ANALYZING THE RELATIONSHIP BETWEEN ELECTRIC VEHICLE SPECIFICATIONS AND DRIVING RANGE

DSA 210 - Introduction to Data Science

Ahmet Durmaz

32569

FALL2025

1. Introduction and Motivation

Electric vehicles have become an increasingly important part of modern transportation systems. As their adoption grows, understanding the factors that influence real-world driving range has become a key concern for both users and manufacturers. Driving range directly affects user experience, charging behavior, and confidence in electric mobility.

Although battery capacity is commonly considered the most important factor determining an electric vehicle's range, it is not the only one. Characteristics such as vehicle weight, motor power, energy efficiency, acceleration performance, and vehicle category can also influence how far an electric vehicle can travel on a single charge. Focusing on only one parameter may therefore lead to an incomplete understanding of range performance.

The motivation of this project is to examine how a combination of technical specifications and manufacturing-related attributes jointly affect the real-world driving range of electric vehicles. By applying data analysis and machine learning techniques learned during the course, this study aims to provide a clearer, data-driven perspective on the main factors shaping electric vehicle range.

2. Data Sources and Data Collection

This project is based on two publicly available datasets that were combined to enrich the analysis and provide a broader perspective.

The first dataset was obtained from the EV Database and contains detailed technical specifications of electric vehicles. This dataset includes information such as usable battery capacity, vehicle weight, drivetrain power, real-world energy efficiency, acceleration performance, vehicle segment, and real-world driving range. These variables form the technical foundation of the analysis.

The second dataset was collected from Kaggle and provides market and manufacturing-related information. It includes details such as manufacturer, model year, country of manufacture, pricing information, safety ratings, and sales figures. While this dataset does not focus on technical performance, it adds important contextual information related to production and market characteristics.

The two datasets were merged using a common identifier, Vehicle_ID. This integration step allowed technical features to be analyzed together with manufacturing and market attributes, resulting in a more comprehensive dataset than would be possible using a single data source.

3. Methodology

The analysis followed a structured data science workflow similar to the one taught in the course, moving from data preparation to exploratory analysis, hypothesis testing, and finally machine learning.

The first step involved data preprocessing. Relevant numerical and categorical variables were selected based on their relevance to driving range. Rows with missing values in critical variables, such as range and key numerical predictors, were removed to ensure the reliability of the analysis. After cleaning, the dataset was suitable for both statistical analysis and modeling.

Exploratory Data Analysis (EDA) was then conducted to better understand the structure and distribution of the data. This stage included examining descriptive statistics, visualizing the distribution of real-world driving range, and exploring relationships between range and other variables through scatter plots, correlation matrices, and boxplots. These visualizations provided initial insights and guided the formulation of statistical hypotheses.

To formally test these relationships, hypothesis testing methods were applied. Pearson correlation tests were used to evaluate linear relationships between driving range and numerical variables, while one-way ANOVA tests were conducted to assess whether mean driving range differed significantly across vehicle segments and countries of manufacture.

In the final stage, supervised machine learning techniques were applied to predict real-world driving range. The problem was formulated as a regression task, with Range_Real as the target variable. Linear Regression was used as a baseline model, while Random Forest Regressor was employed to capture potential non-linear relationships. Categorical variables were encoded using one-hot encoding, and the dataset was split into training and test sets to evaluate model performance.

4. Results and Findings

The analysis revealed substantial variation in real-world driving range across electric vehicles. Correlation analysis showed a strong and statistically significant positive relationship between battery capacity and driving range, confirming the expectation that larger batteries enable longer travel distances.

Acceleration performance exhibited a strong negative relationship with range, indicating a trade-off between performance and efficiency. Vehicles designed for faster acceleration generally tended to have shorter driving ranges. Vehicle weight and drivetrain power showed moderate relationships with range, suggesting that larger and more powerful vehicles often compensate for increased energy consumption by incorporating larger batteries.

Segment-based analysis demonstrated that vehicle class plays a significant role in determining driving range. Luxury and executive vehicle segments tended to achieve higher median ranges, while smaller vehicle categories showed lower values. Country-based analysis also revealed statistically significant differences, suggesting that manufacturing location may indirectly influence range through technological standards, regulations, and market preferences.

Machine learning results supported and extended these findings. Linear Regression provided a solid baseline and confirmed many of the linear relationships observed during EDA. However, the Random Forest model achieved better predictive performance, indicating that non-linear interactions between vehicle features are important. Feature importance analysis further highlighted battery capacity as the most influential predictor, followed by efficiency and drivetrain power.

5. Limitations and Future Work

Despite the strong results, this study has several limitations. The analysis relies on aggregated specification data rather than real-world driving logs, which may not fully capture everyday driving conditions. External factors such as weather, road conditions, and driving behavior were not included. Additionally, the number of vehicles varies across countries, which may introduce sample imbalance.

Future work could address these limitations by incorporating environmental data, performing cross-validation and hyperparameter tuning, or applying more advanced machine learning techniques. Expanding the dataset with real-world usage data could also improve the robustness of the results.

6. Conclusion

This project demonstrates that real-world driving range is influenced by a combination of technical specifications, performance characteristics, vehicle class, and manufacturing-related factors. Battery capacity remains the strongest determinant of range, but other features such as efficiency, acceleration, and vehicle segment also play meaningful roles. The application of machine learning methods further shows that non-linear models can effectively capture these complex relationships. Overall, the project successfully applies data science techniques learned in the course to a real-world problem.

8. References

EV Database. https://ev-database.org

Kaggle. EV Electrical Vehicles Dataset (3K Records 2025)
https://www.kaggle.com/datasets/pratyushpuri/ev-electrical-vehicles-dataset-3k-records-2025