Final Report

Topic A

Mason Ballard
Cybersecurity
Florida State University
Tallahassee, Florida, USA
mgb15c@fsu.edu

Holly Jordan
Computer Science
Florida State University
Tallahassee, Florida, USA
hmj19a@fsu.edu

Chris Pierre Paul
Computer Science
Florida State University
Tallahassee, Florida, USA
cp20fl@fsu.edu

Chris Que
Computer Science
Florida State University
Tallahassee, Florida, USA
coq19@fsu.edu

Sophia Villalonga
Computer Science
Florida State University
Tallahassee, Florida, USA
scv18@fsu.edu

ABSTRACT

As electric vehicles (EVs) become more affordable and popular, there is a growing perception that they offer several benefits over traditional gasoline-powered cars. Some of these benefits include reduced environmental impact, increased cost-effectiveness, and improved reliability in terms of fuel economy and convenience. In this paper, we will analyze a variety of data sources and methods to quantify these perceived advantages and examine their potential limitations and disadvantages. We are going to answer the question: to what extent are electric vehicles more environmentally sustainable, cost-effective, and reliable for a consumer than traditional gasoline-powered cars? We will accomplish this by comparing the environmental impact, cost-effectiveness, fuel economy, and convenience of EVs to that of gas-powered vehicles. We completed our research according to the data science process. After asking questions and getting the data, we explored the data through data cleaning, visualization, and EDA. We then modeled the data through regression, classification, and clustering models. Lastly, we communicated and visualized the results using evaluation metrics, model performance, and the overall final results. By doing this, we aim to provide a comprehensive understanding of the current state of the EV market, its growth prospects, as well as the advantages and disadvantages of EVs compared to gas-powered vehicles. Our analysis will draw on a range of data sources, including industry reports, government statistics, and academic research. All members of our team contributed equally to this project.

1 ASK AN INTERESTING QUESTION

Each year almost 80 million motor vehicles are produced globally. In 2022, the global sale of electric vehicles rose by 60%, accounting for about 10 million sales. In that same year, one in seven cars sold globally was electric [4]. The first electric vehicle, or EV, that was mass-produced was introduced in 1996 by General Motors [6]. Now, there are over 40 major manufacturers that produce electric cars, and that number is projected to rise. There are many reasons for the increase in EV sales across the world. One perceived advantage is that EVs are more environmentally sustainable compared to cars that depend on fossil fuels. According to the United Nations, fossil fuels are the largest contributor to climate change [2]. It is believed that EVs emit significantly fewer pollutants and greenhouse gases than traditional gasoline-powered cars, which can help to reduce the overall carbon footprint. Another perceived advantage is that EVs are more cost-effective. Electric vehicles have lower fuel and maintenance costs compared to gasoline-powered cars which can be significant over the lifetime of the car. EVs are also perceived to be more reliable than traditional cars due to their simpler design and fewer moving parts.

The question we are asking is: to what extent are electric vehicles more environmentally sustainable, cost-effective, and reliable for a consumer than traditional gasoline-powered cars? This question was motivated by the many perceptions surrounding electric vehicles and the desire to gain a better understanding of their overall performance and benefits compared to traditional gasoline-powered cars. By following the data science process, we can provide valuable insights for consumers, policymakers, and the automotive industry on the potential of electric vehicles to address key challenges in global transportation. Analysis of this question can also help to inform future research and development efforts in the electric vehicle industry, leading to further improvements in their sustainability, affordability, and reliability.

Our motivation to research this topic is rooted in our desires of reducing our environmental impact and saving money. With the ever-increasing costs of living, education, and fossil fuels, it is vital that we find ways to cut expenses wherever possible. As college students, we wanted to research a vehicle that can save us money for the entire lifetime of the car. Additionally, we wanted to research the best vehicle to reduce our impact on the environment. With the future in mind, we wanted to find a vehicle to help ensure that future generations can inherit a cleaner, healthier planet. We believe that by exploring the extent to which electric vehicles can save money and reduce environmental impact, we can make informed decisions about our future transportation choices.

2 GET THE DATA

We used datasets that correlate to our overall objective, which pertain to the difference between EV's and gasoline-powered cars in terms of environmental impact, cost-effectiveness, efficiency, and reliability.

One limitation of our research is the availability of datasets and the features measured on which to investigate each section of our research question in addition to the rapid growing nature of the industry leading to volatility and variability of the data depending on when it was collected and using what methods.

2.1 Sources of Data

Our sources of data:

- **CO2 Emissions From Cars**
  - Source: European Environment Agency | Contains data from EU member states
  - Data about CO2 Emissions collected according to the New European Driving Cycle (NEDC) protocol
  - https://www.kaggle.com/datasets/vivovinco/monitoring-of-co2-emissions-from-passenger-cars
- **Car Price Dataset**
  - Data based on different aspects of the cars (i.e fuel economy, engine information)
  - https://www.kaggle.com/code/dronax/car-prices-dataset
- **Alternative-Fuel-Vehicles US**
  - Data based on different aspects of AFV vehicles (ie. fuel economy, fuel type)
  - https://www.kaggle.com/datasets/saketpradhan/alternative-fuel-vehicles-in-the-us
- **Crude-Oil Production**
  - Yearly production value of Crude Oil from 1983 to 2022
  - https://www.macrotrends.net/2562/us-crude-oil-production-historical-chart#:~:text=Interactive%20historical%20chart%20showing%20the%20monthly%20level%20of,February%202022%20is%2011%2C600.00%20thousand%20barrels%20per%20day
- **Monthly-Electricity Statistics**
  - Yearly net electricity production from various sources of nonrenewable energy/other
  - https://www.iea.org/data-and-statistics/data-product/monthly-electricity-statistics
- **Global-Ev-Outlook: EV stock, sales, number of charging points (fast/slow)**
  - https://www.iea.org/data-and-statistics/data-product/global-ev-outlook-2022
- **Number-of-gasoline-stations US**
  - The number of total gas stations in US from 2014-2017
  - https://www.statista.com/statistics/525107/number-of-gasoline-stations-in-the-united-states/
- **Electric Vehicle Ownership Costs: Today's Electric Vehicles Offer Big Saving for Consumers**
  - Dataset describing top picks in ICE vehicles, and different types of EVs including the cost of each, the cost of ownership, operation, maintenance, and fuel cost differences across states
  - https://advocacy.consumerreports.org/wp-content/uploads/2020/10/EV-Ownership-Cost-Final-Report-1.pdf
- **Electric Alternative Fuel Vehicles US in 2023**
  - Dataset describing energy efficiency rate in gas-powered vehicles vs electric vehicles
  - https://www.kaggle.com/datasets/saketpradhan/alternative-fuel-vehicles-in-the-us

3 EXPLORE THE DATA

We each examined datasets pertaining to our selected area of the research question, focusing on environmental impact, cost-effectiveness, and efficiency. We compared relevant data for each focus comparing between gas-powered vehicles and EVs.

To perform our research analysis, we used data from the datasets we found, shown above in section 2.1. Our steps to clean, preprocess and explore the data for each topic of our research question is shown below.

4 MODEL THE DATA

Similar to above, we each created a model based on our chosen area of focus. We will show the specific analysis below, broken up by topic.

Section 5 - Environmental Impact
Section 6 - Cost-Effectiveness
Section 7 - Reliability - Fuel Economy
Section 8 - Reliability - Convenience

5 ENVIRONMENTAL IMPACT

5.1 About

To compare the environmental impact of electric vehicles versus vehicles that run on other types of fuel, we examined the difference in CO2 emissions based on fuel type, from the "CO2 Emissions from Cars" dataset. This dataset used contains +1 million rows and 33 columns. The data was collected from the European Environmental Agency and contains information like the car's manufacturer and vehicle type. It also included things like the engine's power and capacity as all these affect CO2 emissions.

5.2 Explore the Data

Our initial data preparation, exploration, and analysis allowed us to gain a general understanding of the different columns in the data that would be useful for comparing environmental impact between different types of vehicles.

With this information, we loaded the dataset into a pandas dataframe. To make the analysis more efficient, we only loaded the data from the csv file into the dataframe once, and saved it as a binary file using the pandas' function `to_pickle()`, in a file called `setup.py`. After we ran this code, we were able to load the dataframe from the binary file, using pandas' function `pd.read_pickle()` in our main file, called `environment.py`, making our analysis considerably faster and more efficient.

`setup.py`:

```python
import pandas as pd

# Save dataframes as a binary file
df_co2 = pd.read_csv("data/CO2_passenger_cars2.csv", low_memory=False)
df_co2.to_pickle("pickles/df_co2.pkl")
```

`environment.py`:

```
df_co2 = pd.read_pickle("pickles/df_co2.pkl")
```

After importing dataframe from the binary file, we determined that the most useful columns in this dataset pertaining to our specific focus on environmental impact are the columns named `Enedc (g/km)`, and `Ft`. These columns contain data that represent the CO2 emissions in grams per kilometer, as well as the fuel type of each vehicle, respectively.

In order to handle missing values, we calculated the median of the `Enedc (g/km)` column, and the mode of the `Ft` column. Then, we filled in the missing values in these columns with the median and mode that we just calculated, respectively.

```
# Calculate the median and mode
median = df_co2["Enedc (g/km)"].median()
mode = df_co2["Ft"].mode()[0]

# Fill the NaN values with the median / mode
df_co2["Enedc (g/km)"] = df_co2["Enedc (g/km)"].fillna(median)
df_co2["Ft"] = df_co2["Ft"].fillna(mode)
```

Next, we formatted the data in each columns to make the results easier to read and understand. We removed extraneous spaces from both columns. Then, we capitalize the first letter of each fuel type and isolated the unique fuels. This is because multiple instances of the same fuel type were being represented as distinct fuel types due to inconsistent formatting. For instance, "ELECTRIC" and "electric" were being counted as two different fuel types, so to fix this we standardized the format to have the first letter capitalized. This correctly grouped all unique fuel types together to give us accurate results.

```
# Format the data to remove extraneous spaces
df_co2["Enedc (g/km)"] = df_co2["Enedc (g/km)"].apply(
    lambda x: x.strip() if isinstance(x, str) else x
)
df_co2["Ft"] = df_co2["Ft"].apply(lambda x: x.strip() if isinstance(x, str) else x)

# Capitalize the first letter and get unique values
df_co2["Ft"] = df_co2["Ft"].apply(lambda x: x.capitalize() if type(x) == str else x)
```

Before the data preprocessing, the fuel types contained in the fuel types column were:

```
['DIESEL' 'PETROL' 'ELECTRIC' 'Petrol' 'Diesel' 'LPG'
 'Petrol/Electric''Electric' 'PETROL/ELECTRIC'
 'petrol' 'diesel'
 'Diesel/Electric''NG-biomethane' 'NG-BIOMETHANE'
 'electric' 'E85' 'Hydrogen' 'OTHER']
```

After this data preprocessing, the resulting ten fuel types were:

```
['Diesel' 'Petrol' 'Electric' 'Lpg' 'Petrol/electric'
 'Diesel/electric' 'Ng-biomethane' 'E85'
 'Hydrogen' 'Other']
```

Then, we created a new dataframe containing on the emissions and fuel type columns, to simplify subsequent analysis and model building.

```
# Create the stripped down dataframe
co2 = df_co2[["Enedc (g/km)", "Ft"]]
```

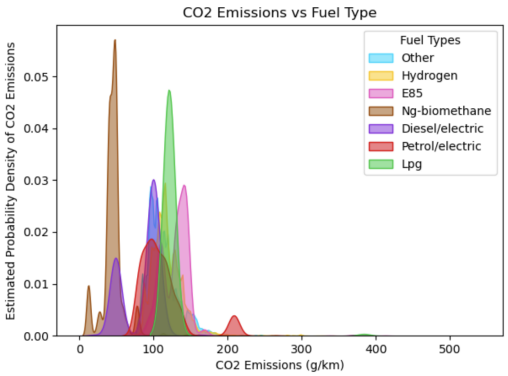
Next, in order to prepare the data for use in a K-Mean clustering and Linear Regression model, we used a dictionary to encode each of the ten formatted fuel types as an integer 1-10.

```
# Create the fuel types dictionary
fuel_dict = {
    "Other": 1,
    "Hydrogen": 2,
    "E85": 3,
    "Ng-biomethane": 4,
    "Diesel/electric": 5,
    "Petrol/electric": 6,
    "Lpg": 7,
    "Electric": 8,
    "Petrol": 9,
    "Diesel": 10
}
# Replace the fuel type strings with their corresponding integers
co2.loc[:, "Ft"] = co2["Ft"].replace(fuel_dict)
```

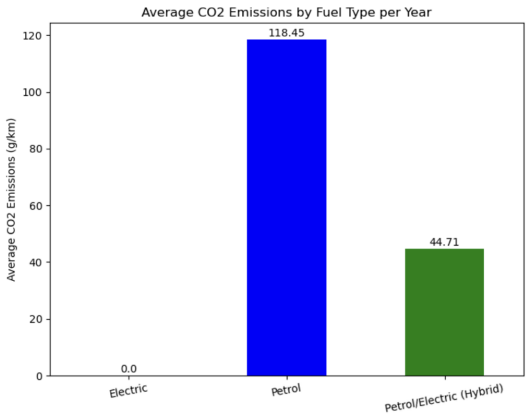
This was a way of manually encoding the categorical variables in the fuel type column as continuous variables. We found this way easier to keep track of the different fuel types for later visualizations of the data. This method of using a dictionary to store the fuel types allowed us to later display the string representations of the integer fuel types on any resulting graph later. We did this by iterating through a reverse of this dictionary. This improved the readability of the graphs, since the audience would not know which fuel type corresponds to which integer if this step was not performed.

```
fuel_dict_rev = {v: k for k, v in fuel_dict.items()}
fuel_types = [fuel_dict_rev[i] for i in range(1, 11)]
...
plt.legend(title="Fuel Types",labels=fuel_types)
```

To further explore the data now that it has been cleaned, we decided to make a density plot to show the distribution of CO2 emissions for each fuel type, and how they compare to each other. The density plot is shown below.



There is overlap with most of the fuel types but "Petrol/electric" and "Hydrogen" have the least estimated density variation. However, as you can see, some fuel types were not shown on the density plot. They include "Electric," and "Petrol". This shows us that for every entry of that fuel type in the dataset, the CO2 emissions are exactly the same, thus there is no variation to plot on the density graph. To visualize these CO2 emissions, we created a bar graph, and added "Petrol/electric" to compare the values. The bar graph is shown below. An interesting result visualized in the bar graph is that electric vehicles have zero CO2 emissions.



Cleaning and preparing the 'CO2 Emissions From Cars' dataset greatly simplified subsequent building of the K-Means Clustering and Linear Regression models, which will be explained in Section 4 of this report. Additionally, these models will allow us to further explore the environmental impact of electric vehicles.

5.3 Model the Data

In recent years, there has been a growing interest in electric vehicles (EVs) due to their potential to significantly reduce CO2 emissions compared to traditional gasoline-powered cars. Additionally, EVs do not emit any greenhouse gases during operation, whereas gas-powered cars emit a considerable amount of CO2, contributing to global climate change. However, it is worth noting that the CO2 emissions associated with an EV depend on the source of electricity used for charging. If the electricity is generated from renewable sources, such as wind or solar, the CO2 emissions of EVs are significantly lower than those of gas-powered cars. Overall, the decision to choose an EV or a traditional car depends on a range of factors, including the consumer's driving needs, location, and access to charging infrastructure.

To further explore this topic, we built a K-Means Clustering model to see CO2 emissions from different fuel types of cars.

K-Means Clustering Model

In this model, we used three clusters ($k = 3$) to group the cars into 3 distinct categories based on their carbon emissions and fuel type.

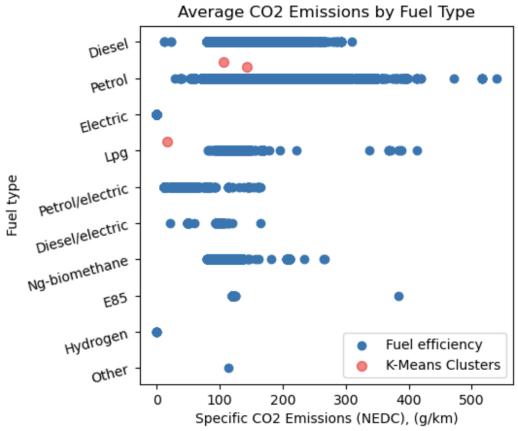
To build the model, we followed this procedure:

(1) Ran the K-Means Clustering algorithm on the dataframe
(2) Found the cluster centers
(3) Create the plot and scatter plot
(4) Add the cluster centers to the plot
(5) Perform various formatting tasks on the plot to ensure the plot is easy to understand
(6) Show the plot and interpret results

This procedure is shown in the code below.

```
# ***** K MEANS CODE *****
# Run K-Means Clustering Algorithm with k=3
kmeans =
↪  KMeans(n_clusters=3,n_init=10).fit(co2[["Enedc
↪  (g/km)", "Ft"]])
centers = kmeans.cluster_centers_

# Create plot and scatter plot
fig, ax = plt.subplots()
ax.scatter(x=co2["Enedc (g/km)"], y=co2["Ft"],
↪  label="Fuel efficiency")

# Add the cluster centers to the plot
ax.scatter(centers[:, 0], centers[:, 1], c="red",
↪  s=50, alpha=0.5, label="K-Means Clusters")

# FORMATTING THE PLOT:
# Set the axis labels
plt.xlabel("Specific CO2 Emissions (NEDC), (g/km)")
plt.ylabel("Fuel type")
plt.title("Average CO2 Emissions by Fuel Type")
plt.legend(loc="lower right")
# Show string of fuel type instead of an integer 1-10
y_tick_labels = [key for key, value in
↪  fuel_dict.items() if value in range(1, 11)]
ax.set_yticks(range(1, 11))
ax.set_yticklabels(y_tick_labels)
y_tick_objs = ax.get_yticklabels()
for tick_obj in y_tick_objs:
    tick_obj.set_rotation(15) # Rotate the y-axis
    ↪  labels
fig.subplots_adjust(left=0.3)

plt.show() # Show the plot
```

The resulting K-Means Clustering plot is shown below.



Interpreting the K-Means Model

The red dots on the graph represent the three centroids of the three clusters that were found by the k-means algorithm. Each centroid represents the mean values of the emissions and fuel type of all the data points assigned to its respective cluster.

Visualizing these centroids on a plot are useful to help identify patterns or trends in the data that might not be immediately apparent from viewing the scatter plot alone. In our plot, we can see that the resulting centroids are near the "Petrol", "Electric" and "Lpg" fuel types, and on the side of lower emissions. This means that these fuel types have the lowest emissions on average. This is in accordance with our discoveries from the initial data analysis in section 3.

This is extremely important information because it can be used to inform and predict. This clustered data can be used to identify which gas types result in the lowest CO2 emissions and which have the highest. This data can be used by the consumer, manufacturer, and even policymakers. For many consumers, the environmental impact of their car is a very important subject. This chart and its clusters can help consumers decide on the types of cars that fit their lifestyle. This is also for the manufacturer to know so they can look into which cars produce the most CO2 and look for ways to limit these emissions. Polic makers can use this information as well to see how cars are effecting global warming and look to implement policies to make cars release a lower amount of hazerdous emissions.

6 COST-EFFECTIVENESS

6.1 About

Though not the most common reason for switching, arguably the second most common reason that people either consider or have switched to an EV is the rising price of gas over the years. Such a concern and subsequent response could be put more aptly as consumer interest in the fiscal impact reduction that might be brought about by using an electric vehicle over a gas-powered vehicle. From these pieces of information, we can see that consumers often assume the largest contributor to their obligate financial burden they take on when owning a car is how it is powered. After all, it is the one people deal with on a bi-weekly, weekly, or even more frequent basis. A majority of the time, electricity is cheaper than a gas tank fill-up depending on where you live and how you are charging if referring to an electric vehicle. But in totality, there are many different factors that contribute to the cost of a vehicle in both upfront costs and long term costs over the lifetime of a car. All of these must come together in order for a consumer to not only make the most responsible fiscal decision but also the most sensible decision for themselves as individuals- choosing the right type of car, what specific make, model, and year of the car, how they choose to insure it, the financing schedule they go with, the financial institution they borrow from, and more.

As we can see, figuring out if switching from a gas-powered vehicle to and electric vehicle is the right decision for a consumer financially in the short-term and long term is a multi-factored and complex evaluation. In relative comparison, the electric vehicle industry and the connected and associated industries (i.e. manufacturing, financing) are relatively young and lacking in the time, and investment to grow in magnitude, efficiency, and competition in the way that internal combustion engine (ICE) vehicles have. Because of this traditional gas-powered vehicles have been studied a great deal more with data that is numerous, varied, and spanning years in a way that can accurately reflect and predict trends on a variety of parameters. This is an area where electric vehicles of all kinds have to catch up. As such, we will briefly go over the available information regarding those factors less researched and therefore having less data to make accurate observations and predictions with, and after that, we will dive into the data we have, analyzing the available, up to date, and verifiable information regarding electric vehicles on their upfront costs, comparing new vehicles and their normalized operation and maintenance costs and how this affects users' lifetime ownership costs based on the vehicle type they choose.

One of the most publicized reasons that consumers may consider electric vehicles include the financial rebates and tax credits. Available federal tax credits that vehicles in the dataset qualify for are taken into account in this research; however, the Biden Administration has recently as of March 2023 introduced new restrictions and allowances to the EV Tax Credit as part of the 2022 Inflation Reduction Act that restrict what vehicles this tax credit can be applied to what and what consumers make take advantage of it [8]. Data on these restrictions and particularly just how many consumers could or already do qualify, for how much, currently utilization rates, and how the transition period of these new restrictions from the old ones will affect each of these things is lacking. Our research and data analysis follows the rules of the new restrictions, but it because they are so particular it seemed useful to quickly go over those restrictions to both educate about how it works as well as highlight the breadth and relatively narrowness of applicability that these restrictions have caused.

Rules regarding income limits, restrictions on which countries the car can be manufactured in (North America), battery component sourcing and size restrictions, and tax debt minimum requirements - all of these factors come together to give certain consumers federal rebates of up to $7,500 in the form of two payments of $3,750 if they choose the correct car. Buyers that qualify include those that have a household income level below $300,000, have a head of household income level less than $225,000, or an individual income level below $150,000 who also owe at least $7,500 in taxes since the rebate comes in the form of a tax credit [3]. Going forward developments are being made to make it where all of this can be

checked and processed at the dealership or at time of purchase to get an instant discount off of the purchase price for the buyer. Once a buyer knows that they qualify, they must then choose a qualifying vehicle, a list of which is on the IRS's website or can be checked via a VIN checker which will help streamline the process for seeing if a vehicle is made in the right year, manufactured in eligible countries, if battery components were sourced from eligible countries and has the right specifications, and if the car under consideration is marked below the price caps. Those still unsure if they qualify or how to get the best bang for their buck are best advised to ask a financial advisor about their specific situation.

Though these qualifications may seem to disqualify many consumers and many vehicles, there are still some options for discounts if a buyer either does not qualify or their chosen car doesn't qualify. There are rebates and credits at the federal, state, and local level for new, used, and leased vehicles. There are rebates and credits available for charging station installations for first time buyers up at $1,000 or 30% of the purchase price [8] though the details of who qualifies, for how much, and how to apply can vary by state. If financing the vehicle, there are certain financial institutions that offer loans geared toward borrowers seeking to buy or invest in a green vehicle offering lower interest rates on these loans. Consolidated data on these rebates, their availability, their amount, and their utilization is lacking, so consumers are currently best served researching each of these facets and how all of these rules apply in their locality.

The next step after purchasing a vehicle and sometimes before is buying insurance. On the whole insurance tends to be more expensive for electric vehicles over gas-powered vehicles due to the higher MSRP and typically higher standard repair costs. However, on the more minute and granular level insurance premium costs vary due to a series reasons. These include but are not limited to the model of the car; the year and condition of the car; the location; the age, driving, and credit history of the driver; and the company of issuance. All of this put together and qualified by things such as price shopping and common discounts like bundling, good student, claims-free, and paperless billing which most companies offer, an electric vehicle can either be much more or less than the national average. Take for example, the Chevy Bolt, an electric vehicle which typically has premiums below the national average, when compared to the Tesla Model X, a luxury electric vehicle whose typical insurance premium is more than double the national average[9]. Data widely available regarding insuring vehicles is lacking as well in addition to the fact that premiums and rates can fluctuate a great deal while also being very subjective to each individual situation. Due to the other incentives available and the ability to shop around for a better premium and rate, automobile insurance should be a consideration when buying and not a deterrent, especially as costs are expected decrease as the industry grows and improves which is likely as evidenced by the reductions and increased accessibility already observed.

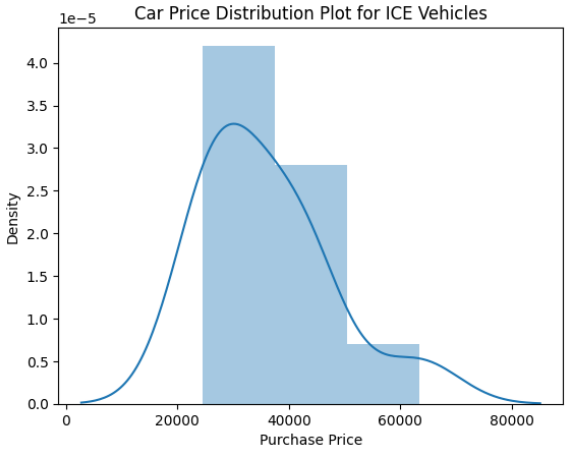
6.2 Explore the Data

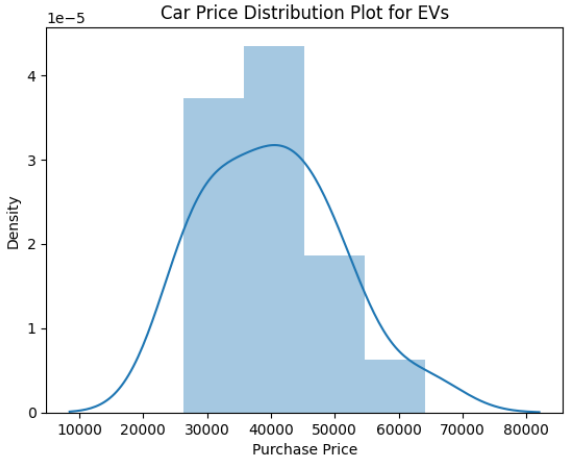
In order to analyze the cost effectiveness of electric vehicles (EVs) over internal combustion engine (ICE) vehicles on the macro level, data was pulled from Consumer Reports Electric Vehicle Ownership Report [5] which included factors such as the purchase price (standardized by the MSRP), fuel efficiency (re. ICE vehicles, HEVs, and PHEVs), electric efficiency (re. BEVs, HEVs, PHEVs), electric range (in miles, re. BEVs, HEVs, PHEVs), and the normalized operation and maintenance costs (which includes things like fuel costs, battery charges, oil changes, battery replacements, etc.) for new cars. Also included were the costs per year for gas and battery charges, aggregating the average savings possible to consumers when separated and compared by state. Vehicles included in the data included a cross selection of the most efficient, best-selling, top-rated, and top in performance vehicles for ICE vehicles, Battery Electric Vehicles (BEVs), Plug-in Hybrid Electric Vehicles (PHEVs), and traditional Hybrid Electric Vehicles (HEVs). This data was cross validated using websites like Kelly Blue Book, Car Edge, and car manufacturers' websites. A MySQL database and schema was maintained on a local machine, instantiated using MySQL Workbench, and accessed using the mysql-connector module from Python. The original purchase price data was first updated to reflect the most up to date prices, and a few pieces of missing data were researched using the sources mentioned before and manually imputed where necessary. No normalization was needed since the original source already did that. From there matplotlib's pyplot was used to plot any future data and findings.

`distplot.py`:

```
import pandas as pd
import matplotlib.pyplot as plt
import mysql.connector
import seaborn as sns

# connect to db
cnx = mysql.connector.connect(user='root',
↪  password='*****',
                              host='127.0.0.1',
                              database='vehicles')

# instantiate cursor to interact with db
cursor = cnx.cursor()

select = None
try:
    cursor.execute('SELECT purchase_price FROM
    ↪  cars_v2 where fuel_type like "ICE";')
    select = cursor.fetchall()
except:
    pass

# put purchase price data for ICE vehicles in
↪  dataframe
ice = pd.DataFrame(select, columns=['Purchase
↪  Price'])

# repeat for EVs
select = None
try:
    cursor.execute('SELECT purchase_price FROM
    ↪  cars_v2 where fuel_type like "%EV";')
    select = cursor.fetchall()
```

```
    except:
        pass
    ev = pd.DataFrame(select, columns=['Purchase Price'])

    ice['Purchase Price'] = pd.to_numeric(ice['Purchase Price'])
    ev['Purchase Price'] = pd.to_numeric(ev['Purchase Price'])

    # create distplots to illustrate how many cars are priced at different prices
    plt.title('Car Price Distribution Plot for ICE Vehicles')
    sns.distplot(ice['Purchase Price'])
    plt.show()

    plt.title('Car Price Distribution Plot for EVs')
    sns.distplot(ev['Purchase Price'])
    plt.show()

```

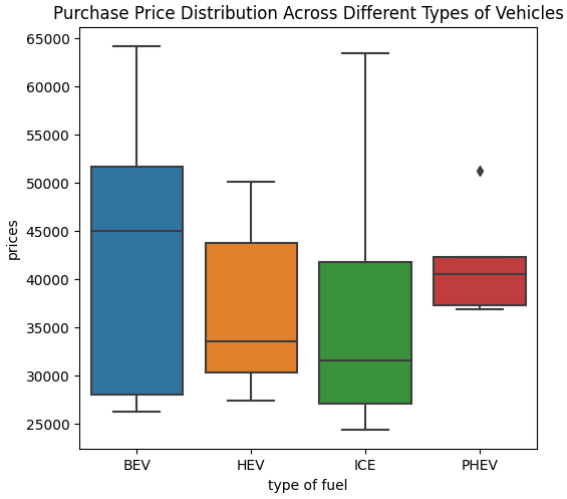




Initially plotting the distribution curve of ICE vehicles' and then EVs' purchase prices, we can compare how these two classes of cars are comparatively priced. We can see that the price spread of both is comparable with the ranges being nearly perfectly overlapping, there are no apparent outliers, and they maintain similar shapes which suggests to us that the pricing structure is similar. We will verify this and go into greater detail about this later using linear regression.

boxplot.py:

```
    select = None
    try:
        cursor.execute('SELECT fuel_type, purchase_price FROM cars_v2;')
        select = cursor.fetchall()
    except:
        pass

    data = pd.DataFrame(select, columns=['Type of Fuel', 'Purchase Price'])
    data.head()

    data['Purchase Price'] = pd.to_numeric(data['Purchase Price'])

    plt.figure(figsize=(20, 12))

    plt.subplot(2,3,1)
    boxplt = sns.boxplot(x = data['Type of Fuel'], y = data['Purchase Price'], data = data)
    boxplt.set(xlabel='type of fuel', ylabel='prices')

    plt.title('Purchase Price Distribution Across Different Types of Vehicles')
    plt.show()

```



Plotting the price spread of our dataset, separating by fuel type, we can see that battery operated electric vehicles have the largest price spread with vehicles priced just below the cheapest ICE cars in our dataset and vehicles that are the most expensive of the entire dataset. PHEVs in comparison have a much smaller price spread likely due to the fact that they are not the most common on the market at the moment and lack data in comparison with

the other type of vehicles, though we can see an outlier on this data for the class. Since it is not on or outside the outer bounds of the dataset, it should not affect calculations and models made later on. On the whole ICE vehicles are not the fuel category of cars with largest price spread nor the smallest. They are not the class with the most expensive vehicles and though containing the cheapest cars in the dataset, they are not significantly cheaper than the lowest of the class of BEVs. In this way, we can observe that traditional gas-powered vehicles may be a safe choice with their operation, maintenance, and upfront costs and how the value to cost ratio might shift over time being more common knowledge and that knowledge being part of this choice's safety, they are not guaranteed nor are they automatically the best deal that a consumer could find if they are willing to do a bit more research just based on the exploratory data analysis.

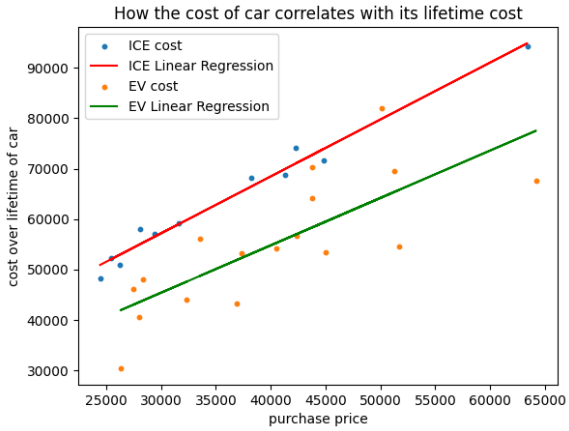
6.3 Model the Data

Modeling our data both confirms some prior observations made using standardized metrics and helps us observe some more key details about our dataset. First, the upfront cost of a car is correlated with the expected price of ownership over the lifetime of the car (estimated at 200,000 miles or 15 years). This means that we can fairly accurately predict the lifetime cost of a car based on its purchase price when the car is bought brand new which would also indicate to consumers that one of the biggest ways to save when buying a new vehicle is to pick the best deal (including incentives and financing deals where available and applicable) on it's upfront purchase price. This is shown below using a linear regression curve for both ICE vehicles and EVs plotted together for comparison.

`lin_reg.py`:

```
    # get ICE cost data
    ice_cost = """
    select
        `purchase_price`,
        (`purchase_price`+`lt_o_m`)
    from cars_v2 where fuel_type like 'ICE';
    """

    cursor.execute(ice_cost)
    select = None
    try:
        select = cursor.fetchall()
    except:
        pass

    if select != None:
        ice_cost = pd.DataFrame(select,
        ↪  columns=['Purchase Price', 'Lifetime
        ↪  Cost'])

    # get EV cost data
    ev_cost = """
    select
        `purchase_price`,
        (`purchase_price`+`lt_o_m`-`rebate`)
    from cars_v2 where fuel_type like '%EV';
    """

    cursor.execute(ev_cost)
    select = None
    try:
        select = cursor.fetchall()
    except:
        pass

    if select != None:
        ev_cost = pd.DataFrame(select,
        ↪  columns=['Purchase Price','Lifetime
        ↪  Cost'])

    fig, ax = plt.subplots()

    x = ice_cost[['Purchase Price']]
    y = ice_cost[['Lifetime Cost']]

    regression_model = LinearRegression()
    regression_model.fit(x, y)
    y_predicted = regression_model.predict(x)

    # model evaluation
    rmse = mean_squared_error(y, y_predicted)
    r2 = r2_score(y, y_predicted)

    # printing values
    print('---- ICE Car Data: -----')
    print('Slope:' ,regression_model.coef_)
    print('Intercept:', regression_model.intercept_)
    print('Root mean squared error: ', rmse)
    print('R2 score: ', r2)

    # data points
    ax.scatter(x, y, s=10, label='ICE cost')

    # predicted values
    ax.plot(x, y_predicted, color='r', label='ICE
    ↪  Linear Regression')

    # repreat for EVs
    x = ev_cost[['Purchase Price']]
    y = ev_cost[['Lifetime Cost']]
    regression_model = LinearRegression()
    regression_model.fit(x, y)
    y_predicted = regression_model.predict(x)
    rmse = mean_squared_error(y, y_predicted)
    r2 = r2_score(y, y_predicted)
    print('---- EV Car Data ----')
    print('Slope:' ,regression_model.coef_)
    print('Intercept:', regression_model.intercept_)
    print('Root mean squared error: ', rmse)
    print('R2 score: ', r2)
    ax.scatter(x, y, s=10, label='EV cost')
    ax.plot(x, y_predicted, color='g', label='EV
    ↪  Linear Regression')

    # Plot details
    ax.set_xlabel('purchase price')
    ax.set_ylabel('cost over lifetime of car')
```

```
    ax.set_title('How the cost of car correlates with its lifetime cost')
    ax.legend()
    plt.show()

```




```
---- ICE Car Data: -----
Slope: [[1.128639]]
Intercept: [23306.26757803]
Root mean squared error:  3570800.7823847416
R2 score:  0.9779424926234267
---- EV Car Data ----
Slope: [[0.9397916]]
Intercept: [17187.75711007]
Root mean squared error:  63452167.47145149
R2 score:  0.5891249674281578
```

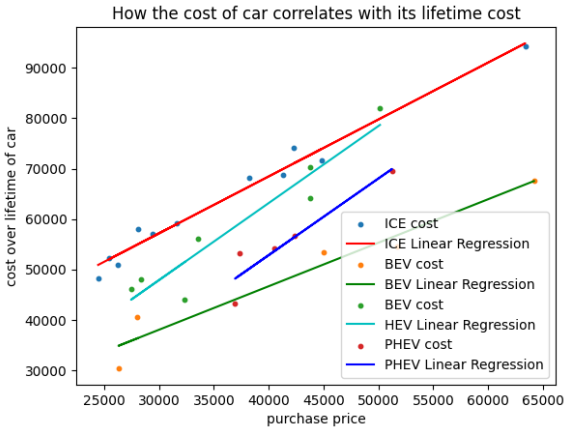
Linear regression shows us the previously mentioned correlation, how tight it is, and how accurate a predictor the purchase price of a car is for ICE cars and EVs respectively. With an R2 score of 0.97794, the correlation between upfront cost and ownership cost for ICE vehicles is highly correlated. This would reflect the consistency in pricing as well the reliability of predicting the ownership cost for the consumer when buying an ICE vehicle. On the other hand, our linear regression curve shows that EVs have an R2 score 0.58912 which while it still indicates a correlation, indicates the increased variance in pricing for EVs. This is likely partially due to the inclusion of federal rebates available for EVs as well as the relative youth of the industry which lacks the history and longevity that would lead to more competitive pricing and consistency. It should also be observed that the slope of the linear regression curve for ICE vehicles, 1.12864, when compared to EVs, 0.93979, is steeper which reflects that as ICE vehicles increase in purchase price, they increase in lifetime ownership costs, while EVs year over year maintain a more stable operation and maintenance cost.

To observe these details with greater granularity and observe how the different types of EVs compare to each other and to ICE vehicles in how upfront costs affect long term costs we can repeat the prior process but plot four linear regression lines rather than grouping the three types of EVs as a collective group.

`lin_reg_bytype.py`:

```
    # ICE DATA
    fig, ax = plt.subplots()
    x = ice_cost[['Purchase Price']]
    y = ice_cost[['Lifetime Cost']]
    regression_model = LinearRegression()
    regression_model.fit(x, y)
    y_predicted = regression_model.predict(x)
    rmse = mean_squared_error(y, y_predicted)
    r2 = r2_score(y, y_predicted)
    print('---- ICE Car Data: -----')
    print('Slope:' ,regression_model.coef_)
    print('Intercept:', regression_model.intercept_)
    print('Root mean squared error: ', rmse)
    print('R2 score: ', r2)
    ax.scatter(x, y, s=10, label='ICE cost')
    ax.plot(x, y_predicted, color='r', label='ICE Linear Regression')

    # BEV DATA
    bev_cost = """
        select
            `purchase_price`,
            (`purchase_price`+`lt_o_m`-`rebate`)
        from cars_v2 where fuel_type like 'BEV';
    """
    cursor.execute(bev_cost)
    select = None
    try:
        select = cursor.fetchall()
    except:
        pass
    if select != None:
        bev_cost = pd.DataFrame(select, columns=['Purchase Price', 'Lifetime Cost'])

    # BEV PLOT
    x = bev_cost[['Purchase Price']]
    y = bev_cost[['Lifetime Cost']]
    regression_model = LinearRegression()
    regression_model.fit(x, y)
    y_predicted = regression_model.predict(x)
    rmse = mean_squared_error(y, y_predicted)
    r2 = r2_score(y, y_predicted)
    print('---- BEV Car Data: -----')
    print('Slope:' ,regression_model.coef_)
    print('Intercept:', regression_model.intercept_)
    print('Root mean squared error: ', rmse)
    print('R2 score: ', r2)
    ax.scatter(x, y, s=10, label='BEV cost')
    ax.plot(x, y_predicted, color='g', label='BEV Linear Regression')

    # HEV DATA
    hev_cost = """
        select
            `purchase_price`,
            (`purchase_price`+`lt_o_m`-`rebate`)
        from cars_v2 where fuel_type like 'HEV';
    """
    cursor.execute(hev_cost)
    select = None
    try:
        select = cursor.fetchall()
    except:
```

```
        pass
    if select != None:
        hev_cost = pd.DataFrame(select,
        ↪ columns=['Purchase Price', 'Lifetime
        ↪ Cost'])

    # HEV PLOT
    x = hev_cost[['Purchase Price']]
    y = hev_cost[['Lifetime Cost']]
    regression_model = LinearRegression()
    regression_model.fit(x, y)
    y_predicted = regression_model.predict(x)
    rmse = mean_squared_error(y, y_predicted)
    r2 = r2_score(y, y_predicted)
    print('----HEV Car Data: -----')
    print('Slope:' ,regression_model.coef_)
    print('Intercept:', regression_model.intercept_)
    print('Root mean squared error: ', rmse)
    print('R2 score: ', r2)
    ax.scatter(x, y, s=10, label='BEV cost')
    ax.plot(x, y_predicted, color='c', label='HEV
    ↪ Linear Regression')

    # PHEV DATA
    phev_cost = """
        select
            `purchase_price`,
            (`purchase_price`+`lt_o_m`-`rebate`)
        from cars_v2 where fuel_type like 'PHEV';
    """
    cursor.execute(phev_cost)
    select = None
    try:
        select = cursor.fetchall()
    except:
        pass
    if select != None:
        phev_cost = pd.DataFrame(select,
        ↪ columns=['Purchase Price', 'Lifetime
        ↪ Cost'])

    # PHEV PLOT
    x = phev_cost[['Purchase Price']]
    y = phev_cost[['Lifetime Cost']]
    regression_model = LinearRegression()
    regression_model.fit(x, y)
    y_predicted = regression_model.predict(x)
    rmse = mean_squared_error(y, y_predicted)
    r2 = r2_score(y, y_predicted)
    print('---- PHEV Car Data: -----')
    print('Slope:' ,regression_model.coef_)
    print('Intercept:', regression_model.intercept_)
    print('Root mean squared error: ', rmse)
    print('R2 score: ', r2)
    ax.scatter(x, y, s=10, label='PHEV cost')
    ax.plot(x, y_predicted, color='b', label='PHEV
    ↪ Linear Regression')

    # TABLE VARIABLES
    ax.set_xlabel('purchase price')
    ax.set_ylabel('cost over lifetime of car')
    ax.set_title('How the cost of car correlates with
    ↪ its lifetime cost')
    ax.legend()
    plt.show()
    plt.savefig('comp_all_reg.png')

```




```
---- ICE Car Data: -----
Slope: [[1.128639]]
Intercept: [23306.26757803]
Root mean squared error:  3570800.7823847416
R2 score:  0.9779424926234267
---- BEV Car Data: -----
Slope: [[0.86260583]]
Intercept: [12163.82997108]
Root mean squared error:  9619627.160275806
R2 score:  0.9410008587710839
----HEV Car Data: -----
Slope: [[1.52763269]]
Intercept: [2059.84143978]
Root mean squared error:  15729927.776301336
R2 score:  0.9078685640975681
---- PHEV Car Data: -----
Slope: [[1.51939746]]
Intercept: [-7935.77577795]
Root mean squared error:  8984987.651623055
R2 score:  0.8741413843172352
```

The first standout observation we can make from this data is that the correlation between initial cost and lifetime cost increases rather dramatically for all types of vehicles with all having an R2 score over 0.87. This shows us in contrast to the prior chart that purchase price and lifetime ownership and maintenance costs are consistent when separated by fuel type. It also can inform consumers that the type of electric vehicle does matter when trying to pick the most economical vehicle even if the consumer is sure that they want to go with an electric vehicle.

Putting that observation together with the price spread data shown in the box plot above, we can first observe the inverse relationship between the cost of battery-powered electric vehicles and their lifetime cost of ownership. With the largest price spread and the highest median price, around $45,000, battery electric vehicles are the only among the broader class of electric vehicles that are purely electric. Where ICE cars run only on gas or petrol, hybrid electric vehicles (HEVs) and plug-in hybrid electric vehicles (PHEVs) use a combination of gas and electricity which can result in an increase in fuel efficiency and less gas fill ups for the consumer. The effect of the car owner still needing gas and maintenance like oil changes affects the lifetime costs of these vehicles, and is reflected in the first chart above in the slope of the HEV linear regression line

(1.52763) and the PHEV linear regression line (1.51939) both being higher than that of ICE vehicles (1.12863). All of this compared to the slope of the linear regression line for battery electric vehicles (0.86260) shows how owning a car than runs purely on electricity can allow consumers to reap the full benefits financially if they are willing and able to make the upfront investment because the lifetime ownership costs for these kinds of vehicles consistently stays lower than that of ICE vehicles shown by the fact that the entire linear regression line of BEVs is below that of ICE vehicles and the significantly lower slope of the BEV linear regression line.

7 RELIABILITY - FUEL ECONOMY

7.1 About

In addition to the above factors, there is also a perceived increase in efficiency of EV's over gasoline-powered cars. This is due to the ability to charge an EV from anywhere there is a power source. However, on longer trips, it might be inconvenient to sit and wait for the vehicle to charge, while adding gas to a traditional car can be performed in a matter of minutes. With these factors in mind, are EV's really more efficient and reliable than gas-powered vehicles?

EV's are powered by batteries that convert electrical energy into motion. EV's are becoming an increasingly popular alternative to gas-powered vehicles, and their increased efficiency, better fuel economy, increased potential to reduce greenhouse gas emissions, and decreased reliance on fossil fuels are some of the key factors that are pushing this trend. EV's are more efficient in converting stored energy into motion than gas-powered vehicles because conventional gas-powered vehicles waste energy in the form of heat. This section of the paper aims to provide an overview of efficiency of electric vehicles, including the different factors that can affect it and ways that EV's might be able to be improve in the years to come.

Not only are EV's more efficient in converting stored energy into motion, but they are also typically equipped with a regenerative braking system that captures some of the kinetic energy lost from braking and stores it back into the batteries as electrical energy. This system further increases the overall efficiency in EV's. Gas-powered vehicles are less efficient than EV's because they have to use energy even while they are not in potion. For example, if a gas-powered vehicle is stopped at a red light, or is in idle while stuck in traffic, the vehicle is still converting the stored energy, meaning energy is wasted despite not being in motion. EV's only convert the energy it has stored when it is in motion. EV's produce zero CO2 emissions making this type of vehicle more environmentally friendly when comparing the amount of CO2 each type of vehicle emits while in use. The overall environmental impact of EV's is determined by where the electrical energy that is used to power them came from. If the electricity used to power EV's came from renewable resources such as, wind turbines or solar panels, the amount of emissions that EV's would produce would be zero. EV's are simply more environmentally friendly when compared to gas-powered vehicles.

Driving conditions pose a significant impact on the efficiency of EV's. Ev's are more efficient in the city due to the slower speeds and stop-and-go traffic, which allows the vehicle to utilize regenerative braking to recover some of the energy that has been lost when coming to a stop. In contrast, gas-powered vehicles do not perform as well as EV's in the city because of the stop-and-go traffic. Gas-powered vehicles use energy even when they are stopped, sitting in idle, even when they are slowing down. This is because fuel is constantly being consumed as long as the engine is on. This makes the fuel efficiency of gas-powered vehicles in the city very poor. Gas-powered vehicles perform well in the highway when they are maintaining a high speed and torque which makes them have better fuel efficiency in the highway. EV's do not perform as well on the highway because it takes a significant amount of energy to maintain a high speed, and with the limited ability to use the regenerative braking feature like it can in the city, the fuel efficiency of EV's drops on the highway.

Overall, EV's are more energy efficiency, environmentally friendly if the energy is sourced from renewable energy, and quickly becoming more popular as a primary mode of transportation. EV's are more efficient than gas-powered vehicles due to their ability to regenerate energy that would have normally been lost. While gas-powered vehicles use energy even when they are not moving, EV's only use energy while in motion. EV's have zero CO2 emissions and perform better in citys over highways. These are all reasons why EV's are becoming more and more popular.

7.2 Explore the Data

The dataset for electric cars from Kaggle ("Electric Alternative Fuel Vehicles US 2023") [7] had different columns that had information for the fuel/energy consumption rates for electric and gas-powered vehicles respectively. To properly analyze the data and make comparisons between the consumption rates of energy for the two types of vehicles, I had to import the pandas library to get tools to manipulate the data properly. I also imported the matplotlib to make proper visualizations to make the comparison easier to see.

```
import pandas as pd
import matplotlib.pyplot as plt

# Read the data from ElectricCarData_Clean.csv
df_ec = pd.read_csv("ElectricCarData_Clean.csv")

```

The next step was to find the average efficiency in watt hours per kilometer for electric vehicles. Since we are making a comparison for the average energy consumption versus the average fuel consumption, I also had to find the average fuel consumption.

```
# Calculate the average efficiency (in WhKm) from ElectricCarData.csv
avg_efficiency = df_ec['Efficiency_WhKm'].mean()

# Calculate the average fuel consumption (in mpg) from ElectricCarData.csv
avg_fuel_consumption = avg_efficiency['Fuel Consumption Comb (mpg)'].mean()

```

Since the two metrics are not equivalent, I needed to convert the rate of fuel consumption (mpg) to watt hours per kilometer

(WhKm). This was done using the following formula that has been written in python.

```
    # Calculate the energy consumption in kWh/100km
    energy_consumption = avg_efficiency / 1000 * 100

    # Calculate the fuel consumption in gallons/100
    ↪ miles
    fuel_consumption = 235.214 / avg_fuel_consumption

    # Calculate the GGE (Gasoline Gallon Equivalent)
    gge = fuel_consumption / 1.33

    # Calculate the difference in energy and fuel
    ↪ consumption
    energy_saving = fuel_consumption -
    ↪ energy_consumption

```

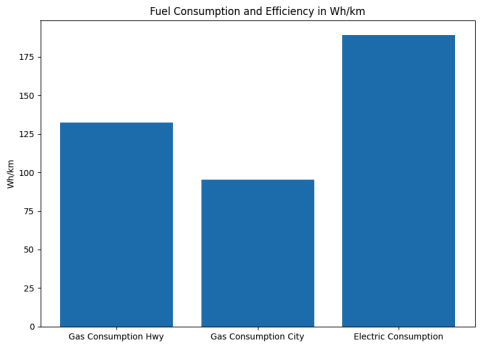
This next portion is what makes the data visualized and emphasizes the drastic difference in energy consumption rates between the two types of vehicles.

```
    # Plot the bar graph to compare the energy and
    ↪ fuel consumption
    fig, ax = plt.subplots()
    ax.bar(["Average energy consumption of EV's",
    ↪ "Average fuel consumption of ICEV's"],
    ↪ [energy_consumption, gge])
    ax.set_ylabel("GGE/100 miles")
    ax.set_title("Comparison of Electricity and Fuel
    ↪ Consumption")
    plt.show()

```

7.3 Model the Data

The following process was relatively similar to the process done before. Started by importing the necessary libraries, reading the data in, found the count for the fuel types, created a bar graph that would correspond with the number of vehicles per type.

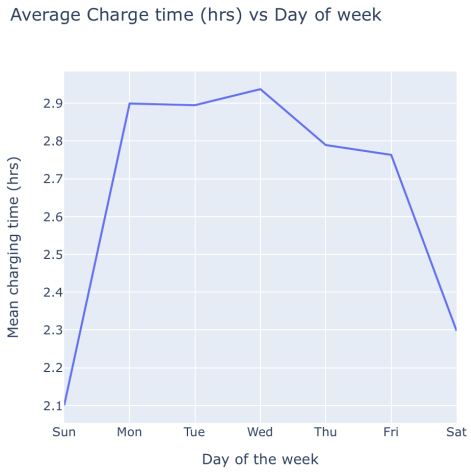
```
    import pandas as pd
    import matplotlib.pyplot as plt

    # Load the CSV data into a Pandas DataFrame
    df = pd.read_csv("AlternateFuelVehicles.csv")

    # Get a count of the number of vehicles for each
    ↪ fuel type
    fuel_counts = df["Fuel"].value_counts()

    # Create a bar chart of the fuel type counts
    plt.bar(["Hybrid Electric", "Electric", "Plug-in
    ↪ Hybrid Electric"], fuel_counts[["Hybrid
    ↪ Electric", "Electric", "Plug-in Hybrid
    ↪ Electric"]])
    plt.title("Number of Vehicles by Fuel Type")
    plt.xlabel("Fuel Type")
    plt.ylabel("Number of Vehicles")
    plt.show()

```



This graph above shows the energy consumption rates of traditional gas-powered vehicles, hybrid vehicles, and electric vehicles all converted to watt hours per kilometer. This was done to make the comparison between gas-powered vehicles and EV's more apparent. It is clear that vehicles that use internal combustion engines(ICE) convert less stored energy into motion compared to hybrid vehicles and electric vehicles. There are three values on the x-axis which are the gas consumption in the highway, gas consumption in the city, and electric consumption. The rate of consumption in the city versus the highway in gas-powered vehicles is drastically different than that of EV's. This highlights the benefits of EV's and their ability to convert more stored energy into movement compared to gas-powered vehicles. It also emphasizes the amount of CO2 emissions that would be saved from going into the atmosphere by emphasizing the amount of energy output EV's have.

```
Average Alternative Fuel Economy City: 84.59 mpg
Average Alternative Fuel Economy Highway: 77.59 mpg
Average Conventional Fuel Economy City: 25.90 mpg
Average Conventional Fuel Economy Highway: 30.09 mpg
```

The Kaggle dataset ("Electric Alternative Fuel Vehicles US 2023") was easy to clean due to the dataset having no missing values making the process of analyzing the data quicker. I was able to find that the average alternate (EV) fuel economy in the city and highway, and the average conventional fuel economy for the city and highway. From this image above, it is clear that alternative fuel has a significantly higher fuel economy than conventional vehicles. This further supports that EV's are significantly more efficient in converting stored energy into motion compared to gas-powered vehicles.



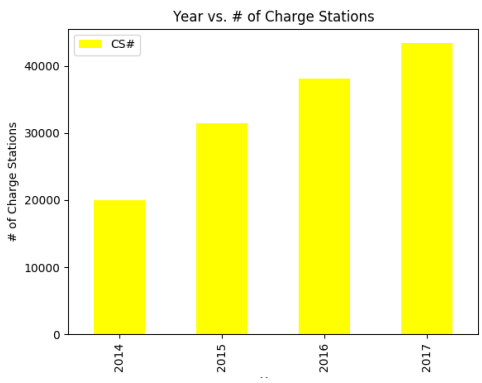
The image above shows the daily average charging times for EV users. We can see that it takes a significant amount of time to get a full charge, which is not a great characteristic if the user wants to be efficient with their time. We know it takes a short period of time to fuel a regular gas-powered car. Typically, it takes less than 5 minutes to get a full tank in a gas-powered vehicle, but in an EV it could take up to 3 hours. I originally planned to make a model to show the daily charging rates for EV's. In regards to making a prediction model, it requires two dependent variables, and although time spent charging is a dependent variable, days of the week is categorical which means it would not be able to be used to make a prediction model. For this reason the model for this portion was omitted.

8 RELIABILITY - CONVENIENCE

8.1 About

EVs are often thought to be more reliable than gasoline-powered cars. While analyzing the reliability of Electric vehicles, it is critical to consider the topics such as comparing the number of Charge Stations and Gas Stations analyzed to reach an appropriate conclusion.

8.2 Explore the Data

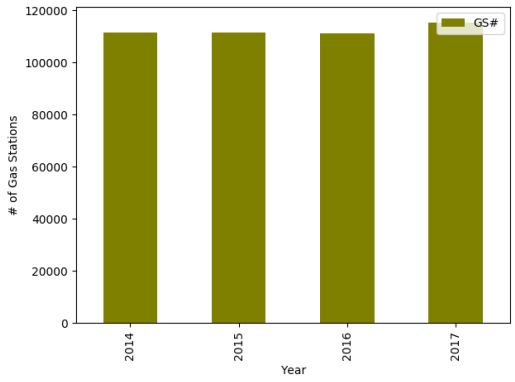


The IEA's ("Global-ev-outlook-2022") dataset [7] contained eight columns: the region, category, parameter, mode, power-train, year, unit, and value of various countries worldwide. As there were various countries, the dataset was first filtered only to contain data from the USA. In order to compare the difference between the number of Charge Stations, the number of fast chargers and the number of slow chargers were summed into one column value indicating the total of the two values per row. Before conducting EDA, the dataset was grouped by each year from 2014-2017 so that the total sum of charge stations (fast and slow) could compare to the total number of gas stations (both visualized).

```
#Reading in the file IEA-EV-data.csv
df = pd. read_csv ('IEA-EV-data.csv")

# Filtered the dataset to only contain data regarding the US
us = df['region'].isin(["USA"])
df = df[us]

# Filtered the dataset to only contain data regarding the US Charging Stations
cs = df ['parameter ].isin(["EV charging points"])
df = df[cs]

# Filter years to be from 2013-2017 for accurate comparison to other dataset
df = df[(df['year' ]
<= 2017) & (df['year' ] >= 2014)]

# Dropping unnecessary column
df = df.drop('mode', axis=1)
df = df .groupby ('year') .agg ({'value': 'sum '})
df .rename (columns={'value': 'CS#'}, inplace=True)
df.reset_index (inplace=True)

# EDA
print (df. shape.
df dtypes, df.isnull0.sumOr df.columns df.
describe df. head(10) sep='\n\n')
```

```
(4, 2)

year      int64
CS#     float64
dtype: object

year    0
CS#     0
dtype: int64

Index(['year', 'CS#'], dtype='object')

              year           CS#
count     4.000000      4.000000
mean   2015.500000  33250.000000
std       1.290994  10085.798597
min    2014.000000  20000.000000
25%    2014.750000  28625.000000
50%    2015.500000  34800.000000
75%    2016.250000  39425.000000
max    2017.000000  43400.000000

   year      CS#
0  2014  20000.0
1  2015  31500.0
2  2016  38100.0
3  2017  43400.0
```

During the process of EDA (the same methods are used for every dataset), the methods ".shape," ".dtypes," "isnull().sum()," ".describe()" to analyze any null values that were present, as well as the summary statistics of the consequential data frame.



Statista's ("The Number of Gasoline Stations in the US") dataset [1] was relatively easy to clean, considering it contained no missing values and was pre-filtered to contain the years 2014-2017). Prior to conducting EDA, the previous y value was the (GS#(Thousands))number of Gas Stations in Thousands, which was converted to the (GS#) number of Gas Stations by multiplying all values by 1000. Subsequently, EDA was performed using identical methods as prior.

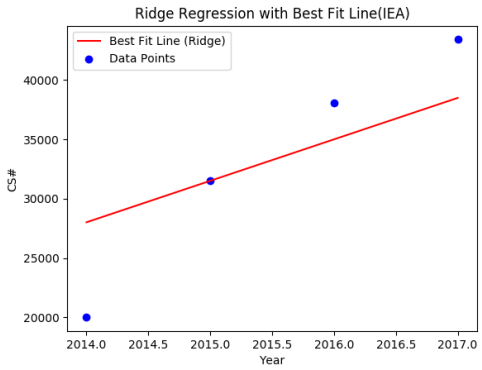
```
#Reading in the file
#statistic_id525107_number-of-gasoline-stations-
#in-the-united-states-2013-2017.xls
df = pd.read_excel('statistic_id525107_number-of-
↪ gasoline-stations
-in-the-united-states-2013-2017.xls'
sheet_name='Data', usecols='B:C', skiprows-4)

# I renamed the columns as the weren't read in
↪ correctly df.rename (columns={'Unnamed: 1':
↪ 'Year',
"Unnamed: 2': 'GS#(Thousands) '}, inplace=True)
#Convert y unit from GS#(Thousand) to GS# (*1000)
df['GS#*] = pd.to_numeric(df ['GS# (Thousands)'],
↪ errors='coerce") .astype(float) * 1000

# EDA
print (df.shape, df dtypes, df.isnull().sum(),
↪ df.columns, df.describe(), df. head(10),
↪ sep="\n\n")
```

```
(4, 2)

Year      int64
GS#     float64
dtype: object

Year    0
GS#     0
dtype: int64

Index(['Year', 'GS#'], dtype='object')

              Year            GS#
count     4.000000       4.000000
mean   2015.500000  112417.500000
std       1.290994    1982.395437
min    2014.000000  111100.000000
25%    2014.750000  111475.000000
50%    2015.500000  111600.000000
75%    2016.250000  112542.500000
max    2017.000000  115370.000000

   Year       GS#
0  2014  111600.0
1  2015  111600.0
2  2016  111100.0
3  2017  115370.0
```

8.3 Model the Data

```
# Train Test Split
X = df ['year' ].values. reshape (-1,1)
y = df [ 'CS#']. values
X_train, X_test, y_train, y_test =
↪ train_test_split(X,y, test_size=0.25,
↪ random_state=42)

# Initialize the Ridge Regression model
ridge = Ridge()

# Define the range of alpha values for
↪ hyperparameter tuning
alpha_range = p.arange(1, 11)

# Lists to store RMSE for different alpha values
rmse_scores = []

# Loop through each alpha value, fit the model,
↪ and calculate RMSE
for alpha in alpha_range:
```

```
    # Fit the Ridge Regression model with the current alpha value
    ridge.set_params(alpha=alpha).fit (X_train, y_train)

    # Make predictions on the test set
    y_pred = ridge.predict (X_test)

    # Calculate RMSE and Append the scores to the lists
    rmse_scores.append( np.sqrt (mean_squared _error (y_test, y_pred)) )

    # Find the index of the alpha value with the lowest RMSE
    best_alpha_idx = np.argmin(rmse_scores)
    best_alpha = alpha_range [best_alpha_idx]

    # Train the Ridge Regression model with the best alpha on the entire dataset
    ridge.set_params (alpha=best_alpha).fit(X, y)

    # Make predictions on the entire dataset
    y_pred_all = ridge.predict (X)

    # Plot the data points and the best fit line
    plt.scatter(X, y, color='blue', label='Data Points") plt.plot(X, y_pred_all, color='red', label='Best Fit Line (Ridge)') plt.xlabel('Year') plt.ylabel( 'CS#')
    plt. title('Ridge Regression with Best Fit Line(IEA)')
    plt. legend()
    plt. savefig('IEA_model. png')

    # Print the best alpha value, RMSE
    print('\nBest Alpha:', best_alpha)
    print ('MSE:
    rmse_scores [best_alpha_idx]**2)
    print('RMSE:' rmse_scores [best_alpha_idx])
```

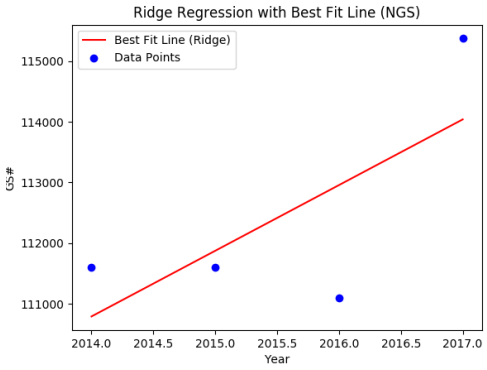



```
Best Alpha: 6
MSE:  39.062499988835847
RMSE: 6.249999999068677
```

In order to create a Ridge Regression model for the IEA's ("Global-ev-outlook-2022") dataset [7], the "year" and "CS#" (Charge Station number) columns were selected as the independent and dependent variables, respectively. Then, 25 percent of the data was split into test data, while the rest became train data fed to the Ridge Regression model. Subsequently, a for loop recreated hyper-parameter tuning that tested various alpha values from 1-10 (6 was the best alpha value for the dataset). The MSE (Mean Squared Error) and RMSE (Root Mean Squared Error) evaluated the Ride Regression model's performance by analyzing the model's prediction of the actual test data values. Under the alpha score of 6, the MSE was 39.06, while the RMSE was 6.249 (indicates the model was very accurate in predictions as it is deficient). In order to visualize the model, each point was graphed compared to the model's best-fit line (portrayed a positive slope meaning the number of charge stations increases at a steady rate of 1000 each year).

```
    # Train Test Split
    X = df[ 'Year']. values.reshape(-1, 1)
    y = df [ 'GS#']. values
    X train, Xtest, y_train, y_test = train_test_split(X,y, test_size=0.25, random_state=42)

    # Initialize the Ridge Regression model
    ridge = Ridge()

    # Define the range of alpha values for hyperparameter tuning
    alpha_range = np.arange(0, 10, 0.1)

    # Lists to store RMSE and R2 scores for different alpha values
    rmse_ scores = [

    # Loop through each alpha value, fit the model, and calculate RMSE and R2 scores
    for alpha in alpha_range:
        # Fit the Ridge Regression model with the current alpha value
        ridge.set_params(alpha=alpha)
        ridge.fit(X_train, y_train)

        # Make predictions on the test set
        y_pred = ridge.predict(X_test)

        # Calculate RMSE and R2 scores
        rmse = np.sqrt(mean_squared _error (y_test, y_pred))

        # Append the scores to the lists
        rmse_scores.append(rmse)

    # Find the index of the alpha value with the lowest RMSE
    best_alpha_idx = np.argmin(rmse_scores)
    best_alpha = alpha_range [best_alpha_idx]
```

```
    # Train the Ridge Regression model with the best alpha on the entire dataset
    ridge.set_params (alpha=best_alpha)
    ridge.fit(X, y)

    # Make predictions on the entire dataset
    y_pred_all = ridge.predict(X)
    # Plot the data points and the best fit line
    plt.scatter(X, y, color='blue', label='Data Points') plt.plot(X, y-pred_all, color='red', label='Best Fit Line (Ridge)") plt.xlabel('Year') plt.ylabel('GS#")
    plt.title('Ridge Regression with Best Fit Line (NGS)')
    plt. legend()
    plt.savefig('NGS_model.png')

    # Print the best alpha value, RMSE, and R2 score
    print(' Best Alpha:' best_alpha)
    print('MSE:
    rmse_scores [best_alpha_idx]**2)
    print('RMSE:', rmse_scores [best_alpha_idx])

```




```
Best Alpha: 0.0
MSE:  156589.79591857793
RMSE: 395.7142857145518
```

The Statista's ("The Number of Gasoline Stations in the US") dataset [1] also utilized a Ridge Regression model to analyze a small dataset better. Primarily, the dataset was split into "year" and "GS#" (Gas Station number) columns were selected as the independent (X) and the dependent variable (Y), respectively, which would split into training and testing data. Hyper-parameter Tuning was applied in the form of a for loop to find the most efficient alpha value (0) based on the evaluation metric of MSE (156589.80) and RMSE (395.71) after fitting the training data and predicting the number of gas stations with the testing data (indicates the model was not accurate in predictions as it is very high). Finally, the visualization process included plotting all existing data points and the best-fit line of the same X values with the predicted values for Y (GS#), which portrayed an increasing number of Gas Stations over some time.

9 COMMUNICATE / VISUALIZE THE RESULTS

In summary, our study revealed that electric vehicles outperform gasoline-powered vehicles in terms of environmental impact, cost-effectiveness, and efficiency. However, gasoline-powered vehicles showed better reliability compared to electric vehicles.

All in all when considering the reliability of EVs versus gas-powered vehicles, selecting the vehicle with the most available resources is what makes the traditional gas vehicle more reliable than its electric counterpart. After analyzing the IEA's ("Global-ev-outlook-2022") dataset [7] and Statista's ("The Number of Gasoline Stations in the US") dataset [1] , the results of our findings confirm that there exist 3.38 times the number of gas stations compared to charge stations from the span of 2014-2017. Therefore, gas vehicles are more reliable in terms of convenience compared to electric vehicles due to the increased availability of gas stations.

Putting together everything that we learned about the cost-effectiveness of electric vehicles over gasoline-powered vehicles, electric vehicles are often comparatively priced in relation to gas-powered vehicles and this margin can both be because of and improved with the inclusion and utilization of available rebates and tax credits. As of right now, the data does not demonstrate that hybrid vehicles (HEVs and PHEVs) save consumers money in the long run, but battery electric vehicles though being the most widely priced, having the largest price spread, can save consumers a great deal over the long term ownership of the vehicle due to it's sole reliance on electricity as it's power source. Data was sparse, quite varied, and inconsistent regarding insurance solutions and financing options for such vehicles, but research that falls outside of the scope of this project has been conducted that suggests that it is possible that consumers could save even more in the long term if buying used.

Electric vehicles are gaining popularity as a primary mode of transportation due to their energy efficiency and environmentally friendly features, especially when the energy source is renewable. EVs regenerate energy that would have been lost, unlike gas-powered vehicles that use energy even when not in motion. EVs have zero CO2 emissions and perform better in cities than on highways due to the traffic found in cities. Through this study, we have learned that EV's are able to convert more stored energy into movement compared to gas-powered vehicles, which is a large reason why EV's are gaining popularity today.

Of course, the right combination of factors is always important to get the best outcome and hopefully, advancements will be made in the field that will allow for improvements, technological advancements, better accessibility, and more competitive pricing as the industry grows and its surrounding infrastructure grows to support it.

REFERENCES

[1] Published by Statista Research Department and Jul 12. 2022. Number of gasoline stations in the u.s. (July 2022). https://www.statista.com/statistics/525107/number-of-gasoline-stations-in-the-united-states.

[2] 2022. Causes and effects of climate change. (May 23, 2022). Retrieved May 2, 2023 from https://www.un.org/en/climatechange/science/causes-effects-climate-change.

[3] Camila Domonoske. 2023. Buying an electric car? you can get a $7,500 tax credit, but it won't be easy. *NPR*, (Apr. 3, 2023). Retrieved May 2, 2023 from https://www.npr.org/2023/01/07/1147209505/electric-car-tax-credit-climate-bill-tesla-volkswagen-ev.

[4] 2022. Global electric car sales have continued their strong growth in 2022 after breaking records last year - news. IEA. (May 23, 2022). Retrieved May 2, 2023 from https://www.iea.org/news/global-electric-car-sales-have-continued-their-strong-growth-in-2022-after-breaking-records-last-year.
[5] Chris Harto. 2020. EV Ownership Cost Final Report. (Oct. 2020). Retrieved May 2, 2023 from https://advocacy.consumerreports.org/wp-content/uploads/2020/10/EV-Ownership-Cost-Final-Report-1.pdf.
[6] 2022. History of the electric car: from the first ev to the present day. (May 23, 2022). Retrieved May 2, 2023 from https://www.autoexpress.co.uk/car-news/electric-cars/101002/history-of-the-ev-from-the-first-electric-car-to-the-present-day.
[7] Iea. [n. d.] Monthly electricity statistics - data product. https://www.iea.org/data-and-statistics/data-product/monthly-electricity-statistics.
[8] Katie Lowery Xiomara Martinez-White. 2023. 2023 EV tax credit: what you need to know. LendingTree. (Apr. 26, 2023). Retrieved May 2, 2023 from https://www.lendingtree.com/auto/ev-tax-credit/.
[9] Jessie See. 2023. Everything you need to know about insuring an electric vehicle. Bankrate. (Mar. 14, 2023). Retrieved May 2, 2023 from https://www.bankrate.com/insurance/car/electric-car-insurance/.