GITHUB LINK TO CODE:

https://github.com/CS418/group-project-big-baller-data

Changes

During our discussion with the professor, we realized that we needed to obtain more data sources than what we had. This is one of the changes that we had to make, so we added another source to our project. We also saw that our data cleaning was ok so no changes were made to that. For this project, we will aim to visualize all of the data and try to connect different sources together in order to find a correlation between them.

Data

1. The first dataset that will be analyzed is obtained from the Bureau of Transportation Statistics (BTS) website. The table provides estimated national average vehicle emissions rates for different vehicle types using gasoline and diesel fuel.
2. The second dataset is obtained from the Fuel Economy website, which is maintained by the U.S. Department of Energy. The table contains data on the fuel economy of different vehicles, including passenger cars and trucks, as well as information on the emissions generated by these vehicles. The original size of the data is 2539 rows and 18 columns.
3. The third dataset is obtained from the Greenhouse Gas Inventory Data Explorer of the U.S. Environmental Protection Agency. The original size of the data is 5 rows and 30 columns. The table provides data on the emissions of greenhouse gasses from different sectors, including transportation, energy, and industry.
4. The fourth dataset is obtained from https://www.epa.gov/compliance-and-fuel-economy-data/data-cars-used-testing-fuel-economy which contains data on the range and efficiency of electric vehicles. The original size of the data is 4298 rows and 67 columns. Each row provides information on a specific electric vehicle, such as its make, model, and range, etc. .

By examining the data from these various sources we are able to make judgments about the environmental effect of electric vehicles and how they stack up against other kinds of vehicles. As automakers, and customers who are interested in lessening the environmental effect of transportation may find this information to be useful.

Problem

In these modern times, we often hear about electric vehicles being the future of the automobile industry because of its many advantages over conventional gas

powered cars. Our aim for this project was to ask the question “Are electric vehicles actually good for the environment, or is their impact the same as gas cars?”

Research questions

Christian:
How do emissions of carbon monoxide, nitrogen oxides, particulate matter, and volatile organic compounds vary across different vehicle types and fuel types, and how have these emissions changed over time?

Hamza:
What factors are most strongly associated with higher fuel efficiency, and how do these factors vary by vehicle type, fuel type, and model year? Can we identify specific vehicle models or manufacturers that consistently have high fuel efficiency ratings?

Paul:
What share of total greenhouse gas emissions in the United States can be attributed to transportation, and how have these emissions changed over time? How do emissions from different transportation modes (e.g. cars, trucks, buses) contribute to overall transportation emissions?

Abanoub:
What is the correlation between fuel economy and vehicle weight in cars used for testing fuel economy?
Is there a difference in fuel economy performance between domestic and foreign cars used for testing?

Data cleaning

1. The first set of data that Christian cleaned is related to the Bureau of Transportation Statistics.

```
data = pd.read_csv('BTSData.csv', header=0, index_col=0)
#dropping the data notes
data = data.drop(data.index[60:70])
# #dropping the brakewear data because it is present in both evs and electric cars
data.drop('Tirewear PM2.5', inplace=True)
data.to_csv('CleanedBTSData.csv', index=False)
```

Overall, we lost rows of data pertaining to factors in emissions that apply to both EVs and gas powered cars as they do not add anything to our research. We kept only the

data that can change between EVs and gas powered vehicles.

2. The second set of data cleaned was Data on Car Fuel Economy and their effects on the environment with Air pollution score and Greenhouse Gas score

```
import pandas as pd
import numpy as np

df = pd.read_excel('/content/drive/MyDrive/CS418/Milestone/all_alpha_23.xlsx')
df.fillna(0, inplace=True)
before = df.shape
df = df.drop(columns=['Trans','Cert Region', 'Stnd','Stnd Description',
                      'Underhood ID', 'City MPG', 'Hwy MPG', 'Cmb MPG', 'SmartWay'])
df = df.reset_index(drop=True)
after = df.shape
print(df)
print("before cleaning:", before)
print("after cleaning:", after)
```

```
before cleaning: (2539, 18)
after cleaning: (2539, 9)
```

After cleaning the data, the column size went from 18 to 9 while the row size did not change. Within the data the electric cars have nan values so it was replaced with 0 to indicate that the electric doesn't have displ and cyl.

3. The third set of data cleaned was Data on Cars used for Testing Fuel Economy:

```
#ABANOUBs clean data

data = pd.read_excel('23-testcar.xlsx')

data = data.drop(columns=['Represented Test Veh Make','Veh Mfr Code',
'Test Vehicle ID','Test Veh Configuration #', '# of Cylinders and Rotors',
'Engine Code','Tested Transmission Type Code','# of Gears','Transmission
Lockup?', 'Drive System Code','Transmission Lockup?', 'Transmission
Overdrive Code','Transmission Overdrive Desc', 'Axle Ratio','N/V
Ratio','Shift Indicator Light Use Cd','Shift Indicator Light Use
Desc','Test Number','Test Originator','Analytically Derived FE?','ADFE
Test Number','ADFE Total Road Load HP','ADFE Equiv. Test Weight (lbs.)',
'ADFE N/V Ratio','Test Originator','FE Bag 4','Police - Emergency
Vehicle?', 'Averaging Group ID','Averaging Weighting Factor','Averaging
Method Cd', 'Averging Method Desc'])
```

```
# save the cleaned data to a new CSV file
data.to_csv('cleaned_data.csv', index=False)
```

```
Before cleaning:  (4298, 67)
After cleaning:  (4298, 38)
```

After cleaning the data, the number of rows stayed the same, but the number of columns decreased from 67 to 38. Therefore, by removing irrelevant or redundant data I was able to improve the accuracy and efficiency of my data analysis.

4.The fourth set of data cleaned was Greenhouse inventory data:

```
import pandas as pd
from matplotlib import pyplot as plt
df = pd.read_csv('data.csv')
# Drop the row with 'Incineration of waste'
df = df[df.index != 'Incineration of waste']
```

Cleaning my data consisted of dropping the row ‘Incineration of waste’ because that would be useless information in our research study. Keeping only the data for combustion and electricity emissions to do a comparison between the two over the years. The size before the cleaning was: 5 rows and 30 columns. The size after the cleaning was 4 rows and 30 columns.

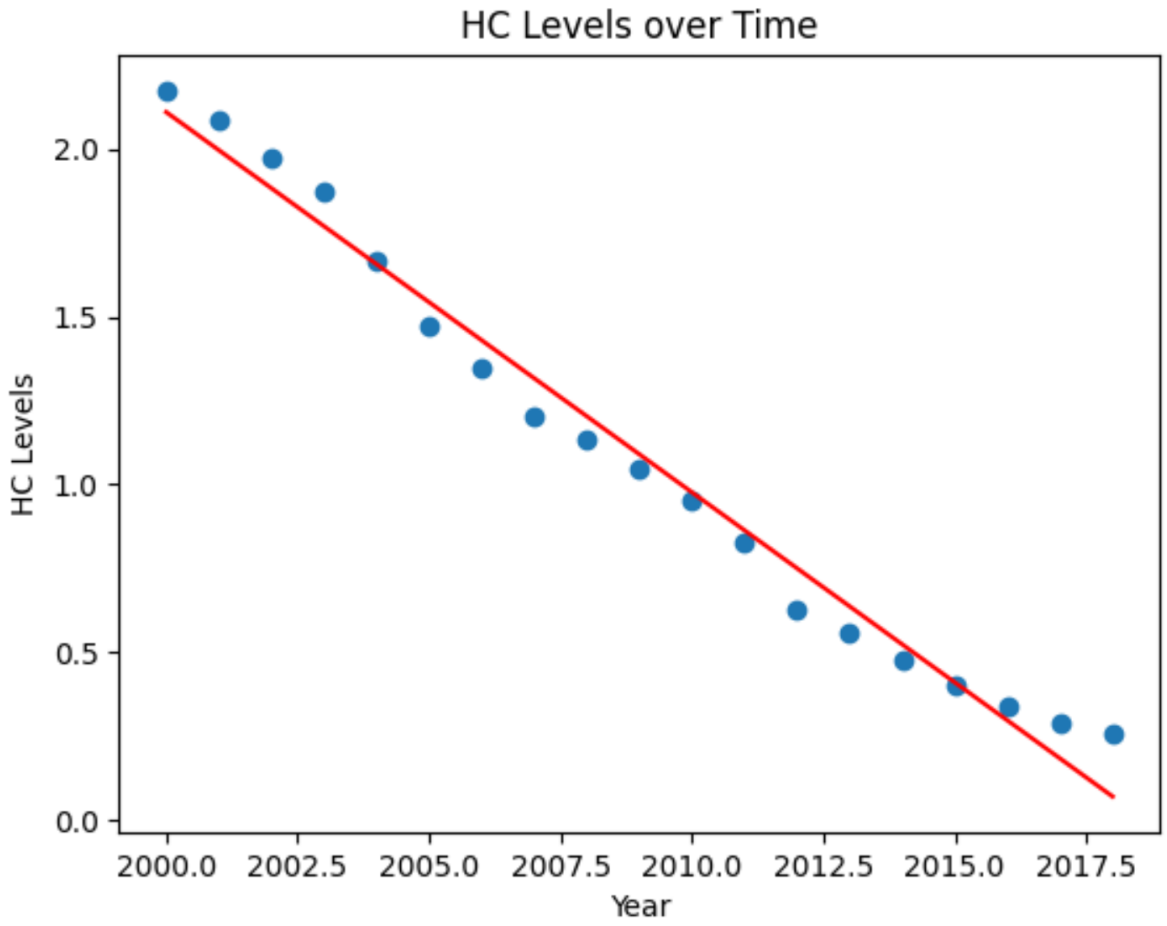
Exploratory data analysis

Christian:

```
row = data.iloc[4]
years = []
hc_levels = []
for col in row.index:
    if col.startswith('(R) '):
        year_str = col[4:]
        if year_str.isdigit():
            years.append(int(year_str))
            hc_levels.append(row[col])

plt.scatter(years, hc_levels)
```

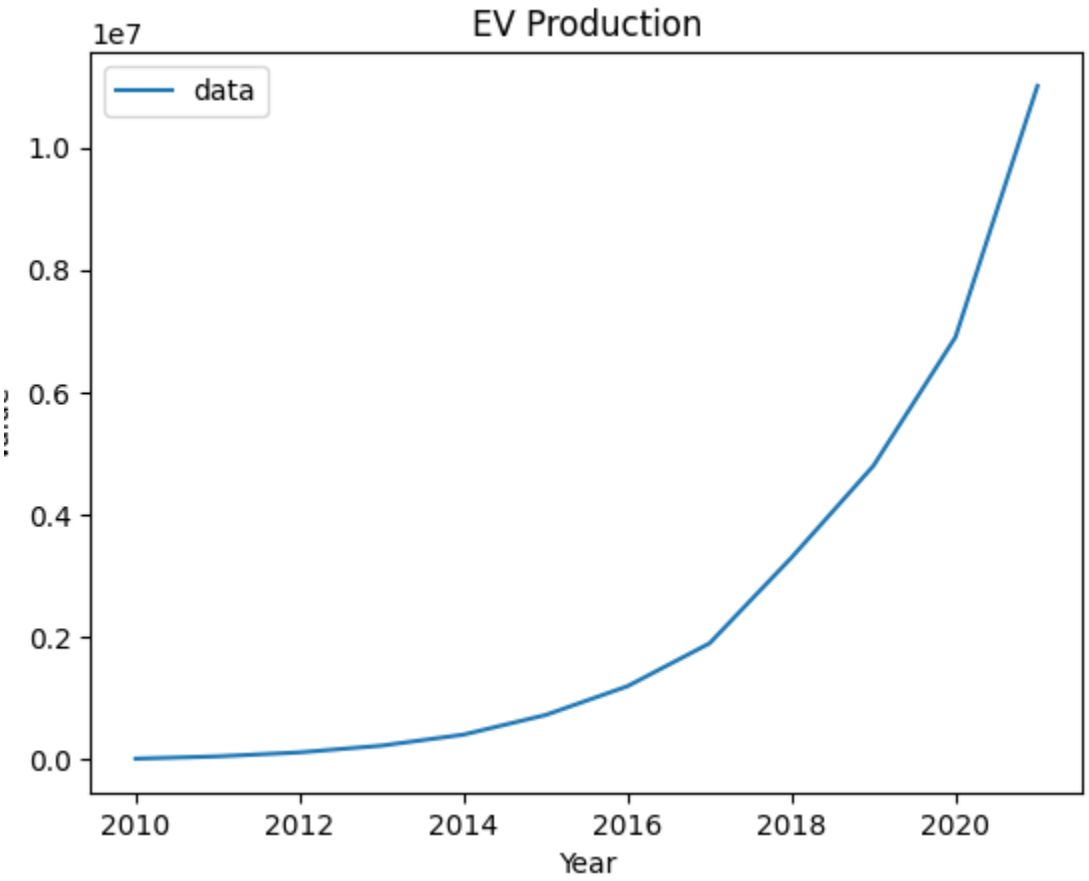
```
coefficients = np.polyfit(years, hc_levels, 1)
line_fn = np.poly1d(coefficients)
plt.plot(years, line_fn(years), color='red')
plt.xlabel('Year')
plt.ylabel('HC Levels')
plt.title('HC Levels over Time')
plt.show()
```



The next set of data shows the increase of EV production over time

```
from scipy.optimize import curve_fit
df = pd.read_csv('EV production data.csv', header=0)
df = df[df['region'] == 'World']
df = df[~df['powertrain'].str.contains('PHEV')]
plt.plot(df['year'], df['value'], label='data')
plt.xlabel('Year')
plt.ylabel('Value')
plt.title('EV Production')
```

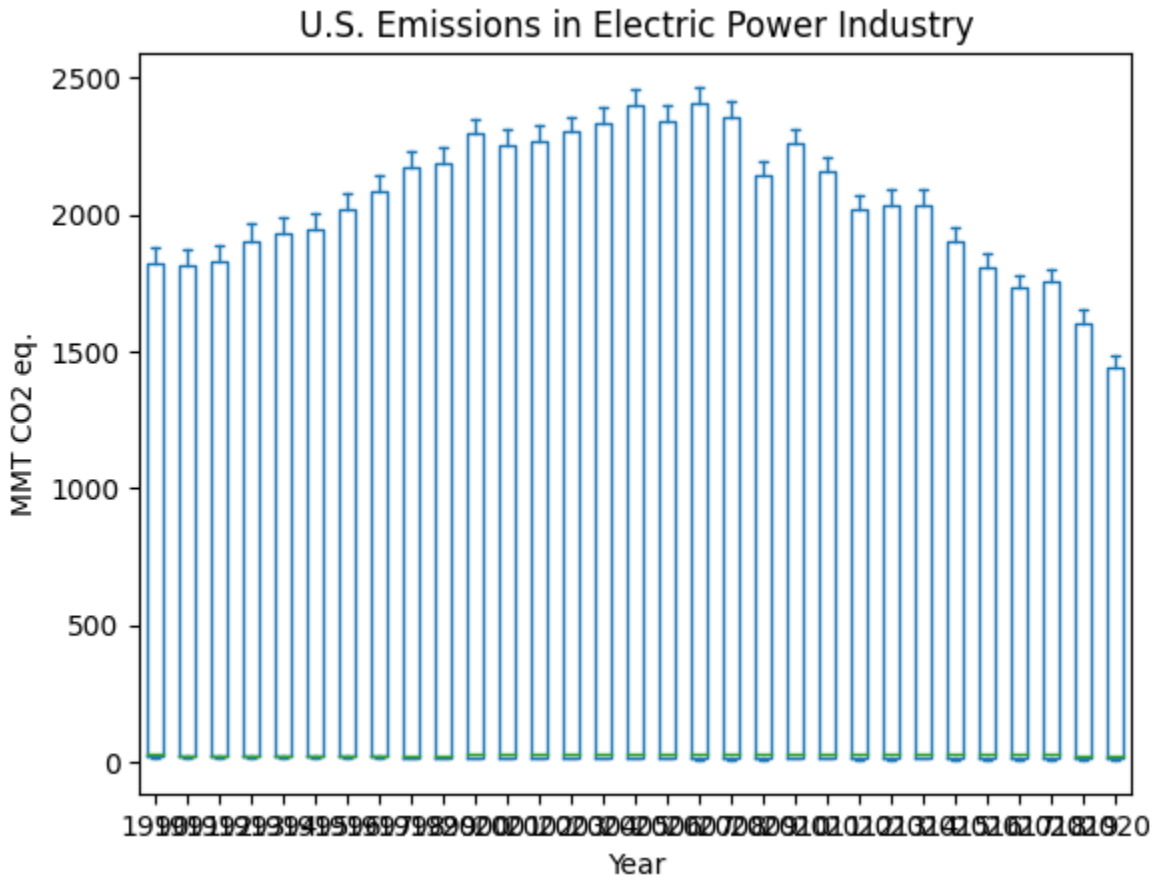
```
plt.legend()
plt.show()
```



One of the preliminary conclusions is that, I would have expected the HC levels to decrease as more EV vehicles were produced near the 2020s but it has stayed the same.

Paul:

```
plt.figure(figsize=(10,10))
df.plot(kind='box')
plt.title('U.S. Emissions in Electric Power Industry')
plt.ylabel('MMT CO2 eq.')
plt.xlabel('Year')
plt.show()
```

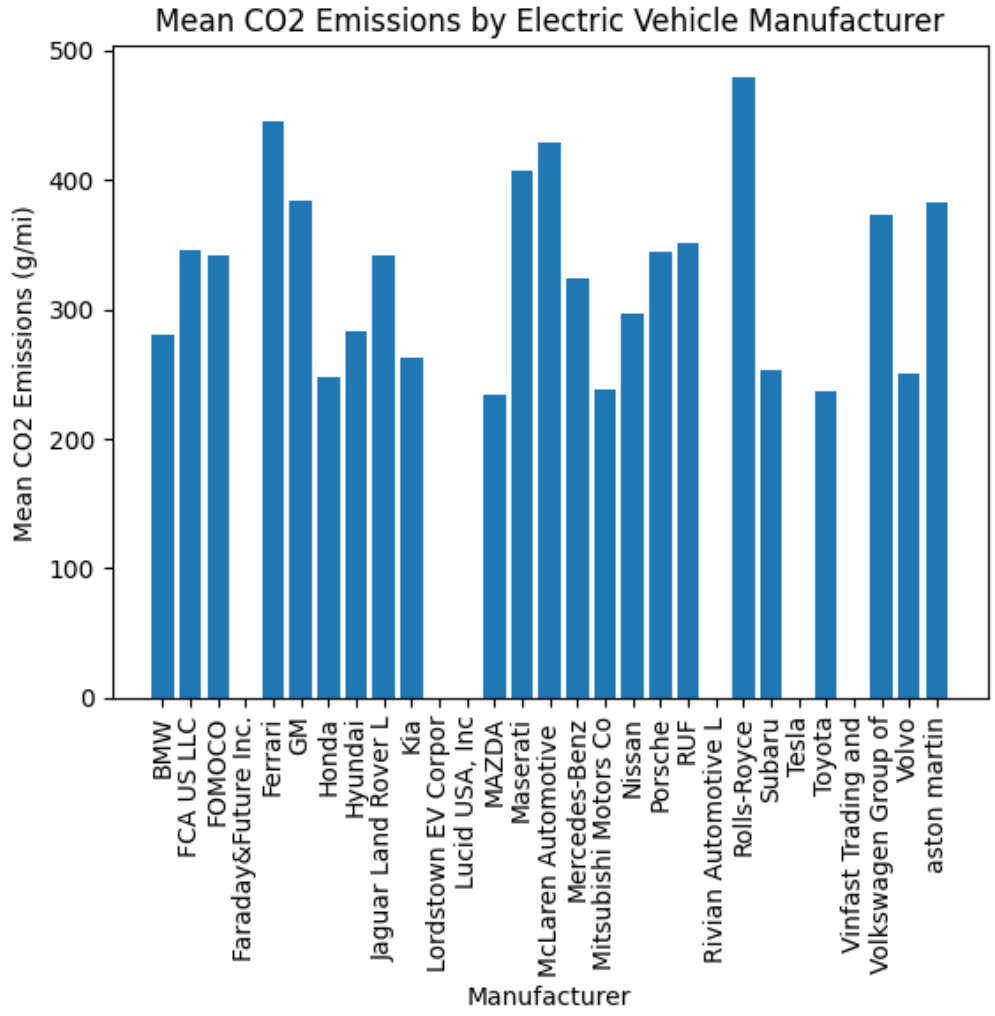


Description: I created a box plot for the whole data frame since it's a small dataframe and I chose a boxplot to throw out any outliers to get really accurate information. The columns for the data are all years from 1990 to 2020 and the rows are the different emission types that are released in the atmosphere.

Abanoub:

```
#plot data
```

```
data2 = pd.read_csv('cleaned_data.csv')
data2 = data2[['Vehicle Manufacturer Name', 'CO2 (g/mi)']]
new_mean = data2.groupby('Vehicle Manufacturer Name').mean()
# bar chart
plt.bar(new_mean.index, new_mean['CO2 (g/mi)'])
plt.xticks(rotation=90)
plt.xlabel('Manufacturer')
plt.ylabel('Mean CO2 Emissions (g/mi)')
plt.title('Mean CO2 Emissions by Electric Vehicle Manufacturer')
# show the plot
plt.show()
```



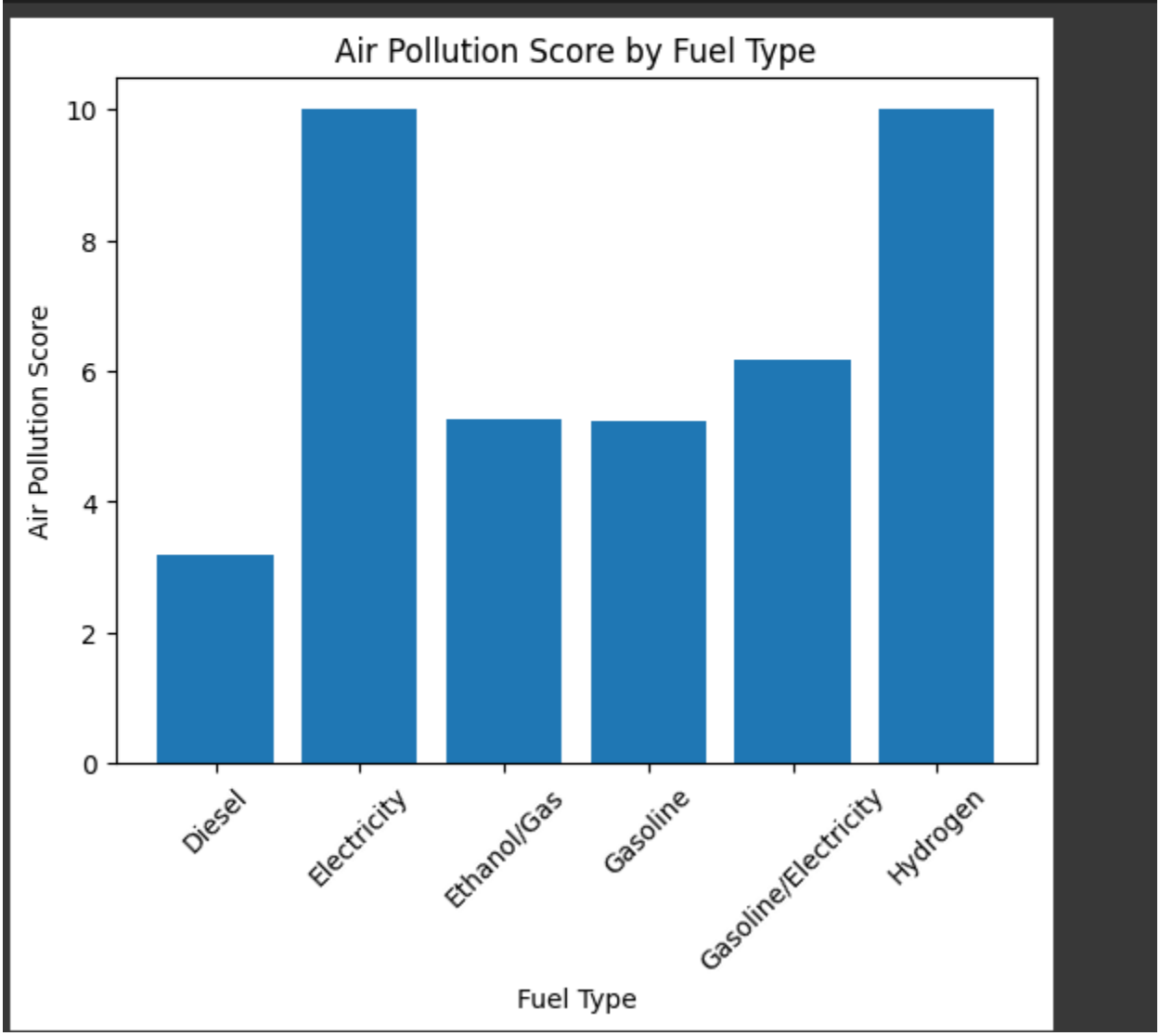
I choose to compare manufacturers and CO2 emissions because it can provide insights into which manufacturers are producing vehicles with lower CO2 emissions. This information is important because comparing manufacturers can help identify trends and patterns in the industry as a whole, such as which types of vehicles are associated with

lower emissions.

Hamza:

```
fuel_data = df.groupby('Fuel')['Air Pollution Score'].mean().reset_index()

plt.bar(fuel_data['Fuel'], fuel_data['Air Pollution Score'])
plt.xlabel('Fuel Type')
plt.ylabel('Air Pollution Score')
plt.title('Air Pollution Score by Fuel Type')
plt.xticks(rotation=45)
plt.show()
```



In this bar graph is a group of each fuel type that a vehicle uses and gets the mean of each fuel type and the air pollution score which the higher the score, the little it affects the environment. Based on this graph, cars that use electricity or hydrogen have little

effect on the environment while Diesel has a great impact on the environment.

Model planning

Data #1: Bureau of Transportation Statistics (BTS) website

- This table will be used for the K-means machine learning task, which aims to group similar vehicles based on their emission rates.

Data #2: U.S. Department of Energy Fuel Economy

- This table will be used for the classification machine learning task, which aims to predict the fuel type of a vehicle based on its fuel economy and other attributes.

Data #3: Greenhouse Gas Inventory Data Explorer of the U.S. Environmental Protection Agency

- The table will be used for the clustering machine learning task, which aims to group similar sectors based on their greenhouse gas emissions.

Data #4: Testing Fuel Economy

- This table will be used for the regression machine learning task, which aims to address the environmental impact of vehicles based on their efficiency and other attributes.

Reflection

The hardest part of the project has been to find the applicable data related to our problem statement. It seems that when researching for topics, we found bits and pieces of what we are looking for but never the exact measurements that we are looking for in order to compare EVs versus gas powered cars. In order to get around this problem, we have to look for other kinds of data that infer what we are looking for and compare multiple csv files in the hopes of finding a correlation. For this reason, we are not able to find concrete results related to our problem statement. It also means that we are not on the timeline that we expected with our project in order to finish on time. We would need to go back and find more data or change our problem statement a bit to something that can be discovered while using the data that we have collected so far.

Next Steps

For the next steps we plan on first looking for correlations between the existing data that we have in hopes to make progress in answering the question of our problem statement. If that doesn't work, we plan on looking for new data on the internet that can help us instead of what we have already collected. The last resort of this would be to completely change our problem statement to something that can be found using what we have cleaned and collected so far in this project.