Group Final Report

1. Introduction. An overview of the project and an outline of the report.

As group pollution solvers, we tried to tackle the issue of the effects of energy consumption by specific types of energy on the deaths caused by pollution. While we were building up the project the first obstacle that we had to overcome was to find a problem to solve. So for our first meeting we came up with topic proposals of our own. After non-heated discussions we decided to ramble with a topic that interested us all, which was the topic of energy consumption and pollution. One question that we had regarding this topic was the fact that we were not sure if we were allowed to use multiple data sets. We talked with the professor and the TA which ended with us, using multiple datasets.

Outline of the project is following:

- Picking up the 5 datasets(**data_health, data_clean, data_pollution, data_GDP, data_energy)**
- Preprocessing, cleaning up the data
- Model Building
- Displaying the Linear Regression findings via GUI(mainly using tinker)

2. Description of the data set.

- **data_health: 2.12_Health_systems.csv**

Health_exp_per_capita_USD_2016 : Numerical,Continuous, health expendation of each location

- **data_clean: cleanFuelAndTech.csv**

First Tooltip: Numerical,Continuous, percentage of clean fuel used in house

- **data_pollution: death-rates-from-air-pollution.csv**

Indoor air pollution (deaths per 100,000):Numerical,Continuous, death caused by indoor air pollution of each country

Outdoor particulate matter (deaths per 100,000):Numerical,Continuous, death caused by outdoor air pollution of each country

- **data_GDP:Country_wise_GDP_from_1994_to_2017.csv**

GDP per capita (in USD):Numerical,Continuous, GDP of each country

- **data_energy: Percentage_of_Energy_Consumption_by_Country.csv**

Oil Consumption – EJ: Numerical,Continuous, oil consumption of each country

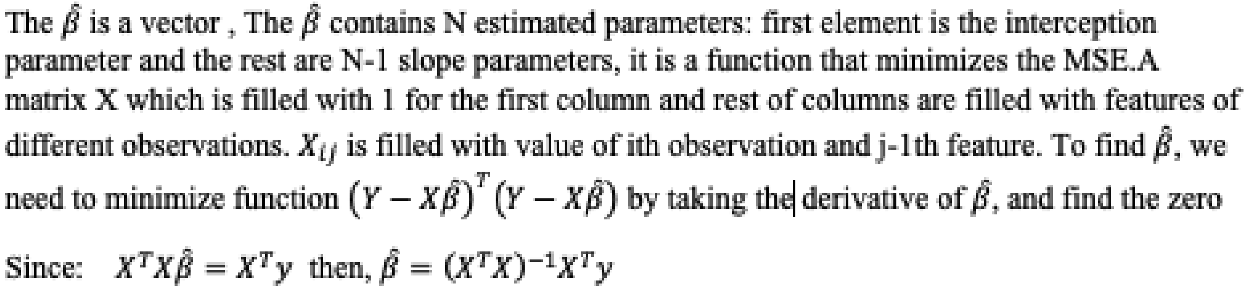
Coal Consumption - EJ :Numerical,Continuous, coal consumption of each country

3. Description of the data mining and learning or cleaning algorithm or other algorithms that you used. Provide some background information on the development of the algorithm and include necessary equations and figures.

PollutionSolvers used the linear regression model to figure out the relationship between fatality rate, fuel usage and some other potential features. The mathematical interpretation of linear regression with multiple features input could be demonstrated by the linear algebra.

Regression function:

$$\hat{Y} = X\hat{\beta}$$



The $\hat{\beta}$ is a vector , The $\hat{\beta}$ contains N estimated parameters: first element is the interception parameter and the rest are N-1 slope parameters, it is a function that minimizes the MSE.A matrix X which is filled with 1 for the first column and rest of columns are filled with features of different observations. X_{ij} is filled with value of ith observation and j-1th feature. To find $\hat{\beta}$, we need to minimize function $\left(Y - X\hat{\beta}\right)^T\left(Y - X\hat{\beta}\right)$ by taking the derivative of $\hat{\beta}$, and find the zero

Since: $X^T X\hat{\beta} = X^T y$ then, $\hat{\beta} = (X^T X)^{-1} X^T y$

The data set features may not show the linear correlation, it could require data transformation and feature generation.For some data, the predicted feature and explanatory features do not show linear relationship. Transforming the predicted feature with log, and the model will be an exponential decay model.

$$\hat{Y} = exp\left(X\hat{\beta}\right)$$

The generated explanatory features could be added into model by adding a feature which is a function of the original feature. For instance, square the original feature and add as a new column of feature for input X. Sometimes the previous feature could have interaction, a term of interaction feature could be added into input X by generating a feature from multiplying or dividing previous features.

It is efficient to use residuals to analyze the performance of the regression; the assumption of ideal linear regression is that the residual for all of the different feature values should have the same variance and mean of 0. For regression with multiple explanatory features, plot regression line with observation is hard. Analysis of residuals with each feature is more feasible.

Residual: $\hat{e} = Y - \hat{Y}$ residual is the real value - predicted value

R-squared Metrics

The R-squared portion of predicting variable's variation in observations could be explained by the model.

Since $MSE = \hat{e}^T\hat{e}/n$ and R-squared= $1 - (\frac{MSE}{Var(Y)})$

Hypothesis testing: F-test

To make selection of a model, F-test is a method to compare two similar models. If model 1 contains all of the explanatory features of model 2, and model 1 contains some more features. Then model 1 is the full model and model 2 is the reduced model. A F-test could identify whether the full model has an advantage on prediction.

Confidence Interval

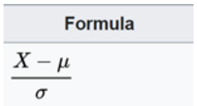
Using the regression model by inputting features, we can get an estimated mean for the predicted feature with specific explanatory features. A confidence Interval can generate a lower bound and an upper bound for this estimated mean.

$$\hat{Y} \pm \sqrt{(X_0(Z^TZ)^{-1}_{X_0})}\hat{\sigma}\Phi^{-1}_{(\alpha)}$$

4. Experimental setup. Describe how you are going to use the data to clean and preprocess. Explain how you will implement the data mining technique in the chosen software and how you will judge the performance. Write a complete report with theoretical description and verify this mathematical concepts with applying it with actual data. Provide enough

information about the codes tat you have written. Write your codes in separate subroutines and call the functions if needed?. Explain each subroutine.

For the pre-processing section, PollutionSolvers used some basic measures like data normalization, missing values imputation, data integration, and Noise identification. Data normalization is the process of minimizing redundancy from relation or set of relation. Selecting the specific datasets or data columns, using the function preprocessing from sklearn:



Formula
$\frac{X - \mu}{\sigma}$

: the formula for normalization of StandardScaler

The purpose for missing values imputation is to fill up the variables that contain missing values. Data integration is an important step for preprocessing, the purpose of it is to merge multiple datasets to let it be read by the model. Finding Noise is also a necessary step. Noise is part of data which is unreasonable and should be detected during preprocessing. When finding noises out, PollutionSolvers can decide if PollutionSolvers are keeping them or dropping them out from the dataset. Finally for cleaning datasets, and dropping the columns and data which is unnecessary for the model.

For the modeling section, during the discussion PollutionSolvers considered many models, and after evaluating the datasets the team decided to use linear regression as the basis for the model. The first and most basic one is that a linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

5. Results. Describe the results of your experiments, using figures and tables wherever possible. Include all results (including all figures and tables) in the main body of the report, not in appendices. Provide an explanation of each figure and table that you include. Your discussions in this section will be the most important part of the report.

Table 1. The best model found

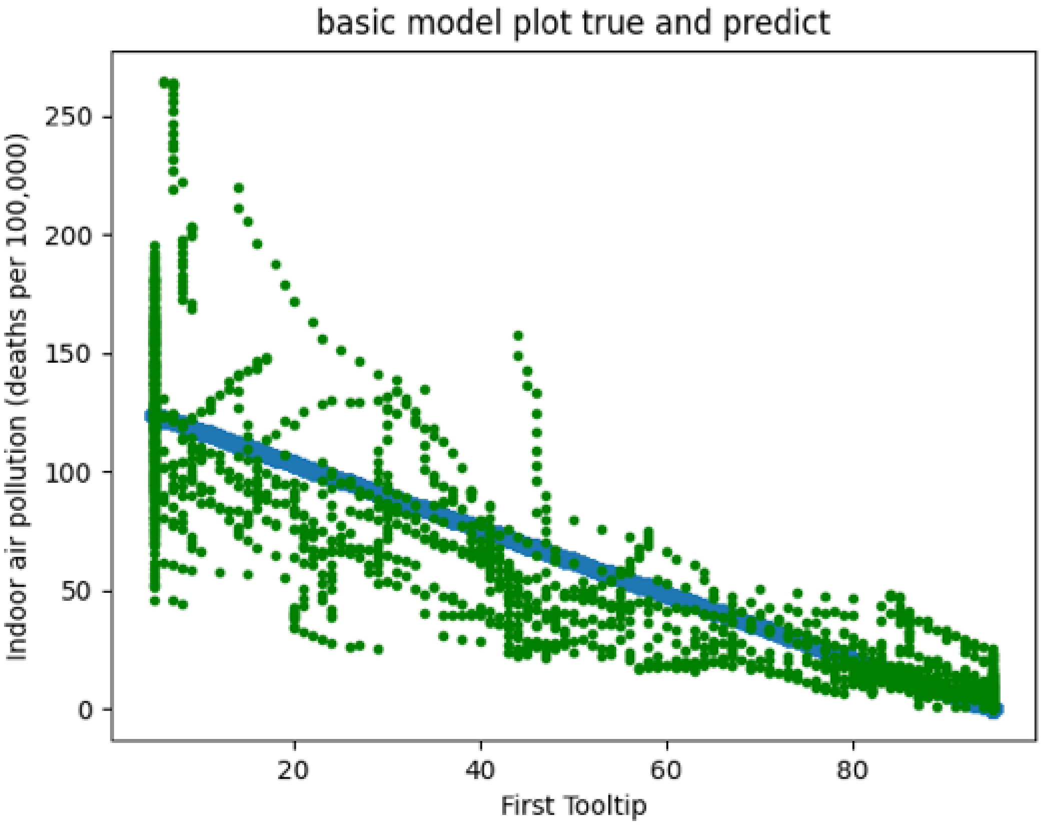
	R-squared	Variables
Indoor log model	0.814	Clean fuel access rate in the population(%), Gdp per Capita(USD)
Outdoor log model	0.528	GDP/fuel(USD/kwh), GDP per capita

During the training process, we received the previous EDA results to find potential important features. TeamPollutionSolvers generated the first basic model and plotted the residual vs the portion of population accessing clean fuel and technology.

The result that was found did not show the constant variance. From the plot, it could be seen that the residual increased with the percentage of people accessing clean fuel and technology. It could be inferred that the team needed to do log transformation to the fatality rate. That will infer that the model with fatality rate and percentage of population access to the clean fuel and technology could be an exponential decay relationship.

Per this situation, the random noise term will have a mean of 0 and constant noise. And the noise could be a scale noise with exponential, that means that the residual is proportional to the estimated mean. For both indoor and outdoor situations, the log transformation is required.

Fig2. The plot of basic model predict and real observations



To explore “Better Model”, we also did some preprocessing parts, including feature generating and data transformation. We first wanted to use the medical expenditure as a second variable, however, the observation of medical expenditure is limited. In the 2016 case study, we found out that the general relationship is that the medical expenditures and GDP are linearly proportioned and the intercept is nearly zero. The R squared of this model is 0.93. After that we used the GDP and clean fuel usage rate in the main indoor model to apply “estimate”. After removing the extreme conditions that the usage rate is over 94. The model showed well with r-squared and residuals.

The best model that PollutionSolvers found is the log model with explanatory variables: percentage of population access clean fuel and GDP per capita, it shows R-squared over 0.82 and the residual plot shows the same variance and mean of zero.

Fig3: Plot the residual, x axis is the percentage of population access clean fuel

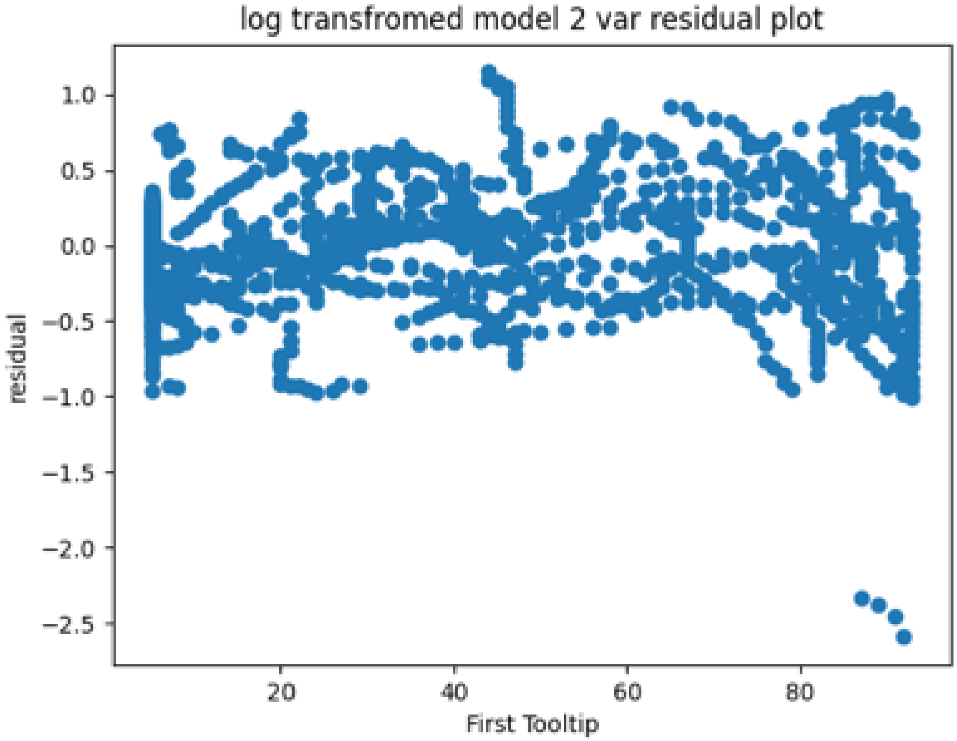


Fig4: Plot the estimated result and true result, x axis is the percentage of population access clean fuel

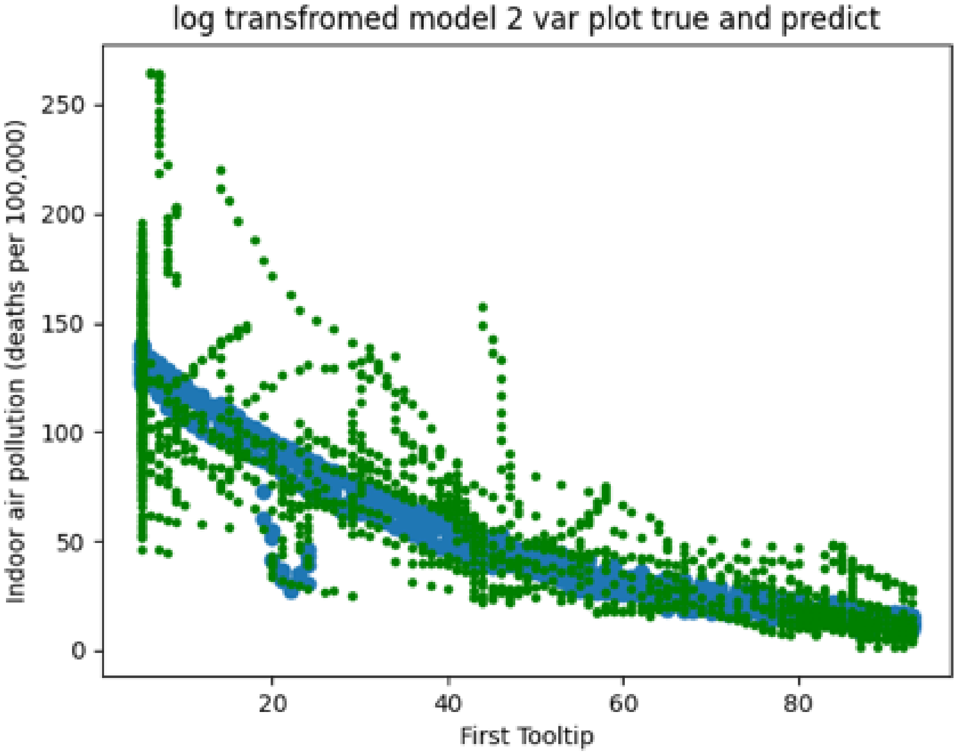
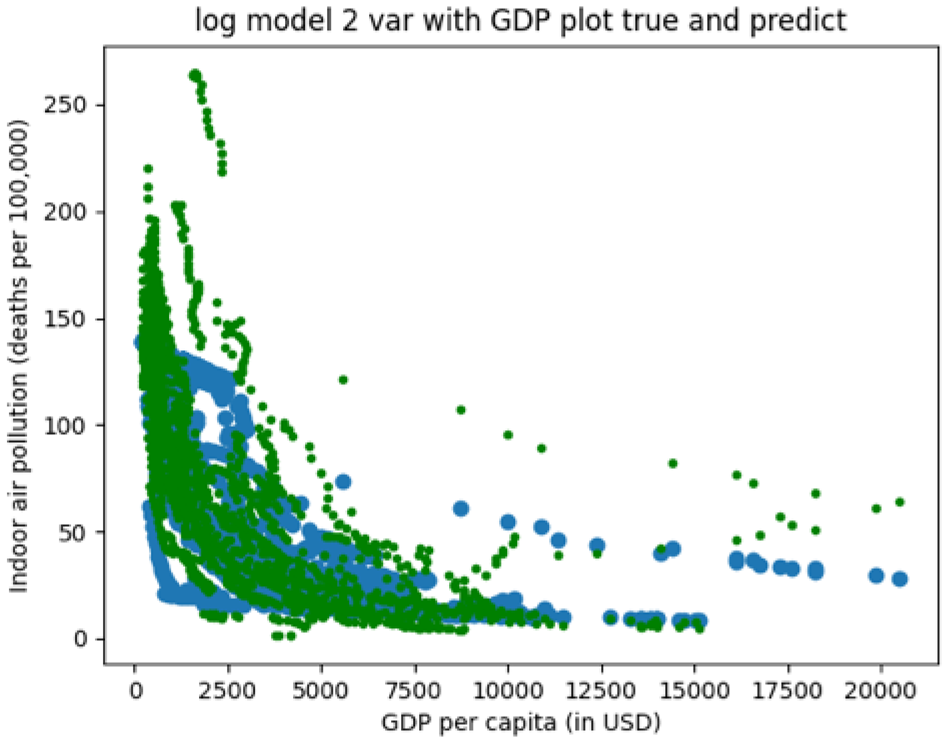


Fig5: Plot the estimated result and true result, x axis is the GDP per capita



The team needed to use the pandas map function to square some variables and generate new features by applying a function with existing features when tried to deal with the outdoor fatality rate. Since all of the features do not work well with prediction of the outdoor fatality rate. The regression model and log regression model has R-squared less than 0.1. The generated feature that average usage energy from coal per capita and The generated feature that average usage energy from coal per capita could not solve the problem. However, generated feature, GDP/fuel in the unit of USD/kwh, with feature GDP per capita make the log model with the R-squared of 0.53

Fig 6: plot the estimated result and real observations of outdoor fatality rate, x axis is GDP per capita

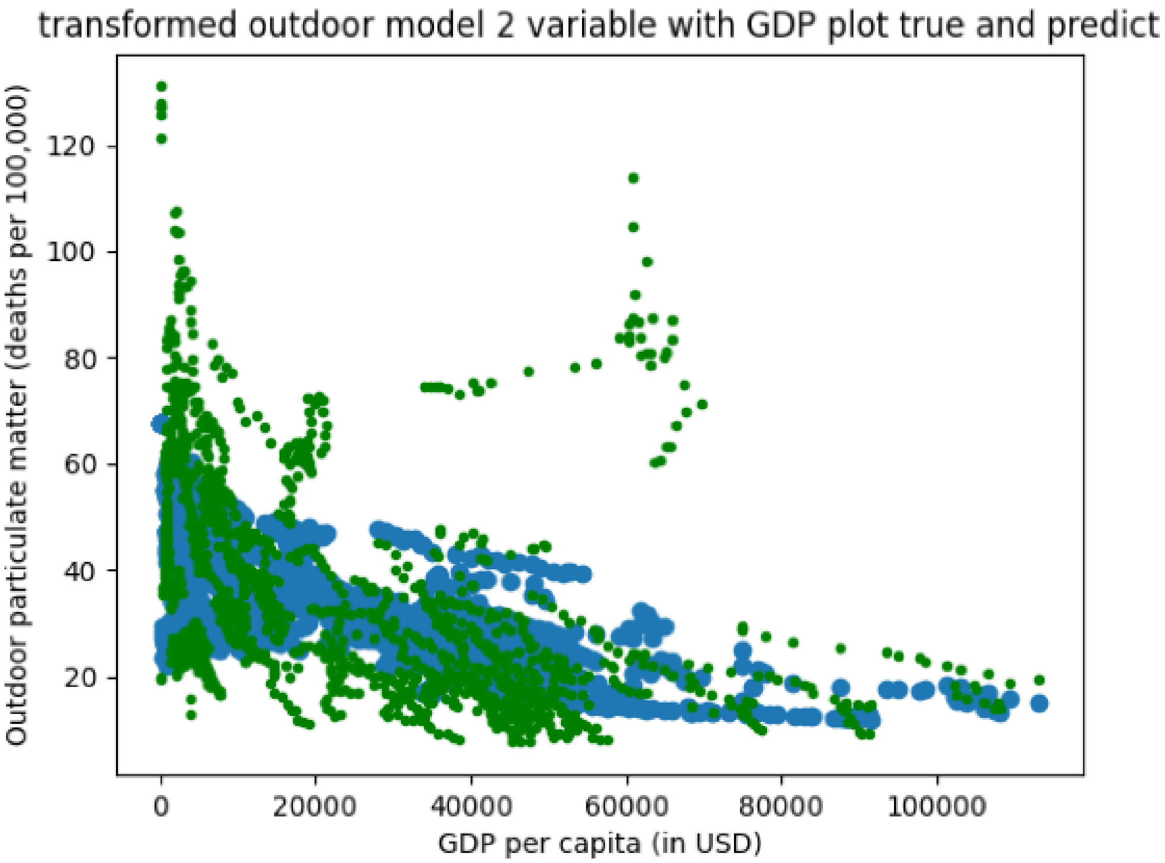
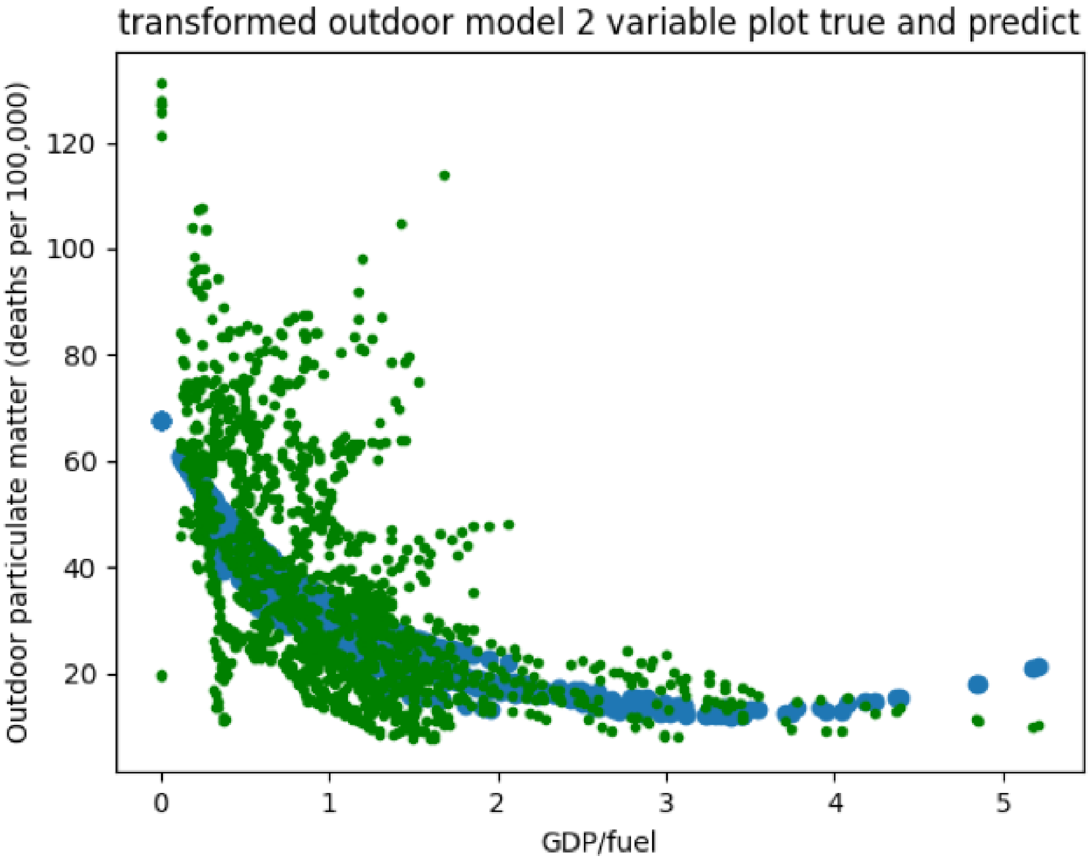


Fig 7: plot the estimated result and real observations of outdoor fatality rate, x axis is GDP/fuel



After we finished the model, we also wrote an estimator class to store the trained model. It could predict the indoor and outdoor fatality rate by entering the percentage of population accessing clean fuel usage, GDP/fuel ratio and GDP per capita. It could also calculate the confidence interval for both estimations. It is imported in the GUI to estimate results.

6. Summary and conclusions. Summarize the results you obtained, explain what you have learned, and suggest improvements that could be made in the future.

From the model of indoor and outdoor, the team found a strong correlation between fatality caused by indoor pollution and percentage of population accessing clean fuel usage. The usage of clean fuel at home can significantly decrease the fatality rate. For the outdoor situation, the model is not very accurate since the R-square is around 0.5. It still points out that the energy saving industries that produce more value with less energy could be good for people's health.

Using linear regression to deal with data with a lot of features could be a challenge, since it requires me to inspect the feature in detail. Doing data transform and feature generation are also important techniques in modeling. The linear regression has constraints on linearity; it cannot predict accurately without suitable preprocessing. The model still needs some improvement since it predict some value with huge discrepancies. The dataset from Kaggle still has the limitation that it could not divide the regions into more subregions. It could not provide enough information to improve more accuracy. Another problem is the observation of some specific features is limited. The linear regression model's prediction in these intervals has large discrepancy. For instance with high GDP per capita and GDP/fuel ratio, the estimation is not accurate. I need to apply some new model and preprocessing techniques to increase the accuracy of extreme cases.

- In the case of indoor pollution, mortality is expressed as a function of exponential decay until the consumption rate of clean fuel reaches very high values. Value> 94%.
- In the case of outdoor pollution, mortality has something to do with GDP / fuel.

According to the results that the team found, PollutionSolvers suggest the following improvements to the various countries.

- For countries with deaths due to the high indoor pollution, clean Fuel should be used more. Some examples of activities that create indoor pollution are cooking, warming up.
- For countries with deaths due to high outdoor pollution, energy density should be changed for activities that involve high energy consumption such as industrial production and transportation.

One thing that the team found out to be problematic was the fact that graphs were not enough for display as referenced by the professor. The team took note on that, and will try to do a better job in their next endeavors.

7. References.

Code:

Pandas Website: https://pandas.pydata.org

Sklearn Website: https://scikit-learn.org/stable/

Mathematical Background:

Linear Model With R, by Julian J. Faraway

8. **A separate appendix should contain documented computer listings.**