Forecasting Monthly Median AQI Levels in the US

Kelly Pham
Springboard, May 2022 Cohort

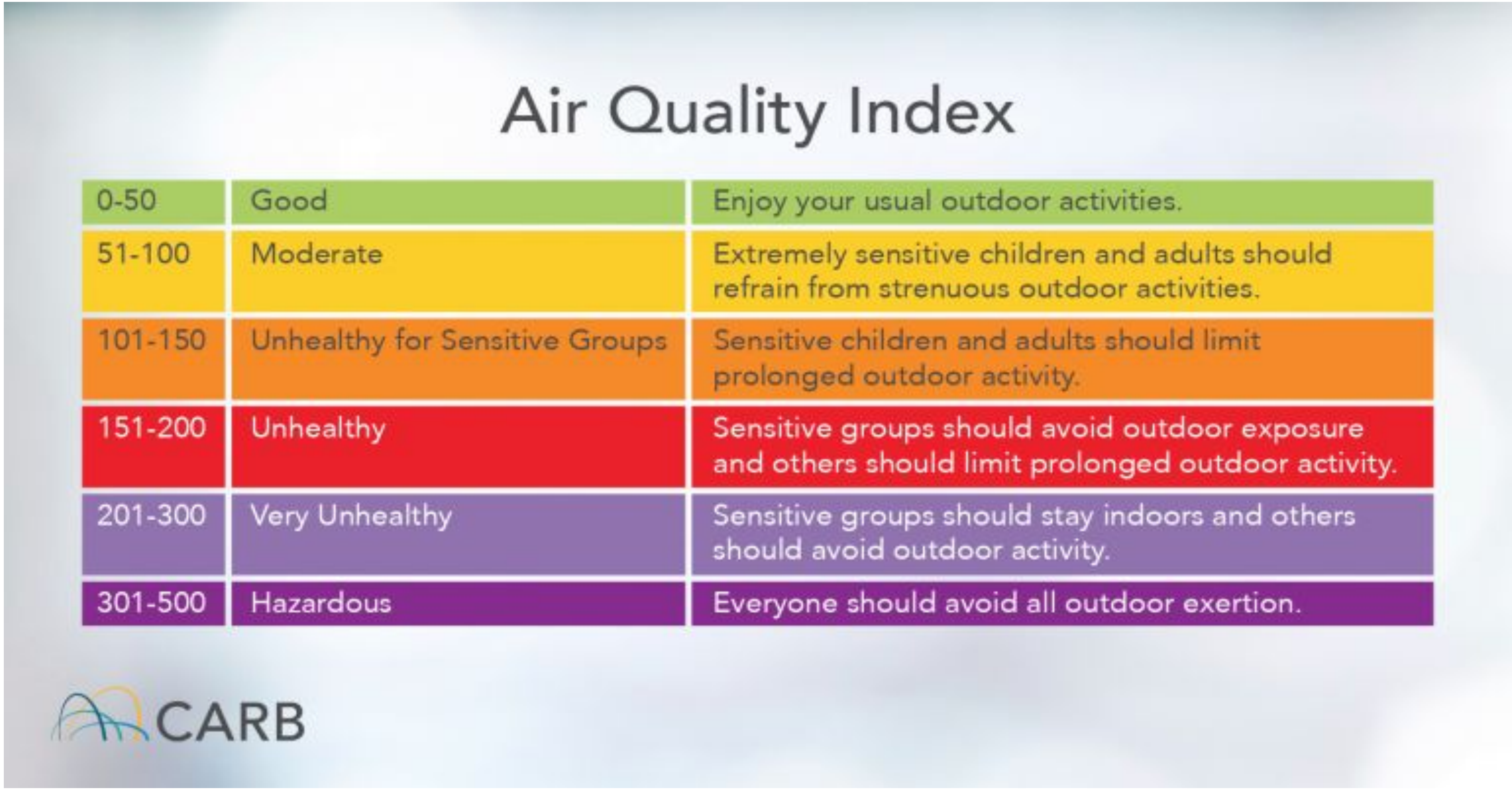
Problem Statement

- Given daily AQI data from the US, can we forecast the monthly median AQI levels within the next year?

- Are there correlations between AQI and features such as latitude, longitude, and population density?

Why care?

- **Air pollution can negatively affect health**
    - Linked to decreased lung function
    - May increase heart attacks
- **Especially for those with pre-existing conditions**
    - Asthma, lung disease
    - Heart disease
- **Climate change**



Air Quality Index

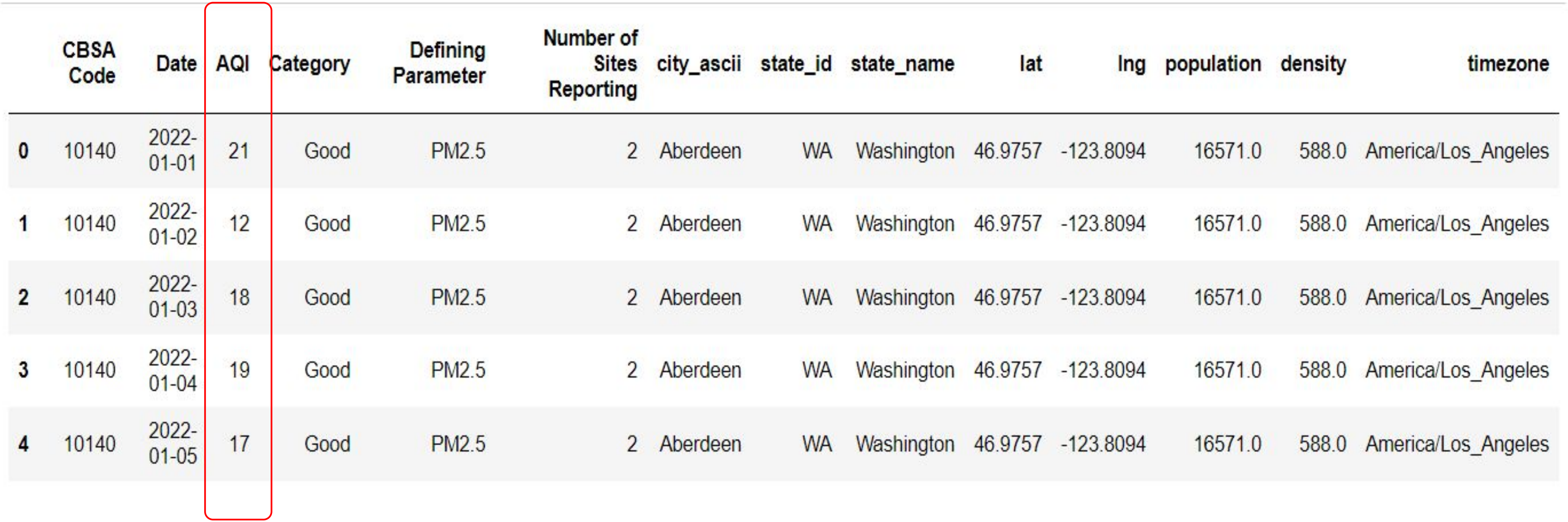
0-50	Good	Enjoy your usual outdoor activities.
51-100	Moderate	Extremely sensitive children and adults should refrain from strenuous outdoor activities.
101-150	Unhealthy for Sensitive Groups	Sensitive children and adults should limit prolonged outdoor activity.
151-200	Unhealthy	Sensitive groups should 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.

CARB

501+ : ‘Beyond the AQI’ – follow guidelines for ‘Hazardous’

The Data

- **Link:** https://www.kaggle.com/datasets/calebreigada/us-air-quality-1980present
- Table of Daily AQI values from stations across the US, 1980-2022
- Daily

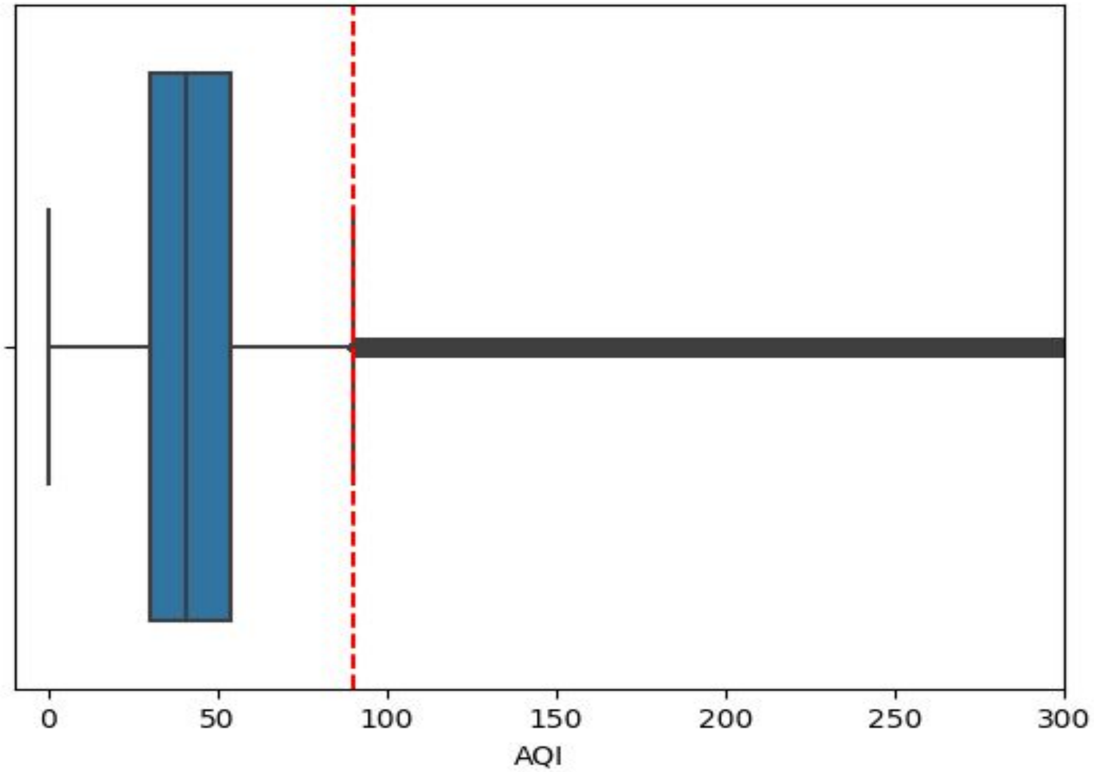


	CBSA Code	Date	AQI	Category	Defining Parameter	Number of Sites Reporting	city_ascii	state_id	state_name	lat	lng	population	density	timezone
0	10140	2022-01-01	21	Good	PM2.5	2	Aberdeen	WA	Washington	46.9757	-123.8094	16571.0	588.0	America/Los_Angeles
1	10140	2022-01-02	12	Good	PM2.5	2	Aberdeen	WA	Washington	46.9757	-123.8094	16571.0	588.0	America/Los_Angeles
2	10140	2022-01-03	18	Good	PM2.5	2	Aberdeen	WA	Washington	46.9757	-123.8094	16571.0	588.0	America/Los_Angeles
3	10140	2022-01-04	19	Good	PM2.5	2	Aberdeen	WA	Washington	46.9757	-123.8094	16571.0	588.0	America/Los_Angeles
4	10140	2022-01-05	17	Good	PM2.5	2	Aberdeen	WA	Washington	46.9757	-123.8094	16571.0	588.0	America/Los_Angeles

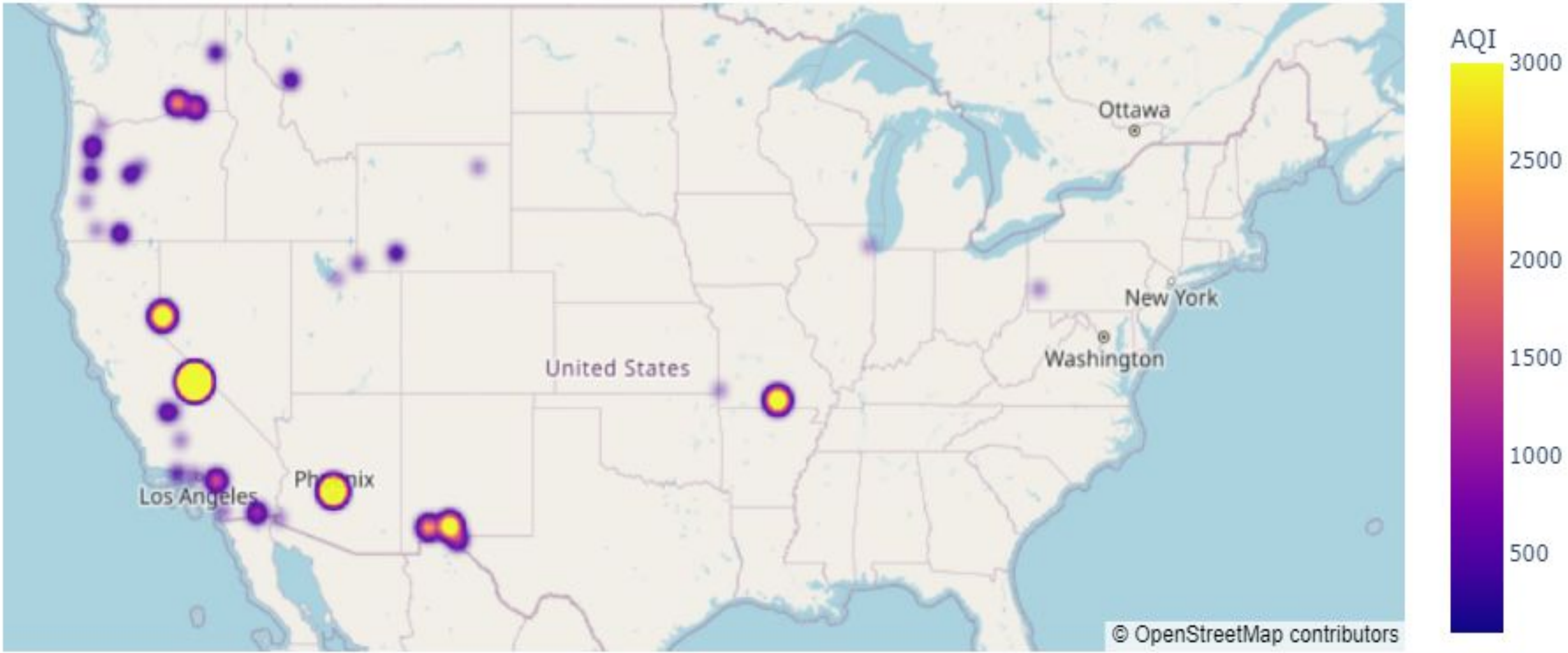
Exploratory Data Analysis

Boxplot of AQI values

- **Median:** 41
- **Minimum:** 0
- **75th-percentile:** 90
- **Lots of outliers!**



Places with AQI >500 from 1980-2022



Defining Parameters of mainland US by region

Defining Parameter

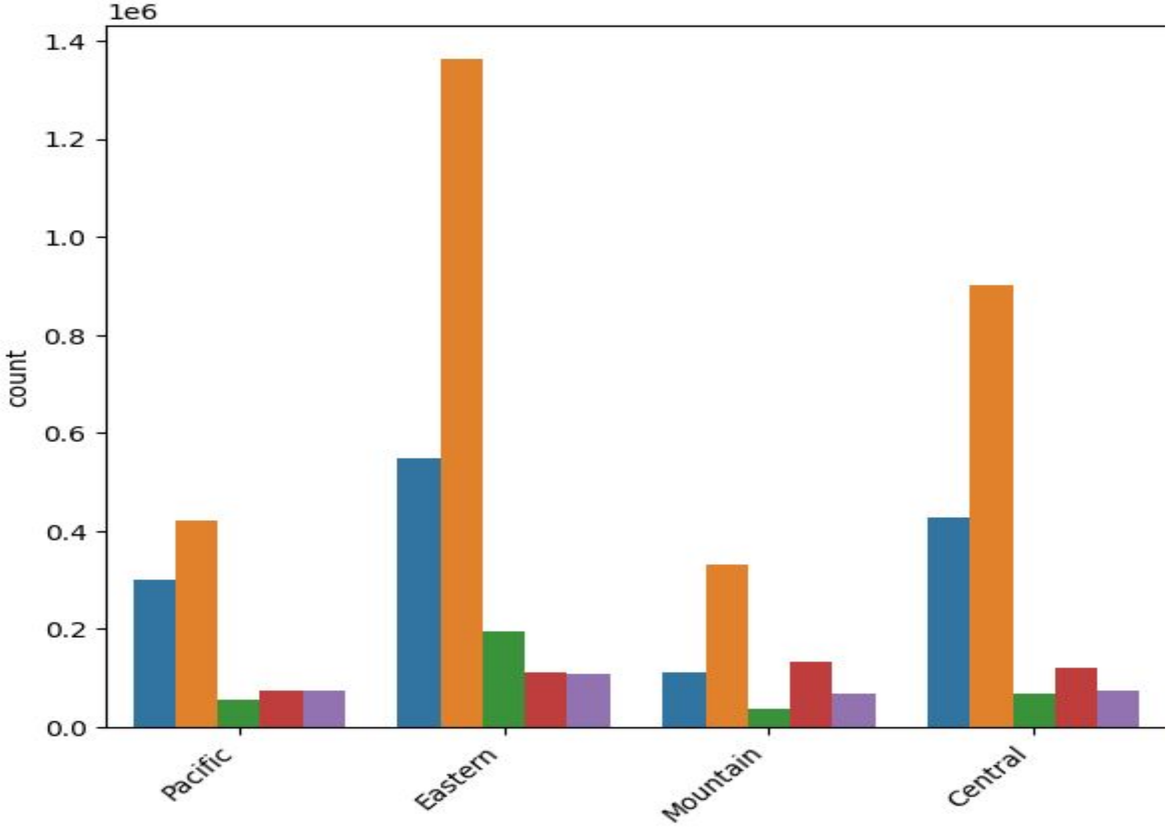
🟦 PM2.5

🟧 Ozone

🟩 NO2

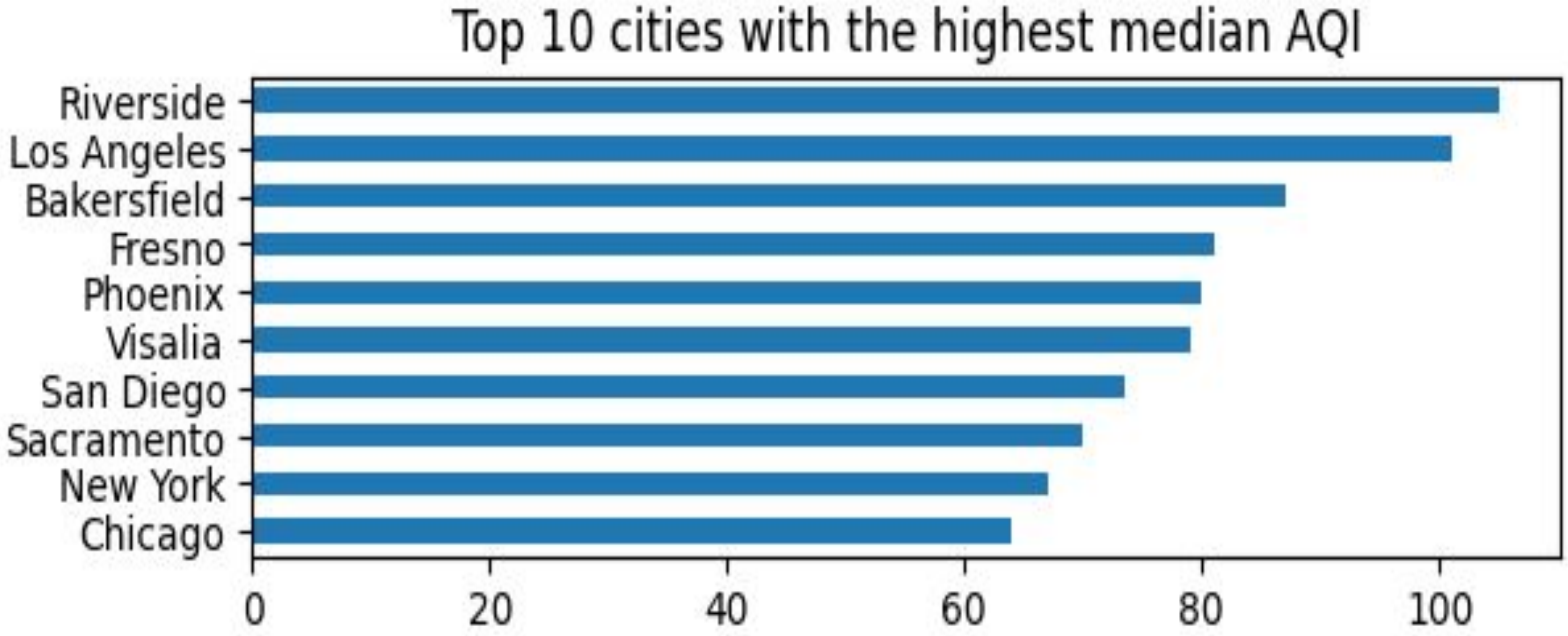
🟥 PM10

🟪 CO



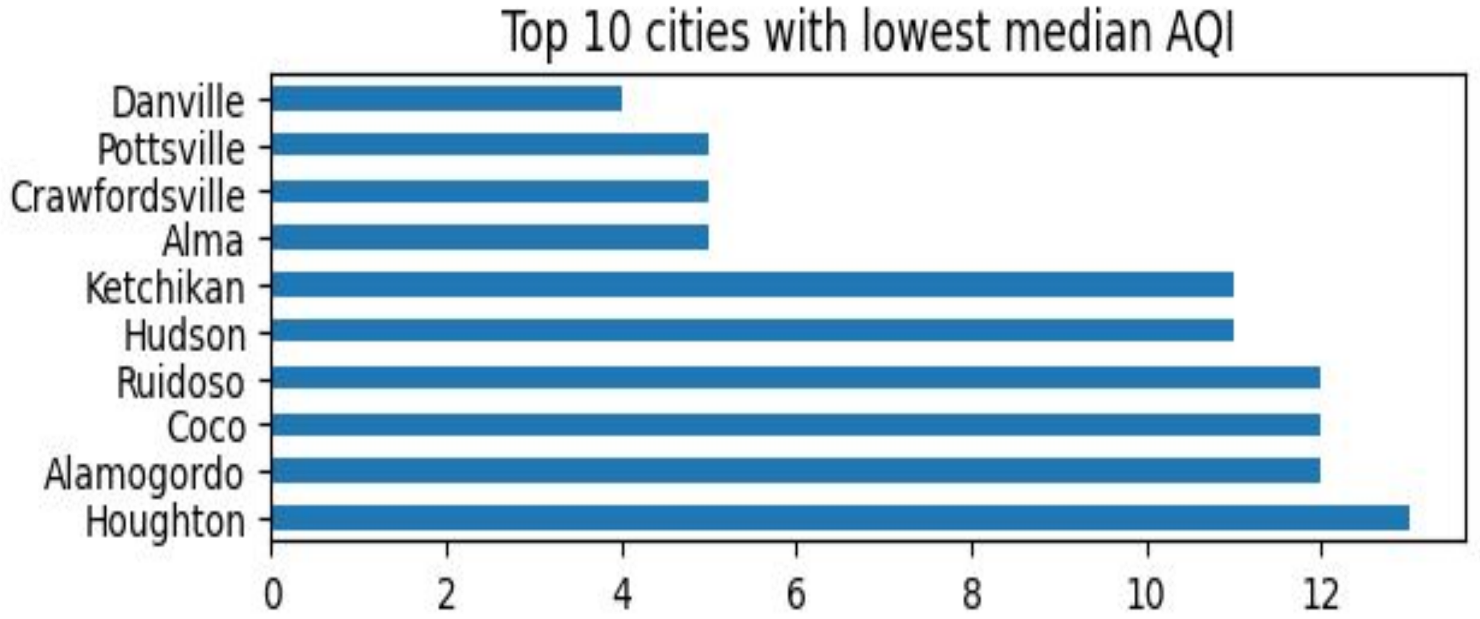
Ozone is #1 for all regions.

PM2.5 is #2, except for Mountain region



(CA)
(CA)
(CA)
(CA)
(AZ)
(CA)
(CA)
(CA)
(NY)
(IL)

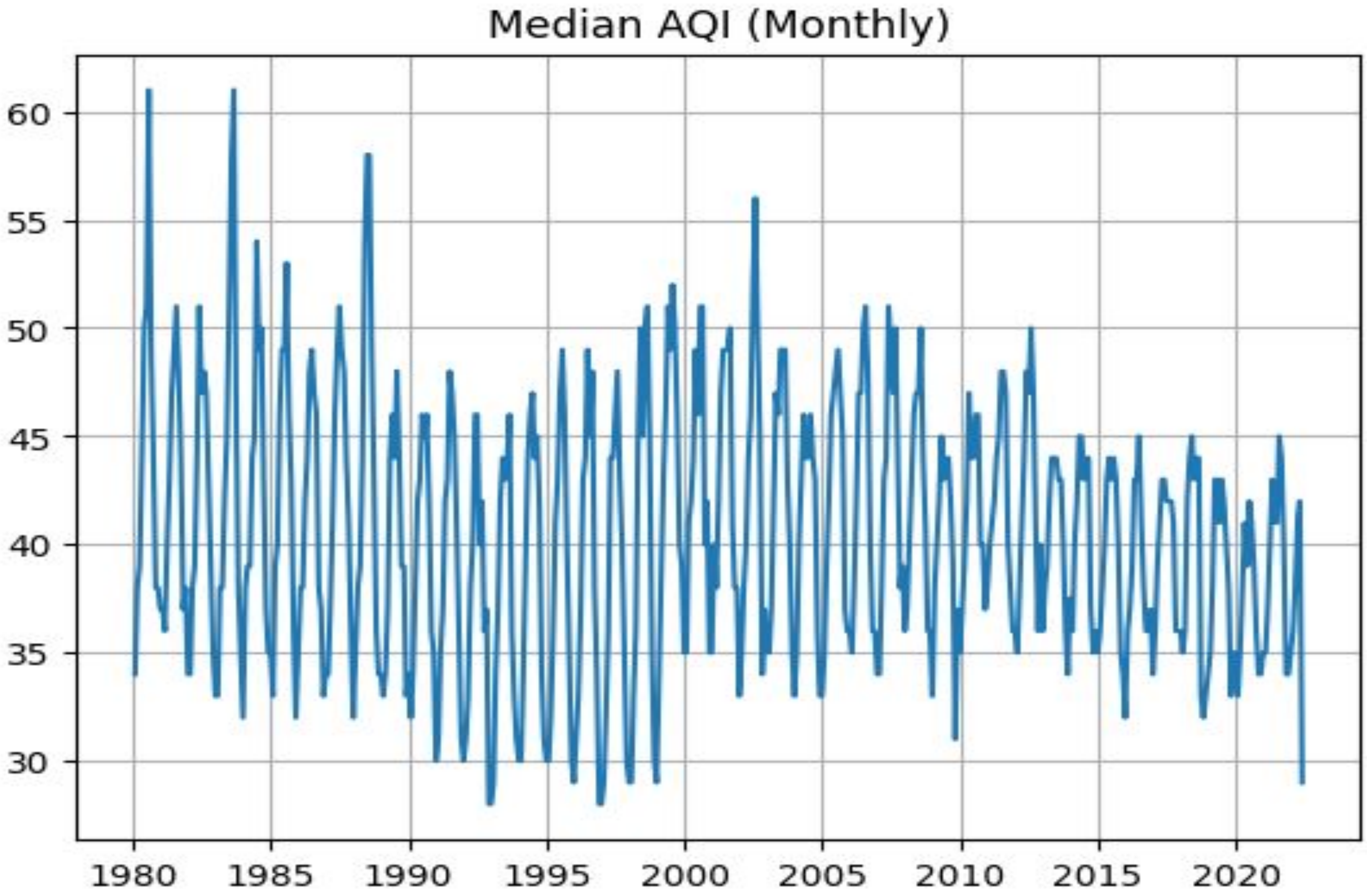
(KY)
(PA)
(IN)
(MI)
(AK)
(NY)
(NM)
(PR)
(NM)
(MI)



Time Series Analysis and Modelling

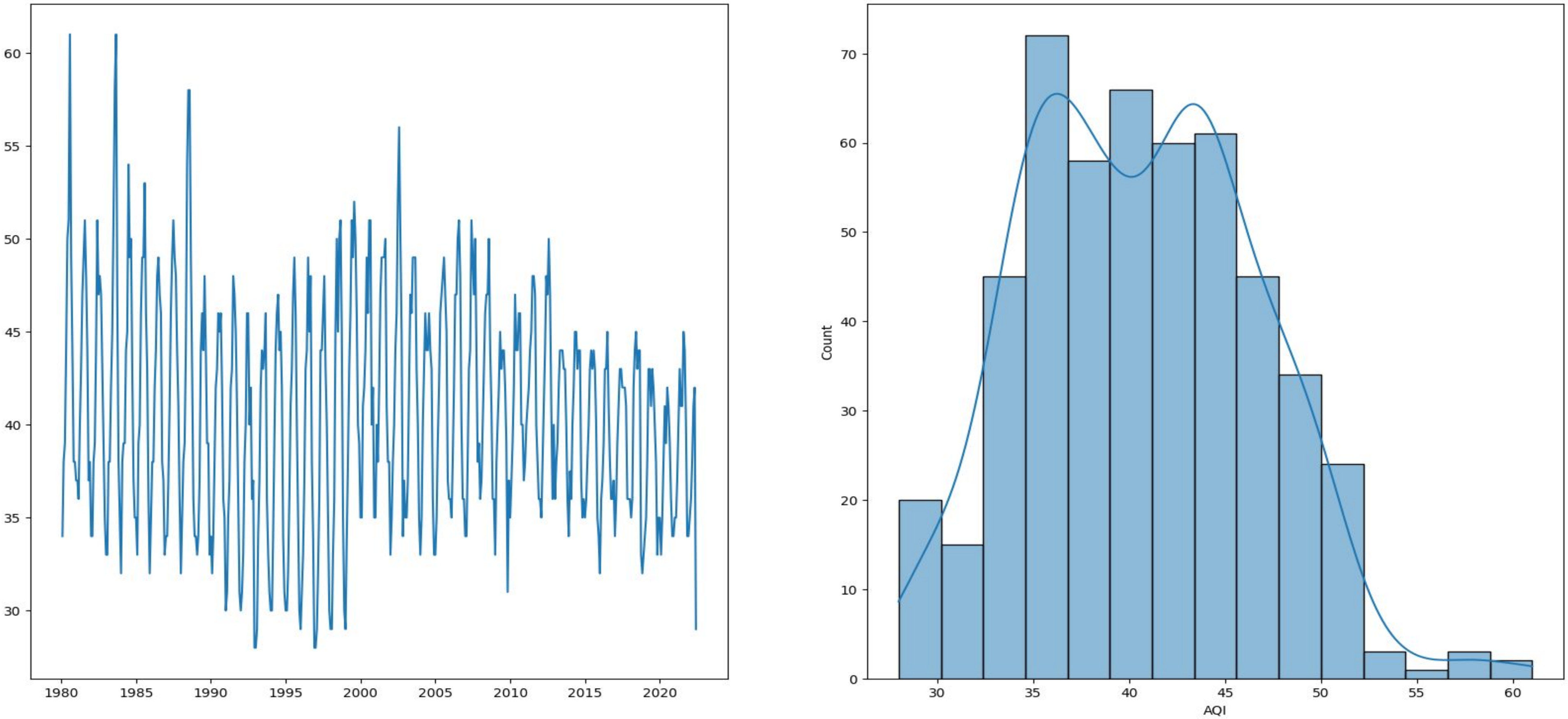
Steps

- Convert data from daily to monthly via resampling
- Train-test-split
- Fit **SARIMA** model
- Fit Facebook **Prophet** model
- Fit **LSTM** model
- Compare models’ RSME scores
- Select best model


Median AQI (Monthly)
60
55
50
45
40
35
30
1980
1985
1990
1995
2000
2005
2010
2015
2020

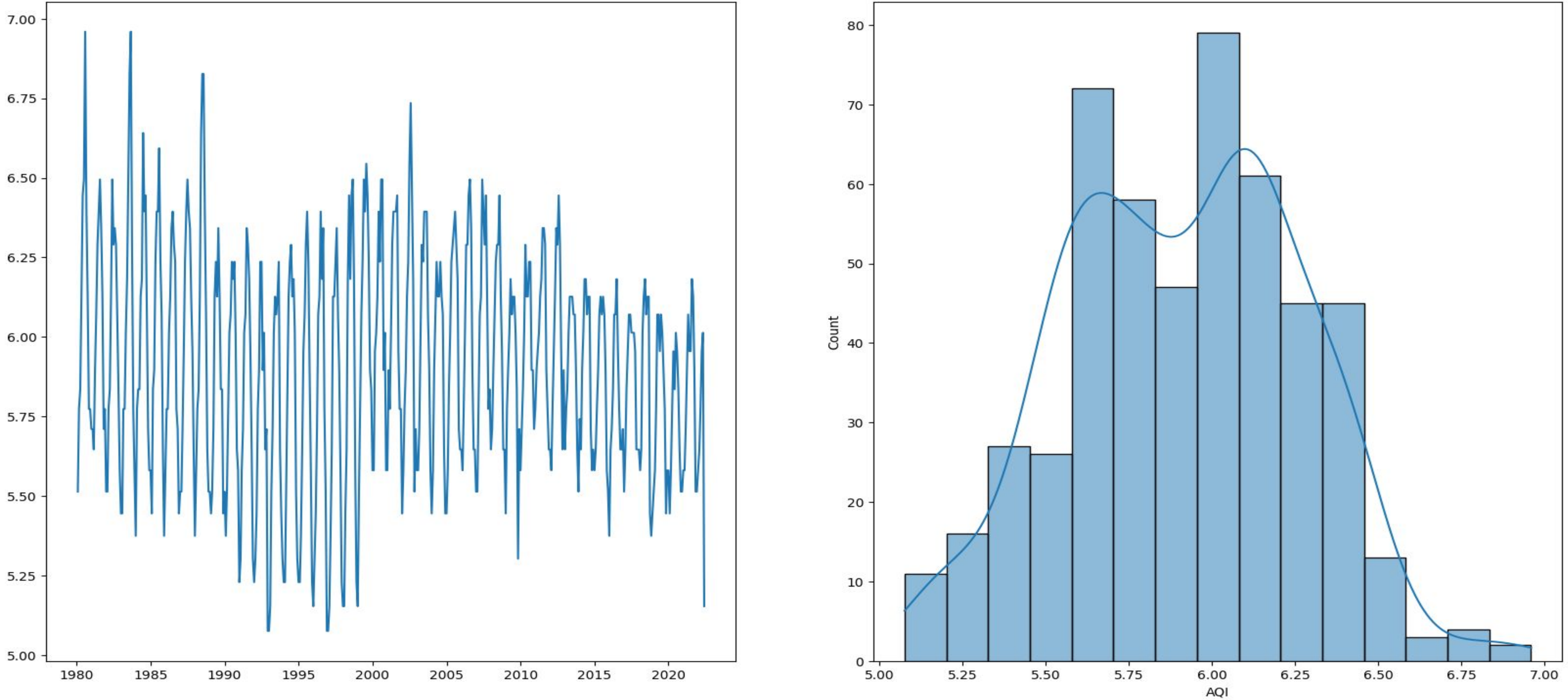
SARIMA (p,d,q) x $(P,D,Q)_s$ Modelling

Untransformed Data and its Histogram



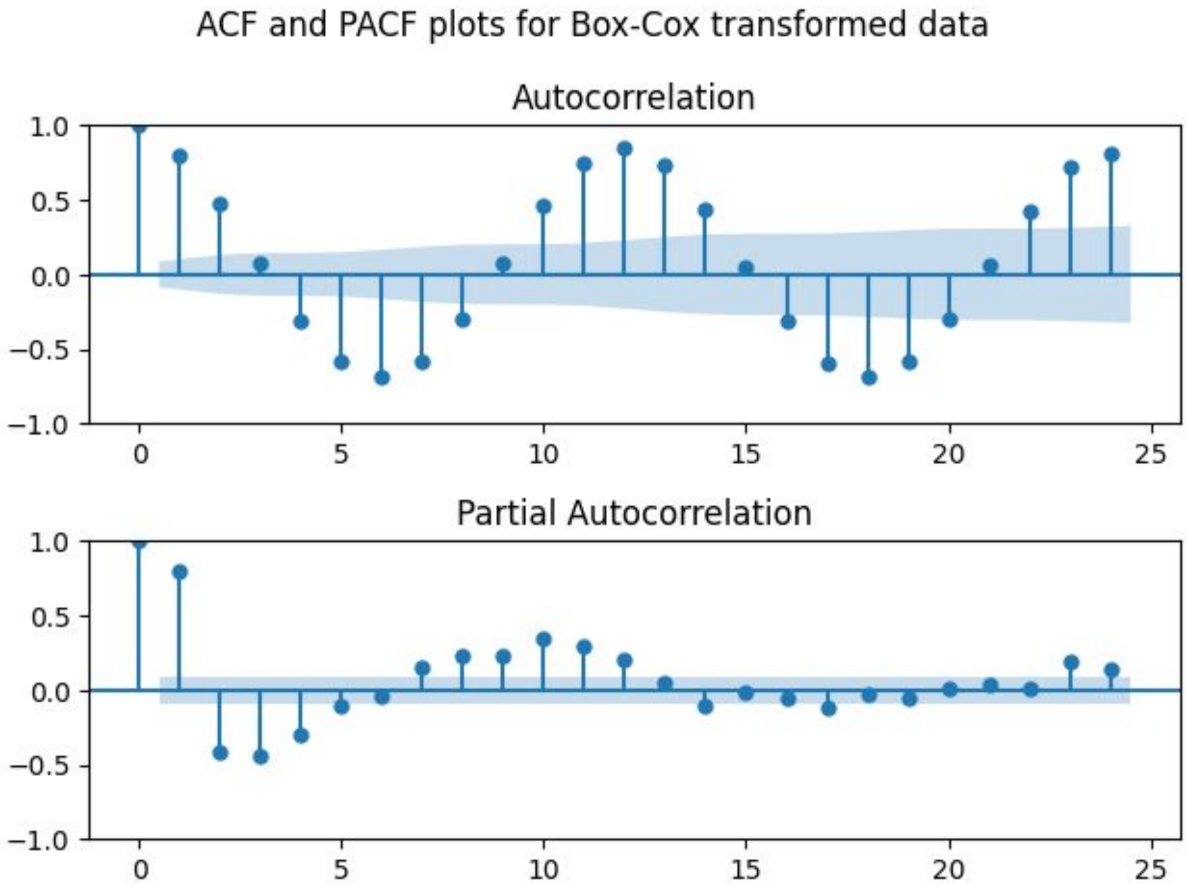
Data not normally distributed. Use Box-Cox transformation to normalize for SARIMA model fitting.

Optimal BoxCox-transformed Data and its Histogram



Box-Cox transformation with lambda = 0.237

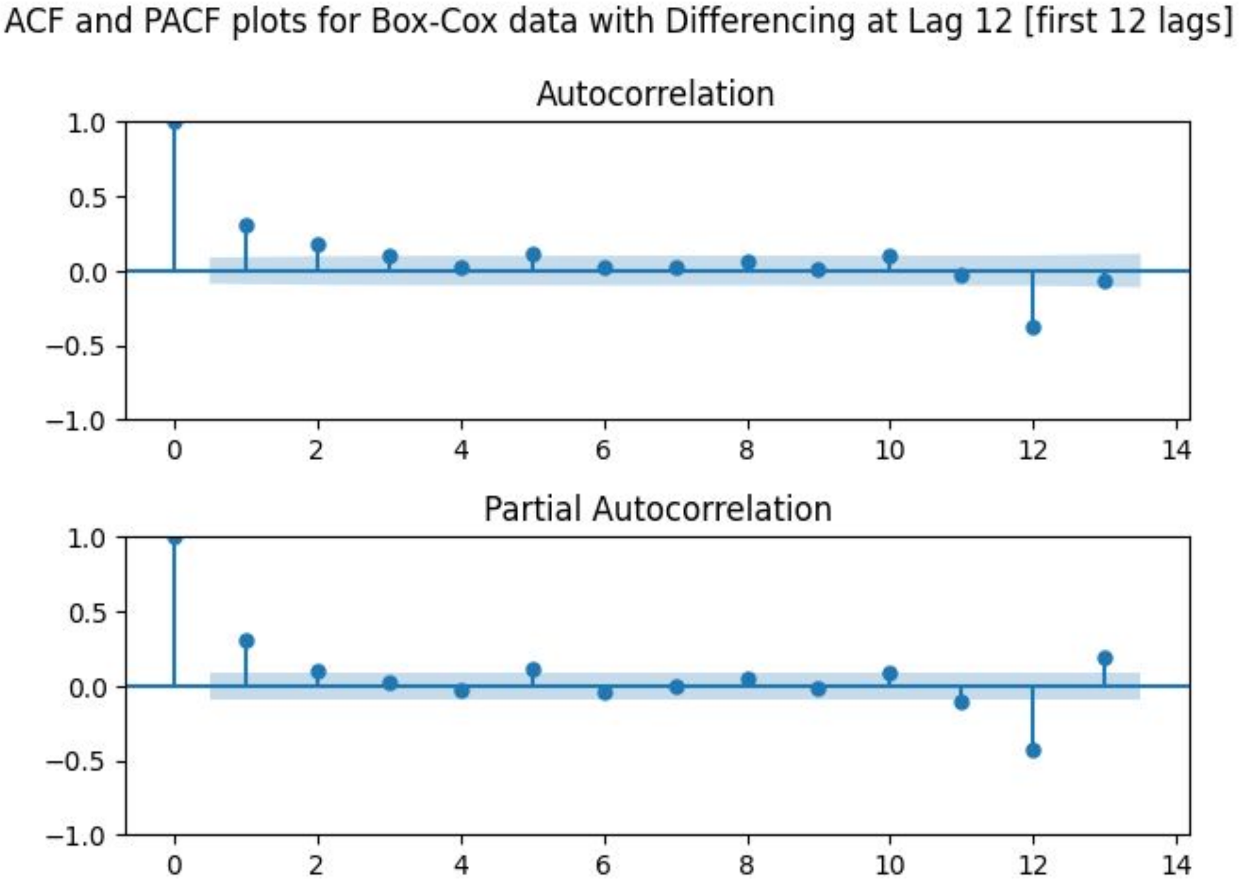
- ACF suggests monthly seasonality
    - Try **s=12**
- Take 1 difference at lag 12
    - Try **D=1**



- No more differencing
  - Try **d = 0**

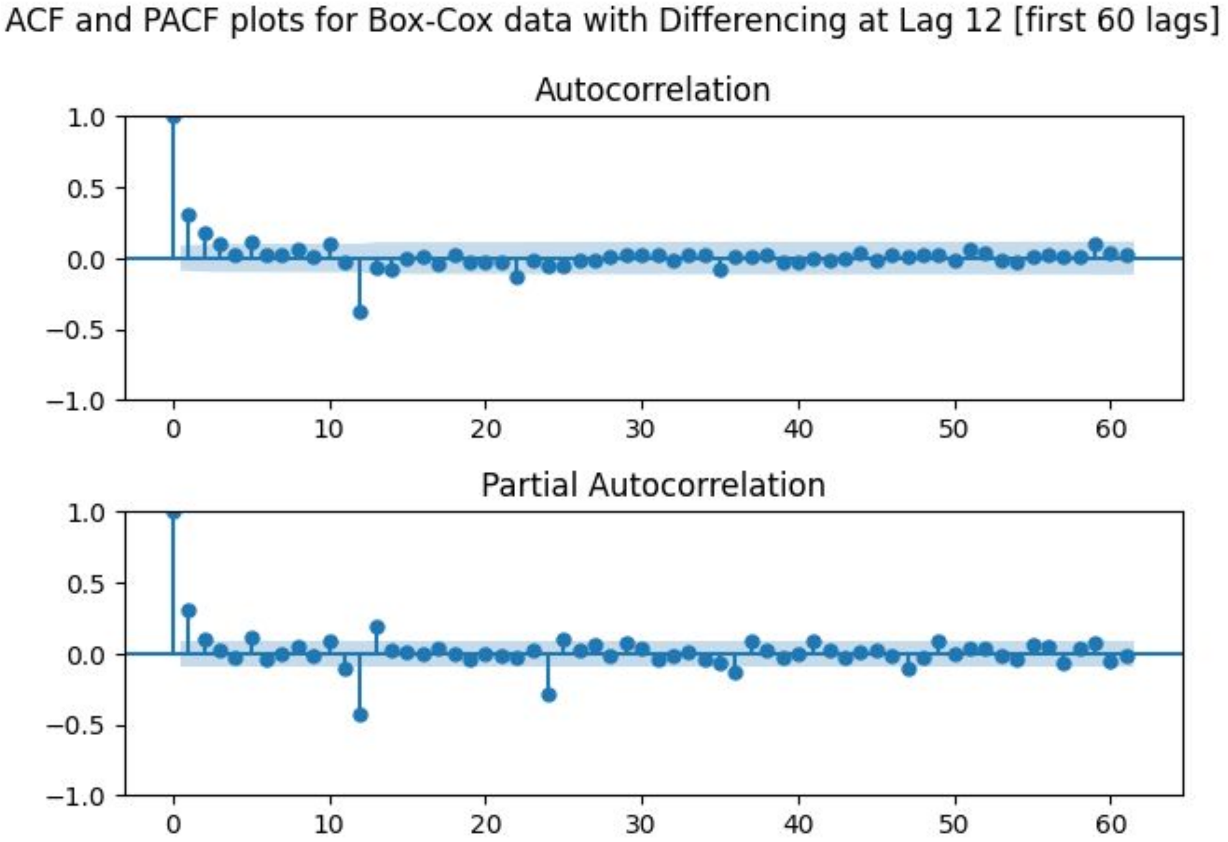
- ACF of first 12 lags:
  - Try **q = 0, 1, 2**

- PACF:
  - Try **p = 0, 1, 2**



- ACF:
  - Try **Q = 0, 1**

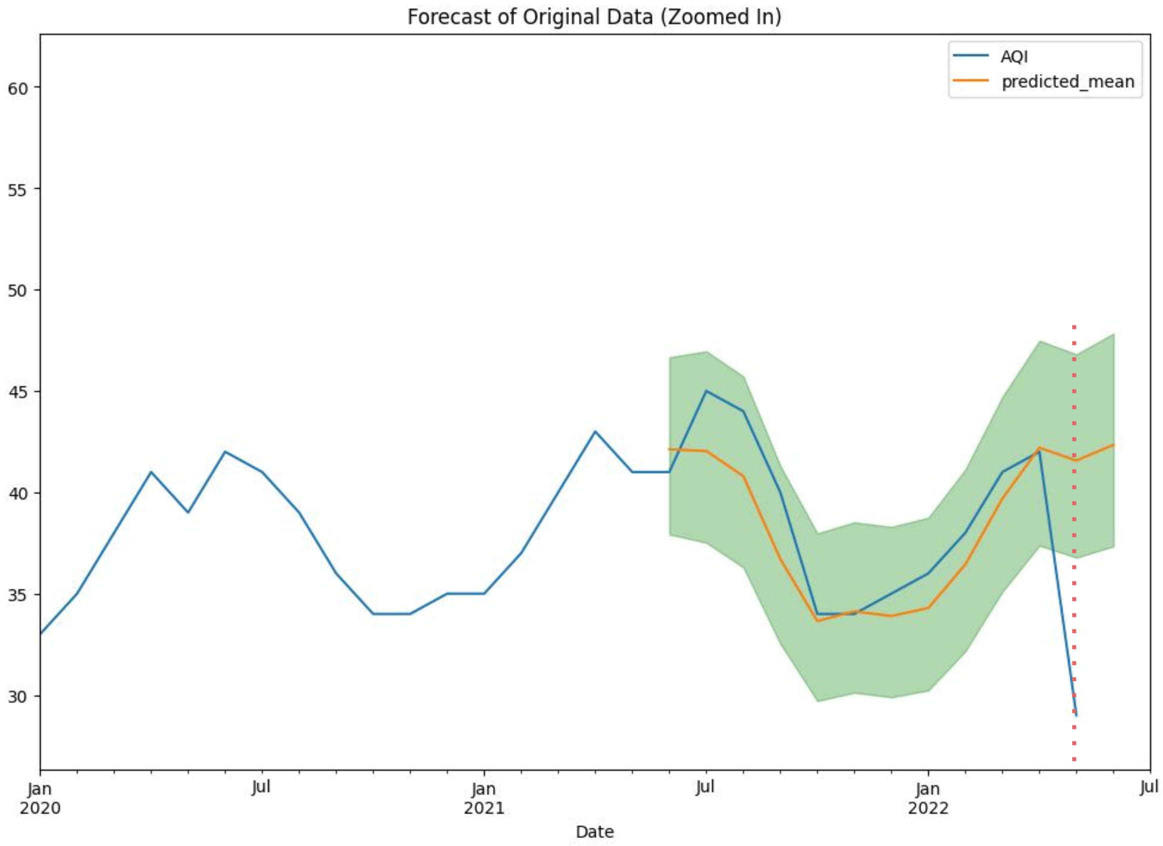
- PACF:
  - Try **P = 0, 1, 2**



Best SARIMA Model:

SARIMA (2,0,1) x $(0,1,1)_{12}$
on Box-Cox data

- RMSE: 4.056
- Problem: last point on test set



Prophet Modelling

Hyperparameter Tuning and Cross-Validation

Hyperparameters

- *changepoint_prior_scale*: [0.001, 0.05, 0.1, 0.5] #default 0.05
- *seasonality_prior_scale*: [0.01, 0.1, 1.0, 10.0] #default 10.0

Cross-validation

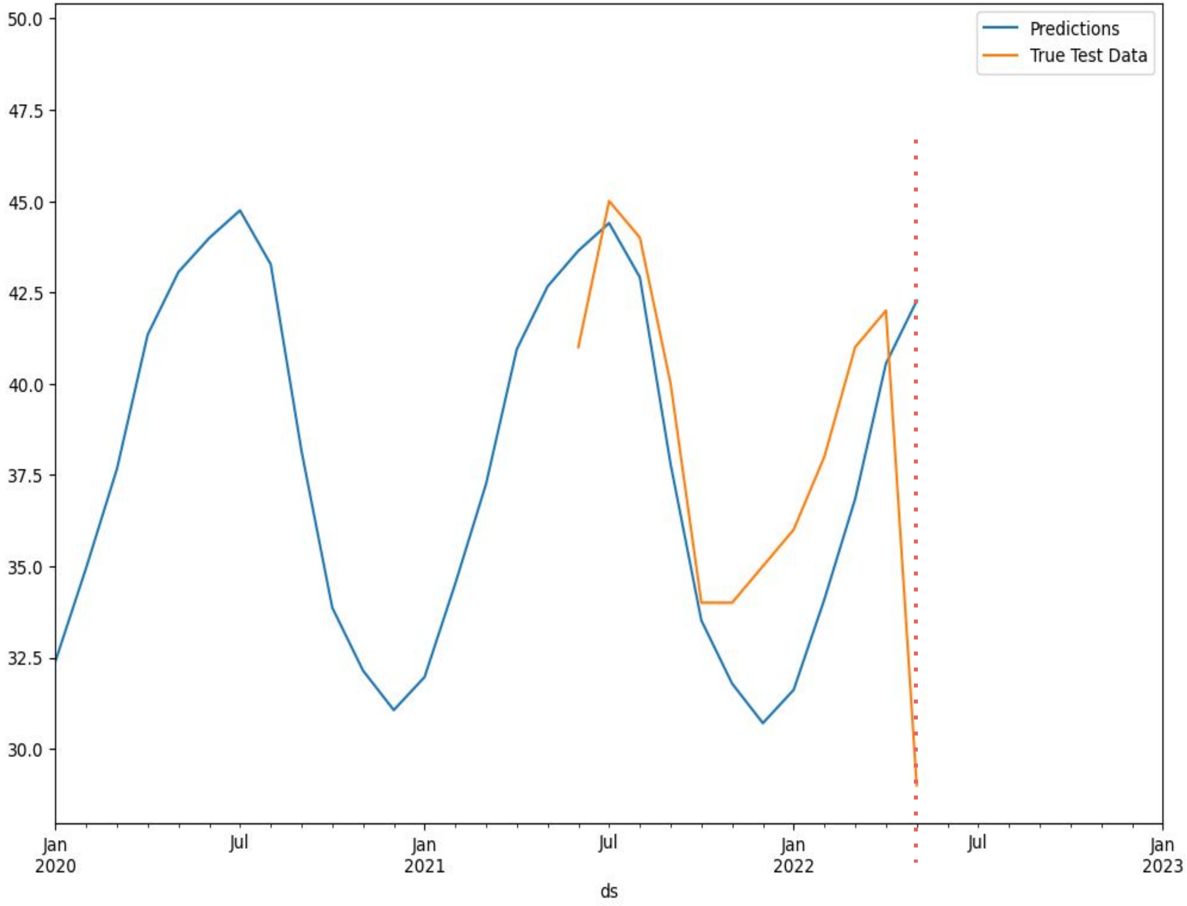
- *initial* = 10957.5 days #30 years
- *horizon* = 360 days #predict the next year
- *period* = 90 days #run the model every 90 days

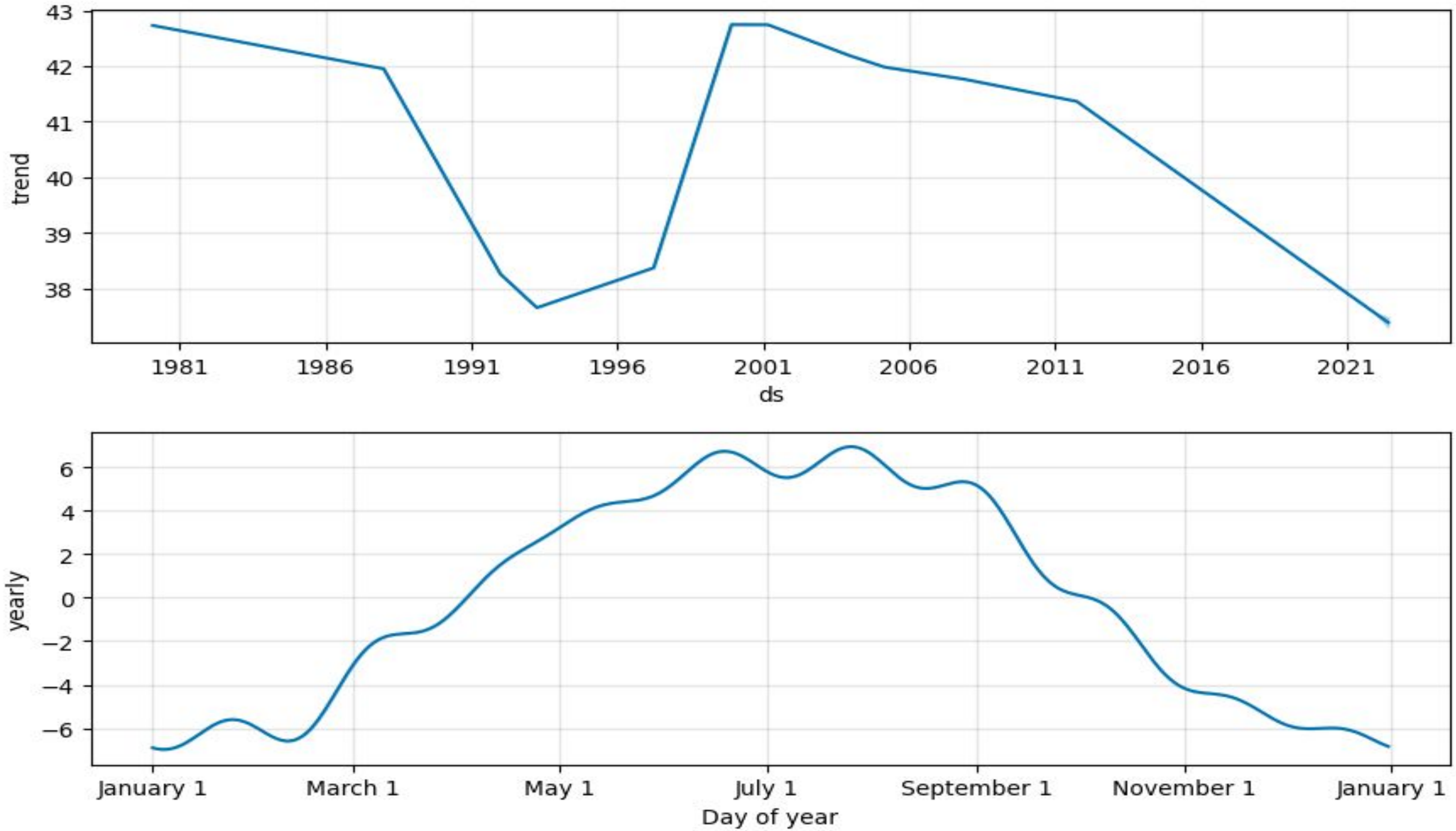
Best Prophet Model

- changepoint_prior_scale = 0.5
- seasonality_prior_scale = 0.01

- RMSE: 2.578

- Also problem with last test point





Trend and yearly components of best prophet model

LSTM Modelling

Preprocessing and Hyperparameter Tuning

Preprocessing:

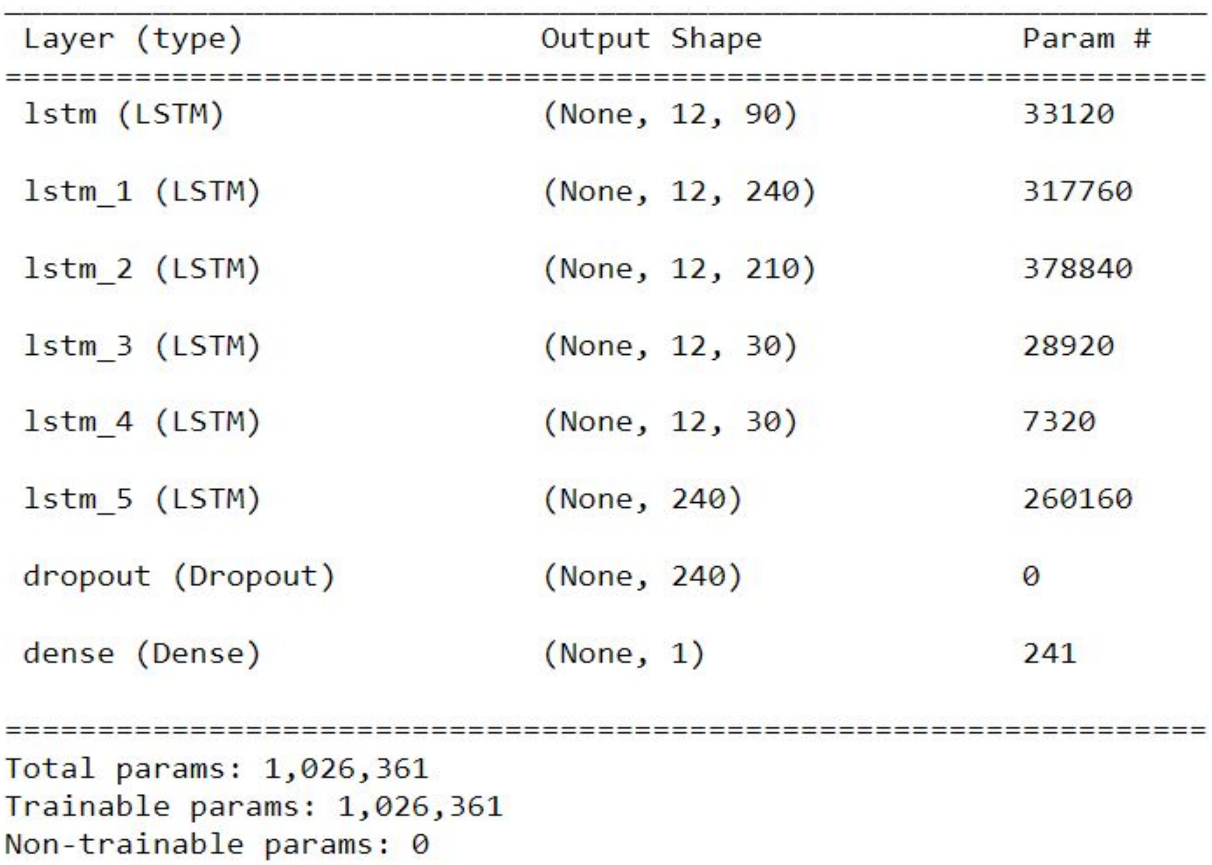
- Scale data with *MinMaxScaler*

Hyperparameters:

Number of neurons:	30- 360, with step 30
Number of layers:	1-4
Best dropout rate:	0 - 0.5, with step 0.1
Activation:	relu, sigmoid

Fit model with 30 epochs

Best LSTM Model Summary

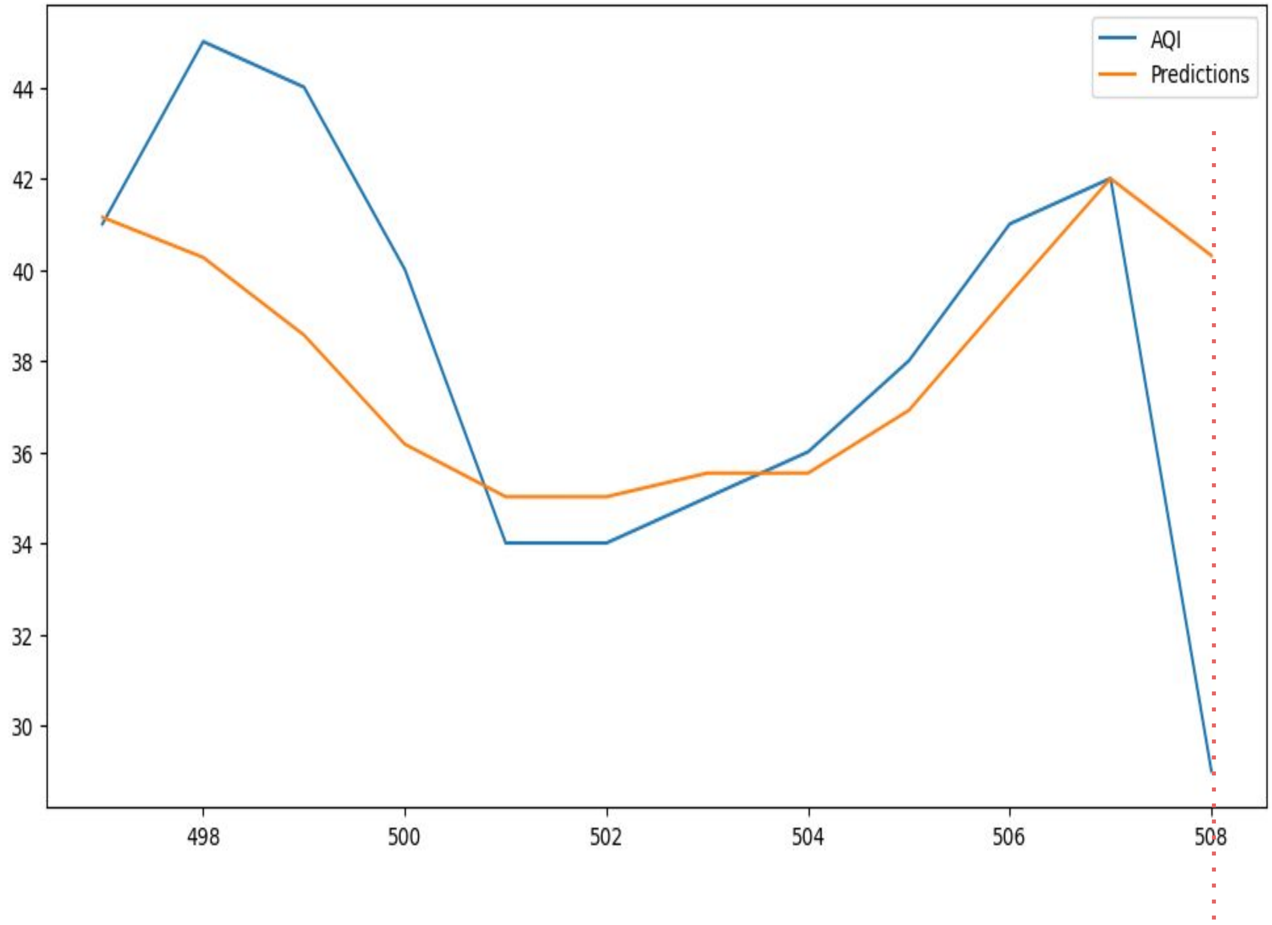


```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 12, 90)            33120

 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
```

Best LSTM Model

- Problem with last test point

- RMSE: 4.089



Comparison of Models

SARIMA

- RMSE: 4.056
- Most interpretable

Prophet

- RMSE: 2.578

LSTM

- RMSE: 4.089
- Least interpretable

Model Similarities:

- All had trouble with predicting the last point in the test set.

- This could be because of unforeseeable events (e.g. Covid19, the economy, war)

Conclusion

- Prophet is the best model to forecast this data
    - Lowest RSME
    - Yearly seasonal component – maxima in July and August

- All models imply unprecedented events that caused the median AQI to drop in May 2022 (last point in test set)
- Top cities with high median AQI are in CA and with have high density
- Relationships between AQI level and region/ population still needs to be explored much further