DECODING THE EV ADOPTION

Amit Singh, Anant M Nambiar, Shweta R Joshi, Shubham Yadav

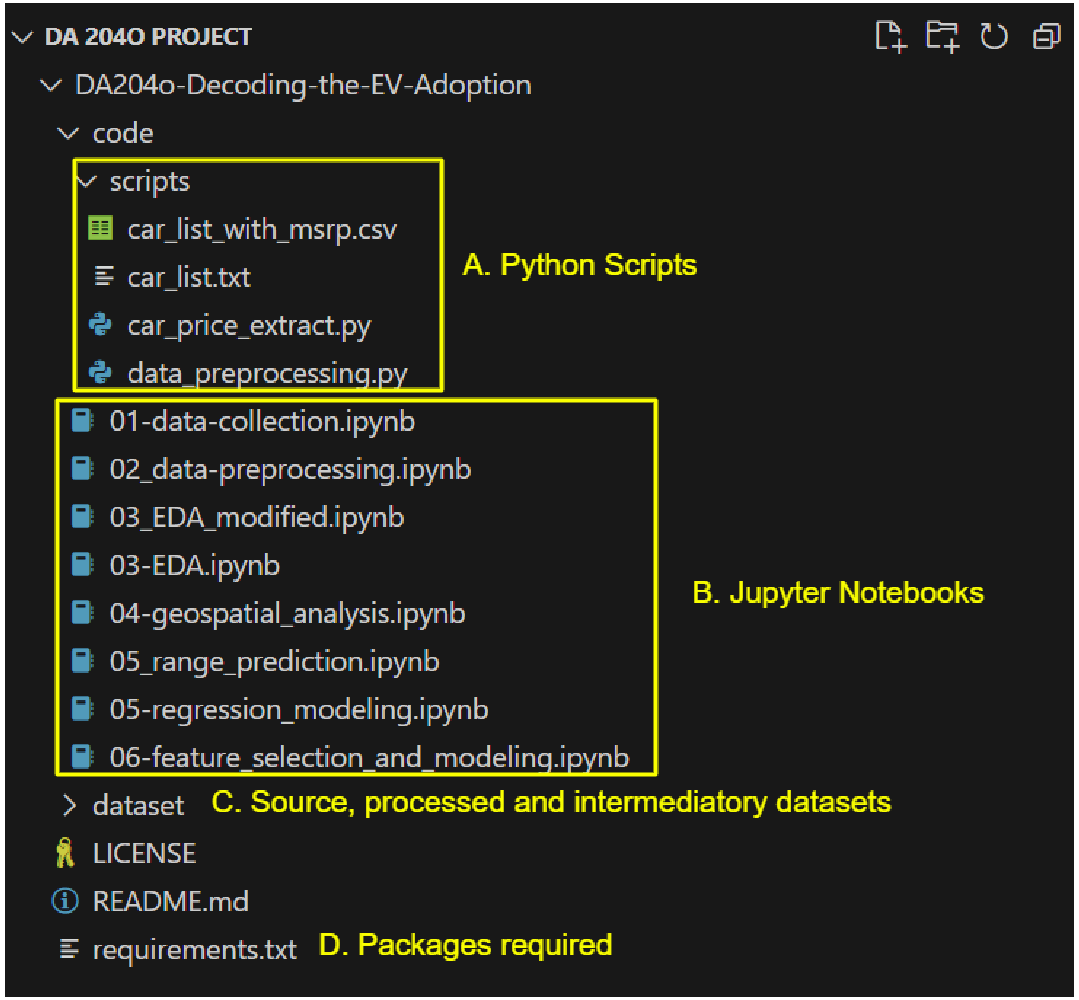
PROJECT OVERVIEW

- The project highlights the need for sustainable transportation options, with electric vehicles (EVs) emerging as a promising solution.
- The dataset is the latest and was obtained from the U.S. government website, data.gov.
- The objective of the project is to use geographic and EV-related data to predict the range and price of EVs, as well as to perform a geographical analysis of EV demand.
- The project aims to identify key factors influencing the adoption of EVs in different regions, providing insights to promote sustainable transportation.
- Machine learning algorithms are employed to uncover hidden patterns in the data, contributing to more informed decision-making.

OBJECTIVES

- Exploratory Data Analysis (EDA)
  - Analyze trends and patterns in EV adoption across different regions.
  - Examine the distribution of EV types and their electric ranges.
  - Identify correlations between features, univariate and bivariate analysis.
- Predictive Modeling
  - Develop machine learning model to predict the EV range and price which would be useful to have an idea of the demand for charging infrastructure in different regions.
- Geospatial Analysis
  - Use geographic information to analyze spatial patterns in EV adoption.
  - Identify regions with high or low EV adoption density.
- Visualization Tools
  - Use maps to show EV adoption density by county or city.

PROJECT STRUCTURE



- Different stages of the project, from data collection, preprocessing, EDA, and finally the model training, are documented as Jupyter notebooks.
- While analyzing the dataset, we observed a lot of missing values for the price. To populate those values, we have Python scripts in the scripts folder to fetch the data from Edmunds dealership.
- The source and intermediate parquet data are stored in the dataset folder.
- The packages required to run these scripts and notebooks are listed in requirements.txt.

DATA OVERVIEW

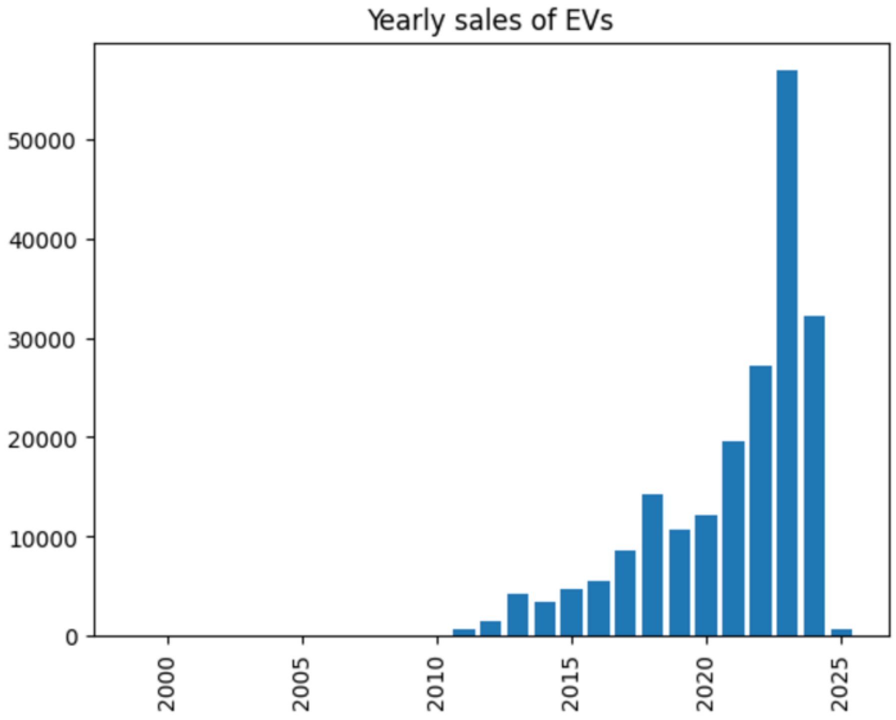
#	Column	Non-Null Count	Dtype
0	VIN (1-10)	210165 non-null	object
1	County	210161 non-null	object
2	City	210161 non-null	object
3	State	210165 non-null	object
4	Postal Code	210161 non-null	float64
5	Model Year	210165 non-null	int64
6	Make	210165 non-null	object
7	Model	210165 non-null	object
8	Electric Vehicle Type	210165 non-null	object
9	Clean Alternative Fuel Vehicle (CAFV) Eligibility	210165 non-null	object
10	Full Vehicle Name	210165 non-null	object
11	Fetched Range	210165 non-null	int64
12	Electric Range	210160 non-null	float64
13	Max Range	210165 non-null	float64
14	Fetched Price	210165 non-null	int64
15	Base MSRP	210160 non-null	float64
16	Legislative District	209720 non-null	float64
17	DOL Vehicle ID	210165 non-null	int64
18	Vehicle Location	210155 non-null	object
19	Electric Utility	210161 non-null	object
20	2020 Census Tract	210161 non-null	float64

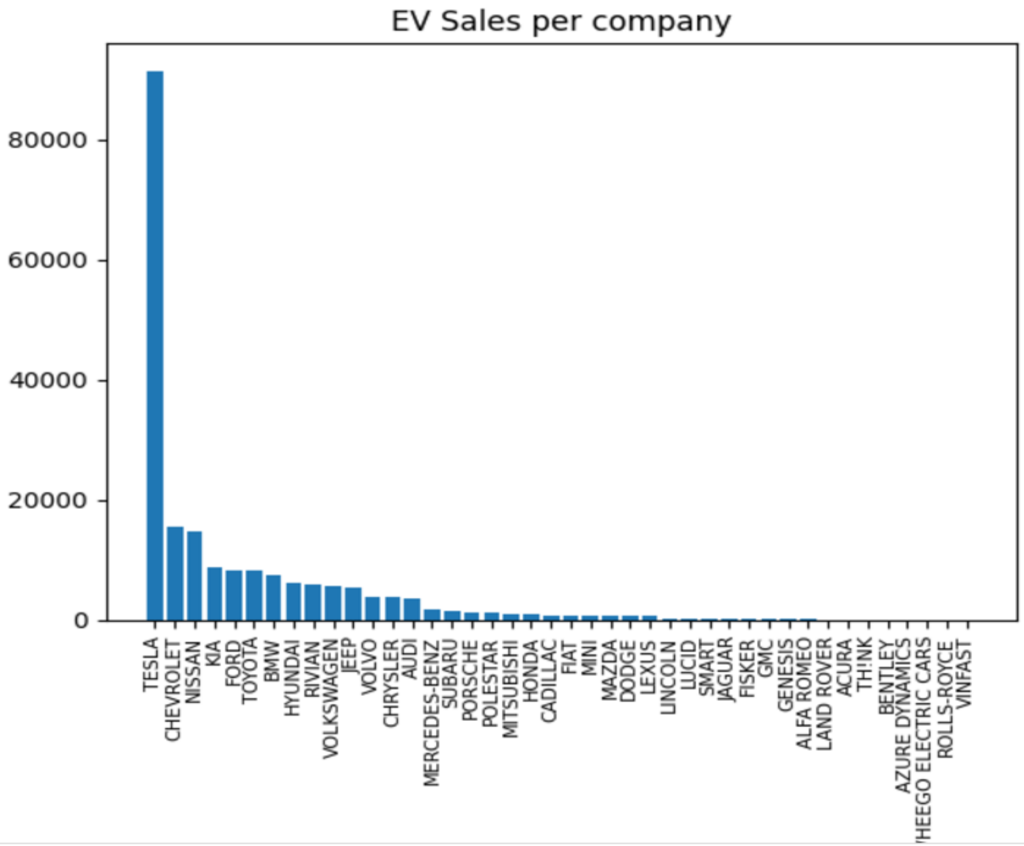
- VIN (1-10): Vehicle Identification Number (categorical)
- County: Registration county (categorical)
- City: Registration city (categorical)
- State: Registration state (categorical)
- Postal Code: Registration postal code (numerical - float)
- Model Year: Vehicle manufacture year (numerical - int)
- Make: Vehicle manufacturer/brand (categorical)
- Model: Specific vehicle model (categorical)
- EVType: Electric vehicle classification (categorical)
- CAFV: Clean Alternative Fuel Vehicle eligibility (categorical)
- Electric Range: Maximum electric-only travel distance in miles (numerical - float)
- Base MSRP: Manufacturer's Suggested Retail Price (numerical - float)
- Legislative District: Registration legislative district (numerical - float)
- DOL Vehicle ID: Department of Licensing identifier (numerical - int)
- Vehicle Location: Specific registration location details (categorical)
- Electric Utility: Power provider for the vehicle's location (categorical)
- 2020 Census Tract: Census geographic area identifier (numerical - float)

DATA PREPROCESSING

- Initial data had very high number of null values present, especially in the Range and Price fields (~48k out of 2.1L)
- Multiple extra columns which were not relevant to our analysis were removed from the dataset - “DOL Vehicle ID”, “2020 Census Tract”
- Range and Price details were also inconsistent across rows for same models
- Data pulled from Edmunds Dealership Data (API endpoints to fetch details by model and year)
- "Clean Alternative Fuel Vehicle (CAFV) Eligibility" column simplified to "True"/"False" values
- Post EDA, outliers (price > $200,000) were removed to better represent the data

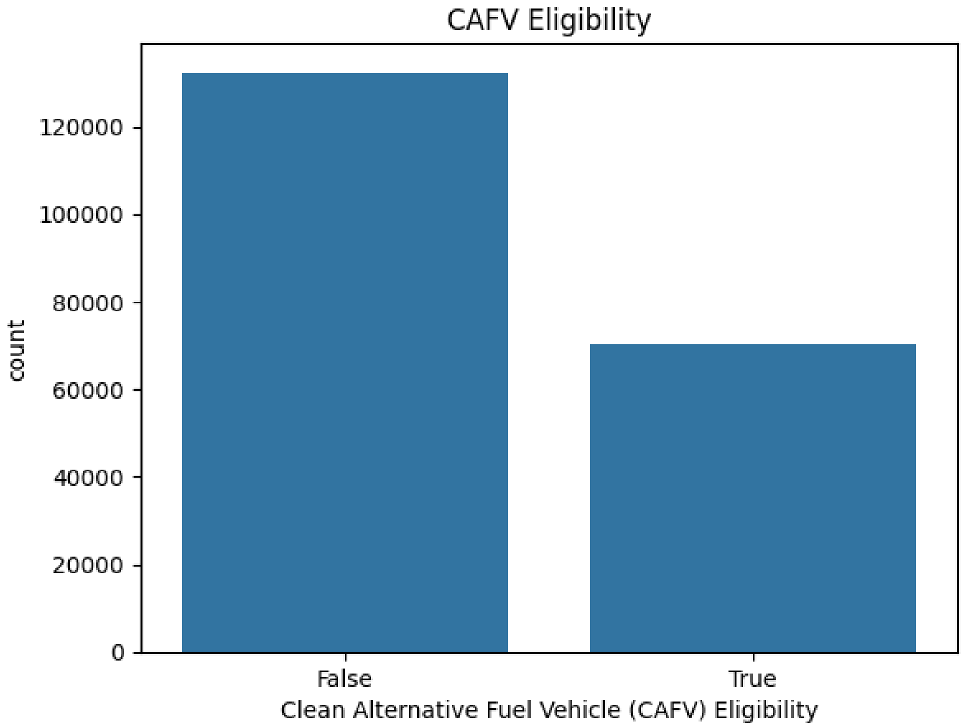
EDA : SALES INSIGHTS

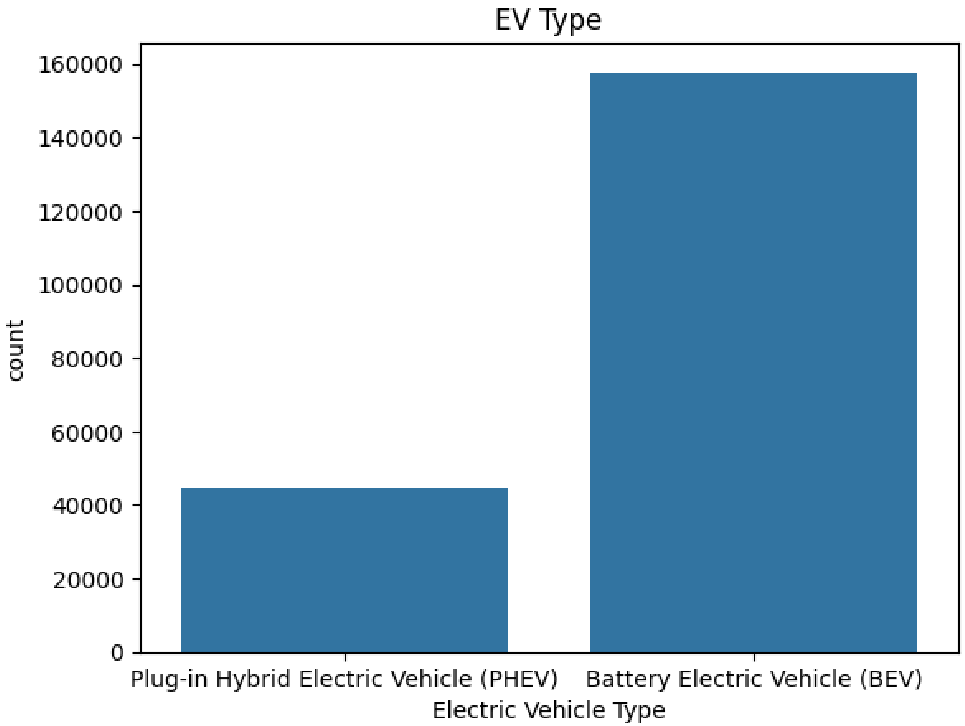




Yearly sales: increase in EV adoption
EV sales per company: Tesla dominates the sales

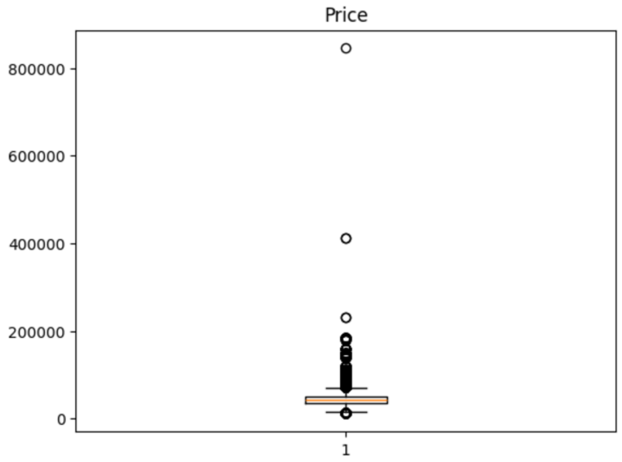
EDA : UNIVARIATE ANALYSIS

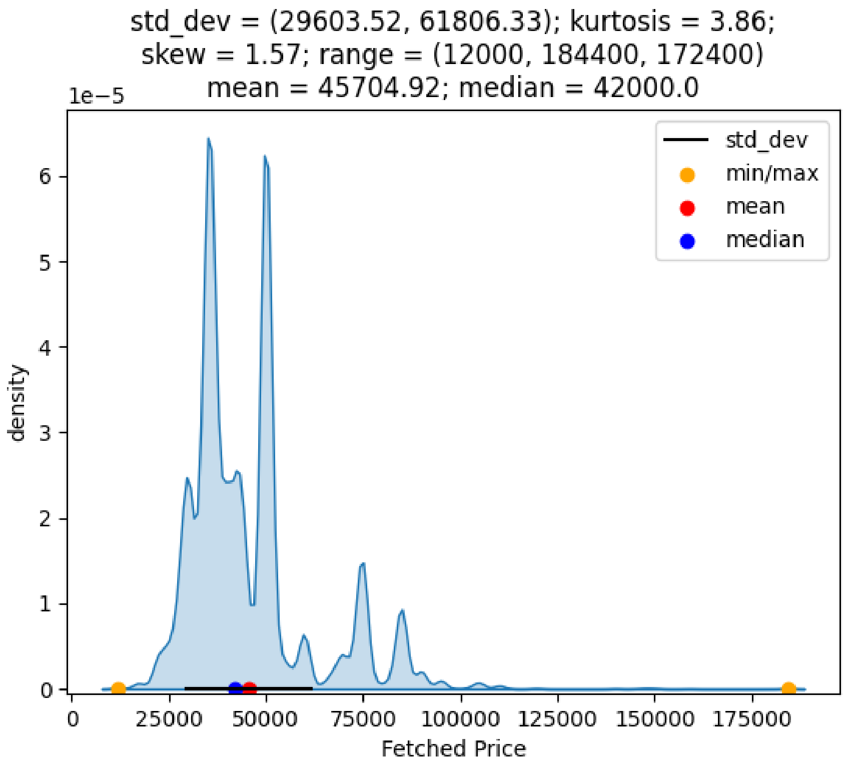


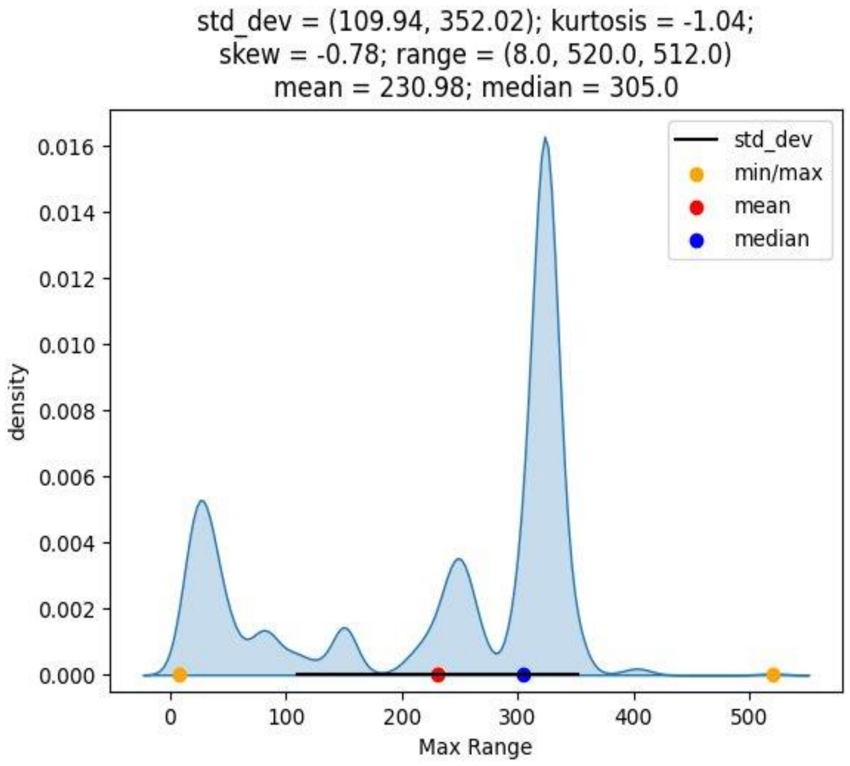


Around 2/3 of the EV models do not have CAFV eligibility.
Purely electric vehicles are more common than hybrid ones.

EDA : UNIVARIATE ANALYSIS

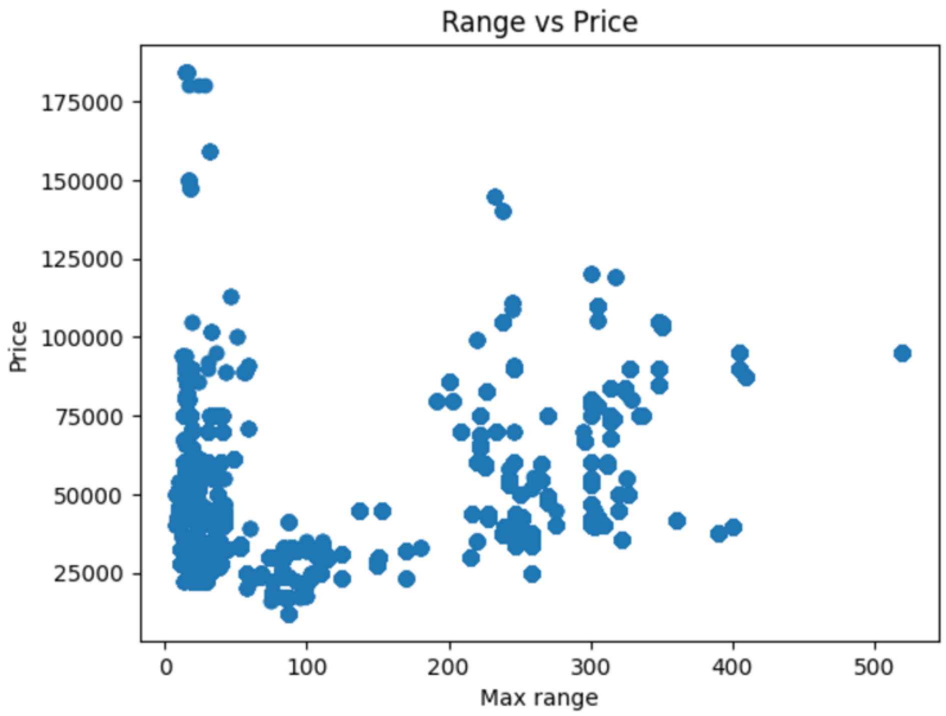


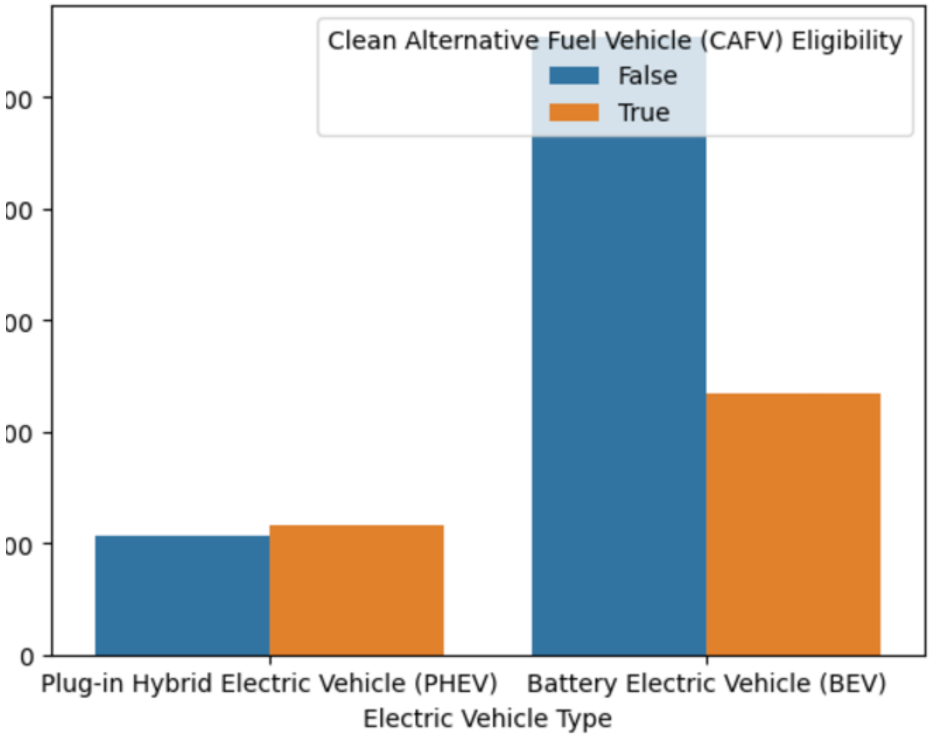




Price feature has outliers. Remove price>$200000

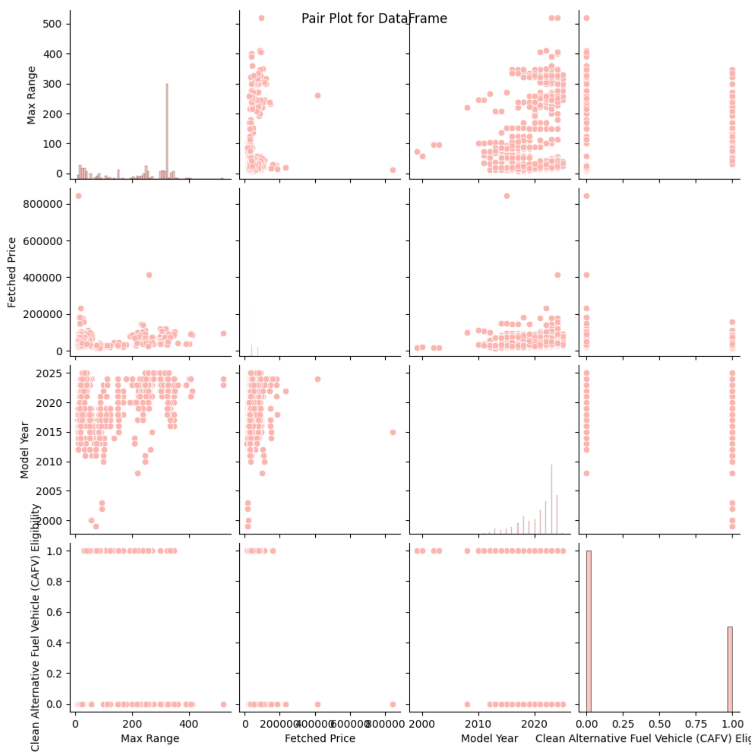
EDA : BIVARIATE ANALYSIS

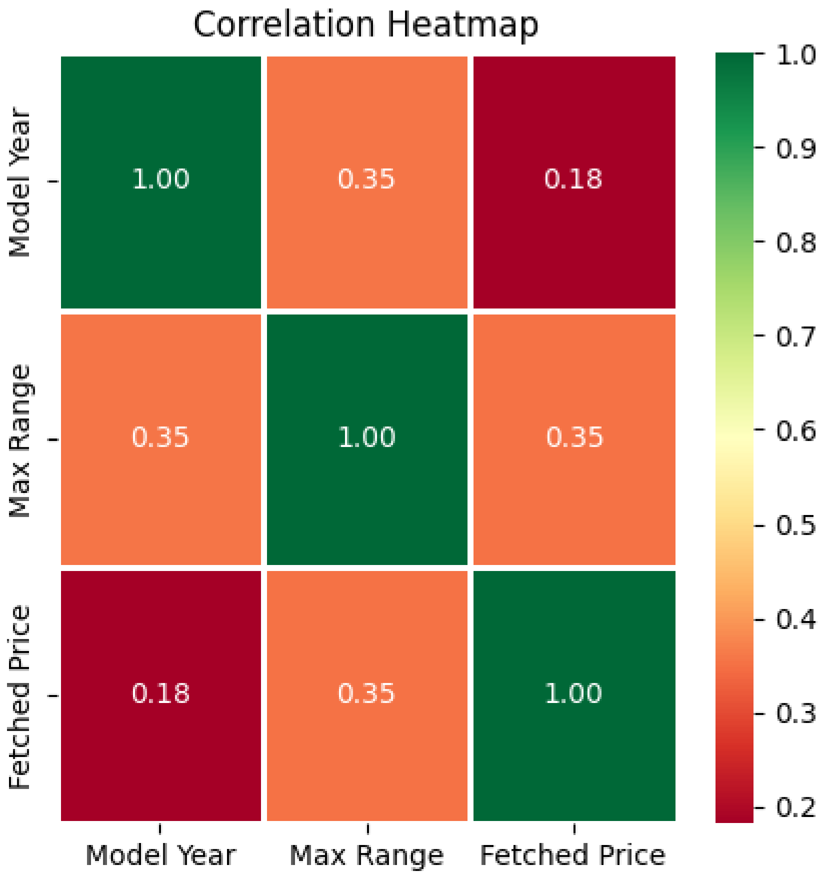




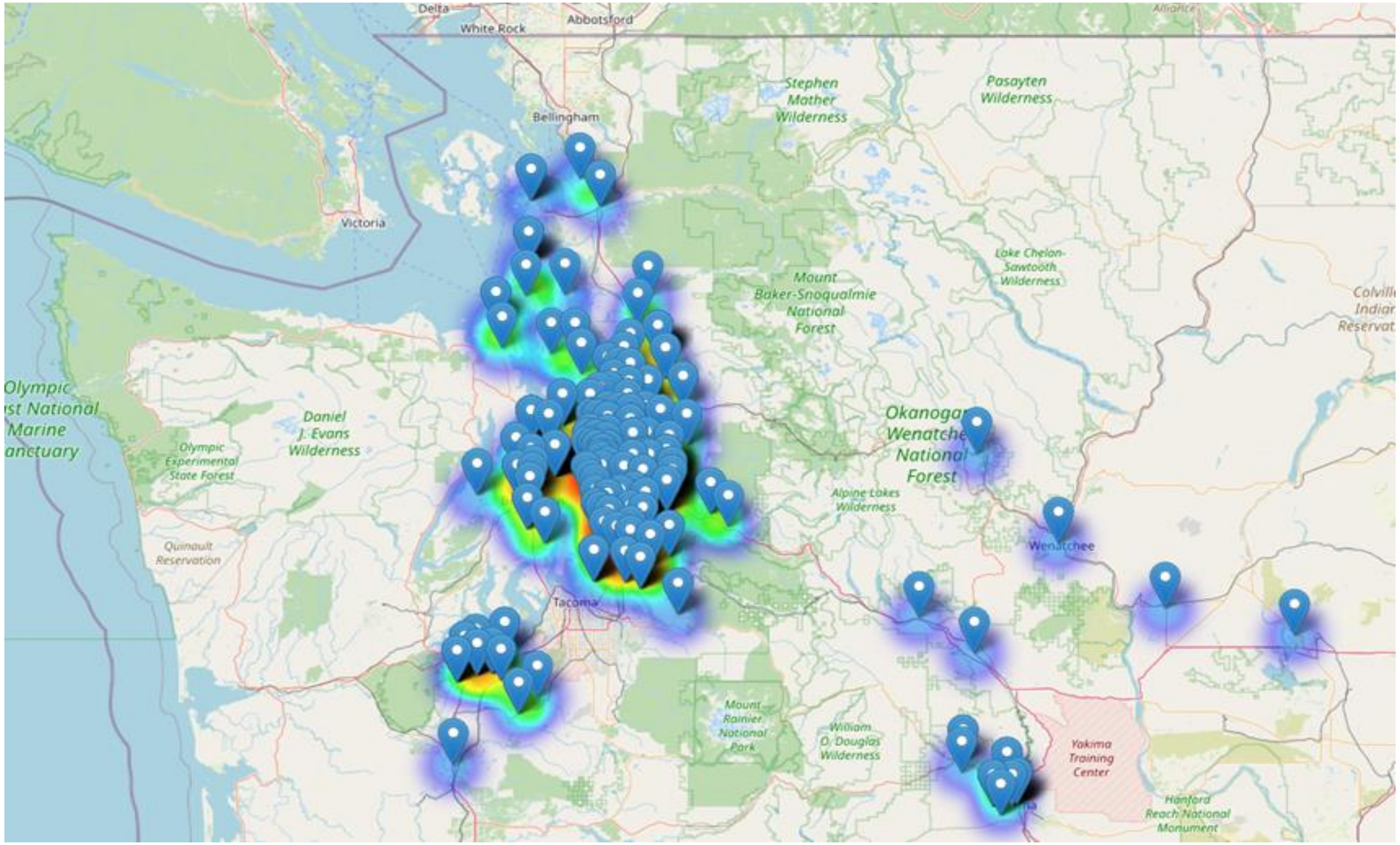
Range vs Price: no strong correlation

EDA : MULTIVARIATE ANALYSIS

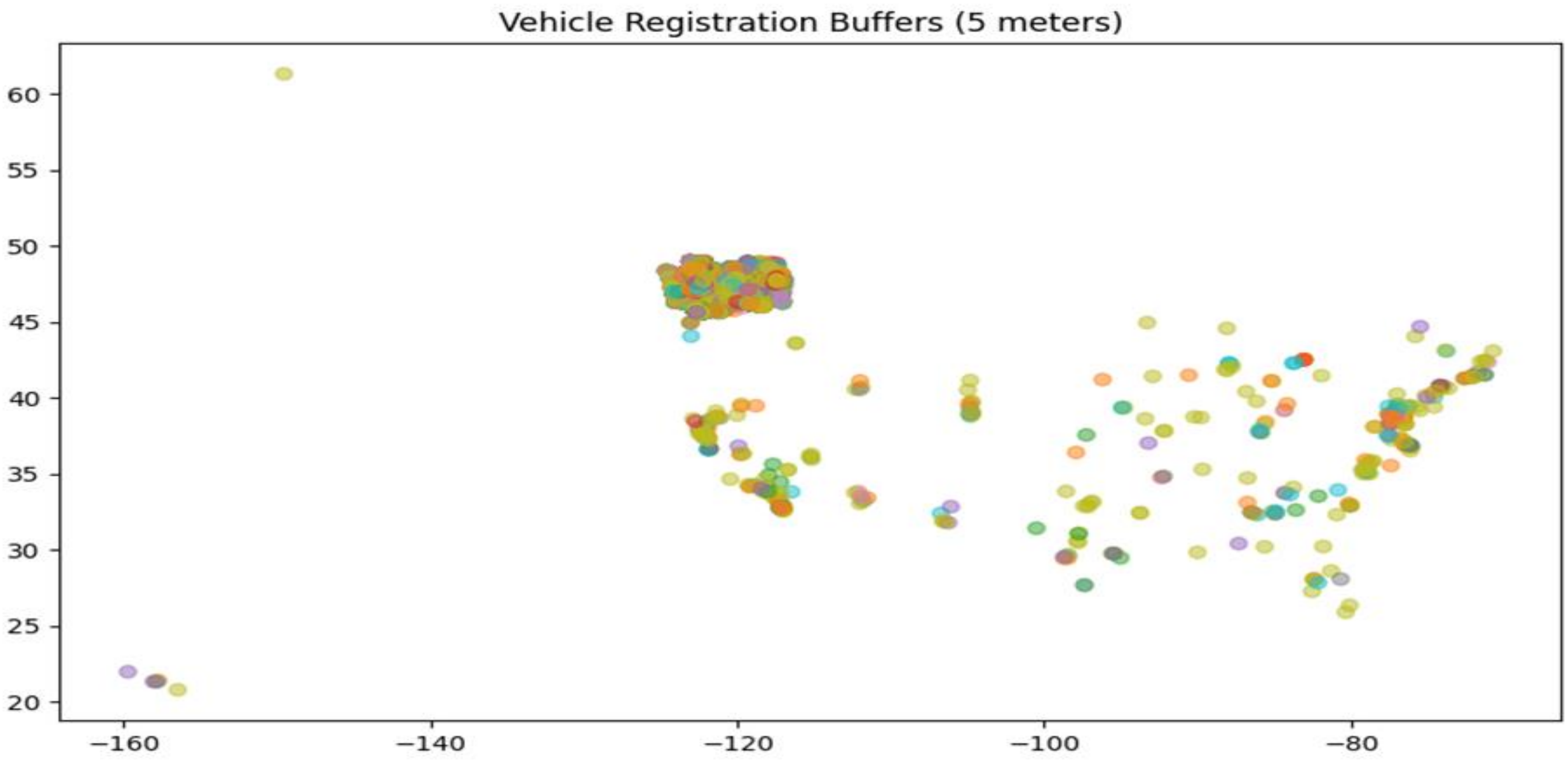




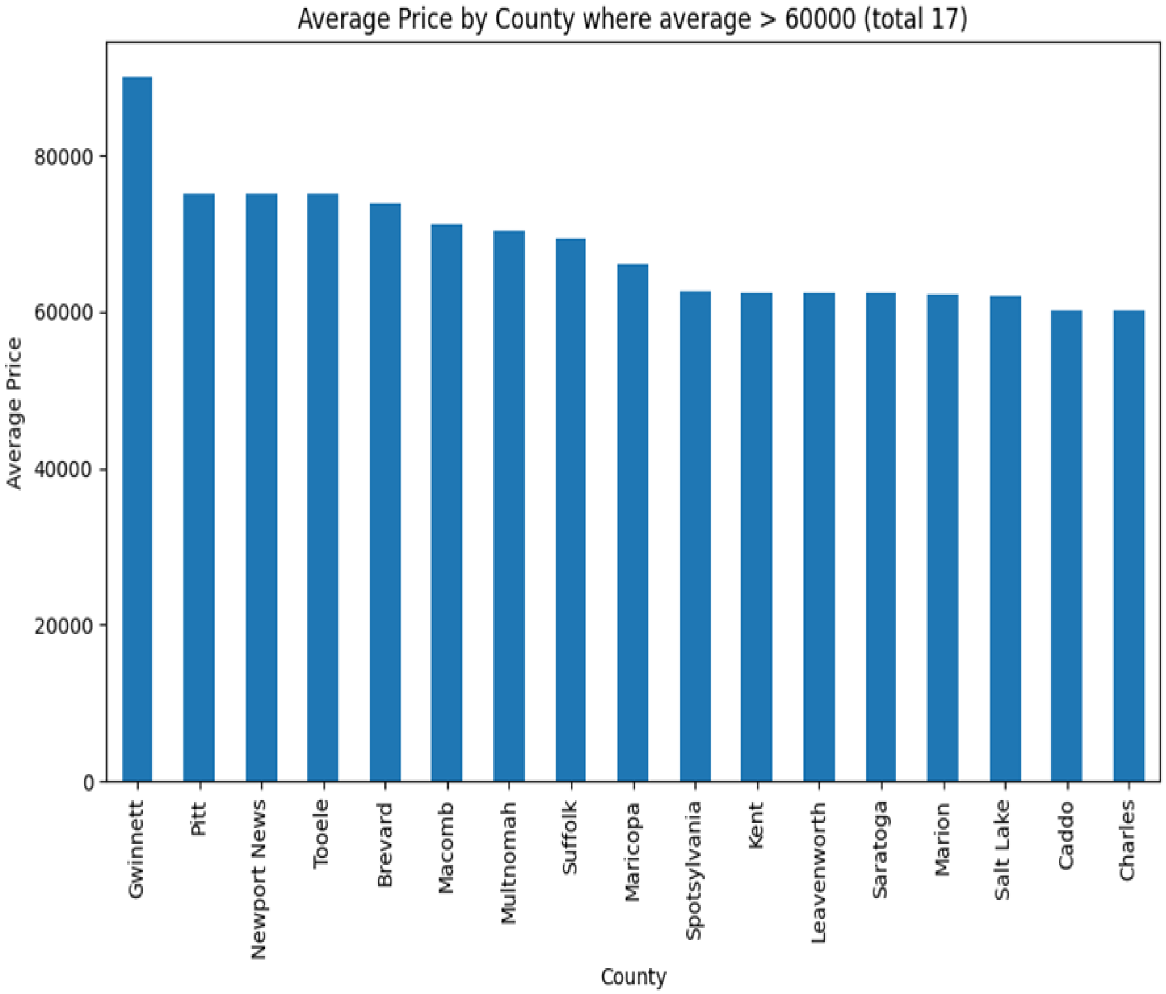
GEOSPATIAL ANALYSIS: HEATMAP(DENSITY)

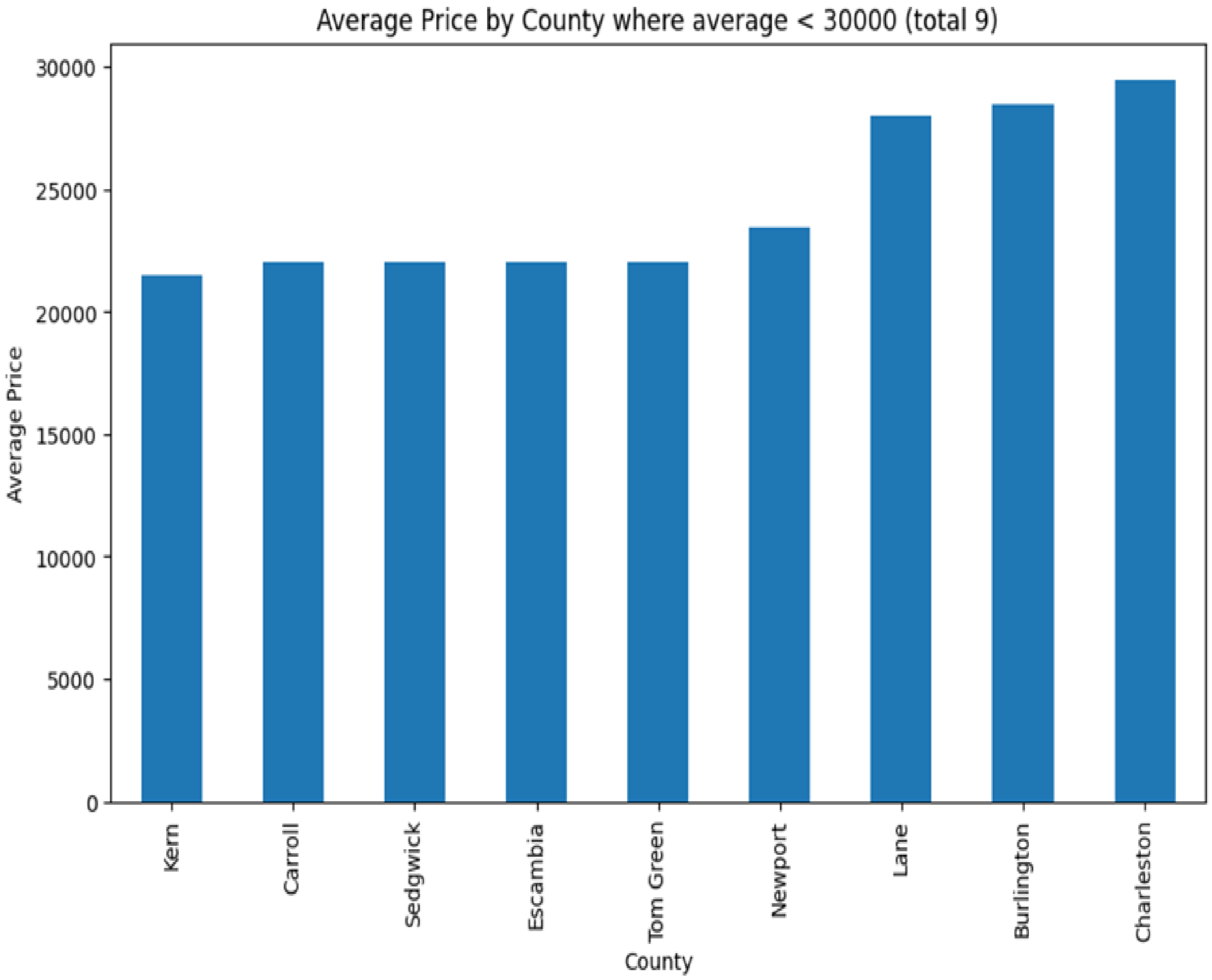


GEOSPATIAL ANALYSIS: BUFFER ANALYSIS



GEOSPATIAL ANALYSIS





PREDICTIVE MODELLING

Random Forest Regression

(Range Prediction)

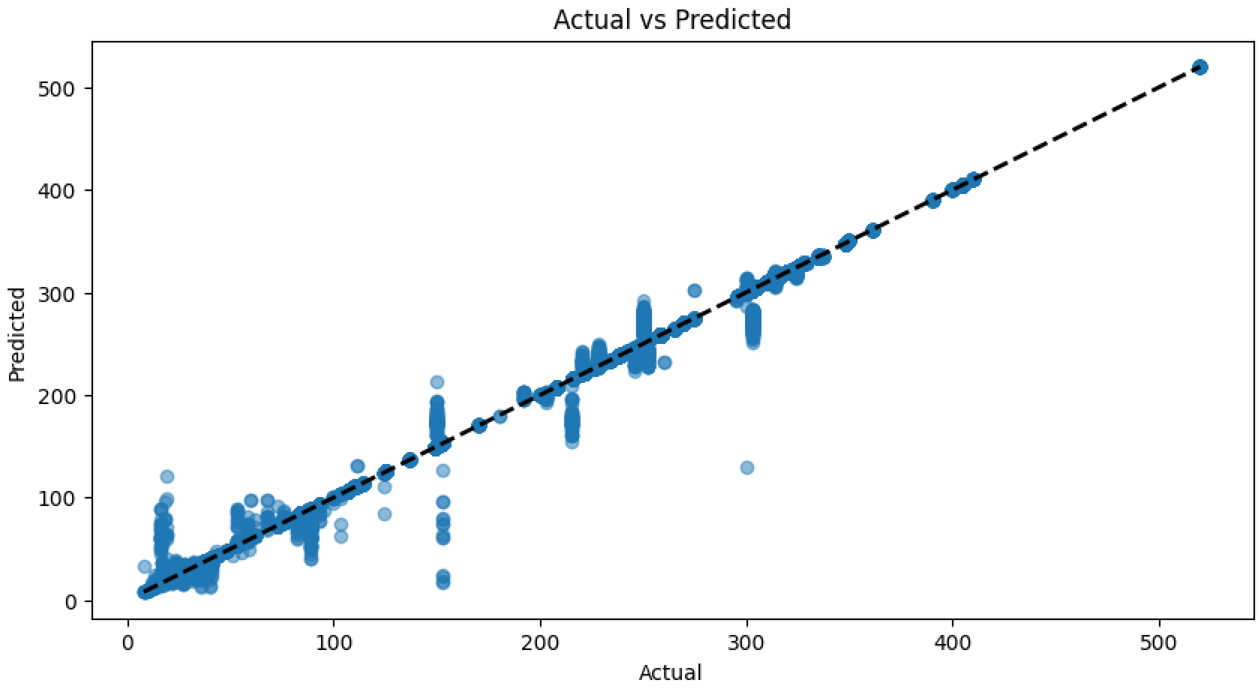
The Random Forest Regressor is an ensemble machine learning algorithm primarily used for regression tasks. It is based on the Random Forest method, which builds multiple decision trees and aggregates their results for prediction.

Gradient Boosting Regression

(Price Prediction)

Gradient Boosting Regression is a powerful ensemble learning technique used for regression tasks. It builds a predictive model by combining the strengths of multiple "weak learners," typically decision trees, to improve accuracy and robustness.

RESULTS : RANGE PREDICTION



Evaluation Matrix

Mean Absolute Error: 1.187043

Mean Squared Error: 30.738393

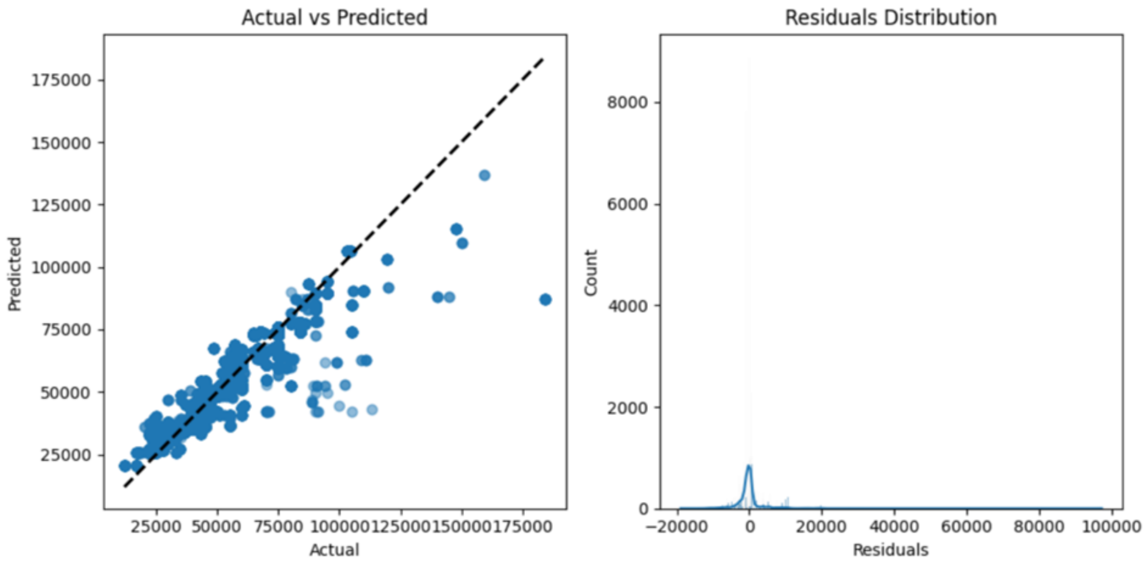
Root Mean Squared Error: 5.544222

R-squared Error: 0.947912

MAPE: 6.626591%

RESULTS : PRICE PREDICTION

```
features = [
    "County",
    "City",
    "State",
    "Model Year",
    "Make",
    "Model",
    "Electric Vehicle Type",
    "Clean Alternative Fuel Vehicle (CAFV) Eligibility",
    "Max Range",
]
target = "Fetched Price"
```



Mean Squared Error: 17370466.769106243
R^2 Score: 0.9331644118865465

THANK YOU

Data Science Canvas

Project: Decoding EV Adoption

Team: Amit Singh, Anant M Nambiar, Shweta R Joshi, Shubham Yadav

Problem Statement

Business Case & Value Added
Which business case should be analyzed and what added value does it generate?

Based on the current trends of sales, companies can use this type of data to tune their pricing models so that the vehicles are priced aggressively as compared to their competitors.

Data Landscape
Which data is required for this and which is already available? Which additional data has to be collected?

'Electric Vehicle Population Data' dataset from data.gov. Additional data for Price of vehicles had to be collected separately from Edmunds API

Model Selection
Which analysis methods can be considered on the basis of the specific data landscape and the business case?

EDA to describe the data and visualise it

Supervised regression models to predict prices and range

Model Requirements
Which model requirements must be complied with in order to obtain a valid model?

Data quality – proper data without missing or inconsistent values, feature relevance, etc

Software & Libraries
Which software should be used? Is there already a standard solution? Which libraries are used?

Pandas
Scikit-learn
Numpy
Matplotlib
Seaborn
Geopandas
Folium

Skills
What skills are needed to provide the data and model development?

Data Collection and Integration

Data Cleaning and Preprocessing

Experience with machine learning libraries

Coding abilities

Ability to translate data insights into actionable business recommendations for stakeholders.

Execution & Evaluation

Model Evaluation
Which indicators require quality control and validation and how should they be interpreted? Is real-time monitoring necessary?

R2, MAPE to calculate accuracies and errors in the test dataset of the model.

Real time monitoring is not required for this

Data Storytelling
What requirements does the target group have for the presentation of the results and how do I effectively communicate this data?

For the general audience, we have visualizations of trends and insights from EDA that would help them make a decision.

For the decision makers, the geospatial analysis gives clarity on which geospatial parts show an increase in EV adoption.
Our pricing prediction model would help them with setting the pricing model for their product depending on its specs

Data Collection & Preparation

Data Selection & Cleansing
Which of the available data is relevant? Do the data have to be cleaned up?

Model Year,Make,Model,EVType,CAFV,Electric Range,Base MSRP are the relevant columns. There is a significant amount of null values in these columns which require to be cleaned up

Data Integration
In which system should the data from different sources be migrated?

Data on range and MSRP is pulled from the Edmunds API and the cleaned dataset is saved and stored for future operations

Data Collection
How and with which methods should additionally required data be collected? What properties has this data to fulfil?

Data regarding range and MSRP can be fetched externally from other sources and imputed into this dataset from the year,make and model of the cars

Explorative Data Analysis
Are there outliers or structures to be considered? Creation of descriptive key figures for the first assessment of the data.

Any outliers in 'price' feature needs to be handled. Price, range, CAFV, model, make, EV Type and geospatial features are the key figures for the assessment of the data