out, i.e., to select the most useful features under L1 constraints.

Taking all three models into consideration, we find that the third feature, previous day's AQI, with weight θ_2 is the most effective predictor when predicting next days' AQI. This is as expected because we expect AQI does not have abrupt changes in consecutive days.

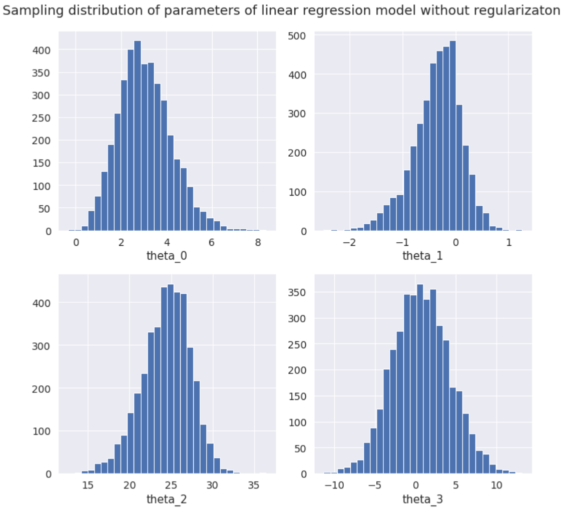


Figure 8: Bootstrap with L1 regularization (Lasso)

5.3.2 Confidence Interval

Confidence interval with 95% confidence and 1000 resamples.

Table 1 verifies that the first and third feature is not useful is predicting next day's AQI. With 95% confidence and 1,000 rounds, both the intervals of θ_1 and θ_3 contain 0, which meaning that they can be eliminated from input without reducing model accuracy greatly.

5.4 Causal Inference

Causal inference focuses on causation and its effect among variables. It is distinct from conditional probability, whose task is to estimate the probability of a variable when some variable is fixed, without caring about the internal relationship between them. For example, given the lawn is wet, we want to guess today's weather. It is likely that it's a rainy day, which means there is some correlation between a wet lawn and rainy day. We can safely predict that today is likely a rainy day if we see a wet lawn, which is a conditional probability.

However, according to our knowledge and experience, if we water the lawn manually, it is clear that the weather will not change just because the lawn becomes wet. In this case, we deliberately set the lawn to be wet, instead of observing the phenomena passively, in order to determine whether a wet lawn causes a rainy day. That is, we actively perform an action to see whether it will lead to changes of variables we are interested in. One thing to note is that we need to figure out the causal relationship, or so-called causal graph, among variables by domain knowledge just as in this example. Causal inference cannot help us to determine such a relationship from data.

The distinction between causal inference and conditional probability leads to some seemingly contradictory results. A simple example would be the relationship among kidney stone size, treatment (A and B), and success in kidney operations [6]. Using our domain knowledge, we know that a larger kidney stone could lower the probability of success. One may observe that for groups of people with both small and large kidney stones, success rate of treatment A is greater than that of B, but taking all groups of kidney size into account, treatment B has a larger success rate. This is because kidney size is the cause of both treatment and operation success, and doctors have preference for treatment A/B when treating with different kidney sizes. To avoid such contradiction, when exploring the causal relationship between education and wealth, we have to control age since it could influence both education and wealth. Without limiting the age group we are studying at, we do not know the reason for the change in wealth from age or education. Age in this example is called a confounder as it is the cause of another two variables.

Given the concept of causal inference and confounder, one of the best ways to explore causal effects between two variables is to conduct a randomized trial. In a randomized trial, we randomly set the variable that is deemed as causation and see changes in another variable. This way, we break any influence by confounders since we take active actions. In real world scenarios, it is usually the case that a randomized trial is impractical, thus, we can only observe some variables without intervention. This is exactly the case for the climate dataset, in which AQI and affecting variables can only be observed. We cannot control the weather to modify temperature/precipitation/wind speed, nor deliberately randomize the amount of air pollutant emit-

	θ_0		θ_1		θ_2		θ_3	
Linear model without regularization	1.06	5.75	-1.35	0.39	18.41	30.01	-5.59	7.62
Linear model with L2	0.92	5.45	-1.45	0.42	18.46	29.35	-5.80	7.99
Linear model with L1	0.99	5.64	-1.42	0.48	18.60	29.92	-6.30	7.79

Table 1: Confidence Interval for features

ted into the atmosphere. Fortunately, one way to conduct causal inference is to control the confounder, i.e., the cause for other variables that have causal relationships. For the climate dataset, if we want to investigate the causal relationship between precipitation and AQI, we need to first build a causal graph, then control confounder variables such as temperature to analyze those two variables.

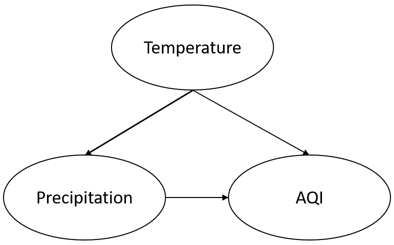


Figure 9: Causal graph w/ temperature as confounder

To explore the causal relationship between precipitation and AQI, we first propose a causal graph. Based on common knowledge, air quality on rainy days is better than on dry days since rain cleans the air by washing away air pollutants such as PM2.5. We therefore draw an arrow from precipitation to AQI to indicate the causal effect. Also, high temperature results in less precipitation for some climate region and more for others. Besides, high temperature can lead to wildfire which produces smoke and drives AQI high. Fig. 9 is a causal graph based on the above analysis.

To eliminate the influence of the confounding variable, temperature, we assign temperature to be in small baskets with the fact that temperature is a continuous numerical variable. In each small range, we observe the relationship between precipitation and AQI. Figures are shown below.

From Fig. 10, we observe that on days with high precipitation (greater than 50 mm), AQI is almost always lower than 100, meaning that the air quality is good. On days with low precipitation , AQI can be low or high but there are more points distributed in the lower left. We conclude that precipitation is indeed a causation of AQI in a way that high precipitation leads to low AQI.

Same analysis can be done for wind speed, precipitation, and AQI. In this case, wind speed is a confounder. High wind speed usually causes precipitation as cold air mass moves fast and forms rain when it confronts warm air mass. So wind speed causes both precipitation and AQI. Besides, wind blows away and dilutes PM2.5 and other air pollutants so that it is a cause for AQI. As stated above, precipitation is the cause for AQI. The causal graph is shown in Fig. 11.

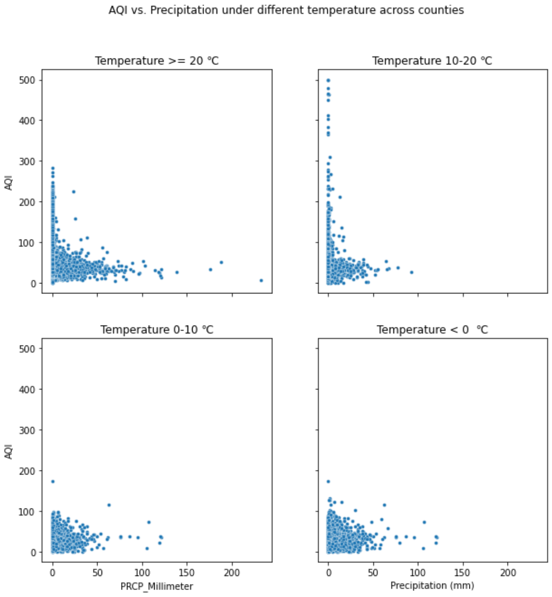


Figure 10: AQI vs. precipitation in different temperature ranges

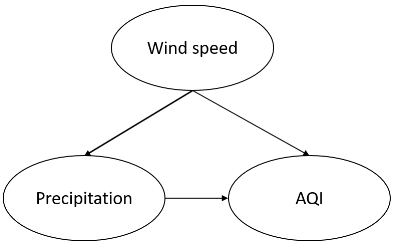


Figure 11: Causal graph w/ windspeed as confounder

Fig. 12 displays a similar distribution with the previous section, that is, on days with high precipitation (greater than 50 mm), AQI is almost always lower than 100, meaning that the air quality is good. On days with low precipitation , AQI can be low or high but there are more points distributed in the lower left. We conclude that precipitation is indeed a causation of AQI in a way that high precipitation leads to low AQI.

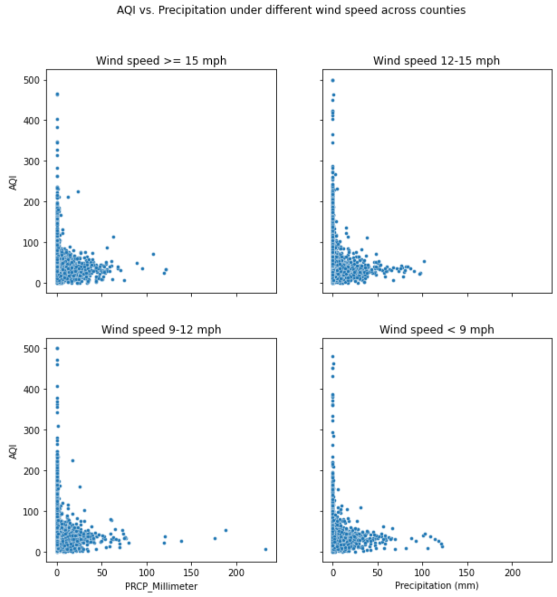


Figure 12: AQI vs. precipitation in different windspeed ranges

6 Interesting Findings

6.1 Individual Dataset

6.1.1 Global Weather

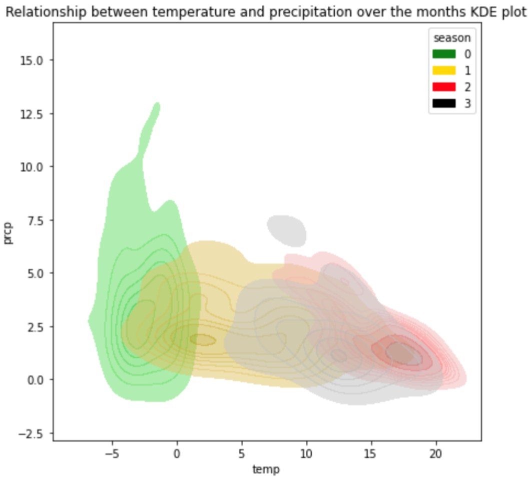


Figure 13: Relationship between temperature and precipitation over the months

From Fig. 13, we can see that there is a relationship between temperature and precipitation. Using the daily averaged temperature and precipitation through Jan 2020 to Oct 2020, we can see that as the temperature increases, the amount of precipitation tends to decrease. We divided 10 months into four seasons and made a scatter plot and KDE plot to show their relationship. Specifically speaking, in Jan and Feb (winter months), when the temperature is cold, we tend to see larger precipitation more often, while near summer, as is shown in plots by red color, we tend to see the least precipitation. So, from this discovery, we can use temperature to infer the range of precipitation in the U.S. region. Moreover, if you bin the above scatter plot by temperature, we can see a different distribution of precipitation (in spring it is spread apart in the vertical axis, and in July and August it is more clustered near 1 in the vertical axis).

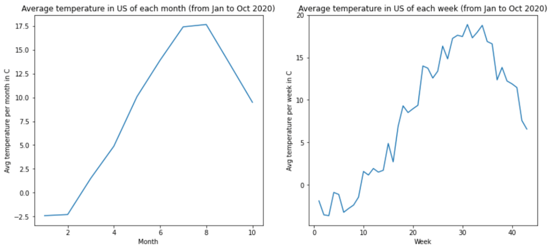


Figure 14: Average temperature in the U.S.

Removing the geographical information, we take the average temperature and precipitation across all regions in the U.S. Fig. 14 and Fig. 15 show the average temperature and precipitation in the U.S. each month and each week, respectively. There is a relationship between the month and the average temperature/precipitation for that month. This is due to how season transitions and although there is fluctuation in weekly based plot, it still keeps the general trend.

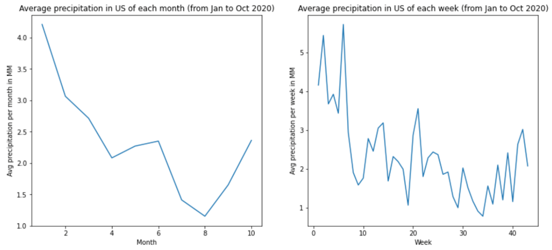


Figure 15: Average precipitation in the U.S.

6.1.2 Daily AQI Dataset

Seasonal Trends for a Region and across Regions There is a strong seasonal trend for AQI in each geographic region. For example, California always has the worst air quality in autumn (around August and September) from 2015 to 2020. One reason behind it could be wildfire, although we do not import data about wildfire season to further verify it. Besides, trends in different regions are distinct, even the opposite. New York State has the best air quality in autumn, in sharp contrast to California's trend.

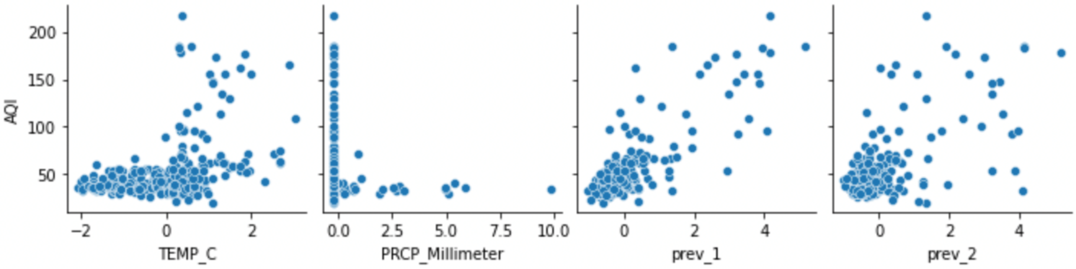


Figure 16: Features vs. AQI

Defining Parameter for AQI PM2.5 is a very popular term among the public and receives much attention. Although it contributes a lot to AQI, ozone is the primary parameter for defining AQI in all counties year-round. Over $175,000$ site-days report ozone as a major air pollutant, where site-day means a monitoring site on a day.

6.2 Cross Datasets

Using the time as the index, we aggregated the Global Weather and AQI datasets and combined the columns together.

Fig. 16 shows a feature's relationship with respect to the daily AQI. The $TEMP\_C$ is a relatively good predictor, as it has a positive linear relationship. But the $PRCP\_Millimeter$ is not a strong predictor, because when we have little precipitation, the AQI varies a lot. Judging by the trend for the yesterday's ($prev\_1$) and the day before yesterday's AQI ($prev\_2$), we will definitely include those in the final model, as we assume that the environment will not have a extreme change and consequently it is reasonable to use AQI value in past few days to predict today's AQI.

Fig. 17 shows a feature's relationship with respect to each other. We can see that $prev\_1$ and $prev\_2$ are highly correlated. So, essentially we can use one column to reach the same effect as the other column. Also from the above two explorations, we can see that there are variables that are not too helpful with the regression task such as precipitation.

7 Evaluation and Limitations

Besides the advantages of simple and easy to compute, our models also have several limitations. Within our capabilities, we have gathered datasets that contain information including temperature, precipitation and AQI on daily bases. However, through literature review, there are other features such as air turbulence, traffic activity, special activity (wildfire and fireworks) that could also influence the AQI in a significant way. Unfortunately, we cannot find a relevant dataset that has the desired granularity. An annual report is way too generalized for our application. So, if we were able to gather the above features, then we believe we could further improve our model's fit.

Also, we are conducting this research topic on a daily scale. Even though a smaller granularity can retain most information of raw data, it also leads to an inevitable overfitting, which may decrease the accuracy of the test set. Besides that, in 2020 there were several wildfires in CA, which could lead to the unusually sharp curves in the summertime. Even though these wildfire did not happen in Alameda, there is a great chance that they did influence air quality, adding uncertainty to the data and making it difficult for our model to predict. If we could find datasets regarding these special events in timestamps and take this information into the model, our fit could be better potentially.

8 Conclusion and Future Work

In conclusion, we build multiple models and carry out inference on AQI datasets. We also design experiments to compare performances of different regression models and find that the multi linear model produce best results. As for inference, we . Our models and inference are useful in predicting next day's AQI. We also discussed the impacts on social ad ethical issues, limitations of our work and possible improvements.

8.1 Potential Societal Impacts and/or Ethical Concerns

Having breathable fresh air for future generations is important, but as the power sector continues to grow and global greenhouse gas emissions increase, we are observing a decline in

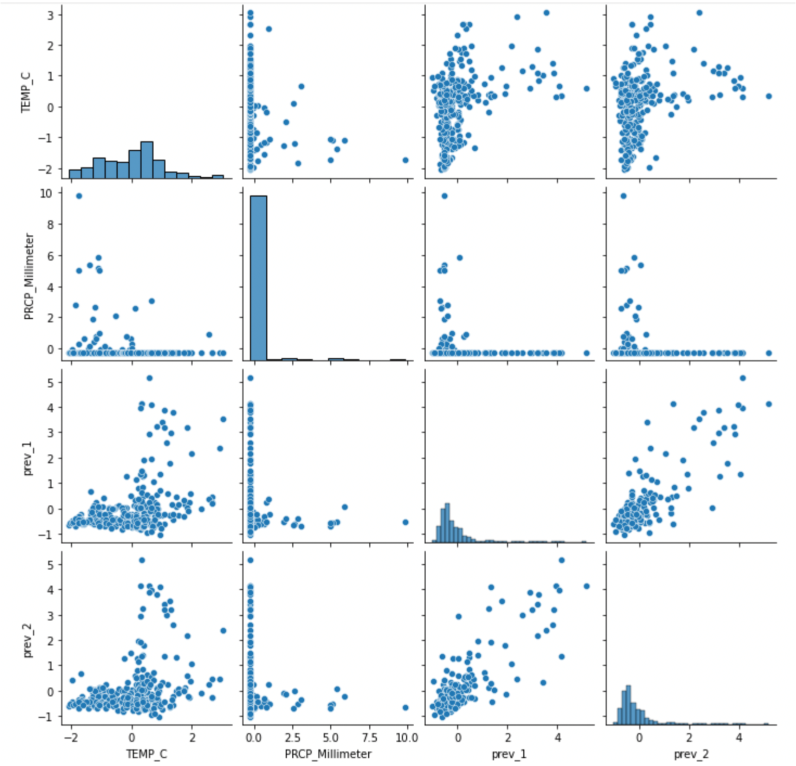


Figure 17: Features pair plot

air quality in some of the world's major cities. Therefore, having an accurate AQI prediction model is critical because it allows government officials and regulators to plan for societal developments such as controlling emissions from power plants, the amount of traffic on the roads, and after-hours electricity use. In addition, an accurate model can yield insights into how each factor affects the AQI. This information would allow interested parties to develop measures to achieve better air quality with a higher success rate. On the other hand, from an ethical point of view, as humans we should be responsible for the greenhouse gas emissions and pollution we produce. Predictive AQI research makes us more aware of how much these factors affect the air we breathe, and pushes each of us to make changes to keep the air healthy and clean. In terms of data collection and analysis, we see no ethical issues with this project. We are using a public dataset from the U.S. government and the data entries do not contain sensitive personal information.

8.2 Future Work

Using a sliding window approach, the current features and models could capture the regression problem on air quality to a certain extent, but not a hundred percent. Some other features such as the level of ongoing traffic, air turbulence could also influence the air quality. However, we cannot find datasets with the same level of granularity as the datasets selected in this project, so we do not cover these features. In the future, we could conduct field research and gather this information by ourselves, following the principles of data science.

Try other granularity in terms of time. In this project, every row in the final dataframe is a daily summary. Although it can mostly retain the preliminary information, it can cause overfitting when we train the models to a certain extent. We could try weekly or monthly to generalize the main trend of AQI and at the same time keep the errors on an acceptable level.

Take previous years' data into our models. We only focus on the 2020 data in this project, which is not very sufficient to train a robust model. If we contain the past 5 years' climate data, the model can learn to be more accurate when doing predictions. If we do so, it is important to select models that could model the cyclic pattern to better fit the periodic pattern in the AQI curve.In the future, we could try using a designed sine/cosine lifting map on our feature.

Acknowledgements

The authors would like to thank Prof. Fernando Perez, Prof. Alvin Wan, and all GSI's and TA's

for their help.

References

[1] Environmental Protection Agency. Air quality and climate change research, 2021.

[2] Huixiang Liu, Qing Li, Dongbing Yu, and Yu Gu. Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Applied Sciences*, 9(19):4069, 2019.

[3] Ni Sheng and U Wa Tang. The first official city ranking by air quality in china—a review and analysis. *Cities*, 51:139–149, 2016.

[4] National Weather Service. Clearing the air on weather and air quality, 2021.

[5] Air Quality Assessment Division. Technical assistance document for the reporting of daily air quality - the air quality index (aqi). Technical report, U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC, September 2018.

[6] Moritz Hardt. Lecture 13: Causal inference i, 2020.