**Data Description and Sourcing**

Han Li, Kun Xu

Pro. Do

MA541 Course Project

Winter - 2024

Table of Contents

[1. Introduction 4](#_Toc184406124)

[2. Data Description and Sourcing 4](#_Toc184406125)

[3. Data Cleaning 4](#_Toc184406126)

[4. Exploratory Data Analysis 4](#_Toc184406127)

[5. Hypothesis Testing 7](#_Toc184406128)

[Objective 7](#_Toc184406129)

[Method 7](#_Toc184406130)

[Results 8](#_Toc184406131)

[Conclusion 8](#_Toc184406132)

[Interpretation 8](#_Toc184406133)

[6. Regression Analysis 8](#_Toc184406134)

[Feature Selection 10](#_Toc184406135)

[Regression Equation 10](#_Toc184406136)

[Results 10](#_Toc184406137)

[Interpretation 10](#_Toc184406138)

[7. Conclusions and Recommendations 11](#_Toc184406139)

[8. Code and Appendix 11](#_Toc184406140)

[Data cleaning and exploratory analysis 11](#_Toc184406141)

[Data cleaning and exploratory analysis 12](#_Toc184406142)

[Hypothesis Testing 12](#_Toc184406143)

[Regression 13](#_Toc184406144)

[References 17](#_Toc184406145)

[Footnotes 18](#_Toc184406146)

[Tables 19](#_Toc184406147)

[Figures 20](#_Toc184406148)

# 1. Introduction

Electric vehicles (EVs) are increasingly prevalent, emphasizing the importance of understanding charging behaviors and patterns. This report analyzes the Electric Vehicle Charging Patterns dataset to explore correlations, trends, and significant factors influencing EV charging behaviors. The analysis aims to provide actionable insights into optimal charging practices and the effects of external factors like temperature.

# 2. Data Description and Sourcing

Data Source:

<https://www.kaggle.com/datasets/valakhorasani/electric-vehicle-charging-patterns>

The dataset was sourced from Kaggle and contains 1,320 entries with 20 variables, including numeric fields (e.g., Battery Capacity, Energy Consumed, Charging Cost) and categorical fields (e.g., User Type, Charger Type). The data required some cleaning to ensure reliability.

* **Numeric:** 10 variables (Battery Capacity, Energy Consumed, Charging Duration, Charging Rate, Charging Cost, State of Charge, State of Charge, Distance Driven, Temperature, Vehicle Age).
* **Battery Capacity (kWh):** Total battery capacity of the vehicle in kilowatt-hours.
* **Energy Consumed (kWh):** Total energy consumed during the charging session, measured in kilowatt-hours.
* **Charging Duration (hours):** Total time taken to charge the vehicle, measured in hours.
* **Charging Rate (kW):** Average power delivery rate during the charging session, measured in kilowatts.
* **Charging Cost (USD):** Total cost incurred for the charging session, measured in US dollars.
* **State of Charge (Start %):** Battery charge percentage at the start of the charging session.
* **State of Charge (End %):** Battery charge percentage at the end of the charging session.
* **Distance Driven** (since last charge) (km): Distance traveled since the last charging session, measured in kilometers.
* **Temperature (°C):** Ambient temperature during the charging session, measured in degrees Celsius.
* **Vehicle Age** (years): Age of the electric vehicle, measured in years.
* **Categorical/Text:** 10 variables (User ID, Vehicle Model, Charging Station ID, Charging Station Location, Charging Start Time, Charging End Time, Time of Day, Day of Week, Charger Type and User Type).
* **User ID**: Unique identifier for each user.
* **Vehicle Model:** Model of the electric vehicle being charged (e.g., Tesla Model 3, Nissan Leaf).
* **Charging Station ID:** Unique identifier for the charging station used.
* **Charging Station Location:** Geographic location of the charging station (e.g., New York, Los Angeles).
* **Charging Start Time:** Timestamp indicating when the charging session began.
* **Charging End Time:** Timestamp indicating when the charging session ended.
* **Time of Day:** Time segment when the charging occurred (e.g., Morning, Afternoon).
* **Day of Week:** Day of the week when the charging occurred (e.g., Monday, Tuesday).
* **Charger Type:** Type of charger used (e.g., Level 1, Level 2, DC Fast Charger).
* **User Type:** Classification of user based on driving habits (e.g., Commuter, Long-Distance Traveler).

# 3. Data Cleaning

Invalid entries were identified where the State of Charge at the start was greater than at the end. These were removed. Columns with missing values, like Energy Consumed, were addressed by filtering out incomplete records. No duplicate values were found.

# 4. Exploratory Data Analysis

Exploratory data analysis involved descriptive statistics and visualization:

1. Correlation Heatmap: Visualized relationships between numeric variables, revealing strong correlations between variables like Charging Rate and Energy Consumed.

A diagram of heatmap

Description automatically generated

2. Scatter Plot: Highlighted the relationship between Temperature and Energy Consumed.

A graph with orange dots

Description automatically generated

3. Boxplot: Compared Charging Costs across User Types.

A chart of a number of colored boxes

Description automatically generated with medium confidence

# 5. Hypothesis Testing

## Objective

To test whether there is a significant difference in charging costs between Commuters and Long-Distance Travelers.

## Method

An **independent two-sample t-test** was conducted:

* **Null Hypothesis (H0H\_0H0​)**: There is no significant difference in mean charging costs between Commuters and Long-Distance Travelers.
* **Alternative Hypothesis (H1H\_1H1​)**: There is a significant difference in mean charging costs between Commuters and Long-Distance Travelers.
* **Significance Level (α\alphaα)**: 0.05.

## Results

1. **Descriptive Statistics**:
   * Mean charging cost for Commuters: **21.45 USD**, standard deviation: **9.78 USD**.
   * Mean charging cost for Long-Distance Travelers: **25.33 USD**, standard deviation: **11.26 USD**.
   * Sample sizes: **546** (Commuters) and **432** (Long-Distance Travelers).
2. **T-test Output**:
   * t-value = **-4.672**.
   * p-value = **< 0.0001**.

## Conclusion

Since the p-value is less than the significance level (α=0.05\alpha = 0.05α=0.05), the null hypothesis is rejected. **There is a statistically significant difference in charging costs between Commuters and Long-Distance Travelers.**

## Interpretation

* Long-Distance Travelers have significantly higher charging costs compared to Commuters.
* Recommendation: Evaluate pricing strategies to ensure fairness and consider the behavioral differences between these user groups.

# 6. Regression Analysis

Linear regression was used to model Energy Consumed based on Temperature. The regression results are as follows:

1. Regression Equation: Energy Consumed = -0.01399 \* Temperature + 43.101

2. R² Score: -0.0023 (indicates a weak relationship between the variables).

3. Mean Squared Error (MSE): 453.84.

4. Coefficients: -0.01399.

5. Intercept: 43.101.

Below is the regression scatter plot with the fitted line:

A graph with blue and red dots

Description automatically generated

## Feature Selection

The following variables were used to predict energy consumption:

* **Temperature**
* **Charging Duration**
* **Charging Rate**
* **Vehicle Age**

## Regression Equation

Energy Consumed=−0.01935×Temperature+0.8901×Charging Duration−0.0570×Charging Rate−0.0436×Vehicle Age+43.042\text{Energy Consumed} = -0.01935 \times \text{Temperature} + 0.8901 \times \text{Charging Duration} - 0.0570 \times \text{Charging Rate} - 0.0436 \times \text{Vehicle Age} + 43.042Energy Consumed=−0.01935×Temperature+0.8901×Charging Duration−0.0570×Charging Rate−0.0436×Vehicle Age+43.042

## Results

* **R² Score**: -0.0021 (The model fails to explain variations in energy consumption.)
* **Mean Squared Error (MSE)**: 476.25
* **Regression Coefficients**:
  + **Temperature**: -0.01935
  + **Charging Duration**: 0.8901
  + **Charging Rate**: -0.0570
  + **Vehicle Age**: -0.0436
* **Intercept**: 43.042

## Interpretation

* **Charging Duration** has the most significant positive impact on energy consumption.
* **Temperature**, **Charging Rate**, and **Vehicle Age** have minimal influence on energy consumption.
* The overall model performance is weak, suggesting the need for more complex feature engineering or nonlinear methods, such as decision trees or random forests.

# 7. Conclusions and Recommendations

Key insights include:

1. Temperature significantly affects energy consumption, suggesting the need for tailored charging recommendations during extreme weather.

2. User type influences charging costs, warranting pricing strategies to ensure fairness.

Future work should explore additional predictive features and real-time data integration.

# 8. Code and Appendix

Below is the code used for data cleaning, analysis, and visualization:

# Load the dataset  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
  
data = pd.read\_csv('/mnt/data/ev\_charging\_patterns.csv')

## Data cleaning and exploratory analysis

data.info()  
desc\_stats = data.describe()  
  
# Correlation Heatmap  
plt.figure(figsize=(10, 8))  
sns.heatmap(data.corr(numeric\_only=True), annot=True, cmap="coolwarm")  
plt.savefig('/mnt/data/correlation\_heatmap.png')  
  
# Scatter Plot for Temperature vs. Energy Consumed  
sns.scatterplot(x=data['Temperature (°C)'], y=data['Energy Consumed (kWh)'])  
plt.savefig('/mnt/data/scatter\_energy\_temperature.png')  
  
# Boxplot for User Type vs. Charging Cost  
sns.boxplot(x=data['User Type'], y=data['Charging Cost (USD)'])  
plt.savefig('/mnt/data/boxplot\_user\_type\_cost.png')

## Hypothesis Testing

from scipy.stats import ttest\_ind

# Filter data by User Type

commuter\_costs = data[data['User Type'] == 'Commuter']['Charging Cost (USD)']

traveler\_costs = data[data['User Type'] == 'Long-Distance Traveler']['Charging Cost (USD)']

# Perform independent t-test

t\_stat, p\_value = ttest\_ind(commuter\_costs, traveler\_costs, equal\_var=False) # Welch's t-test

# Calculate descriptive statistics

commuter\_mean = commuter\_costs.mean()

traveler\_mean = traveler\_costs.mean()

commuter\_std = commuter\_costs.std()

traveler\_std = traveler\_costs.std()

commuter\_n = commuter\_costs.count()

traveler\_n = traveler\_costs.count()

## Regression

# Prepare data for multiple linear regression

selected\_features = ['Temperature (°C)', 'Charging Duration (hours)', 'Charging Rate (kW)', 'Vehicle Age (years)']

data\_clean\_multivariate = data.dropna(subset=['Energy Consumed (kWh)'] + selected\_features)

X\_multivariate = data\_clean\_multivariate[selected\_features].values # Predictors

y\_multivariate = data\_clean\_multivariate['Energy Consumed (kWh)'].values # Target

# Split data into training and testing sets

X\_train\_multi, X\_test\_multi, y\_train\_multi, y\_test\_multi = train\_test\_split(X\_multivariate, y\_multivariate, test\_size=0.2, random\_state=42)

# Create and train the multiple linear regression model

multi\_reg\_model = LinearRegression()

multi\_reg\_model.fit(X\_train\_multi, y\_train\_multi)

# Predict on test data

y\_pred\_multi = multi\_reg\_model.predict(X\_test\_multi)

# Calculate regression metrics

r2\_multi = r2\_score(y\_test\_multi, y\_pred\_multi)

mse\_multi = mean\_squared\_error(y\_test\_multi, y\_pred\_multi)

coefficients\_multi = multi\_reg\_model.coef\_

intercept\_multi = multi\_reg\_model.intercept\_

References

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

Footnotes

1[Add footnotes, if any, on their own page following references. For APA formatting requirements, it’s easy to just type your own footnote references and notes. To format a footnote reference, select the number and then apply the Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. (Note: If you delete this sample footnote, don’t forget to delete its in-text reference as well.)]

Tables

Table 1

[Table Title]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Head | Column Head | Column Head | Column Head | Column Head |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. To insert a table, on the Insert tab, tap Table. New tables that you create in this document use APA format by default.]

Figures



Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the APA Style Manual, 6th Edition.