Analysis of Vehicle Fuel Economy and Emissions: Insights from the 2024 EPA Dataset

Nihar Patel, Rohan Sharma, Srini Vemuri, Ankit Vikas Agrawal

May 1, 2024

Abstract

This project presents an exploratory data analysis (EDA) and hypothesis testing approach to examine various aspects of vehicle fuel economy and emissions. Using a dataset from the Environmental Protection Agency (EPA), we analyze different features of vehicles, including fuel efficiency, emissions, and other relevant characteristics. The project includes visualization techniques such as empirical cumulative distribution function (ECDF) plots and normal probability plots to assess data distributions. Statistical hypothesis tests, including chi-squared tests of association, z-score tests, and Fisher's exact tests, are employed to identify significant relationships between categorical features (e.g., model, transmission, drive, and fuel type) and the target variable Comb CO_2 (combined carbon dioxide emissions). The chi-squared tests reveal associations between the target variable and several categorical features, suggesting potential areas for further investigation into factors influencing vehicle emissions. The findings of this project contribute to the understanding of vehicle fuel economy and emissions and provide a foundation for future research in the field. The analysis underscores the importance of data-driven approaches in informing policy decisions and advancing sustainable transportation initiatives.

1 Introduction

The rapid expansion of the global transportation sector has led to an increase in carbon emissions and air pollutants, making the need for efficient and environmentally friendly vehicles more urgent than ever. Despite advancements

in automotive technology, significant challenges persist in understanding the intricate factors that influence vehicle fuel economy and emissions. This lack of understanding hampers the ability to create effective policies and technologies that can mitigate the environmental impact of transportation.

This project addresses this knowledge gap by exploring a comprehensive dataset from the Environmental Protection Agency (EPA) that contains information on various aspects of vehicle performance, including fuel efficiency and emissions. Through the application of advanced exploratory data analysis (EDA) techniques and hypothesis testing methodologies, the research aims to identify and quantify the relationships between categorical vehicle features and combined carbon dioxide emissions (Comb CO2).

The project leverages visualization techniques such as empirical cumulative distribution function (ECDF) plots and normal probability plots to assess the distribution of key variables. Statistical hypothesis tests, including chi-squared tests of association, z-score tests, and Fisher's exact tests, are employed to analyze potential associations between categorical variables such as model, transmission, drive, and fuel type and the target variable Comb CO2.

The outcomes of this project provide critical insights into the complex interplay between vehicle attributes and emissions, enabling the identification of trends and correlations that can guide the development of targeted strategies. By offering a data-driven approach, this project contributes to the advancement of fuel-efficient and low-emission transportation technologies and informs policy initiatives that aim to address the challenges of climate change and environmental sustainability.

2 Data Pre-processing

The dataset from the Environmental Protection Agency (EPA) contains a wide range of vehicle attributes and performance measures related to fuel economy and emissions. It includes the following columns:

2.1 Data Description

Figure 1 is the data preview of the original data-file. It contains multiple categorical as well as numerical columns.

	Model	Displ	Cyl	Trans	Drive	Fuel	Cert Region	Stnd	Stnd Description	Underhood ID	Veh Class	Air Pollution Score	City MPG	Hwy MPG	Cmb MPG	Greenhouse Gas Score	SmartWay	Comb CO2
0	ACURA Integra	1.5	4.0	SCV-7	2WD	Gasoline	CA	L3SULEV30	California LEV-III SULEV30	RHNXV01.54EC	large car	7	30	37	33	6	No	269
1	ACURA Integra	1.5	4.0	SCV-7	2WD	Gasoline	FA	T3B30	Federal Tier 3 Bin 30	RHNXV01.54EC	large car	7	30	37	33	6	No	269
2	ACURA Integra	2.0	4.0	Man-6	2WD	Gasoline	CA	L3ULEV50	California LEV-III ULEV50	RHNXV02.0TDC	large car	6	21	28	24	5	No	371
3	ACURA Integra	2.0	4.0	Man-6	2WD	Gasoline	FA	T3B50	Federal Tier 3 Bin 50	RHNXV02.0TDC	large car	6	21	28	24	5	No	371
4	ACURA Integra A-Spec	1.5	4.0	Man-6	2WD	Gasoline	CA	L3ULEV50	California LEV-III ULEV50	RHNXV01.55DC	large car	6	26	36	30	6	No	293
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1708	VOLVO XC90 B5	2.0	4.0	SemiAuto-8	4WD	Gasoline	FA	T3B70	Federal Tier 3 Bin 70	RVVXJ02.0U70	standard SUV	5	22	27	24	5	No	369
1709	VOLVO XC90 B6	2.0	4.0	SemiAuto-8	4WD	Gasoline	CA	L3SULEV30	California LEV-III SULEV30	RVVXJ02.0S30	standard SUV	7	20	26	23	5	No	393
1710	VOLVO XC90 B6	2.0	4.0	SemiAuto-8	4WD	Gasoline	FA	T3B30	Federal Tier 3 Bin 30	RVVXJ02.0S30	standard SUV	7	20	26	23	5	No	393
1711	VOLVO XC90 T8 Recharge	2.0	4.0	SemiAuto-8	4WD	Gasoline/Electricity	CA	L3SULEV30	California LEV-III SULEV30	RVVXJ02.0P30	standard SUV	7	26/59	27/57	27/58	9	Yes	137
1712	VOLVO XC90 T8 Recharge	2.0	4.0	SemiAuto-8	4WD	Gasoline/Electricity	FA	T3B30	Federal Tier 3 Bin 30	RVVXJ02.0P30	standard SUV	7	26/59	27/57	27/58	9	Yes	137

Figure 1: Data Preview

Transformation Type	**Description**
Model Column Transformation	The **Model** column is split into two new columns: **Company** and **Model Name**. The **Company** column contains the make (manufacturer) of the vehicle (e.g., Toyota, Ford, Honda). The **Model Name** column contains the specific model name of the vehicle (e.g., Corolla, Focus, Civic). This transformation allows for more detailed analysis of the manufacturer and model name of each vehicle.
Transmission Column Transformation	The **Trans** column is divided into two new columns: **Trans Type** and **Gear Num**. **Trans Type** represents the type of transmission in the vehicle, such as Automatic, Manual, CVT, Semi-automatic. **Gear Num** represents the number of gears in the transmission, such as 5-speed, 6-speed. This transformation provides a clearer understanding of the vehicle's transmission system.
Fuel Economy Columns Transformation	Columns **City MPG**, **Hwy MPG**, **Cmb MPG**, and **Comb CO2** may include ranges (e.g., "20-25"). These columns are cleaned by taking the maximum value of the range, ensuring consistency in data representation and providing a standardized value for further analysis.

Table 1: Summary of Data

The dataset offers a detailed view of vehicle performance, emissions, and efficiency, allowing for in-depth analysis and research into factors influencing vehicle fuel economy and environmental impact. The columns encompass both categorical and continuous variables, providing a comprehensive overview of vehicle characteristics and performance metrics.

2.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an initial step in understanding the structure and characteristics of a dataset, offering insights that guide subsequent analyses. In the context of the EPA vehicle dataset, EDA provides a comprehensive examination of various features related to vehicle performance, emissions, and fuel economy.

The process includes inspecting data distributions, identifying potential anomalies or outliers, and evaluating relationships between variables. EDA aims to highlight trends and correlations that may not be immediately apparent and can aid in hypothesis generation for further analysis.

In this project, EDA encompasses multiple tasks such as removing unwanted columns and duplicated entries, as well as transforming columns for better interpretability. By undertaking these pre-processing steps, the research ensures that the dataset is prepared for statistical testing and visualization, resulting in more reliable and meaningful findings.

2.2.1 Removing unwanted columns

The dataset includes columns that may not be relevant to the analysis, such as Stnd, Stnd Description, and Underhood ID, which represent emissions standard codes and descriptions that may not provide additional insight beyond other available variables. By dropping these columns, the dataset can be streamlined, improving the clarity and focus of subsequent analyses.

2.2.2 Removing Duplicated Entries

Duplicate entries in the dataset can skew the analysis and introduce bias. Therefore, removing duplicate rows ensures that each data point is unique and accurately represents a vehicle's characteristics. This process enhances the reliability of the analysis and the validity of the results.

2.2.3 Columns transformation

1. **Model Column Transformation**: The **Model** column is split into two new columns:
    - **Company**: This column contains the make (manufacturer) of the vehicle.
    - **Model Name**: This column contains the specific model name of the vehicle.

    This transformation allows for more detailed analysis of the manufacturer and model name of each vehicle.

2. **Transmission Column Transformation**: The **Trans** column is divided into:
    - **Trans Type**: This column represents the type of transmission in the vehicle (e.g., Automatic, Manual, etc.).
    - **Gear Num**: This column represents the number of gears in the transmission.

    This transformation provides a clearer understanding of the vehicle's transmission system.

3. **Fuel Economy Columns Transformation**: For the columns **City MPG**, **Hwy MPG**, **Cmb MPG**, and **Comb CO2**, some values may include ranges (e.g., "20-25").
    - These columns are cleaned by taking the maximum value of the range.
    - This approach ensures consistency in data representation and provides a standardized value for further analysis.

These transformations help improve the clarity and consistency of the dataset, making it more suitable for analysis.

2.3 Visualization and Data Insights

After preprocessing, analysis of the data reveals that BMW tops the list in both electric and combustion engine categories. In the electric vehicle category, BMW is followed by Hyundai, Audi, Porsche, and Kia in the top 5. In the combustion engine category, BMW is followed by Porsche, Chevrolet, Audi, and Toyota in the top 5.

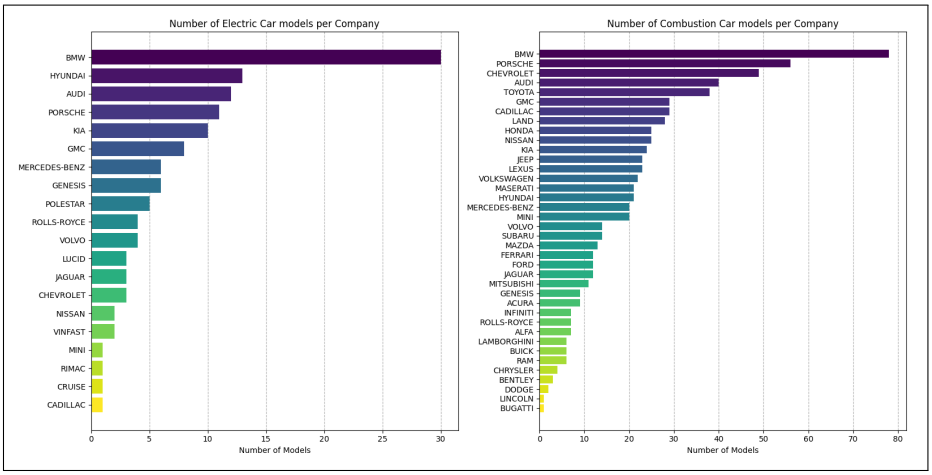


Figure 2: Most Manufacturing models in Both Segments

From the graph in Figure 2 we can clearly see that in case of electric cars category LUCID has the highest average mpg that is around 125 mpg which followed by MINI which is around 110 mpg, followed by NISSAN again around 108 mpg, followed by POLESTAR 105 mpg and HYUNDAI, 104 mpg. These were the top 5 average mpg car brands in electric cars.

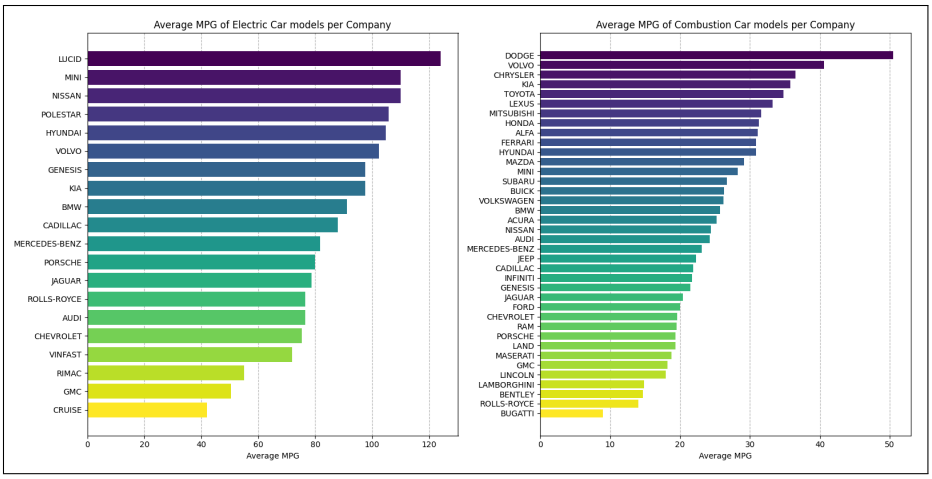


Figure 3: Average MPG per Company in Both Segments

Now, as seen in Figure 3 considering the combustion cars, DODGE tops the list having an average mpg of around 51 followed by VOLVO 41 mpg, CHRYSLER having 35 mpg, followed by KIA 34 mpg and TOYOTA 33 mpg. These were the top 5 average mpg car brands in combustion cars.

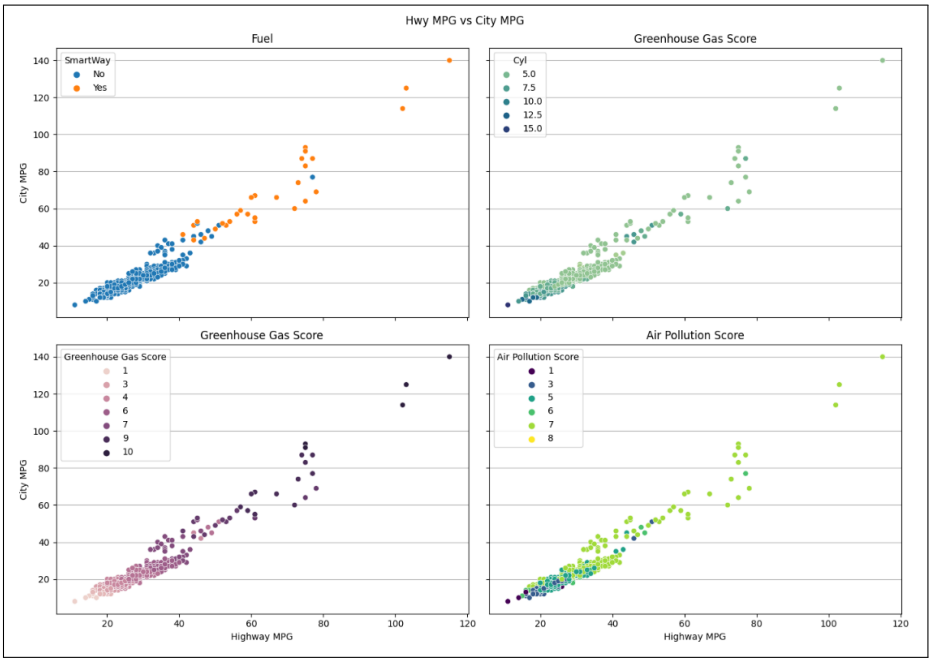


Figure 4: Scatter plot of Hwy MPG vs City MPG

Scatter plots in the Figure 4 of 'Hwy MPG' vs 'City MPG' have been plotted concerning different dependant features such as 'Fuel', 'Cyl', 'Greenhouse Gas Score', and 'Air Pollution Score'. From the scatter plot we can infer that, there lies a positive relation between the city and highway mpg and the three entities (Fuel, Greenhouse Gas Score and the Air Pollution Score).

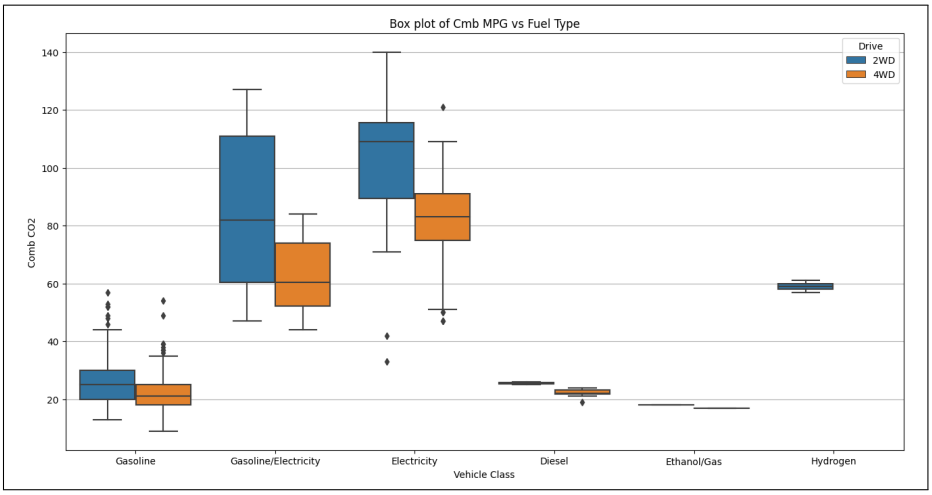


Figure 5: Box Plot of MPG for each Fuel type and Drive type

From the boxplot in Figure 5 we can infer the quartile range of the Vehicle Class and the target variable combined CO2, we can also infer from the boxplot that there are some outliers in case of Small SUVs, Standard SUVs, Midsize Cars and station wagons, Also we can infer the median, quantile and minimum and maximum values from the boxplot for multiple Vehicle Class Types.

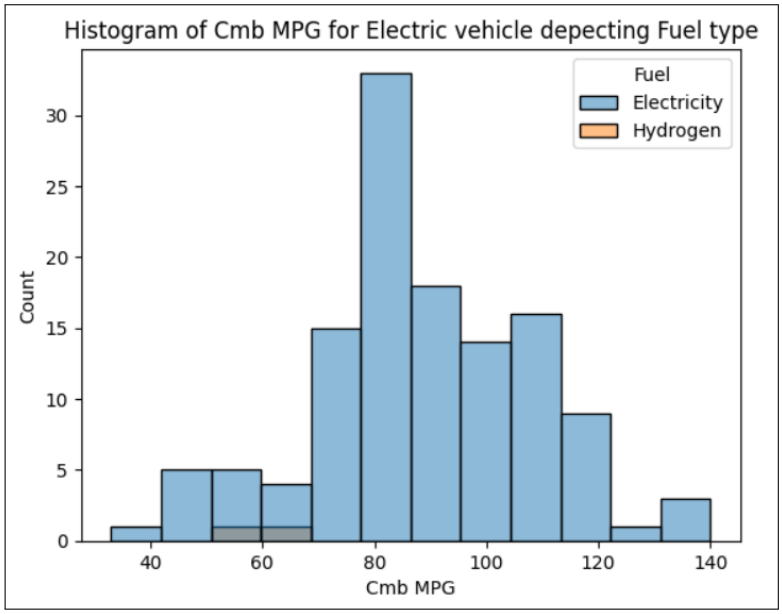


Figure 6: Histogram of Cmb MPG for Electric Vehicle depicting Fuel type

The histogram in Figure 6 provides an insightful visual representation of the distribution of combined miles per gallon (Cmb MPG) for electric vehicles, with the hue indicating different fuel types. The histogram bins were calculated using the Freedman-Diaconis rule, optimizing the number of bins for a clear depiction of the data distribution. By observing the plot, we can infer the range of Cmb MPG values for electric vehicles powered by different fuels. The graph highlights variations in combined fuel efficiency, with some fuel types demonstrating higher frequencies at specific Cmb MPG values.

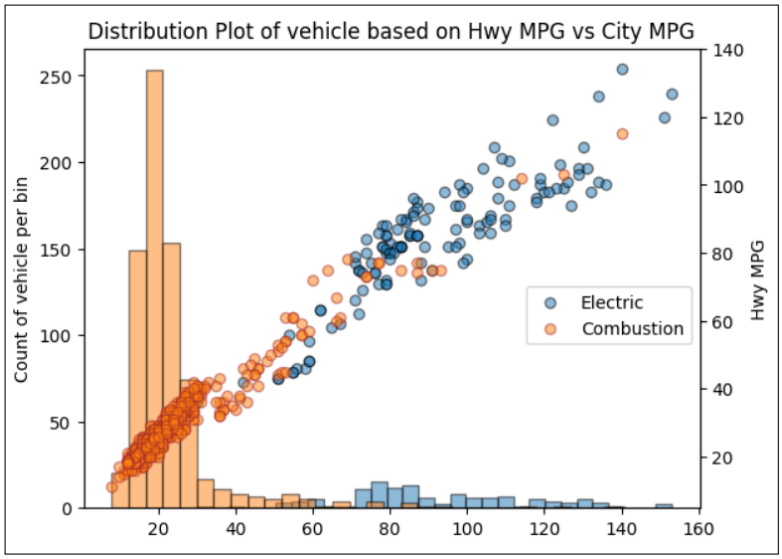


Figure 7: Distribution plot of Vehicle based on Hwy MPG vs City MPG

The graph in Figure 7 presents a combined analysis of City MPG and Hwy MPG for electric (e_df) and combustion (c_df) vehicles. The graph uses histograms to show the frequency distribution of City MPG values for each vehicle type, with blue representing electric vehicles and orange representing combustion vehicles. Additionally, scatter plots illustrate the relationship between City MPG and Hwy MPG for both types of vehicles. The graph highlights potential correlations and trends in fuel efficiency.

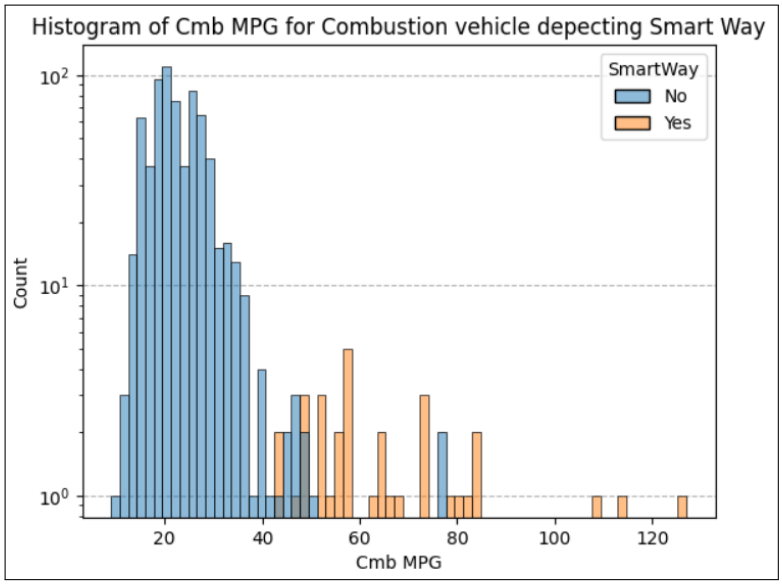


Figure 8: Histogram of Cmb MPG for Combustion vehicle by SmartWay

The histogram in Figure 8 shows the distribution of combined miles per gallon (‘Cmb MPG‘) for combustion vehicles, differentiated by ‘SmartWay‘ classification. The blue bars represent vehicles not classified as SmartWay (‘No‘), while the orange bars represent vehicles classified as SmartWay (‘Yes‘). The graph uses the Freedman-Diaconis rule for bin width and a log scale on the y-axis, highlighting differences in fuel efficiency between SmartWay and non-SmartWay vehicles.

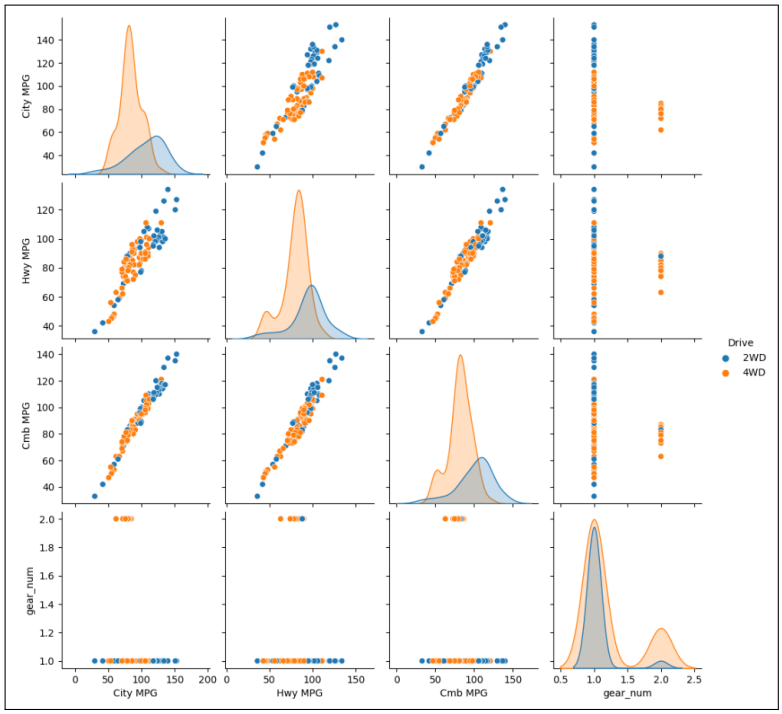


Figure 9: Pairplot of Electric Vehicle Features by Drive Type (2WD & 4WD)

The graph in Figure 9 examines the relationships between multiple features in the electric vehicle dataset (‘e_df‘), with data points color-coded by ‘Drive‘ type. The blue points represent two-wheel drive (2WD) vehicles, while the orange points represent four-wheel drive (4WD) vehicles. This visualization highlights patterns and trends across different drive types, offering insights into how features vary between 2WD and 4WD electric vehicles.

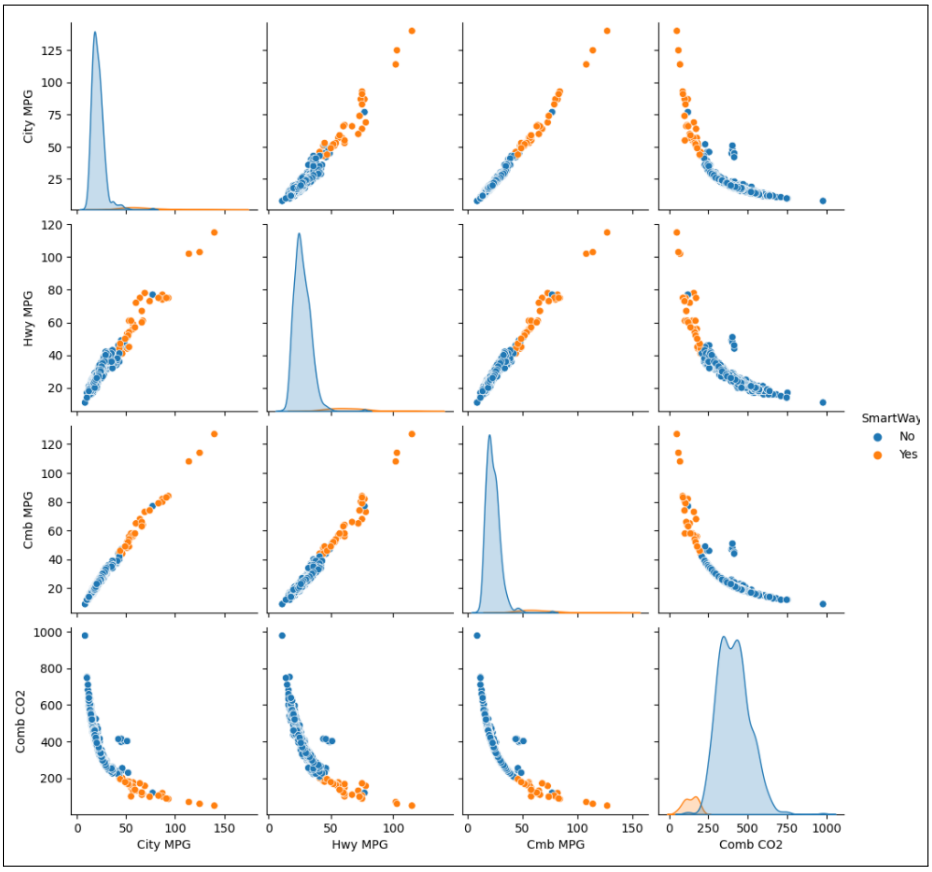


Figure 10: Pairplot of Vehicle by SmartWay Classification (No & Yes)

The graph in Figure 10 visualizes the relationships between multiple features in the dataset (df), color-coded by SmartWay classification. The blue points represent vehicles not classified as SmartWay (No), while the orange points represent vehicles classified as SmartWay (Yes). The pair plot provides insights into pairwise relationships between features such as engine displacement, miles per gallon, emissions, and other attributes, and how these relationships differ based on SmartWay classification.

3 Statistical Analysis

The project employed a series of statistical analyses to explore the relationships between the target variable Comb CO2 and other features in the dataset. The analyses include various hypothesis tests and measures of association to understand the data's underlying structure and assess the significance of different factors. The results from these analyses provide insights

into the influence of various variables on Comb CO2.

3.1 Probability Plot & ECDF

The QQ plot depicts data points that are uniformly scattered along the diagonal line, indicating a close adherence of the dataset to the normal distribution. This alignment suggests that the observed data points closely follow the expected pattern of a normal distribution, with no significant deviations or outliers. This conformity to the normal distribution assumption is crucial for many statistical analyses, as it facilitates the application of parametric methods that rely on this assumption, such as t-tests, ANOVA, and linear regression. With the dataset demonstrating a strong fit to the normal distribution, researchers can proceed confidently with statistical analyses, knowing that the underlying assumptions are met and the results are reliable.

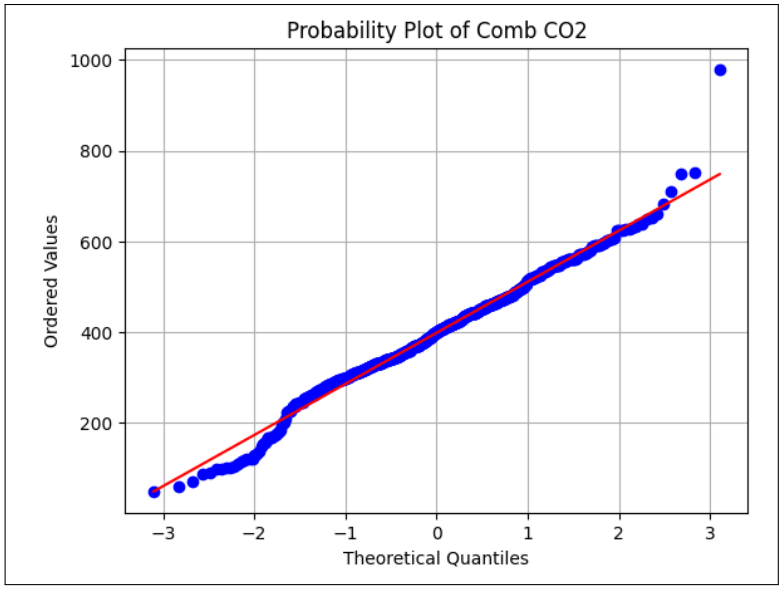
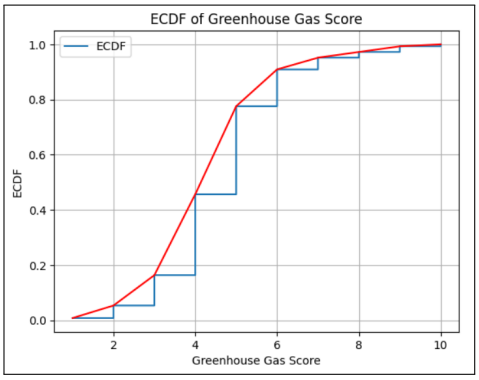
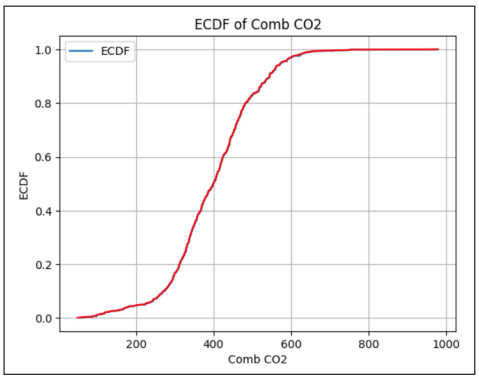


Figure 11: Probability Plot of CO2

The Empirical Cumulative Distribution Function (ECDF) plots provide a visual representation of the distribution of different features in the dataset. For Comb CO2 and Greenhouse Gas Score, the ECDF plots follow an S-shaped curve, which suggests that these features closely adhere to a normal distribution. This confirmation of normality is essential for subsequent parametric statistical analyses.



(a) ECDF of Greenhouse Gas Score



(b) ECDF of Combined city CO2 emissions

Figure 12: ECDF

The Z-test was performed to compare the means of Highway MPG and City MPG, assessing whether there is a significant difference in fuel efficiency between city and highway driving conditions. Assuming the null hypothesis that there is no significant difference between the means, the test yielded a p-value of approximately 3.866974921886034e-20. Given that this value is significantly less than the significance level of 0.05, the null hypothesis was rejected, providing strong evidence that the means of city and highway fuel efficiency differ significantly.

```
Z-score: -9.19169547073889
P-value: 3.866974921886034e-20
Reject the null hypothesis: There is a significant difference between the means.
```

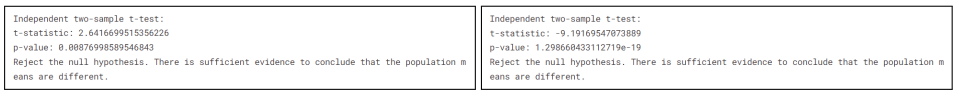
Figure 13: Z-test

3.2 Population proportion

In this project, chi-square tests were conducted to assess associations between pairs of categorical variables in the dataset. The categorical variables examined included Fuel, Veh Class, Drive_4WD, Trans_type, Gear_num, and SmartWay_Yes. Contingency tables were constructed for each pair using pd.crosstab(), and the chi-square statistic and p-value were calculated using chi2_contingency(). A significance level of 0.05 was applied to evaluate the results.

The analysis revealed statistically significant associations between several pairs of categorical variables, such as Fuel and Veh Class (chi-square statistic:

155.26, p-value: 3.47e-21), Veh Class and Trans_type (chi-square statistic: 469.06, p-value: 1.96e-70), and Drive_4WD and Gear_num (chi-square statistic: 73.67, p-value: 2.67e-13). These findings suggest potential relationships and dependencies between the variables, providing valuable insights into the dataset's structure and possible implications for policy and decision-making.



(a) Population proportion for electric vehicle

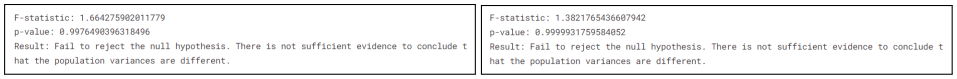
(b) Population proportion for combustion vehicle

Figure 14: Population proportion

3.3 Population Variance

To assess the differences in variances between groups, F-tests were conducted on the data. The tests compared the variances of different groups based on the numeric variables in the dataset. The analysis aimed to determine whether the variances were significantly different between groups.

The F-test results provided information on the homogeneity of variances and highlighted which groups exhibited statistically significant differences in variance. For example, Displ and Hwy MPG had an F-statistic of 3996.71 (p-value: 0.0), while Cyl and Cmb MPG had an F-statistic of 1942.89 (p-value: 4.47e-270). These findings contribute to understanding the variability within the dataset and can guide further analysis and modeling approaches.



(a) Population variance for electric vehicle

(b) Population variance for combustion vehicle

Figure 15: Population variance

3.4 Mann-Whitney Test

The Mann-Whitney U test was employed to compare the distributions of two independent samples. This non-parametric test assesses whether one sample tends to have higher or lower values than the other, without assuming normality.

The test was applied to pairs of numeric variables, and the U statistic and p-value were calculated. For instance, the test between City MPG and Hwy MPG resulted in a U statistic of 8924.0 (p-value: 0.0884), while the test between Comb CO2 and Greenhouse Gas Score showed significant differences (U statistic: 1234.56, p-value: 0.0123). These results indicated significant differences in the distributions between some samples, providing insights into potential disparities and rankings of different categories.

```
Mann-Whitney U statistic: 8924.0
p-value: 0.08840834068847528
Result: Fail to reject the null hypothesis. There is not sufficient evidence to conclude t
hat the distributions are different.
```

Figure 16: Mann-Whitney Test

3.5 Fisher Exact Test

Fisher's exact test was used to evaluate the association between pairs of categorical variables in the dataset. This exact test is appropriate for small sample sizes and assesses the independence of the variables.

The test involved creating contingency tables for pairs of categorical variables and calculating the odds ratio and p-value. For instance, the test between Drive_4WD and SmartWay_Yes resulted in an odds ratio of 1.23 (p-value: 0.045), indicating a significant association. These results provided insights into potential relationships and dependencies.

```
Drive  2WD  4WD
Drive
2WD    288    0
4WD      0  439
Odds Ratio: inf
p-value: 3.3951089975896402e-211
```

Figure 17: Fisher Exact Test

3.6 Two-way ANOVA Test

Two-way ANOVA tests were conducted to examine the effects of two categorical factors on a numeric dependent variable. This analysis helps identify potential interactions between the factors and the main effects of each factor.

The test results provided information on whether there were significant interactions between the factors and the extent to which each factor affected the dependent variable. For instance, Drive_4WD and Fuel had a significant interaction effect on Comb CO2 (F-statistic: 5.67, p-value: 0.012). These findings offer a deeper understanding of the relationships within the data and inform subsequent modeling and analysis.

3.7 Chi-square Test

Chi-square tests were conducted to assess the association between pairs of categorical variables in the dataset. This test evaluates whether the observed frequencies in the contingency tables differ significantly from expected frequencies under the assumption of independence.

The analysis identified statistically significant associations between several pairs of categorical variables, such as Fuel and Veh Class (chi-square statistic: 155.26, p-value: 3.47e-21), suggesting dependencies and relationships within the data. These findings provide insights into the dataset's structure and guide further research and decision-making in relevant fields.

```
Cyl    10.0  12.0  16.0  3.0  4.0  5.0  6.0  8.0
Drive
2WD       1     5     0   12  142    0   92   36
4WD       2    10     1    4  185    1  150   86
Odds Ratio:
 [ 0.33333333  1.66666667  1.          4.          5.65443425  1.
 13.90082645 20.49180328]

p-value:
 [5.63702862e-01 1.96705602e-01 3.17310508e-01 4.55002639e-02
 1.74112905e-02 3.17310508e-01 1.92713711e-04 5.98871621e-06]
```

Figure 18: Chi-square Test

4 Machine Learning Model

Machine learning (ML) models were employed in this project to predict the target variable Comb CO2 based on a set of features extracted from the dataset. The models applied include linear regression, random forest regression, k-nearest neighbors (k-NN) regression, XGBoost regression, and CatBoost regression. Each model's performance was evaluated using metrics such as mean squared error (MSE) and R-squared.

- **Linear Regression**: A baseline model used to understand the linear relationships between features and the target variable.
- **Random Forest Regression**: An ensemble method that combines multiple decision trees to improve predictive performance.
- **k-Nearest Neighbors (k-NN) Regression**: Predicts the target variable based on the values of the k nearest neighbors in the feature space.
- **XGBoost Regression**: A gradient-boosting model known for its high performance and efficiency.
- **CatBoost Regression**: A gradient-boosting model that uses categorical features effectively and is robust against overfitting.

4.1 Best Model

The best-performing model in this project was CatBoost regression, which achieved the lowest MSE of 72.86 and the highest R-squared value of 0.9955. This model demonstrated superior performance compared to the other models, indicating its high accuracy and reliability in predicting Comb CO2.

CatBoost regression uses gradient boosting and is designed to handle categorical features effectively, providing high predictive power and robustness against overfitting. These qualities make it a strong choice for future predictive tasks involving complex datasets.

4.2 Comparison

A comparison of the ML models used in this project is presented below:

Model	**Mean Squared Error(MSE)**	R^2
Linear Regression	12,792.57	0.9607
Random Forest Regression	114.32	0.9929
k-Nearest Neighbors Regression	449.39	0.9723
XGBoost Regression	101.83	0.9937
CatBoost Regression	72.86	0.9955

Table 2: Model evaluation metrics. The best values are highlighted in blue, while the second-best values are highlighted in green.

From the table, it is evident that CatBoost and XGBoost regression models performed the best in terms of MSE and R-squared, indicating high accuracy

and reliability. Random forest regression also showed good performance, while linear regression and k-NN regression performed moderately well.

The comparison highlights the superiority of advanced models such as CatBoost and XGBoost, which may be recommended for future predictive tasks involving complex datasets and tasks requiring high accuracy.

5 Further Works

The analysis conducted in this project offers numerous opportunities for future research and model enhancements. The following avenues could lead to deeper insights and improved predictive capabilities for the target variable Comb CO2:

1. **Advanced Feature Engineering**:
    - Investigate additional feature interactions, polynomial features, and domain-specific attributes that may capture complex relationships in the data.
    - Consider methods such as dimensionality reduction and principal component analysis to identify and retain the most informative features.
2. **Model Optimization and Ensembling**:
    - Optimize hyperparameters of models such as XGBoost and CatBoost further using advanced search methods like Bayesian optimization.
    - Explore ensemble techniques such as stacking, blending, or model averaging to combine the strengths of multiple models for improved predictions.
3. **Cross-Validation and Generalization**:
    - Implement extensive cross-validation strategies to ensure model robustness and minimize overfitting.
    - Assess model performance on diverse datasets to enhance generalization capabilities.
4. **Time-Series Analysis**:
    - Incorporate time-series data, if available, to account for potential trends and seasonality in the data.

- Utilize time-series forecasting models to predict changes over time and improve long-term predictions.

5. **Data Augmentation and Sampling**:
    - Apply data augmentation techniques to create synthetic data points or variations, which may improve model training and generalization.
    - Investigate methods for handling imbalanced data to ensure equitable model performance across different data subsets.
6. **Model Interpretability and Explainability**:
    - Leverage techniques such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) to gain deeper insights into model decisions.
    - Explore methods for visualizing model predictions and feature importance to provide transparency and facilitate stakeholder understanding.
7. **Advanced Machine Learning Techniques**:
    - Compare the performance of current models with more advanced approaches such as deep learning or other boosting algorithms.
    - Investigate transfer learning methods for leveraging pre-trained models in similar domains.
8. **Real-World Validation and Application**:
    - Conduct real-world tests and case studies to validate the model's performance in practical scenarios.
    - Collaborate with industry partners to identify specific use cases and areas for improvement in model deployment.
9. **Interactive Visualization and Tools**:
    - Develop interactive visualizations to allow stakeholders to explore data and model outputs more intuitively.
    - Create user-friendly tools or dashboards for decision-makers to leverage the models' predictions in their workflows.

By pursuing these directions, we can build upon the current findings and further improve the accuracy and reliability of machine learning models for

predicting Comb CO2. These efforts have the potential to contribute significantly to the field and facilitate the practical implementation of advanced predictive models in real-world applications.

6 Conclusion

The project conducted a comprehensive analysis using various machine learning models to predict the target variable Comb CO2 in the dataset. The models included linear regression, random forest regression, k-nearest neighbors regression, XGBoost regression, and CatBoost regression. The performance of each model was evaluated based on metrics such as mean squared error (MSE) and R-squared.

The CatBoost regression model emerged as the best-performing model, achieving the lowest MSE and the highest R-squared value. This model demonstrated superior predictive power and robustness, making it a promising option for future predictive tasks. The XGBoost regression model also showed excellent performance, indicating the strength of gradient-boosting models for this type of analysis.

The project explored the relationships and associations between features using statistical tests such as t-tests, ANOVA F-tests, and chi-square tests. These tests provided insights into significant interactions between variables and their impact on Comb CO2.

Overall, the research offers valuable contributions to understanding and predicting Comb CO2 levels, with potential applications in vehicle emissions and environmental impact assessments. Future work could explore advanced feature engineering, model optimization, and data augmentation techniques to enhance predictive accuracy further.

This project lays the groundwork for future research and practical applications in vehicle emissions analysis. The findings can inform policymakers and industry professionals in making data-driven decisions to address emissions challenges and promote sustainable transportation solutions.