THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

A Detailed Look in to Energy Consumption and Pollution Related Deaths

Presentation by PollutionSolvers - Zhongyang Hu, Hsueh-Yi Lu, Ali Can Mutlu,

Data Set Information

1. Data Sets taken from Kaggle
2. Some Main features: GDP per Capita, Death caused by Indoor Air Pollution per 100,000 people , Death caused by Outdoor Particulate Matter per 100,000 people, Medical Expenditure, Clean Fuel Usage Rate in the Population, Energy Consumption from fuel….
3. Data Date: 1974-2017
4. #Observations: +2000

Dataset Summary

data_health: **2.12_Health_systems.csv**

Source: Kaggle

Variables:

Country_Region: Qualitative,Categorical

World_Bank_Name : Qualitative,Categorical

Health_exp_pct_GDP_2016 : Numerical,Continuous

Health_exp_public_pct_2016 : Numerical,Continuous

Health_exp_out_of_pocket_pct_2016 : Numerical,Continuous

Health_exp_per_capita_USD_2016 : Numerical,Continuous

Completeness_of_birth_reg_2009-18: Numerical,Continuous

per_capita_exp_PPP_2016: Numerical,Continuous

External_health_exp_pct_2016: Numerical,Continuous

Physicians_per_1000_2009-18: Numerical,Continuous

Nurse_midwife_per_1000_2009-18: Numerical,Continuous

Specialist_surgical_per_1000_2008-18: Numerical,Continuous

Dataset Summary

data_clean: **cleanFuelAndTech.csv**

Source: Kaggle

Variables:

Location: Qualitative,Categorical

Indicator: Qualitative,Categorical

Period: Numerical, Categorical

First Tooltip: Numerical,Continuous

Dataset Summary

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

Source: Kaggle

Variables:

Entity: Qualitative,Categorical

Code: Qualitative,Categorical

Year: Numerical, Categorical

Air pollution (total) (deaths per 100,000): Numerical,Continuous

Indoor air pollution (deaths per 100,000):Numerical,Continuous

Outdoor particulate matter (deaths per 100,000):Numerical,Continuous

Outdoor ozone pollution (deaths per 100,000):Numerical,Continuous

Dataset Summary

data_GDP: **Country_wise_GDP_from_1994_to_2017.csv**

Source: Kaggle

Variables:

Country: Qualitative,Categorical

Year: Numerical, Categorical

GDP (in USD): Numerical,Continuous

GDP change (%):Numerical,Continuous

GDP per capita (in USD):Numerical,Continuous

Population: Numerical,Continuous

Pop. change (%): Numerical,Continuous

GDP Real (in USD):Numerical,Continuous

Dataset Summary

data_energy: **Percentage_of_Energy_Consumption_by_Country.csv**

Source: Kaggle

Variables:

Country: Qualitative,Categorical

Code: Qualitative,Categorical

Year: Numerical, Categorical

Coal Consumption - EJ :Numerical,Continuous

Gas Consumption - EJ Numerical,Continuous

Geo Biomass Other - TWh :Numerical,Continuous

Hydro Generation – TWh:Numerical,Continuous

Nuclear Generation – TWh: Numerical,Continuous

Solar Generation – TWh: Numerical,Continuous

Wind Generation –TWh: Numerical,Continuous

Oil Consumption – EJ: Numerical,Continuous

Data Pre-Processing

```
class preprocessing():

def indoor2016(self):
    df1=self.dfselection(self.setf2,self.data_health,'World_Bank_Name')
    df2=self.dfselection(self.setf2,self.data_clean,'Location')
    df3=self.dfselection(self.setf2,self.data_pollution,'Entity')
    df4=self.dfselection(self.setf2,self.data_GDP,'Country')
    df2=df2[df2['Period']==2016]
    df3=df3[df3['Year']==2016]
    df4=df4[df4['Year']==2016]
    df1=df1.rename(columns={"World_Bank_Name": "Location"})
    df3=df3.rename(columns={'Entity':'Location'})
    df4=df4.rename(columns={'Country':'Location'})
    out=pd.merge(df1,df2,how='outer',on='Location')
    out=pd.merge(out,df3,how='outer',on="Location")
    out=pd.merge(out,df4,how='outer',on='Location')
    out=out.drop(columns='Province_State')

    return out
```

Indoor2016():

Select datasets
Year = 2016
Rename columns
Merge datasets

Data Pre-Processing

```
def indoor(self):

    df2=self.dfselection(self.setf2,self.data_clean,'Location')
    df3=self.dfselection(self.setf2,self.data_pollution,'Entity')
    df4=self.dfselection(self.setf2,self.data_GDP,'Country')
    df2 = df2.rename(columns={"Period": "Year"})
    df3 = df3.rename(columns={'Entity': 'Location'})
    df4 = df4.rename(columns={'Country': 'Location'})
    out = pd.merge(df2, df3, how='outer', on=['Location','Year'])
    out = pd.merge(out, df4, how='outer', on=['Location','Year'])

    return out
```

indoor():

Select datasets

Rename columns

Merge datasets by Location and Year

Data Pre-Processing

```
def outdoor(self):


df3=self.dfselection(self.setf,self.data_pollution,'Entity')
    df4=self.dfselection(self.setf,self.data_GDP,'Country')
    df3 = df3.rename(columns={'Entity': 'Location'})
    df4 = df4.rename(columns={'Country': 'Location'})
    out = pd.merge(df3, df4, how='outer', on=['Location',
'Year'])
    return out
```

Outdoor():

Select datasets

Rename columns

Merge datasets by Location and Year

Data Pre-Processing

```
def Merge_energy(self):
    df5 = self.dfselection(self.setf2, self.data_energy, 'Entity')
    df3 = self.dfselection(self.setf2, self.data_pollution, 'Entity')
    df4 = self.dfselection(self.setf2, self.data_GDP, 'Country')
    df3 = df3.rename(columns={'Entity': 'Location'})
    df4 = df4.rename(columns={'Country': 'Location'})
    df5 = df5.rename(columns={'Entity': 'Location'})
    out = pd.merge(df5, df3, how='outer', on=['Location', 'Year'])
    out = pd.merge(out, df4, how='outer', on=['Location', 'Year'])
    out = out.drop(columns='Wind Generation -TWh')
    out = out.drop(columns='Solar Generation - TWh')
    out = out.drop(columns='Nuclear Generation - TWh')
    out = out.drop(columns='Hydro Generation - TWh')
    out = out.drop(columns='Geo Biomass Other - TWh')
    out = out.drop(columns='Gas Consumption - EJ')
    return out
```

Merge_energy():

Select datasets

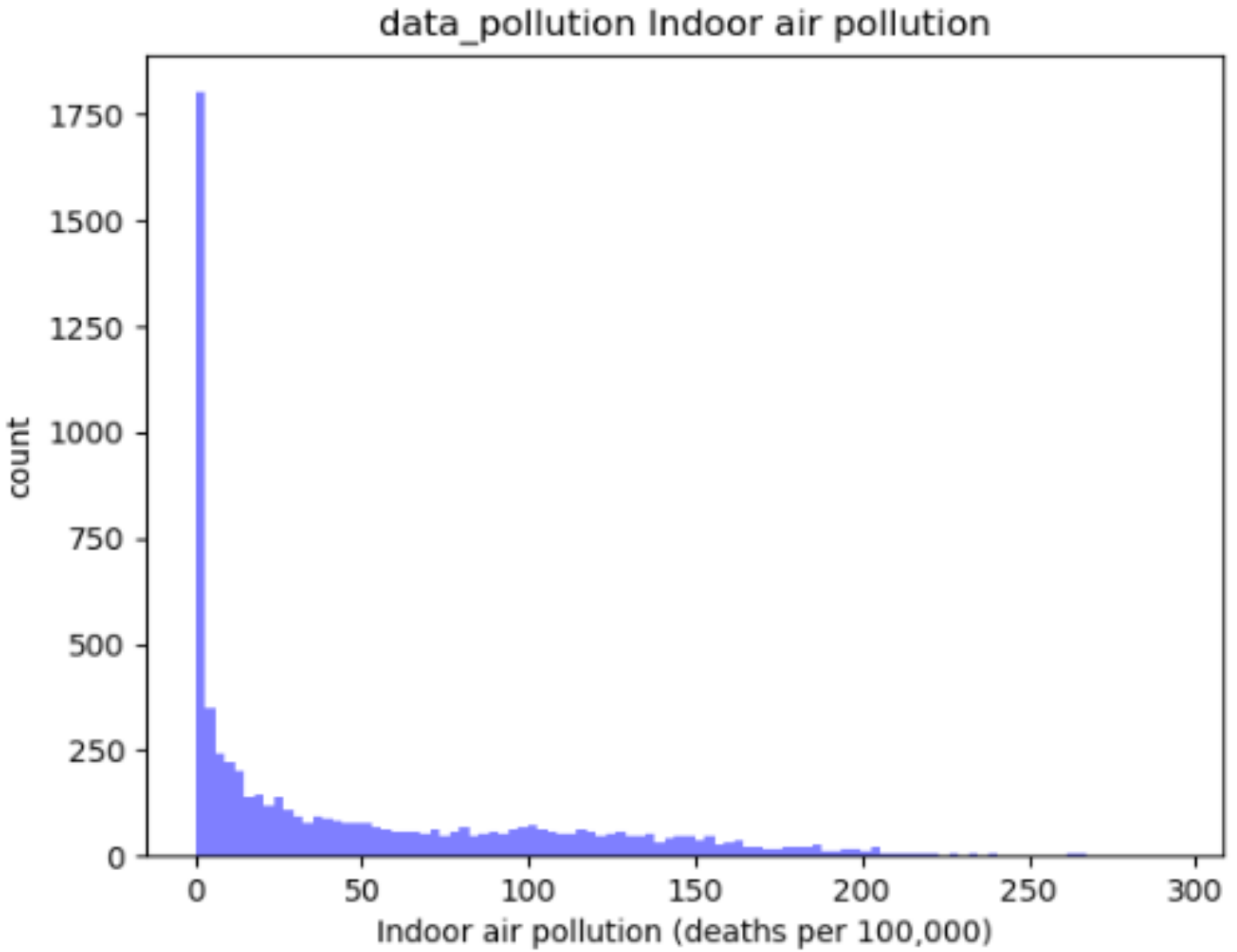
Rename columns

Merge datasets

Drop unnecessary columns

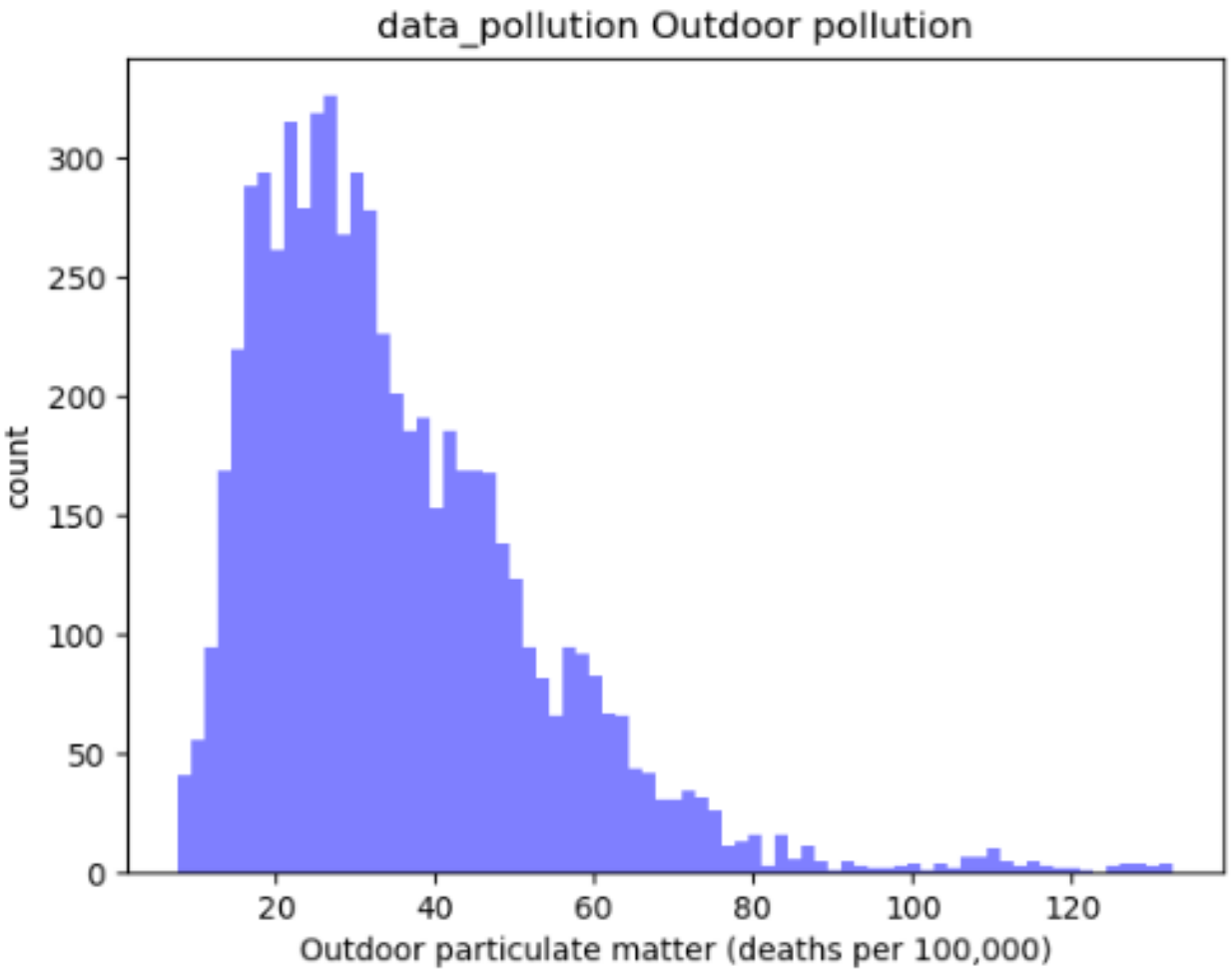
EDA

```
num_bins = 100
n, bins, patches = plt.hist(dp['Indoor air pollution (deaths per 100,000)'], num_bins, facecolor='blue', alpha=0.5)
plt.xlabel('Indoor air pollution (deaths per 100,000)')
plt.ylabel('count')
plt.title('data_pollution Indoor air pollution')
plt.show()
```



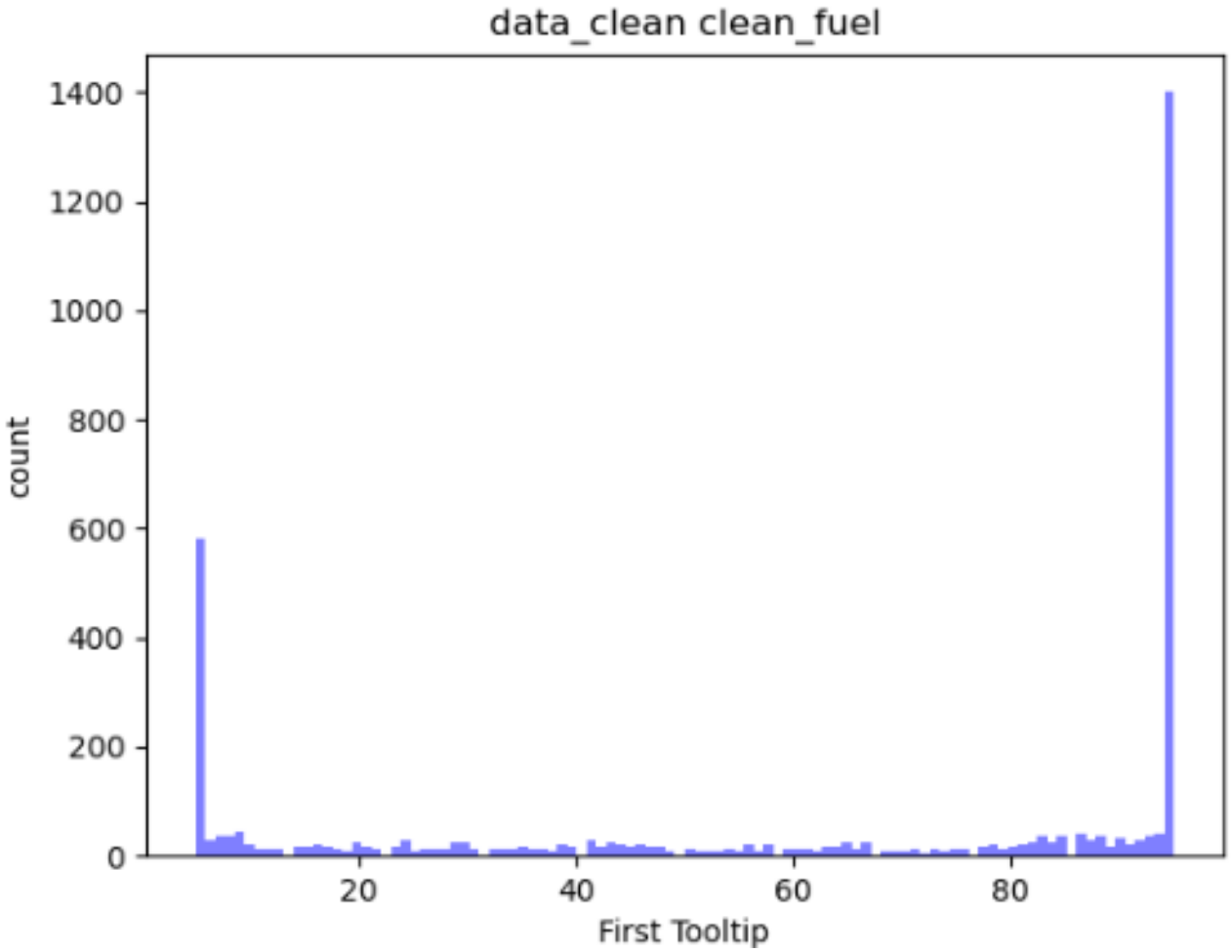
EDA

```
dp = data_pollution_p
num_bins = 75
n, bins, patches = plt.hist(dp['Outdoor
particulate matter (deaths per 100,000)'],
num_bins, facecolor='blue', alpha=0.5)
plt.xlabel('Outdoor particulate matter (deaths
per 100,000)')
plt.ylabel('count')
plt.title('data_pollution Outdoor pollution')
plt.show()
```



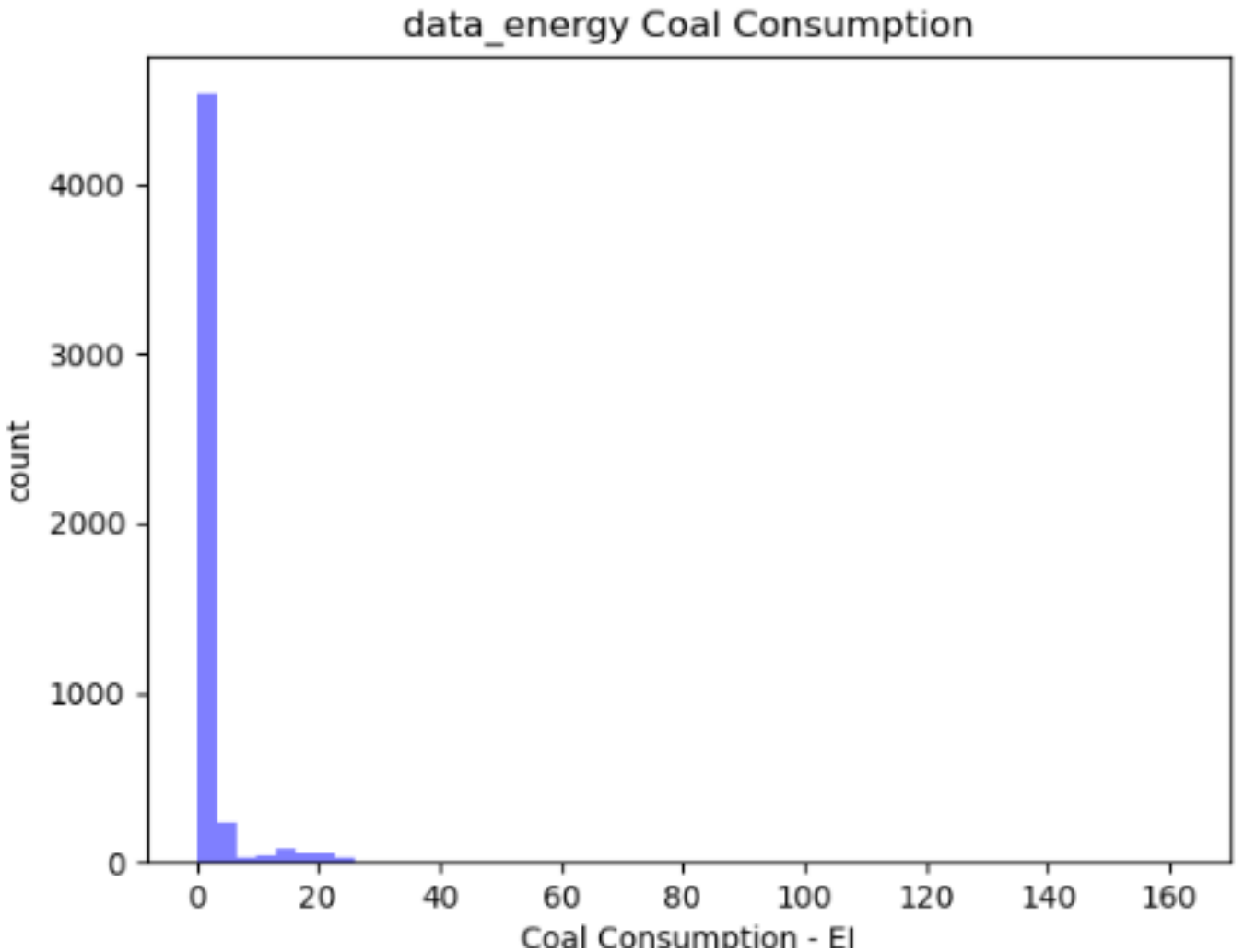
EDA

```
dc = data_clean_p
num_bins = 100
# the histogram of the data
n, bins, patches = plt.hist(dc['First Tooltip'],
num_bins, facecolor='blue', alpha=0.5)
plt.xlabel('First Tooltip')
plt.ylabel('count')
plt.title('data_clean clean_fuel’)
plt.show()
```



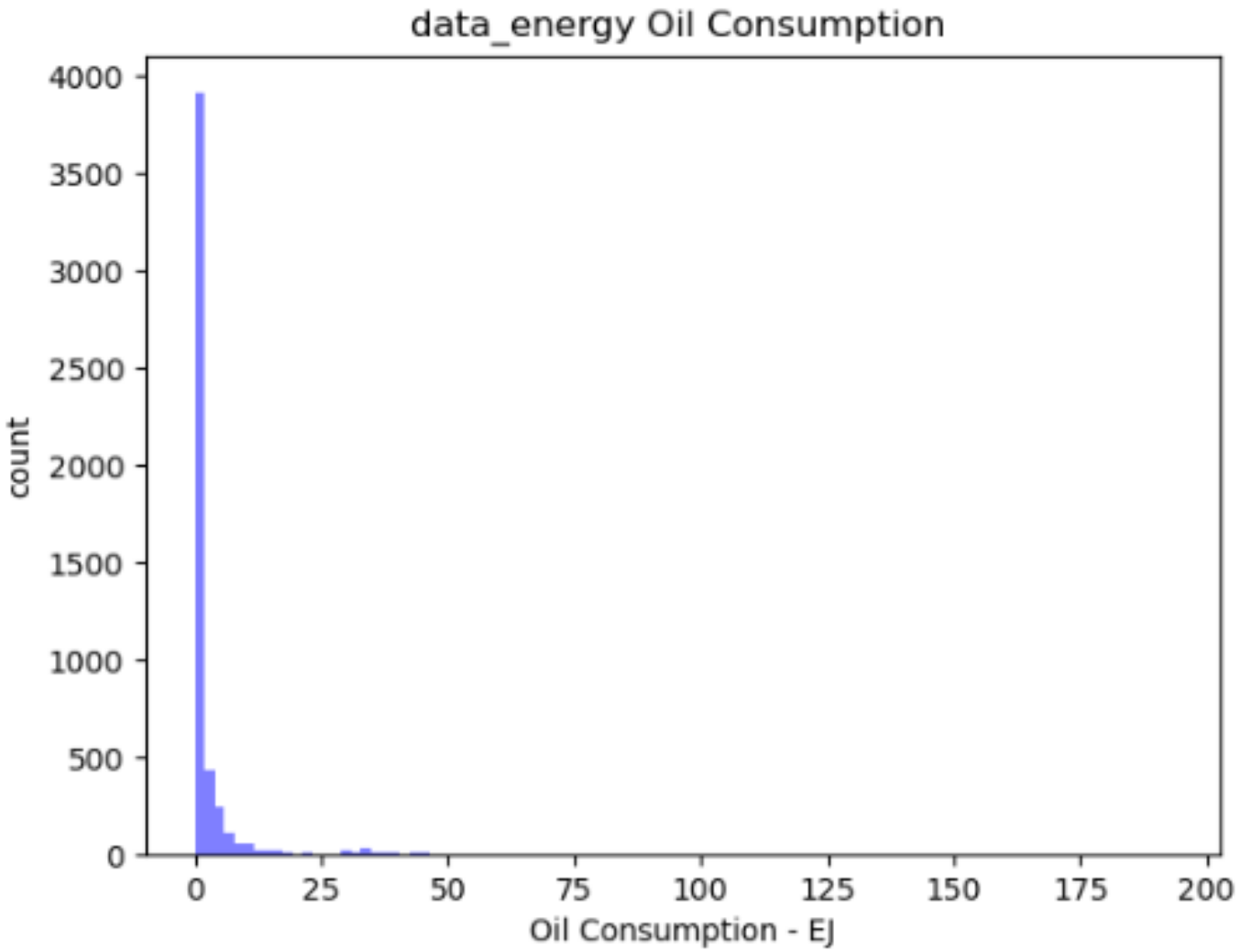
EDA

```
de = data_energy_p
num_bins = 50
n, bins, patches = plt.hist(de['Coal
Consumption - EJ'], num_bins,
facecolor='blue', alpha=0.5)
plt.xlabel('Coal Consumption - EJ')
plt.ylabel('count')
plt.title('data_energy Coal Consumption')
plt.show()
```



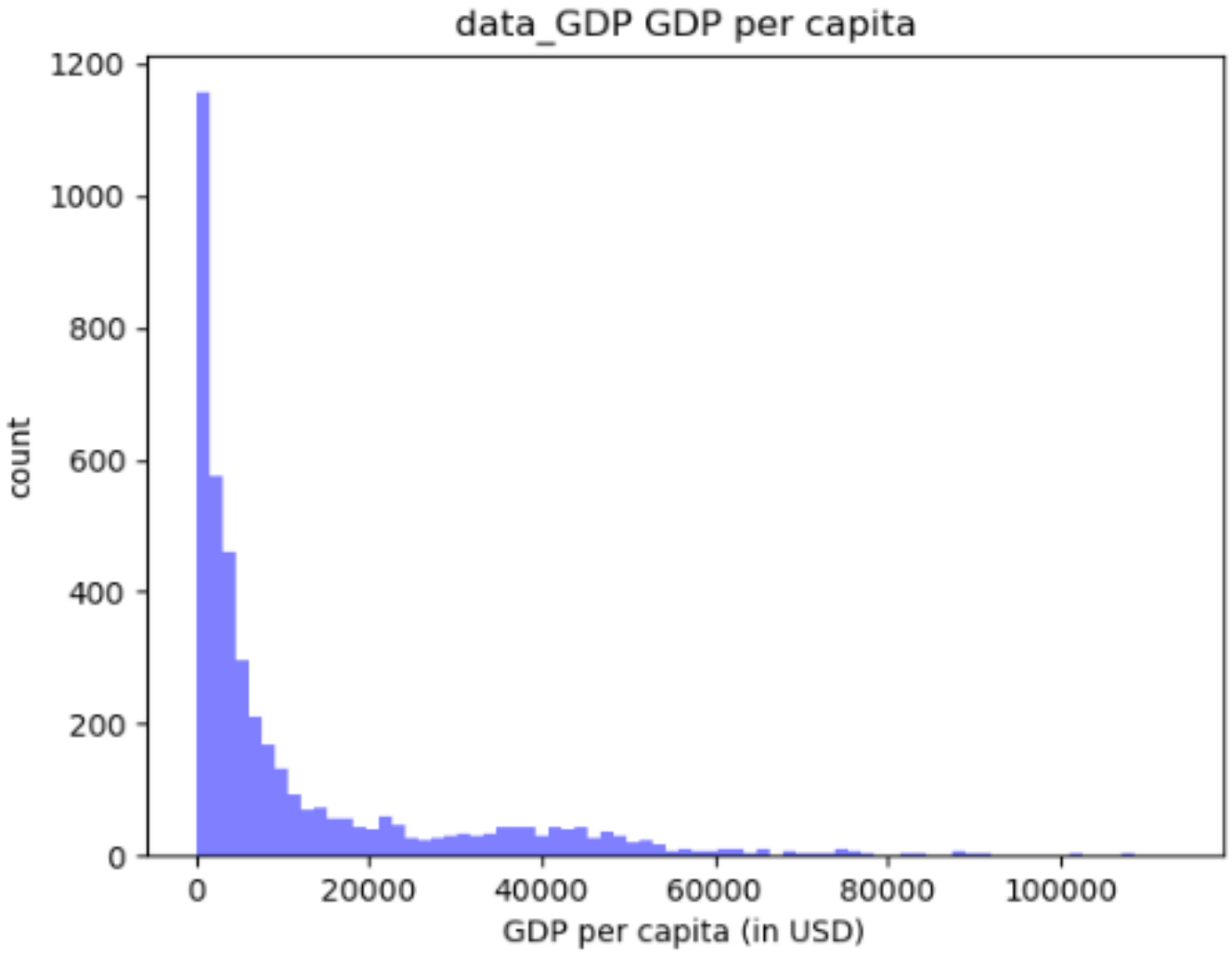
EDA

```
num_bins = 100
n, bins, patches = plt.hist(de['Oil
Consumption - EJ'], num_bins,
facecolor='blue', alpha=0.5)
plt.xlabel('Oil Consumption - EJ')
plt.ylabel('count')
plt.title('data_energy Oil Consumption')
plt.show()
```

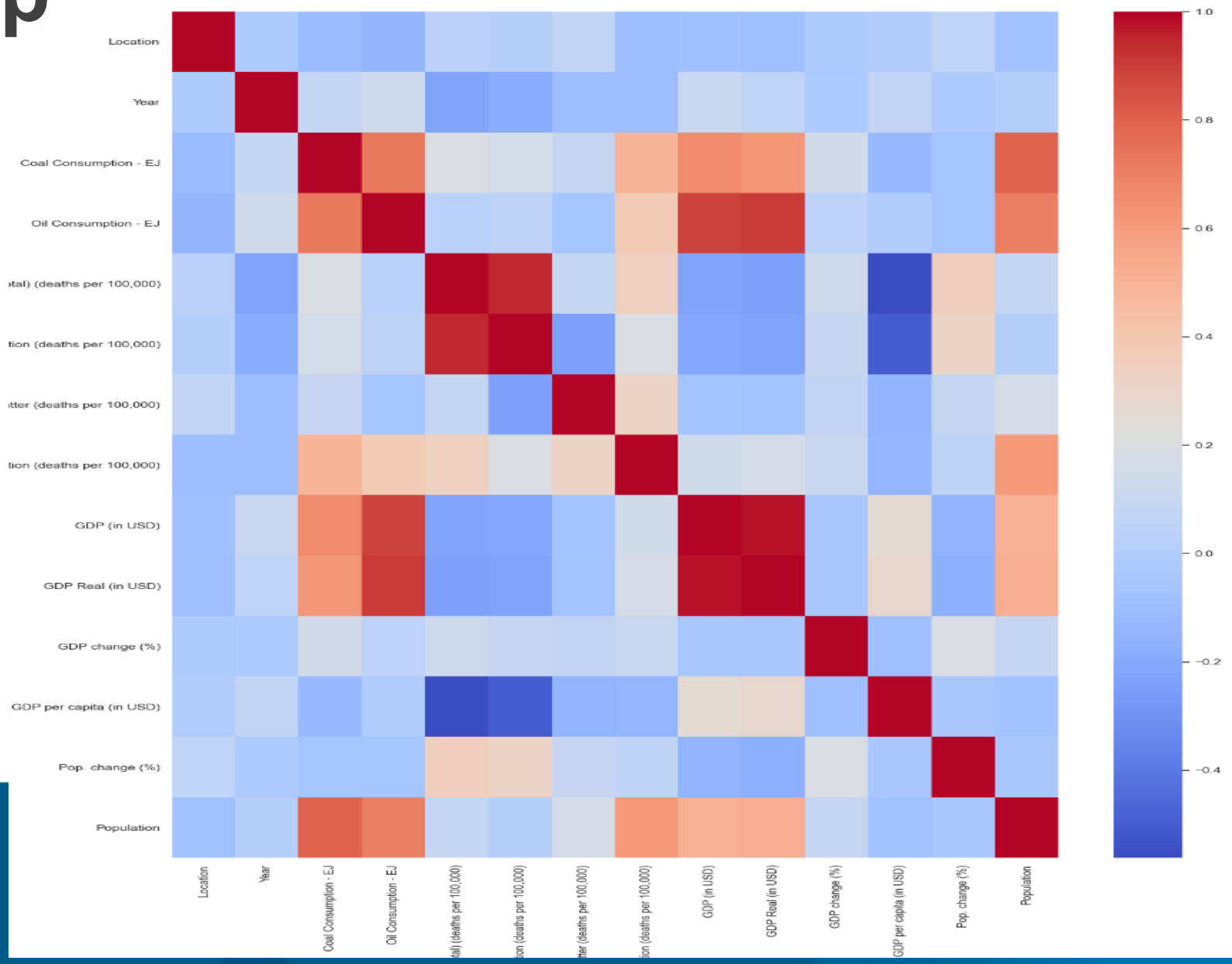


EDA

```
dg = data_GDP_p
num_bins = 75
n, bins, patches = plt.hist(dg['GDP per capita (in USD)'], num_bins, facecolor='blue', alpha=0.5)
plt.xlabel('GDP per capita (in USD)')
plt.ylabel('count')
plt.title('data_GDP GDP per capita')
plt.show()
```



Heat map



GEORGE
SHINGTON
VERSITY
HINGTON, DC

Introduction of the Model

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

The $\hat{\beta}$ is a 1*N vector , The $\hat{\beta}$ contain N estimated parameters: first element is interception parameter and the rest are N-1 slope parameters

X is a vector with first element of 1, and N-1 explanatory features.

Use matrix operation to estimate the $\hat{\beta}$

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 i'th observation and j-1th feature

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

Analysis of the Model

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

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

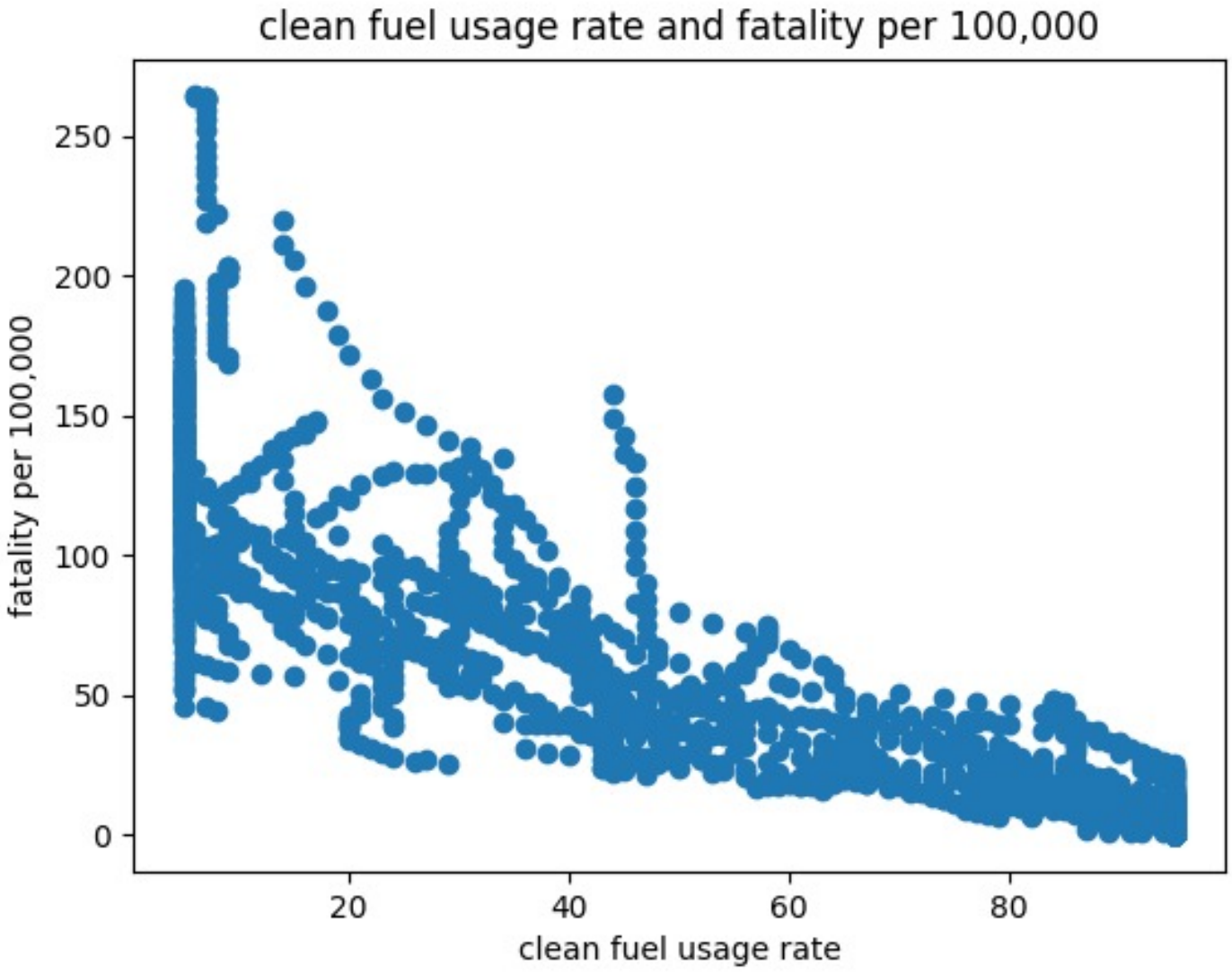
Analysis of the Model

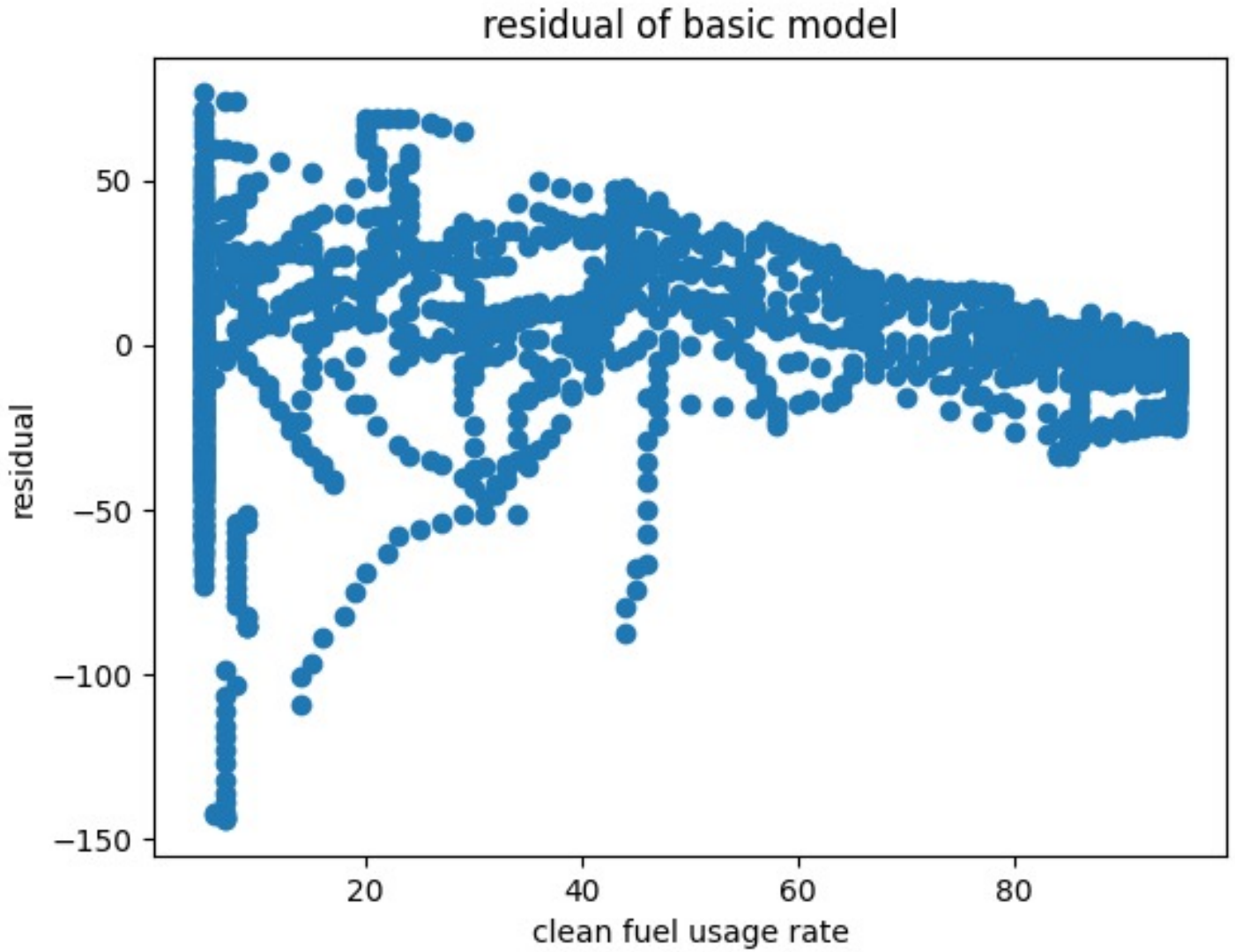
R-squared suppose portion of predicting variable's variation in observations could be explained by the model.

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

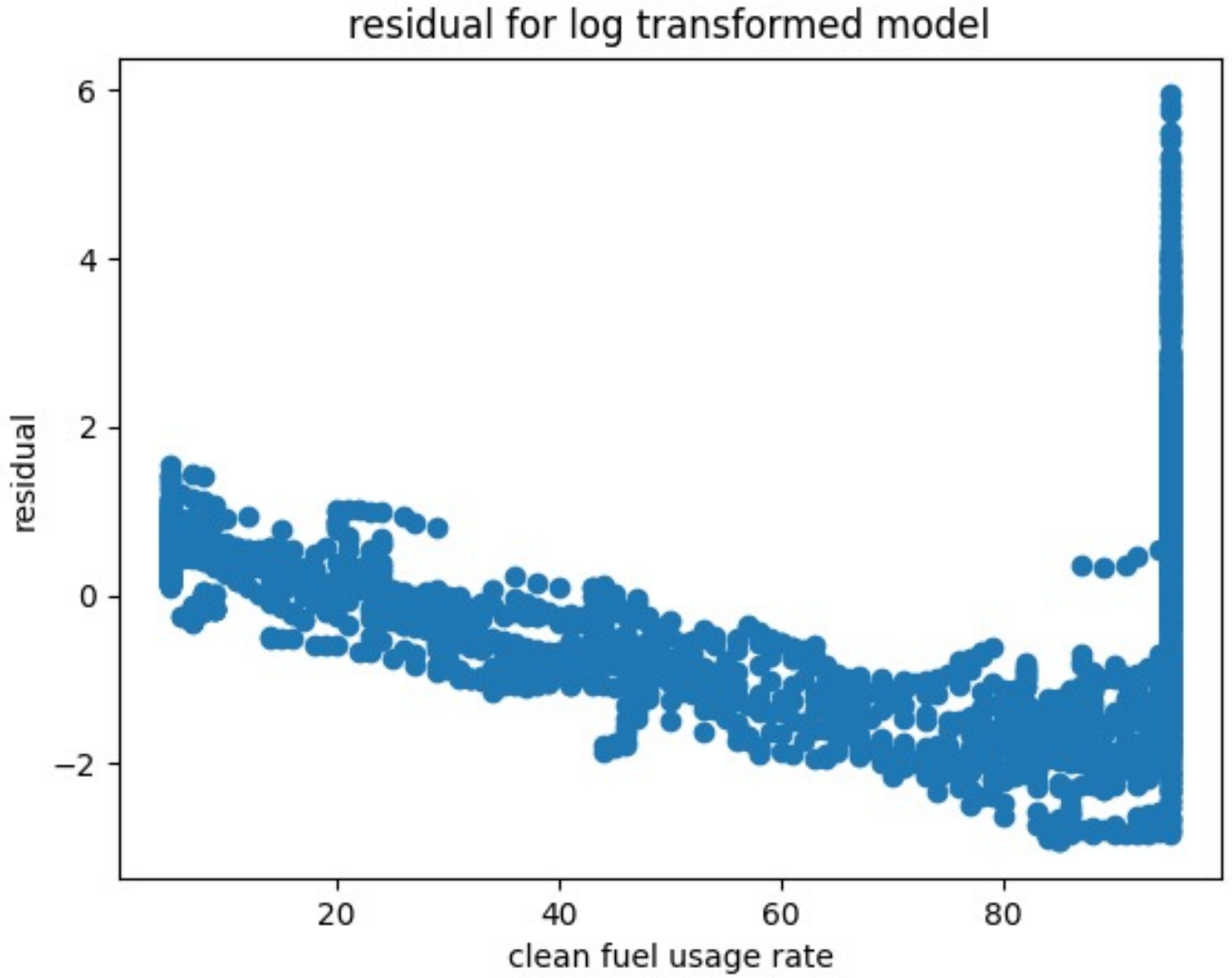
GUI GRAPHS DISPLAY

Findings – Basic Model Indoor Pollution

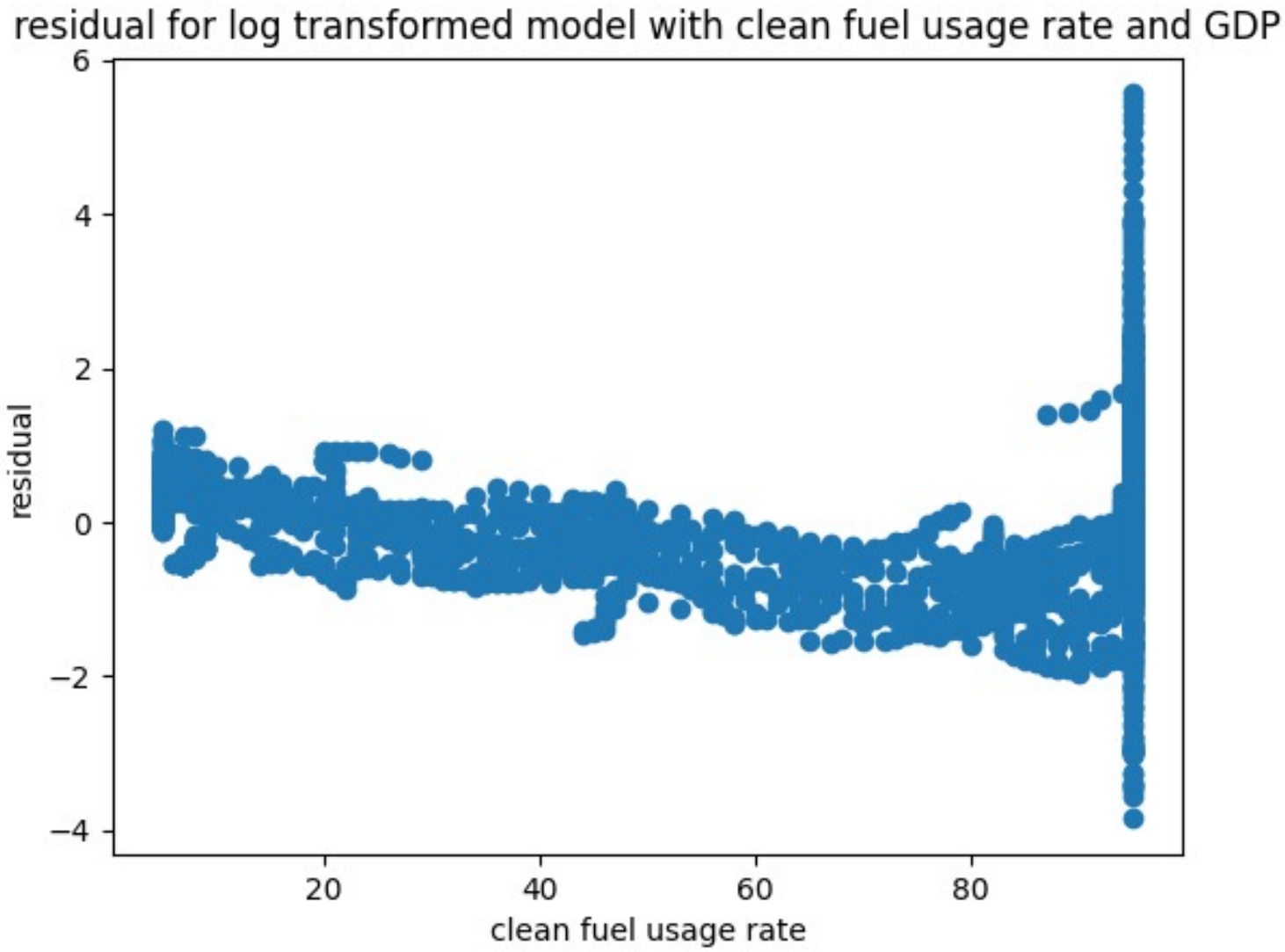




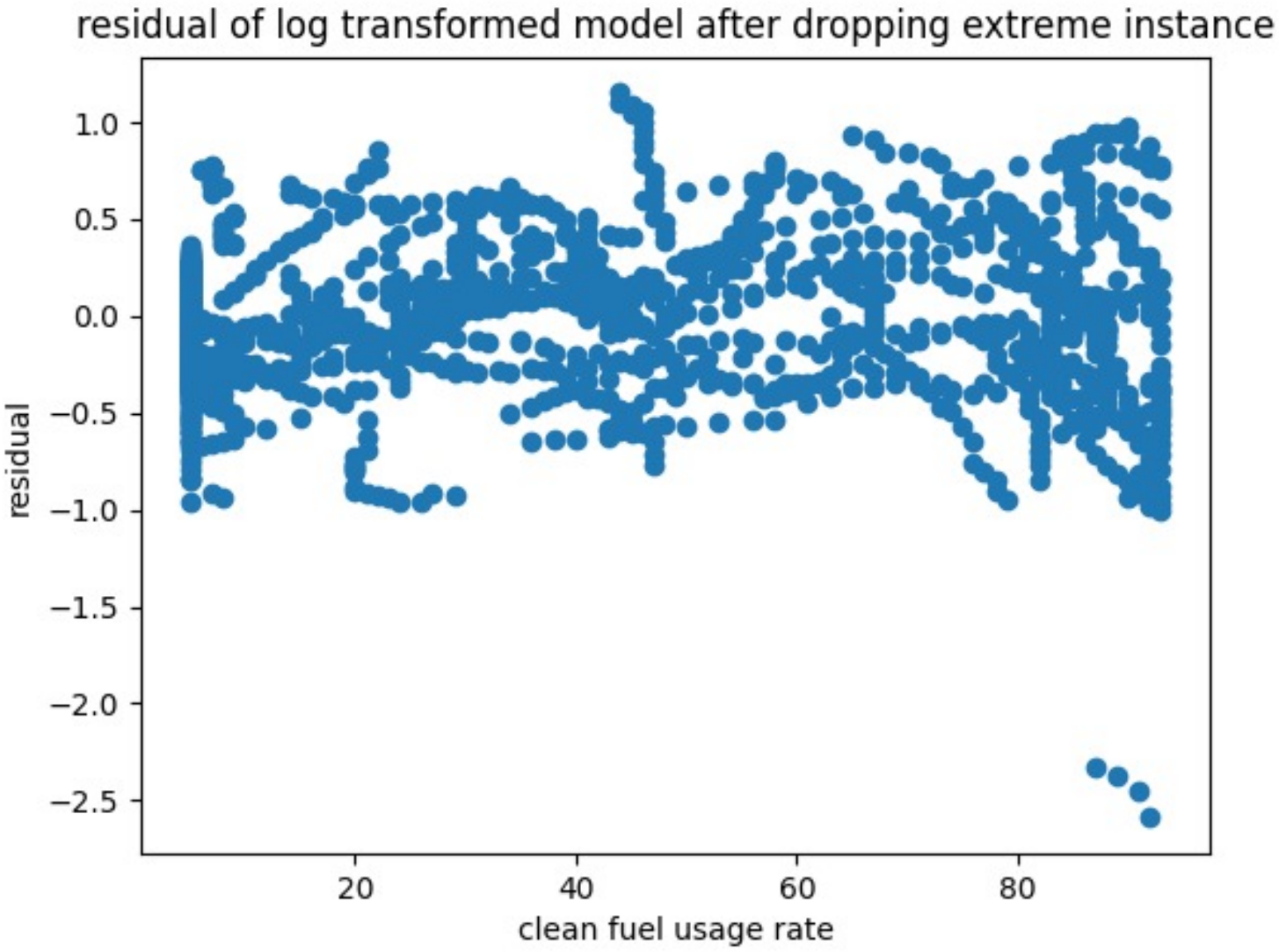
Transform the Model with 1 Feature

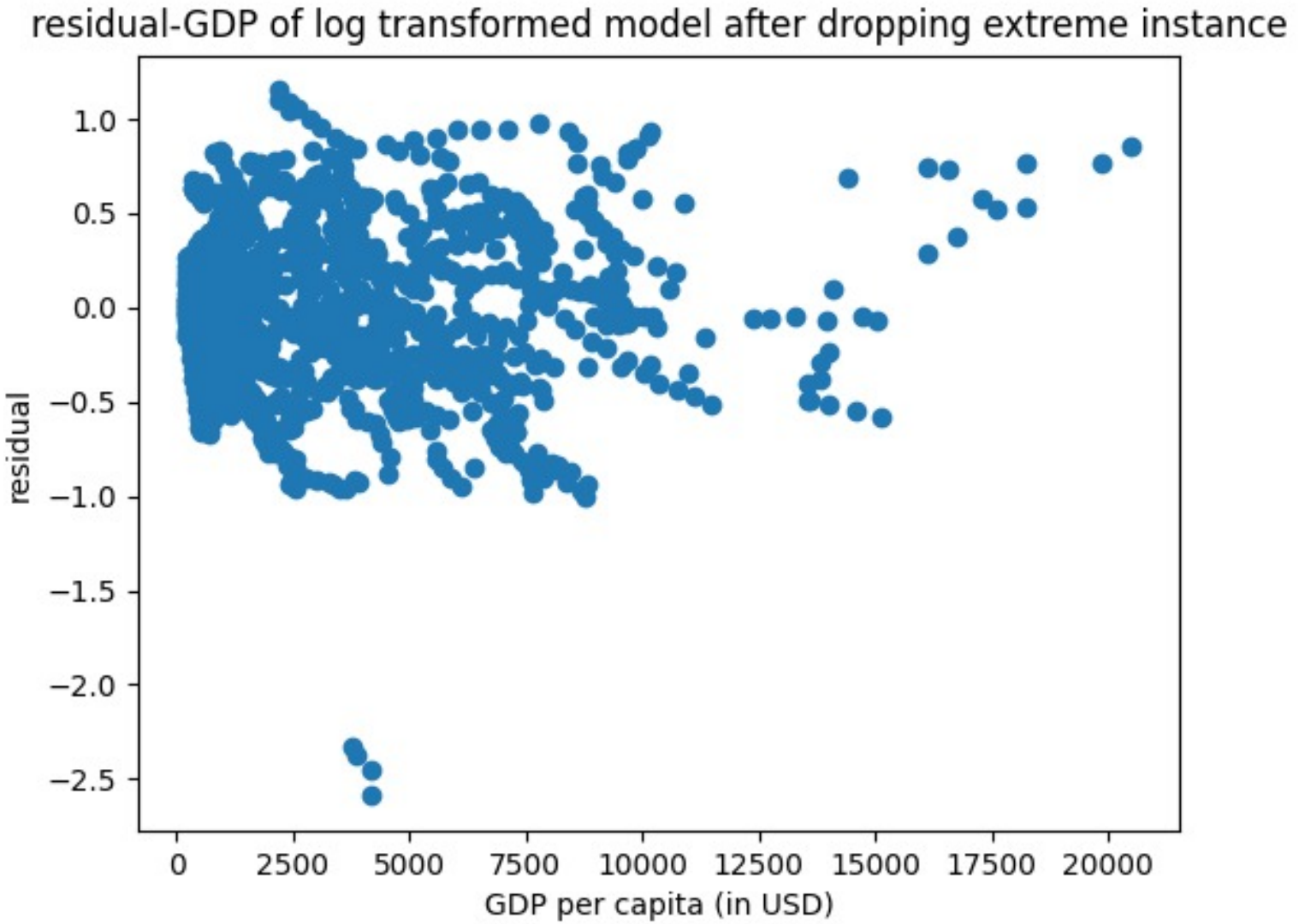


Transform the Model with 2 Features

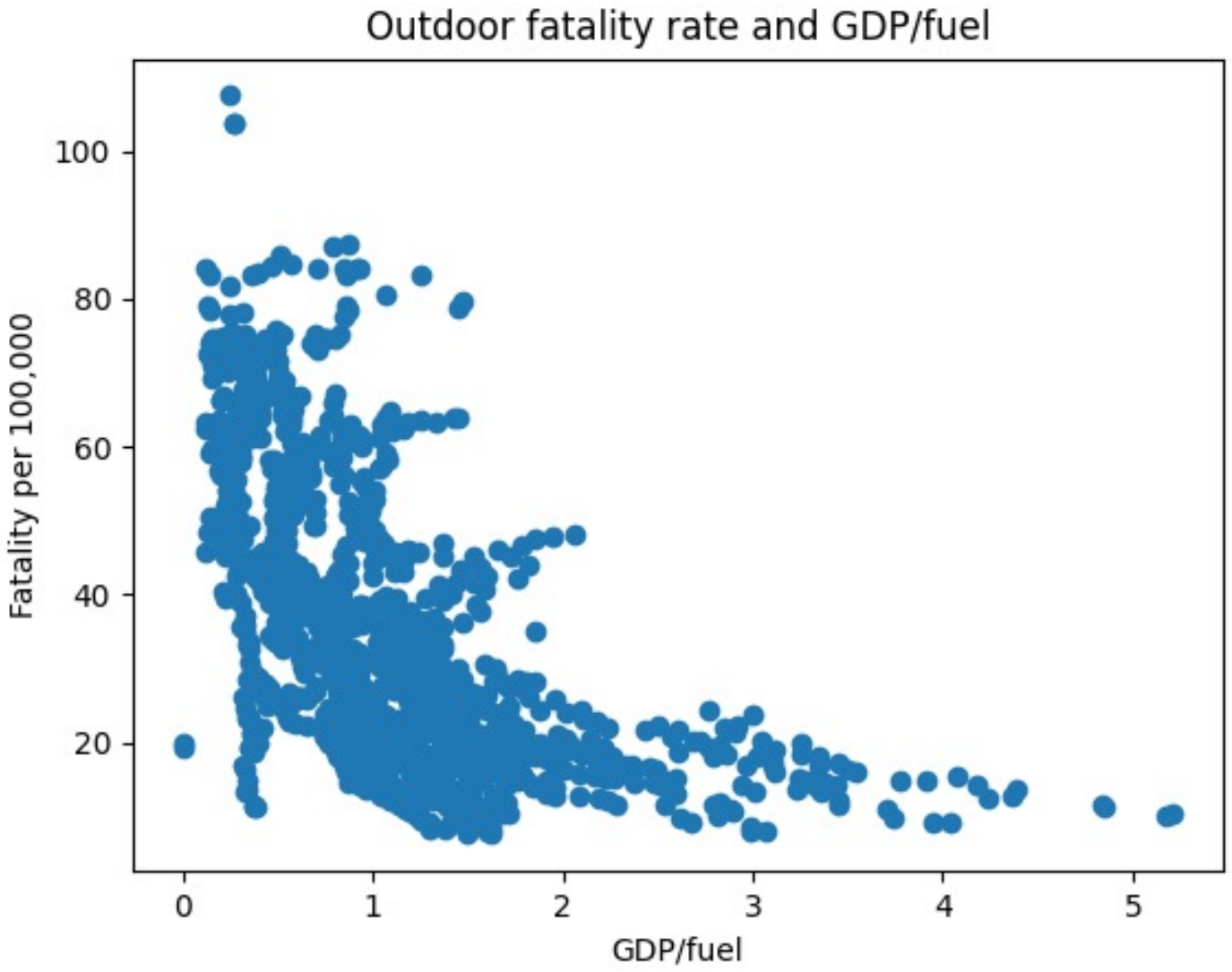


Best Model with R^2 = 0.82

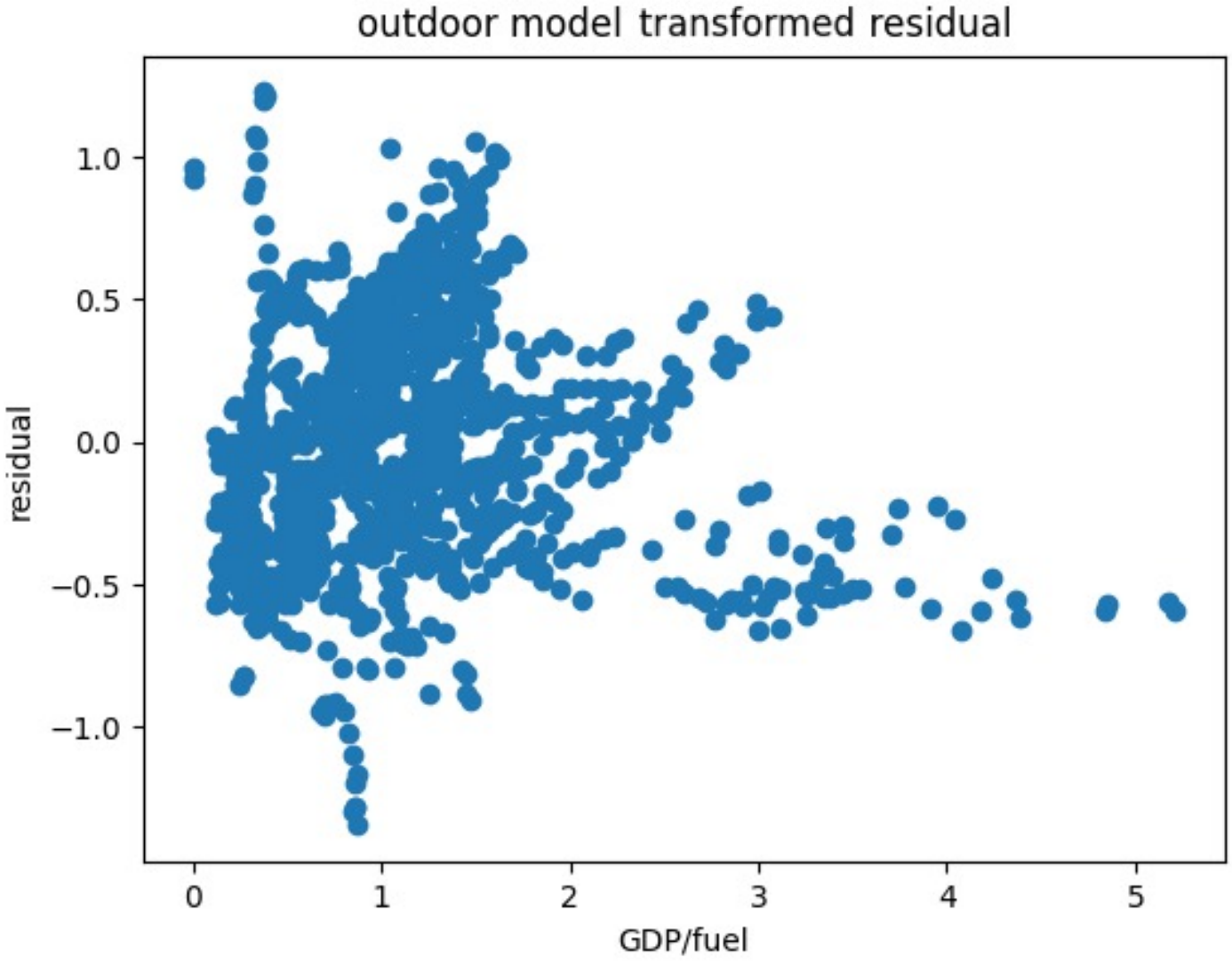




Some Correlation With Generated Feature



Residual of the Transformed Model with Generated Feature



Results

- Best Model for indoor pollution with R^2 is 0.82
- Best Model for outdoor pollution with R^2 is 0.51
- Indoor model shows constant variance for residual
- Outdoor model has a high regression, residual not constant where the GDP/Fuel ratio is high
- This could induced by the small amount of observation data

Conclusion

- For indoor pollution, the fatality rate shows as a function of exponential decay with clean fuel usage rate until the extreme high value. The value > 94%.
- For outdoor pollution, the fatality rate shows some relation with GDP/fuel.

Suggestions

- Indoor Pollution
  - Clean Fuel
    - Cook
    - Warm Up
- Outdoor pollution
  - Energy Density
    - Industrial Production
    - Transportation