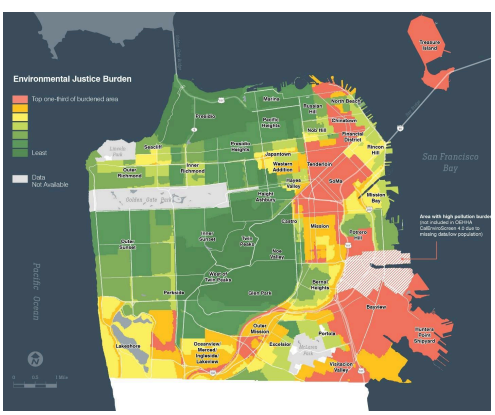
# Statistical Linear Modeling for Hourly $PM_{2.5}$ Multi-Step Time Series Forecasting in Eastern San Francisco

Eleanor Kim

## Abstract

This study explores the application of statistical linear modeling techniques for forecasting hourly $PM_{2.5}$ levels in Eastern San Francisco. Over the past four years, San Francisco has faced challenges with poor air quality, particularly in neighborhoods such as Chinatown, the Tenderloin, and Bayview-Hunters Point. Using data collected by Brightline Defense, an environmental advocacy group, this research aims to develop predictive models to understand the spatiotemporal dynamics of $PM_{2.5}$ and identify influential factors such as humidity, temperature, seasonality, and neighborhood effects. Through exploratory data analysis, feature engineering, parameter estimation, and sensitivity analysis, various modeling approaches including ordinary least squares, weighted least squares, ridge, lasso, and outlier detection methods are evaluated. The findings highlight the importance of including both temperature and humidity covariates in the model, the significance of neighborhood and seasonal effects, and the impact of extreme outliers on model performance. Ultimately, this study contributes to the understanding of air quality dynamics in Eastern San Francisco and informs potential interventions to mitigate the adverse effects of poor air quality on public health and environmental justice.

*Figure 1. Environmental Justice Communities from SF Planning[1] (Left)*
*Figure 2. Brightline's Air Quality Monitoring Network Map (Right)*




[1] "Environmental Justice Framework and General Plan Policies." *Environmental Justice Framework and General Plan Policies. SF Planning*, sfplanning.org/project/environmental-justice-framework-and-general-plan-policies.

## Introduction to Topic

Over the past four years, San Francisco has experienced record levels of particulate matter in its troposphere as a result of increased carbon emissions and extreme wildfire events. The city's unique location on a peninsula with three sides surrounding the Bay and the Pacific Ocean is prone to thermal inversion[2], wherein a cool layer of air traps pollutants close to the surface, producing hazy air quality. Portions of the city experience disproportionate effects of poor air quality as a result of environmental and socioeconomic factors. The eastern side of San Francisco, which includes the neighborhoods of Chinatown, the Tenderloin, South of Market (SoMa), and Bayview-Hunters Point bears the brunt of the city's poor air quality such that communities in these neighborhoods are termed by the state as *disadvantaged communities* by Senate Bill 535[3]. For almost four years, Brightline Defense[4], an environmental justice advocacy non-profit located in the SoMA neighborhood of San Francisco, has been monitoring the air quality in Eastern San Francisco to draw attention to the environmental burdens of these disadvantaged communities. Brightline's air quality monitoring program is funded in part by California's Assembly Bill 617[5] Community Air Protection Program. The network of around twenty sensors leverages the technology of Clarity's[6] low-cost solar powered air quality monitors to gather data on $PM_{2.5}$[7] (2.5 micrometers in diameter) across Eastern San Francisco.

## Project Questions

I have been a data analyst for Brightline's air quality monitoring program for about three years, mostly summarizing and comparing the $PM_{2.5}$ levels in these neighborhoods with informative visuals. Prior to this project, never have I attempted modeling the air quality using the 3+ years of granular data on $PM_{2.5}$ I work with. Producing a linear model to forecast $PM_{2.5}$, while not directly a practical tool, could provide valuable insights into the spatiotemporal dynamics of $PM_{2.5}$ across various disadvantaged communities in the city and how factors such as temperature, humidity, and seasonality affect these dynamics. Many studies[8,9,10] exist that explore a forecasting model for $PM_{2.5}$ using a multi-step time series approach, wherein their models include a fixed number of "lag steps" of $PM_{2.5}$ to predict the next value on the time scale. Using this multi-step time series forecasting model approach, I have proposed the following four questions which I will use to guide my data analysis.

---

[2] "The Marine Layer." *National Oceanic and Atmospheric Administration*, www.noaa.gov/jetstream/ocean/marine-layer.
[3] *Oehha.ca.Gov*, oehha.ca.gov/calenviroscreen/sb535.
[4] *Brightline Defense*, www.brightlinedefense.org/.
[5] "California Air Resources Board." *Community Air Protection Program | California Air Resources Board*, ww2.arb.ca.gov/capp.
[6] "Low-Cost Air Quality Monitoring & Measurement: Clarity Movement Co.." *A*, www.clarity.io/.
[7] "California Air Resources Board." *Inhalable Particulate Matter and Health (PM2.5 and PM10) | California Air Resources Board*, ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health.
[8] Gregório, J.; Gouveia-Caridade, C.; Caridade, P.J.S.B. Modeling PM2.5 and PM10 Using a Robust Simplified Linear Regression Machine Learning Algorithm. *Atmosphere* 2022, *13*, 1334. https://doi.org/10.3390/atmos 13081334
[9] Zaini, N., Ean, L.W., Ahmed, A.N. *et al.* PM2.5 forecasting for an urban area based on deep learning and decomposition method. *Sci Rep* 12, 17565 (2022). https://doi.org/10.1038/s41598-022-21769-1
[10] Kefei Zhang, Xiaolin Yang, Hua Cao, Jesse Thé, Zhongchao Tan, Hesheng Yu, Multi-step forecast of PM2.5 and PM10 concentrations using convolutional neural network integrated with spatial–temporal attention and residual learning, Environment International, Volume 171, 2023, 107691, ISSN 0160-4120, https://doi.org/10.1016/j.envint.2022.107691.

- Can a predictive model for $PM_{2.5}$ be produced with high accuracy?
- How can the model's accuracy be optimized with respect to the number of lag steps included in the model?
- What are the individual contributions of factors such as humidity, temperature, seasonality, and neighborhood?
- How do anomalously high $PM_{2.5}$ values affect the model?

## Data

This analysis will explore hourly-averaged $PM_{2.5}$ data from twenty sensors located across five neighborhoods (Bayview-Hunters Point, Chinatown, Potrero Hill, SoMa, and the Tenderloin) starting in August 2020 up until March 2024. The cleaned dataset contains 546,743 observations indexed over a unique air quality sensor and hourly timestamp. Data is publicly available on Brightline's website[11]. The full data can be found in Forecasting-PM2.5-in-SF/sf_pollution.csv.

*Figure 3. R Data Frame Summary Output*

```
'data.frame':   546743 obs. of  10 variables:
 $ ID                           : chr  "AYHT7CF7" "ATWJ3V74" "ALZ4PSJB" "A2BPKVSX" ...
 $ Datetime                     : POSIXct, format: "2024-02-15 23:00:00" "2024-02-15 23:00:00" "2024-02-15 23:00:00"
 $ PM2.5_Hour_MassConc_Calibrated: num  7.39 5.86 5.93 9.42 5.86 4.04 4.03 3.61 6.35 6.32 ...
 $ Temperature                  : num  12.1 13.1 11.7 11.1 11.6 ...
 $ Humidity                     : num  83.4 81.2 86.2 89 85.7 ...
 $ Latitude                     : num  37.8 37.8 37.8 37.7 37.7 ...
 $ Longitude                    : num  -122 -122 -122 -122 -122 ...
 $ Device_Name                  : chr  "Fitness SF" "Howard & 9th" "BAAQMD Co-Location" "The Box Shop" ...
 $ Neighborhood                 : chr  "SoMa" "SoMa" "Potrero Hill" "BVHP" ...
 $ Season                       : chr  "Winter" "Winter" "Winter" "Winter" ...
```

## Exploratory Data Analysis

Before diving into the analysis, it is necessary to understand the empirical distribution of the data and what covariates might be worth considering for the model. Across the entire dataset, the median $PM_{2.5}$ level is 6.27 $\mu g/m^3$ with a standard deviation of 8.62. The distribution is highly right skewed with a maximum value of 241,426.27 $\mu g/m^3$, as seen in Figure 4. Very high pollutant levels occurred during extreme wildfire events in North California that severely impacted the Bay Area. I will later explore the impact of extreme outliers on our model during the sensitivity analysis. Furthermore, controlling for neighborhoods allows us to see if the spatial effects on $PM_{2.}$. The following histograms are the empirical distribution of $PM_{2.5}$, subsetted by neighborhood. It appears that Potrero Hill and Chinatown experience better air quality on average compared to SoMa, Tenderloin, and Bayview Hunters Point.

[11] "Environmental Justice Data." *Brightline Defense*, www.brightlinedefense.org/environmental-justice-data.

*Figure 4. R $PM_{2.5}$ Summary Output*

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00    3.96    6.27    8.44   10.02  241.42
```

*Figure 5. Histogram of $PM_{2.5}$*

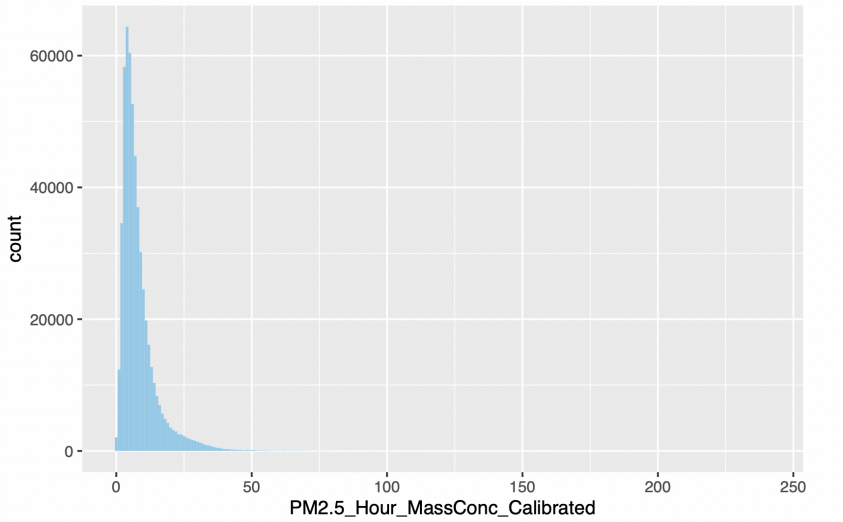
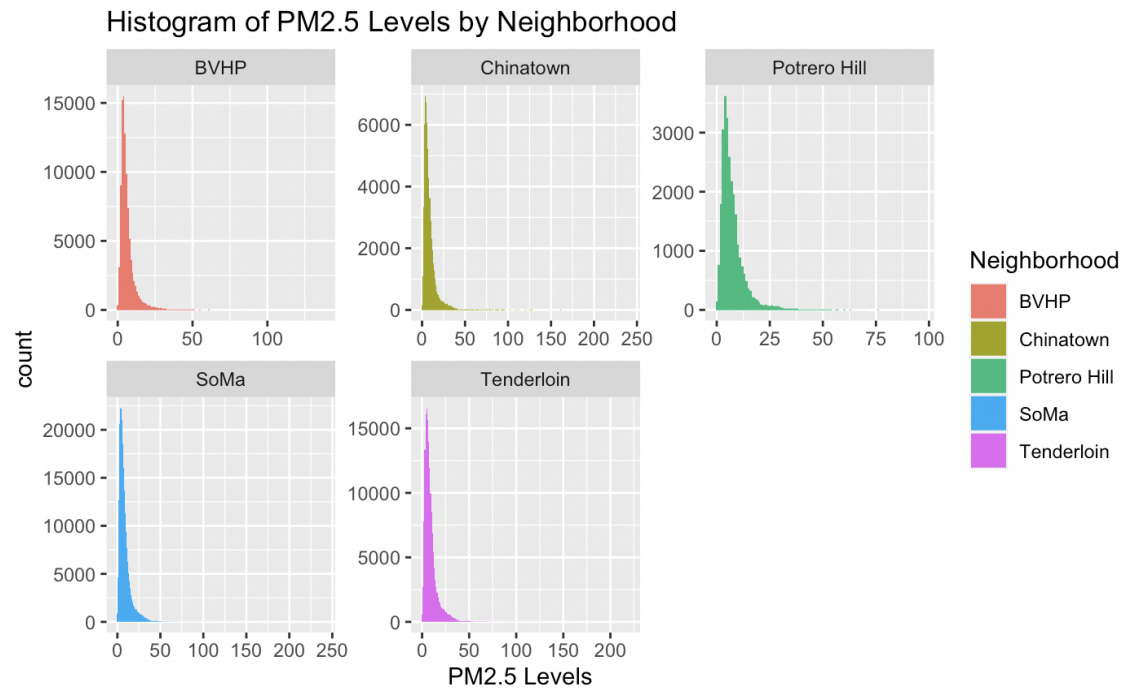



*Figure 6. Histogram of $PM_{2.5}$ for each Neighborhood*



Next I explore the time series from August 2020 to March 2024, including a rough visual of the seasonality of $PM_{2.5}$. A smaller time scale of a 2024 time series only is shown as well for more visible granularity. According to the Bay Area Air Quality Management District[12], we should expect to see worse air pollution in the winter due to wood burning in homes. High pollution during the summer is typically attributed to car smog and wildfire effects.

*Figure 7. Time Series of $PM_{2.5}$ August 2020 - March 2024*

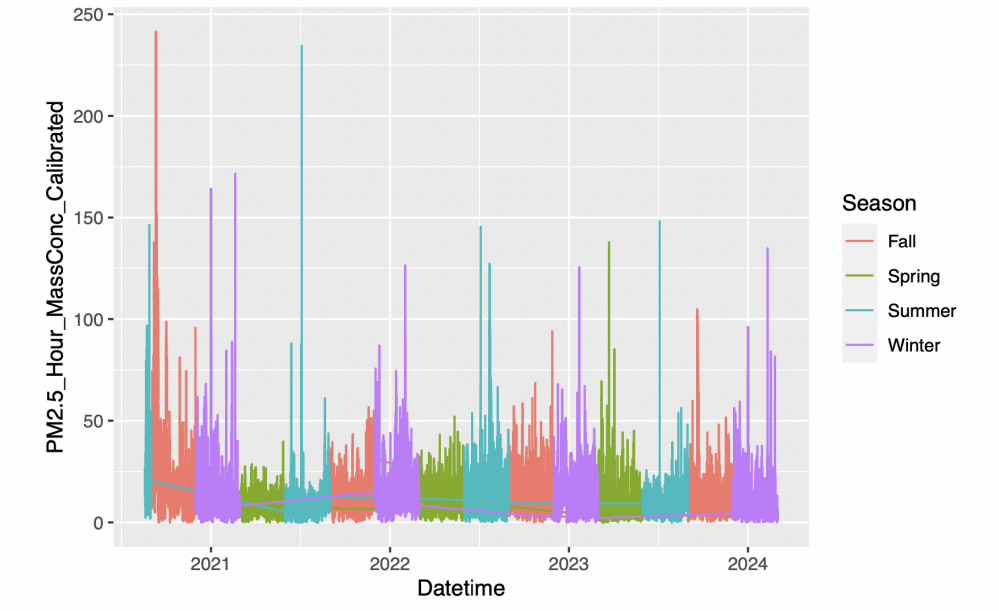
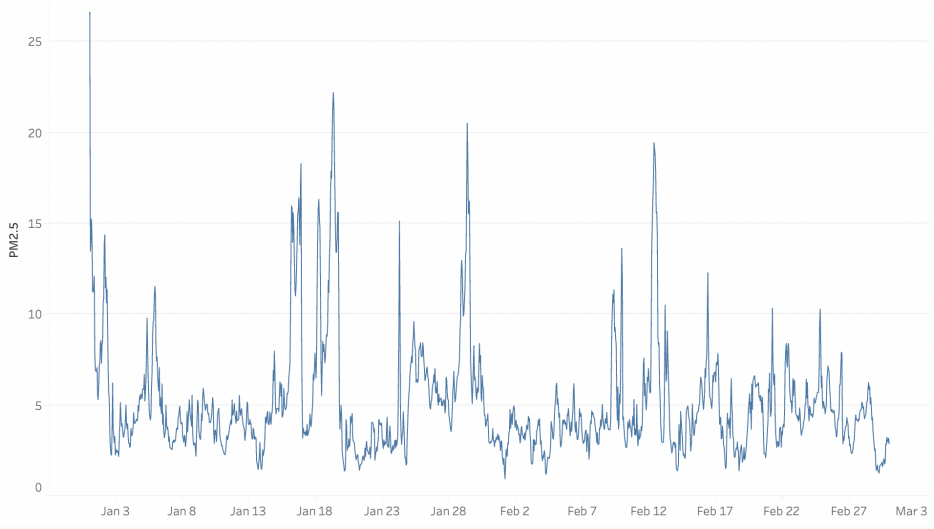



*Figure 8. Time Series of $PM_{2.5}$ January - March 2024*




[12] "Wood Smoke Pollution." *Bay Area Air Quality Management District A Healthy Breathing Environment For Every Bay Area Resident*, www.baaqmd.gov/rules-and-compliance/wood-smoke.

And finally, I explore the relationship between $PM_{2.5}$ and meteorological conditions by plotting $PM_{2.5}$ against temperature and humidity. It seems that there is roughly a positive correlation between humidity and $PM_{2.5}$, while a rough negative correlation between temperature in $PM_{2.5}$. This inverse relationship makes sense because temperature and pressure are inversely related. However, there is high heteroskedasticity among each variable.

*Figure 9. Scatter Plot of $PM_{2.5}$ vs. Humidity*

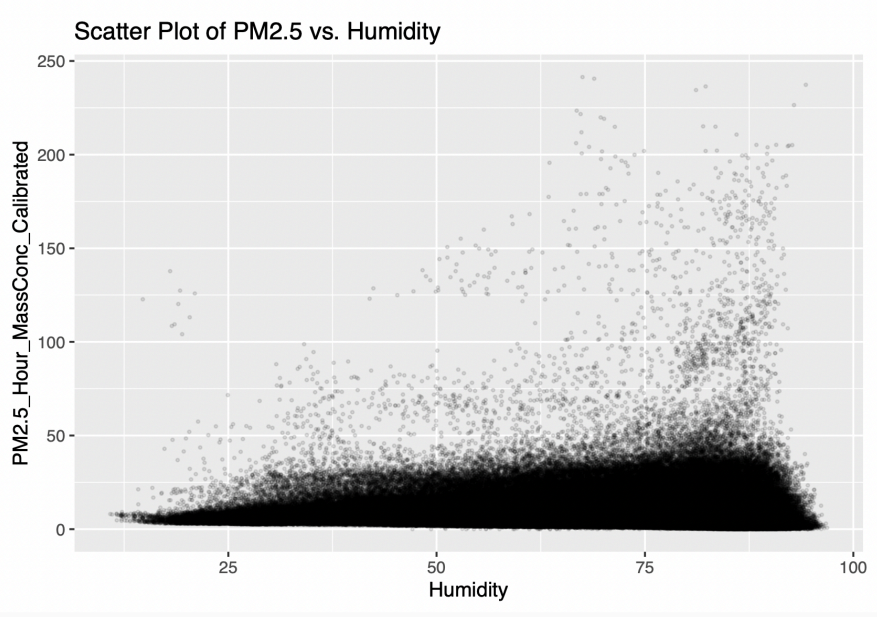
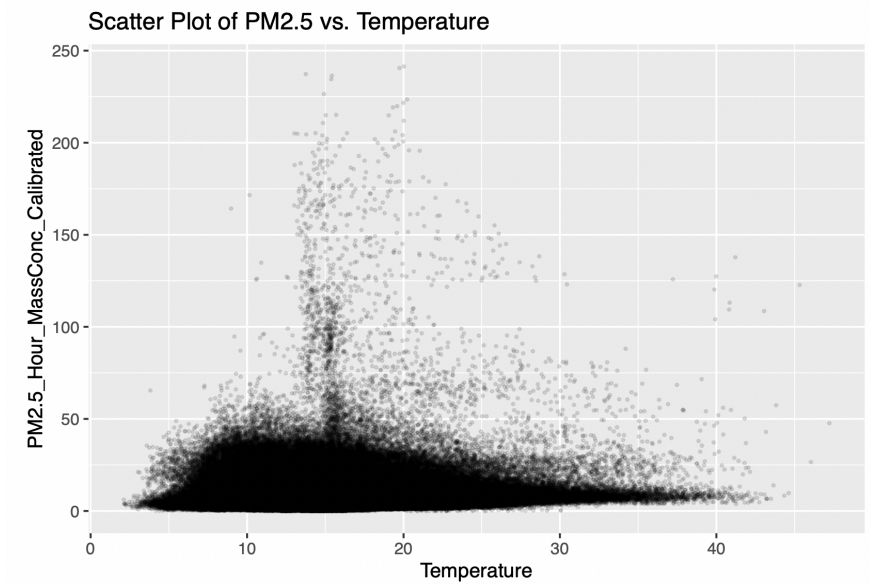



*Figure 10. Scatter Plot of $PM_{2.5}$ vs. Temperature*



## Methodology and Results

My analysis is composed of three sections: feature engineering, parameter estimation, and sensitivity analysis. When I train the model on sample data, I am optimizing our coefficients with respect to our sample. An accurate model must also optimize performance on any sample from the population. In order to evaluate the performance of the model throughout the process I will randomly split 80% of the data into a training subset and the remaining 20% into a testing subset. In order to evaluate the models, I will consider the following objective criteria: mean absolute percentage error (MAPE), adjusted $R^2$, and Akaike Information Criterion (AIC). A good model will minimize the MAPE, maximize adjusted $R^2$, and minimize AIC. The MAPE metric evaluates the average deviation in percentage of actual values from the model's predictions and the adjusted $R^2$ evaluates the proportion of variability in the predicted $PM_{2.5}$ explained by the covariates in the model. The AIC is frequently used in time series analysis to evaluate the model's fit on the data by adding a penalty term for the complexity of the model. I will both train and test my model on these criteria. In feature engineering and selection, my baseline model is ordinary least squares. During the parameter estimation section, however, I explore how including weights, ridge coefficients, and lasso coefficients might improve the model's performance. Below is a summary of methodology from the course textbook[13] that will be utilized in this project by associated chapters:

- Chapter 3: OLS with Multiple Covariates
- Chapter 4: The Gauss-Markov Model and Theorem
- Chapter 7: Frisch-Waugh-Lovell Theorem
- Chapter 11: Leverage Scores and Leave-One-Out Formulas
    - 11.3 Applications of the leave-one-out formulas
        - Cook's Distance
- Chapter 13: Perils of Overfitting
    - 13.2 Variance inflation factor
    - 13.3 Bias-variance trade-off
    - 13.4 Model Selection criteria
        - RSS, $R^2$, adjusted $R^2$
        - Information criteria
        - Cross-Validation (CV)
- Chapter 14: Ridge Regression
- Chapter 15: Lasso
- Chapter 19: Weighted Least Squares

[13] arXiv:2401.00649

# I. Feature Engineering and Selection

### *A. Optimization with Respect to Lag Steps*

In this section I start by determining the number of lag steps to include in the training model that will optimize the performance of the model with respect to my model selection criteria. The lag steps in my model are the indexed back $PM_{2.5}$ readings on the hourly scale. If my model contains just two lag steps, then the main predictors are $PM_{2.5}$ from an hour before and from two hours before to predict the next $PM_{2.5}$ reading.

My method involves iterating over a range of lag steps and also across multiple random seeds to determine which number of lag steps on average produces the highest performing model. In my code I start by looping over a random seed, then I loop across different lag steps. I created a function that generates a random training set from the lagged data frame, fit an ordinary least squares model of $PM_{2.5}$ on the lagged $PM_{2.5}$, and then calculate the selection criteria based on the model's perdictions. I repeat this process over 10 lag steps and also across 20 different random seeds. I normalize my selection criteria values, and take a weighted average of them to produce a "performance score." The number of lag steps that produced the highest performance score the most number of times across the random seed iterations is my "optimal" number of lag steps. I find that the optimal number of lag steps to include in my model is six. During some EDA I discovered that if I include too many lag steps, there is higher prediction error (MAPE), but too few lag steps produces lower fit ($R^2$). The optimal number is a balance of the two.

Once I have the number of lag steps to include in my model, I randomly generate a training subset and testing subset from my data which includes six features of lagged $PM_{2.5}$. The remaining process of feature selection involves determining the individual contributions of humidity, temperature, seasonality, and neighborhood to my predictions. If the contributions are significant then they are worth including in my model. The more features I include, I expect the fit to be better, but the accuracy could be less stable (bias-variance tradeoff).

### *B. Neighborhood Effects - Full Regression and Partial Regression*

I start by identifying the contributions of including neighborhood dummies in my model. The full regression model I evaluate predicts $PM_{2.5}$ on six lag steps plus four neighborhood dummy variables using ordinary least squares. The coefficient estimate for each neighborhood dummy variable reflects its association with the outcome variable. While accounting for the effects of the other predictors in the model. This method produces coefficients that allow us to infer the effect of each neighborhood. In other words, we can determine disparities in $PM_{2.5}$ among the five neighborhoods, including where we tend to see higher $PM_{2.5}$ and lower $PM_{2.5}$. Based on the linear model summary output, the neighborhood effects are significant and their inclusion in the model slightly improves the objective criterion values calculated on the full simple regression without neighborhood effects.

Next I use the Frisch-Waugh-Lovell Theorem to calculate the partial regression coefficients by holding other variables constant. This method explicitly isolates the effect of each neighborhood by regressing the residuals from the full model on each neighborhood dummy separately. As a result, the coefficients obtained from these separate regressions represent the unique contribution of each neighborhood to the outcome variable while controlling for other predictors. The partial regression analysis using separate linear regression models for each neighborhood dummy variable after obtaining residuals from the full model is essentially an implementation of Frisch-Waugh-Lovell theorem. By regressing the residuals from the full model on each neighborhood dummy variable separately, I am isolating the effect of each neighborhood while controlling for the effects of other variables. As it turns out, none of the neighborhood dummy variables have a statistically significant association with the variation in the residuals from the full regression model. This lack of significance implies that, after accounting for other predictors in the model, including lagged variables and possibly other factors like seasonality or temperature, the variation in $PM_{2.5}$ levels that is not explained by these predictors is not significantly associated with the different neighborhoods. This doesn't necessarily mean that $PM_{2.5}$ predictions do not have varied effects by neighborhood as we've seen in the full regression analysis.

### *C. Seasonality Effects - Full Regression and Partial Regression*

To determine if seasonality is an important predictor, meaning the model varies in the summer versus the winter, I essentially repeat the same exact process from the neighborhood analysis. This time I include three season dummy variables in my full model. I come to the same conclusion as neighborhoods: seasonality is significant, as each seasonal dummy coefficient is statistically significant with p-values that are practically zero. In my application of Frisch-Waugh-Lovell in the partial regression I find that none of the seasonal dummy variables have a statistically significant association with the variation in the residuals from the full regression model. This lack of significance implies that, after accounting for other predictors in the model, including lagged variables and possibly other factors like humidity or temperature, the variation in $PM_{2.5}$ levels that is not explained by these predictors is not significantly associated with the different seasons However, $PM_{2.5}$ predictions do have varied effects by season.

### *D. Humidity and Temperature Effects - Feature Selection*

As seen in the EDA, there doesn't seem to be a strong clear relationship between $PM_{2.5}$ and temperature or humidity due to high variation in the data. There are four combinations of options for our model with respect to including these covariates: include neither temperature nor humidity variables in the model, include only humidity, include only temperature, or include both humidity and temperature covariates. After running OLS on each model, the training results show significant coefficients for all models. Evaluating these models on the selection criteria

provides a more holistic approach to the performance of these models. The table reveals that the model that minimizes MAPE, maximizes adjusted $R^2$ and minimizes AIC the most is the model which includes both humidity and temperature I find that for every 1% increase in humidity, the predicted PM2.5 value increases by 0.003 $\mu g/m^3$, holding all else constant. For every 1C° increase in temperature, the predicted PM2.5 value decreases by 0.006 $\mu g/m^3$, holding all else constant.

*Figure 11. Model Selection Criteria for Humidity and Temperature*

| Model_Includes | MAPE | Adj_R2 | AIC |
|---|---|---|---|
| Neither | 16.25 | 0.85774 | 392308 |
| Humidity | 16.43 | 0.85783 | 392253 |
| Temperature | 16.36 | 0.85781 | 392271 |
| Humidity and Temperature | 16.44 | 0.85785 | 392246 |

### *E. Final Feature Selection*

The features that will be included in the model are six lag steps, three season dummies, four neighborhood dummies, temperature, and humidity for a total of 15 predictors. The index *i* ranges from 1 to 20 for each air quality sensor, while *t* is indexed over an hourly timestamp from.

$$\begin{aligned} PM2.5_{t,i} = &\beta_0 + \beta_1\, PM2.5_{t-1,i} + \beta_2\, PM2.5_{t-2,i} + \beta_3\, PM2.5_{t-3,i} + \beta_4\, PM2.5_{t-4,i} \\ &+ \beta_5\, PM2.5_{t-5,i} + \beta_6\, PM2.5_{t-6,i} + \beta_7\, Humidity_{t,i} + \beta_8\, Temperature_{t,i} \\ &+ \beta_9\, Spring_{t,i} + \beta_{10}\, Summer_{t,i} + \beta_{11}\, Winter_{t,i} + \beta_{12}\, Chinatown_{t,i} \\ &+ \beta_{13}\, Potrero\ Hill_{t,i} + \beta_{14}\, SoMa_{t,i} + \beta_{15}\, Tenderloin_{t,i} + \epsilon_{t,i} \end{aligned}$$

## II. Parameter Estimation

The next aspect of the model to explore is the estimates for the coefficients. So far, the model has been evaluated using OLS coefficients. The conditions and considerations for the accuracy and precision of these estimates include the following:

- Linearity
- Homoskedasticity
- Independence of Errors
- Normality of Residuals
- Variance Inflation Factors
- Multicollinearity
- Outliers
- Model Fit

When the OLS model is fitted against actual values from the training set, there appears to be heteroskedasticity and potential non-linearity. The Durbin-Watson test for independence of errors found that we cannot reject the null that there is no auto-correlation. Furthermore, while adjusted $R^2$ is high, residuals do not appear to be normal and there exist extreme outliers and influential points. The Variation Inflation Factors of the model also point to high multicollinearity from the inclusion of highly correlated lag steps. It's possible that ordinary least squares is not the best method for parameter estimation. Heteroscedasticity in the residuals indicates that a Weighted Least Squares (WLS) model may be a better choice than OLS. WLS accounts for heteroscedasticity by assigning different weights to each observation based on the variance of the residuals. I'll also consider Ridge, Lasso, Weighted Ridge (WR), and Weighted Lasso (WL) methods in order to reduce overfitting and account for high multicollinearity among lag step predictors. Ridge and Lasso regressions were performed with 10-fold cross-validation. The lambda values were found by minimizing cross-validated error. Weights were calculated by taking the inverse of squared residuals to adjust for heteroskedasticity.

*Figure 12. Normal Q-Q Plot for OLS Model (Left)*

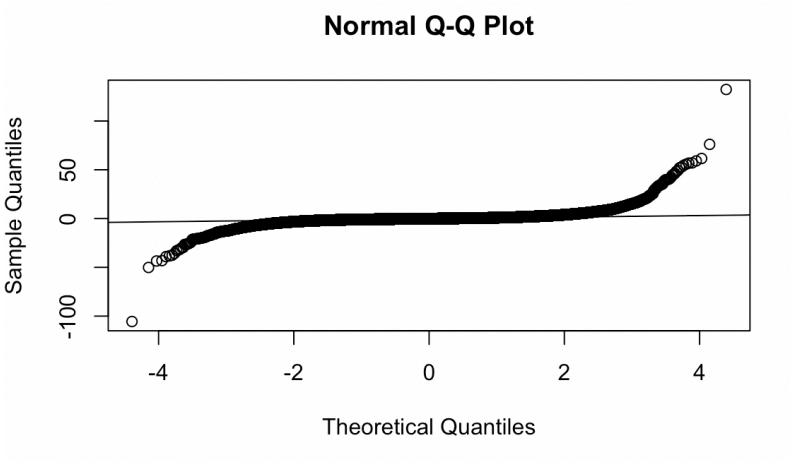
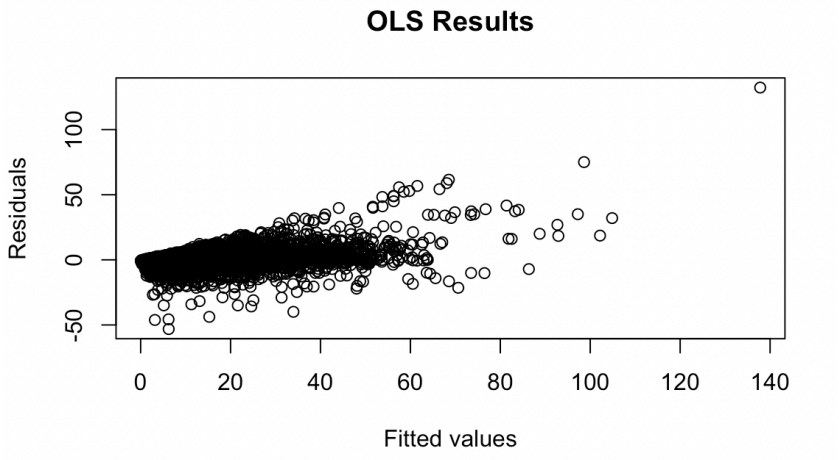



*Figure 13. Residual Plot for OLS Model*

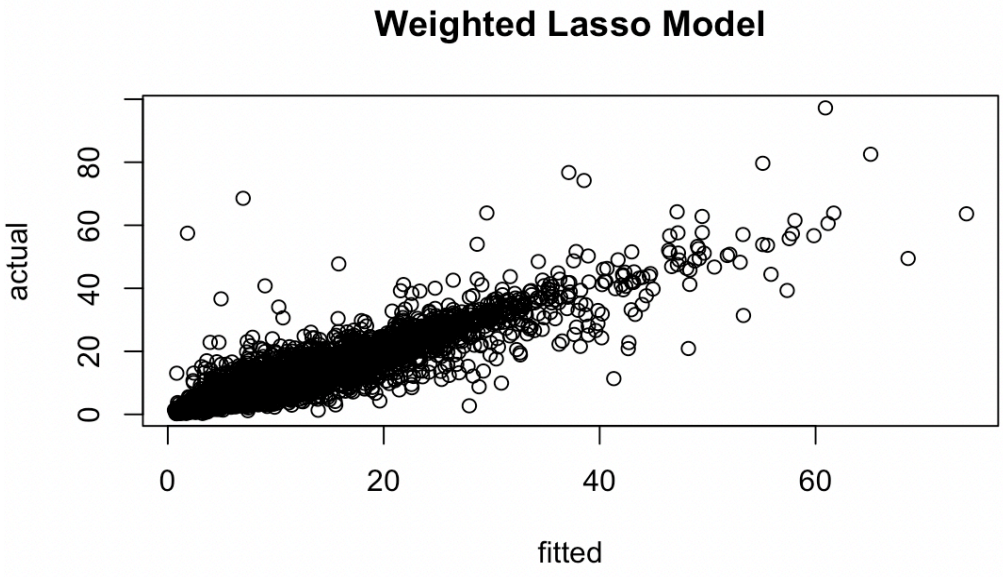


After checking conditions, calculating coefficients, and visually inspecting the residual plots for each model, predictions from the testing set are used to calculate the model performance criteria. Results from each model are in the table below. Each model appears to perform relatively well, but the model with the lowest APE, Adjusted $R^2$ is the Weighted Lasso model.

*Figure 14. Model Selection Criteria for Regression Analysis*

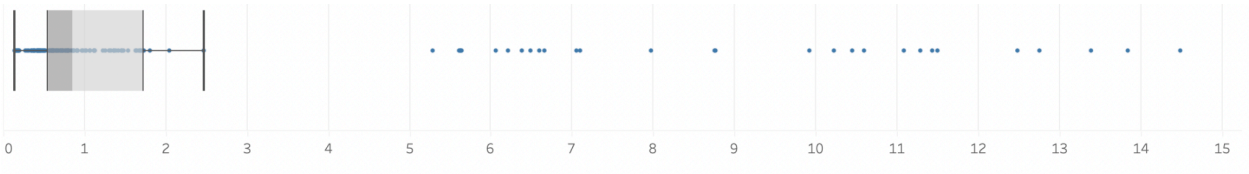
| Model | MAPE | Adj_R_2 | AIC |
|---|---|---|---|
| OLS | 16.37 | 0.8580 | 31145 |
| WLS | 16.37 | 1.0000 | 31145 |
| Ridge | 16.54 | 0.8769 | 31350 |
| Lasso | 16.39 | 0.8779 | 31164 |
| Weighted Ridge | 16.48 | 0.8774 | 31249 |
| Weighted Lasso | 16.35 | 0.8779 | 31172 |

*Figure 15. Residual Plot for Weighted Lasso Regression*



## III. Sensitivity Analysis

*Figure 16. Box Plot of $PM_{2.5}$*




As seen in the box plot and figures above, there are high outliers in the data that seem to influence the model significantly. During September through October 2020 there was an unusually hazardous wildfire season wherein air pollution in the Bay reached record highs. You might recall the bright orange sky that residents awoke to on September 19, 2020[14] experience in the Bay Area and covered internationally by news outlets. In the past few years, California has thankfully not experienced such extreme wildfire events, and so many of these high outliers are sequestered in this small time frame from earlier in the dataset. It would be a reasonable choice to exclude extreme outliers in the data that affect the accuracy of the predictions. Rather than throwing out data from timestamps when wildfire smoke affected the City, we can analytically calculate the individual effects that each observation has on the model; in other words the extreme outliers in the model. Two different methods will be considered in the determination of highest contributing outliers: one involving taking out the top 10% of Cook's distances and the other involving taking out the top 10% of observations that contribute the most to the OLS coefficients under the Leave-One-Out (LOO) formula. In the first method, Cook's distance for each observation is found by calculating the leverage scores and standardized residuals. In the second method, the coefficient contribution of each observation is calculated from the difference between the full OLS estimator and the leave-i-out OLS estimator. The largest calculated differences are the largest contributors. Because the top 10% values are taken, approximately 9,000 observations are thrown out from the model under each method.

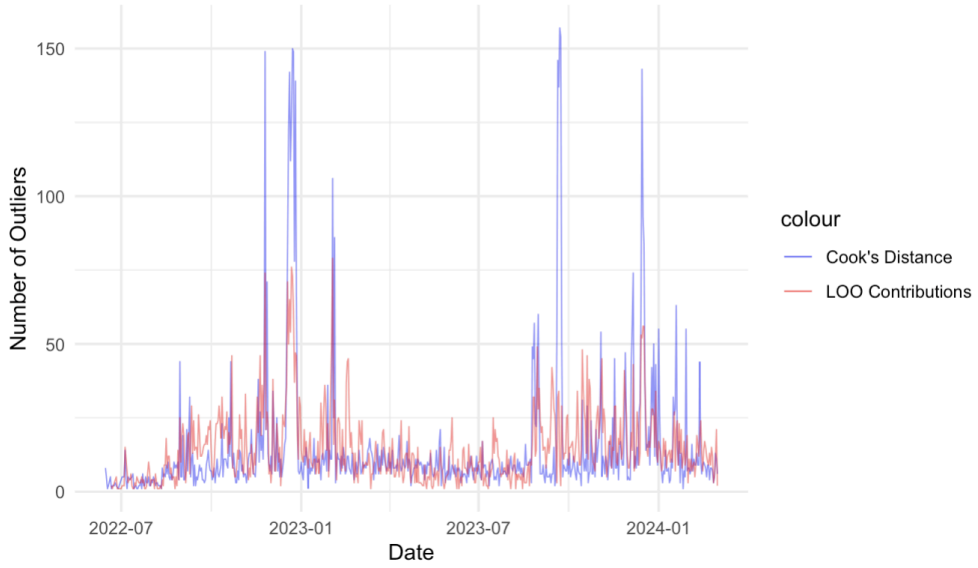
$$\text{cook}_i = \text{standr}_i^2 \times \frac{h_{ii}}{p(1-h_{ii})}$$

$$\hat{\beta}_{[-i]} = \hat{\beta} - (1-h_{ii})^{-1}(X^{\mathrm{T}}X)^{-1}x_i\hat{\varepsilon}_i$$

Taking the corresponding timestamp from each outlier and plotting them on a time-series graph we can compare the outliers from either method. It seems that the highest extreme values are better captured by Cook's distance, while the outliers that are high relative to their adjacent counterparts are better captured by the LOO method. It seems that using the LOO method would be better at capturing more contextualized outliers which is better for the predictive model. If we are predicting $PM_{2.5}$ on extreme outliers, we should still get a pretty close extreme prediction.

[14] Hamedy, Saba. "In California's Smoke-Filled Horizon, It's Become Hard to Breathe." *CNN*, Cable News Network, 11 Sept. 2020, www.cnn.com/2020/09/10/us/california-air-quality-wildfires-trnd/index.html.

*Figure 17. Time Series Comparison for Outliers*




## Discussion

In the feature selection process, it is decided that including both temperature and humidity produces more accurate and precise OLS coefficient estimates, relative to their exclusion from the model. The full regression analysis indicated that seasonality and neighborhood effects are significant, so dummy variables for both categories should be included in order to distinguish the individual effects across seasons and across neighborhood location. To calculate the optimal number of lag steps to include in the model, which serve as the primary predictors of the next step of $PM_{2.5}$, there needs to be a bias-variance tradeoff consideration. By evaluating training models across multiple random seeds and calculating different selection criteria for each model, I take a more holistic approach to calculating the optimal lag steps. Feature selections are established at the start of the analysis. Moving forward, all models that are evaluated consider six steps back of $PM_{2.5}$, temperature, humidity, seasonal effects, and neighborhood effects.

In the parameter estimation estimation section I consider the performance of OLS, WLS, Ridge, Lasso, WR and WL under MAPE, adjusted $R^2$ and AIC. No single model stands out among the rest as the "best" performer on the test set, but it seems like the Weighted Lasso Model performs slightly better than the rest with respect to MAPE and adjusted $R^2$. Included in the model options is whether or not to exclude outliers from the data and if so, under which outlier classification method?

Running OLS without the outliers calculated using Cook's distance and without the outliers calculated using the Leave-One-Out formula on the training model, I can also add to Table 2 the performance criteria from these models. As expected, removing outliers significantly decreases the mean absolute percentage error, indicating more accurate predictions under OLS. Previously it was concluded from the time series comparison that the LOO method is better at identifying

the relative outliers. Even if there are extremely high outliers, they may still play an important role in dictating a close model fit, which we can see in the higher adjusted $R^2$ using LOO. The AIC is reduced for both outlier exclusion methods as well. Because the fit is much better under LOO, the OLS model that excludes the top 10% of observations that contribute the highest to the estimates is the highest performing. In many cases, it is not a good idea to simply throw out observations that have high contributions to the model because it violates our assumption of theoretically randomly sampled data. However, in this context, the extreme outliers from our data are likely to have occurred from sensor malfunctions (which are known to occur) or extreme wildfire events from the past. Excluding these outliers might make our sample a better representation of the population I am estimating.

The so-called "best" model under this analysis is the OLS model that excludes the top 10% of observations that contribute the highest to the estimates. The residual plot is shown below. Showing a strong positive relationship between the actual values and the fitted values of $PM_{2.5}$. The coefficient estimates from this model are also shown in the formula below.

$$\begin{aligned} PM2.5_{t,i} = &-2.43 + 0.75\,PM2.5_{t-1,i} + 0.15\,PM2.5_{t-2,i} + 0.026\,PM2.5_{t-3,i} + 0.04\,PM2.5_{t-4,i} \\ &+ 0.007\,PM2.5_{t-5,i} + 0.0048\,PM2.5_{t-6,i} + 0.029\,Humidity_{t,i} + 0.075\,Temperature_{t,i} \\ &- 0.29\,Spring_{t,i} - 0.72\,Summer_{t,i} - 0.22\,Winter_{t,i} - .44\,Chinatown_{t,i} \\ &- 0.28\,Potrero\ Hill_{t,i} - 0.29\,SoMa_{t,i} - 0.34\,Tenderloin_{t,i} + \epsilon_{t,i} \end{aligned}$$

*Figure 18. Residual Plot for "Best" Model*

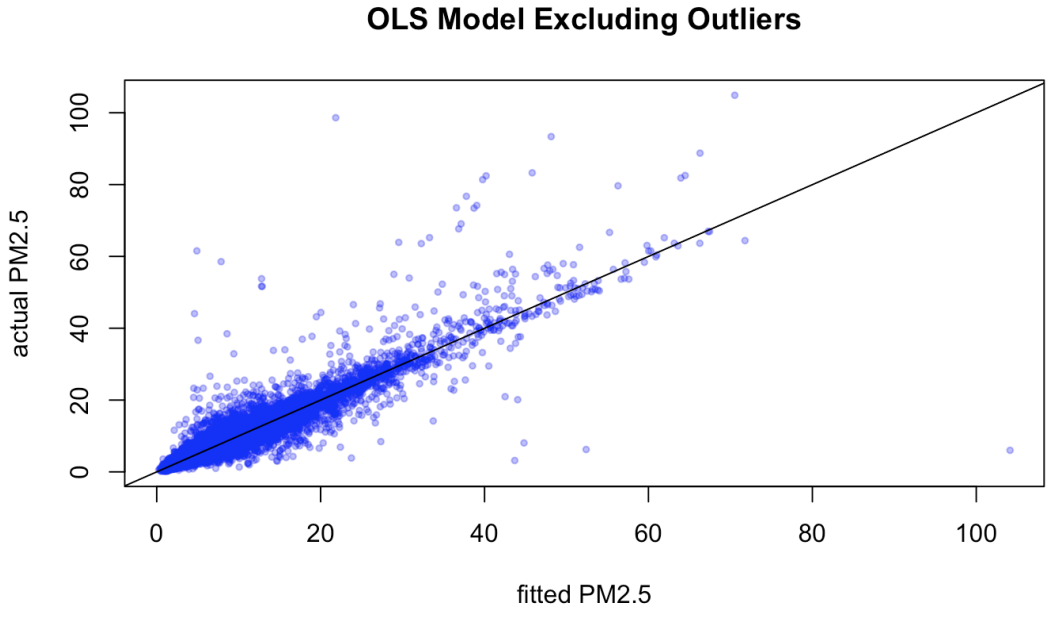
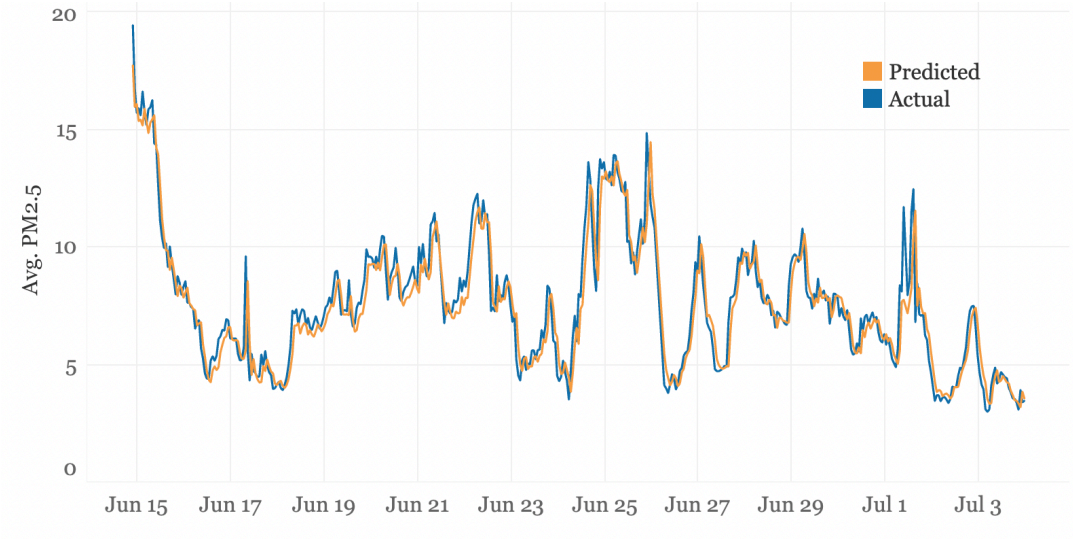


*Figure 19. Sample of Actual vs. Predicted $PM_{2.5}$ under "Best" Model*




**Conclusion**

This model demonstrates that PM2.5 at a given air quality sensor can be predicted using the previous six hourly PM2.5 values with an average of 85% accuracy. These predictions are strengthened by accounting for variations in temperature and humidity. Furthermore, there are differential effects by neighborhood and season included in the model for higher precision. Since PM2.5 can be predicted with high accuracy using lagged values, the behavior of PM2.5 does not appear to be extremely volatile. This project provides potentially valuable insights into the factors influencing hourly $PM_{2.5}$ levels in Eastern San Francisco and demonstrates the effectiveness of statistical linear modeling techniques for forecasting air quality. By incorporating temperature, humidity, seasonality, and neighborhood effects into predictive models, we have gained a deeper understanding of the spatiotemporal dynamics of $PM_{2.5}$ in disadvantaged communities. The evaluation of various modeling approaches has highlighted the importance of considering both the accuracy and precision of model estimates, as well as the impact of outliers on model performance. Moving forward, this study sets the stage for targeted interventions aimed at improving air quality and addressing environmental justice concerns in Eastern San Francisco. The high accuracy of the model demonstrates the modelability of $PM_{2.5}$ and encourages further analysis of air quality in San Francisco through an equity lens. Additionally, it underscores the need for continued monitoring and analysis to mitigate the adverse effects of poor air quality on public health and community well-being.

**Code**

Full methods, results, and outputs can be found in the R markdown file in Forecasting-PM2.5-in-SF/Stat230_Final_Code.Rmd