Evaluation of Various Machine Learning Approaches to Predict PM2.5 Concentrations

Presented by: Preetham Pangaluru
Mentor: Dr. Jonathan Jiang

How Data Was Obtained

PM2.5 Data Set:

- Using API to access data from OpenAQ with a base url of "https://api.openaq.org/v1/measurements"

- The PM2.5 Data set was from Jan 01, 2018 – Dec 31, 2021 and was made into daily, monthly, and seasonal average data as well to better predict trends in each of those temporal periods

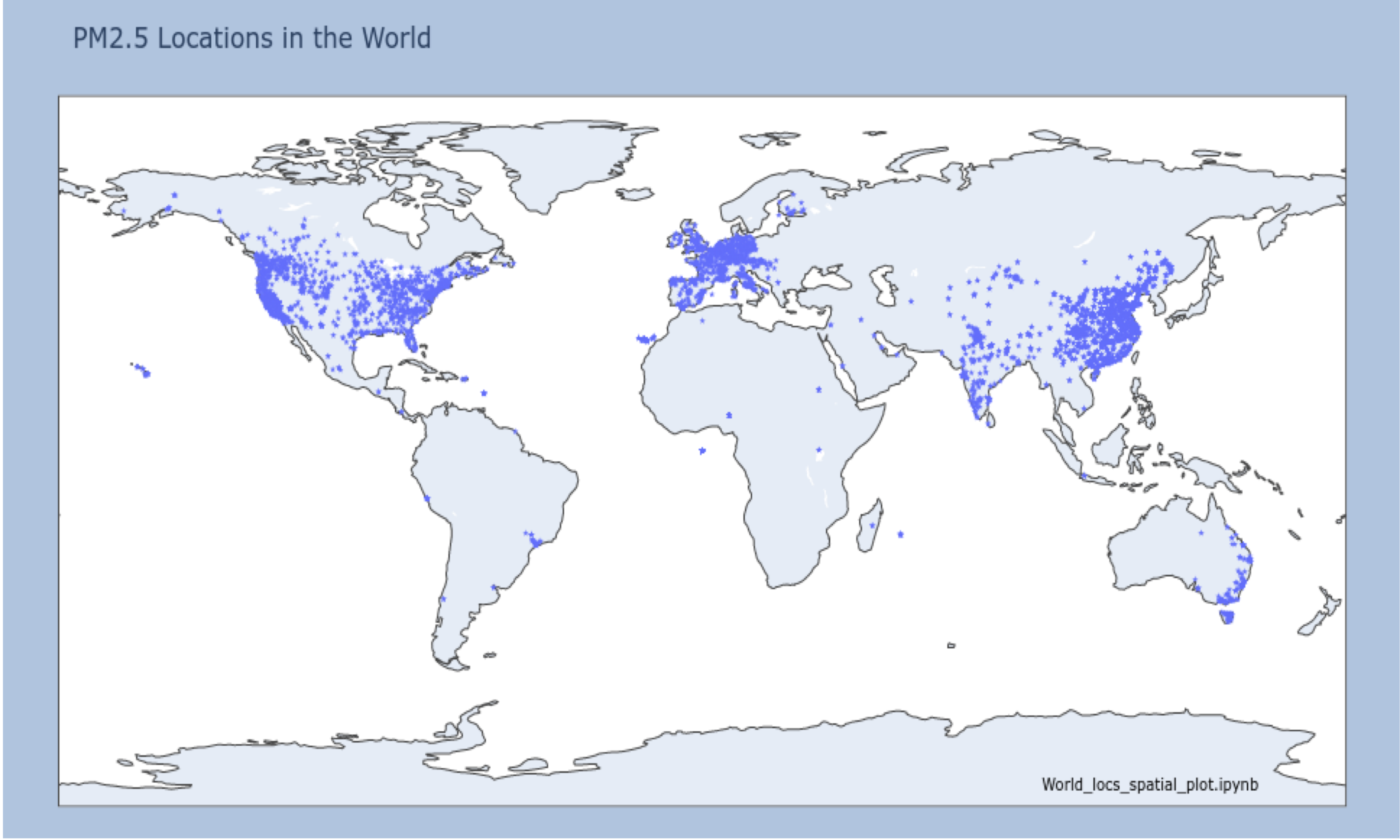


Fig. 1: PM2.5 locations (stations) around the world

PM2.5 Locations Around the United States

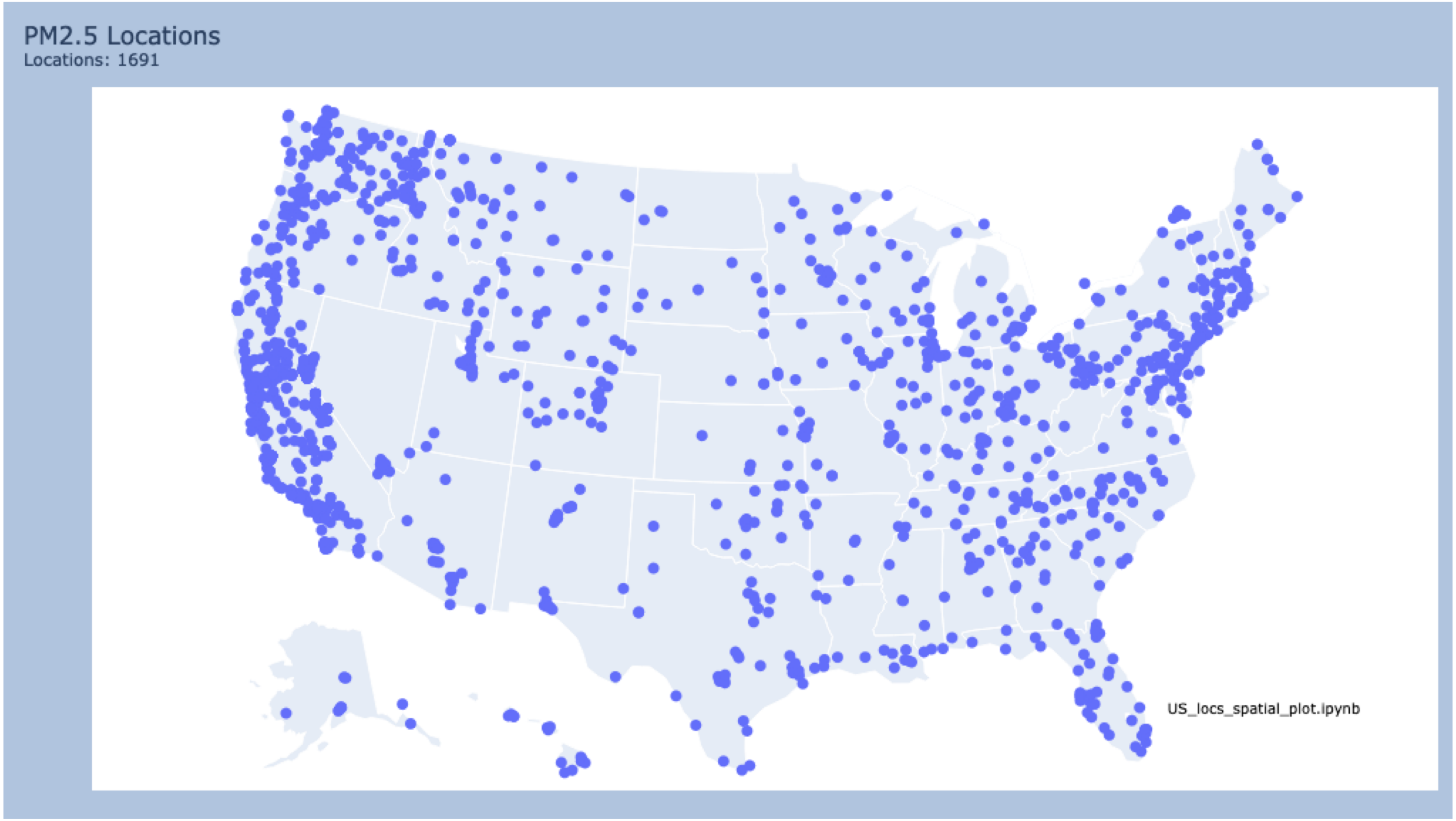


Fig. 2: There are 1691 PM2.5 locations around the US

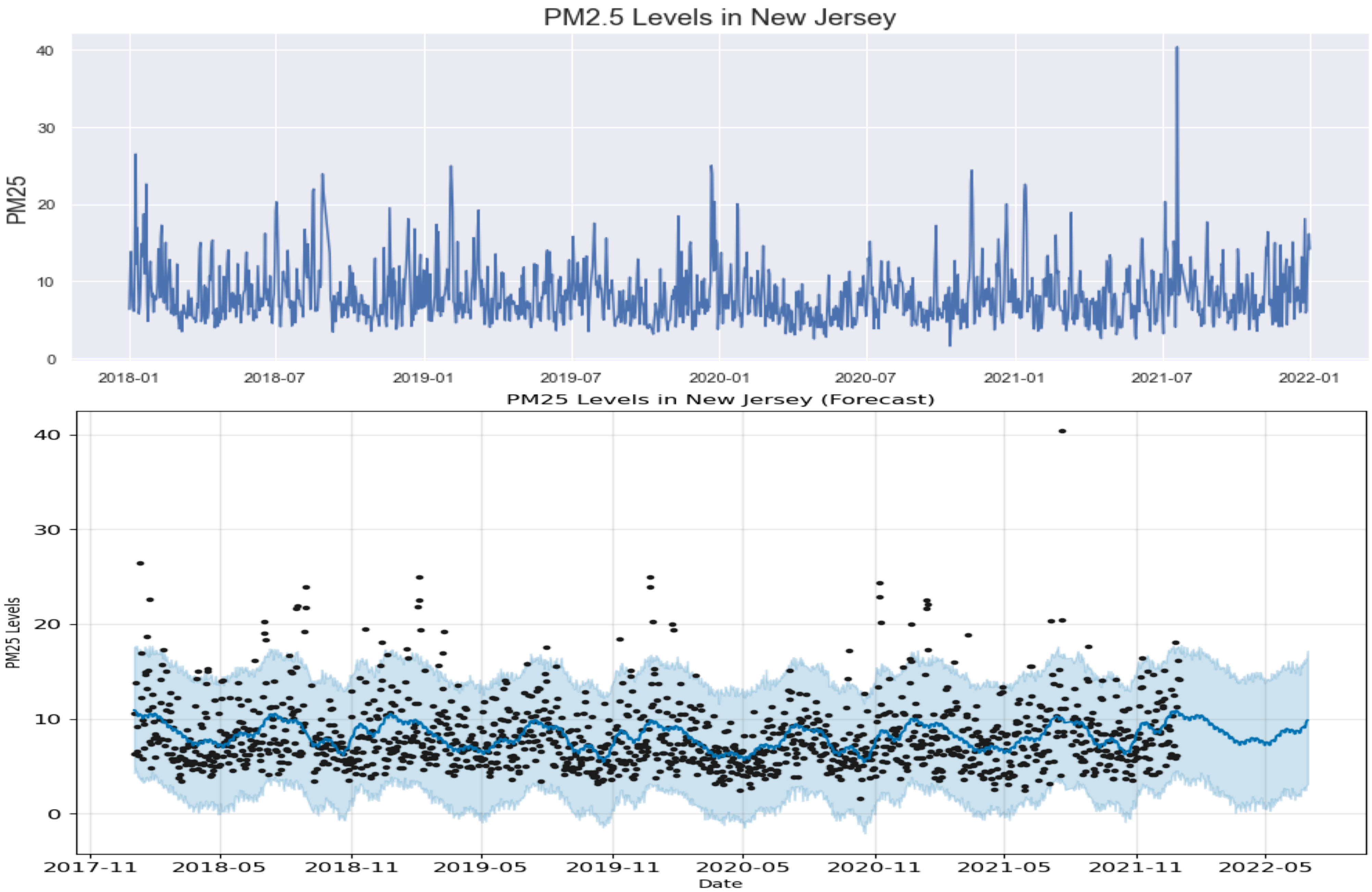
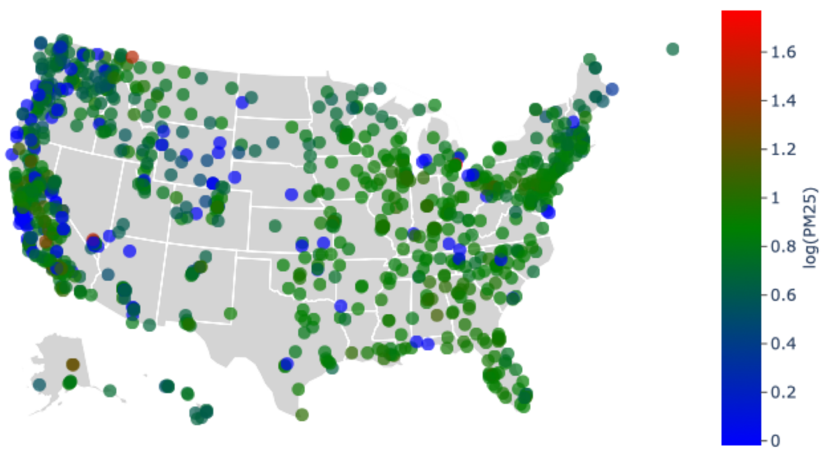


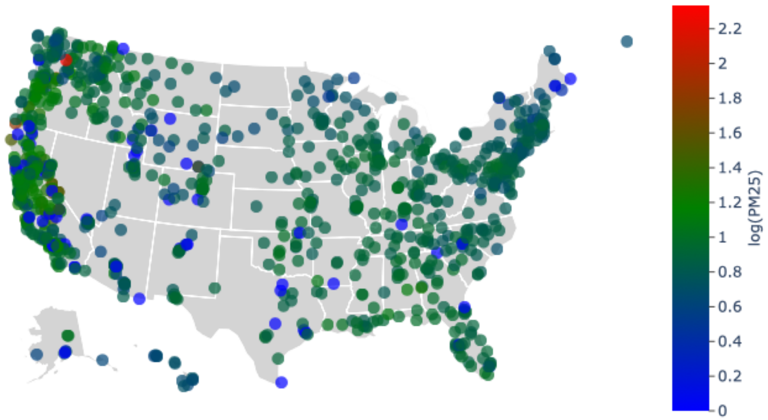
Fig.3: Daily PM2.5 time series (top) and bottom forecast predictions.

Before and During Lockdown: A Visual Comparison of PM2.5

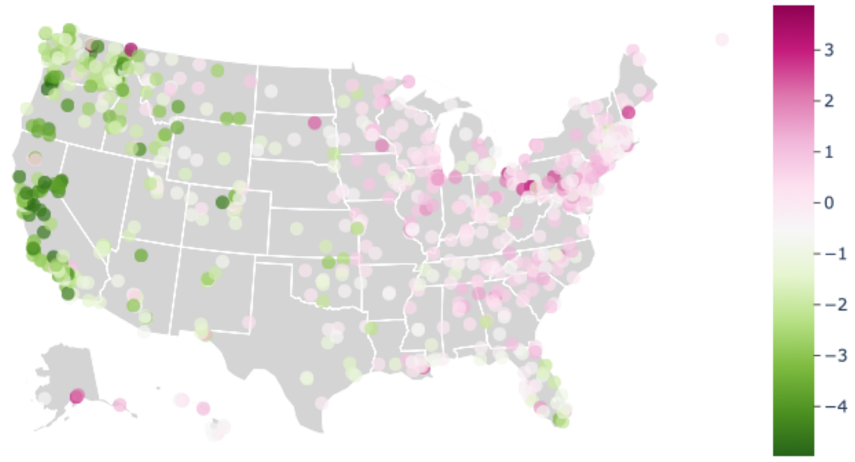
PM2.5 Locations in USA Before Lockdown (01/08/2019 - 03/19/20) 8 AM - 6 PM



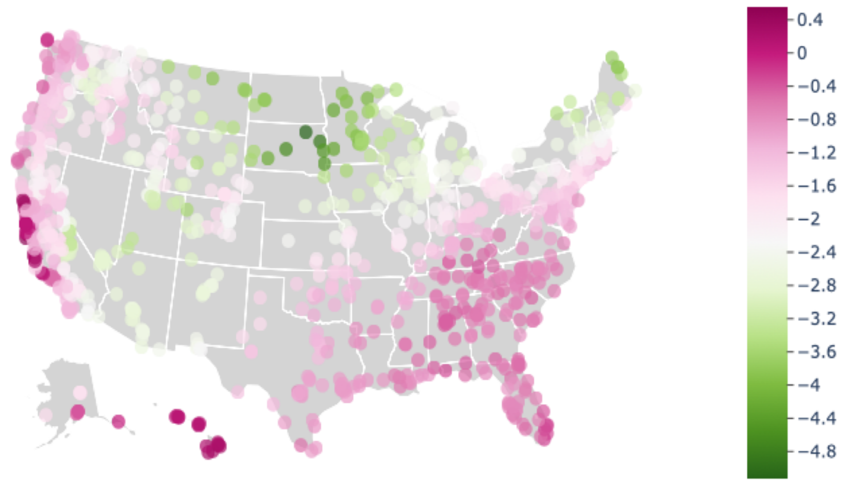
PM2.5 Locations in USA During Lockdown (03/20/20 - 03/20/21) 8 AM - 6 PM



PM2.5 Locations in USA (Pre-Lock) Selected Period 8 AM - 6 PM



Temperature Locations in USA (Pre-Lock) Selected Period



mnlyclimatology_USmap.ipynb

Fig. PM2.5 monthly climatology over each location during Jan 2018 – Dec 2021.

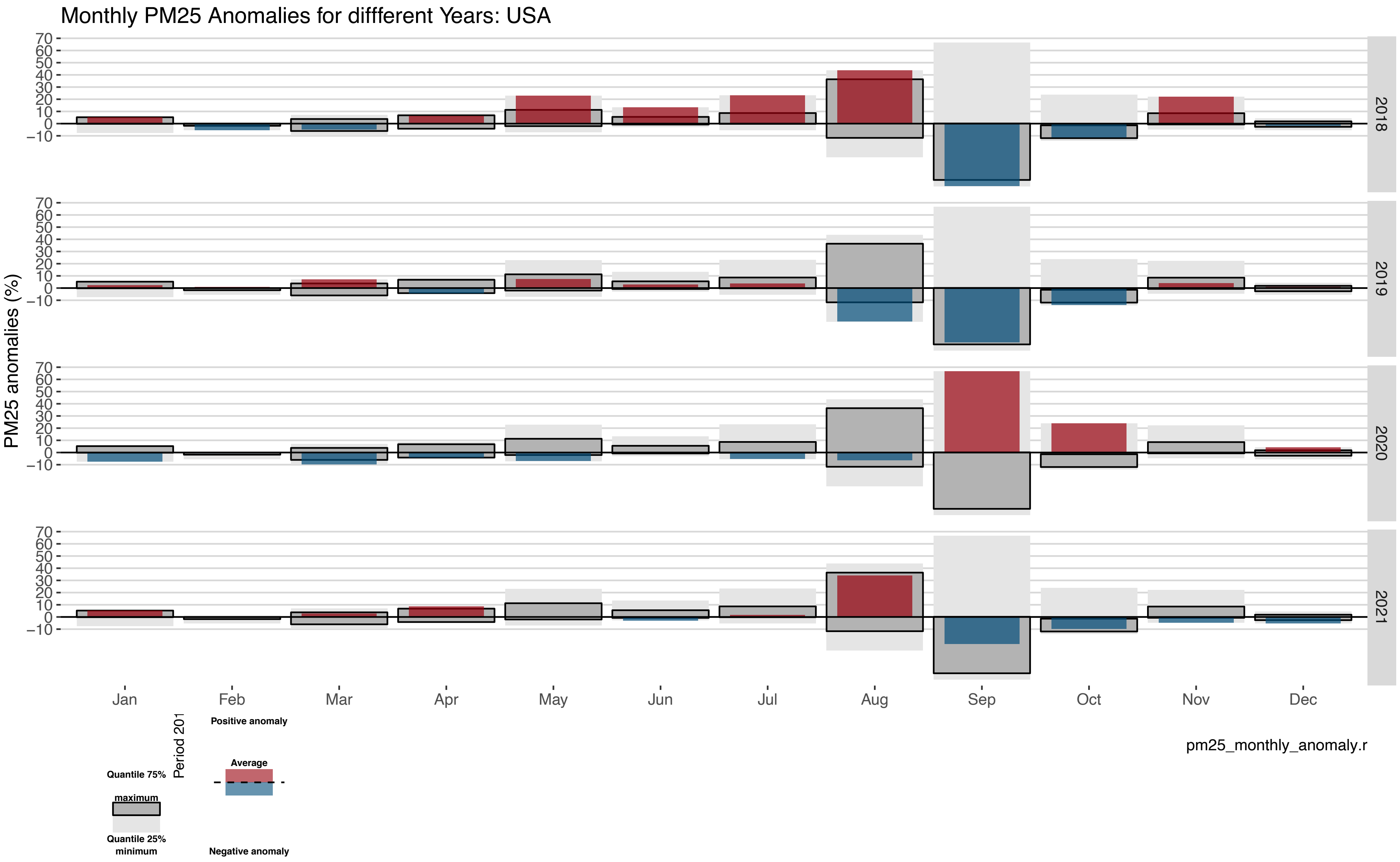


Fig.: Monthly PM2.5 anomalies for different years (2018 – 2021)

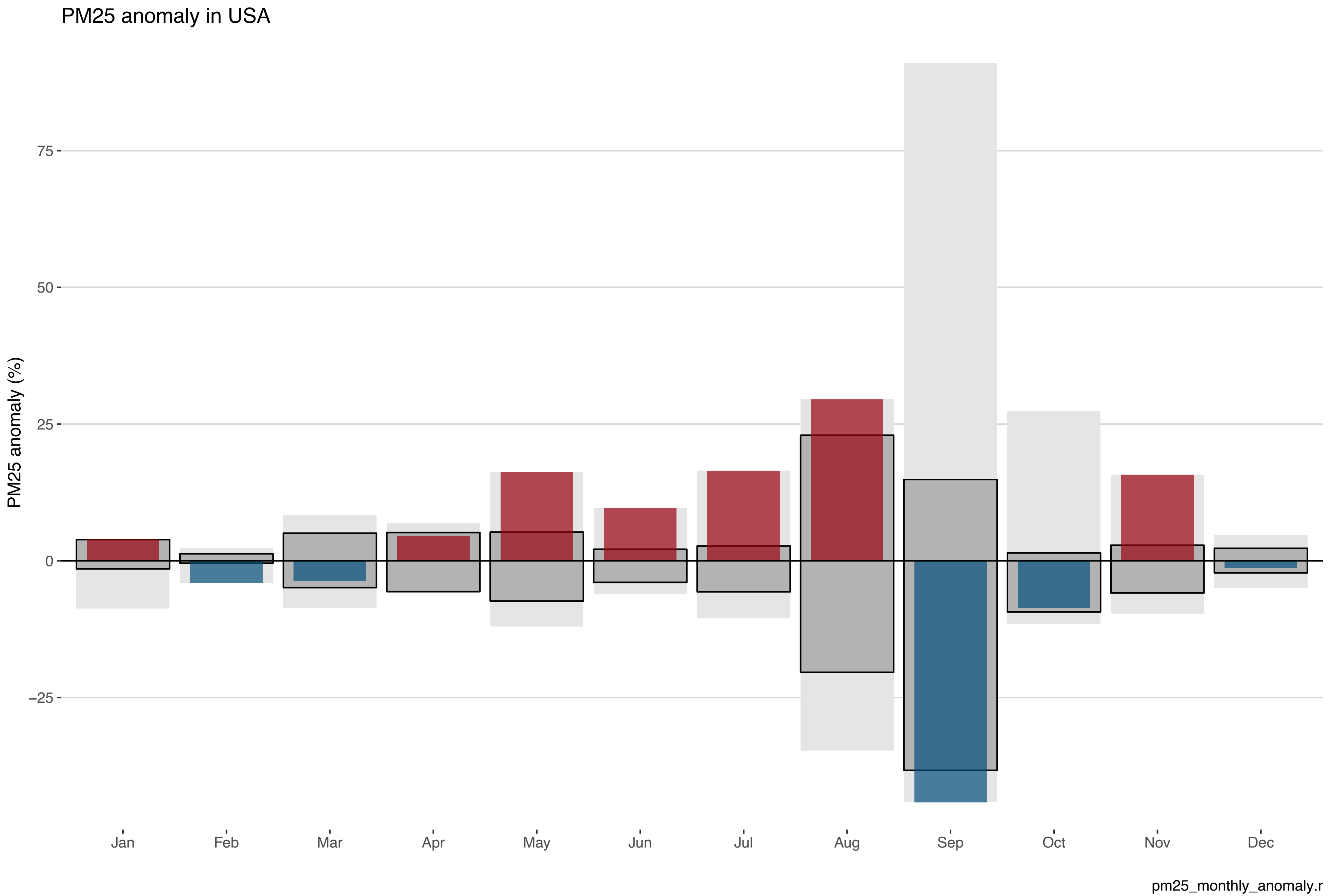
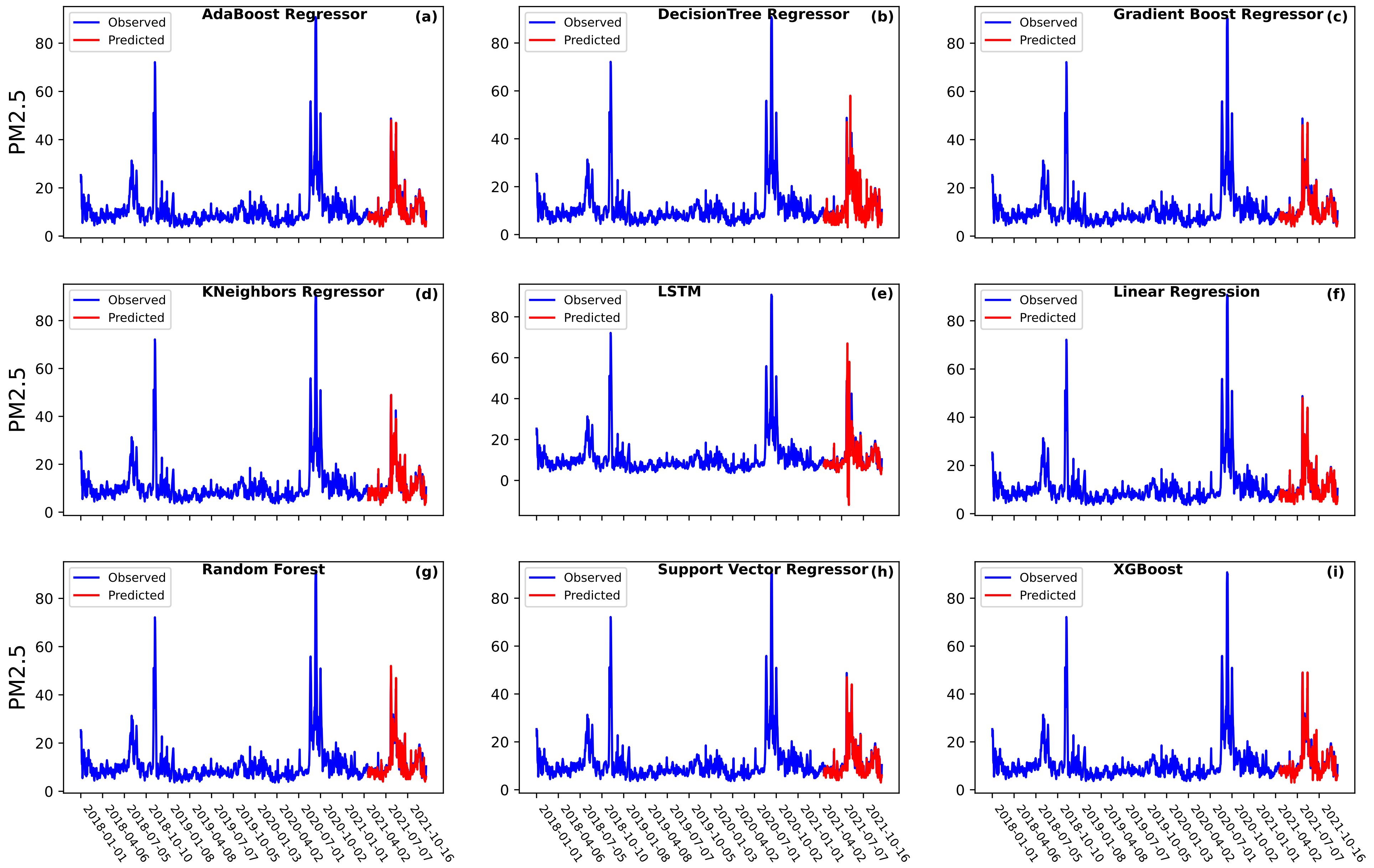
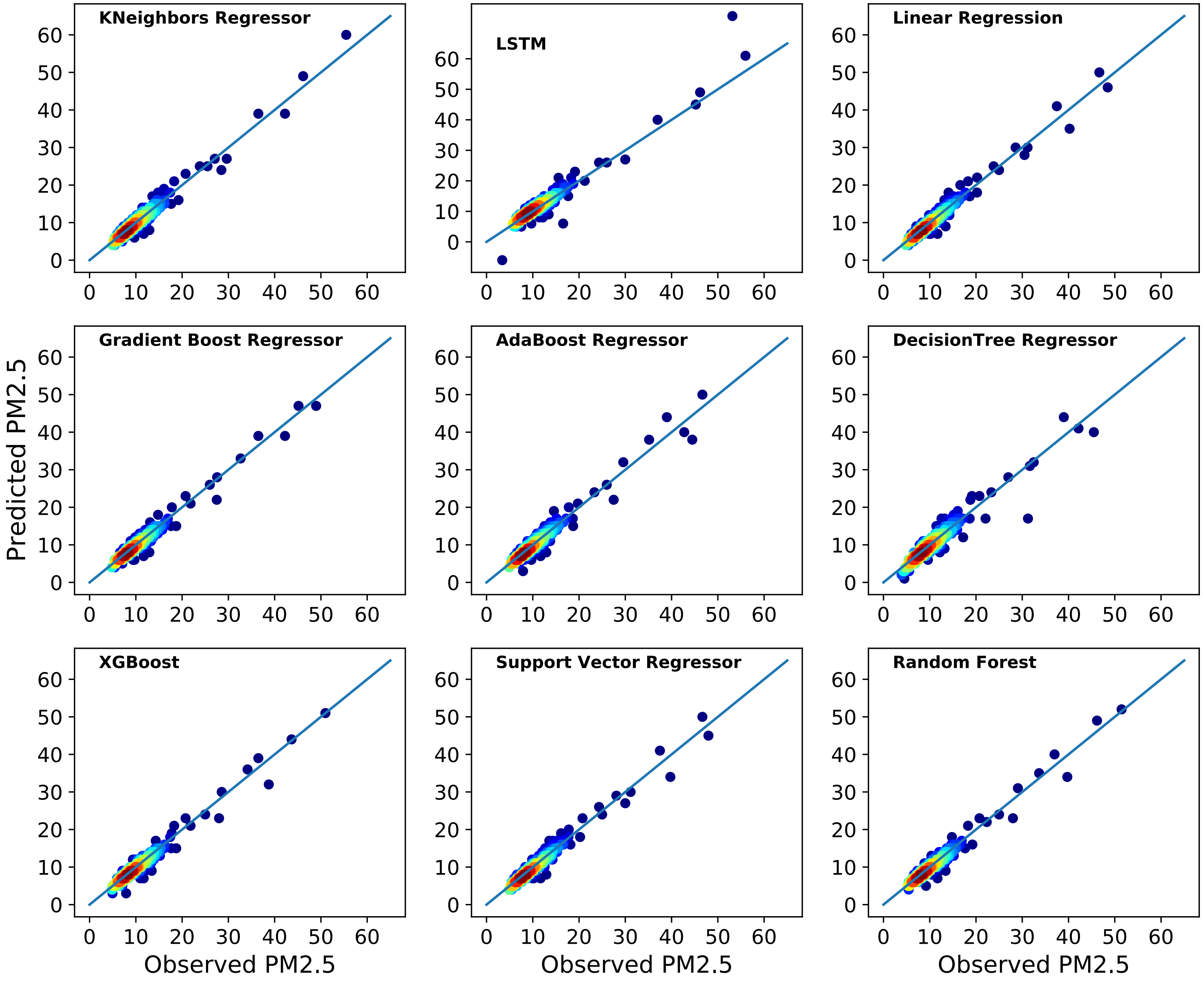


Fig.: USA monthly anomalies during the period from 2018 to 2021.



PM25_dly_dfrntmdls_obs_pred_lineplt.ipynb

Fig.: California PM2.5 daily time series (blue) and predictions (red) for different ML models



PM25_dly_dfrntmdls_scatterplt.ipynb

Fig.: PM2.5 observed and predicted scatter plot for different ML models

Table Demonstrating Error Metric Comparison of ML Models

California							
Model	RMSE	MAE	MSE	R2	NSE	NORM	PBIAS
Linear Regression	3.883	2.309	15.077	0.688	0.70	60.156	4.24
Decision Tree	5.136	3.109	26.387	0.454	0.533	79.58	3.44
Gradient Boost Regressor	3.822	2.394	14.606	0.698	0.683	59.207	5.21
AdaBoost Regressor	3.961	2.316	15.693	0.676	0.683	61.369	4.653
XG Boost	3.898	2.501	15.197	0.686	0.681	60.393	4.342
KNeighbors Regressor	3.919	2.379	15.358	0.683	0.677	60.711	4.515
LSTM	7.487	3.359	56.058	0.158	0.255	115.991	6.00
Random Forest	3.121	2.122	11.348	0.899	0.211	38.671	2.989
SVM	3.125	2.145	12.129	0.857	0.280	39.161	3.011

Conclusion

- Machine Learning models are relevant to the study of PM2.5 concentrations across the USA and other parts of the world

- Our variety of developed ML models provide a robust and accurate measure of PM2.5 levels In USA and can even possibly be extended to other forms of particulate matter and pollution

- New ML models and techniques have allowed us to further predictive ability when it comes to PM2.5 levels

- The SVM, RF, and LSTM models were the best at capturing temporal dependencies in time series data, Increasing accuracy of PM2.5 forecasting

- Fall and winter were especially harder to predict because they happened to have higher levels of air pollution and especially, PM2.5 concentrations

Thanks For Watching