



Final Report

LUNG CANCER AND AIR QUALITY ANALYSIS

No Error

APR 17, 2020



CONTENTS

Part I. Background Knowledge and Assumption	3
Part II. Data Sources and Data Preparation	3
2.1 Data Source and Initial Quality	3
2.2 Data Wrangling Using Trifacta Wrangler	5
2.3 Final Data Quality	6
2.4 Data Transformation	6
Part III. Descriptive Statistics and Target Variable	7
3.1 Descriptive Statistics	7
3.2 Scatter Plot Matrix	12
3.2.1 Original Numerical Target Variable for Regression	12
3.2.2 Transformation for Classification	15
3.3 Conclusion: Target Variable Chosen and Justification	19
Part IV. Regression Model with Numeric Target Variable	20
4.1 Simple Multi Regression	20
4.2 Feature Engineering in Multiple Regression Model	25
4.2.1 Hierarchical Clustering	23
4.2.2 KMeans Clustering	23
4.3 Applied Feature Engineering to Multiple Regression Model	25
4.4 Model Comparison	29
4.4.1 Performance in Training Set	31
4.4.2 Performance in Testing Set	31
4.5 Interpretation of Multi Regression Model	30
Part V. Classifications Model with Categorical Target Variable	31
5.1 Logistic Regression	32
5.2 KNN Classification	33
5.3 Decision Trees Classification	34
5.4 Interpretation of Classification methods	35
Part VI. Data Robot Modeling	36
6.1 Light Gradient Boosted Trees Regressor with Early Stopping	36
6.2 Random Forest Classifier	40
6.3 Model Comparisons	43
Part VII. Conclusions and Reflections	44
7.1 Conclusion and Insights	44
7.2 Reflections	45

Part I. Background Knowledge and Assumption

The effects of future climate change on public health are an active and growing area of research, the impacts of climate extremes on future air quality and associated health implications are also under analysis.

Lung cancer is the first cancer killer of both men and women in the United States. It forms in the tissue of the lung, usually in the cells lining air passages. In addition to cigarette smoke, there are many environmental exposures that can raise the risk of lung cancer death. This is particularly the case for PM_{2.5}. Epidemiological studies indicate that PM_{2.5} has substantially greater toxicity than other larger particles: PM_{2.5} is associated with greater increases in daily mortality than other particles, and are of greater public health concern, suggesting that size is not the only indicator of PM-related health effects. Besides, it's worth noting that it is impossible to get a total picture of the environmental effects on health by only measuring a single environmental exposure. Rather, various environmental exposures occur simultaneously, to engender poor health upshots including lung cancer.

However, most people still don't know that pollution is a risk factor for lung cancer. That's why it is important we need to analyze the relationship between air quality and lung cancer.

Our assumption for this analysis is that PM 2.5 would be a leading factor of lung cancer mortality and the PM 2.5 level may have a positive relationship with mortality rate, which means, higher PM_{2.5} would cause greater increase in lung cancer mortality.

In this project, we will apply machine learning tools including regression, clustering and classification to investigate how largely these environmental factors can affect lung cancer incidence and suggest that the government and organizations should pay more attention on causes of lung cancer mortality.

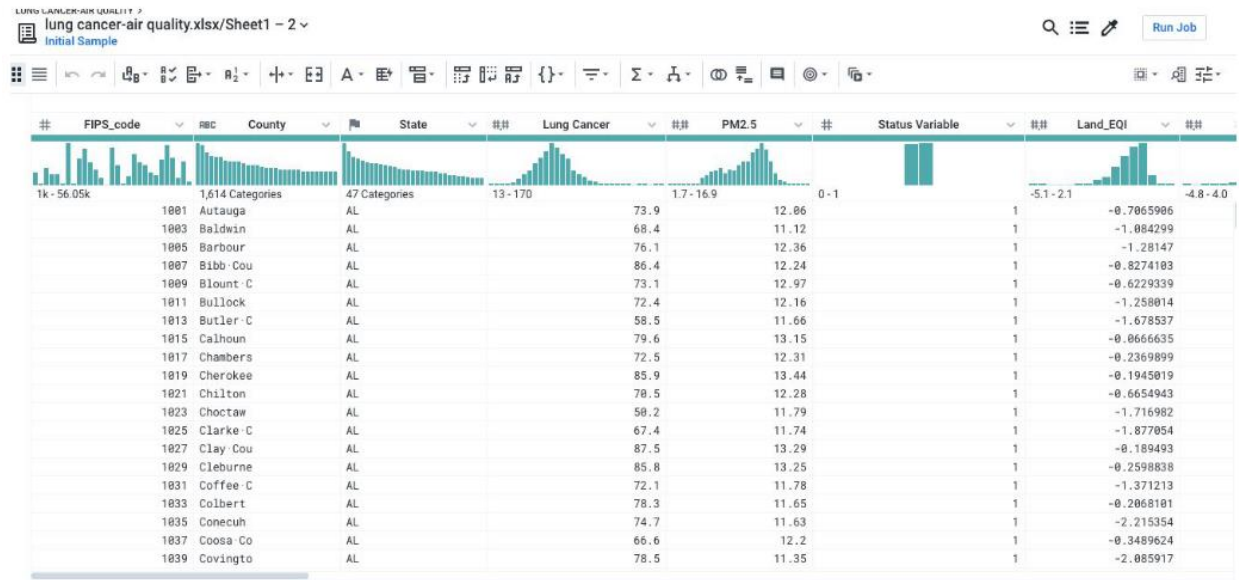
Part II. Data Sources and Data Preparation

2.1 Data Source and Initial Quality

We found the "Air Quality-Lung Cancer Data" from the Harvard Dataverse website. From its description, we know that the original data comes from two different sources. The Population-

based lung cancer incidence rates data were abstracted from the National Cancer Institute state cancer profiles, which is a national county-level database collected by state public health surveillance systems. And the domain-specific county-level environmental quality index (EQI) data were abstracted from the United States Environmental Protection Agency (USEPA) profile. Such data sources are reliable, so we perform our data wrangling process on this data using Trifacta Wrangler.

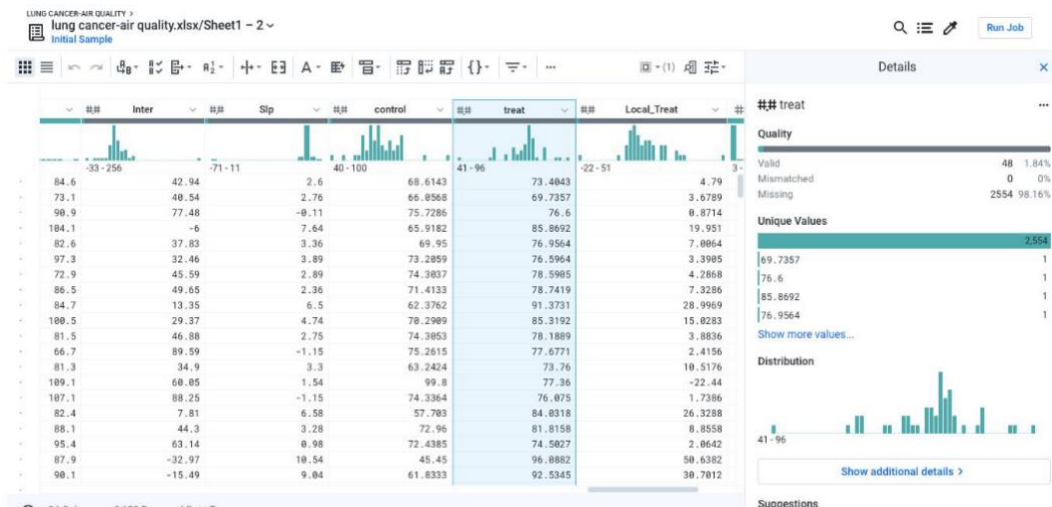
Screenshot 1



However, out of 3144 counties in the United States this dataset has available information for 2602 counties, some state's data is not available. Besides, some variables have large percentage missing values and are not relevant to our analysis goal, so we may need to remove them.

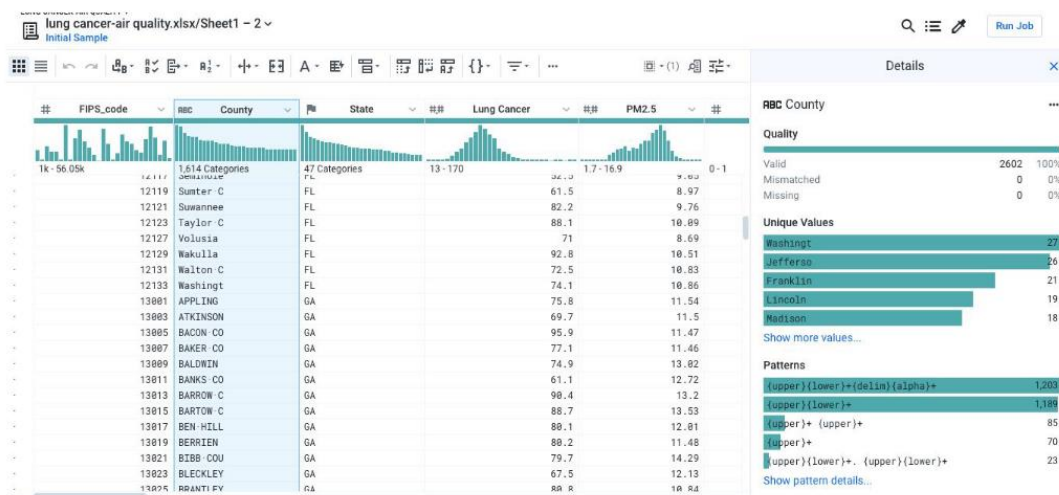
For instance, value "treatment" only contains 1.84% valid values.

Screenshot 2



Moreover, we notice that some county names are not in the same format(some are uppercase in all letters), which should be modified:

Screenshot 3



Then we start our data wrangling process.

2.2 Data Wrangling Using Trifacta Wrangler

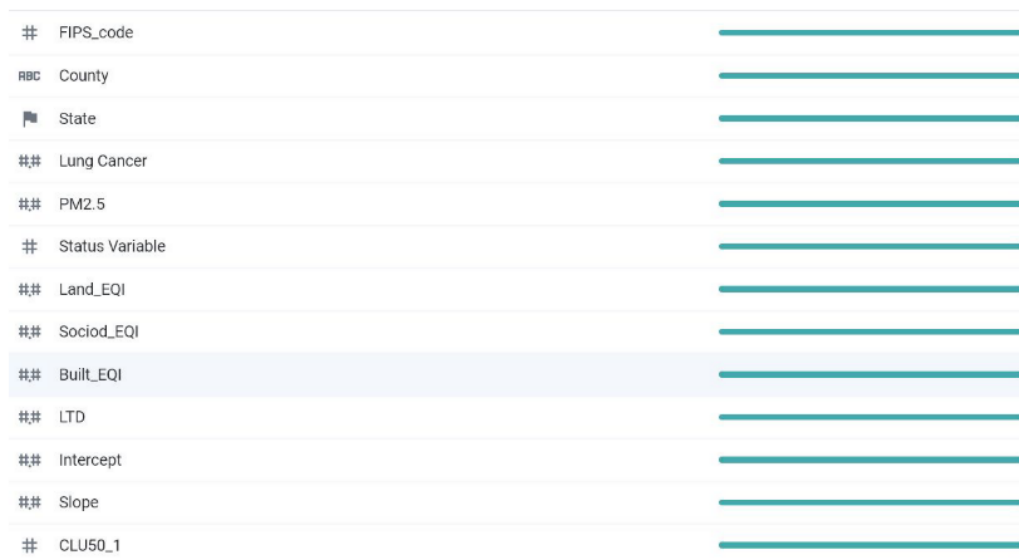
Firstly, Lung cancer mortality information in states namely Kansas, Michigan, Minnesota, and Nevada is not available, so we deleted the rows of these states. Then, we deleted the columns of Inter, Slip, control, treat, and Local_Treat since the valid data in the five columns only occupy less than 2% of the rows and information is not important for our goal. The Union county,

Florida is an outlier in terms of mortality, so we deleted it from the dataset. Finally, we changed all the county names in uppercase to lowercase with the first letter capitalized. There are columns such as intercept and slope, which seem to be irrelevant, but they help to divide the dataset into several clusters and are helpful for understanding the distribution of data.

2.3 Final Data Quality

The final dataset has 100% valid values and no mismatching values or missing values. And the values in some columns are more consistent and relevant, and there are fewer outliers. All the values are in the right formats, so the final dataset has better uniformity. Since most of the data are from the official resources, they are likely to be accurate and reliable. Additionally, the dataset concludes the most used air pollution index and air quality index, it is relatively complete.

Screenshot 4



##	FIPS_code	
ABC	County	
🇺🇸	State	
##	Lung Cancer	
##	PM2.5	
##	Status Variable	
##	Land_EQI	
##	Sociod_EQI	
##	Built_EQI	
##	LTD	
##	Intercept	
##	Slope	
##	CLU50_1	

2.4 Data Transformation

It's necessary to transform and standardize variables when using some classification methods. We will introduce how we transform data in the next section.

Part III. Descriptive Statistics and Target Variable

3.1 Descriptive Statistics

The dataset has 29 variables. However, we ignored irrelevant ones, only keeping “Lung Cancer” as target variable and other 16 variables as candidate predictors. A description of each important variable and descriptive statistics are given below.

Table 1 -- Variables Description

Lung Cancer	Lung cancer mortality(average number of people killed by lung cancer per 100,000 people)
CN, Disel	The concentration of Cyanide and diesel exhaust in the air
SO2, NO2, O3, CO, CS2	sulfur dioxide, nitrogen dioxide, ozone, carbon monoxide and carbon disulfide concentrations in air
PM2.5	The concentrations of particulate with aerodynamic diameter < 2.5 μm
PM10	The concentrations of particulate with aerodynamic diameter < 10 μm
EQI, Water_EQI, Land_EQI, Built_EQI, Sociod_EQI, Air_EQI	Environmental Quality Index(EQI) and its five domain indices(water, land, built, air and sociodemographic), presenting cumulative environmental quality. For each index, higher values correspond to poorer environmental quality

Table 2 -- Descriptive Statistics

Numerical Variables					
	min	mean	median	max	std
Lung Cancer	12.9	69.17091	68.6	169.9	17.418
PM2.5	1.7	10.12524	10.65	16.91	2.341308
PM10	1.22	11.90521	11.53	39.55	4.843881
SO2	1.000001	233.5402	143.0972	12180.53	407.2483
NO2	1.012354	592.9689	457.0564	8661.627	571.2364
O3	1.64	5125.998	4574.93	80276.76	4305.516
CO	1.112084	604.8169	356.3005	24815.74	798.7597
CS2	0	0.00545	0.000323	2.3	0.073157
Air_EQI	-2.81905	0.207845	0.230291	2.78984	0.825925
Water_EQI	-1.64127	0.042511	0.250895	1.478177	0.985815
Land_EQI	-5.11581	0.033273	0.174875	2.094526	0.845161
Built_EQI	-3.99272	0.079856	0.179253	3.883786	0.863834
Sociod_EQI	-4.80999	0.009265	0.00463	3.979472	0.991924
EQI	-3.22	0.122344	0.13	2.85	0.88512
CN	0.00012810	0.02427167	0.01540605	1.34934220	0.04475452
Disel	0.0168773	0.3931278	0.2969613	8.8151163	0.3885572

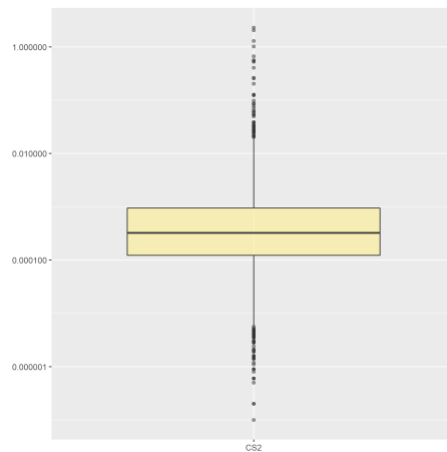
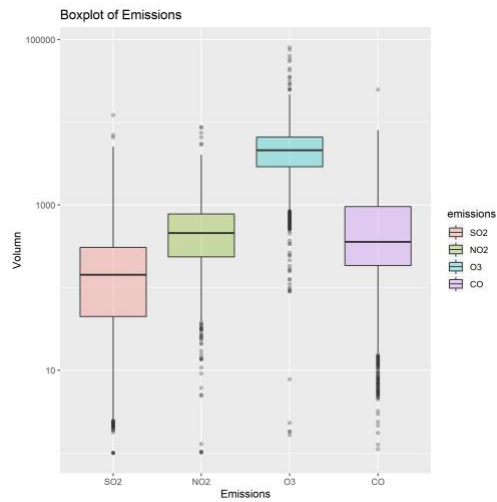
Lung Cancer Mortality

a) Distribution

Plot 1 -- Histogram and Boxplot of Lung Cancer Mortality

SO₂, NO₂, O₃ and CO are four major emissions leading to poor outdoor air quality. Generally, the concentration of CS₂ is not as high as other 4 emissions. However, even a small amount of CS₂ present in the air will result in great pollution.

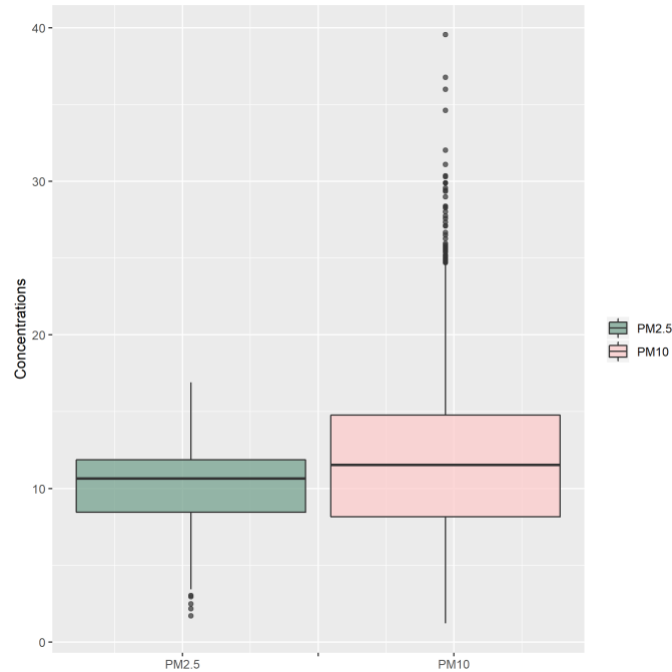
Plot 3 -- Boxplot of Emissions



For better review, we transform the Y axis into log scale. We can tell that the emissions across different areas vary largely.

Particles

Plot 4 -- Boxplot of Particles

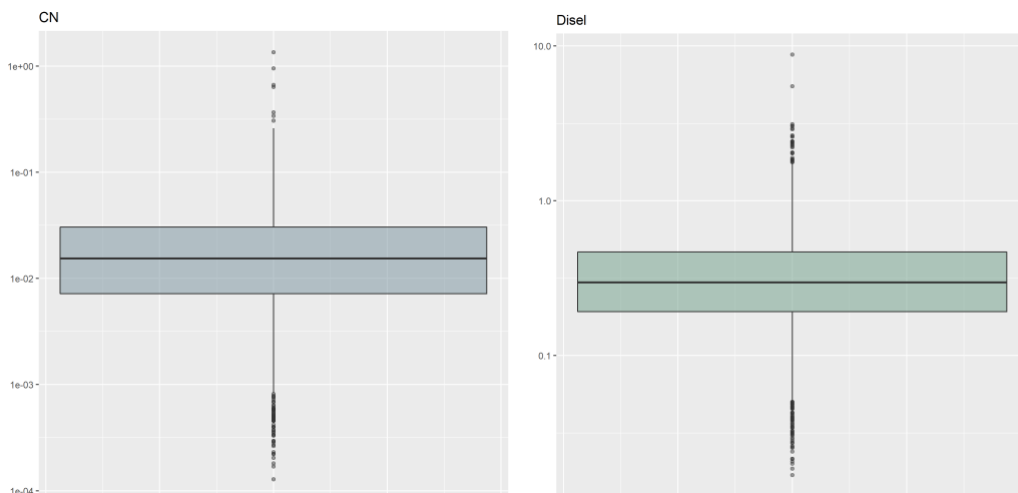


The concentration of PM10 is slightly right skewed and we can see a few outliers even larger than 26. By contrast, the distribution of PM2.5 has a thick left tail with 70% of data exceeding 9, which indicates needed transformation.

It should be noted that PM10 includes PM2.5 so it's not appropriate to put both of them into the same regression model.

Other Substances in the Air

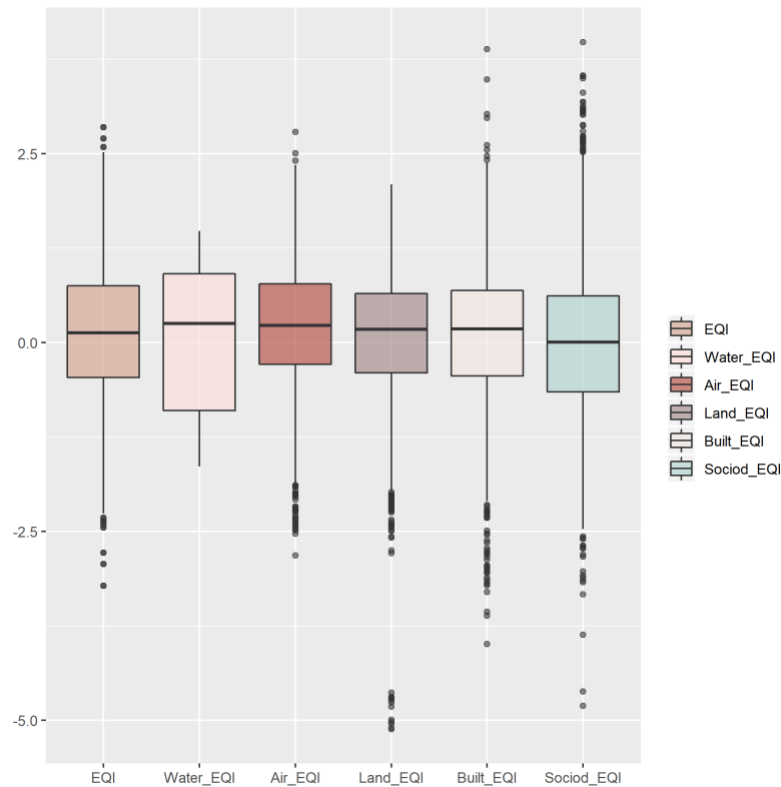
Plot 5 -- Boxplot of Cyanide and Diesel Exhaust



We take log scale both for CN and Disel, which means that both variables are not Gaussian-like distributed.

Environmental Quality Index

Plot 6 -- Boxplot of EQI



Environmental Quality Index demonstrates environment quality from different perspectives, ranging from -5 to 5. We can see that water quality doesn't have that huge variation as the other indicators.

Since EQI is an overall environmental quality indicator generated by other five indices, it will not be included in the model to avoid confounding effects. This is the same for Air EQI, which can be indicated by PM2.5.

3.2 Scatter Plot Matrix

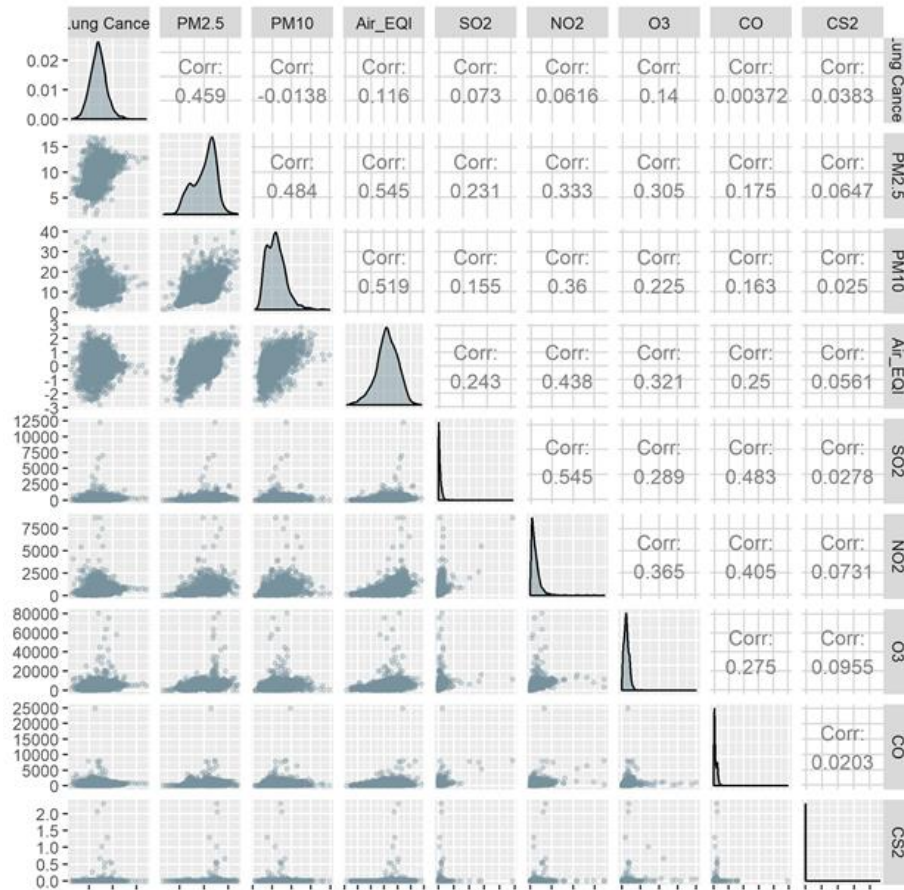
3.2.1 Original Numerical Target Variable for Regression

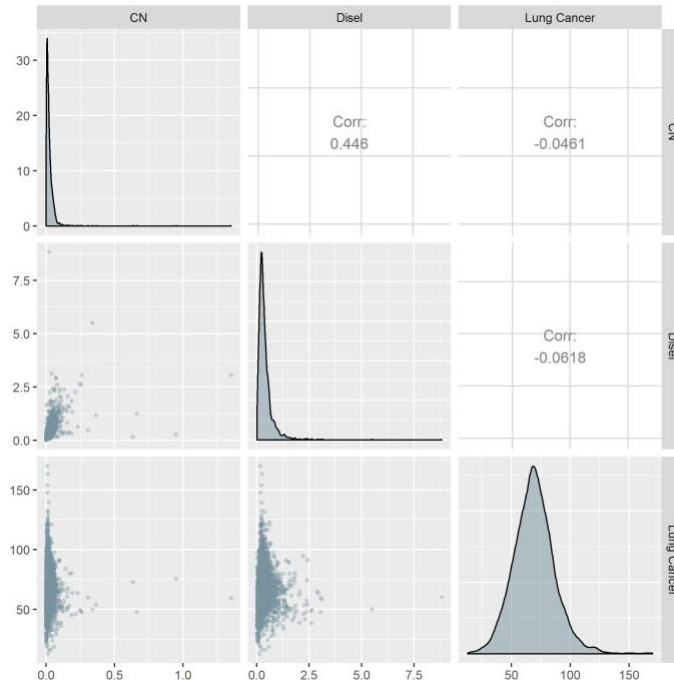
To further make sure which variables should be selected, we apply the scatter plot matrix to show the potential linear or curvilinear relationship between variables, and to detect multicollinearity. We also plot a scatter plot matrix for all transformed variables.

Due to too many candidate predictors, we create 3 matrices, two for air quality indices and another for other environmental indicators.

Air quality indicators

Plot 7 -- Scatter Plot Matrix of Lung Cancer and Air Quality Indicators



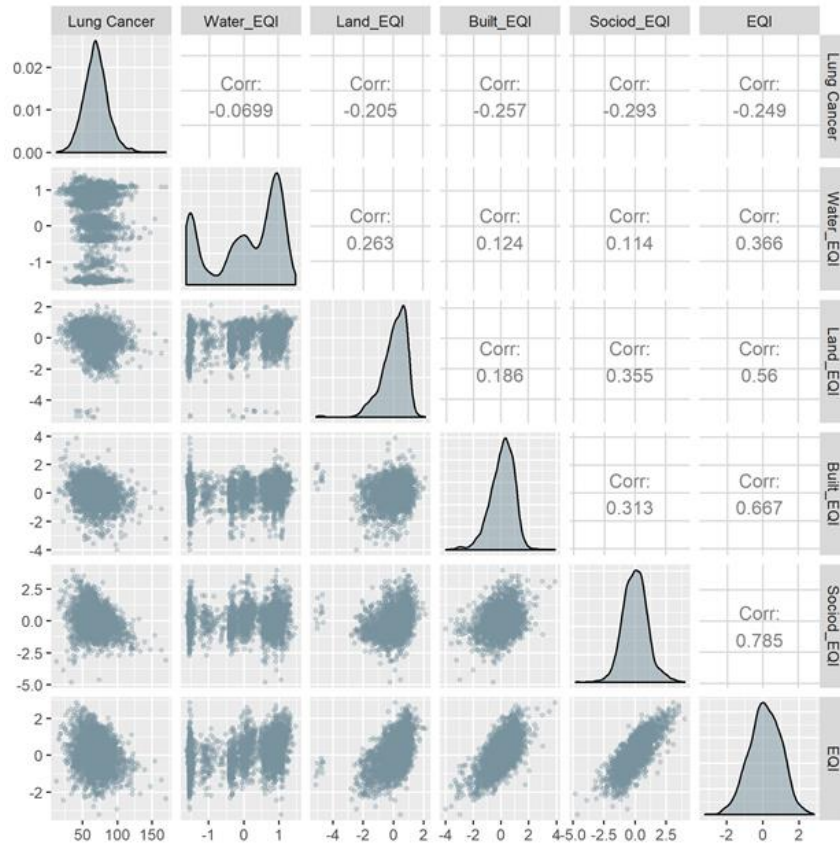


We can see that lung cancer mortality has a relatively strong correlation with the concentrations of PM_{2.5} while the correlation between lung cancer death rate and PM₁₀ is nearly 0. This is the same story for Air EQI, with a weak linear relationship with mortality. Besides this, we also discover quite a strong correlation between PM_{2.5} and PM₁₀, and between PM_{2.5} and Air EQI. Out of both insignificant correlation and reason to avoid confounding effects, once we choose to include PM_{2.5} into regression, PM₁₀ and Air EQI are supposed to be dropped. Moreover, the scatter plot matrix demonstrates a left skewed distribution for PM_{2.5}, suggesting a necessary transformation when creating the model.

In terms of emissions, there's a medium correlation among SO₂, NO₂, O₃ and CO, probably due to same or similar sources. We can also tell that the correlation between lung cancer mortality and these emissions is almost 0 and the scatter plots fail to show any obvious linear or curvilinear relationship. This is the same for CN and Diesel. But we still want to put them into a regression model to see if they are statistically significant.

Other environmental indicators

Plot 8 -- Scatter Plot Matrix of Lung Cancer and Other Environmental Indicators



Land EQI, Built EQI and Sociodemographic EQI have moderate correlation with lung cancer mortality while the relationship between Water EQI and mortality is far from obvious. Like emissions, we will regress lung cancer mortality on these 4 indices regardless of weak correlations.

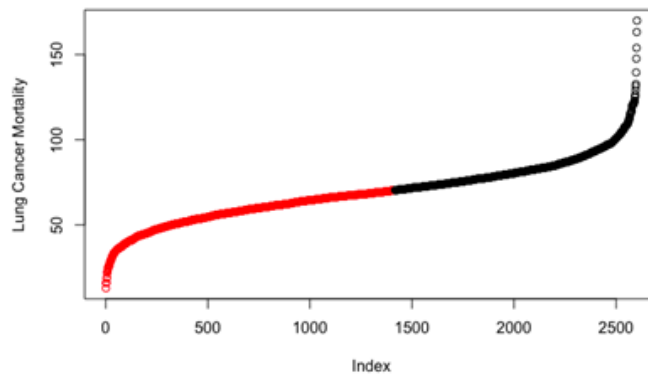
When it comes to EQI, as an overall indicator, it is calculated by other EQI indices so a strong linear relationship with them is justified. In order to remove perfect multicollinearity, overall EQI will not be shown in the regression.

3.2.2 Transformation for Classification

Due to purpose of classification, transformation is made:

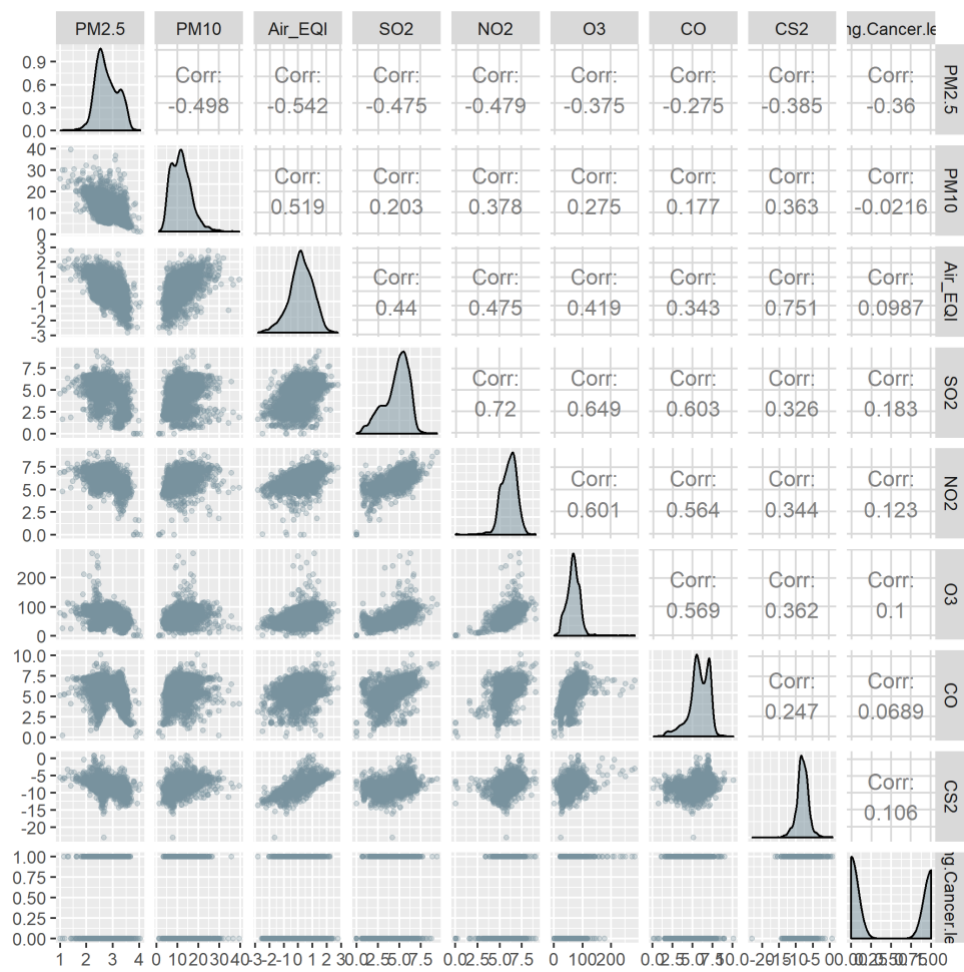
First, we transfer the target variable into a categorical variable. We cluster the target variable to transfer it from a continuous variable to a dummy variable. The value of Lung Cancer Mortality Rate which is less than 70.3 is divided into cluster 0, and we assume cluster 0 means low risk of Mortality. The value of Lung Cancer Mortality Rate which is bigger than 70.4 is divided into cluster 1, and we assume that cluster 1 means high risk of Mortality.

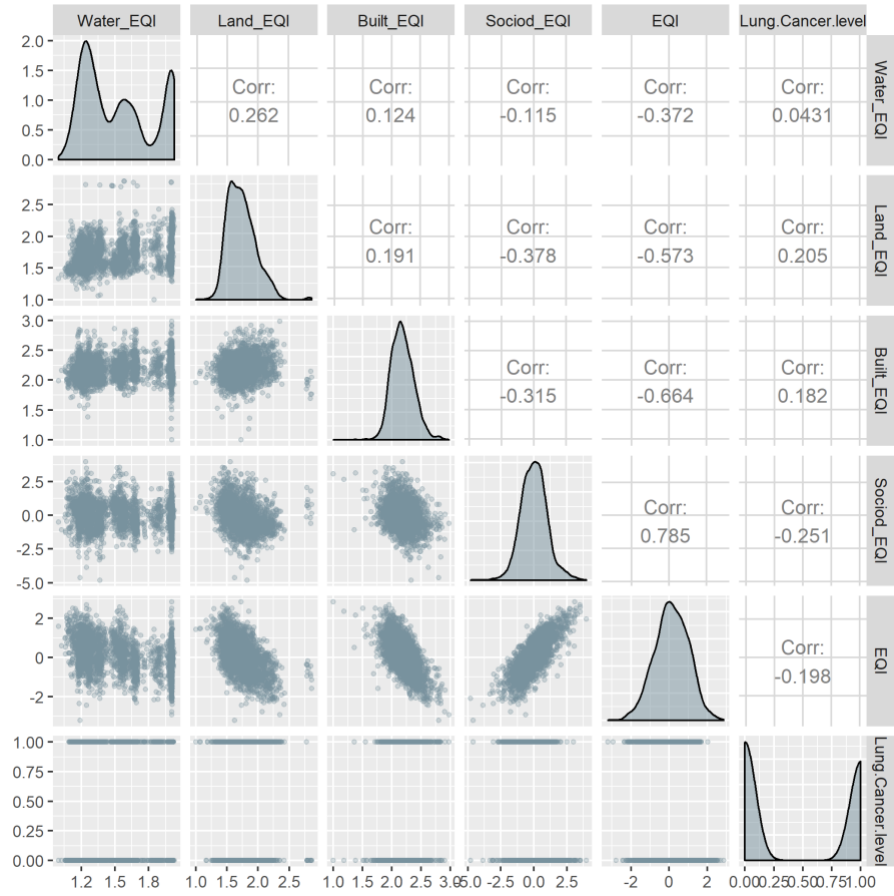
Plot 9

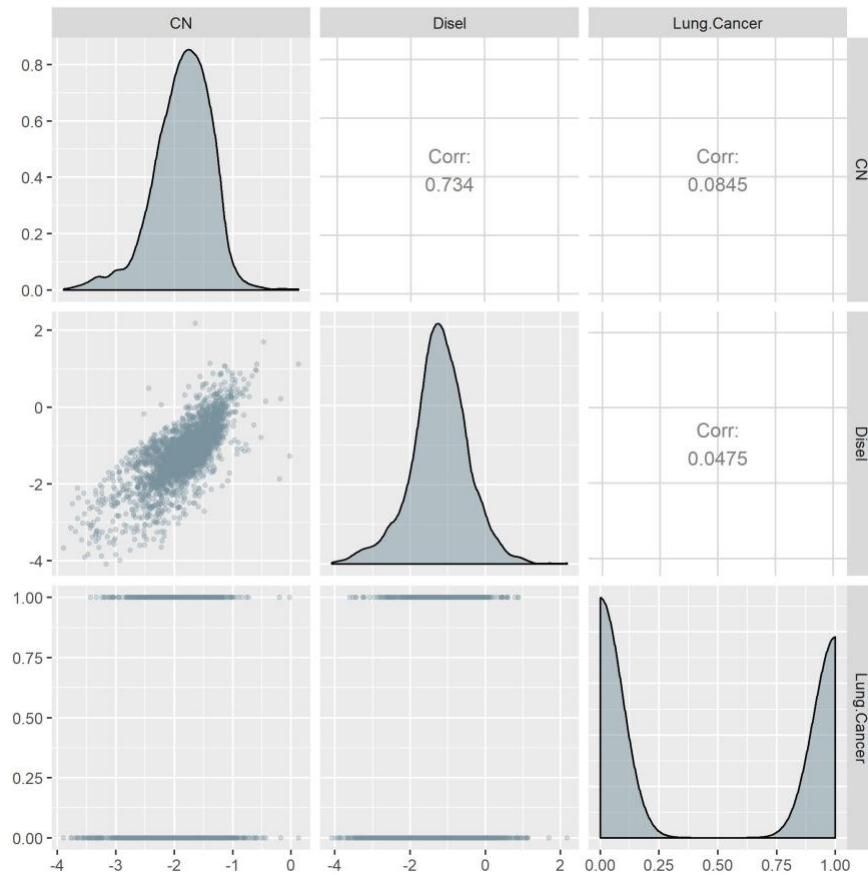


Then, In order to make other variables more Gaussian, we test the skewness of each predictor and make some transformation. For example, square root for moderate skew, log for greater skew while reverse for severe skew. After the transformation, the distribution of the data improved. Below are the scatter plot matrix after transformation:

Plot 10 - Scatterplot Matrix After Transformation







We can see that after transformation, the distribution of variables all looks more normally distributed.

3.3 Conclusion: Target Variable Chosen and Justification

To examine the connection between lung cancer survival and environmental factors, we take Lung Cancer (lung cancer mortality) as the outcome variable not only because the meaning of the variable is well matched with our purpose but also because it's not biased, with normality and no extremely unusual values. Even though there are a few leveraging points largely deviating from the mean, they are crucial for our analysis. Due to the purpose of classification, we transformed the target variable into dummy variables.

As for numerical candidate regressors indicating environmental condition, we divide them into two groups, respectively air quality indicators and other environmental indicators. For those demonstrating air quality, we only pick PM2.5 and 5 harmful emissions, removing PM10 and Air EQI to avoid confounding effects.

In terms of other environmental indices, we exclude overall EQI to prevent perfect multicollinearity.

Also, to make the variables more gaussian distributed, transformations were made.

Part IV. Regression Model with Numeric Target Variable

We want to explore the connections between lung cancer mortality and environment indicators. Firstly, we made a multi regression model by selecting variables, removing insignificant variables, and removing outliers (4.1). Secondly, we did feature engineering on predictive variables by making clustering labels to supplement them (4.2). Finally, we added these two columns of new labels to the original dataset and made multi regression again (4.3).

4.1 Simple Multi Regression

In our dataset, there is data from 47 states and 2602 counties. Since we want to explore the relationship between lung cancer mortality and environmental indicators, we regard each county in each state as a sample. So the size of our dataset, which is represented by n , is 2602.

To be mentioned, most outliers in our dataset have been removed in the data wrangling process. So for this time, we decided to choose the model first and then to delete omitted outliers.

Based on our conclusion, Lung Cancer Mortality is an ideal target variable. For important variables, after removing PM10 which has little impact on lung cancer mortality and removing Air EQI as well as overall EQI to avoid multicollinearity, we had 10 candidate variables. At first, we made a simple linear regression including all these variables and checked the VIF (variance inflation factor) of this regression model. Since we have deleted variables that could cause multicollinearity, the VIF result of each variable is around 1 to 2, which is quite ideal. It means multicollinearity doesn't exist in these variables.

Then, through the subset method, the model with 7 variables is the best model. It has the lowest prediction error but the p-value of Water_EQI is higher than what we expect. Under this consideration, we decided to delete this predictor variable, so our final model has 6 predictor variables, which are PM2.5, O3, CO, Land_EQI, Sociod_EQI, Built_EQI. Three of these variables are related to air pollution, and the other three are related to other environmental indicators.

In the Normal Probability Plot of Residual plot, there are several points that have a quantile lower than -3 or bigger than 3, so we could identify them as outliers and remove them next. We also checked residuals vs leverage and found that there are several leverage points at the right side of the plot. One point whose label is 2303 has a high residual and it can be regarded as an influential point. So it needs to be removed. We repeated this process twice until adjusted R-

Squared doesn't change and the RMSE changes only from 13.4 to 13.2, and we think that is acceptable. At this time, R-square of our model is 0.3151, which is not very high.

plot 11 - Summary of first model with 6 explainable variables

```
Call:
lm(formula = Lung_Cancer_Mortality ~ ., data = important_variables_6)

Residuals:
    Min       1Q   Median       3Q      Max
-47.355  -8.887   0.001   8.729  57.454

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.754e+01  1.268e+00  29.598 < 2e-16 ***
PM2.5        3.065e+00  1.286e-01  23.829 < 2e-16 ***
Land_EQI     -1.006e+00  3.555e-01  -2.829  0.0047 **
Sociod_EQI   -4.477e+00  3.181e-01 -14.077 < 2e-16 ***
Built_EQI    -1.862e+00  3.451e-01  -5.396 7.42e-08 ***
O3           3.533e-04  7.106e-05   4.972 7.05e-07 ***
CO          -1.667e-03  3.630e-04  -4.593 4.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.13 on 2589 degrees of freedom
Multiple R-squared:  0.3166,    Adjusted R-squared:  0.3151
F-statistic: 199.9 on 6 and 2589 DF,  p-value: < 2.2e-16
```

Then we add higher terms in the formula to try to improve the model performance. We try backwards search with both AIC and BIC to attempt to find a smaller, more reasonable model.

We try backwards search with both AIC and BIC to attempt to find a smaller, more reasonable model. Calculating the LOOCV RMSE for each, we see that the model chosen using AIC performs the best. That means that it is both the best model for prediction, since it achieves the good LOOCV RMSE, but also the best model for explanation, as it is also small. The items that used to build model is shown below:

```
Lung.Cancer ~ PM2.5 + Land_EQI + Sociod_EQI + Built_EQI + O3 + CO + Water_EQI + I(PM2.5^2)
+ I(O3^2) + I(CO^2) + I(Sociod_EQI^2) + I(Built_EQI^2) + I(Water_EQI^2) + PM2.5:Land_EQI +
PM2.5:Sociod_EQI + PM2.5:Built_EQI + PM2.5:CO + PM2.5:Water_EQI + Sociod_EQI:O3 +
Sociod_EQI:CO + Sociod_EQI:Water_EQI + Built_EQI:Water_EQI + CO:Water_EQI
```

After including the higher-order terms, the adjusted R-Squared value changes from 0.3151 to 0.396. The calculated RMSE is 13.2. The R-squared has been improved from 0.3151 to 0.396.

```
Residual standard error: 13.3 on 2560 degrees of freedom
Multiple R-squared:  0.404,    Adjusted R-squared:  0.396
F-statistic: 49.6 on 35 and 2560 DF,  p-value: <0.00000000000000002
```

```
#RMSE
sqrt(mean(autompg_mod_back_aic$residuals^2))
```

```
[1] 13.2
```

4.2 Feature Engineering in Multiple Regression Model

The next step we try to add two new variables by clustering existing variables and extract some features. Since we want to explore how environmental indicators can impact target variable: lung cancer mortality, we need to exclude this target variable from our clustering model.(This is what we did wrong in our last assignment.). Before running the clustering model, we prepared data with the selected candidate variables, removed the target variable and normalized them in order to get rid of scale differences.

4.2.1 Hierarchical Clustering

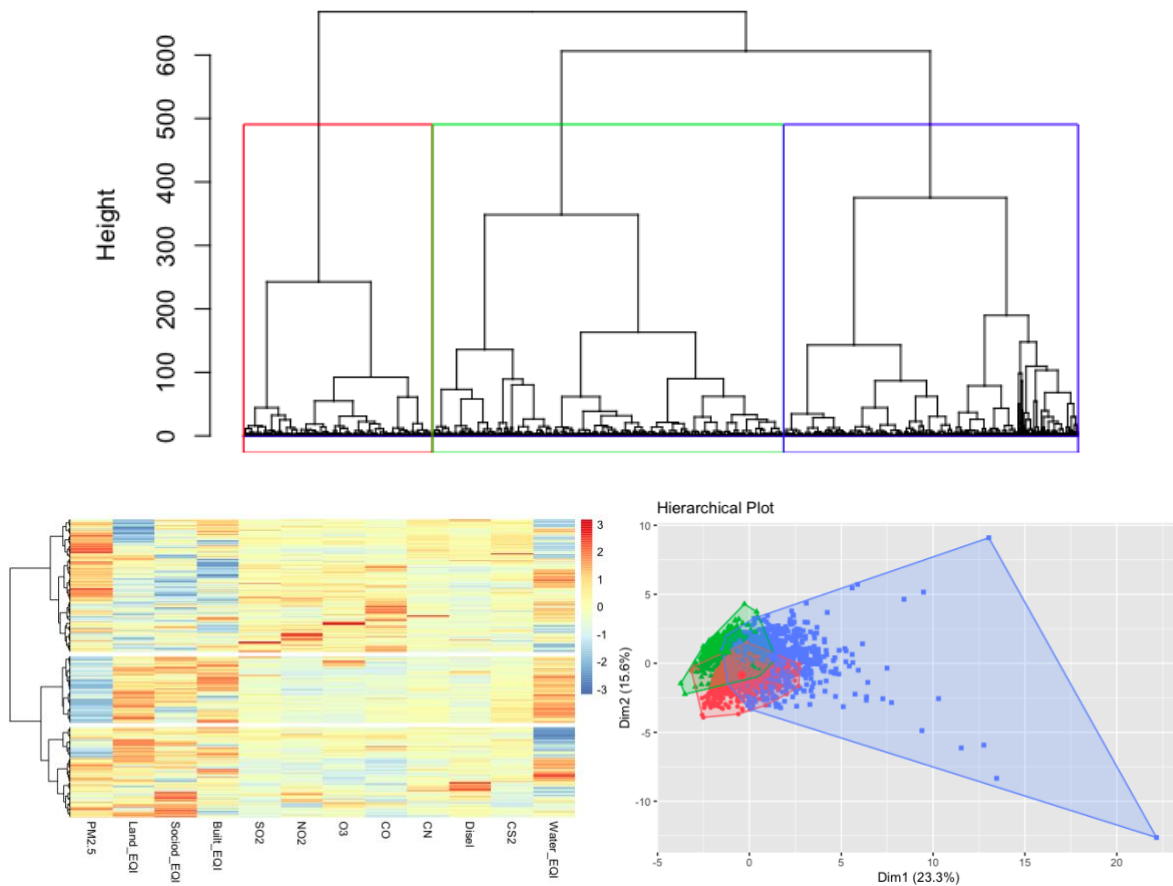
According to the descriptive analysis, we deleted some of the variables to avoid high collinearity(correlation >0.5). Then we changed the unique FISP code to the names of rows and normalized all the variables to get rid of scale differences.

After plotting models using methods- "single", "complete", "average","median", "centroid", "ward.D", we found the model using "ward.D" performed the best. So we chose the "ward.D" method model. Then, we used the NbClust package in R to choose the best "k". This package would try many different evaluation methods to vote the number of clusters that is supported by more methods. So we can see from the results that k equal to 3 is the best choice.

```
*****
* Among all indices:
* 3 proposed 2 as the best number of clusters
* 4 proposed 3 as the best number of clusters
* 2 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 2 proposed 7 as the best number of clusters
* 2 proposed 9 as the best number of clusters
* 1 proposed 14 as the best number of clusters
* 2 proposed 18 as the best number of clusters
* 1 proposed 19 as the best number of clusters
* 2 proposed 32 as the best number of clusters
* 1 proposed 36 as the best number of clusters
* 1 proposed 37 as the best number of clusters
* 1 proposed 50 as the best number of clusters

***** Conclusion *****
* According to the majority rule, the best number of clusters is 3
```

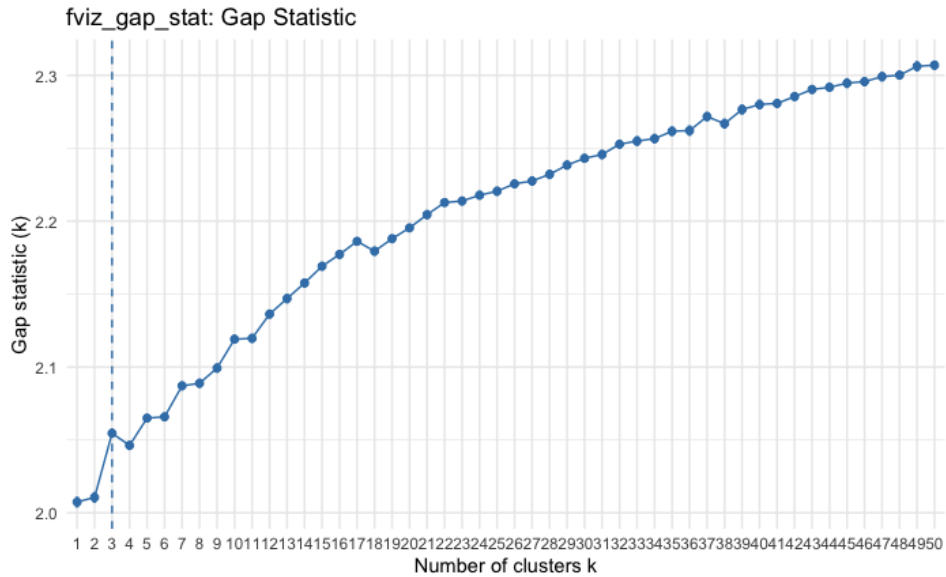
Then we plotted a hierarchical dendrogram and some other visualizations. From the dendrogram we could see that three clusters look quite reasonable.



These are the characteristics of each cluster. In general, we can conclude that each cluster is the combination of the characteristics of different environmental indicators.

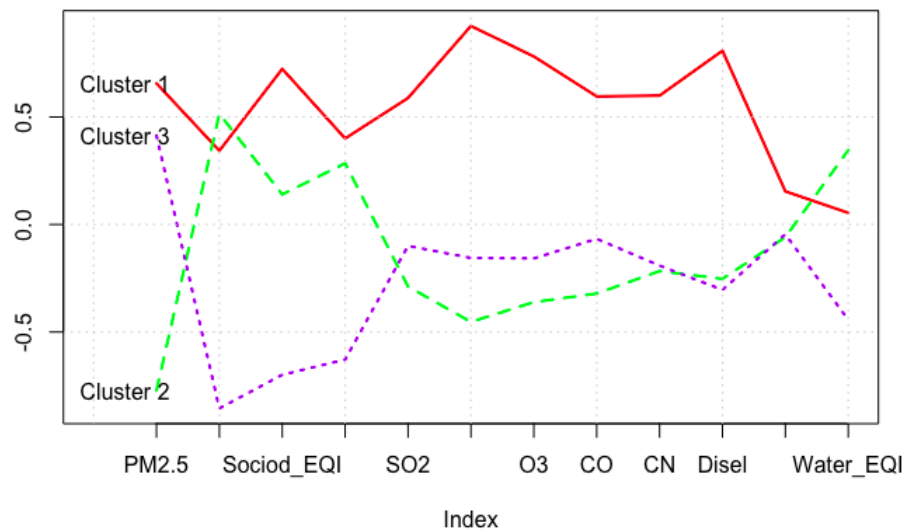
4.2.2 KMeans Clustering

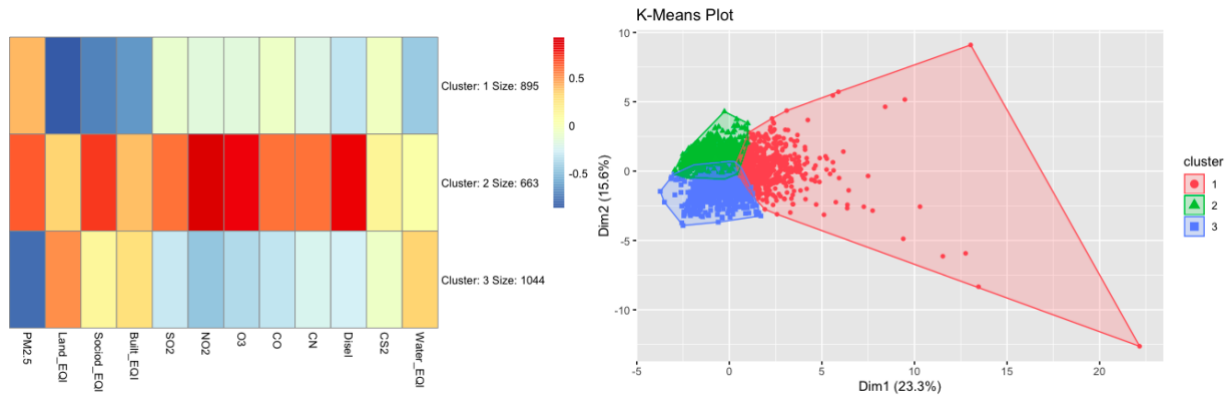
Then we used variables above to do the k-means clustering again. Firstly we ran the clustering to find out the best “k”, we tried gap statistic methods in deciding the k. The gap statistic would compare the total intracluster variation for different values of k with their expected values under null reference distribution of the data. The results are shown below:



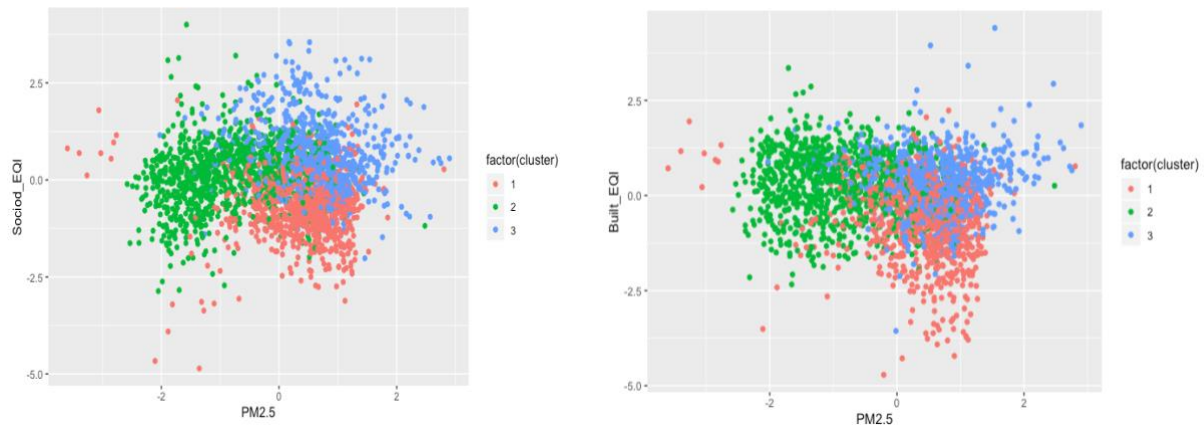
From this plot, we notice $k=3$ have the best gp statistic performance, which is consistent with our conclusion about hierarchical clustering. Then we ran the k-means model with k equals to 3.

From the line and cluster plots, we can see the distribution of the three clusters. And cluster 1 and cluster 3 have the largest distance.





Then we make some scatter plots to show the cluster differences between variables.



From the plots we notice that the three clusters are not completely distinct, but we can still find some similarities within each cluster and differences among them. For instance, cluster 1 shows relatively high PM 2.5 and low Sociod/Built EQI. While cluster 3 and 2 show higher Sociod/Built, cluster 3 has higher PM 2.5 level, in this case, cluster 3 may cause higher mortality.

4.3 Applied Feature Engineering to Multiple Regression Model

Then we add clustering labels to our original regression model to see what would happen. When we did our previous model, we used all of our data to run a regression model and got R-squared of 0.396. But now to be more accurate, we decide to split our dataset into training data and testing data. This is our training model before and after adding Hie_Cluster and KMeansCluster:

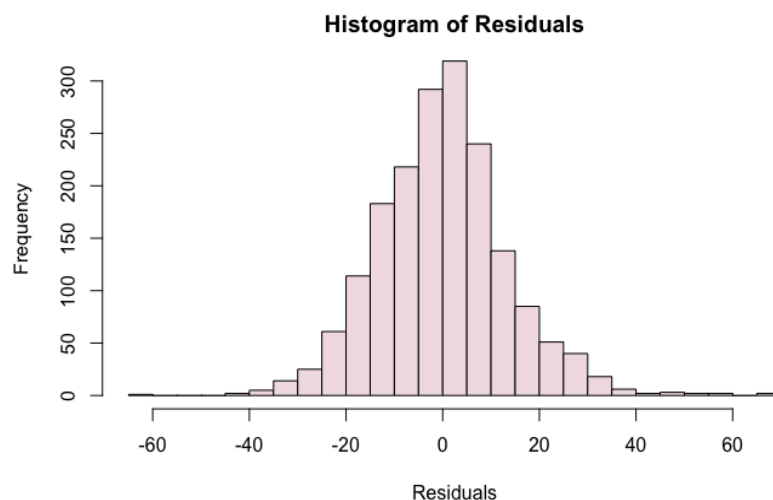
```
Lung.Cancer ~ PM2.5 + Land_EQI + Sociod_EQI + Built_EQI + O3 + CO + Water_EQI + I(PM2.5^2)
+ I(O3^2) + I(CO^2) + I(Sociod_EQI^2) + I(Built_EQI^2) + I(Water_EQI^2) + PM2.5:Land_EQI +
PM2.5:Sociod_EQI + PM2.5:Built_EQI + PM2.5:CO + PM2.5:Water_EQI + Sociod_EQI:O3 +
Sociod_EQI:CO + Sociod_EQI:Water_EQI + Built_EQI:Water_EQI + CO:Water_EQI + Hie_Cluster
+ KMeansCluster, data = training
```

We ran our original model with a training dataset at first. For training dataset, the R-squared is 0.4006, and adjusted R-squared is 0.3929. When we used this model to predict the testing dataset, the RMSE is 13.5139, and R-squared is 0.3927.

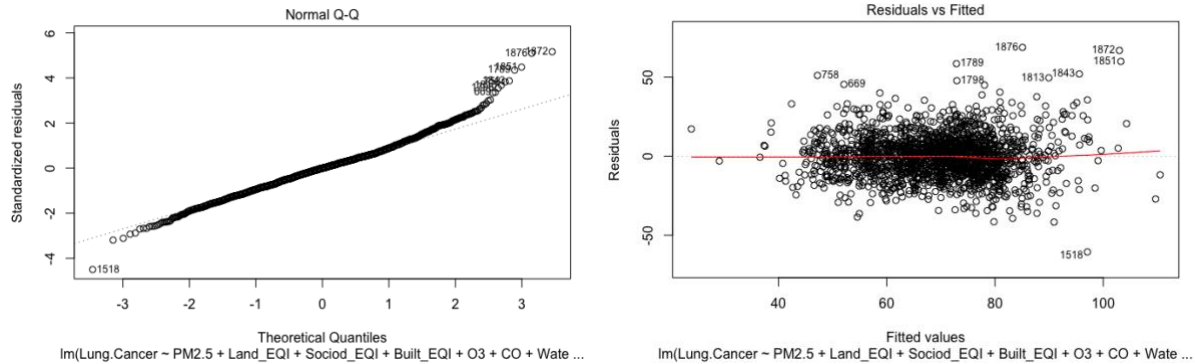
Then same with the original model, we use the training set in the new model with clustering variables. After running the model once, we notice that some of the regressors are not significant, so we remove them out. It should be worth noting that If $X1 \cdot X2$ is significant, we need to retain both $X1$ and $X2$ individual predictors, even if the p-value of $X1 > 0.01$ or p-value of $X2 > 0.01$

After ruling out insignificant predictors, the model gets improved with the adjusted R-squared for training set increasing from 39.29% to 39.79%. Then we check the VIF of the model. VIF for all the predictors are smaller than 10, while most of them are lower than 5, which means severe multicollinearity doesn't exist.

Then we check the residuals and want to remove outliers.

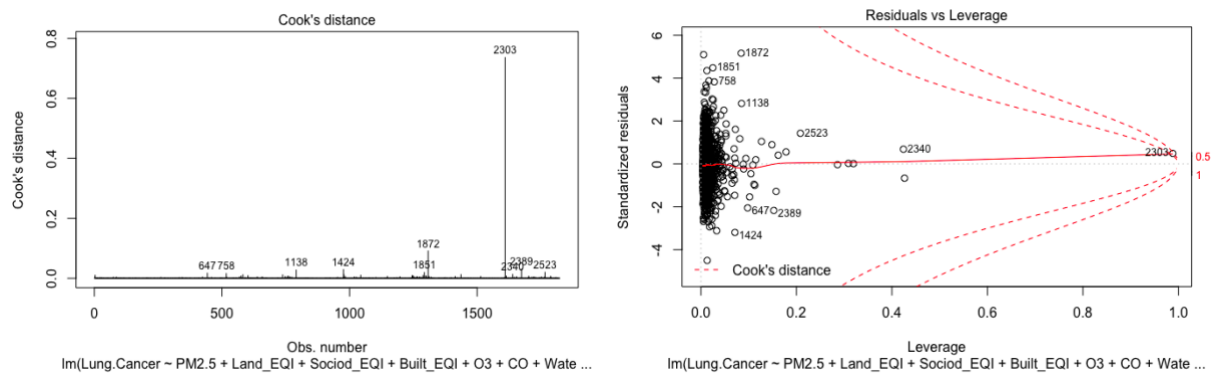


This plot shows the residuals of the regression model. We can see that the residuals are basically normally distributed, with several points higher than 50.



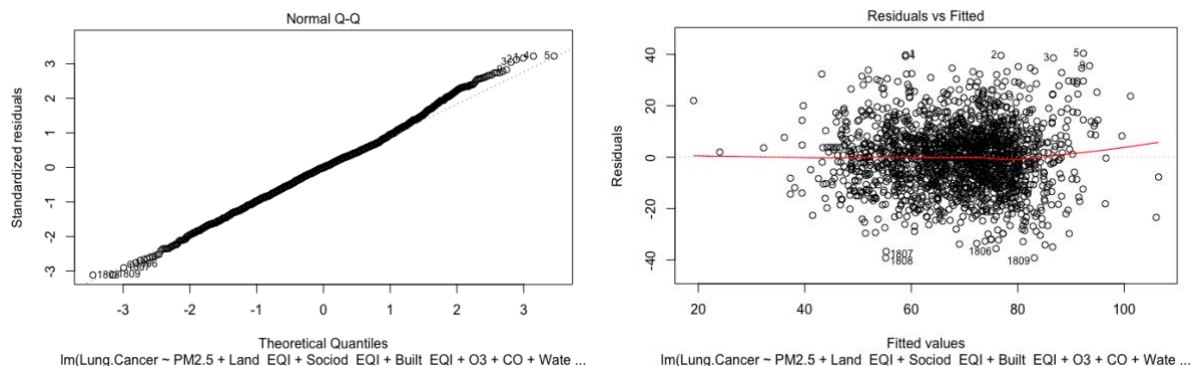
The residual vs fitted value plot demonstrates that most errors distribute randomly and evenly in two sides of X-axis. However, there are still some points in the upper and lower right part.

This is the same story for the QQ plot. In our case most of the residuals track the diagonal line, while some residuals are larger than standardized, which means that there are some “outliers” in the data that are difficult to be explained by our model. We plan to remove these points since our model is very likely to be affected by these outliers.



We also intend to find out leverage points. Not all leverage points are influential in linear regression analysis. Even though data have extreme values, they might not be influential to determine a regression line. For this consideration, we visit the Cook's Distance and residuals versus leverage plot. These plots help us to find influential cases. We could see although 2303 has a high leverage, but its residuals is around 0 and its cook's distance is also not higher than 1. So we keep this point in our model.

Based on what we found above, 15 observations were removed as outliers and $15/1796 = 0.0046$ only account for 0.83% of the observations in the training set. We ran the model again. For the performance of the training sample, the adjusted R-Squared increases from 39.79% to 41.56%. The Normal Q-Q plot and Residuals vs. Fitted plot look ideal now.



Our final model turned to be:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.765e+01	6.078e+00	4.550	5.72e-06	***
PM2.5	6.523e+00	1.156e+00	5.643	1.95e-08	***
Land_EQI	-2.520e+00	1.652e+00	-1.525	0.127388	
Sociod_EQI	3.581e+00	1.564e+00	2.290	0.022115	*
Built_EQI	4.321e+00	1.848e+00	2.339	0.019467	*
O3	8.803e-04	1.747e-04	5.039	5.17e-07	***
CO	-1.220e-02	3.089e-03	-3.950	8.13e-05	***
Water_EQI	-9.593e+00	1.630e+00	-5.886	4.71e-09	***
I(PM2.5^2)	-2.094e-01	5.793e-02	-3.616	0.000308	***
I(O3^2)	-1.073e-08	2.736e-09	-3.921	9.14e-05	***
I(CO^2)	1.967e-07	4.479e-08	4.391	1.20e-05	***
I(Sociod_EQI^2)	-9.390e-01	2.033e-01	-4.618	4.14e-06	***
I(Built_EQI^2)	-1.265e+00	2.508e-01	-5.043	5.06e-07	***
I(Water_EQI^2)	-1.347e+00	4.613e-01	-2.919	0.003557	**
Hie_Cluster2	-3.811e+00	1.498e+00	-2.544	0.011038	*
Hie_Cluster3	1.706e+00	1.054e+00	1.619	0.105692	
KMeansCluster2	1.343e+00	1.199e+00	1.120	0.262775	
KMeansCluster3	5.486e-01	1.308e+00	0.420	0.674840	
PM2.5:Land_EQI	1.361e-01	1.606e-01	0.848	0.396777	
PM2.5:Sociod_EQI	-8.027e-01	1.612e-01	-4.979	7.02e-07	***
PM2.5:Built_EQI	-5.751e-01	1.773e-01	-3.244	0.001200	**
PM2.5:CO	6.447e-04	2.866e-04	2.249	0.024617	*
PM2.5:Water_EQI	9.735e-01	1.567e-01	6.212	6.50e-10	***
Sociod_EQI:O3	-3.086e-04	9.673e-05	-3.191	0.001444	**
Sociod_EQI:CO	1.906e-03	5.253e-04	3.628	0.000294	***
Sociod_EQI:Water_EQI	-9.815e-01	3.654e-01	-2.686	0.007294	**
Built_EQI:Water_EQI	-5.995e-01	3.803e-01	-1.577	0.115066	
CO:Water_EQI	-9.816e-04	4.460e-04	-2.201	0.027887	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

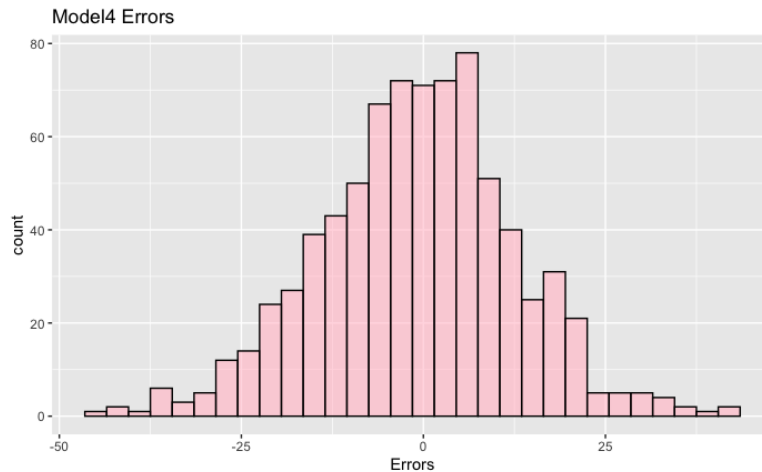
Residual standard error: 12.73 on 1781 degrees of freedom
Multiple R-squared: 0.4244, Adjusted R-squared: 0.4156
F-statistic: 48.63 on 27 and 1781 DF, p-value: < 2.2e-16

Last, we apply our new model into the testing set to see if there's any overfitting issue in our model. If overfitting exists, the performance in the validation set would be much worse. The R-

squared is 0.3956 and RMSE is 13.49711. The performance of the model for the testing set is only slightly different from the result for the training set. We can confirm that there's no overfitting in our new model.

R-Squared: 0.3956353
RMSE: 13.49711

The errors in the validations set seem to be basically normally distributed, also indicating that the new model performs well without overfitting issues.



4.4 Model Comparison

Last, we compare the new model with our original one to see if including cluster variables improves the model.

4.4.1 Performance in Training Set

Original model:

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>	
0.4005583	0.3928946	13.60384	52.26635	7.753429e-181	
df <int>	logLik <dbl>	AIC <dbl>	BIC <dbl>	deviance <dbl>	df.residual <int>
24	-7333.317	14716.63	14854.34	332931	1799

New model:

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>
0.4243596	0.4156329	12.73446	48.6276	2.805187e-191

df <int>	logLik <dbl>	AIC <dbl>	BIC <dbl>	deviance <dbl>	df.residual <int>
28	-7155.411	14368.82	14528.34	288818.7	1781

It's obvious to see that the new model performs better than the original one. Both R-Squared and adjusted R-Squared moderately increase, so that we confirm that the additional predictors are improving the model's performance. At the same time, the new model's RMSE is lower than the old one. This shows that the new model reduces the variance of our parameter which corroborates our conclusion that the new model does a better job modeling lung cancer mortality. Besides, larger F-statistic in the new model suggests that it provides a better "goodness-of-fit". Finally, we can see that the new model has lower AIC and BIC, so that it can be considered of better quality.

4.4.2 Performance in Testing Set

Original model:

R-Squared: 0.3927198
RMSE: 13.51388

New model:

R-Squared: 0.3956353
RMSE: 13.49711

In terms of the testing set, the new model also performs better than the old one, with smaller RMSE and larger R-Squared.

4.5 Interpretation of Multi Regression Model

From our regression result, we found the relationship about lung cancer mortality and multiple environmental equity indicators.

To be specific, for every unit increase of PM2.5, O3 level, it's an increase of lung cancer mortality in the target, among them, the PM 2.5 effect is the highest one, 1% increase of PM 2.5 level would increase the lung cancer mortality 6.23%, which means particle pollutants have positive relationship with the lung cancer mortality.

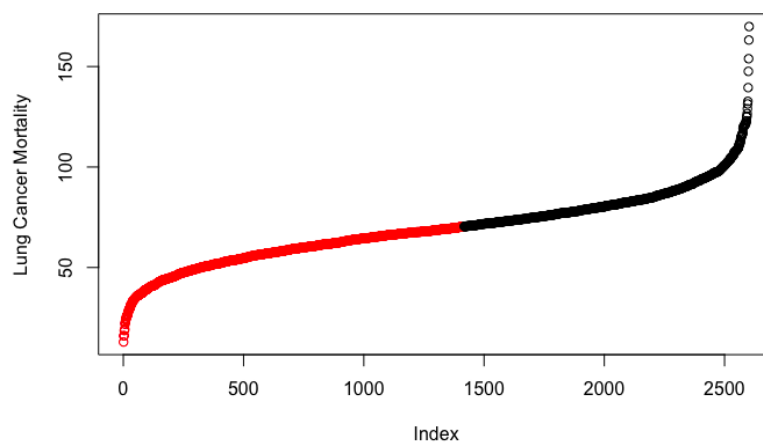
For environment indicators, for every unit increase of Sociod-EQI(which is the second important feature in our data robot modeling), Built-EQI, the mortality rate increases. For instance, when Sociod-EQI increases by 1%, the mortality will decrease by 3.58%. It makes sense because poorer environmental quality has a higher EQI number and may increase lung cancer motability.

Overall, our training model can explain 42.2% of the data. When we majorly concern the lung cancer mortality and PM 2.5 relationship, we see significantly high positive relation.

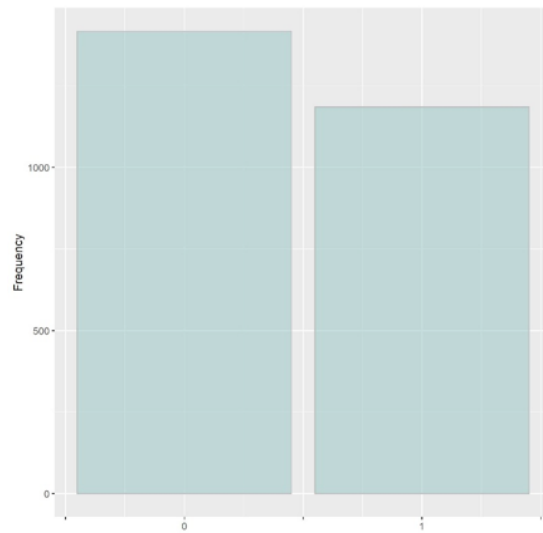
Part V. Classifications Model with Categorical Target Variable

In order to further explore the connection between lung cancer survival and environmental factors, we then apply classification models including logistic regression, KNN and classification trees in the machine learning part.

First, we need to transfer the target Variable into the categorical variable. We cluster the target variable to transfer it from a continuous variable to a dummy variable. The value of LungCancer Mortality Rate which is less than 70.3 is divided into cluster 0, and we assume cluster 0 means low risk of Mortality. The value of LungCancer Mortality Rate which is bigger than 70.4 is divided into cluster 1, and we assume that cluster 1 means high risk of Mortality.

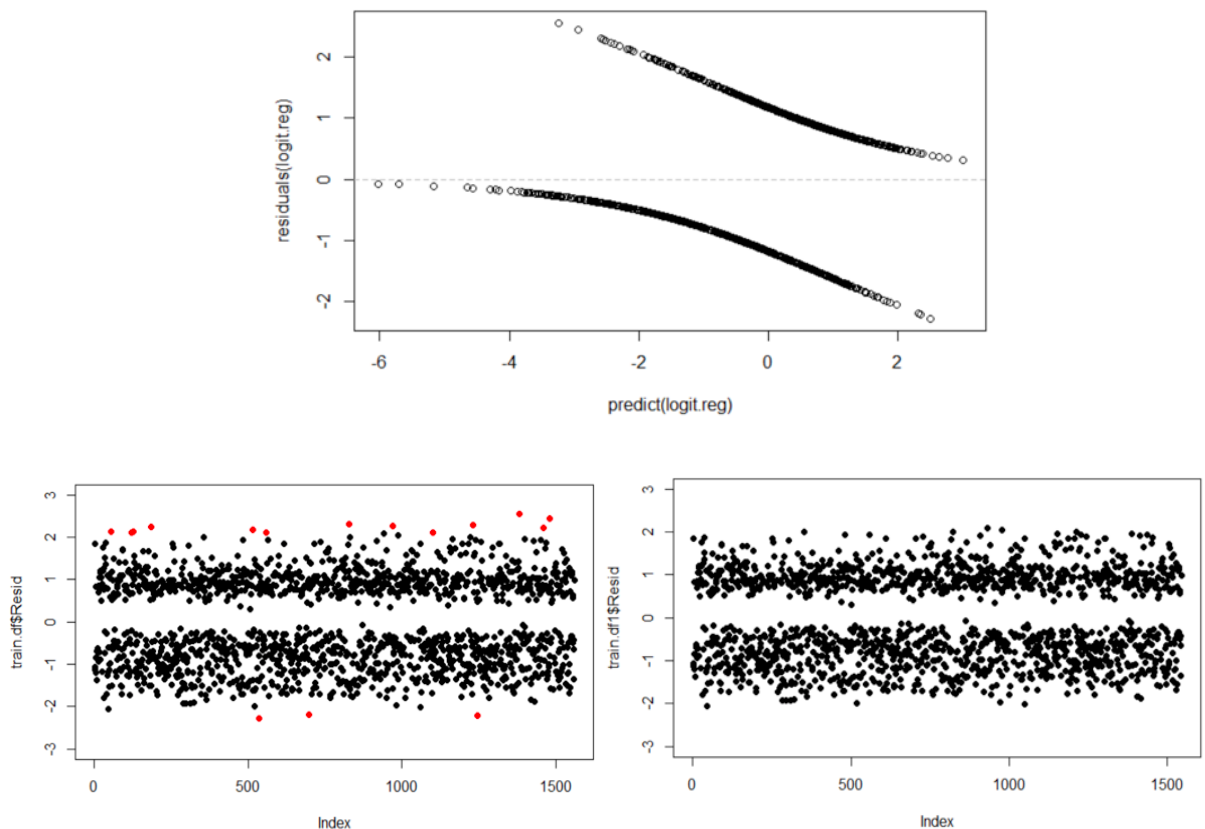


After the clustering, the bar plot of lung cancer mortality is shown below:



5.1 Logistic Regression

We first use the residual plot to see whether there are outliers and use 2 standard deviations as standard to find out the outliers. There are 11 outliers and we delete them from the database.



We divide our dataset into train data(60%) and test data(40%).Then we do the logistic regression on the training data after deleting the outliers. Look at the p values, we find the p values of the coefficients of the Built_EQI and Land_EQI are higher than 0.5. So the coefficients of the Built_EQI and Land_EQI are not statistically significant. We exclude the Built_EQI and Land_EQI from our model.

```
Call:
glm(formula = Lung.Cancer.level ~ PM2.5 + O3 + CO + Land_EQI +
     Sociod_EQI + Built_EQI + Water_EQI + I(PM2.5^2), family = binomial,
     data = train.df1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1034  -0.9073  -0.3073   0.9215   2.2151

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.060786   2.387641  -5.470 0.000000044962286179 ***
PM2.5        12.806966   1.817552   7.046 0.000000000001837755 ***
O3           0.007999   0.003767   2.124  0.03369 *
CO          -0.192769   0.068218  -2.826  0.00472 **
Land_EQI     0.547780   0.314720   1.741  0.08177 .
Sociod_EQI  -0.689517   0.078414  -8.793 < 0.0000000000000002 ***
Built_EQI    -0.095479   0.333610  -0.286  0.77472
Water_EQI    -0.822694   0.212680  -3.868  0.00011 ***
I(PM2.5^2)   -2.761708   0.339053  -8.145 0.000000000000000378 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We use the logistic regression model to predict the test data.The Accuracy shows that the overall predicted accuracy of the model is 0.72.The accuracy, sensitivity, specificity, and F1 scores all perform well, so there is no overfitting problem.

For each predictor, compare the odd ratio against 1. Less than 1 means a decrease of 1.0-odds ratio percent. Greater than 1.0 means an increase of odd ratio-1.0 percent vs. the base category.Exponentiating the PM2.5 coefficient tells us the expected increase in the odds of high Lung.Cancer.level for each unit of PM2.5. The ratio for high Lung.Cancer.level increase in the variable PM2.5 is 364749. This interpretation can be extended to other variables.

The additional benefits of interpreting odds instead of original number is that interpretations like the above are always true for any value of independent variables. But, the change in probability for a unit increase in a specific predictor is not constant, which depends on the specific values of the predictor.

5.2 KNN Classification

For KNN models, we divide our dataset into train data(60%) and test data(40%).There are two categories in our target variables, which is what we want our model to predict.

We do the KNN in R, by cross-validation, we determined the parameter k should be 21. Then we check the confusion matrix:

```

              Reference
Prediction high low
high    337 129
low     137 437

```

As a result, the accuracy of our KNN model is 74.42%, the sensitivity of this method is 71.10%, the specificity of this method is 77.21%, and the F1 Score is 71.70%.

5.3 Decision Trees Classification

Based on the descriptive analysis, we choose the lung cancer mortality rate as the target variable and keep the candidate variables, we also add higher terms($PM2.5^2$). Then we change the target to category variable based on the 70.3 threshold. In order to avoid overfitting and check the model's predictive accuracy, we split the data into train (60%) and valid.

We use cross validation to run the classification trees repeatedly. From the complexity-parameter table of cross-validation errors of increasing depth grown on the lung cancer data, we can simply choose the tree with the lowest cross-validation error (xerror). In our case, the tree is row number 8 with nsplit 10.

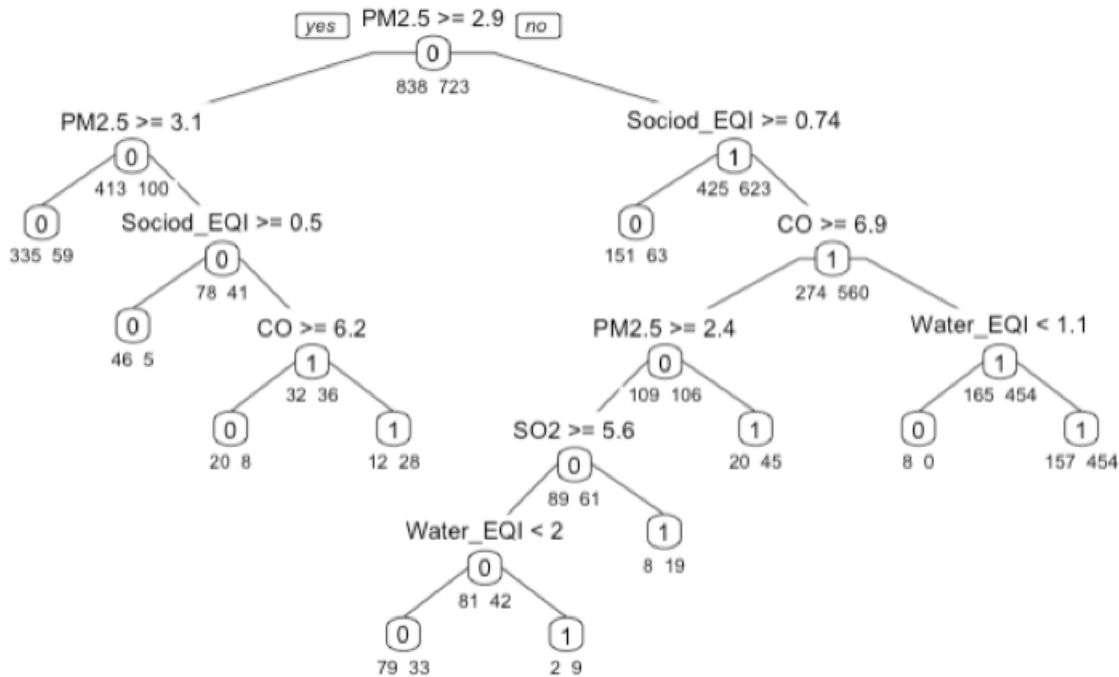
```

n= 1561
      CP nsplit rel error  xerror   xstd
1  0.27385892    0 1.000000 1.00000 0.027249
2  0.12171508    1  0.726141 0.73167 0.025866
3  0.01936376    2  0.604426 0.64454 0.025007
4  0.01521438    4  0.565698 0.64869 0.025053
5  0.01106501    5  0.550484 0.60719 0.024569
6  0.00968188    6  0.539419 0.58506 0.024289
7  0.00737667    7  0.529737 0.58230 0.024252
8  0.00599355   10  0.507607 0.57953 0.024216
9  0.00414938   15  0.473029 0.61272 0.024637
10 0.00391886   21  0.448133 0.63347 0.024882
11 0.00368834   33  0.390041 0.63485 0.024898
12 0.00359613   36  0.378976 0.63485 0.024898
13 0.00322729   43  0.349931 0.65007 0.025068
14 0.00276625   46  0.340249 0.65145 0.025083
15 0.00242047   74  0.262794 0.67358 0.025318
16 0.00230521   78  0.253112 0.67635 0.025346
17 0.00207469   81  0.246196 0.68050 0.025388
18 0.00184417   95  0.217151 0.68880 0.025471
19 0.00158072   98  0.211618 0.72476 0.025806
20 0.00138313  127  0.152144 0.72614 0.025818
21 0.00103734  152  0.116183 0.74274 0.025960
22 0.00092208  156  0.112033 0.74412 0.025971
23 0.00069156  167  0.099585 0.74827 0.026005
24 0.00001000  169  0.098202 0.75657 0.026072

```

Then we will use the lowest CP found in step 3 to regrow the tree and prune it. After that, we notice the length of nsplit is 11.

We plot the tree result, we can see the most important classification variable is PM 2.5.



Finally, we predict on valid data using the classification tree, and make the confusion matrix. From the matrix result, we see the accuracy of our model is 73.2%, while the sensitivity is 74.96%, the specificity is 71%, the F1 score is 0.7567.

5.4 Interpretation of Classification methods

We compare the accuracy, sensitivity, specificity and F1 score of three models. We find that the performances of the three models are very close to each other, indicating that they reinforce each other.

	Logistic Regression	KNN	Classification Tree
Accuracy	72.14%	74.42%	73.20%
Sensitivity	73.73%	71.10%	74.96%
Specificity	70.21%	77.21%	71.00%
F1 Score	74.38%	71.70%	75.67%

We can find that KNN performs best, with the KNN slightly higher than the number of the other two models. Compared with logistic regression and classification, the proportion of observed negatives that were predicted to be negatives is higher in the KNN model.

During the process, we built up three different models using variables selected and transformed. 60% data was partitioned into a training set while the rest was regarded as a validation set. As for the hyper parameters (like the K value in KNN and the number of splits in the tree model), we utilized cross validation. We used the validation test to report the accuracy, sensitivity, specificity and F1 score.

Based on our result, we noticed that environmental quality has a great impact on lung cancer mortality.

In the linear regression model that we performed, air quality is important in our model - PM 2.5 level shows significantly high positive relationship between pollution and mortality rate. In the decision tree model and logistic model we conducted this time, the same result is shown, indicating that air quality can be a vital determinant of lung cancer mortality.

Part VI. Data Robot Modeling

6.1 Light Gradient Boosted Trees Regressor with Early Stopping

We first upload the original data to DataRobot, in this data, our target variable-lung cancer mortality is continuous. In this case, we run regression modeling. From all the 13 models DataRobot provides, we arrange them based on holdout value, since the first model has a suspicious note, we decided to choose the second model and it is a Light Gradient Boosted Trees Regressor with Early Stopping.

Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
RandomForest Regressor Tree-based Algorithm Preprocessing v1 M43 BP79 80.02% RECOMMENDED FOR DEPLOYMENT	DR Reduced Features M9 100.0% +	12.7969 *	12.1358 *	11.9262 *
Light Gradient Boosted Trees Regressor with Early Stopping Tree-based Algorithm Preprocessing v1 M12 BP82 FAST & ACCURATE	Informative Features 63.99% +	13.1012	12.3080	11.7390
RandomForest Regressor Tree-based Algorithm Preprocessing v1 M41 BP79	DR Reduced Features M9 80.02% +	13.0375 *	12.2373 *	11.7456
AVG Blender Average Blender M45 M34+9 MOST ACCURATE	Multiple Feature Lists 63.99% +	13.1396	12.2739	11.8263

Light Gradient Boosted Trees Re...

63.99% Sample Size | Informative Features

BP82 M12

Change Model

RMSE (Validation):
13.1012

RMSE (Cross Validation):
12.3080

RMSE (Holdout):
11.7390

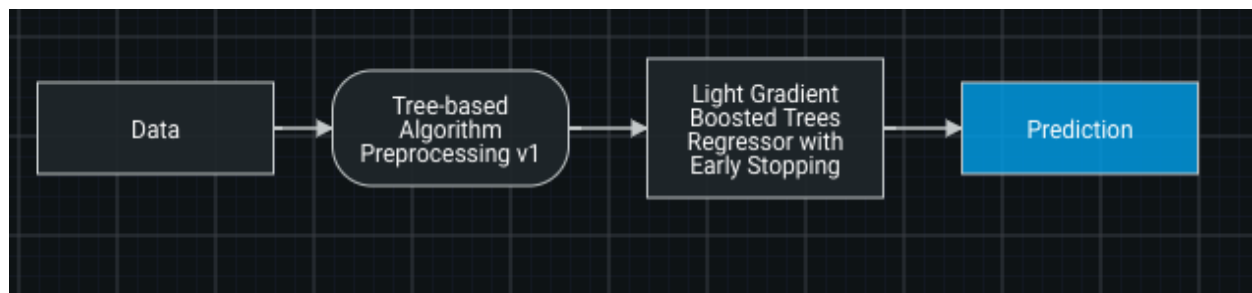
Gini Norm (Validation):
0.6825

Gini Norm (Cross Validation):
0.6987

Gini Norm (Holdout):
0.7325

Prediction Time:
515.64 ms

The model blueprint and overview shows that the model first uses a tree based algorithm, then it uses the LightGBM implementation of Gradient Boosted Trees. It uses least squares loss by default. LightGBM is a gradient boosting framework designed to be distributed and efficient with the following advantages: Faster training speed and higher efficiency, Lower memory usage, Better accuracy, Parallel learning supported, Capable of handling large-scale data.

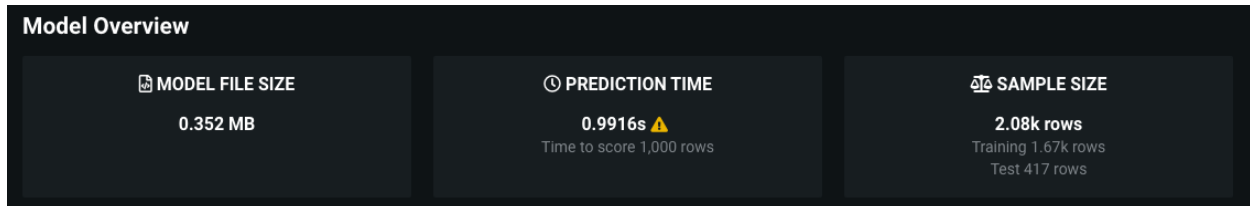


They are very similar in concept to random forests, in that they fit individual decision trees to random re-samples of the input data, where each tree sees a bootstrap sample of the rows of a the dataset and N arbitrarily chosen columns where N is a configural parameter of the model.

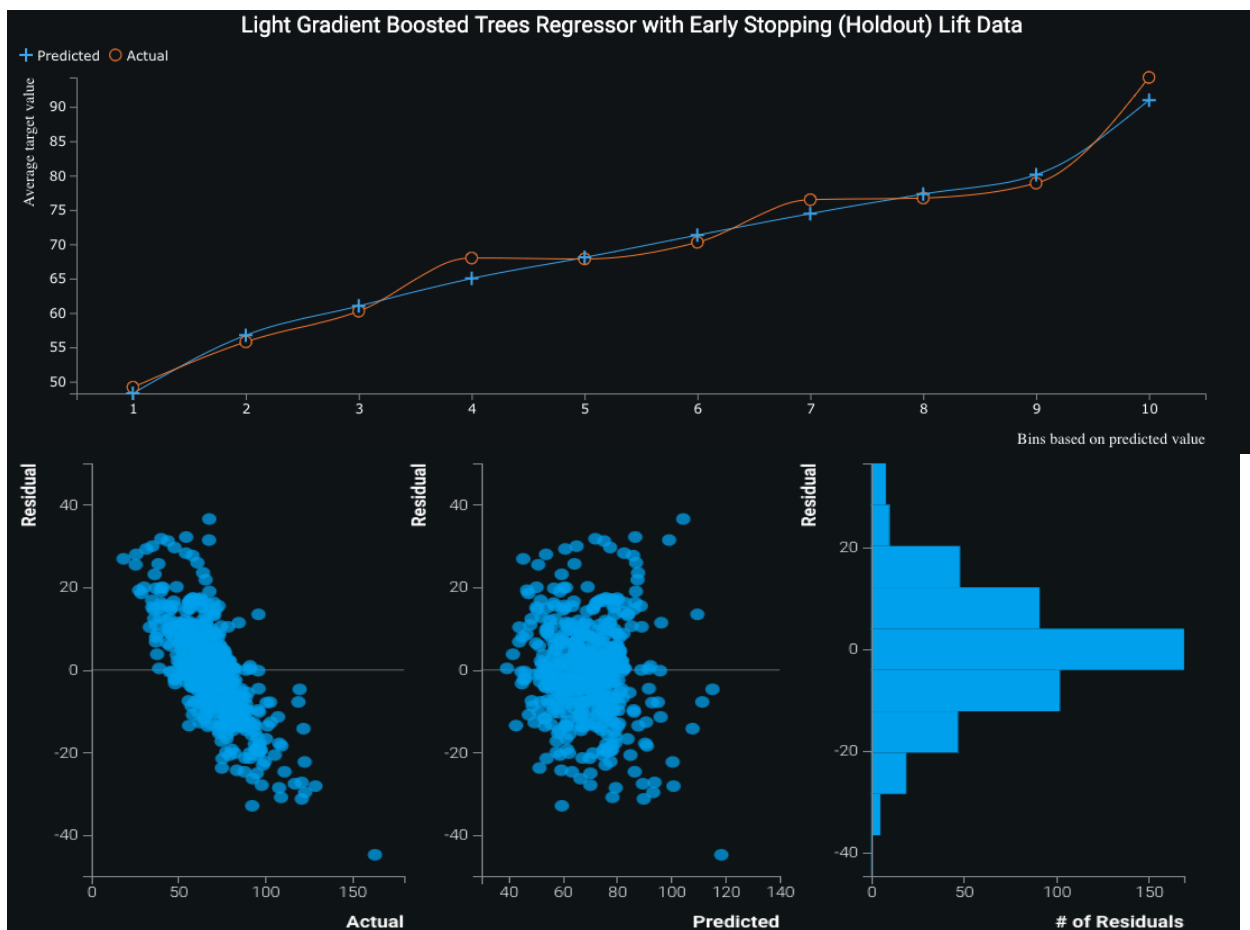
GBMs differ from random forests in a single major aspect: rather than fitting the trees in parallel, the GBM fits each successive tree to the residual errors from all the previous trees combined. This is advantageous, as the model focuses each iteration on the examples that are most difficult to predict (and therefore most useful to get correct).

And early stopping is used to determine the best number of trees where overfitting begins. In this manner GBMs are usually capable of squeezing every last bit of information out of the training set and producing the model with the highest possible accuracy.

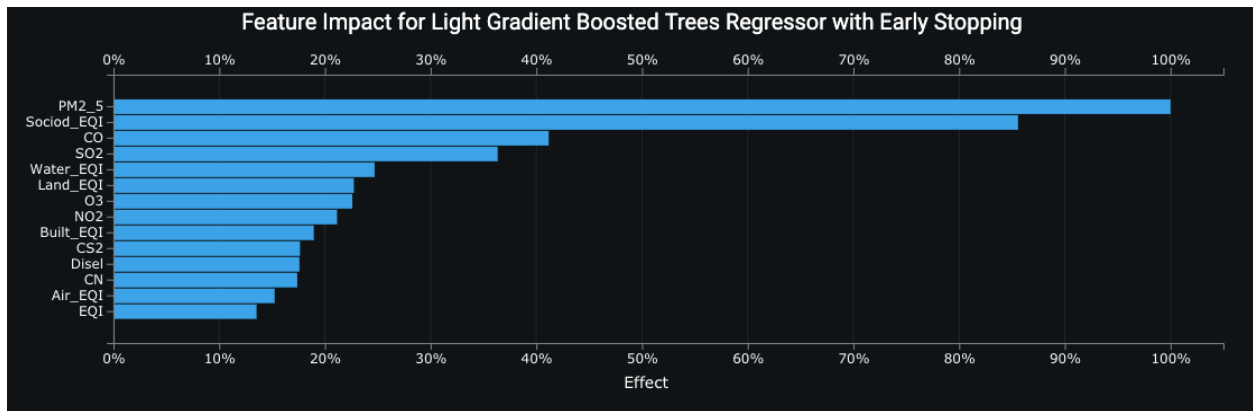
From the model overview, we know that the prediction time is fast and only takes 0.9916 seconds, the model uses 2080 rows as training data while the rest 416 rows as test data.



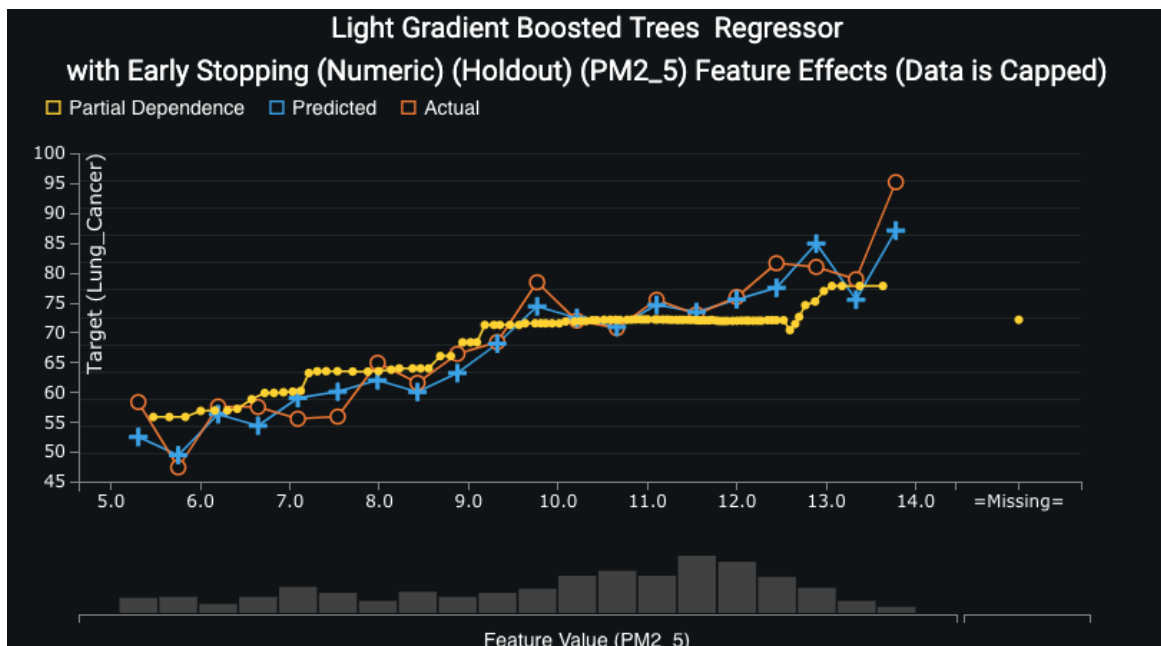
In terms of the prediction performance, from the holdout lift data plot, we see the two lines have great fitness. And the residual plot shows that the histogram distribution is bell shaped. Overall, the model has good prediction performance.



We then take a closer look at the model, the feature impact plot shows the feature importance levels in building the model, PM 2.5 has 100% impact in the modeling, which fits our assumption. Sociod_EQI has almost 86% impact and other features all have less than 45% impact on the model. Such results show that the two variables may be more useful in modeling.



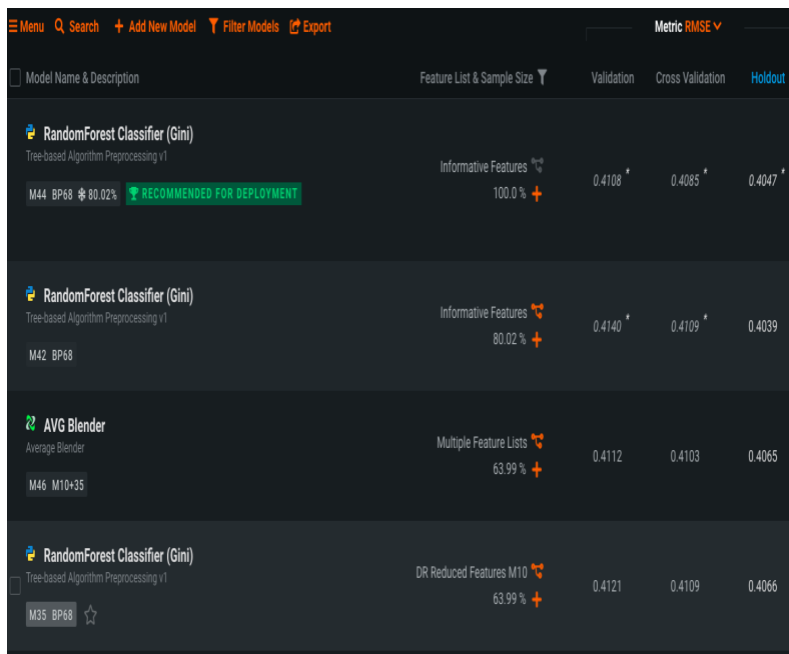
Moreover, when it comes to the relationship between features and the target variable, (a partial dependence plot illustrates the marginal effect that a given feature has on the predicted value), we noticed that PM 2.5 has a positive relationship with lung cancer mortality. It makes sense, because higher PM 2.5 may cause the increase of lung cancer mortality, that fits our results of other models.



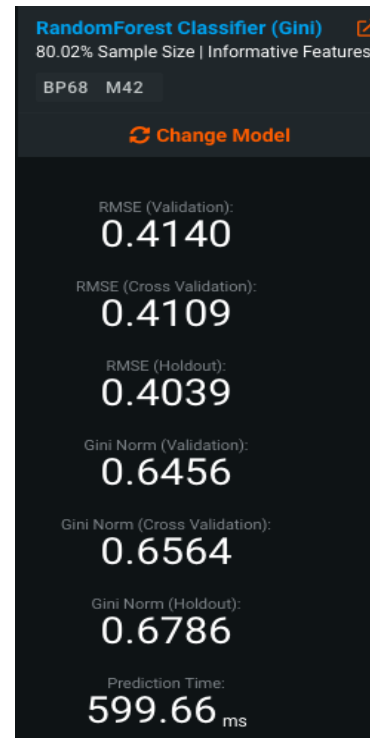
6.2 Random Forest Classifier

We then upload the transformed data with lung cancer mortality as a categorical variable to DataRobot. In this case, we run classification modeling. From all the 14 models DataRobot provides, we arrange them based on holdout value, since the first model has a suspicious note, we decided to choose the second model and it is a random forest classifier..

The holdout RMSE of the model is 0.4039, Gini Norm is 0.6786.



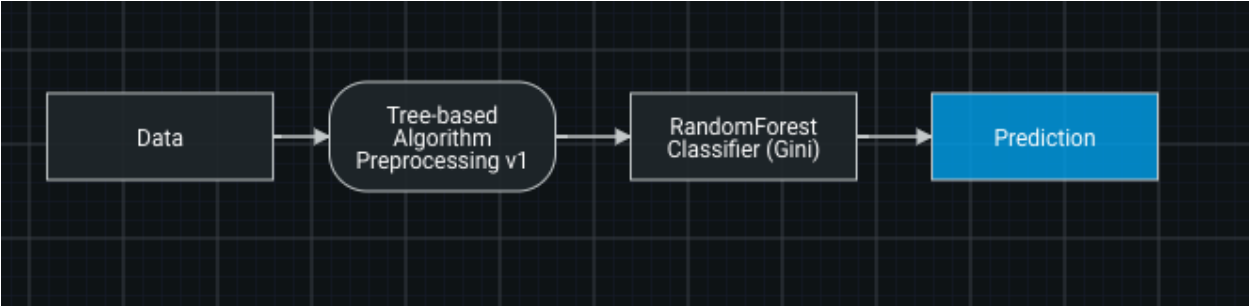
Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
RandomForest Classifier (Gini) Tree-based Algorithm Preprocessing v1 M44 BP68 80.02% RECOMMENDED FOR DEPLOYMENT	Informative Features 100.0% +	0.4108 *	0.4085 *	0.4047 *
RandomForest Classifier (Gini) Tree-based Algorithm Preprocessing v1 M42 BP68	Informative Features 80.02% +	0.4140 *	0.4109 *	0.4039
AVG Blender Average Blender M46 M10+35	Multiple Feature Lists 63.99% +	0.4112	0.4103	0.4065
RandomForest Classifier (Gini) Tree-based Algorithm Preprocessing v1 M35 BP68 ☆	DR Reduced Features M10 63.99% +	0.4121	0.4109	0.4066



RandomForest Classifier (Gini)	
80.02% Sample Size Informative Features	
BP68 M42	
RMSE (Validation):	0.4140
RMSE (Cross Validation):	0.4109
RMSE (Holdout):	0.4039
Gini Norm (Validation):	0.6456
Gini Norm (Cross Validation):	0.6564
Gini Norm (Holdout):	0.6786
Prediction Time:	599.66 ms

The model blueprint and overview shows that the model first uses a tree based algorithm then it uses a random forest classifier to make the prediction. Random forests are an ensemble method where hundreds (or thousands) of individual decision trees are fit to bootstrap re-samples of the original dataset. Ensembling many re-sampled decision trees serves to reduce their variance, producing more stable estimators that generalize well out-of-sample. Random forests are extremely hard to over-fit, are very accurate, generalize well, and require little tuning, all of which are desirable properties in a predictive algorithm.

From the model overview, we know that the prediction time is 2.2123 seconds, the model uses 2080 rows as training data while the rest 417 rows as test data.



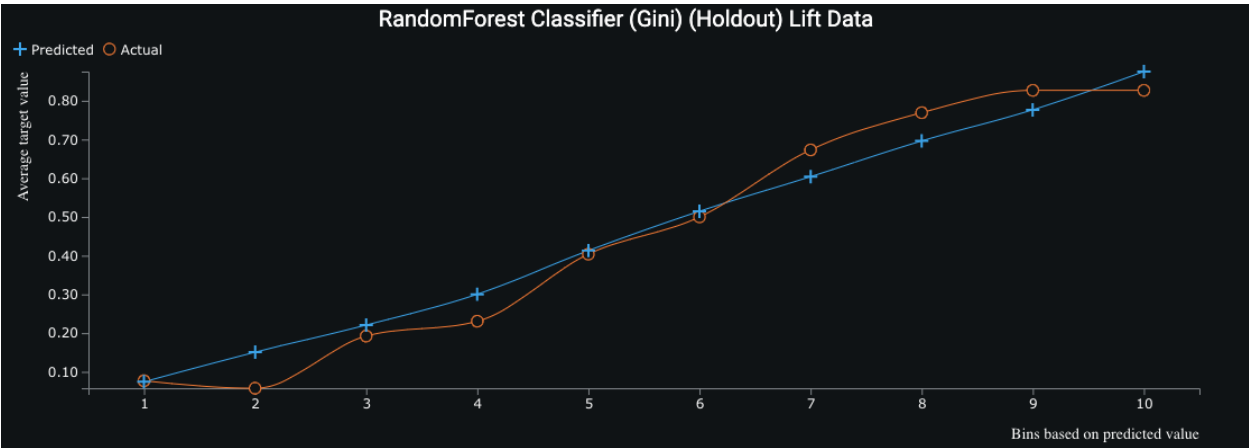
Model Overview

MODEL FILE SIZE
14.193 MB

PREDICTION TIME
2.2123s ⚠️
Time to score 1,000 rows

SAMPLE SIZE
2.08k rows
Training 1.67k rows
Test 417 rows

In terms of the prediction performance, from the holdout lift data plot, we see the two lines have similar overall trends.

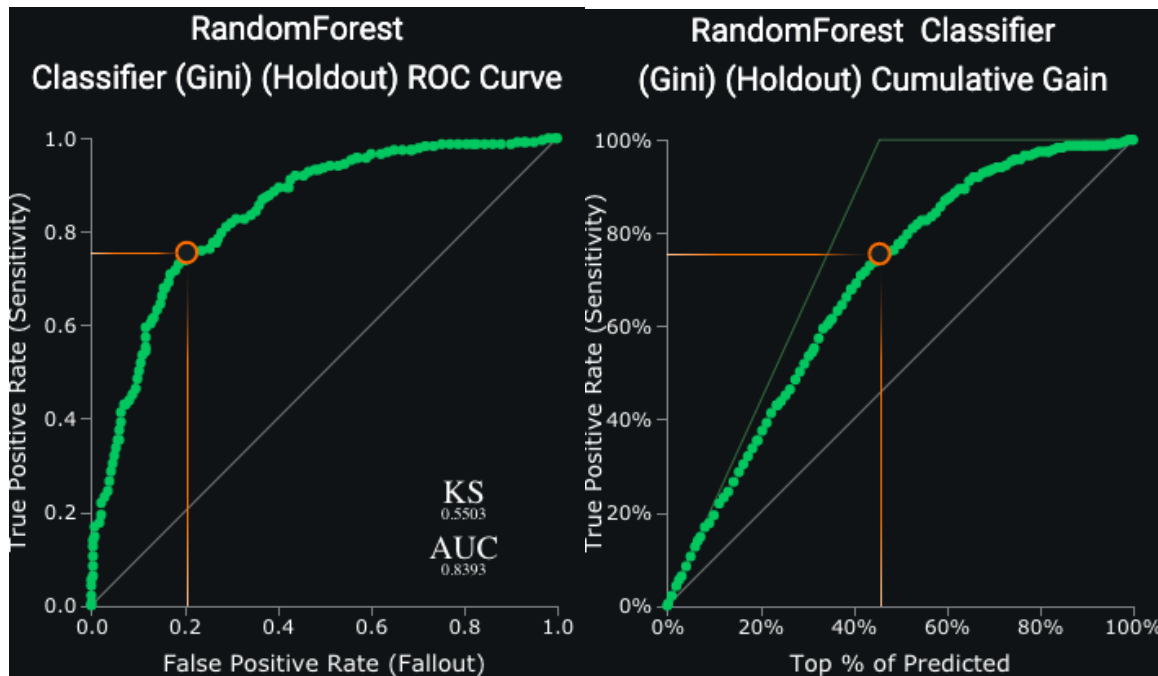


Then we check the model evaluation part, the confusion matrix shows that the accuracy is 0.7769, sensitivity is 0.7553, specificity is 0.7951 and the F1 score is 0.7553. Compared with our own classification models, this one performs the best.

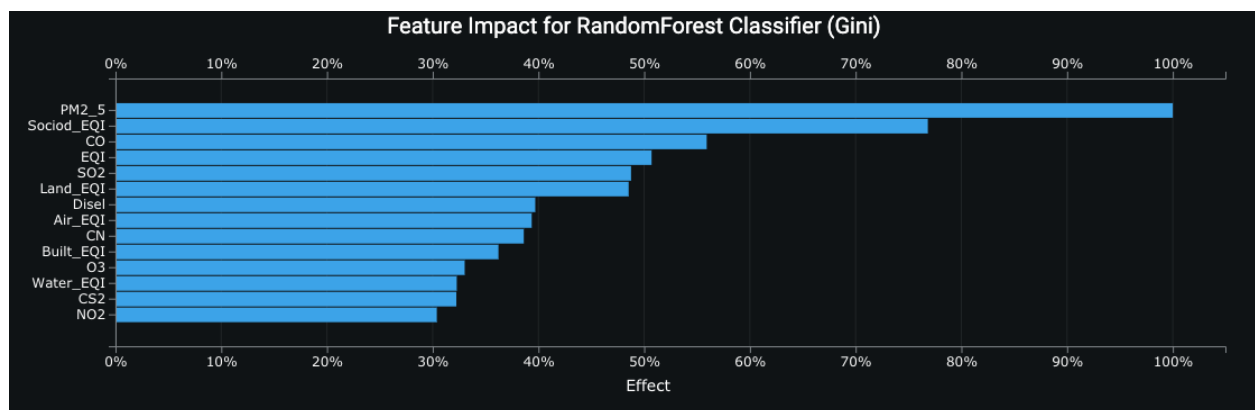
Selection Summary Export								Predicted		
F1 Score	True Positive Rate (Sensitivity)	False Positive Rate (Fallout)	True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Negative Predictive Value	Accuracy	Matthews Correlation Coefficient	Actual	-	+
									225 (TN)	58 (FP)
									58 (FN)	179 (TP)
0.7553	0.7553	0.2049	0.7951	0.7553	0.7951	0.7769	0.5503		283	237
									283	520

From the ROC curve and the cumulative chart, we noticed that the AUC is 0.8393, which is a high number indicating good measure of separability.

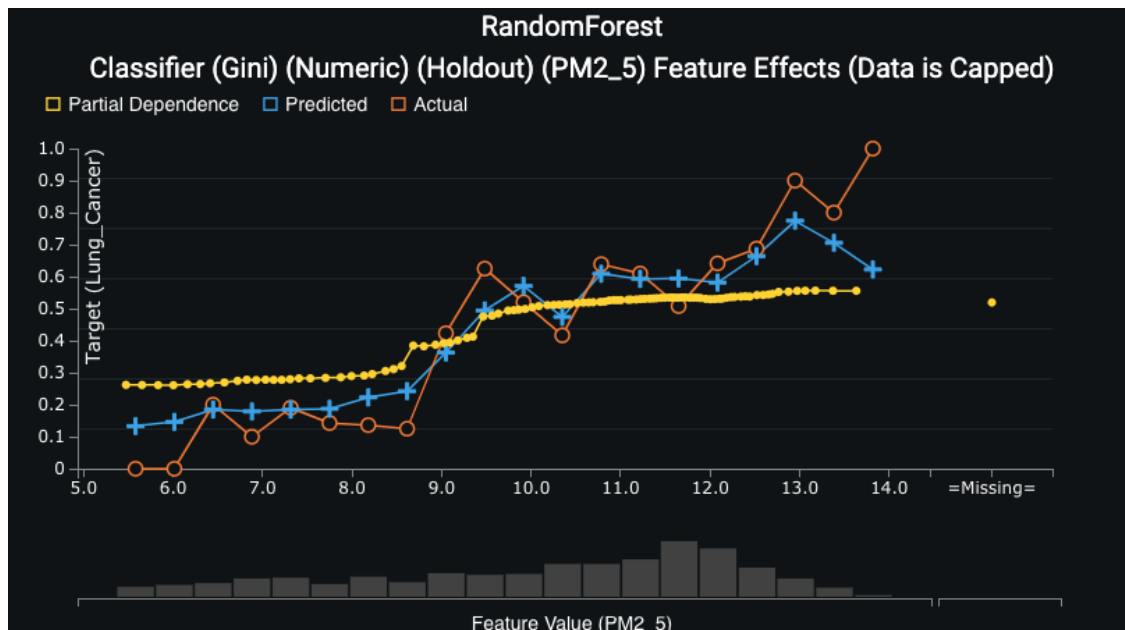
The cumulative gains curve is an evaluation curve that assesses the performance of the model and compares the results with the random pick. From our plot, it shows if we can target the 45% of the observations, the model will ensure that almost 80% of the true positive rate.



We then take a closer look at the model, the feature impact plot shows the feature importance levels in building the model, PM 2.5 has 100% impact in the modeling, which fits our assumption. Sociod_EQI has almost 77% impact, the third is CO. Other features all have less than 50% impact on the model.



Moreover, when it comes to the relationship between features and the target variable, we noticed that PM 2.5 has a positive relationship with lung cancer mortality. It makes sense, because higher PM 2.5 may cause the increase of lung cancer mortality.



6.3 Model Comparisons

We finally compared our best hand-crafted models with DataRobot ones, the results show that DataRobot models perform better both in regression and classification models. It may be because DataRobot uses random forest concepts which are more efficient and accurate.

6.3.1 Regression Models

	Multi Regression with Clusters	Light GBM
R ²	39.56%	54.77%
RMSE	13.4971	11.7390

6.3.2 Classification Models

	KNN	Random Forest
Accuracy	74.42%	77.69%
Sensitivity	71.10%	75.53%
Specificity	77.21%	79.51%
F1 Score	71.70%	75.53%

Part VII. Conclusions and Reflections

7.1 Conclusion and Insights

In this project, we aim to explore the effect of environment factors on lung cancer mortality rate.

Data Cleaning and Preparation

We firstly performed data wrangling using Trifacta Wrangler. Then we did visualization, description of each important variable and descriptive statistics to have a basic understanding of our data and variables and look at the scatter plot matrix to decrease multicollinearity.

For the Regression Model with Numeric Target Variable, we looked at the residual vs fitted value plot and QQ plot to detect outliers and delete them.

For the Classifications Model with Categorical Target Variable, we firstly transform and standardize variables, and then look at residual plot to detect outliers and delete them.

To check the overfitting issue in the models, we compared the performances of validation/testing data and train data.

Best Models

For the Regression Model with Numeric Target Variable, based on the R-squared and adjusted R-squared, RMSE, F-statistic, AIC and BIC, the new model including cluster variables performed better, and including cluster variables and interactive terms can improve the model.

For Classifications Model with Categorical Target Variable, the KNN performs the best based on the Accuracy, Sensitivity, Specificity and F1 Score.

DataRobot Models

From 13 regression models that DataRobot provides, we chose Light Gradient Boosted Trees Regressor with Early Stopping as our best model. PM 2.5 has 100% impact in the modeling and Sociod_EQI has the second greatest impact on the model.

For 14 classification models after transformation, we chose the random forest classifier model as our best one. PM 2.5 has 100% impact in the modeling, and Sociod_EQI has the second greatest impact.

Conclusion

Based on our results, we noticed that air quality is a vital determinant of lung cancer mortality. The PM 2.5 level shows a significantly high positive relationship with mortality rate and is the leading factor of lung cancer mortality. And models from DataRobot perform better than Hand-Crafted Models.

7.2 Reflections

7.2.1 Reflection of Standardization

When we do standardization, we found that some variables are highly skewed, we need to make some changes to these variables based on the direction and extent of skew.

7.2.2 Reflection of Regression Model

From our process of improving the regression model, we found adding interactive terms is a good way to improve model performance significantly. Also, feature extraction such as adding clustering labels to the original model can also help improve model performance in some way.

For adding cluster labels part, at first, we used all the variables including target variables to make clustering. However, our teacher assistant said there is something wrong. Because the target variable is what our regression model wants to explain, its features can't be used in the clustering process. We came to realize the logic behind it and made changes to our model in the final report.

7.2.3 Reflection of Using Data Robot

Additionally, the Random forest model performs the best, which fits our expectation since this model runs fast, always has a relatively higher accuracy and is widely used. We also found that the process part of DataRobot is helpful with modeling and can be used to select important variables efficiently.

DataRobot can run multiple models at the same time and present the results directly, which greatly decreases the time and difficulty of modeling. So the data analysts should pay more attention to data collecting and preparation, parameter choosing and insights from analysis and data-informed decisions.