Project Final Report MGT 6203 – Team 40

Brian Nutwell; bnutwell3 | Sanat Talmaki; stalmaki3| Kseniya Ruta; kruta3
Matthew Kesselman; mkesselman3 | Sergey D'Alessandro; sdalessandro3

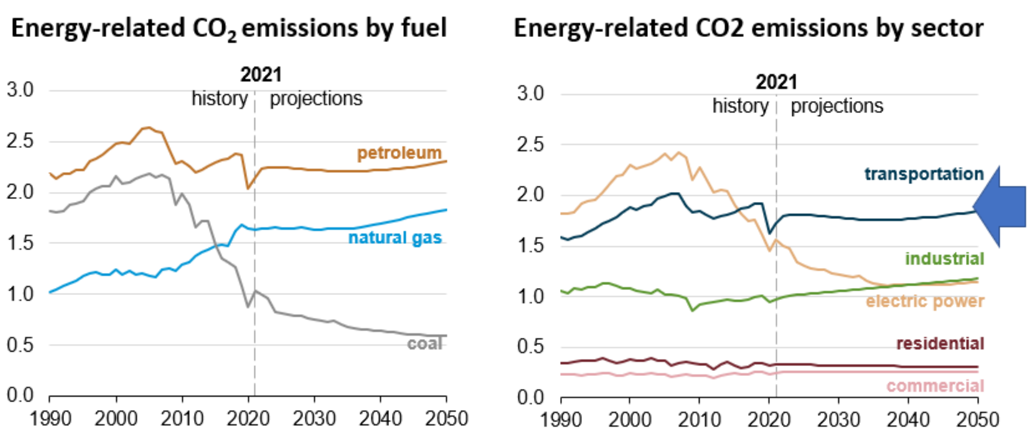
Topic, Business Justification, Problem Statement

The transportation sector generates the largest share of greenhouse gas emissions in the United States. Most governments, including the United States, actively promote sustainable development by setting stringent targets. One of the most effective methods governments use to encourage sustainable development is offering incentives to individuals and businesses.

But why should somebody want their business to go green? Our assertion is that, while initial investments in green energy may be costly, they can save businesses thousands of dollars in the long run. Economic benefits of embracing sustainable business practices include enjoying tax benefits and grants from federal and state governments.

Other benefits of being a more sustainable business include:

- Attracting climate change-conscious customers
- Improving brand image and aligning with shareholder expectations for sustainability
- Reduced carbon footprint, which indirectly and directly contributes to improved public health and reduced pollution
- Increasing competitive pressure on other businesses in your segment, to demonstrate their own sustainability commitments
- Increasing ability to comply with increasingly severe carbon and climate regulations.



Source: Annual Energy Outlook 2022 (AEO2022) – U.S. Energy Information Administration

Figure 1 – Key drivers of CO2 emissions are petroleum (source) and transportation (sector)

Thus, there is much business value to be gained by going green. The transportation sector is a key battleground in this pursuit, as it has become the dominant US emissions-producing activity-- see *Figure 1*. Encouraging the purchase of efficient vehicles which use renewable / low-emissions energy has become a high priority.

It is, however, difficult to directly compare and price the emissions of different vehicle types. Though electric vehicles are increasingly popular and produce no direct carbon emissions themselves, they do demand power from the local energy grid. Local energy grids have dramatically different emissions based on type of energy source they use (NREL, 2016) which affects an EV's relative impact on emissions (Saber et al, 2011). Even some sources considered "renewable", such as biomass or waste-to-energy, produce emissions comparable with fossil fuels (EIA.gov, 2021).

A consumer or business desiring to minimize their carbon footprint should consider this full stack, but information to make direct comparisons is not readily available. This lack of transparency in the full cost of vehicle usage leads to our problem statement:

Create a tool that helps consumers and businesses identify the most emissions-friendly and cost-effective vehicle(s) to purchase, based on purchaser location and driving preferences.

To address this challenge, we propose to build a decision support which connects several domains of information: CO2 emissions from various energy sources, local grid power composition, new vehicle characteristics, and user location & needs.

Data and Data Wrangling

Dataset collection and processing is shown in *Figure 2* below.

Dataset	#Rows	#Cols	Purpose
US Energy Grid	26	200	CO2 / MWh regression
State Electricity Profile	50	10	Power mix reference
Global Powerplant list	34396	36	Local energy mix clustering
Zip code / lat-long lookup	33144	3	Find nearest cluster
Vehicles	113	10	Primary output table
Energy prices by region	50	2	Calculate EV usage costs
Fuel prices by region	50	2	Calculate ICE usage costs

Figure 2 -- Data sources , size, and purpose

Explicit cleaning steps included:

- Identifying the essential columns among many attributes in the US Energy Grid dataset.
- Transforming the state electricity profile dataset. It includes 2 separate sheets: energy mix (%) by source, and total net energy generation, MWh. We joined these 2 data sets to calculate total energy by source, as an input for our regression model.
- Trimming the global powerplant list to only powerplants located in the USA, and selecting only the columns of interest.
- Sourcing vehicle information. A single downloadable source of current model vehicles and attributes was surprisingly difficult to locate. The team manually scraped 113 unique

vehicles from public websites including kbb.com and truecar.com. Data points collected include vehicle type, make, model, price, and fuel usage (city/highway).

- Units conversion to mi/kWh from MPGe for the vehicle dataset.

Analysis Methodology and Discussion

The concept for the data and analysis pipeline and workflow is shown in *Figure 3* below.

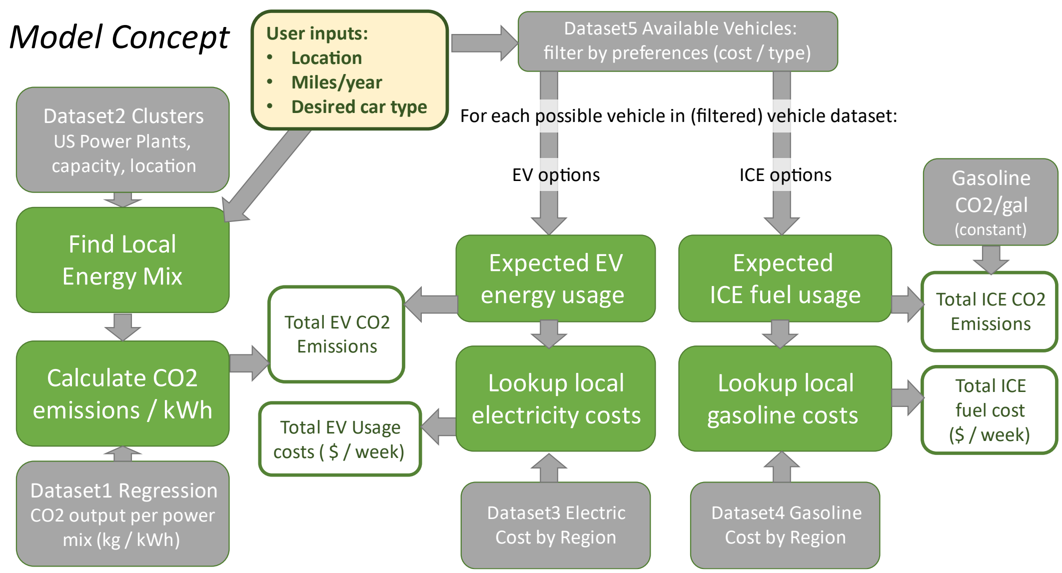


Figure 3 -- Project datasets (grey), calculation steps (green), and output (white)

The key elements in this analysis pipeline include:

- We performed a **Linear regression** on the US Energy Grid (27 regions) to identify emissions per energy usage from each possible source (coal, oil, gas, nuclear, wind, hydro etc.).
  - We iterated the regression to choose the most appropriate parameters from the 247 initial columns. After testing two different approaches and analyzing the outcome, we narrowed down on 12 columns most suitable to build the regression model (initial model). After removing outliers and performing a cross-validationm we kept a subset of 4 independent variables (final model). We also forced the intercept to zero, to avoid a large data offset when using the coefficients for small-capacity grid regions. See *Figure 4*.

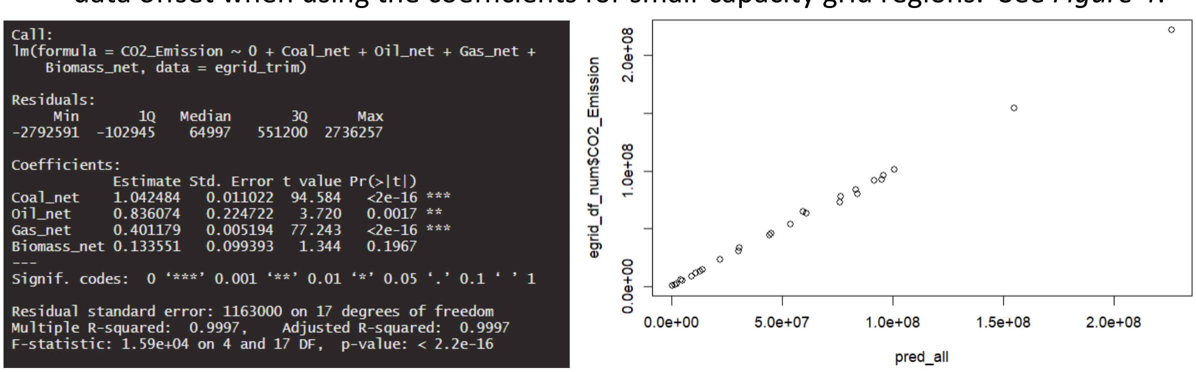
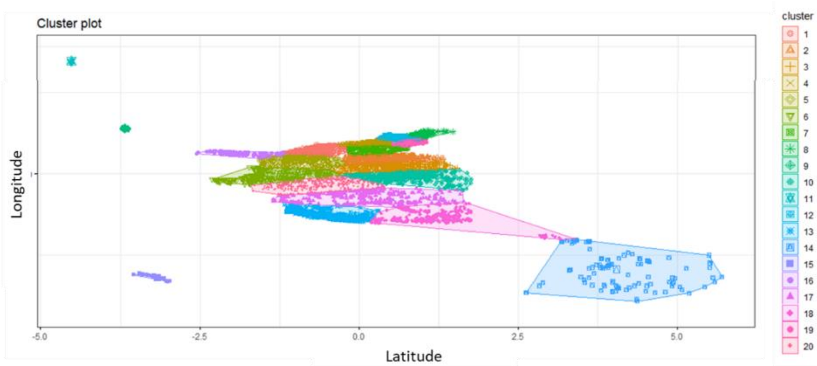


Figure 4 -- Linear Regression fitting with 16 datapoints / 4 predictors (left) and correlation with all 26 datapoints (right)

- Although the R-squared is quite high, indicating overfitting, the regression was validated by comparison to known CO2 emissions / MWh coefficents for specific sources from industry literature, and by forecasting emissions for any combination of energy sources – see the discussion in Sample Model Output below. The biomass term in particular is not judged as statistically significant in this regression – but based on actual physical phenomena, it would be incorrect to treat biomass emissions as zero. So we are leaving it in the prediction model.
- Another consideration is that powerplant emissions are highly regulated and monitored by the US government – it is logical to suppose that emissions/MWh for coal plants (for example) are controlled to meet a specific regulatory target, which would reduce the variance significantly.

- We **Clustered** the 10,000 local powerplants in the US into 200 clusters by latitude and longitude, and calculated the total and % energy generation by fuel type for each cluster. An example geographic clustering output is shown in *Figure 5*.



cluster	primary_fuel	generation_gwh_2019_per_fuel_type
1	Coal	1569.54802
1	Gas	103.84200
1	Oil	0.17044
1	Solar	9.34100
1	Wind	2088.33272
2	Gas	41.85746
2	Hydro	638.10700
2	Solar	60.09700
2	Wind	260.67100
3	Cogeneration	1132.09100
3	Gas	30885.68859
3	Oil	3.01000
3	Solar	715.37900

Figure 5 -- Example geographic clustering (left) and sample of power mix by cluster lookup file (right)

- We **Associated** each of the 33,000 ZIP codes in the US with the nearest geographic cluster. We then built code to created a **lookup file** for power grid composition by ZIP code. This file is referenced by a simple lookup function that could be called directly from the UI. Generating the large reference file took ~90 minutes of processing time in advance, but

```
# Function that can be called with zip code to return power grid characteristics
#  Returns:
#     - if resptype = "coeff" (default) --> Power coefficent (numeric) in metric tons CO2 / MWh
#     - if resptype = "mix"  --> Power mix by source for local cluster (dataframe) in %
get_my_power_mix <- function(inputzip, resptype = "coeff") {

  #Determine the cluster the user belongs to
  my_coord <- zip_df[zip_df$ZIP==inputzip,2:3] # User's coordinates

  #Find the cluster to which the new object is closest (in terms of the sum of squares of distances)
  my_cluster <- which.min( sapply( 1:200, function( x ) sum( ( my_coord - km$centers[x,])^2 ) ) )

  #Subsetting the data set based on the user's cluster
  local_per_source <- local[local$cluster==my_cluster,]
```

```
# create a lookup function to reference the coefficient file
# input zip code, get back a coefficient in mt CO2 / MWh
coefflookup <- function(zipcode) {
  ziprow <- zipcoeff %>%
    filter(zip == zipcode)
  cf = ziprow$Coefficient
  return(cf)
```

X	Coefficient	zip
15436	0.38458	47272
15437	0.38458	47273
15438	0.38458	47274
15439	0.38458	47280
15440	0.38458	47281
15441	0.38458	47282
15442	0.38458	47283
15443	0.21312	47302
15444	0.21312	47303
15445	0.21312	47304
15446	0.21312	47305
15447	0.21312	47306
15448	0.21312	47320
15449	0.71071	47324

Figure 6 -- **get_my_power_mix()** function (fragment shown at top left) was used to generate lookup file (right), which is referenced by **coefflookup()** function (bottom left)

allows the UI to quickly retrieve coefficient values during user interaction rather than running coefficient calculations on the fly. Headers for utility functions are shown in *Figure 6*, along with a sample of ZIP codes associated with different clusters.

- We created a similar **lookup file** by ZIP code for fuel and energy prices.
- **User interaction** is accomplished through an Rshiny dashboard, which allows for model exploration with a reactive UI. Code for an example reactive table is shown in *Figure 7*. Reactive elements include required UI states, as well as **dplyr** commands to manipulate the vehicle table and generate output in real time. These tables are connected through an architecture of *input pane* → *reactive data processing element* → *output data element* → *output pane*.
- Finally, we **display** the resulting coefficients and matching /recommended vehicle lists in the UI (see User Interface section).

```
# Build the dataset for EV vehicles based on inputs
datasetEV <- reactive({
  req(input$vehType)
  req((input$chargerAccess == "Yes" | input$weeklyMiles < 300) & input$freqRoadTrips=="Rarely")
  req(input$zipCode %in% zipcoeff$zip)
  req(input$hwyMiles)
  filter(veh_disp, VehicleType %in% input$vehType) %>%
    filter(FE_Unit == "MPKWh") %>%
    #filter(List_Price <= input$maxPrice) %>%
    mutate(WeeklyCost = input$weeklyMiles / (FE_City * (1- input$hwyMiles/100) + FE_Highway * input$hwyMiles/100)
            * powerpricelookup(as.numeric(input$zipCode),"MPKWh")) %>%
    mutate(WeeklyCO2_kg =  input$weeklyMiles / (FE_City * (1- input$hwyMiles/100) + FE_Highway * input$hwyMiles/100)
            * coefflookup(as.numeric(input$zipCode))) %>%
    arrange(WeeklyCO2_kg)
})
```

Figure 7 -- Example code snippet for Rshiny reactive data table (list of EVs with calculated cost & emissions)

Sample Power Grid Emissions Model Output

Figure 8 shows the results of our grid emissions analysis using four different ZIP codes, each with a unique power source mix. As expected, higher-fossil-fuel power mixes produce a much higher CO2 emissions coefficient than high-renewables regions; the exact levels are driven by the specific mix of power sources.

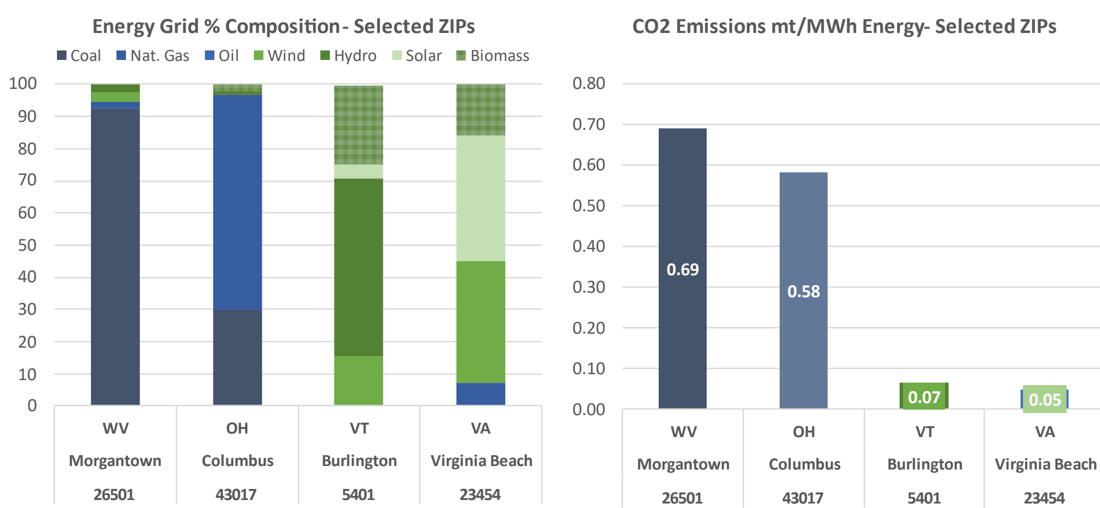


Figure 8 – Energy grid source mix (left) drives CO2 emissions characteristics (right)

One interesting observation: although the Vermont ZIP code in the table above appears to be entirely renewable energy sources, it has a significant % of “biomass” energy. This source, while it is renewable, does create carbon emissions equivalent to some fossil fuels and thus is not completely clean.

User Interface

The final UI is shown in *Figure 9*.

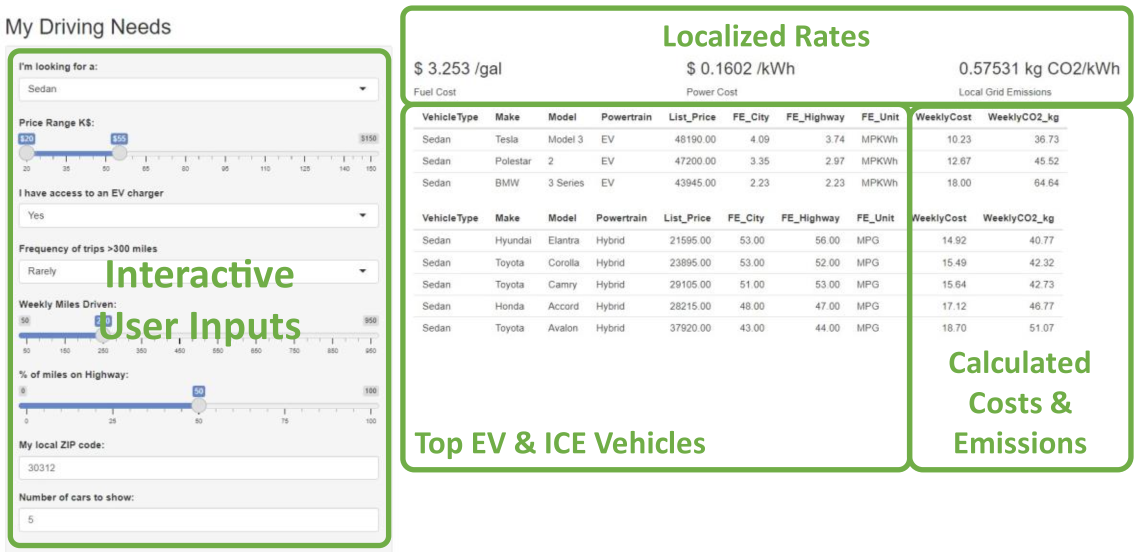


Figure 9 -- Final project working UI (green splined text added for context).

The interface is fully interactive, reflecting the user’s selections as filters while calculating information in real time. Available filters and impacts include:

- Filtering vehicle list by vehicle type and list price range (max/min), and maximum number of vehicles to display
- “EV charger access” and “road trip frequency” questions. (Users who frequently take long trips, or who have high weekly mileage needs AND don’t have access to a charger, may not be good candidates for electric vehicle usage with the current state of battery technology and charging infrastructure. The tool will not recommend EV’s to users with these combinations of inputs.)
- Weekly mileage and % of miles on highway are used to calculate fuel usage, costs, and emissions
- ZIP code, for referencing the user’s local power grid and fuel/energy costs

In addition, we built a variant UI targeted at business fleet customers who may wish to purchase multiple vehicles at a time and evaluate the overall impact (see *Figure 10*). Other functionality is as described above.

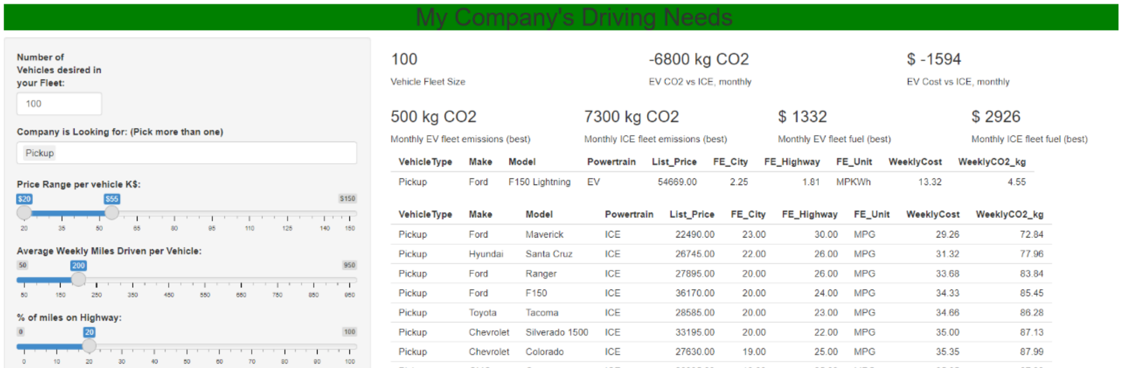


Figure 10 -- Variant UI for business fleet users

The entire project is constructed in an R Markdown notebook (one file each for the Consumer and Business UI variants), using the following R packages: **dplyr, readxl, factoextra, shiny, shinydashboard**. Excel files are used for various lookup tables and reference values as mentioned.

Results and Conclusions

In *Figure 11* below we can see the calculated weekly cost vs. emissions of some representative vehicles, from two different ZIP codes. Several key points are apparent, as highlighted in the numbered blue circles.

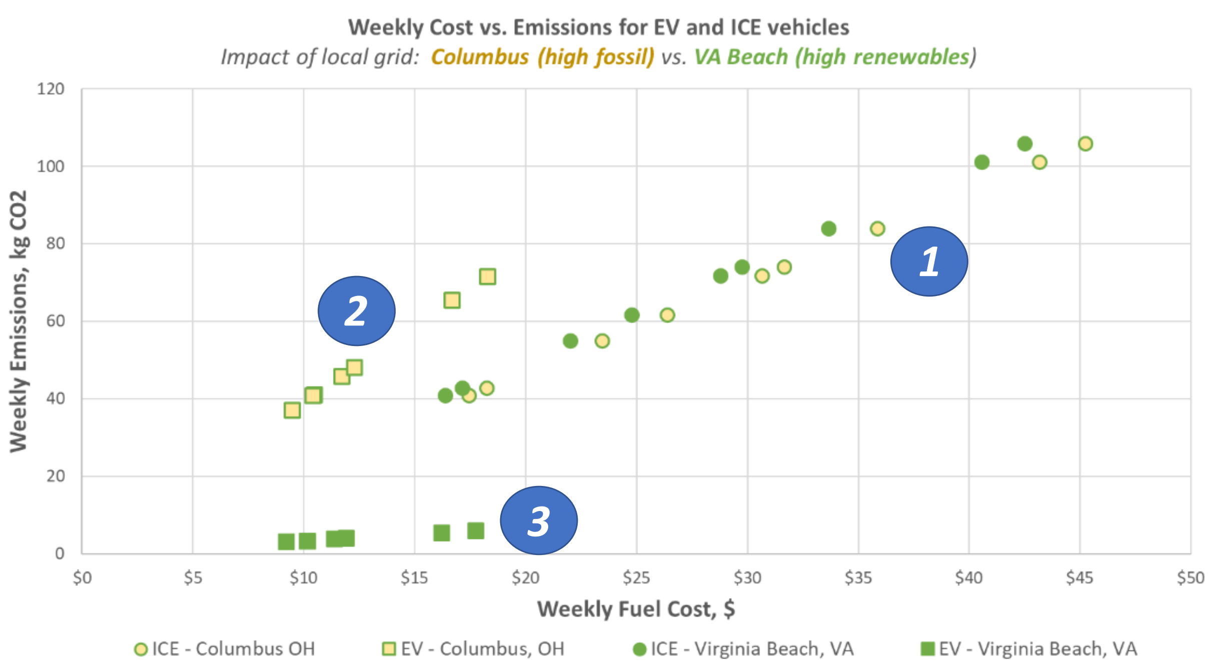


Figure 11 -- Weekly Operating Cost (x) vs. Emissions (y) for selected vehicles and locations

1. Internal combustion (ICE) vehicles (circles) have overall higher fuel costs and emissions
2. In a high-fossil fuel grid (yellow circles and squares), EV effective emissions substantially overlap with the most-efficient ICE vehicles.
3. In a high-renewable grid (green circles and squares), EV effective emissions are dramatically lower than even the best ICE vehicles.

These are tangible recommendations for a complex decision involving factors that are otherwise difficult for a consumer or business to analyze independently.

Figure 12 shows a sample vehicle recommendation list for several regions across the US. The scenario used for comparison is for a hatchback vehicle under $40,000, with up to 250 miles driven per week, and a 50-50 split between city and highway. This scenario is selected as it likely represents a typical customer's needs and a common purchasing decision for a large percentage of vehicle buyers who spend around $40,000 on their purchase (Autotrader, 2021).

Equally, the five regions selected represent a broad mix of energy sources ranging from over 90% coal-based power in Morgantown, West Virginia to over 80% renewables in Virginia Beach, Virginia. While EVs are often cheaper to run, they are not always the most carbon-efficient option, depending on where one lives and charges their EV. *Figure 12* shows hybrid vehicles having lower carbon emissions than EVs in locations such as Columbus, Ohio and Morgantown, West Virginia that have more non-renewable energy sources such as coal and natural gas. This tool helps a user identify the cheapest and most carbon efficient vehicle based on their location and matching their driving needs.

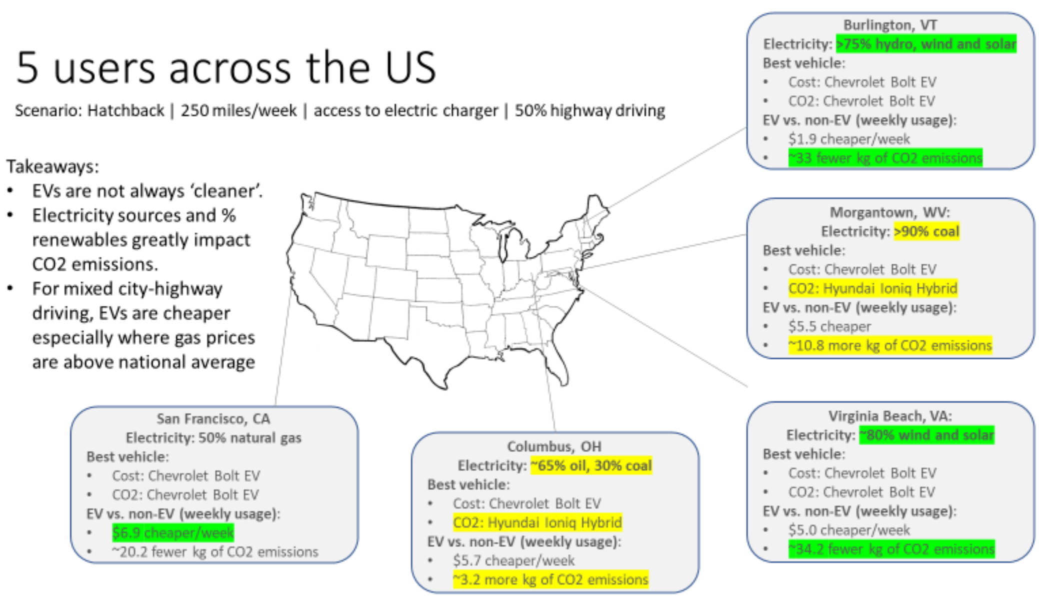


Figure 12 – Vehicle recommendations for the same use case in 5 regions across the US, with varying energy sources mix. In low renewables regions, hybrid powertrains have lower carbon footprint than EV counterparts.

Finally, as seen in *Figure 13* (detail of data from *Figure 10*), potential business benefits are substantial. For a fleet of 100 pickup trucks driving 200 miles/week, the fleet owner could **reduce their monthly fuel expenses by $1600 or 54%**, while also avoiding the generation of nearly **seven metric tons of CO2** every month, by selecting an electric vehicle in a high-renewable grid area.

100	-6800 kg CO2	$ -1594
Vehicle Fleet Size	EV CO2 vs ICE, monthly	EV Cost vs ICE, monthly

500 kg CO2	7300 kg CO2	$ 1332	$ 2926
Monthly EV fleet emissions (best)	Monthly ICE fleet emissions (best)	Monthly EV fleet fuel (best)	Monthly ICE fleet fuel (best)

VehicleType	Make	Model	Powertrain	List_Price	FE_City	FE_Highway	FE_Unit	WeeklyCost	WeeklyCO2_kg
Pickup	Ford	F150 Lightning	EV	54669.00	2.25	1.81	MPKWh	13.32	4.55

VehicleType	Make	Model	Powertrain	List_Price	FE_City	FE_Highway	FE_Unit	WeeklyCost	WeeklyCO2_kg
Pickup	Ford	Maverick	ICE	22490.00	23.00	30.00	MPG	29.26	72.84

Figure 13 – Detail of business impact for ZIP 23454

In conclusion, our project managed to build a tool that provides clarity to a very unclear but critical situation —the true impact of greenhouse gas emissions by vehicle use in the United States for **both** ICE and EV vehicles. The tool we have developed, powered by an assortment of data linking, cleaning, analyses, and user inputs, can provide visibility into this challenging question. The best choice will be specific to the purchase location, vehicle needs, and driving modes.

This transparency creates business value by allowing consumers and businesses to make better decisions for their costs, their brand, their future, and ideally, the planet.

Potential Expansions and Further Work

To further our efforts, we would look for more data sets to expand the accuracy and breadth of our recommendations. These would include:

- Paid API feeds of hyper-localized data for shifting fuel and electricity costs
- Additional vehicle information from paid services (like data on vehicle-specific ranges)
- Data on EV tax credits and finance credits, to calculate the true cost to own and operate a vehicle
- Data on auxiliary costs such as the embedded energy of producing a new car and recycling/disposal costs, to calculate the true climate impact of a purchase decision (EV vs. ICE as well as new vs. used)

We would also consider expanding our analysis outside of the United States, so any user or business throughout the world can understand the true impact of their vehicle purchases.

Appendix

References

1. Emissions Associated with Electric Vehicle Charging: Impact of Electricity Generation Mix, Charging Infrastructure Availability, and Vehicle Type, National Renewable Energy Laboratory, 2016. Source: https://afdc.energy.gov/files/u/publication/ev_emissions_impact.pdf
2. Plug-in Vehicles and Renewable Energy Sources for Cost and Emissions Reductions, Saber, A.Y. and Venayagamoorthy, G.K., IEEE Transactions on Industrial Electronics Vol. 58 Issue 4, 2011
3. Greenhouse Gas Emissions from a Typical Passenger Vehicle, U.S. Environmental Protection Agency (EPA), 2018, Source: https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P100U8YT.pdf
4. How waste-to-energy plants work, 2021, Source: https://www.eia.gov/energyexplained/biomass/waste-to-energy-in-depth.php#:~:text=Waste%2Dto%2Denergy%20plants%20burn,and%20products%20made%20from%20wood.
5. The Average New Car Price is Now Over $40,000, Autotrader.com, February 2021, Source: https://www.autotrader.com/car-news/the-average-new-car-price-is-now-over-40000

Data Sources

1. **CO2 emissions and power mix by US energy grid subregion**: “egrid2020_data_metric.xlsx”
    a. Description: Contains data for each of the 27 energy grid subregions in the US. The dataset contains CO2 emitted by each subregion as well as the power mix (i.e., sources of power) used by each subregion. This is used in the regression model with CO2 as the dependent variable and power mix as the independent variables.
    b. Source: https://www.epa.gov/egrid/download-data
2. **Global power plants database**: “global_power_plant_database.csv” (2021 version of dataset is used)
    a. Description: This dataset contains a list of 34,937 global power plants matched to their zip/postal code. It is used to cluster power plants to determine the local energy profile, where local refers to zip code, which is a user input for the application.
    b. Source: https://datasets.wri.org/dataset/globalpowerplantdatabase
3. **Electricity costs data by state**: “Electricity rates by state.xlsx”
    a. Description: Electricity costs for July 2021 and July 2022, and YoY change in costs. Electricity costs are expressed in cents/KWh.
    b. Source: https://www.saveonenergy.com/electricity-rates/
4. **Gas and diesel prices by state**: “Gas and diesel prices by state.xlsx”

a. Description: Prices of three grades of gas (regular, mid-grade, premium) and diesel fuel. Data pulled on 10/17/2022. Prices expressed in $/gallon.
b. Source: https://gasprices.aaa.com/state-gas-price-averages/

5. **Vehicles data**: "Vehicles_data.xlsx"
   a. Description: This data consists of the vehicle type (sedan/SUV/pickup/hatchback/etc.), make, model, price, powertrain type – EV/Hybrid/ICE (internal combustion engine), fuel economy for city, highway and combined cycle in both MPG for ICE and hybrid, and MPKWh (miles per kilo-watt hour) for EV.
   b. Source: Compiled manually from KBB.com and Truecar.com websites.
6. **Latitude, Longitude and state data**: "zip_code_database.csv"
   a. Description: This dataset maps every zip code to its corresponding US state. This is used to convert user input of zip code to find the corresponding gas and electricity costs.
   b. Source: https://www.unitedstateszipcodes.org/zip-code-database/
7. **CO2/gal and its gasoline equivalent**: "CO2 per gallon.xlsx"
   a. Description: CO_2/gallon emitted for gasoline and diesel fuel per EPA standards. This would be used together with fuel efficiency of a vehicle and estimated miles driven per year to compute total CO_2 emitted (tons or grams) per year.
   b. Source: https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P100U8YT.pdf