EECS 498-009/598-014 Data Science Projects: READS
Final Report

Abstract

The goal of this project is to create a tool that the National Vehicle Fuel Emissions Lab (NVFEL) can use to calculate the difference in fuel consumption between autonomous vehicles (AVs) and those that are human-operated. To these ends, we've been given a vehicle energy dataset (VED) which contains time series data of nearly 400,000 miles driven by 383 vehicles driven in the Ann Arbor area and is to be used to create predictive models for the tool. This project has been divided up into three classes of deliverables: minimum viable product (MVP) deliverables, expected deliverables, and stretch goals, all of which are related to estimating the potential fuel consumption reduction from AVs. To tackle these, we've created a dashboard that relies on several multiple regression models in order to estimate fuel consumption with respect to several parameters -- most notably driver aggressiveness. On the whole, our group is proud to have completed or mostly completed all of the MVP and expected deliverables. That being said, our solution is not without limitation. In particular, we found that our dataset is often sparse with data that is needed to calculate fuel consumption directly, requiring that we estimate it through proxy variables. Furthermore, because our dataset is primarily composed of internal combustion engine (ICE) vehicles and hybrid electric vehicles (HEVs), we excluded plug-in hybrid electric vehicles (PHEVs) from our tool. Further elaboration on what was and was not achieved can be found throughout the remainder of this document.

Accomplishments

Major Goals/Deliverables

As was mentioned in the abstract above, this project has several main deliverables divided into three categories: minimum viable product (MVP), expected deliverables, and stretch goals. The deliverables that have been selected to be implemented by the team are listed below.

1. Minimum Viable Product Deliverables

1.1 Calculate driver aggressiveness, average speed, and fuel consumption for each trip
1.2 Determine a level of driver aggressiveness that would represent an autonomous vehicle
1.3 Determine the potential reduction in fuel consumption as function of vehicle weight, engine size, and level and electrification.
1.4 Estimate the potential total reduction in fuel consumption over the entire data set

1.5 Develop a simple tool to display fuel consumption and potential reduction as a function of vehicle speed for both individual vehicles and fleets of particular power, weight, and hybridization

2. Expected Product Deliverables

2.1 Incorporate the effect of other data variables, particularly external temperature and air conditioning use. Determine to what extent these affect fuel consumption and the potential decrease due to autonomous vehicles.
2.2 Divide the driving by trips or trip segments into a "trip type" range of urban, rural, or freeway driving based on average speeds and frequency of stops
2.3 Expand the simple tool to display fuel consumption and potential reduction as a function of temperature and trip type
2.4 Allow the user to set a definition for the autonomous level of aggressive driving, as either an absolute value, or a "percentage" of the original

3. Stretch Goals

3.1 Use data set time and location, combined with weather data to determine wind and weather for each trip

- **3.1.1** Determine to what extent precipitation and wind magnitude and direction affects fuel consumption and potential reduction due to AVs.
- **3.1.2** Add the wind and precipitation effects to the simple tool

3.2 Refine the tool to allow a mix of trip types as input
3.3 Refine "trip type" by considering the posted speed limits for each trip segment

Specific Objectives

Most deliverables were able to be completed, a few were partially completed, and a few were not able to be completed due to various constraints. The completion of these deliverables are outlined below, along with notes on items that we feel need elaboration.

Minimum Viable Product Deliverables:

Completed: (1.1), (1.3), (1.4), (1.5)
Partially Completed: None
Not Completed: (1.2)

1.2) Determine a level of driver aggressiveness that would represent an autonomous vehicle:

Determining a level of driver aggressiveness has proven to be difficult due to the fact that the VED does not contain any AV data. To counter this, we investigated various third-party datasets but all had shown no promise with respect to our calculation needs -- lacking the time series spatial or velocity data that we'd need to calculate aggressiveness. As explained in the Midterm Report in Changes/Problems, we decided to assign AV aggressiveness to a low percentile of our dataset for subsequent product deliverables.

1.4) Estimate the potential total reduction in fuel consumption over the entire data set

We report on the potential fuel reduction in the "Significant Findings" section of this report. Note that we split this analysis by level of electrification because they behave fundamentally differently.

1.5) Develop a simple tool to display fuel consumption and potential reduction as a function of vehicle speed for both individual vehicles and fleets:

As per the request of the NVFEL in increasing readability and interpretation of the simple tool, this deliverable was modified to displaying the **fuel rates** of vehicles and potential reduction in fuel rates as a function of vehicle speed for both individual vehicles and fleets instead of displaying fuel consumption. It should be noted that due insufficient data for PHEVs (Plug-In Hybrid Vehicles) and BEVs (Battery Electric Vehicles), there is no analysis for those vehicles.

Expected Product Deliverables:

Completed: (2.1), (2.3), (2.4)
Partially Completed: (2.2)
Not Completed: None

2.1) Incorporate the effect of other data variables, particularly external temperature and air conditioning use. Determine to what extend these affect fuel consumption and the potential decrease due to autonomous vehicles:

Due to the VED providing sparse values for air conditioning use, as well as a lack of methods to estimate the power utilized by air conditioning on a trip, air temperature was used as a proxy for air conditioning use as a feature, as a high air temperature might indicate high air conditioning use, and vice versa.

2.2) Divide the driving by trips or trip segments into a "trip type" range of urban, rural, or freeway driving based on average speeds and frequency of stops:

Trips were classified into discrete categories for grouped regressions by Average Speed and Speed Variance over the course of a trip. However, it was quickly discovered that these grouped regressions performed less well than a single larger regression without discretization of trip-type. Speed Variance was used as a proxy for the frequency of stops as a high Speed Variance would indicate large amounts of stopping and starting, however, this variable was entirely uncorrelated with fuel consumption, and discretely splitting Speed Variance into a categorical variable did not improve performance. As a result of the above experimental results, dividing the driving by trip-type was discarded.

2.3) Expand the simple tool to display fuel consumption and potential reduction as a function of temperature and trip type:
As explained in the notes for Deliverable (1.5) in the above MVP section, the simple tool displays fuel rates rather than fuel consumption. In addition, due to our findings implying the lack of impact air temperature has with respect to our dataset as well as fuel consumption (explained more thoroughly in subsequent sections), air temperature was not included as a feature in our model and thereby not included in the simple tool. As per the recommendation of the NVFEL, due to this lack of impact, air temperature was not investigated as a factor impacting the fuel consumption or driver aggressiveness of HEV vehicles.

Stretch Goal Deliverables:

Completed: (3.2)
Partially Completed: (3.1.1), (3.1.2)
Not Completed: (3.1.2), (3.3)

3.1.1) Determine to what extent precipitation and wind magnitude and direction affects fuel consumption and potential reduction due to AVs:
As explained in the Design Plan, wind magnitude and direction was not extracted nor analyzed as a factor affecting the fuel consumption and potential reduction due to many complications regarding the aerodynamics of car in the face of different wind speeds and directions. Precipitation, however, was analyzed as factor in form of rain and snow levels.

3.1.2) Add the wind and precipitation effects to the simple tool:

While neither wind nor precipitation was added in our final models, we did at one point include precipitation in our tool, but found it to have extremely negligible effect on fuel consumption.

3.2) Refine the tool to allow a mix of trip types as input:

The Dashboard was designed to allow a virtually infinite sequence and combination of inputs. As discussed in deliverable 2.2, discretization of trips by type was disregarded as it did not improve performance. Instead the model treats input variables as continuous and as a result, the dashboard accepts continuous values as an input for trip types. By enabling the manual addition of vehicles and the trips these vehicles partake in, any combination of input variables can be considered. Furthermore, the fleet generation capability of the dashboard enables any number of trips and vehicles to be sampled from a normal distribution with mean and variance specified by the user. By enabling a continuous distribution of variables, users can specify a much broader spectrum of trips and vehicles.

3.3) Refine "trip type" by considering the posted speed limits for each trip segment:

Due to time constraints, we were not able to achieve this goal.

Major Activities

The work performed to complete this project can be divided into three categories: preprocessing/data extraction, a backend component, and a frontend component. The frontend component of the project is primarily concerned with the user interface of the tool, while the backend is concerned with the model that the tool uses to predict fuel consumption reduction over a trip, given various factors.

Preprocessing/Data Extraction

A great deal of the project consisted of processing raw trip data and extracting pertinent information. This has consisted of the following tasks:

1. **Dividing the dataset into individual trips taken by each vehicle, and by vehicle type**
2. **Extracting non-existent features from external sources (e.x. APIs)**
    a. Air Temperature (F)
    b. Precipitation Levels (mm)
3. **Calculating and/or preprocessing metrics and features for each trip:**
    a. Total Distance traveled (km)
    b. Average Speed + Variance (km/hr)
    c. Driver Aggressiveness Scores
        i. Positive Kinetic Energy (PKE)
        ii. Power Factors
    d. Fuel Rate (Gallons per Mile) + Fuel Consumption (L)
    e. Vehicle Weight (lb)
    f. Vehicle Engine Displacement (L)

(1) Dividing the Dataset

As described above, the dataset was divided into individual trips taken by each vehicle, and all trips were analyzed together based on their vehicle's hybridization. It is worth noting that although the VED description mentions PHEVs (Plug-in Hybrid Electric Vehicle) and BEVs (Battery Electric Vehicle), the total number of PHEVs and BEVs was too low to allow for sufficient analysis in relation to the features outlined in (3) above. Because of this, only ICE (Internal Combustion Engine) vehicles and HEVs (Hybrid Electric Vehicles) from the VED were analyzed in this project.

(2) Extracting Features from External Sources

- Air Temperature (F), Precipitation Level (mm): Due to the need for historical weather data, it was not feasible to make numerous API calls to the chosen API, OpenWeatherMaps. Instead, a dataset outlining various weather features including air temperature and precipitation levels for every hour from November 1st, 2017 to November 9th, 2018 were retrieved from OpenWeatherMaps. Then, for each trip, the total trip time as well as the trip date was calculated and used to index into the weather dataset. From this, the air temperature and precipitation levels (as a total of rain and snow) were extracted for each trip.

(3) Calculating and/or Preprocessing Metrics and Features for each Trip
- Total Distance (km), Average Speed (km/hr), Speed Variance (km/hr): No extra calculations were needed to include these features. In terms of calculating the total distance of each trip, the Haversine formula was used to calculate the distance between data points within each trip, whose sum yielded the overall total distance of each trip.

- Driver Aggressiveness (PKE (units) + Power Factors (units)): In terms of PKE, some preprocessing/calculations had to be completed to find the change in velocity between the data points in each trip. These values were then used with the total distance of the trip to calculate a PKE value for the trip. In terms of Power Factors, the same route was used in calculating velocity and acceleration, which was then used to calculate the root mean square of several data points.

- Fuel Rate (Gallons per Mile) +Total Fuel Consumption (L): The total fuel consumption for each trip was calculated by first computing the instantaneous fuel consumption (or fuel rate) for each data point within a trip. These rates would then be summed up to yield a total fuel consumption value for each trip. The algorithm and features needed to calculate the instantaneous fuel consumption and overall fuel consumption are referenced in the Design Plan, in Concepts Considered, section 2, method 2.

- Vehicle Weight (lb), Vehicle Engine Displacement (L): No extra calculations were needed to include these features, only a few steps to extract the numerical values from their labels in the dataset.

Backend Components

A significantly large portion of the project dealt with investigating the individual and collective relationships between some of the extracted features and fuel consumption with respect to driver aggressiveness. These investigations would ultimately provide information for final models used for the frontend portion. This has consisted of the following tasks:

1. Running various linear regression models to investigate the impact and credibility of both PKE and Power Factors as accurate driver aggressiveness metrics.

2. Running various linear regression models to investigate the relationship between various features, driver aggressiveness, and overall trip fuel consumption.

3. Creating a fleet generation module allowing the user to mimic any city's drive cycles based on a given mean and variance for various features.

(1) Running Linear Regression Models to Choose Driver Aggressiveness Metric
This was done by conducting simple linear regressions on our ICE data -- one with PKE-defined aggressiveness as the independent variable, and one with power factors. In both cases, the dependent variable was the fuel consumption of the trip. Because PKE-defined aggressiveness completely outperformed power factor aggressiveness (the latter showed no correlation), we felt confident enough to make our decision after this assignment.

(2) Investigating the Impact of Features
The basic blueprint for these investigations consisted of running multiple linear regression models with an input of various preprocessed features outlined above (more importantly, including a metric for driver aggressiveness), and the output being total fuel consumption. The R-squared values of these models would then be compared to each other to make some assumptions of which models might be the best to consider within the prototype, and ultimately be used to calculate a potential reduction in fuel consumption. The results of these investigations/models are outlined in the Significant Findings section.

(3) Fleet Generation Module
Although the drive cycles of a city cannot be fully and accurately simulated, we provided the user the option of generating a fleet given the mean and variance for various vehicle specific features such as weight and displacement, as well as various trip-specific features, such as average speed. These mean and variance values were then used to pull random values from a normal distribution and generate a variety of vehicles and the trips they embark on.

A key assumption of linear regressions is that of multivariate normality: that the residuals are normally distributed. On top of this, when modeling the empirical data, it was seen that a single gaussian distribution was the most likely statistical distribution to fit to each of the input variables. As a result, our dashboard is designed to sample trips and vehicles from a normal distribution specified by the user. This upholds both the assumption of multivariate normality as well as the distribution of the empirical data.

Frontend Components

The frontend portion of the project satisfies the following goals:

1. Display fuel consumption and potential reduction as a function of vehicle speed for

both individual vehicles and fleets.
2. Display fuel consumption and potential reduction as a function of trip type
3. Allow the user to specify the characteristics of a vehicle and its trips as well as create a fleet of vehicles given the option of specifying each characteristic.

To this end, the dashboard contains two sections, 1 for individual vehicle interpretation and 1 for fleet interpretation. As the dashboard is too large to fit into one image, images of individual portions of the dashboard will be featured in the following sections.

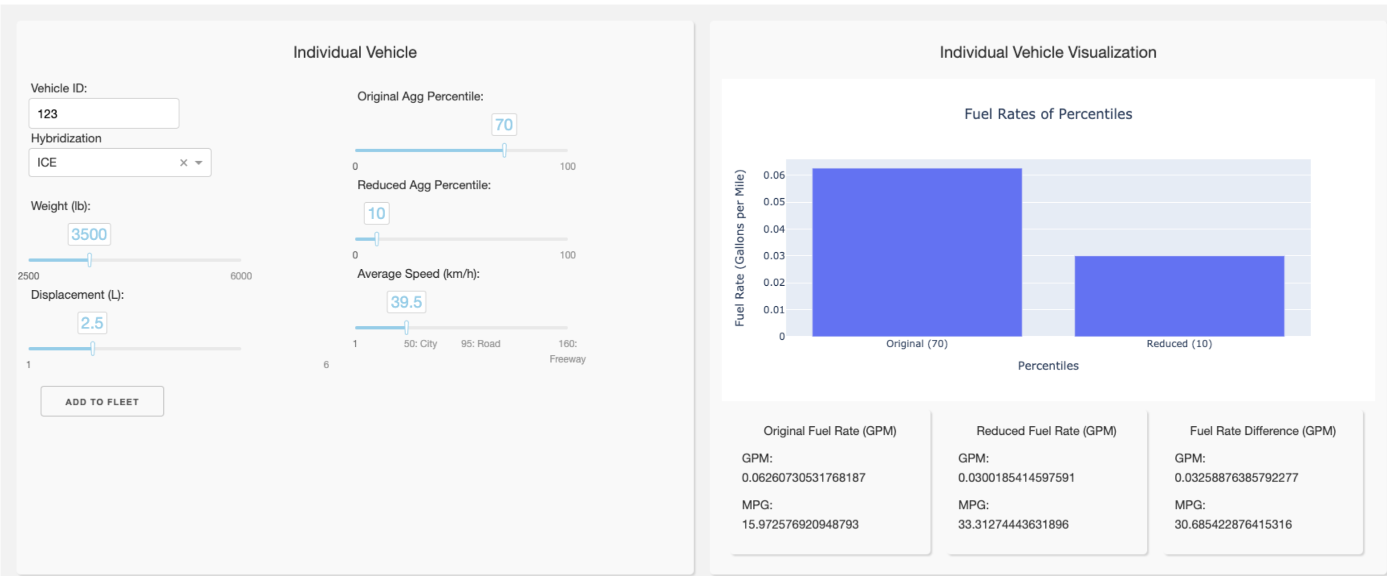

Figure 1: Individual Vehicle Portion of Dashboard

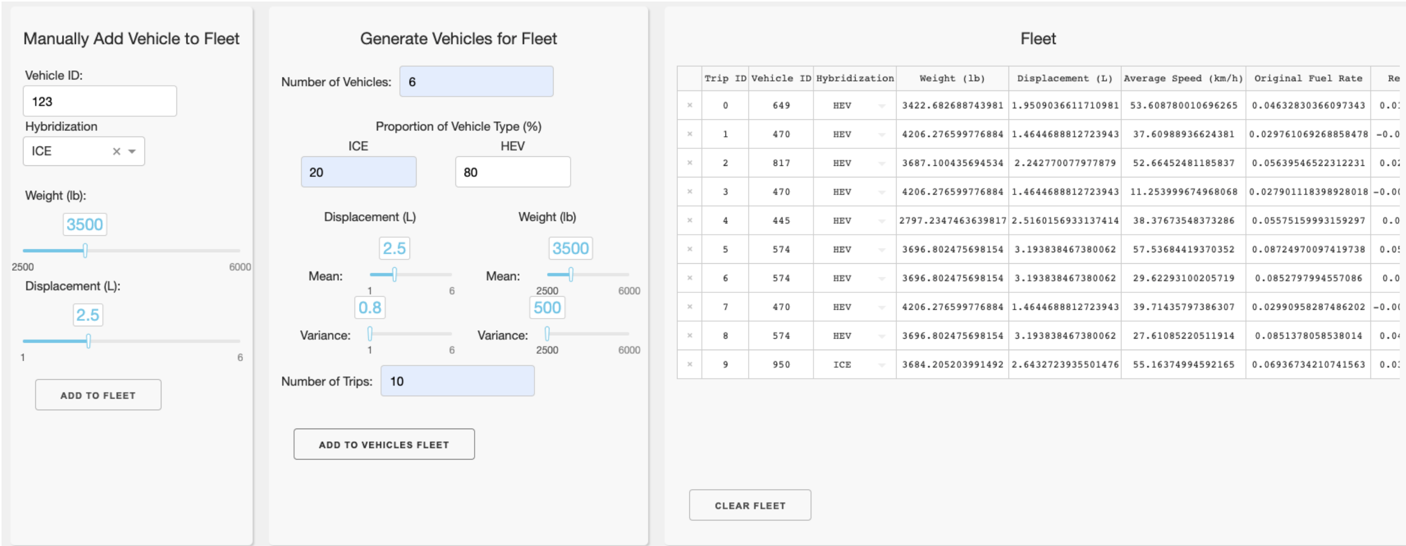

Figure 2: Individual Vehicle Portion of Dashboard

Figure 3. Fleet Visualizations + Adjustments

Individual Vehicles

Validation
The validation of the dashboard consisted of a mix of feedback discussions with the NVFEL as well as outlining a large list of combinations of inputs along with the respective outputs, and checking whether the dashboard met each standard.

Significant Findings

The following sections outline major findings made while working on the project.

Data Exploration

Internal Combustion Engine (ICE) Vehicles
ICEs were by far the most well-represented vehicle in the dataset, making up about 70% of the trips. Moreover, they were so we were able to extract data that reasonably modeled fuel consumption with respect to driver aggressiveness.

Hybrid Electric Vehicles (HEV)
A great deal of fuel consumption could not be calculated for HEVs. This is because a significant portion of the time series data simply did not have fuel data nor proxy variables to estimate the fuel data. This made much of the data too noisy to perform any reasonable analysis on. Because of this, over 90% of the 9,000 HEV trips were omitted, leaving only about 500 "clean" trips. Because the smaller dataset yielded more meaningful results, that is what was used for training our final model.

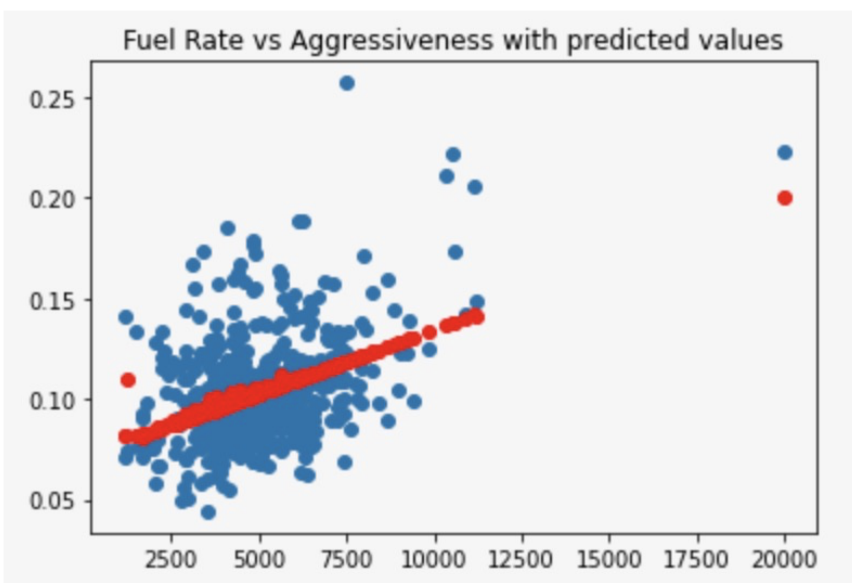

Figure 4: Analysis performed on "clean" datapoints

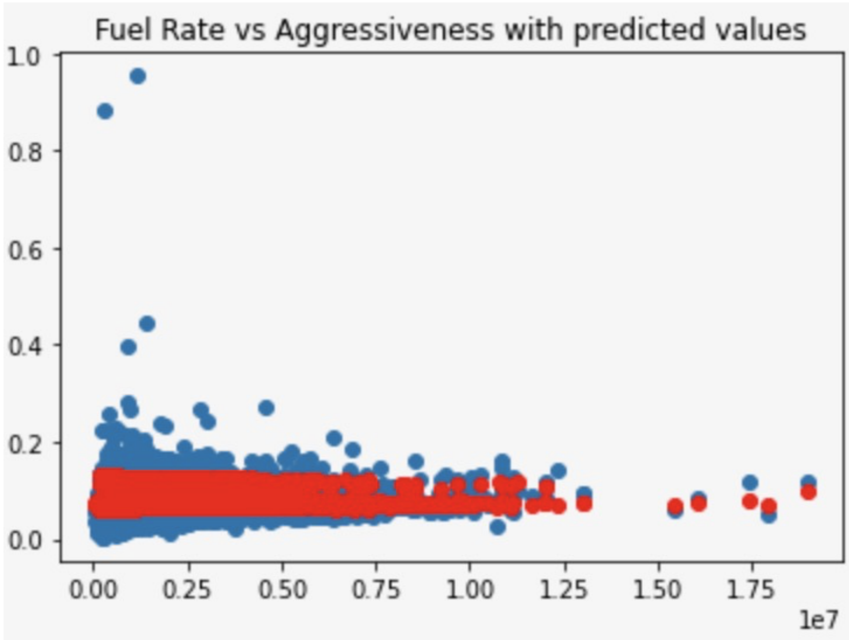

Figure 5: Analysis performed on full HEV dataset

Plug-In Hybrid Electric Vehicles (PHEV)

PHEVs had the same issue as HEVs and were even less abundant in the dataset. There was only one trip that was “clean” in the PHEV dataset.

Conclusions

The dataset was somewhat wanting in terms of consistency. Because fuel consumption is the crux of this project, the lack of fuel data availability was a significant hindrance to the project.

Determining a Driver Aggressiveness Metric

There were two metrics considered for measuring the aggressiveness of a vehicle: PKE and power factors. Our decision to use PKE-defined aggressiveness was largely influenced by the resulting figures 1 and 2 below:

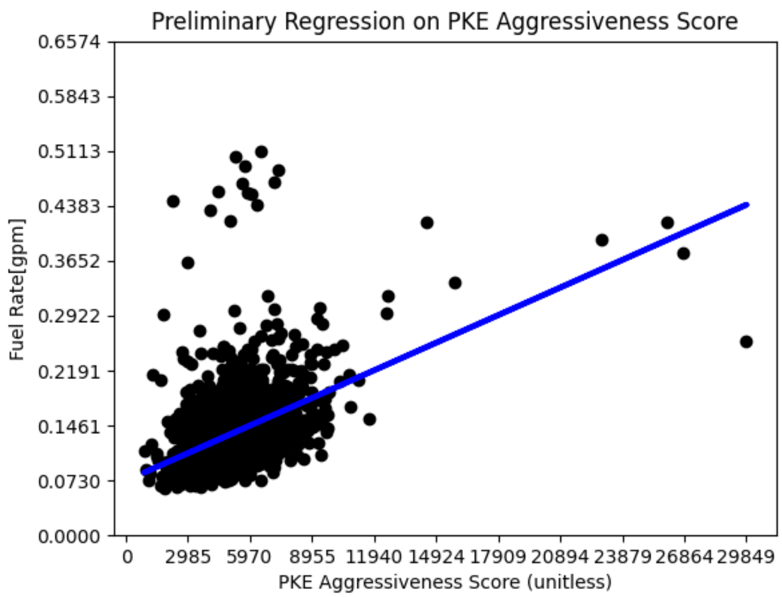

Figure 6: Regression performed on validation data (15% of total dataset) using PKE aggressiveness. $R^2 = 0.1637$

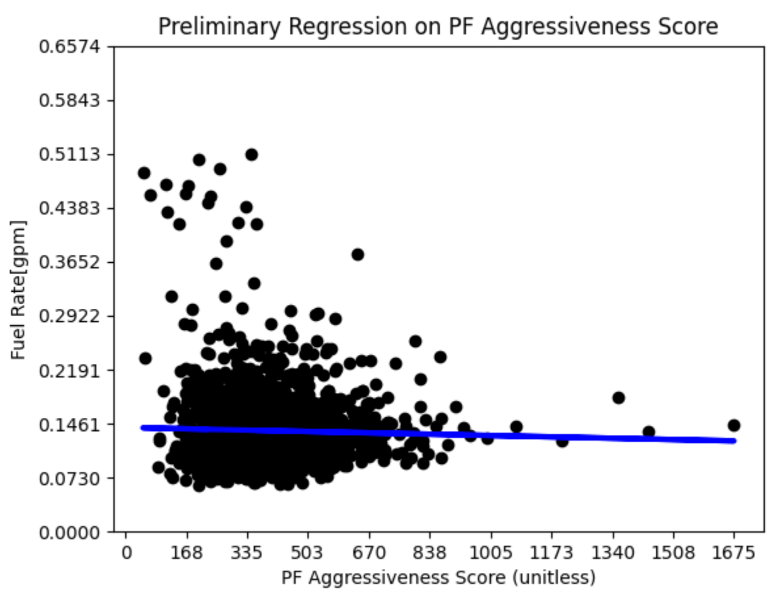

Figure 7: Regression performed on validation data (15% of total dataset) using PF aggressiveness. $R^2 = 0.0025$

Conclusions

Upon evaluating this result, we concluded that the disparity in performance between the two metrics was overwhelming; this test was sufficient evidence to use PKE as our aggressiveness metric for all models used by the dashboard.

Investigating the Impact of Features on a Linear Regression Model

The impact of individual features on ICE and HEV fuel consumption trips was performed in the same manner. For each independent variable in the fitted linear regression, the impact was calculated as a multiple of the coefficient and the mean value of the input. This allowed us to determine the relative impact of a variable on fuel consumption. Further, analysis was also completed by normalizing each of the input variables one at a time with respect to the variable's variance, in order to see the effect of variance on the results. Finally, an analysis was performed where the min-max values for each input variable were passed into the regression to determine the total effect of that variable on the model, this min-max analysis was also done using iterative percentile splits: (10,90), (20,80) … (40,60). This allowed us to determine the effect of extrema on the model's performance.

Internal Combustion Engine (ICE) Vehicles

A confluence of the above values were used to select the features that were used in the ICE regression. Results from first method of determining impact: mean * coefficient is reported in table 1 below.

Variable	Impact % (mean value * coefficient)
Weight	21.634
Displacement	37.04
Aggressiveness	3.29
Aggressiveness * Weight	10.64
Aggressiveness * Displacement	7.29
Displacement / Weight	18.60

Table 1: Impact of each variable on ICE regression

The variables above were selected as throughout every impact measurement they were consistently the highest. As seen, Aggressiveness itself is a small portion of the model's predictive power, but when interacting with both weight and displacement, aggressiveness captures a significant portion of the impact on fuel consumption. This result aligns with both our expectations and previous research. Weight and displacement independently, and when combined as Displacement / Weight to act as a proxy for power to weight ratio, are the highest impact variables in our regression. Speed Average and its interaction with Aggressiveness were considered, and are included in our model, but from an impact perspective they do not affect the actual values of the regression, rather they serve as useful and statistically significant control variables.

Hybrid Electric Vehicles (HEV)

For hybrid electric vehicles the story is very similar to ICE vehicles, as the impact for each component is dominated by displacement, weight, and their interaction with each other. Table 2 below illustrates results from using the first method of impact analysis for the largest-impact variables in our model.

Variable	Impact (mean value * coefficient)
Aggressiveness	10.11
Weight	6.86
Displacement	35.79
Aggressiveness*Weight	3.52
Aggressiveness*Displacement	1.89
Weight*Weight	12.99
Displacement/Weight	28.39

Table 2: Impact of each variable on HEV regression

As mentioned in the ICE vehicle section, these variables were selected for the regression as they performed the best across all splits and combinations of regressions, and were consistently high-impact across every metric. An interesting note here is that the squared value of Weight is an incredibly impactful metric for our regression, whereas this variable was not impactful in the ICE vehicle regression.

Plug-In Hybrid Electric Vehicles (PHEV)

As PHEVs had very few trips, they were not considered in our analysis.

Aggressiveness and Fuel Consumption

The major finding supported by our analysis is that if an autonomous, gasoline-powered vehicle is programmed to drive significantly less aggressively than the median trip in our dataset, we do not see significant fuel reduction. Note, though, that this dataset is limited in size and may not be representative of all vehicles and drivers on the road. Moreover, the data is quite noisy due to our estimation of fuel consumption, implying that these results may not be valid. Results from our analysis over the whole ICE dataset is shown in table 1 below.

	Aggressiveness (PKE)	**Fuel Consumed (gal)**	**Percent Reduction**
Real Data	5202	8040.25	0
AV (10th percentile aggressiveness)	3377.4	7888.51	1.9

Table 3: Fuel Consumption w.r.t. Driver Aggressiveness Over ICE dataset

Other Accomplishments

Most of our accomplishments have been directly related to the project deliverables. Those that are not have been listed briefly below:

1. Using a Neural Network instead of a linear regression model (poor results)

Team Member Contributions

Each team member's contributions are listed below (note that Zubin and Isha's primary roles have largely switched with respect to the design plan):

Gabe

Primarily responsible for data cleaning and analysis. Contributed to the following tasks:

1. Dividing dataset into trips + grouping by vehicle type (ICE, HEV)
2. Computed PKE and Power Factor aggressiveness score for each trip
3. Computed Fuel Consumption for each trip
4. Added other features to each trip
    a. Weight
    b. Displacement
5. Evaluated the strength of PKE and Power Factor aggressiveness as independent variables for regression analysis
6. Performing regression analysis grouped by weight and displacement

Isha

Primarily responsible for dashboard development and external information retrieval.

1. Extracted following features to each trip:
    a. Air Temperature

    b. Precipitation Level
2. Evaluated strength of Air Temperature and Precipitation Levels as independent variables and as part of interaction variables for regression analysis
3. Designed and adjusted layout of dashboard for ease of use
4. Implemented display and integration of finalized models with dashboard for:
    a. Individual Vehicles
    b. Fleet of Vehicles
5. Implemented display and integration of fleet generation module with dashboard

Zubin

Primarily responsible for model development.

1. Implemented the fleet generation module as outlined in the Frontend section of Accomplishments
2. Evaluated strength of Distance, Weight, Displacement, Average Speed, and Speed Variance as independent variables for regression analysis
3. Evaluated the impact of interaction variables on various regression models.

Changes/Problems

Changes in Approach

We generally adhered to the specifications laid out in the design plan, with a few alterations, as outlined below:

1. Determining a Level of Driver Aggressiveness that represents an AV

2. Displaying Fuel Rate instead of Fuel Consumption on the dashboard

3. Omitting PHEVs and BEVs From the Dashboard

As stated previously, we did not include PHEVs in our tool due to the lack of data for these vehicles.

Problems or Delays Experienced and Corrective Actions Taken

One of the primary delays experienced was in determining a level of driver aggressiveness that represents an AV. This mainly stemmed from the fact that the VED did not contain any AV data, and third-party datasets that were investigated did not show promise. This was because at a minimum, both temporal and spatial data are needed to calculate driver aggressiveness, but no third-party datasets investigated contained either of these. As a result, AV aggressiveness was assigned to a low percentile of the dataset as described in the design plan

<u>Impact</u>

Impact on the Environment

The dashboard is designed to advance the Environmental Protection Agency's (EPA's) goal of "improving air quality." It does this by providing NVFEL members with an estimation of fuel reduction based on driver aggressiveness. This, in turn, can be used to evaluate fuel consumption reduction of an AV, which allows them to better regulate these vehicles and prepare for a future with AVs.

Impact on Individual Project Members

Gabe: Consciously thought about design decisions that had to be made for a project and the tradeoffs associated with my choices. This contrasts with most of the academic work I've done because most of my projects have involved adhering to a specification. I also gained a great deal of experience with feature engineering and feature extraction, something that I didn't feel was heavily practiced in classes I've taken before. In addition to this, I gained greater familiarity with using Python's Pandas and Numpy libraries for data manipulation. These are all things that I think will serve my professional career well, due to the fact I feel all of these skills are highly relevant in industry.

Isha: Similar to Gabe, learned to think about design choices as well as the respective tradeoffs with each design decision. This was very prevalent in the dashboard construction phase, as each component had to be placed in a way that would be easy for interpretability by the user. I was also able to gain quite a bit of experience with the Plotly Dash framework and gain more understanding into various statistical methods. I believe that these skills will help me quite a bit at the beginning of my professional career.

Zubin:
This project was exceedingly helpful in growing my project management skills. It is rare that I have the opportunity to continually seek feedback from a client and produce client-facing reports, meeting agendas, and documents. I found the modelling work exciting and reinforced my statistical skills as a result, but definitely grew the most in my project management role.