|  |
| --- |
| TECHNICAL REPORT |

|  |
| --- |
| **Distributed and Scalable Data Engineering**  **(DSCI-6007)** |

|  |  |
| --- | --- |
| SPRING 23 |  |



Contents

[Project Name 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Technical Report 3](#_Toc96341551)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

[Submitted on: 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

[Abstract 4](#_Toc96341554)

[Methodology 5](#_Toc96341555)

[Results Section 9](#_Toc96341556)

[Discussion](#_Toc96341557) 11

[Conclusion 11](#_Toc96341558)

[Contributions/References 12](#_Toc96341559)

|  |  |  |
| --- | --- | --- |
| Executive summary The electric vehicle (EV) market, driven by technology and environmental concerns, benefits from Exploratory Data Analysis (EDA) and feature engineering for informed decision-making. Key findings encompass global EV trends, regional disparities, and consumer preferences. The analysis also highlights the impact of charging infrastructure, environmental considerations, and government policies. This approach provides a holistic understanding for strategic navigation in the growing EV market, essential for industry players to stay competitive and contribute to sustainable transportation. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **Sivaji Reddy Raju - Team leader**  **Vineeth Paradesi – Data Scientist**  **Komal Tankashala – Data Engineer** | **Questions?**  Contact: vpara4@unh.newhaven.edu |  |

|  |
| --- |
| Electric Vehicle EDA: Market Insightswith Feature Engineering |

|  |  |
| --- | --- |
| **Project Title:** Electric Vehicle EDA: Market Insights with Feature Engineering <https://github.com/SivajiR45/DSCI6007_Team11/tree/main> |  |
| Highlights of Project  * The EDA process would reveal insights into the data, such as distribution, trends, outliers, and correlations among variables. This phase is essential for hypothesis formulation and deciding on appropriate modeling techniques. * The project may suggest areas for future exploration, improvements to the current models, or application of the findings to other datasets or domains.  Submitted on: 05/12/2023 |

## 

## Abstract

This project employs Exploratory Data Analysis (EDA) and feature engineering to uncover key insights in the dynamic electric vehicle (EV) market. It focuses on global EV trends, regional variations, and consumer preferences, emphasizing the impact of charging infrastructure, environmental factors, government policies, and predictive modeling. The study facilitates informed decision-making for stakeholders, contributing to a strategic understanding of market dynamics and supporting the shift towards sustainable transportation.

Data source Link

Through the following link, web scraping is being used to collect data.

<https://github.com/kvellai/Capstone_EV_Adoption_Prediction/blob/main/AfterMerge_Dataset.csv>

## 

## Methodology

We used CRISP- DM Methodology which includes the following:

Business Understanding

As the Electric vehicles population is fast growing these years, there is a need to acknowledge that and make necessary preparations for the future. This includes service stations, charging points, sale of automobile parts and customizations etc., Although the increase in population of electric vehicles is not the same in every state of the United States, we can prioritize making the above mentioned facilities in the states which have more electric vehicles accordingly.

As the Electric vehicles population is fast growing these years, there is a need to acknowledge that and make necessary preparations for the future. This includes service stations, charging points, sale of automobile parts and customizations etc., Although the increase in population of electric vehicles is not the same in every state of the United States, we can prioritize making the above mentioned facilities in the states which have more electric vehicles accordingly.

Data Understanding:

The dataset contains a variety of columns, each representing different attributes. Here's an explanation of each column based on their names and the initial data overview:

bichoice: Likely a binary choice variable, possibly indicating a specific decision or preference (e.g., choosing between two options).

range: This could refer to a range of values for a particular attribute, possibly numerical in nature. In certain contexts, it could relate to distance (e.g., the range of a vehicle).

home\_chg: Potentially indicates home charging, possibly related to electric vehicles or electronic devices.

work\_chg: Similar to 'home\_chg', this could represent charging at work or a related activity performed at the workplace.

town: This might represent a categorical or numerical variable related to towns, perhaps indicating the size of the town or a town identifier.

highway: This could indicate information related to highways, such as proximity, usage frequency, or a related numerical measure.

gender: A categorical variable indicating the gender of individuals in the dataset.

state: This is likely a categorical variable indicating the state (geographical region) associated with each data point.

Region: A broader geographical categorization than 'state', possibly indicating a region within a country or a global region.

education: Likely a categorical variable indicating the education level of individuals.

employment: This could represent the employment status or type of employment of individuals in the dataset.

hsincome: This might represent household income, potentially a numerical value indicating the income level.

hsize: Likely referring to household size, this could be a numerical variable indicating the number of individuals in a household.

housit: Potentially a categorical variable indicating housing situation or type.

residence: This might indicate the type or location of residence, possibly a categorical variable.

all\_cars: Likely a count of all cars owned or used by an individual or household.

ev\_cars: Specifically indicating the number of electric vehicles (EVs), this could be a count variable.

home\_parking: This could indicate the availability or type of home parking, possibly a numerical or categorical variable.

home\_evse: Potentially related to electric vehicle supply equipment at home, possibly indicating presence, type, or capacity.

work\_parking: Similar to 'home\_parking', but related to parking availability or type at the workplace.

work\_evse: As with 'home\_evse', but related to electric vehicle supply equipment at work.

buycar: This might indicate intentions or preferences related to purchasing a car, possibly a categorical variable.

zipcode: A numerical variable representing geographic zip codes.

dmileage: Likely indicating distance or mileage, possibly related to vehicle usage or travel habits.

long\_dist: Could represent long-distance travel frequency or preference, possibly a numerical variable.

Age\_category: Categorical variable indicating the age group or category of individuals.

RUCA: This could be an acronym or a specific categorical variable, the meaning of which depends on the context of the dataset.

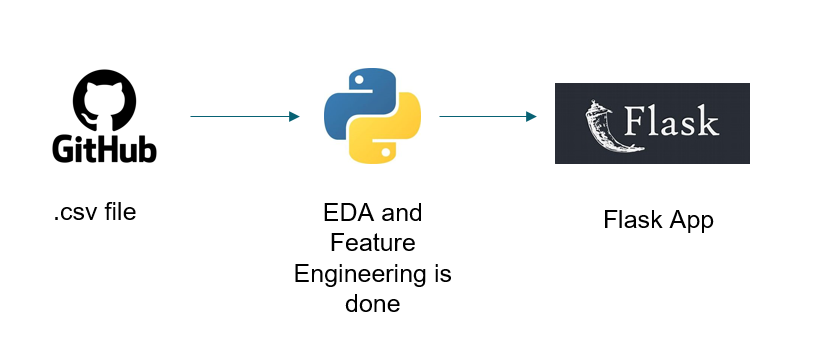
Data Preparation

"Data preprocessing" and "Understanding the shape of the dataset," suggest that the initial phase of the project involves exploring and preparing the data for analysis. This typically includes steps like handling missing values, encoding categorical variables etc.,

**Modeling and Evaluation**

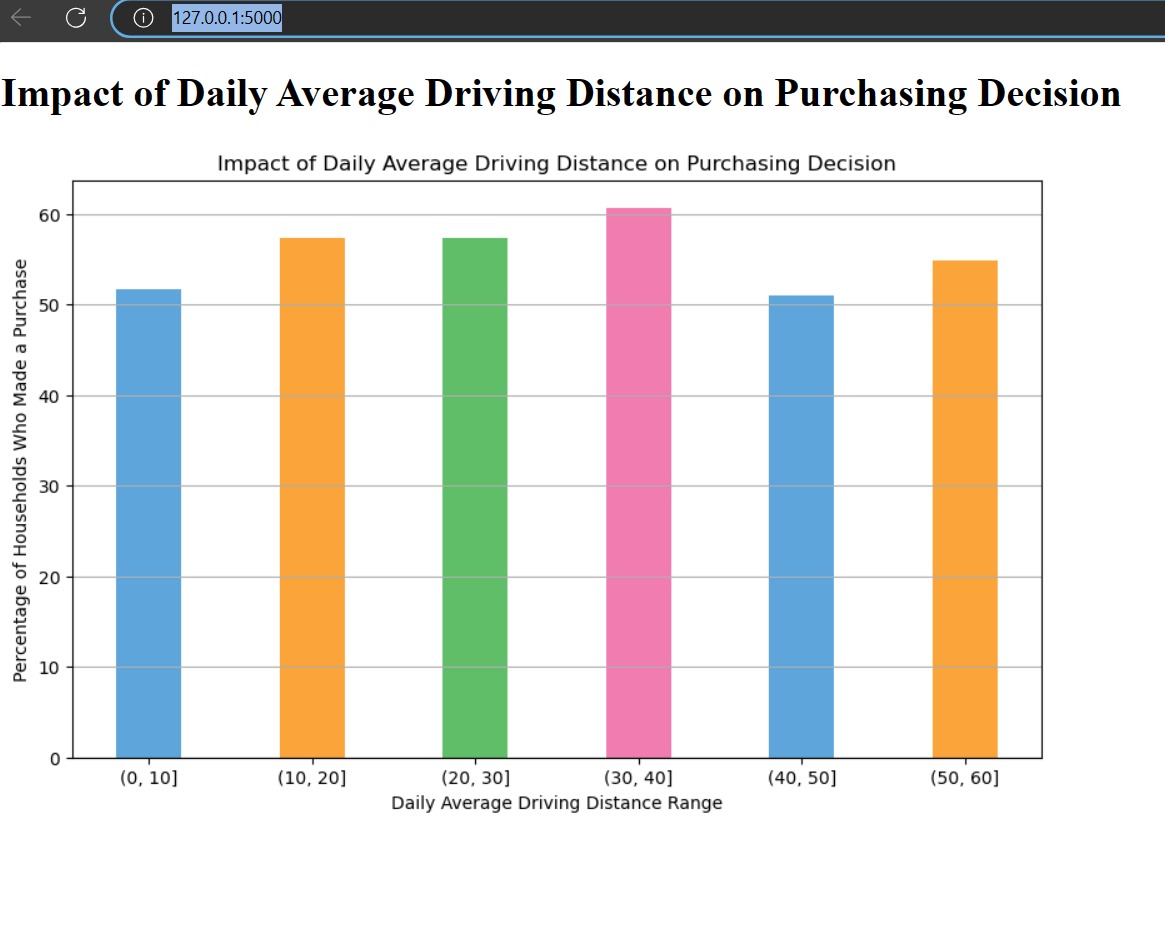
We used a Flask application to deploy the results

## Data Pipeline Architecture



## Results Section

The output is deployed using flask



## Discussion

We had an idea to get so much information from the given dataset by doing exploratory data analysis. We were able to extract 11 kinds of analysis from the dataset and we can do much more. We did feature engineering to create new columns and compared them as well to obtain insights on electric charging stations.

There is so much future scope for this project. The data is prepared and ready for preparing any type of complex models.

## Conclusion

In conclusion, by using EDA and Feature engineering, We found

1. householders’ residence area classification (city/suburb/rural) influence their electric vehicle purchase decision
2. a charging point at work option influence a household’s decision to purchase EV
3. individuals who have an electric outlet facility at their home parking space show more interest to buy an electric vehicle
4. U.S. having the highest number of households expressing interest to purchase an electric car in the next three years
5. annual income impact a household’s choice of purchasing an electric vehicle
6. students prefer to buy electric vehicles when compared to others
7. having an off-street/at home parking option influence a householder’s EV purchasing decision
8. the extent of long distance trips taken affect a householder’s decision to purchase an electric vehicle
9. household’s daily average driving distance impact their choice of buying an EV
10. household already owning an electric vehicle prefer to buy another EV
11. households currently owning a fuel vehicle show higher interest in purchasing an EV as a next car
12. We created a new Columns with the already existing columns in the data set and then plotted the comparison graph of both the newly created columns in the data set.

Git hub Link

<https://github.com/SivajiR45/DSCI6007_Team11>

## 

## Contributions/References

* <https://ieeexplore.ieee.org/abstract/document/5875917>
* <https://www.kaggle.com/code/vencerlanz09/electric-cars-eda-with-feature-engineering>