

Springboard Data Science Capstone #3: **Forecasting Monthly Median AQI Levels in the US**

Kelly Pham


February 5th, 2023

Problem Statement:

Given daily AQI data from the USA, can we forecast the monthly average air quality for the next year? Is there a trend in AQI? Moreover, are there correlations between air quality, and features such as latitude, longitude, or population density? What is the most common defining parameter for the most polluted cities?

AQI or Air Quality Index is the primary way to measure the current quality of the air. AQI values range from 0-500 with 0 being perfectly healthy and 500 being extremely hazardous. However, it is possible for values to exceed 500. AQI levels past this range is called 'Beyond the AQI'.

Air Quality Index		
0-50	Good	Enjoy your usual outdoor activities.
51-100	Moderate	Extremely sensitive children and adults should refrain from strenuous outdoor activities.
101-150	Unhealthy for Sensitive Groups	Sensitive children and adults should limit prolonged outdoor activity.
151-200	Unhealthy	Sensitive groups should avoid outdoor exposure and others should limit prolonged outdoor activity.
201-300	Very Unhealthy	Sensitive groups should stay indoors and others should avoid outdoor activity.
301-500	Hazardous	Everyone should avoid all outdoor exertion.



AQI values are derived from moving averages/current values of PM2.5 (particulate matter), PM10, Ozone, Carbon Monoxide, Sulfur Dioxide, and Nitrogen Dioxide levels.

We humans all rely on clean air to survive. Monitoring air quality is important because polluted air can be bad for our health—and the health of the environment; increases in air pollution have been linked to decreases in lung function and increases in heart attacks. High levels of air pollution according to the EPA Air Quality Index directly affect people with asthma and other types of lung or heart disease.

The Dataset:

The data is a table of daily AQI values from stations across the US from 1980-2022 acquired from kaggle.com (<https://www.kaggle.com/datasets/calebreigada/us-air-quality-1980present>). It contains 5.72M rows and 15 columns. The columns are as follows:

- **Index**
- **CBSA Code** - CBSA = a U.S. geographic area defined by the Office of Management and Budget (OMB) that consists of one or more counties.
- **Date** - The day of measurement.
- **AQI** - The average air quality index (AQI) value for the day.
- **Category** - The category of air quality ranging from "Good" to "Hazardous".

- **Defining Parameter** - One of PM2.5 (particulate matter), PM10, Ozone, Carbon Monoxide, Sulfur Dioxide, or Nitrogen Dioxide which has the highest concentration.
- **Number of Sites Reporting** - The number of stations used to make the data aggregation.
- **city_ascii** - Name of the city where the measurement was taken.
- **state_id** - Abbreviation of the state where the measurement was taken.
- **state_name** - The state where the measurement was taken.
- **lat** - The latitude where the measurement was taken.
- **lng** - The longitude where the measurement was taken.
- **population** - The population of the region where the measurement was taken.
- **density** - The population per square kilometer where the measurement was taken.
- **timezone** - The time zone of the region where the measurement was taken

Data Wrangling and EDA:

There are no missing values.

We converted the 'Date' column into a datetime object to prepare the data for time series analysis. The 'timezone' column initially had 17 unique values, but they are repetitive. For example, 'America/New York' and 'America/Chicago' are the same time zones. We regroup them to just 7 distinct values: Pacific, Mountain, Central, Eastern, Puerto Rico, Hawaii-Aleutian, and Alaska.

Outlier detection:

Average AQI is 46.67 (Good), median is 41, but the maximum value is 20646.

We investigate the top columns with the largest AQI values:

	CBSA Code	Date	AQI	Category	Defining Parameter	Number of Sites Reporting	city_ascii	state_id
3334599	13860	2001-05-02	20646.0	Hazardous	PM10	6	Bishop	CA
3034632	13860	2003-02-02	16515.0	Hazardous	PM10	8	Bishop	CA
1742171	13860	2011-12-01	13276.0	Hazardous	PM10	14	Bishop	CA
3334516	13860	2001-02-08	12056.0	Hazardous	PM10	6	Bishop	CA
3334515	13860	2001-02-07	10856.0	Hazardous	PM10	6	Bishop	CA

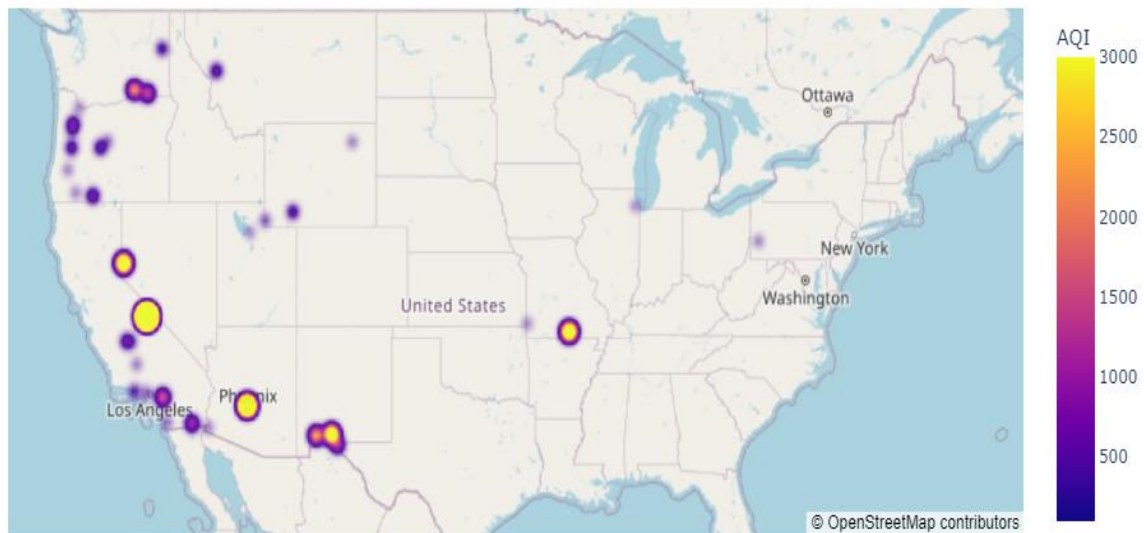
The top 5 rows all happen to be in Bishop, CA, and mostly in the early 2000's.

A Google search lead to an LA Times article from 2001 called '*7 Wildfires Burn 80,000 Acres*'. The article states that major fires were reported in Nevada, some which were burning partially in

California, forcing evacuations. Bishop, CA borders Nevada. Perhaps these fires contributed to the abnormally high AQI values in Bishop, CA.

This begs the question: are certain regions more prone to AQI levels over 500?

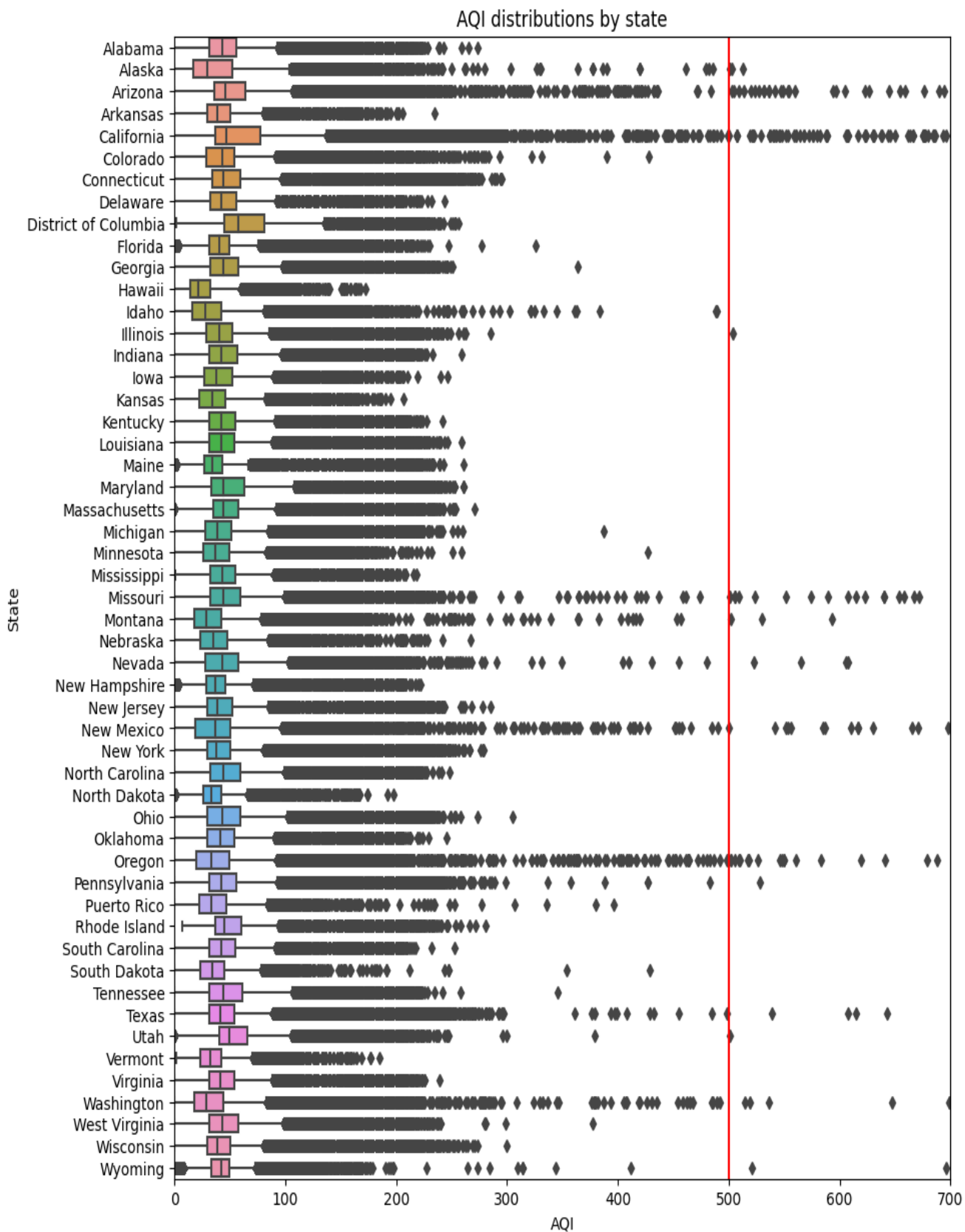
Places with AQI >500 from 1980-2022



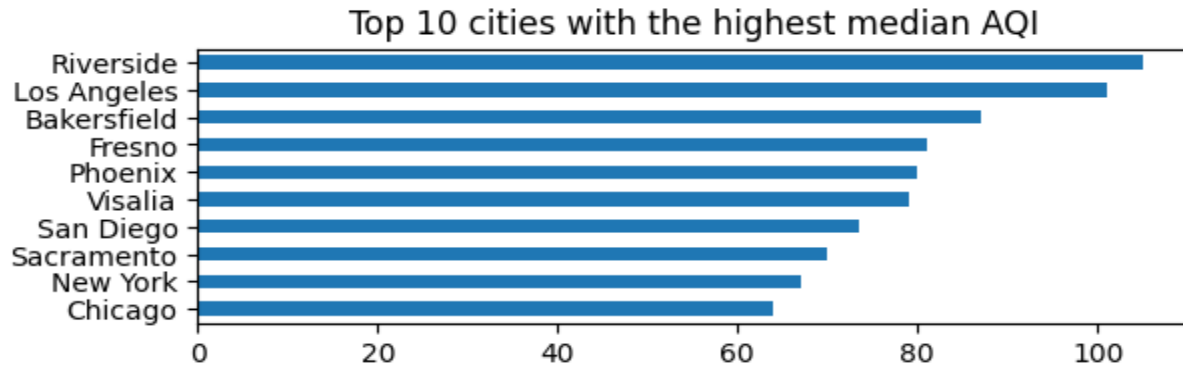
By observation, it seems likely; majority of the places are in the West and Southwest, which are drier areas prone to drought and wildfires. West Plains, MO, which is in the Midwest.

AQI distribution by state:

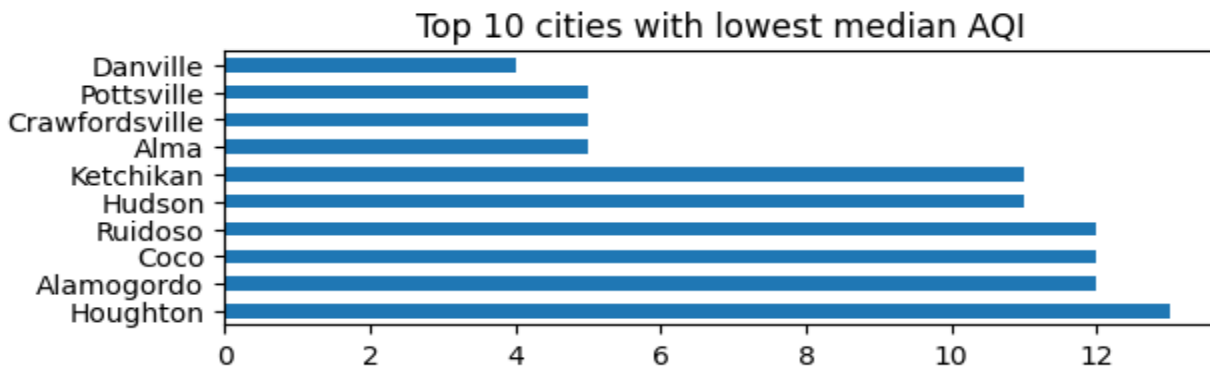
The next page is a visualization of the distribution of AQI levels by state, with the red line denoting the value 500; anything above that is beyond the AQI. The median AQI for all states are all below 100, and there are plenty of upper outliers in every state. There are a few lower level outliers in the states of Florida, Maine, Massachusetts, New Hampshire, North Dakota, Utah, Vermont, and Wyoming that would be worth analyzing in the future as well.



Analysis of Regions/States and Median AQI:

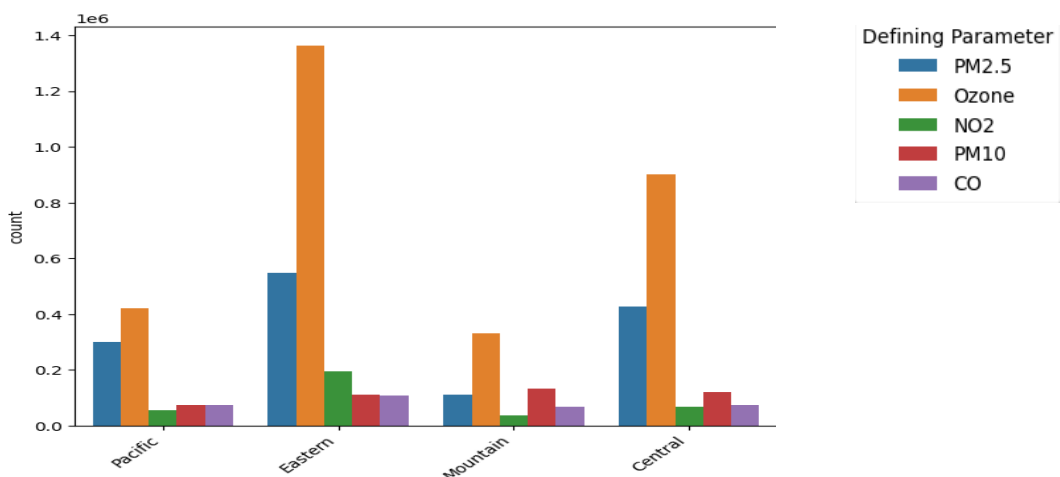


Seven out of the top 10 cities with the highest median AQI are from California. Furthermore, these cities are within the 50 largest cities in the state. Phoenix, New York, and Chicago are the largest cities in their states, respectively.



The 10 cities with the lowest median AQI are unknown cities with smaller populations than the top 10 list. The smallest city on this list is Ruidoso, NM, with a population density of 188 people /km². The largest is Pottsville, PA with density of 1255 people / km². None of these cities are in California.

Defining Parameters of Mainland USA by Region:



Ozone is the defining parameter in all the US, with PM2.5 in second place in all but the Mountain region.

Time Series Analysis and Modelling:

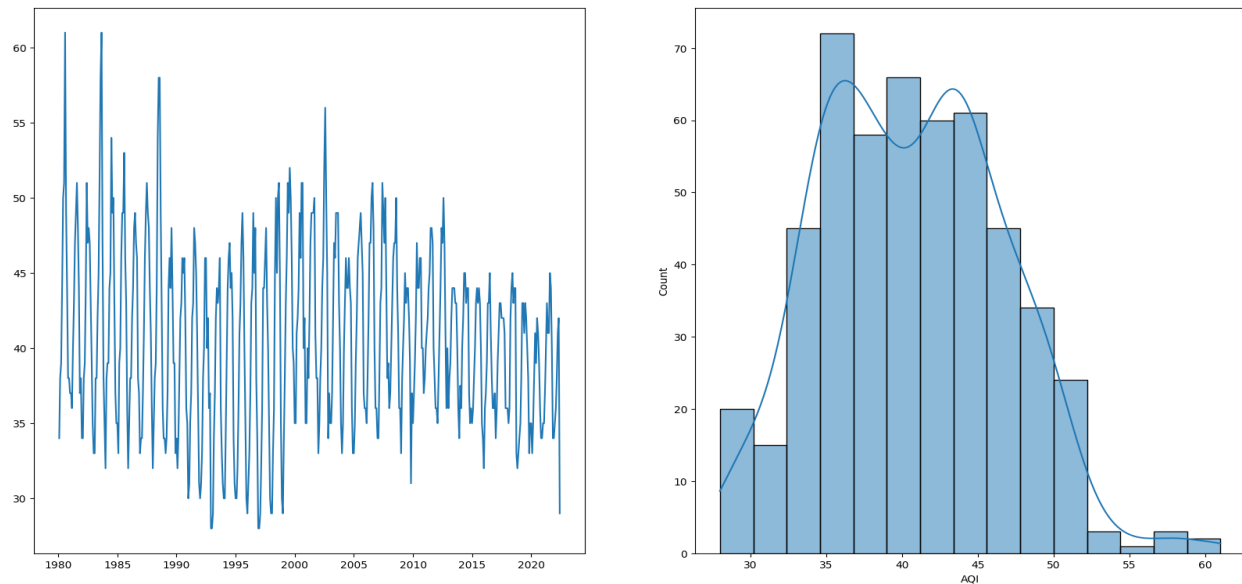
Now it is time to forecast the monthly AQI levels in the US. That is, we do the following:

1. Convert the data from daily to monthly via resampling with the median.
2. Drop all columns except for 'Date' and 'AQI'.
3. Take the last 12 data points to be the test set; the rest will be the train set.
4. Fit 3 models – SARIMA, Facebook Prophet, and LSTM.
5. Compare the models' RSME scores.
6. Select the best model.

SARIMA (p,d,q) x (P,D,Q)_s:

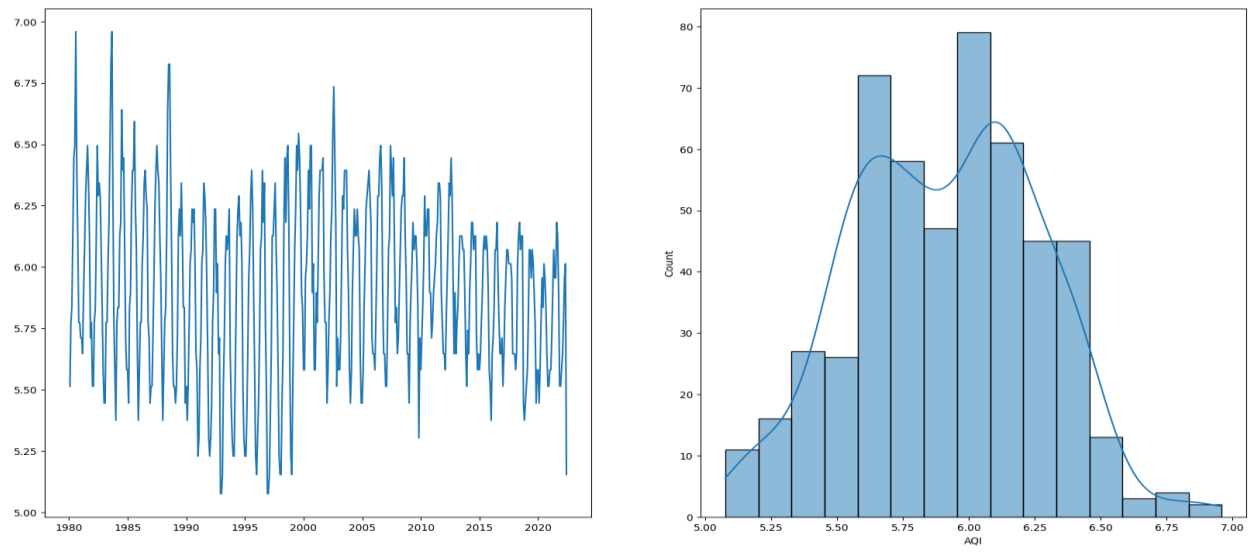
To find optimal p , d , q , P , D , Q , and s values for a SARIMA model, we first check if the data is normal and stationary by plotting it.

Untransformed Data and its Histogram

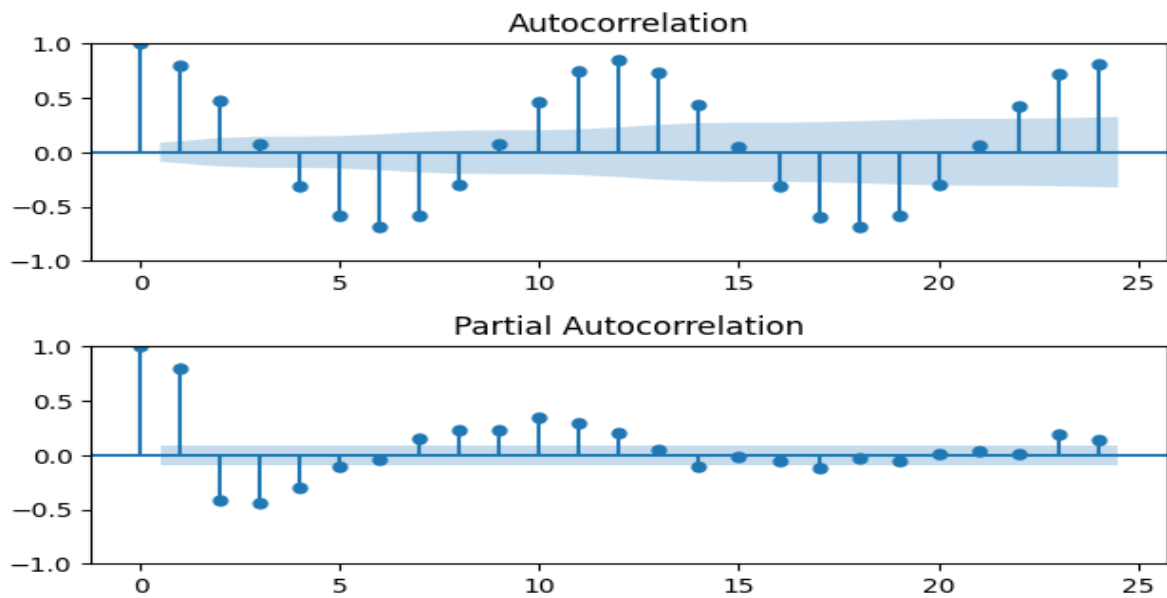


The histogram is right-skewed, so we apply a Box-Cox transformation to normalize it. We plot the transformed data and its ACF/PACF plots to check if it is stationary.

Optimal BoxCox-transformed Data and its Histogram



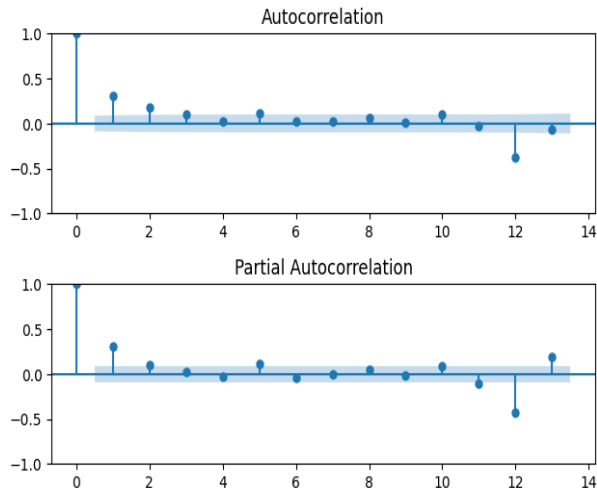
ACF and PACF plots for Box-Cox transformed data



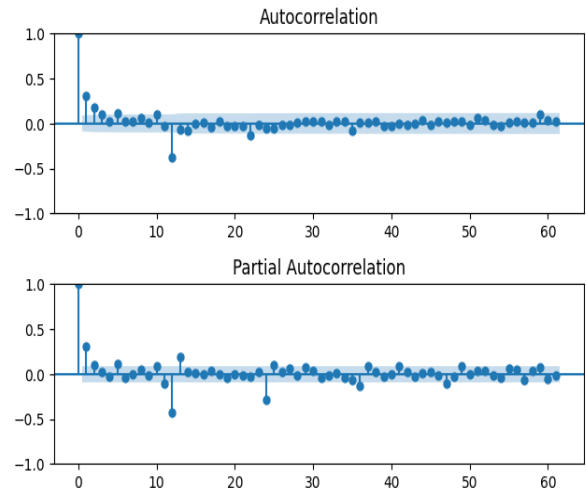
The ACF plot suggests yearly seasonality; we difference the data at lag 12 to make it stationary.

We plot and analyze the ACF and PACF graphs of this differenced data to pick optimal parameters for our model.

ACF and PACF plots for Box-Cox data with Differencing at Lag 12 [first 12 lags]



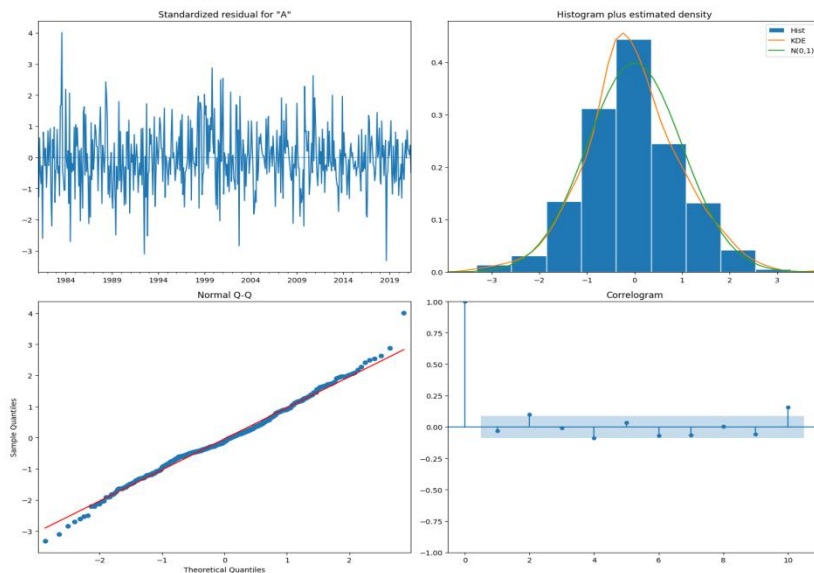
ACF and PACF plots for Box-Cox data with Differencing at Lag 12 [first 60 lags]



Based on the non-seasonal part of the ACF, it is best to try $q = 0, 1$, or 2 . The non-seasonal PACF suggests trying $q = 0, 1$, or 2 as well. The seasonal components of the graphs on the right hand side suggest to try $Q = 0$ or 1 and $P = 0, 1$, or 2 . We set $s = 12$ because of yearly seasonality and $D = 1, d = 0$ because we only took one seasonal difference.

Automating this process, the best values (values that give the lowest AIC score) are: $p = 2, q = 1, P = 0, Q = 1$. Its AIC score is -593.191 .

Now we perform diagnostics on the residuals:



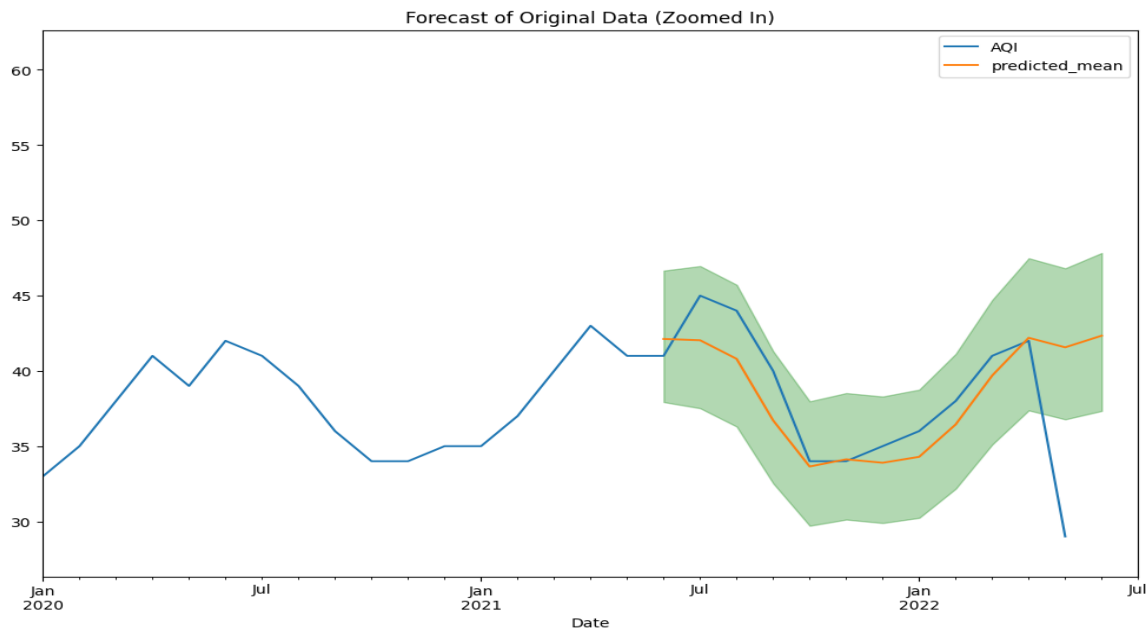
The value of the correlogram at lag 10 is nonzero, which means the residual does not represent white noise.

We try $q = 10$ instead to obtain a $SARIMA(2,0,10) \times (1,0,1)_{12}$ model with a lower AIC score of -600.159 .

However, the model is complex, and it fails the Ljung-Box test. Zero is within confidence interval of some of the coefficients, we set those to zero.

Tweaking the coefficients did allow the new model to pass the Ljung-Box test, but it increased the AIC score to -577.534 . Thus, we stick with the original and simpler $SARIMA(2,0,1) \times (1,0,1)_{12}$ model.

We now run the model on the test set and undo the Box-Cox transformation see how well the model did. The predictions on the untransformed test set are plotted below:



The predictions are all within the confidence interval; they are a good fit overall, except the very last point in the test set. Perhaps this is due to unforeseen circumstances, such as Covid19, weather changes, or the economy. The RMSE score for this model is 4.056.

Facebook Prophet:

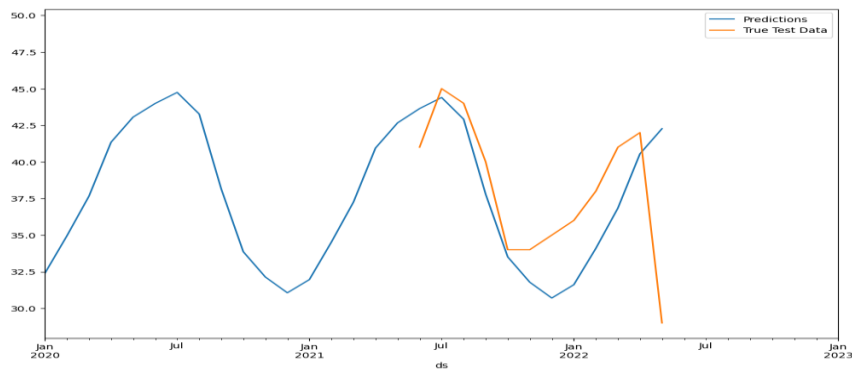
Since Prophet does not require stationary data, we will not transform the data. To find the best Facebook Prophet model, we tune the following hyper-parameters:

- *Changepoint_prior_scale* = 0.001, 0.05, 0.1, or 0.5
- *Seasonality_prior_scale* = 0.01, 0.1, 1.0, or 10.0

The *changepoint_prior_scale* hyper-parameter determines the flexibility of the trend, and its default value is 0.05. Values too small will under-fit the data, and values too large can over-fit. The second hyper-parameter is similar with default value of 10.0.

We perform cross-validation with an initial period of 30 years, and run the model every 90 days to predict the horizon of 360 days. This method implies that the best hyper-parameter values are *changepoint_prior_scale* = 0.5 and *seasonality_prior_scale* = 0.01.

The predictions on the test set are plotted below:



The predictions are similar to the ones from the SARIMA model – good fit except for the last point.

The RMSE score for this model is 2.578, which is better than the SARIMA model.

LSTM:

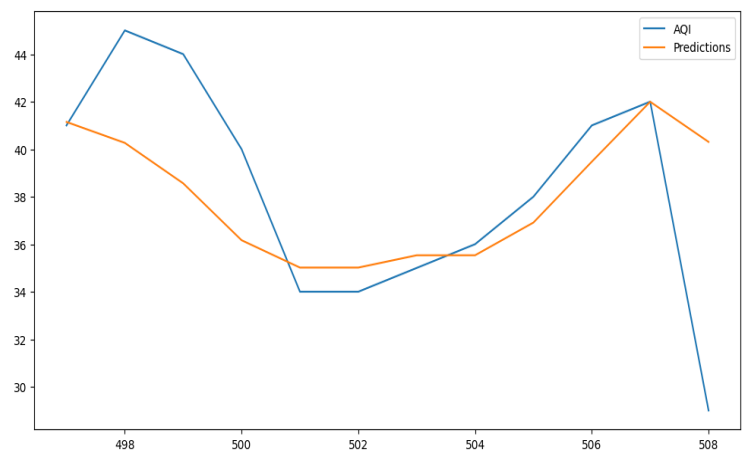
We preprocess the train data with the MinMaxScaler, and then fit the scaler on both the train and test sets. To select the best neural network, we build a sequential model and tune the following hyper-parameters:

- *Number of neurons:* 30-360, with step 30
- *Number of layers:* 1-4
- *Best dropout rate:* 0-0.5, with step 1
- *Activation:* relu or sigmoid

Running the hyper-parameters with 30 epochs, the following is the best model:

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 12, 90)	33120
lstm_1 (LSTM)	(None, 12, 240)	317760
lstm_2 (LSTM)	(None, 12, 210)	378840
lstm_3 (LSTM)	(None, 12, 30)	28920
lstm_4 (LSTM)	(None, 12, 30)	7320
lstm_5 (LSTM)	(None, 240)	260160
dropout (Dropout)	(None, 240)	0
dense (Dense)	(None, 1)	241

=====
Total params: 1,026,361
Trainable params: 1,026,361
Non-trainable params: 0



There are 6 LSTM layers, instead of the maximum of 4, as intended during the hyper-parameter tuning, so the code will need to be checked in the future.

The predictions of the LSTM model have the same issue as the other two models: the prediction of the last point was not accurate. The RMSE score for the LSTM model is 4.087, the highest of the three models.

Model Selection:

The table below shows the models and their RMSE scores.

SARIMA	4.056
Prophet	2.578
LSTM	4.087

Based on the RMSE scores, Facebook Prophet is the best model for forecasting this dataset.

Conclusion:

- EDA suggests that regions in the west coast, especially California, are more prone to extremely AQI values. More analysis is needed to determine whether region and state are correlated to AQI levels.
- Facebook Prophet is the most accurate model for this dataset, based on RMSE values.
- All 3 models predicted the last point of the test set (May 2022) inaccurately – this suggests unforeseeable events that cannot be captured by our models.

Further Work:

- Double check the code for the LSTM.
- Analyze the data by state to observe if the trends are different.