MATH 448 - FINAL PROJECT REPORT: PREDICTING POLLUTION IN CA

Student ID: 921460888

Name: Zain Mirza

Table of contents (click to see section)

I. Executive Summary 2
II. Introduction 2
III. The data 3
IV. Data Cleansing + Preprocessing 3
1. Filter the Data 3
2. Remove Outliers 4
3. Data Summary + Visualization 4
V. Model selection 5
1. Linear Regression Model 5
2. Ridge Regression Model 9
3. Lasso Regression Model 10
4. Best-Subset Selection 10
5. PCR Method 11
6. PLS Method 12
7. Regression Tree 12
VI. Conclusion 13
VII. Future Direction 14

I. Executive Summary

In recent years, the state of California has succumbed an enormous amount of damage from greenhouse gases. These gases have created such a problem in the state of California that it has drawn attention towards new technologies such as electrical vehicles and solar power. All these new elements brought into society to keep away what is primarily harming our environment, the pollution. Thankfully, it's because of these new technologies that we are capable of tracking observed data to understand patterns and measurements such as Air Quality Index (AQI). Looking at pollution data allows us to better comprehend our atmosphere and regional environment. The data which was used to conduct this project was published through Kaggle. A user put together a dataset of collected quantitative data provided by the US Environmental Protection Agency (EPA).

For this project, I used multiple statistical methods to accurately predict major gas pollutants which influence pollution the most. My main idea for this project was to find which response variables are most significant in increasing a region's pollution.

I used many different models to predict such as Linear Regression, Ridge Regression, and Lasso. I also included a Best-Subset selection model to help classify the optimal number of variables to best predict the response variables. A PCR/PLS test was also conducted in this project, as well as a Regression Tree. Each of these models were compared to identify which tended the best accuracy with the lowest Root Mean Square Error (RMSE).

After using the methods that are mentioned above, it can be concluded that the Linear Regression model is the best model for the given dataset. From this, it's clear that each of the variables hold a relatable correlation with one another.

As far as future work for this project, I believe that it might be interesting to further investigate and understand deep learning models for this project. Furthermore, I believe adding more variables would be a bonus in the dataset and would lead to more clear models and better predictions.

II. Introduction

In todays' California society, many people are concerned with the recent inconsistency of the weather. California temperature has reached new all-time highs as well as lows. With the pandemic already being a giant headache, this is an added dose especially for residents in California. Data is an important asset that is used as a tool and sense of guidance for making plans of action. As long as the topic of pollution is in discussion, we will continue to collect and use data to understand the proper steps that are needed to help protect and keep our environment clean and safe to breathe.

Regression is a statistical task which determines the strength and character of the relationship between some dependent *y* variable and a series of predictor variables. This form of analysis is key with the variables and numerical data given in the set. This will also play an important role as we try and identify patterns amongst each variable in the database.

III. The data

The title of the dataset is "U.S Pollution Data" and the website that sourced the set is public and free, so no permission was needed permission to use this data.

The data includes a total of 22 variables, most of which was numeric including only few categorical. But this was not an issue during my pre-processing stages. The dataset is designed revolving around 4 major gas pollutants: Nitrogen Dioxide (NO2), Sulphur Dioxide (SO2), Carbon Monoxide (CO), and the Ozone (O3). The data collected was for every day in every County and State from the year 2000 to March 2021. There are a total of 1.4 million observations

IV. Data Cleansing + Preprocessing

1. Filter the Data

The first thing that was needed to fix in the set was of course filtering out the states. The dataset incorporated every state in the United States, and we had to filter it in order to only display data in California. This required some use of Excel primarily because there were so many observations in the set and I would constantly find myself running into issues with RStudio not being able to load the data.

Next, I wanted to filter the set so that it shows me monthly observed data. The data given was for every single day so I managed to keep it uniform by making sure I only observed the first day of each month.

2. Remove Outliers

Finally, I removed any variables that were deemed to be insignificant in reaching my objective. These outliers include *address, city,* as well as others such as the *max/min first hour AQI/Mean* for the given chemical pollutant.

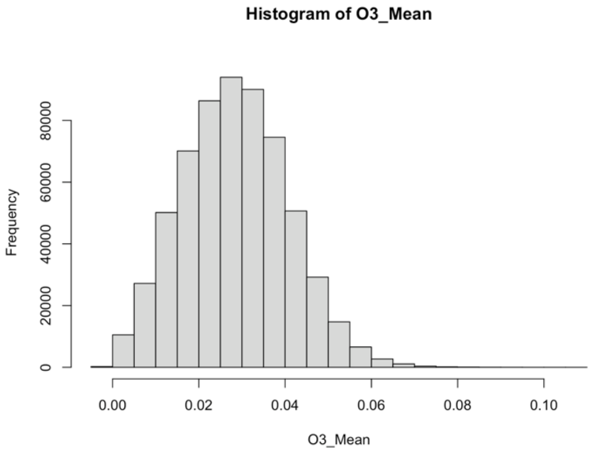
Successfully performing these actions led me to a new and more concise dataset of about 595,000 observations. Also being able to jump from 22 variables down to 6 variables. From here I was able to upload this data into RStudio, splitting the training and test data into the 80/20 ratio.

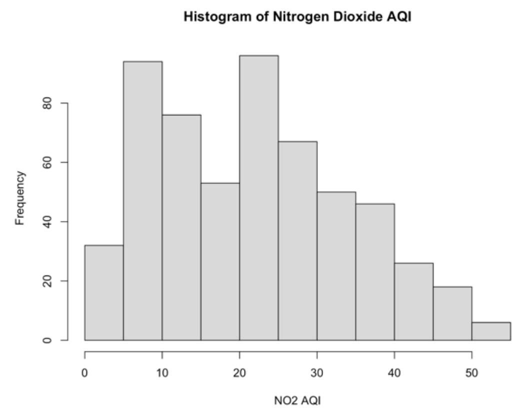
3. Data Summary + Visualization

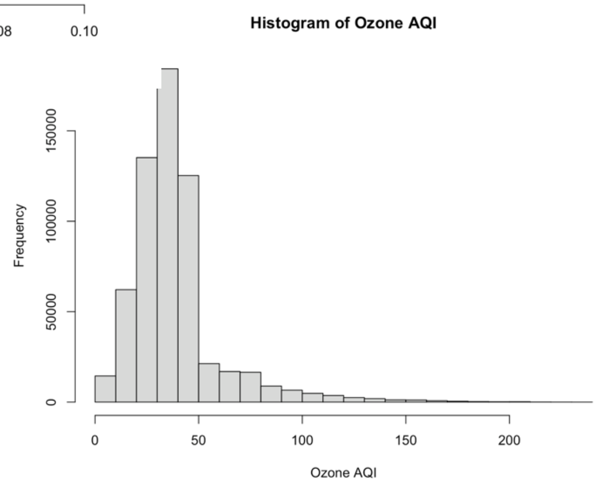
Here is a summary of the cleansed data uploaded into R. You can see in the summary all the variables that were used in this project.

```
> summary(PData2)
      Date                          Year          Month             Day            State              County         
 Min.   :2000-01-01 00:00:00   Min.   :2000   Min.   : 1.000   Min.   : 1.00   Length:608699      Length:608699     
 1st Qu.:2006-07-18 00:00:00   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   Class :character   Class :character  
 Median :2012-01-07 00:00:00   Median :2012   Median : 7.000   Median :16.00   Mode  :character   Mode  :character  
 Mean   :2011-07-16 14:35:09   Mean   :2011   Mean   : 6.509   Mean   :15.74                                        
 3rd Qu.:2016-09-21 00:00:00   3rd Qu.:2016   3rd Qu.: 9.000   3rd Qu.:23.00                                        
 Max.   :2021-10-31 00:00:00   Max.   :2021   Max.   :12.000   Max.   :31.00                                        
     City              O3_Mean             O3_AQI          CO_Mean            CO_AQI           SO2_Mean       
 Length:608699      Min.   :-0.000706   Min.   :  0.00   Min.   :-0.4375   Min.   :  0.000   Min.   : -2.5083  
 Class :character   1st Qu.: 0.019647   1st Qu.: 27.00   1st Qu.: 0.1792   1st Qu.:  2.000   1st Qu.:  0.1875  
 Mode  :character   Median : 0.028235   Median : 35.00   Median : 0.2625   Median :  5.000   Median :  0.6667  
                    Mean   : 0.028477   Mean   : 39.11   Mean   : 0.3373   Mean   :  5.377   Mean   :  1.5234  
                    3rd Qu.: 0.036765   3rd Qu.: 44.00   3rd Qu.: 0.4208   3rd Qu.:  7.000   3rd Qu.:  1.7727  
                    Max.   : 0.107353   Max.   :237.00   Max.   : 7.5083   Max.   :201.000   Max.   :321.6250  
    SO2_AQI           NO2_Mean          NO2_AQI     
 Min.   :  0.000   Min.   : -4.629   Min.   :  0.00  
 1st Qu.:  0.000   1st Qu.:  4.978   1st Qu.: 10.00  
 Median :  1.000   Median :  9.542   Median : 20.00  
 Mean   :  5.569   Mean   : 11.738   Mean   : 22.12  
 3rd Qu.:  6.000   3rd Qu.: 16.304   3rd Qu.: 31.00  
 Max.   :200.000   Max.   :140.650   Max.   :133.00  
```

Below includes some of the histograms for the numerical variables that were used in the regression analysis. Each variable is different as far as its distribution of value goes. This was unique to me as I grew more interest in wanting to explore for the best response and predictor variables.







V. Model selection

Throughout each process, the objective remained consistent, which is to determine which model is best for accurately predicting for the Nitrogen Dioxide Air Quality Index, otherwise the NO2 AQI. According to the EPA, the same agency that collected the dataset, NO2 levels are classified as the most harmful gas pollutant to the atmosphere. It was because of this I chose this variable as the response.

1. Linear Regression Model

The objective of a linear regression model is to predict the value of an output variable (response) based on the value of an input (predictor) variables. The idea behind the model is to look at primarily whether a set of predictor variables do a good job in predicting a response variable? It also illustrates which specific variables are significant predictors to the response.

Linear Regression will be the first method used for the project because it is the simplest method and computes a clean easy to interpret model. The results from the model are shown below:

The first model includes all predictor variables, including the N02 mean for each observation.

```
lm(formula = NO2_AQI ~ County + CO_AQI + CO_Mean + NO2_AQI +
    NO2_Mean + SO2_AQI + SO2_Mean, data = PData2.train)

Residuals:
    Min      1Q  Median      3Q     Max
-70.917  -3.499  -0.872   2.776  95.216

Coefficients:
                              Estimate Std. Error  t value Pr(>|t|)
(Intercept)                  10.052542   0.178318   56.374  < 2e-16 ***
CountyAdair                  -8.565525   0.276902  -30.933  < 2e-16 ***
CountyAdams                  -1.435332   0.190287   -7.543 4.60e-14 ***
CountyAlameda                -5.960421   0.197727  -30.145  < 2e-16 ***
CountyAlexandria City        -3.690406   0.217149  -16.995  < 2e-16 ***
CountyAllegheny              -2.453564   0.201974  -12.148  < 2e-16 ***
CountyAnoka                  -3.557453   0.219791  -16.186  < 2e-16 ***
CountyAroostook              -7.404339   0.199333  -37.146  < 2e-16 ***
CountyAthens                 -5.963518   0.464600  -12.836  < 2e-16 ***
CountyBaltimore              -2.675603   0.199386  -13.419  < 2e-16 ***
CountyBeaver                 -2.933979   0.231441  -12.677  < 2e-16 ***
CountyBerks                  -3.905520   0.212173  -18.407  < 2e-16 ***
CountyBernalillo             -2.075768   0.206111  -10.071  < 2e-16 ***
CountyBexar                  -2.827143   0.233895  -12.087  < 2e-16 ***
CountyBlair                  -2.807746   0.230989  -12.155  < 2e-16 ***
CountyBlount                 -9.227339   0.214067  -43.105  < 2e-16 ***
CountyBoyd                   -2.099911   0.280017   -7.499 6.43e-14 ***
CountyBronx                  -4.215527   0.193354  -21.802  < 2e-16 ***
CountyBucks                  -2.307578   0.210833  -10.945  < 2e-16 ***
CountyBurleigh               -5.502088   0.234491  -23.464  < 2e-16 ***
CountyCambria                -4.448934   0.197347  -22.544  < 2e-16 ***
CountyCamden                 -3.623430   0.195566  -18.528  < 2e-16 ***
CountyCass                   -3.699186   0.214402  -17.254  < 2e-16 ***
CountyCharleston             -7.570935   0.238078  -31.800  < 2e-16 ***
CountyCherokee               -3.668213   0.235368  -15.585  < 2e-16 ***
CountyClark                  -1.777190   0.226888   -7.833 4.78e-15 ***
CountyContra Costa           -5.441731   0.182362  -29.840  < 2e-16 ***
CountyCook                   -2.114402   0.193036  -10.953  < 2e-16 ***
CountyCumberland             -2.524341   0.203656  -12.395  < 2e-16 ***
CountyCuyahoga               -3.888025   0.192363  -20.212  < 2e-16 ***
CountyDallas                 -2.910353   0.194092  -14.995  < 2e-16 ***
CountyDauphin                -2.439186   0.309837   -7.872 3.48e-15 ***
CountyDaviess                -3.405688   0.388144   -8.774  < 2e-16 ***
CountyDeKalb                 -2.111665   0.206286  -10.237  < 2e-16 ***
CountyDenver                 -3.093895   0.194829  -15.880  < 2e-16 ***
CountyDistrict of Columbia   -2.790007   0.190914  -14.614  < 2e-16 ***
CountyDuchesne               -7.222905   1.174403   -6.150 7.74e-10 ***
CountyEast Baton Rouge       -2.469474   0.192404  -12.835  < 2e-16 ***
CountyEl Paso                -0.450244   0.189458   -2.376  0.01748 *
CountyErie                   -1.650461   0.213172   -7.742 9.78e-15 ***
CountyEssex                  -3.323113   0.202550  -16.406  < 2e-16 ***
CountyFairbanks North Star   -5.035121   0.243675  -20.663  < 2e-16 ***
CountyFairfax                -2.289479   0.192549  -11.890  < 2e-16 ***
CountyFairfield               0.710527   0.267877    2.652  0.00799 **
CountyFayette                -1.427751   0.289310   -4.935 8.02e-07 ***
CountyForsyth                -0.537887   0.337305   -1.595  0.11079
CountyFremont                -6.622107   0.389280  -17.011  < 2e-16 ***
CountyFresno                 -3.908122   0.197383  -19.800  < 2e-16 ***
CountyGarrett                -8.951258   0.218305  -41.003  < 2e-16 ***
CountyHamilton               -4.285143   0.208001  -20.602  < 2e-16 ***
CountyHampton City           -6.883500   0.221303  -31.104  < 2e-16 ***
CountyHarris                 -3.466675   0.185953  -18.643  < 2e-16 ***
CountyHartford               -3.856785   0.229305  -16.819  < 2e-16 ***
CountyHaywood                -3.799088   0.469985   -8.083 6.31e-16 ***
CountyHenderson              -0.621781   0.382241   -1.627  0.10381
CountyHenrico                -4.631508   0.209819  -22.074  < 2e-16 ***
CountyHillsborough           -4.487655   0.219893  -20.408  < 2e-16 ***
CountyHinds                  -4.348137   0.242398  -17.938  < 2e-16 ***
CountyHonolulu               -4.812887   0.187814  -25.626  < 2e-16 ***
CountyHumboldt               -7.137462   0.193162  -36.951  < 2e-16 ***
CountyImperial                0.508195   0.194054    2.619  0.00882 **
CountyJackson                -9.122477   0.315344  -28.929  < 2e-16 ***
CountyJefferson              -2.513048   0.200589  -12.528  < 2e-16 ***
CountyKay                    -7.696717   0.293672  -26.209  < 2e-16 ***
CountyKent                   -1.769054   0.263956   -6.702 2.06e-11 ***
CountyKern                    0.769881   0.639062    1.205  0.22832
CountyKing                   -2.289748   0.217500  -10.528  < 2e-16 ***
CountyLackawanna             -0.339624   0.232382   -1.461  0.14388
CountyLancaster              -2.419225   0.230791  -10.482  < 2e-16 ***
CountyLaramie                -4.205406   0.208193  -20.200  < 2e-16 ***
CountyLawrence               -4.138034   0.231474  -17.877  < 2e-16 ***
CountyLinn                   -7.103134   0.225143  -31.549  < 2e-16 ***
CountyLitchfield             -7.480139   0.231746  -32.277  < 2e-16 ***
CountyLos Angeles            -4.430973   0.182300  -24.306  < 2e-16 ***
CountyLuzerne                -1.560621   0.321502   -4.854 1.21e-06 ***
CountyMaricopa               -0.078017   0.185758   -0.420  0.67449
CountyMarion                 -3.081842   0.198555  -15.521  < 2e-16 ***
CountyMcCracken              -0.924976   0.379081   -2.440  0.01469 *
CountyMcLennan               -6.384612   0.207365  -30.789  < 2e-16 ***
CountyMecklenburg            -2.425318   0.189323  -12.810  < 2e-16 ***
CountyMedina                 -3.316567   0.416826   -7.957 1.77e-15 ***
CountyMeigs                  -1.229228   0.453641   -2.710  0.00673 **
CountyMilwaukee              -2.644552   0.246182  -10.742  < 2e-16 ***
CountyMinnehaha              -4.971538   0.206642  -24.059  < 2e-16 ***
CountyMonroe                 -4.951947   0.414361  -11.951  < 2e-16 ***
CountyMontgomery             -2.089748   0.231677   -9.020  < 2e-16 ***
CountyMultnomah              -6.003894   0.201095  -29.856  < 2e-16 ***
CountyNew Castle             -4.104559   0.219186  -18.726  < 2e-16 ***
CountyNew Haven              -2.051879   0.199938  -10.263  < 2e-16 ***
CountyNorthampton            -2.686907   0.229900  -11.687  < 2e-16 ***
CountyOklahoma               -4.267768   0.223205  -19.120  < 2e-16 ***
CountyOrange                 -3.702113   0.186426  -19.858  < 2e-16 ***
CountyOttawa                 -4.324873   0.274516  -15.755  < 2e-16 ***
CountyPhiladelphia           -5.444619   0.196837  -27.661  < 2e-16 ***
CountyPima                   -1.892160   0.192016   -9.854  < 2e-16 ***
CountyPolk                   -4.223908   0.202925  -20.815  < 2e-16 ***
```

```
CountyPrince George's      -4.873743   0.207500  -23.488  < 2e-16 ***
CountyProvidence           -5.185274   0.205045  -25.289  < 2e-16 ***
CountyPulaski              -1.116074   0.188464   -5.922 3.18e-09 ***
CountyQueens               -3.810452   0.192417  -19.803  < 2e-16 ***
CountyRichland             -5.256003   0.702623   -7.481 7.41e-14 ***
CountyRiverside            -3.523848   0.186891  -18.855  < 2e-16 ***
CountyRoanoke              -5.595621   0.231067  -24.216  < 2e-16 ***
CountyRockingham           -7.438489   0.225421  -32.998  < 2e-16 ***
CountyRutland              -4.852561   0.230824  -21.023  < 2e-16 ***
CountySacramento           -3.578095   0.189341  -18.898  < 2e-16 ***
CountySaint Clair          -3.789492   0.196000  -19.334  < 2e-16 ***
CountySaint Louis          -1.547622   0.225901   -6.851 7.35e-12 ***
CountySalt Lake            -3.436557   0.192127  -17.887  < 2e-16 ***
CountySan Bernardino       -1.154484   0.187233   -6.166 7.01e-10 ***
CountySan Diego            -2.793705   0.186291  -14.996  < 2e-16 ***
CountySan Francisco        -5.310146   0.207509  -25.590  < 2e-16 ***
CountySanta Barbara        -6.989545   0.182638  -38.270  < 2e-16 ***
CountySanta Clara          -5.226569   0.196799  -26.558  < 2e-16 ***
CountySanta Cruz           -5.156835   0.209529  -24.611  < 2e-16 ***
CountyScott                -5.169572   0.196752  -26.275  < 2e-16 ***
CountyShelby               -6.504102   0.857174   -7.588 3.26e-14 ***
CountySolano               -5.252967   0.191166  -27.479  < 2e-16 ***
CountySt. Louis City       -2.010335   0.194624  -10.329  < 2e-16 ***
CountySteuben              -8.588419   0.266273  -32.254  < 2e-16 ***
CountySuffolk              -4.602935   0.187913  -24.495  < 2e-16 ***
CountySumner               -4.792392   0.444472  -10.782  < 2e-16 ***
CountySweetwater           -2.366475   0.345569   -6.848 7.49e-12 ***
CountyTravis               -5.014050   0.326434  -15.360  < 2e-16 ***
CountyTulsa                -3.782201   0.202211  -18.704  < 2e-16 ***
CountyUinta                -6.086594   0.303887  -20.029  < 2e-16 ***
CountyUintah               -6.183662   0.282607  -21.881  < 2e-16 ***
CountyUnion                -7.772483   0.396353  -19.610  < 2e-16 ***
CountyVentura              -4.561486   0.245148  -18.607  < 2e-16 ***
CountyWake                 -4.627754   0.216194  -21.406  < 2e-16 ***
CountyWashington           -3.561263   0.205692  -17.314  < 2e-16 ***
CountyWashoe               -2.766484   0.206353  -13.407  < 2e-16 ***
CountyWayne                -2.543540   0.258470   -9.841  < 2e-16 ***
CountyWestmoreland         -4.484367   0.231716  -19.353  < 2e-16 ***
CountyWyandotte            -2.147325   0.193942  -11.072  < 2e-16 ***
CountyYork                 -2.482377   0.201875  -12.297  < 2e-16 ***
CO_AQI                      0.649651   0.005048  128.702  < 2e-16 ***
CO_Mean                   -11.463936   0.090664 -126.445  < 2e-16 ***
NO2_Mean                    1.377978   0.001549  889.670  < 2e-16 ***
SO2_AQI                     0.076671   0.001465   52.319  < 2e-16 ***
SO2_Mean                   -0.289669   0.006536  -44.319  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.687 on 486818 degrees of freedom
Multiple R-squared:  0.8485,    Adjusted R-squared:  0.8484
F-statistic: 1.947e+04 on 140 and 486818 DF,  p-value: < 2.2e-16
```

From this model you can clearly see every County in California is factored in as predictor values. Something to note for AQI and in this model is that a negative coefficient indicates a decreased NO2 AQI value. Based on the measure of AQI a lower index is preferred over a higher index. A lower index illustrates a cleaner and more breathable atmosphere. For example, if you see from the model Kern County depicts a positive coefficient given the NO2 AQI as the response. This can be explained because Kern County is in none other than Bakersfield, CA. Bakersfield is known for its high air pollution given the area’s high-emission industries. The climate conditions of the region also factor into play as it causes polluted air to become trapped in the valley. From this first model we can see that the R^2 value is very high at .85. Depicting that this model fits the data quite accurately. This is because the NO2 mean is included in this model. In my perspective, by allowing this to be a predictor variable, it leads to a higher R^2 value.

In the second model you will see that I removed the NO2 mean variable because I wanted to predict for the NO2 AQI without any biasness in the predictor variables. Below lies the results from the second model:

```
lm(formula = NO2_AQI ~ County + CO_AQI + CO_Mean + O3_AQI + O3_Mean +
    SO2_AQI + SO2_Mean, data = PData2.train)

Residuals:
     Min       1Q   Median       3Q      Max
-196.488   -5.485   -0.744    4.893  115.565

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                2.188e+01  2.771e-01   78.959  < 2e-16 ***
CountyAdair               -1.478e+01  4.270e-01  -34.627  < 2e-16 ***
CountyAdams                9.876e-02  2.935e-01    0.337 0.736490
CountyAlameda             -6.181e+00  3.050e-01  -20.263  < 2e-16 ***
CountyAlexandria City      1.702e+00  3.348e-01    5.083 3.72e-07 ***
CountyAllegheny           -2.101e+00  3.115e-01   -6.745 1.53e-11 ***
CountyAnoka               -7.807e+00  3.389e-01  -23.035  < 2e-16 ***
CountyAroostook           -1.518e+01  3.071e-01  -49.436  < 2e-16 ***
CountyAthens              -1.662e+01  7.163e-01  -23.200  < 2e-16 ***
CountyBaltimore           -4.028e+00  3.075e-01  -13.100  < 2e-16 ***
CountyBeaver              -4.562e+00  3.571e-01  -12.777  < 2e-16 ***
CountyBerks               -5.755e+00  3.273e-01  -17.583  < 2e-16 ***
CountyBernalillo           3.006e-01  3.180e-01    0.945 0.344549
CountyBexar               -8.136e+00  3.606e-01  -22.559  < 2e-16 ***
CountyBlair               -6.021e+00  3.563e-01  -16.898  < 2e-16 ***
CountyBlount              -1.674e+01  3.300e-01  -50.734  < 2e-16 ***
CountyBoyd                -3.644e+00  4.320e-01   -8.434  < 2e-16 ***
CountyBronx                1.604e+00  2.981e-01    5.382 7.38e-08 ***
CountyBucks               -3.827e+00  3.253e-01  -11.766  < 2e-16 ***
CountyBurleigh            -9.661e+00  3.616e-01  -26.718  < 2e-16 ***
CountyCambria             -1.031e+01  3.043e-01  -33.876  < 2e-16 ***
CountyCamden              -9.675e-01  3.016e-01   -3.208 0.001337 **
CountyCass                -8.399e+00  3.306e-01  -25.406  < 2e-16 ***
CountyCharleston          -1.585e+01  3.669e-01  -43.204  < 2e-16 ***
CountyCherokee            -5.643e+00  3.631e-01  -15.543  < 2e-16 ***
CountyClark               -7.217e-01  3.500e-01   -2.062 0.039229 *
CountyContra Costa        -9.942e+00  2.812e-01  -35.360  < 2e-16 ***
CountyCook                 1.610e+00  2.977e-01    5.409 6.36e-08 ***
CountyCumberland          -6.017e+00  3.141e-01  -19.159  < 2e-16 ***
CountyCuyahoga            -3.141e+00  2.967e-01  -10.587  < 2e-16 ***
CountyDallas              -7.812e-01  2.993e-01   -2.610 0.009063 **
CountyDauphin             -3.782e+00  4.779e-01   -7.914 2.49e-15 ***
CountyDaviess             -3.240e+00  5.987e-01   -5.412 6.24e-08 ***
CountyDeKalb              -5.904e+00  3.183e-01  -18.551  < 2e-16 ***
CountyDenver               6.915e+00  3.000e-01   23.049  < 2e-16 ***
CountyDistrict of Columbia -5.476e+00  2.944e-01  -18.599  < 2e-16 ***
CountyDuchesne            -1.290e+01  1.811e+00   -7.122 1.07e-12 ***
CountyEast Baton Rouge    -3.566e+00  2.968e-01  -12.015  < 2e-16 ***
CountyEl Paso              2.053e+00  2.922e-01    7.024 2.16e-12 ***
CountyErie                -6.534e+00  3.287e-01  -19.878  < 2e-16 ***
CountyEssex                3.974e+00  3.121e-01   12.732  < 2e-16 ***
CountyFairbanks North Star -1.106e+01  3.759e-01  -29.432  < 2e-16 ***
```

```
CountyFairfax             -7.924e+00  2.968e-01  -26.693  < 2e-16 ***
CountyFairfield           -1.997e+00  4.132e-01   -4.832 1.35e-06 ***
CountyFayette              1.067e+00  4.462e-01    2.390 0.016827 *
CountyForsyth             -8.941e+00  5.204e-01  -17.183  < 2e-16 ***
CountyFremont             -1.174e+01  6.003e-01  -19.562  < 2e-16 ***
CountyFresno              -6.749e+00  3.048e-01  -22.143  < 2e-16 ***
CountyGarrett             -1.576e+01  3.365e-01  -46.835  < 2e-16 ***
CountyHamilton            -4.220e+00  3.208e-01  -13.155  < 2e-16 ***
CountyHampton City        -1.439e+01  3.411e-01  -42.174  < 2e-16 ***
CountyHarris              -4.094e+00  2.869e-01  -14.271  < 2e-16 ***
CountyHartford            -8.090e+00  3.536e-01  -22.879  < 2e-16 ***
CountyHaywood             -2.810e+01  7.238e-01  -38.824  < 2e-16 ***
CountyHenderson            2.491e+00  5.895e-01    4.225 2.38e-05 ***
CountyHenrico             -8.410e+00  3.236e-01  -25.992  < 2e-16 ***
CountyHillsborough        -1.105e+01  3.389e-01  -32.614  < 2e-16 ***
CountyHinds               -8.103e+00  3.738e-01  -21.677  < 2e-16 ***
CountyHonolulu            -1.301e+01  2.893e-01  -44.984  < 2e-16 ***
CountyHumboldt            -1.768e+01  2.974e-01  -59.454  < 2e-16 ***
CountyImperial            -6.937e+00  2.992e-01  -23.187  < 2e-16 ***
CountyJackson             -1.599e+01  4.863e-01  -32.888  < 2e-16 ***
CountyJefferson           -5.796e+00  3.094e-01  -18.735  < 2e-16 ***
CountyKay                 -1.116e+01  4.530e-01  -24.628  < 2e-16 ***
CountyKent                -1.003e+00  4.071e-01   -2.463 0.013782 *
CountyKern                 8.504e+00  9.855e-01    8.628  < 2e-16 ***
CountyKing                -5.794e-01  3.355e-01   -1.727 0.084211 .
CountyLackawanna          -1.910e+00  3.585e-01   -5.329 9.89e-08 ***
CountyLancaster           -7.071e+00  3.561e-01  -19.858  < 2e-16 ***
CountyLaramie             -7.322e+00  3.212e-01  -22.795  < 2e-16 ***
CountyLawrence            -7.008e+00  3.570e-01  -19.631  < 2e-16 ***
CountyLinn                -1.615e+01  3.469e-01  -46.559  < 2e-16 ***
CountyLitchfield          -1.398e+01  3.573e-01  -39.136  < 2e-16 ***
CountyLos Angeles          3.652e+00  2.808e-01   13.007  < 2e-16 ***
CountyLuzerne             -5.249e+00  4.959e-01  -10.586  < 2e-16 ***
CountyMaricopa             3.607e+00  2.865e-01   12.589  < 2e-16 ***
CountyMarion              -5.399e+00  3.062e-01  -17.630  < 2e-16 ***
CountyMcCracken           -2.928e-01  5.847e-01   -0.501 0.616618
CountyMcLennan            -1.312e+01  3.196e-01  -41.053  < 2e-16 ***
CountyMecklenburg         -6.429e+00  2.920e-01  -22.014  < 2e-16 ***
CountyMedina              -1.129e+01  6.427e-01  -17.572  < 2e-16 ***
CountyMeigs               -2.451e+01  6.991e-01  -35.061  < 2e-16 ***
CountyMilwaukee            4.027e-01  3.797e-01    1.061 0.288807
CountyMinnehaha           -9.247e+00  3.186e-01  -29.022  < 2e-16 ***
CountyMonroe              -5.956e+00  6.391e-01   -9.320  < 2e-16 ***
CountyMontgomery          -2.282e+00  3.575e-01   -6.383 1.74e-10 ***
CountyMultnomah           -9.461e+00  3.102e-01  -30.501  < 2e-16 ***
CountyNew Castle          -3.366e+00  3.381e-01   -9.956  < 2e-16 ***
CountyNew Haven            1.171e+00  3.083e-01    3.798 0.000146 ***
CountyNorthampton         -5.121e+00  3.547e-01  -14.437  < 2e-16 ***
CountyOklahoma            -9.104e+00  3.442e-01  -26.449  < 2e-16 ***
CountyOrange              -7.654e+00  2.875e-01  -26.627  < 2e-16 ***
CountyOttawa              -4.472e+00  4.235e-01  -10.560  < 2e-16 ***
CountyPhiladelphia         1.992e+00  3.034e-01    6.566 5.17e-11 ***
```

```
CountyPolk                -7.327e+00  3.129e-01  -23.414  < 2e-16 ***
CountyPrince George's     -8.295e+00  3.200e-01  -25.920  < 2e-16 ***
CountyProvidence          -8.172e+00  3.162e-01  -25.844  < 2e-16 ***
CountyPulaski             -5.821e+00  2.906e-01  -20.034  < 2e-16 ***
CountyQueens               4.446e+00  2.964e-01   15.000  < 2e-16 ***
CountyRichland            -7.566e+00  1.084e+00   -6.981 2.92e-12 ***
CountyRiverside           -3.842e+00  2.889e-01  -13.297  < 2e-16 ***
CountyRoanoke             -1.196e+01  3.562e-01  -33.586  < 2e-16 ***
CountyRockingham          -1.415e+01  3.474e-01  -40.720  < 2e-16 ***
CountyRutland             -8.571e+00  3.560e-01  -24.077  < 2e-16 ***
CountySacramento          -9.045e+00  2.920e-01  -30.975  < 2e-16 ***
CountySaint Clair         -6.508e+00  3.023e-01  -21.528  < 2e-16 ***
CountySaint Louis         -4.967e+00  3.485e-01  -14.253  < 2e-16 ***
CountySalt Lake            3.142e+00  2.961e-01   10.610  < 2e-16 ***
CountySan Bernardino       4.419e+00  2.888e-01   15.298  < 2e-16 ***
CountySan Diego           -3.281e+00  2.874e-01  -11.415  < 2e-16 ***
CountySan Francisco       -2.881e+00  3.201e-01   -9.002  < 2e-16 ***
CountySanta Barbara       -1.325e+01  2.815e-01  -47.070  < 2e-16 ***
CountySanta Clara         -5.917e+00  3.036e-01  -19.490  < 2e-16 ***
CountySanta Cruz          -1.490e+01  3.227e-01  -46.187  < 2e-16 ***
CountyScott               -8.653e+00  3.034e-01  -28.520  < 2e-16 ***
CountyShelby              -1.304e+01  1.322e+00   -9.865  < 2e-16 ***
CountySolano              -1.033e+01  2.947e-01  -35.066  < 2e-16 ***
CountySt. Louis City      -2.257e+00  3.002e-01   -7.520 5.47e-14 ***
CountySteuben             -1.579e+01  4.105e-01  -38.458  < 2e-16 ***
CountySuffolk             -4.119e-01  2.897e-01   -1.422 0.155152
CountySumner              -9.431e+00  6.855e-01  -13.758  < 2e-16 ***
CountySweetwater          -1.038e+01  5.328e-01  -19.487  < 2e-16 ***
CountyTravis              -8.095e+00  5.034e-01  -16.079  < 2e-16 ***
CountyTulsa               -7.496e+00  3.118e-01  -24.039  < 2e-16 ***
CountyUinta               -1.184e+01  4.688e-01  -25.255  < 2e-16 ***
CountyUintah              -2.139e+01  4.353e-01  -49.136  < 2e-16 ***
CountyUnion               -1.423e+01  6.112e-01  -23.276  < 2e-16 ***
CountyVentura             -5.269e+00  3.781e-01  -13.933  < 2e-16 ***
CountyWake                -1.078e+01  3.333e-01  -32.352  < 2e-16 ***
CountyWashington          -8.203e+00  3.172e-01  -25.860  < 2e-16 ***
CountyWashoe               1.775e+00  3.183e-01    5.578 2.44e-08 ***
CountyWayne                7.508e-01  3.986e-01    1.884 0.059632 .
CountyWestmoreland        -5.759e+00  3.574e-01  -16.114  < 2e-16 ***
CountyWyandotte           -4.230e+00  2.991e-01  -14.143  < 2e-16 ***
CountyYork                -2.395e+00  3.114e-01   -7.691 1.46e-14 ***
CO_AQI                     1.460e+00  7.719e-03  189.123  < 2e-16 ***
CO_Mean                   -4.196e+00  1.396e-01  -30.067  < 2e-16 ***
O3_AQI                     2.274e-01  1.032e-03  220.372  < 2e-16 ***
O3_Mean                   -4.001e+02  1.953e+00 -204.865  < 2e-16 ***
SO2_AQI                    1.295e-01  2.262e-03   57.259  < 2e-16 ***
SO2_Mean                   2.992e-01  1.002e-02   29.851  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.771 on 486817 degrees of freedom
Multiple R-squared:  0.6395,	Adjusted R-squared:  0.6394
F-statistic:  6125 on 141 and 486817 DF,  p-value: < 2.2e-16
```

From this model you can see that the R^2 value has decreased (.64), which is logically sensible as we are decreasing the number of variables and observations in our model. However, by removing the biased variable we can see our coefficients have also changed and we can see, especially as we look at each gas pollutant variable. In this model we can see other counties representing a positive coefficient. Los Angeles and San Bernardino County are all showing positive coefficients. This gives me indication that these counties produce high levels of NO2 AQI levels. This can be explained as Los Angeles is listed as the most polluted zip code in California. Where the County tends to the burning of fossil fuels, especially by vehicles, ships, planes and manufacturing, as well as its recent wildfires. San Bernardino is also highly polluted given its dry and hot climate.

Below includes the RMSE value from the second model (less than 1). It is clear that the model depiction is quite accurate given this value. The takeaway from this is that because there are so many counties across California as well as the other 3 gas pollutants being included in my predictor, the leading cause is an accurate model representing the true values of NO2 AQI.

```
> sqrt(mean(lm.pred - PData2.test$NO2_AQI)^2)
[1] 0.006074433
```

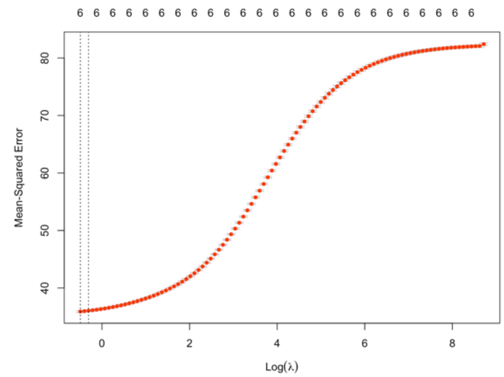
2. Ridge Regression Model

The goal of Ridge Regression is to minimize the RSS. Minimizing our RSS will lead to better prediction accuracy by introducing the shrinkage penalty. The shrinkage penalty will shrink the coefficient estimates towards and approximate zero value. Conducting a Ridge Regression on the training data set is also less prone to overfitting. To perform Ridge Regression, I used a cross-validation to figure out which “tuning parameter” lambda results in the smallest RMSE.

My findings led me to conclude the best lambda value is 0.610. This value gave me an RMSE of 6.037. Our prediction indicates that by using the Ridge Regression Model, we are approximately 6.037 levels of AQI away from the test value.

Below includes the coefficients for the major variables given the best lambda as well as a plot of the cross-validation errors for all lambdas.

```
CO_AQI          4.783983e-01
CO_Mean         6.411196e+00
O3_AQI          8.436547e-02
O3_Mean        -2.447117e+02
SO2_AQI         5.756222e-02
SO2_Mean        4.100404e-01
```



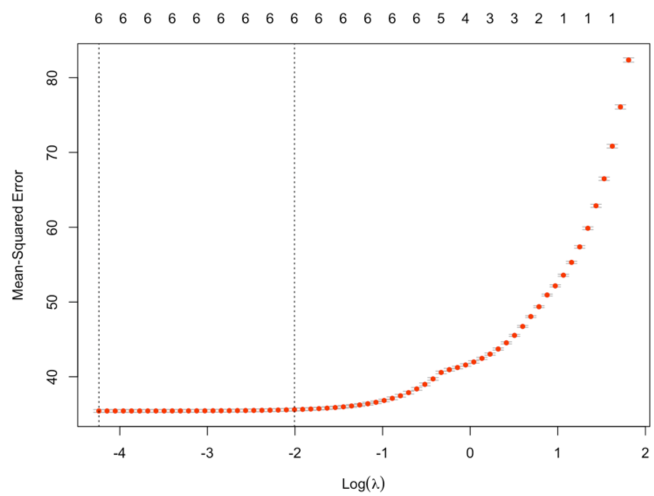
3. Lasso Regression Model

The Lasso model is like the previous Ridge Regression. However, Lasso is a sparse regression model because it shrinks the coefficient estimates towards zero and only a small number of estimates are non-zero. The advantages of using Lasso Regression is that it solves the overfitting issue using the Linear Models. Also, it works well with a large number of predictor variables.

In order to determine the best lambda value, I used another cross-validation method giving me a best lambda of 0.014. This lambda value led me to derive the RMSE which was 6.0001.

Below includes the coefficients for the major variables given the best lambda as well as a plot of the cross-validation errors for all lambdas.

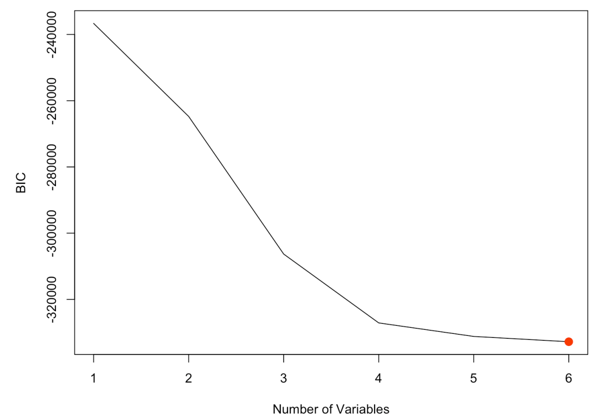
```
CO_AQI           0.52895084
CO_Mean          5.39658966
O3_AQI           0.12042642
O3_Mean       -313.84716706
SO2_AQI          0.04955011
SO2_Mean         0.43350972
```



4. Best-Subset Selection

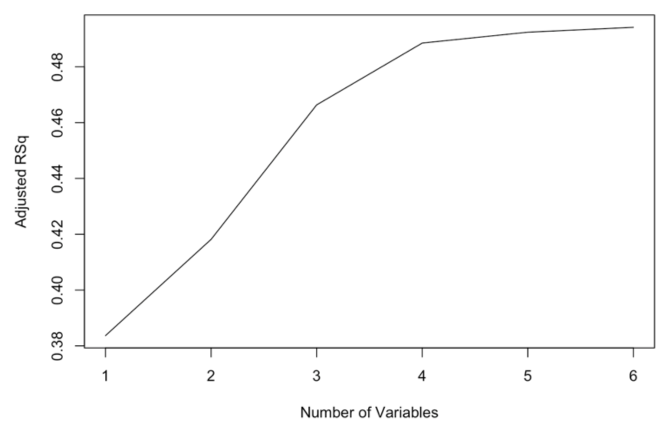
The best-subset selection value helps aim to find the best possible predicted outcome for our response variable (NO2 AQI). For determining the best subset, I went ahead and followed the proper steps under the Bayesian Information Criterion (BIC).

Below lies the best-subset selection given the dataset:



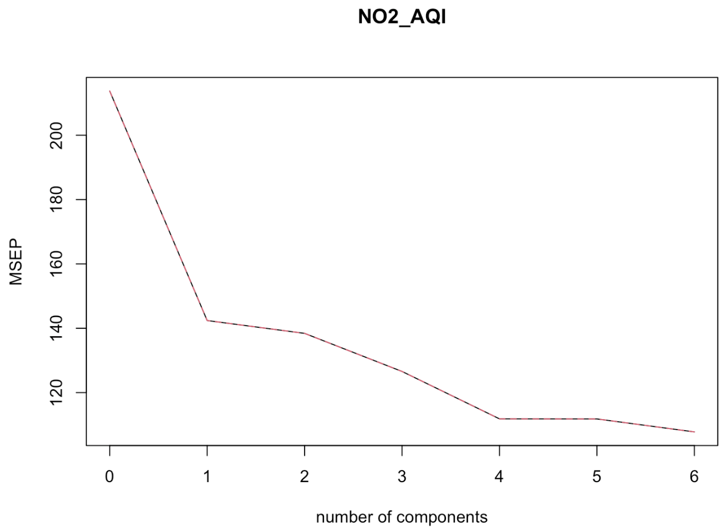
According to the graph, the most optimal number of variables to include in my model is 6.

Below is another graph illustrating the adjusted R^2 value given the number of variables: From here, you can see that 6 variables leads to the highest R^2 value.



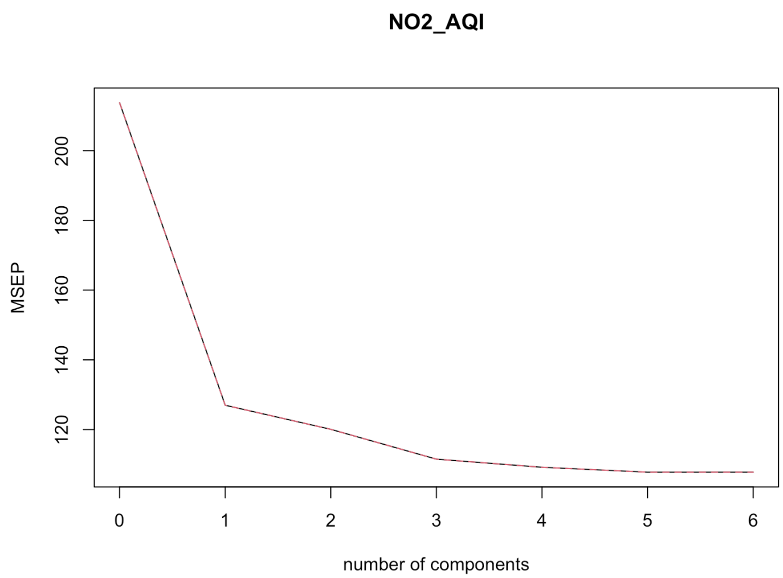
5. PCR Method

The PCR Method is a dimensional reduction method. PCR is used for computing regression when the explanatory variables are highly correlated because it converts these variables into a set of linearly uncorrelated variables. The downside of using this method is PCR does not incorporate the response variable. Therefore, there is no guarantee the directions that best explain the predictors will also be the best to predict the response. Just like the Ridge and Lasso models previously, cross-validation approach was performed to find which M value leads to the lowest RMSE. The lowest RMSE value occurs when M = 6. The RMSE is also 10.42.



6. PLS Method

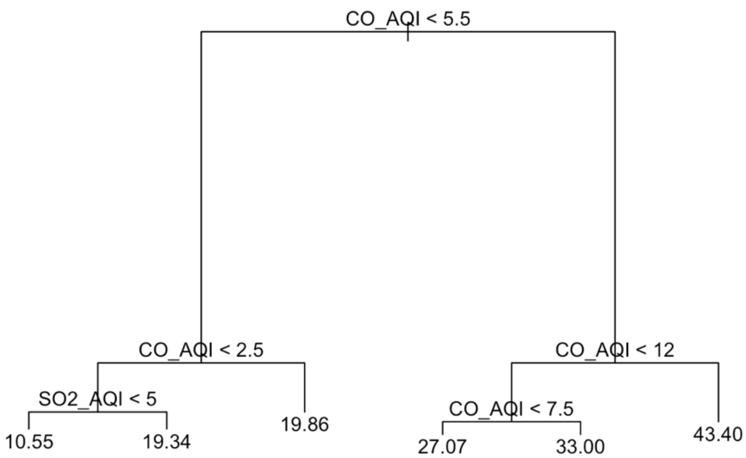
The PLS Method, unlike PCR, is incorporated with the response variable. Because of this, we now have a high chance to predict the response. After performing the proper cross-validation, I concluded with M = 6 again. The RMSE for when M = 6 is 17.84.



7. Regression Tree

A Regression Decision Tree is a non-parametric supervised learning method and is the easiest model to interpret and understand. However, it is prone to overfitting. Thus, we prune the

Decision tree to prevent the training data from overfitting. Below includes the results from the model:




```
Regression tree:
tree(formula = NO2_AQI ~ CO_AQI + CO_Mean + O3_AQI + O3_Mean +
    SO2_AQI + SO2_Mean, data = PData2, subset = train)
Variables actually used in tree construction:
[1] "CO_AQI"  "SO2_AQI"
Number of terminal nodes:  6
```

As illustrated in the Regression tree, the variables used post pruning were the Carbon Monoxide AQI (CO AQI) and Sulphur Dioxide AQI (SO2 AQI). The RMSE value from this model is 12.32.

VI. Conclusion

Here we have a comparison chart between the different classes of regression models used for this project:

Model	Metric	Result
Linear Regression	RMSE	**.0067 (<1)**
Ridge Regression	RMSE	6.037
Lasso Regression	RMSE	6.0001
PCR	RMSE	10.42
PLS	RMSE	17.84
Regression tree	RMSE	12.32

Based on the above results, Linear and Lasso Regression are most optimal for best predicting the NO2 AQI accurately as they certainly hold the lowest RMSE values.

VII. Future Direction

In the future, I wish to further extend my exploration into this dataset by using other variables as the response variable. Also maybe removing any Counties that are not as significant to see whether this helps reduce chances of overfitting in my models. By understanding deeper learning within the dataset, I believe I can successfully be able to find specific correlations between the response and the predictor variable(s).

DATA SOURCE:

https://www.kaggle.com/datasets/sogun3/uspollution

APPENDIX:

```
install.packages("tidyverse")
install.packages("dplyr")
install.packages("readxl")
library("readxl")
library(tidyverse)
library(dplyr)

PData = read_excel("PData.xltx")
summary(PData)

PData2 = read_excel("PData2.xlsx")
summary(PData2)
NO2_Mean = PData2$NO2_Mean
hist(NO2_Mean)

Ozone = PData2$O3_AQI
hist(Ozone, main = "Histogram of Ozone AQI", xlab = "O3 AQI")
```

```
Nitrogen = PData$NO2_AQI
hist(Nitrogen, main = "Histogram of Nitrogen Dioxide AQI", xlab = "NO2 AQI")

CO = PData2$CO_AQI
hist(CO, main = "Histogram of Carbon Monoxide AQI", xlab = "CO AQi")



#Creating a split(train and test data set)
train = sample(dim(PData2)[1], dim(PData2)[1]*.8)
test = -train
PData2.test = PData2[test, ]
PData2.test
PData2.train = PData2[train, ]

#linear regression model
lm.fit0 = lm(NO2_AQI ~ County + CO_AQI + CO_Mean + NO2_AQI + NO2_Mean + SO2_AQI +
SO2_Mean, data = PData2.train)
summary(lm.fit0) #Low R^2, O3 is not a good response variable
lm.fit = lm(NO2_AQI ~ County + CO_AQI + CO_Mean + O3_AQI + O3_Mean + SO2_AQI +
SO2_Mean, data = PData2.train)
summary(lm.fit) #NO2 Mean is very significant as predictor for 03_AQI
lm.fit1 = lm(CO_AQI ~ O3_AQI + O3_Mean + NO2_AQI + NO2_Mean + SO2_AQI + SO2_Mean,
data = PData2.train)
summary(lm.fit1)

lm.pred = predict(lm.fit, PData2.train)
summary(lm.pred)
mean((lm.pred - PData2.train$NO2_AQI)^2) #train MSE
```

```
sqrt(mean(lm.pred - PData2.train$NO2_AQI)^2) #train RMSE very close, good fit model for
prediction

lm_coef <- summary(lm.fit)$NO2_AQI
lm_coef

### Lasso and Ridge ###
#######################

library(ISLR)
library(glmnet)

train = sample(dim(PData2)[1], dim(PData2)[1]*.8)
test = -train
PData2.test = PData2[test, ]
PData2.test
PData2.train = PData2[train, ]


#model.matrix()automatically transforms qualitat var into dummy var #glmnet() can only take
numerical inputs.
x=model.matrix(NO2_AQI ~ County + CO_AQI + CO_Mean + O3_AQI + O3_Mean + SO2_AQI +
SO2_Mean, PData2)[,-1]
y=PData2$NO2_Mean
dim(x)

test<-(-train)
y.test=y[test]
```

```
#generating grid lambidas
grid=10^seq(10,-2,length=100)
ridge_mod=glmnet(x,y,alpha=0,lambda=grid)


######################
### Ridge Regression ###
######################

ridge_mod=glmnet(x[train,],y[train],alpha=0,lambda=grid, thresh=1e-12) # thresh controls
coordinate descent convergence

#we can use CV to choose the tuning parameter lambda
set.seed(1)
cv_out=cv.glmnet(x[train,],y[train],alpha=0)
plot(cv_out)#cross validation errors  (y) for all lambidas(x)

bestlam=cv_out$lambda.min #yelds best lambida
bestlam

ridge.pred=predict(ridge_mod,s=bestlam,newx=x[test,])
ridge_MSE= mean((ridge.pred-y.test)^2)
ridge_RMSE = sqrt(mean((ridge.pred-y.test)^2))
ridge_RMSE
##RMSE for CO_AQI response is 3.78
##RMSE for NO2_AQI is 6.01

out=glmnet(x,y,alpha=0)
ridge_coef=predict(out,type="coefficients",s=bestlam)
```

```
ridge_coef


##############
### Lasso  ###
##############

lasso_mod=glmnet(x[train,],y[train],alpha=1,lambda=grid, thresh=1e-12) # thresh controls
coordinate descent convergence

set.seed(10)
cv_out_lasso=cv.glmnet(x[train,],y[train],alpha=1)
plot(cv_out_lasso)#cross validation errors  (y) for all lambidas(x)

bestlam=cv_out_lasso$lambda.min #yields best lambida
bestlam

lasso_pred=predict(lasso_mod,s=bestlam,newx=x[test,])
lasso_MSE= mean((lasso_pred-y.test)^2)
lasso_MSE
lasso_RMSE = sqrt(mean((lasso_pred-y.test)^2))
lasso_RMSE
#lasso RMSE is 5.97 for NO2_AQI response


out=glmnet(x,y,alpha=1)
predict(out,type="coefficients",s=bestlam)

lasso_coef=predict(out,type="coefficients",s=bestlam)
```

```
lasso_coef
lasso_coef[lasso_coef!=0]

#####BEST SUBSET
library(leaps)
regfit.full=regsubsets(NO2_AQI ~ County + CO_AQI + CO_Mean + O3_AQI + O3_Mean +
SO2_AQI + SO2_Mean ,PData2.train)
summary(regfit.full)

regfit.full=regsubsets(NO2_AQI ~ County + CO_AQI + CO_Mean + O3_AQI + O3_Mean +
SO2_AQI + SO2_Mean, data=PData2.train,nvmax=13)
reg.summary=summary(regfit.full)

names(reg.summary)
reg.summary$rsq

plot(reg.summary$adjr2,xlab="Number of Variables",ylab="Adjusted RSq",type="l")
which.max(reg.summary$adjr2) #which gives largest adjusted R2
points(13,reg.summary$adjr2[13], col="red",cex=2,pch=20)
plot(reg.summary$cp,xlab="Number of Variables",ylab="Cp",type='l')
which.min(reg.summary$cp)
points(10,reg.summary$cp[13],col="red",cex=2,pch=20)
which.min(reg.summary$bic)
plot(reg.summary$bic,xlab="Number of Variables",ylab="BIC",type='l')
points(6,reg.summary$bic[6],col="red",cex=2,pch=20)

#built-in plot command can be used to display the selected variables
#for the best model with a given number of predictors, ranked
#BIC, Cp, adjusted R2, or AIC..
```

```
plot(regfit.full,scale="r2")
plot(regfit.full,scale="adjr2")
plot(regfit.full,scale="Cp")
plot(regfit.full,scale="bic")

#see the coefficient estimates estimated with the model
coef(regfit.full,6)

### PCR ###
###########
library(pls)

pcr_fit <- pcr(NO2_AQI ~ CO_AQI + CO_Mean + O3_AQI + O3_Mean + SO2_AQI + SO2_Mean,
data=PData2.train, scale = TRUE, validation="CV")
summary(pcr_fit)

validationplot(pcr_fit, val.type="MSEP")

pcr_pred <- predict(pcr_fit, PData2.test, ncomp=6)
pcr_mean <- mean((pcr_pred - PData2.test$NO2_AQI)^2)
pcr_mean
pcr_RMSE = sqrt(mean((pcr_pred - PData2.test$NO2_AQI)^2))
pcr_RMSE

### PLS ###
###########
pls_fit <- plsr(NO2_AQI ~ CO_AQI + CO_Mean + O3_AQI + O3_Mean + SO2_AQI + SO2_Mean,
data=PData2.train, scale = TRUE, validation="CV")
summary(pls_fit)
```

```
validationplot(pls_fit, val.type="MSEP")

pls_pred <- predict(pls_fit, PData2.train, ncomp = 5)
pls_mean <- mean((pls_pred - PData2.test$NO2_AQI)^2)
pls_mean

pls_RMSE = sqrt(mean((pls_pred - PData2.test$NO2_AQI)^2))
pls_RMSE

###Regression Tree
library(tree)
tree.PData2 = tree(NO2_AQI ~ CO_AQI + CO_Mean + O3_AQI + O3_Mean + SO2_AQI +
SO2_Mean, PData2 ,subset = PData2.train)
summary(tree.PData2)
#in regression, deviance is the RSS

plot(tree.PData2)
text(tree.PData2,pretty=0)
cv.PData2=cv.tree(tree.PData2)
plot(cv.PData2$size,cv.PData2$dev,type='b')
prune.PData2=prune.tree(tree.PData2,best=6)
plot(prune.PData2)
text(prune.PData2,pretty=0)

lm_MSE
ridge_MSE
lasso_MSE
tree_MSE
```

pcr_mean

pls_mean

Predicting Pollution in CA

- Zain Mirza

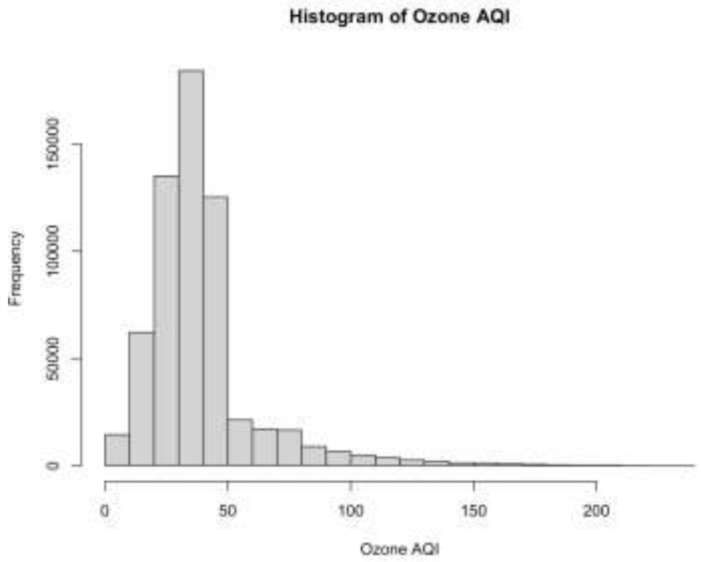
Data Source

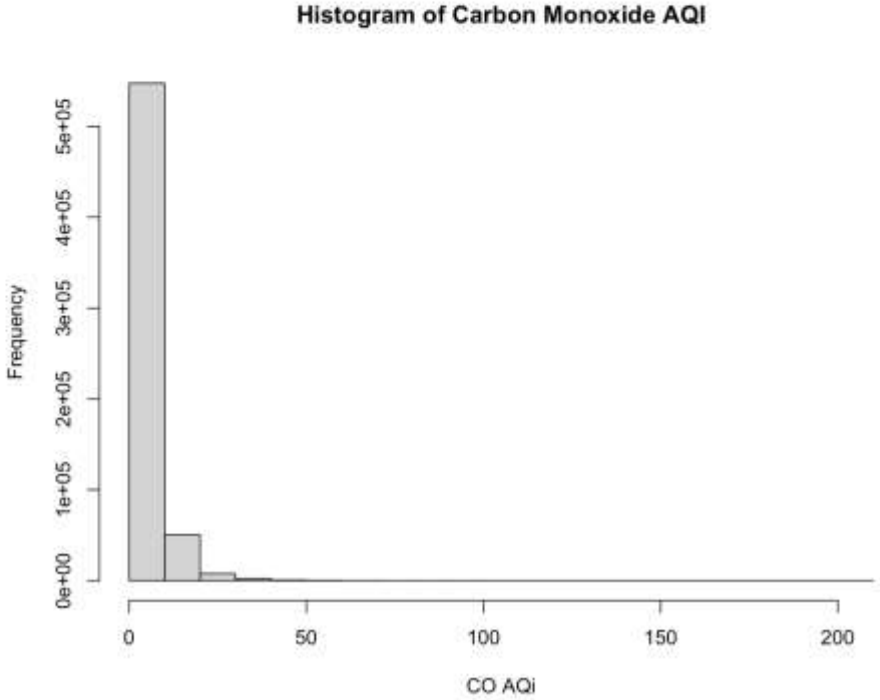
- Obtained:
  - Kaggle resources (updated 3 months ago.)
- Variables:
  - 22 (important ones*** - CO/NO2/O3/SO2 Mean + AQI)
- Variable Type:
  - Numeric
  - Categorical
- Observations:
  - Over 1.4 Million

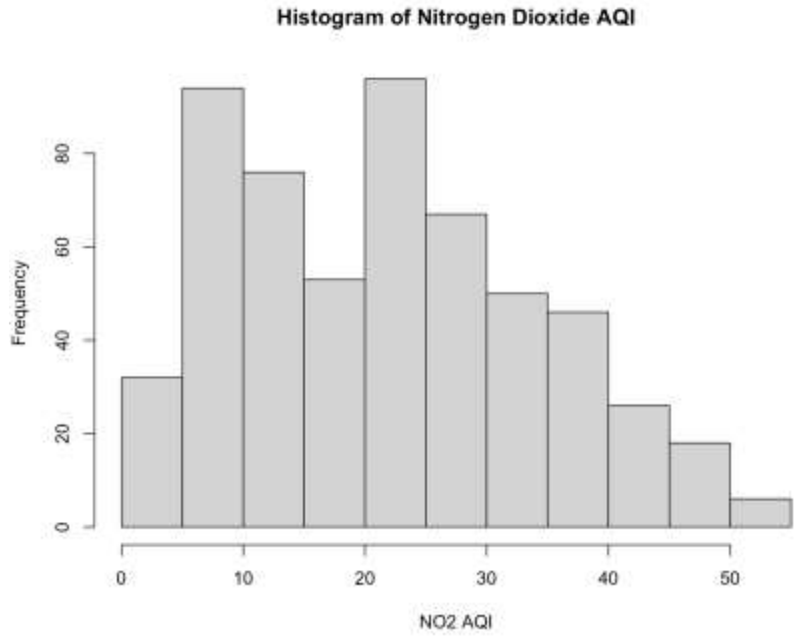
Data Cleansing + Splitting

- 1: Adjust focus to one state (CA)
- 2: transitioning from daily averages to monthly
- 3: Removing insignificant variables.
- Conclusion: dataset concise to about 595k observations
- Training and Test:
  - Split the data 80% train and 20% test.

Overview of Cleansed Dataset







R2084

Date	Year	Month	Day	State	County	City	O3_Mean	O3_AQI	CO_Mean	CO_AQI	SO2_Mean	SO2_AQI	NO2_Mean	NO2 AQI
1/1/00	2000	1	1	California	Contra Costa	Concord	0.019412	27	0.572222	9	0.636364	3	14.782609	25
2/1/00	2000	2	1	California	Contra Costa	Concord	0.016118	21	0.633333	8	0.5	1	14.434783	29
3/1/00	2000	3	1	California	Contra Costa	Concord	0.022118	28	0.779167	13	2.636364	9	17.347826	34
4/1/00	2000	4	1	California	Contra Costa	Concord	0.027529	45	0.6	10	1.227273	3	11.173913	31
5/1/00	2000	5	1	California	Contra Costa	Concord	0.027294	34	0.7375	8	1.772727	3	14.478261	23
6/1/00	2000	6	1	California	Contra Costa	Concord	0.041118	71	0.4875	8	1.727273	6	16.478261	27
7/1/00	2000	7	1	California	Contra Costa	Concord	0.020765	22	0.279167	3	1.090909	3	6.826087	9
8/1/00	2000	8	1	California	Contra Costa	Concord	0.053929	112	0.74375	11	3.111111	14	22.105263	36
9/1/00	2000	9	1	California	Contra Costa	Concord	0.005588	8	0.454167	7	0.181818	1	13.652174	19
10/1/00	2000	10	1	California	Contra Costa	Concord	0.035471	46	0.645833	10	1.681818	11	16.956522	36
11/1/00	2000	11	1	California	Contra Costa	Concord	0.002941	6	1.325	22	3.263158	17	25.142857	37
12/1/00	2000	12	1	California	Contra Costa	Concord	0.005412	9	1.675	31	2.15	8	20.180476	31
1/1/00	2000	1	1	California	Contra Costa	Bethel Island	0.019059	20	0.505556	8	1.227273	4	9.826087	13
2/1/00	2000	2	1	California	Contra Costa	Bethel Island	0.013118	19	0.35	5	0.090909	1	9.608696	21
3/1/00	2000	3	1	California	Contra Costa	Bethel Island	0.028471	35	0.333333	5	1.409091	6	8.608696	15
4/1/00	2000	4	1	California	Contra Costa	Bethel Island	0.034529	45	0.270833	3	0.363636	4	4.347826	15
5/1/00	2000	5	1	California	Contra Costa	Bethel Island	0.033765	38	0.4375	6	1.045455	4	8.826087	14
6/1/00	2000	6	1	California	Contra Costa	Bethel Island	0.048471	80	0.358333	7	1.909091	0	13.043478	26
7/1/00	2000	7	1	California	Contra Costa	Bethel Island	0.030412	35	0.2	2	0.681818	1	5.043478	8
8/1/00	2000	8	1	California	Contra Costa	Bethel Island	0.061765	122	0.458333	6	2.681818	7	13.26087	22
9/1/00	2000	9	1	California	Contra Costa	Bethel Island	0.007	16	0.325	5	1	1	0.681818	14
10/1/00	2000	10	1	California	Contra Costa	Bethel Island	0.032765	48	0.495833	8	3.318182	9	10.695652	23
11/1/00	2000	11	1	California	Contra Costa	Bethel Island	0.010953	18	0.516667	7	2.772727	13	15.086957	24
12/1/00	2000	12	1	California	Contra Costa	Bethel Island	0.007	12	0.741667	16	2.318182	6	15.130435	24
1/1/00	2000	1	1	California	Contra Costa	San Pablo	0.022882	26	0.433333	8	1.521739	3	13.695652	24
2/1/00	2000	2	1	California	Contra Costa	San Pablo	0.019235	22	0.525	8	4.454545	16	12.826087	16
3/1/00	2000	3	1	California	Contra Costa	San Pablo	0.022882	24	0.616667	9	3.045455	14	14.521739	28
4/1/00	2000	4	1	California	Contra Costa	San Pablo	0.025471	44	0.466667	8	1.318182	3	17.347826	30
5/1/00	2000	5	1	California	Contra Costa	San Pablo	0.026824	31	0.433333	6	2.454545	7	12.391304	25
6/1/00	2000	6	1	California	Contra Costa	San Pablo	0.028471	30	0.575	9	3.454545	9	18.173913	29
7/1/00	2000	7	1	California	Contra Costa	San Pablo	0.021059	21	0.2	2	1.181818	3	2.956522	8
8/1/00	2000	8	1	California	Contra Costa	San Pablo	0.028412	35	0.4	6	3.409091	7	13.043478	20
9/1/00	2000	9	1	California	Contra Costa	San Pablo	0.015412	17	0.270833	3	1.090909	3	9.086957	12
10/1/00	2000	10	1	California	Contra Costa	San Pablo	0.028529	34	0.3125	6	0.5	1	7.086957	15
11/1/00	2000	11	1	California	Contra Costa	San Pablo	0.005529	10	0.795833	11	2.545455	10	20.217391	29
12/1/00	2000	12	1	California	Contra Costa	San Pablo	0.003882	7	0.833333	11	2.636364	10	18.217391	28
1/1/00	2000	1	1	California	Contra Costa	Pittsburg	0.015824	18	0.088889	10	2.272727	8	16.217391	25
2/1/00	2000	2	1	California	Contra Costa	Pittsburg	0.013529	19	0.5375	7	0.5	1	13.565217	28
3/1/00	2000	3	1	California	Contra Costa	Pittsburg	0.025176	29	0.745833	15	2.227273	7	17.608696	31
4/1/00	2000	4	1	California	Contra Costa	Pittsburg	0.023882	44	0.366667	10	0.363636	3	14.521739	42
5/1/00	2000	5	1	California	Contra Costa	Pittsburg	0.040667	38	0.383333	6	1	6	10.304348	16
6/1/00	2000	6	1	California	Contra Costa	Pittsburg	0.043294	67	0.445833	8	2.954545	13	16.608696	29
7/1/00	2000	7	1	California	Contra Costa	Pittsburg	0.028118	28	0.3	3	0	0	4.086957	7
8/1/00	2000	8	1	California	Contra Costa	Pittsburg	0.058059	103	0.6875	10	3.045455	10	17.130435	29
9/1/00	2000	9	1	California	Contra Costa	Pittsburg	0.012176	14	0.358333	5	1	1	9.652174	14
10/1/00	2000	10	1	California	Contra Costa	Pittsburg	0.036765	47	0.0625	10	5.863636	21	16.869565	41
11/1/00	2000	11	1	California	Contra Costa	Pittsburg	0.008176	11	1.175	19	5.105263	14	25.391304	35
12/1/00	2000	12	1	California	Contra Costa	Pittsburg	0.008118	13	1.520833	25	1.818182	4	24.782609	35
1/1/00	2000	1	1	California	Imperial	Calexico	0.035647	38	0.494444	9	0.391304	4	9.304348	23
2/1/00	2000	2	1	California	Imperial	Calexico	0.016588	33	0.995833	57	0.186957	1	25.173913	52
3/1/00	2000	3	1	California	Imperial	Calexico	0.025294	44	0.645833	27	0.208696	1	16.086957	51
4/1/00	2000	4	1	California	Imperial	Calexico	0.028353	50	0.766667	30	1	11	20.521739	63
6/1/00	2000	6	1	California	Imperial	Calexico	0.021909	32	0.6	8	1.875	6	18.6875	30
7/1/00	2000	7	1	California	Imperial	Calexico	0.046235	90	0.416667	7	1.086957	4	8.478261	14
8/1/00	2000	8	1	California	Imperial	Calexico	0.044353	119	0.704167	9	1.869565	4	10.73913	25
9/1/00	2000	9	1	California	Imperial	Calexico	0.030647	37	0.2	2	1.043478	3	6.25	8
10/1/00	2000	10	1	California	Imperial	Calexico	0.022412	49	2.8125	68	9.708333	23	29.291667	56
11/1/00	2000	11	1	California	Imperial	Calexico	0.014235	31	1.0875	52	4.130435	17	21.565217	42
12/1/00	2000	12	1	California	Imperial	Calexico	0.008471	20	1.470833	50	2.318182	11	33.318182	44
11/1/00	2000	11	1	California	Kern	Bakersfield	0.013706	29	0.795833	15	2.434783	9	25.608696	38
1/1/00	2000	1	1	California	Los Angeles	Burbank	0.016529	26	1.3	34	0	0	25.869565	39
2/1/00	2000	2	1	California	Los Angeles	Burbank	0.011235	20	2.091667	38	2.521739	9	42.782609	57
3/1/00	2000	3	1	California	Los Angeles	Burbank	0.016765	30	1.1375	18	1.173913	3	35.478261	49
4/1/00	2000	4	1	California	Los Angeles	Burbank	0.046	105	0.4375	8	0	0	18.652174	55
5/1/00	2000	5	1	California	Los Angeles	Burbank	0.038588	119	1.570833	22	0.043478	1	60	89
6/1/00	2000	6	1	California	Los Angeles	Burbank	0.032706	77	0.891667	13	0	0	47	72

Sheet1

Ready Filter Mode 75%

Objective

- Determine which model is best for accurately predicting AQI.
- We will look at NO2_AQI as the response because it is classified as the most harmful according to EPA(Environmental Protection Agency) .

Model Methods

- Linear Model
- Ridge Regression
- Lasso
- Best-subset selection
- PCR
- PLS
- Regression Tree

Linear Regression

- 2 Models:
  - First model includes the NO2 mean for the given first day of the month.
  - Second model removes the NO2 mean, solely relying on other predictor variables.
  - Both included a low test RMSE value of .0574
- Conclusion: Linear Regression claims all predictors are significant to the response variable. R^2 value shows significant increase when including NO2 mean to model.

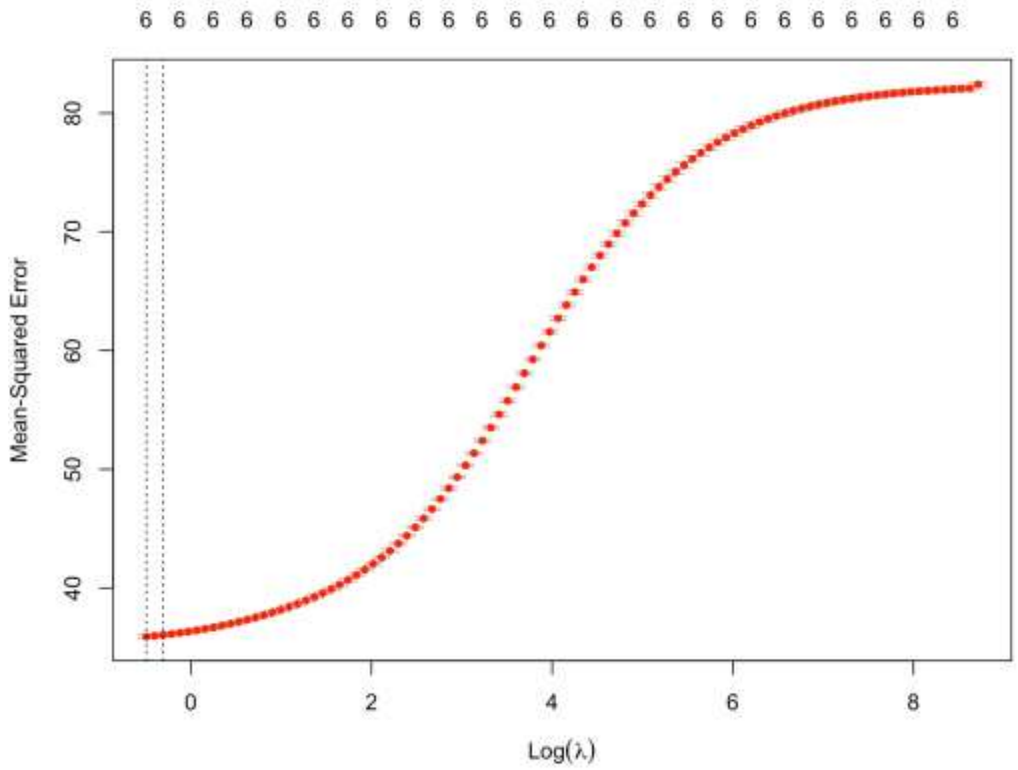
```
lm(formula = NO2_AQI ~ CO_AQI + CO_Mean + O3_AQI + O3_Mean +
    SO2_AQI + SO2_Mean + NO2_Mean, data = PData2.train)

Residuals:
    Min      1Q  Median      3Q     Max
-73.103  -3.633  -0.983   2.894  95.750

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.818e+00  3.003e-02   60.52   <2e-16 ***
CO_AQI       7.879e-01  4.839e-03  162.83   <2e-16 ***
CO_Mean     -1.342e+01  8.433e-02 -159.19   <2e-16 ***
O3_AQI       5.987e-02  6.601e-04   90.70   <2e-16 ***
O3_Mean      4.250e+01  1.299e+00   32.73   <2e-16 ***
SO2_AQI      6.420e-02  1.391e-03   46.16   <2e-16 ***
SO2_Mean    -2.838e-01  6.082e-03  -46.66   <2e-16 ***
NO2_Mean     1.458e+00  1.371e-03 1063.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.691 on 486951 degrees of freedom
Multiple R-squared:  0.8481,	Adjusted R-squared:  0.8481
F-statistic: 3.885e+05 on 7 and 486951 DF,  p-value: < 2.2e-16
```

```
lm(formula = NO2_AQI ~ CO_AQI + CO_Mean + O3_AQI + O3_Mean +
    SO2_AQI + SO2_Mean, data = PData2.train)

Residuals:
     Min       1Q   Median       3Q      Max
-251.248   -7.259   -1.209    6.354  109.446

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.639e+01  4.873e-02  336.25   <2e-16 ***
CO_AQI       1.957e+00  8.592e-03  227.77   <2e-16 ***
CO_Mean     -1.036e+01  1.537e-01  -67.42   <2e-16 ***
O3_AQI       2.867e-01  1.139e-03  251.70   <2e-16 ***
O3_Mean     -4.882e+02  2.186e+00 -223.33   <2e-16 ***
SO2_AQI      1.281e-01  2.534e-03   50.55   <2e-16 ***
SO2_Mean     4.434e-01  1.102e-02   40.24   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.38 on 486952 degrees of freedom
Multiple R-squared:  0.4952,	Adjusted R-squared:  0.4952
F-statistic: 7.961e+04 on 6 and 486952 DF,  p-value: < 2.2e-16
```

Ridge Regression

- Best lambda given @ .610
- Test RMSE = 6.037

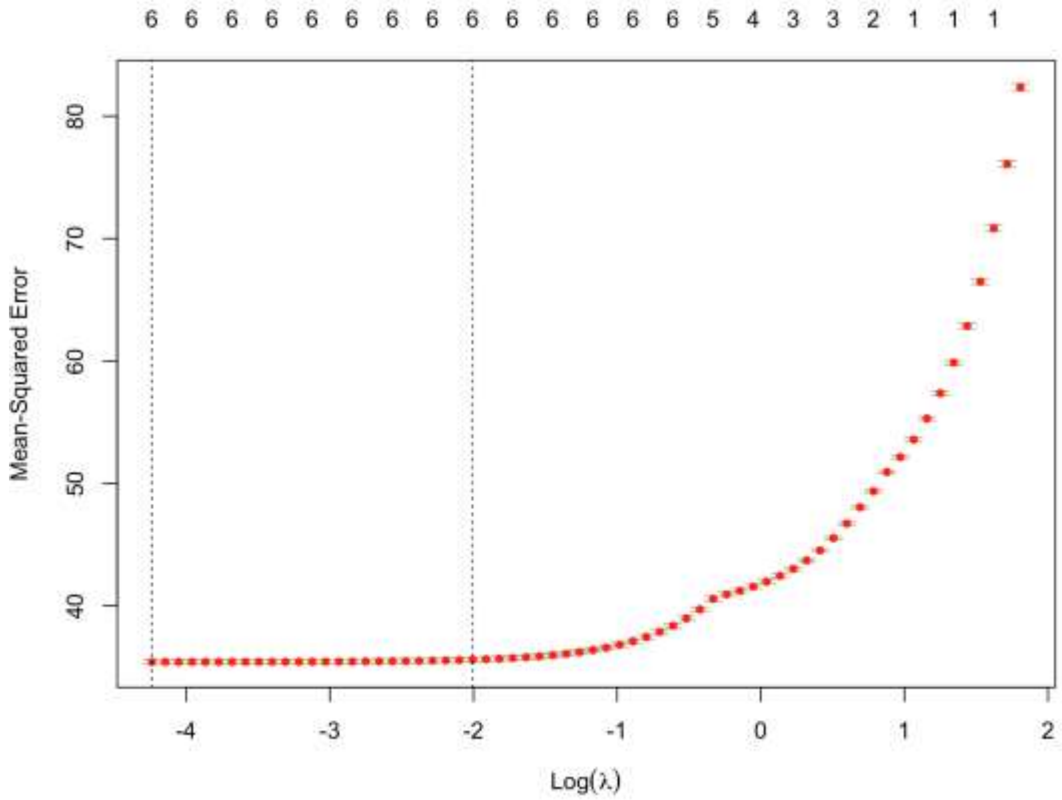
CO_AQI	4.783983e-01
CO_Mean	6.411196e+00
O3_AQI	8.436547e-02
O3_Mean	-2.447117e+02
SO2_AQI	5.756222e-02
SO2_Mean	4.100404e-01



LASSO

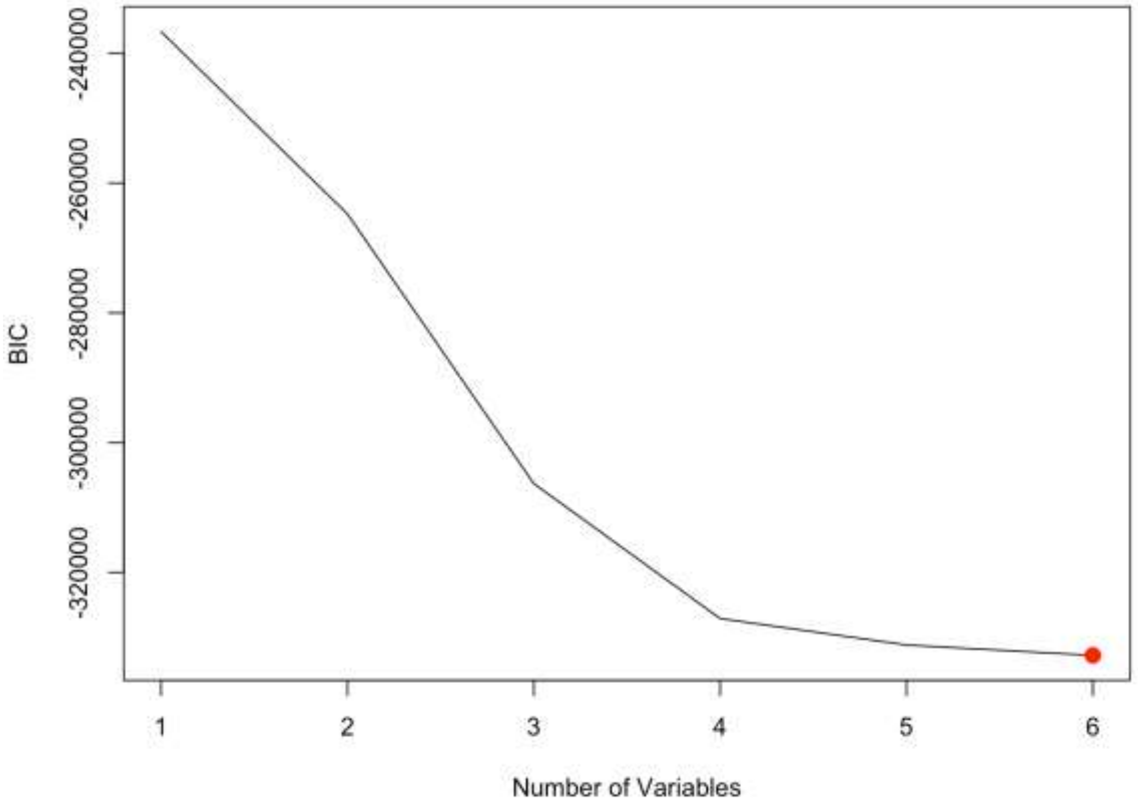
- Best lambda given @ .014
- Test RMSE = 6.0001

CO_AQI	0.52895084
CO_Mean	5.39658966
O3_AQI	0.12042642
O3_Mean	-313.84716706
SO2_AQI	0.04955011
SO2_Mean	0.43350972

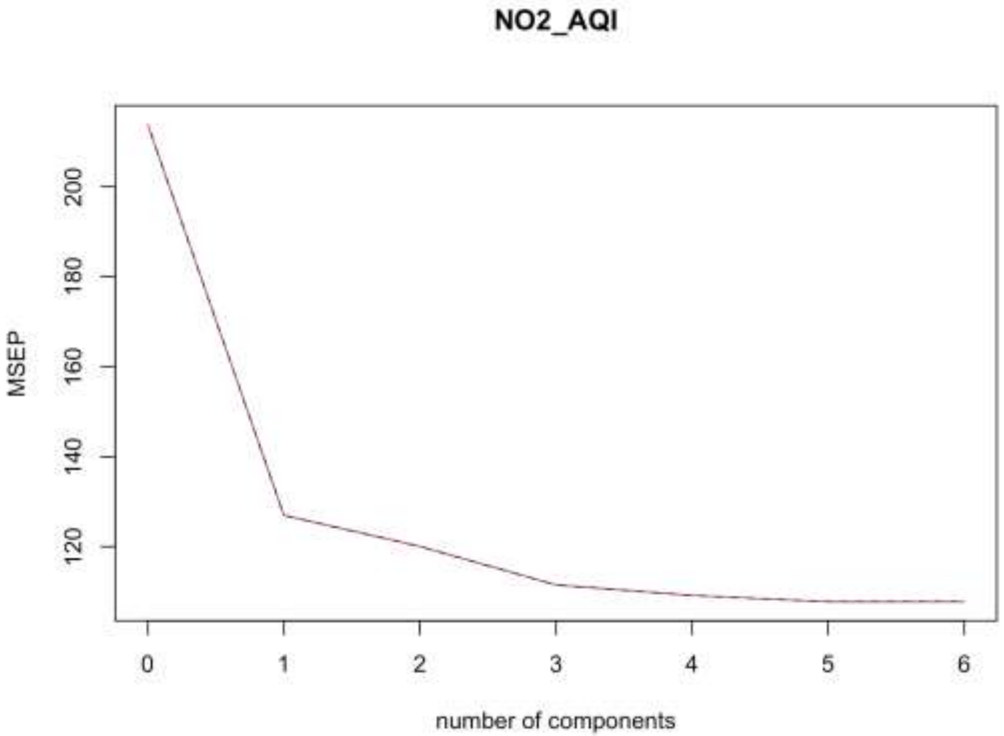


Best Subset Selection

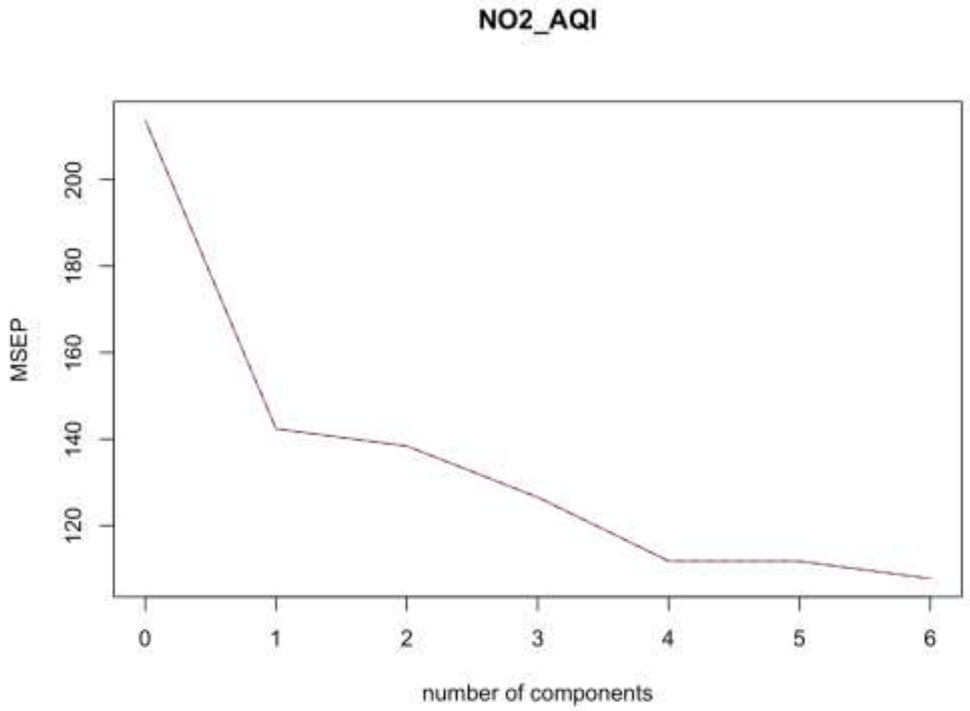
- Bayesian Information Criterion
  - Model = Included all 6 variables.



PCR & PLS Test



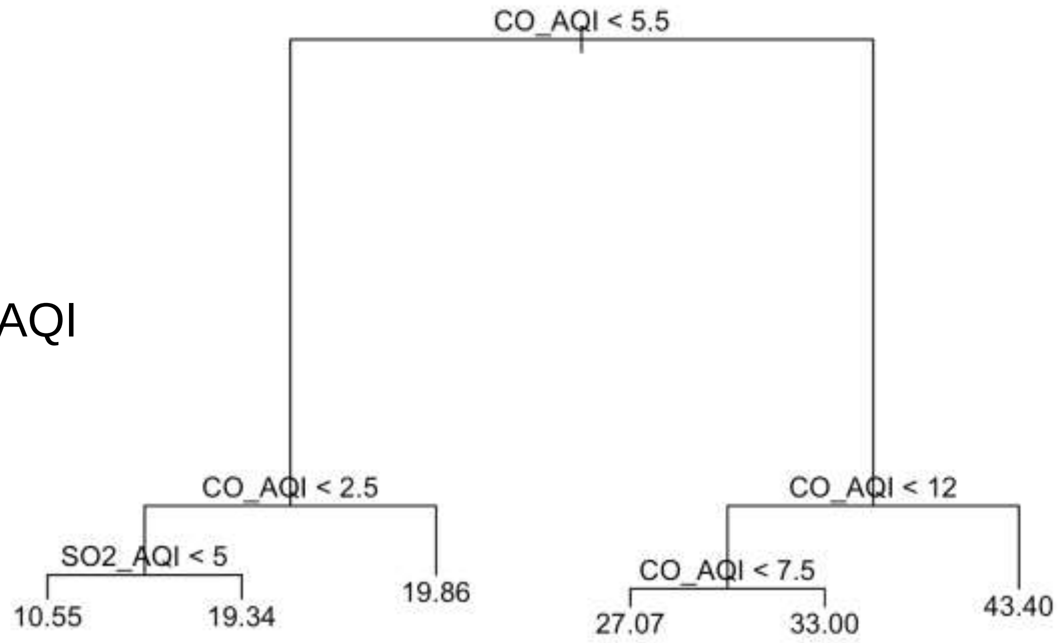
- PLS
- RMSE = 17.84
@ 5 components



- PCR
- RMSE = 10.42 @ 6 components

Regression Tree

- Test RMSE = 12.32
- Variables used post – pruning: CO_AQI & SO2_AQI




```
Regression tree:
tree(formula = NO2_AQI ~ CO_AQI + CO_Mean + O3_AQI + O3_Mean +
    SO2_AQI + SO2_Mean, data = PData2, subset = train)
Variables actually used in tree construction:
[1] "CO_AQI"  "SO2_AQI"
Number of terminal nodes:  6
Residual mean deviance:  120.2 = 58540000 / 487000
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-43.400  -7.550  -1.072   0.000   6.450 108.100
```

Best Model

- Linear Model:
  - RMSE = <1
  - R^2 = .5 (.84 w/ Mean of NO_2)
- Ridge:
  - RMSE = 6.037
- LASSO:
  - RMSE = 6
- Best Subset:
  - 6 variables (all of them)
- PCR
  - RMSE = 10.42
- PLS
  - RMSE = 17.84
- Regression Tree
  - RMSE = 12.32

❖ Linear Model is preferred!

Conclusion

- To further extend my research I plan to use other response variables to extrapolate insights and findings hidden beneath the dataset.
- By doing this I think I will successfully be able to find specific correlations between the response and predictor variables.

Thank You for Listening