**Project Proposal:** Evaluating Electric Vehicles and their Real Carbon Footprint

**Student’s Name:** Jake Cook

**Advisor:** Dr. Sridhar Ramachandran

## Kaggle Competitions Investigated

### Competition 1

**Competition Name:** Passenger Screening Algorithm Challenge

**Competition URL**: [Challenge Overview Page](https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/overview)

**Description:** The Passenger Screening Algorithm Challenge, hosted by the Department of Homeland Security, aimed to enhance the accuracy of the threat recognition algorithms used in airport security screenings by the U.S. Transportation Security Administration (TSA). The dataset contains pictures from volunteers who agreed to have their images used for this competition and were collected from airport scanners. The goal is to “identify the presence of simulated threats under a variety of object types, clothing types, and body types.” I chose to investigate this competition for two reasons; the first was that it was from a credible institution, The Department of Homeland Security, and the second reason I chose this competition was because of the total award amount of 1.5 million dollars. The parameters are straightforward: predict the probability of a threat in each of the 17 body zones for every scan in the dataset. First place in this competition received a whopping 500,000 dollars!

**Learnings:** Some of the things I learned from this competition’s leaderboard came from a notebook submitted by Brian Farrar. In this notebook, I learned things like properly citing and describing your code blocks in order to allow someone who is not that familiar with coding to begin to understand your work. Another thing I saw in this notebook that encouraged me was under the “For What its Worth” section. In this section, Brian explains that he realizes that his solution is a “brute force-ish way to segment the data,” but he felt as if it was the best solution. This gave me some encouragement because I, too, tend to brute force things and then, much later down the road, realize that it could have been accomplished in a simpler and shorter way. Lastly, from this competition’s leaderboard, you understand how much detail there needs to be in your notebook when attempting to win a prize of this magnitude.

### Competition 2

**Competition Name:** Carvana Image Masking Challenge

**Competition URL**: [Competition Overview Page](https://www.kaggle.com/competitions/carvana-image-masking-challenge)

**Description:** The Carvana Image Masking Challenge, hosted by the used car seller Carvana, aims to streamline its photo studio process. For every car Carvana sells, they capture 16 images from different angles in their rotating photo studio. The issue is that with certain vehicles, the paint may be the same color or close to the color of the background; auto-editing will not decipher the difference, forcing the implementation of manual editing. The project's parameters are straightforward, "develop an algorithm that automatically removes the photo studio background" and is evaluated on the mean dice coefficient. This competition caught my eye because I recently sold a car to Carvana, which allowed me to quit my job and become a full-time student for my senior year. The dataset is full of vehicle images, each with 16 photos from different angles.

**Learnings:** This competition's leaderboard has two of the three prize winners' notebooks available, the first-place and third-place solutions. The first-place solution gives an alternate approach to the readMe or notebook description file. In this readMe, the team provides instructions on how to properly run this on your own machine, whereas in the previous notebook from the Homeland Security competition, there was a long and detailed description of the code. A crossover between the two would be great for my own project. From the third-place solution, I learned about the Torch library, PyTorch, and how it could be applied to a project like this, as I would have thought to use sklearn as the first-place submission did.

### Competition 3

**Competition Name:** Two Sigma: Using News to Predict Stock Movements

**Competition URL**: [Competition Overview Page](https://www.kaggle.com/competitions/two-sigma-financial-news)

**Description:** This competition, hosted by Two Sigma, aims to use the news cycle and news analytics to predict stock prices. The goal of this project is to return a confidence level for each stock based off the news data. Similar to how we do predictions in our school work, the higher the confidence value the higher the correlation of stock price to news data, the scale for this confidence is -1.0 to 1.0. Negative values are if you are predicting the stock to have a negative return while positive values are if you predict the stock will have a positive return. I chose this competition because I, like many others, have taken a swing at the stock prediction program. I developed mine during my time in the Data Science bootcamp at Southern Methodist University. The goal of my program was to predict the closing price of bitcoin for how ever many days in advance the user requested. But adding the news data is a very interesting idea and gives me an idea of artists streaming data and the news cycle, predicting how this could help or ruin an artists streams.

**Learnings:** Since the solutions to the award winners weren’t public, I went to the code section and sorted the notebooks by the most votes. From that, I chose DJ Sterling’s notebook to take a closer look at. A rather simple thing that I learned from this notebook is to remember to inspect the tail of the dataset. I often get in the routine of doing .head() to get a quick glance at things like formatting but never do .tail(). Incorporating this habit into my workflow improves how I approach data exploration and validation.

## Chosen Dataset

**Dataset Name:**

Dataset 1: emissions.csv

Dataset 2: vehicles.csv

**Source:**

Dataset 1: United States Energy Information Administration(EIA)

Dataset 2: United States Department of Energy

**Source Link**:

Dataset 1: [Energy Plant Data](https://www.eia.gov/electricity/data/emissions/)

Dataset 2: [Vehicle Emission Data](https://www.fueleconomy.gov/feg/ws/index.shtml#vehicle)

**Potential Insights:**

Out of high school and into my freshman year of college in 2012, I was initially a Mechanical Engineering major. At the time, electric vehicles were starting to plant themselves as a suitable option for transportation. As a racecar enthusiast, I had my own biases in why I disliked this new trend of vehicles. In my English 1301 class, I wrote a research paper on electric cars and whether they could match up to the performance and sustainability of fossil-fueled transportation. Fast forward to 2020, I enrolled in a Data Science Bootcamp and began to learn all about this thing we call Data Science. After this boot camp, I decided to return to college, changing my major to Data Science, hoping one day to use my new skillset to validate my previous assumptions in that research paper. The goal of this project is to present the actual carbon footprint of an electric vehicle by using the data from the EPA on Energy Plant CO2 emissions, data containing fuel economy information on various cars from the US Department of Energy, and creating a database that holds information from the top three electric manufacturers in order to see an electric version of the CO2 emissions since the dataset from the US Department of Energy shows a -1 value for CO2. We know this isn’t true, considering that CO2 is generated by energy plants that produce electricity for the grid to charge the car. With a bit of math, we can figure out what the CO2 per kilowatt produced is, which, in turn, we can get the CO2 in grams per mile of these electric vehicles. From this research, I hope to uncover the truth about electric vehicle emissions and see if, in fact, they have less of a carbon footprint than their competitors.

## Project Scope

**Objectives:**

The goal of this project is to determine the true carbon footprint of electric vehicles (EVs) by calculating CO2 emissions per mile, taking into account the emissions generated by energy plants producing electricity for EV charging. This addresses a real-world need to provide accurate and transparent sustainability metrics for EVs; just because there isn’t a tailpipe doesn’t mean there isn’t a footprint. The project will contribute to a better understanding of EVs for the consumer who is under the impression that their EV has less or no environmental impact when it comes to emissions. This project will be done using R and RStudio, as I don’t have a dedicated R project in my portfolio.

**Deliverables:**

* Detailed calculations of CO2 emissions per kilowatt-hour of electricity produced
* A comprehensive database consolidating data from energy plants, vehicles, and EV manufactures
* Charts and graphs comparing CO2 emissions of efficient fossil fuel cars, i.e. Toyota Corolla, versus EVs
* A report summarizing findings, methods, and insights
* An interactive dashboard that allows the user to select which EV they would like to compare to in order to aid the buying process.

**Milestones:**

Week 1: Finalize project scope, gather initial datasets, and define metrics.

Week 3: Complete the data cleaning and preprocessing.

Week 5: Perform initial analyses and calculations for CO2 emissions per kilowatt-hour.

Week 7: Develop visualizations and conduct in-depth comparisons.

Week 9: Finalize report and validate findings.

Week 10: Submit final deliverables and present results via video presentation.

**Tasks:**

1. Collect and clean data from emissions.csv and vehicles.csv.
2. Research and gather additional information on energy plant CO2 emissions and EV manufacturer data.
3. Calculate CO2 emissions per kilowatt-hour of electricity produced.
4. Calculate CO2 emissions per mile for EVs using electricity consumption data.
5. Compare EV emissions to traditional vehicle emissions using the Department of Energy fuel economy data.
6. Decide to create per state emissions or obtain a national average CO2 number.
7. Create visualizations to present findings.
8. Write a detailed report summarizing insights and conclusions.
9. Create video presentation.

**Resources:**

* R libraries: dplyr, ggplot2, tidyr, readr, and caret.
* A dedicated computer for data analysis and computation.
* Cloud resources or local storage for dataset management.
* Documentation from the US Department of Energy and EPA for additional context.

## Research Plan

**Techniques and Methods:**

* Statistical analysis to calculate average emissions per kilowatt-hour.
* Data visualization for comparing emissions across different vehicle types.
* Regression analysis to predict emissions based on vehicle specifications and energy consumption.

**Application:**

* Use statistical techniques to calculate CO2 emissions for energy production and vehicle usage.
* Visualize the comparison of EV and traditional vehicle emissions through bar charts and scatterplots.
* Perform regression to estimate the impact of specific factors (e.g., vehicle type, energy source) on overall emissions.

**Hypothesis Statement:** While EVs produce no direct CO2 emissions, their total carbon footprint—including energy plant emissions—is still significantly lower than that of fossil-fueled vehicles.

## Research Milestone Timeline

**Week 1 from submission date of this document:**

Tasks to be completed this week:

* Finalize project scope and objectives.
* Collect initial datasets (emissions.csv and vehicles.csv).
* Define key success metrics for the analysis.
* By the end of this week: A clear plan and access to relevant datasets.

**Week 2 from submission date of this document:**

Tasks to be completed this week:

* Clean and preprocess the emissions.csv and vehicles.csv datasets.
* Identify and document data gaps.
* Research additional data sources for energy plant CO2 emissions.
* By the end of this week: Datasets cleaned and data gaps identified.

**Week 3 from submission date of this document:**

Tasks to be completed this week:

* Begin exploratory data analysis (EDA) to understand dataset trends.
* Start calculating CO2 emissions per kilowatt-hour.
* Document initial observations.
* By the end of this week: Preliminary EDA and emission calculations completed.

**Week 4 from submission date of this document:**

Tasks to be completed this week:

* Refine calculations for CO2 emissions per kilowatt-hour.
* Begin calculating CO2 emissions per mile for EVs.
* Cross-check data accuracy with external sources.
* By the end of this week: CO2 emission calculations refined and validated.

**Week 5 from submission date of this document:**

Tasks to be completed this week:

* Develop initial visualizations to compare EV and traditional vehicle emissions.
* Identify trends and outliers in the data
* Prepare progress report summarizing findings so far.
* By the end of this week: Visualizations and a progress report completed.

**Week 6 from submission date of this document:**

Tasks to be completed this week:

* Conduct deeper analysis, focusing on top EV manufacturers.
* Develop a database consolidating all key metrics.
* Validate findings through peer review or feedback.
* By the end of this week: Database created and findings validated.

**Week 7 from submission date of this document:**

Tasks to be completed this week:

* Enhance visualizations for presentation.
* Finalize comparisons between EV and traditional vehicle emissions.
* Prepare draft of the final report.
* By the end of this week: Draft report and advanced visualizations completed.

**Week 8 from submission date of this document:**

Tasks to be completed this week:

* Conduct a detailed review of all findings and methods.
* Adjust calculations or visualizations based on feedback.
* Begin creating interactive dashboards (if feasible).
* By the end of this week: Finalized findings and draft interactive dashboards.

**Week 9 from submission date of this document:**

Tasks to be completed this week:

* Finalize the report and all deliverables.
* Conduct a final review of all visualizations and calculations.
* Prepare presentation materials (if required).
* By the end of this week: Deliverables finalized and ready for submission.

**Week 10 from submission date of this document:**

Tasks to be completed this week:

* Submit all deliverables.
* Present findings (if required).
* By the end of this week: Project completed and submitted.

## Potential Scope Creep

* Expanding the analysis to include additional datasets, such as international energy or vehicle data, which could delay completion, possibly focusing on Texas Energy Plants to prevent this.
* Adding features like interactive dashboards, requiring additional time and resources.
* Over-analyzing minor trends or outliers, diverting focus from core objectives.

## Ethics Statement

As a student, I pledge to sincerely hold myself to the timeline and complete the project in 10 weeks. I understand the importance of time management and commitment to my work. I will strive to uphold the highest standards of academic integrity throughout this project.