Project Report: Analysis of Electric Vehicle (EV) Charging Station Usage

1. Introduction

This project investigates Electric Vehicle (EV) charging station data to uncover trends and relationships between charging session duration, greenhouse gas (GHG) savings, and energy consumption. The dataset is analyzed using statistical methods and visualizations to provide insights into charging patterns, identify correlations between key variables, and make predictions using regression models.

The statistical computing environment R was used for this analysis, leveraging its capabilities for data manipulation, statistical analysis, and visualization. The report outlines the various techniques applied and the results obtained, alongside an explanation of how R was used to implement each step.

---

2. Data Overview

The dataset was imported into R from a CSV file using the `read.csv()` function. This file contains information on various charging sessions at EV stations in a California city. After loading the dataset, an initial inspection was conducted using the `str()` function to display the structure of the data, ensuring the correct data types were assigned to each variable.

The following columns were retained for analysis:

- **End.Date**: The date a charging session was completed.
- **Station.Name**: The name of the charging station where the session occurred.
- **Charging.Time (hh:mm)**: The duration of the charging session, recorded in hours, minutes, and seconds.
- **GHG.Savings (kg)**: The amount of greenhouse gas saved during the session, measured in kilograms.
- **Energy (kWh)**: The energy consumed during the charging session, measured in kilowatt-hours.
- **Total Time (hh:mm)**: The duration the user keeps their EV in the station, recorded in hours, minutes, and seconds.

These variables were chosen for their relevance to analyzing charging behavior, GHG savings, and energy usage.

3. Data Preprocessing

Data preprocessing ensures the data is in a format suitable for statistical analysis. In this step, the date and duration fields were converted to usable formats, and missing values were handled.

3.1. Date Conversion

The `End.Date` column was converted from a character type to a Date object in R using the `as.Date()` function. This conversion was essential to enable time-based analyses, such as visualizing how GHG savings changed over time.

```
ev_data$`End.Date` <- as.Date(ev_data$`End.Date`)
```

3.2. Duration Conversion

The charging duration was initially provided as a string in "hh:mm

" format. To make it usable for statistical analysis, the duration was converted into total minutes. The following steps were applied:

1. The `strsplit()` function was used to split the duration into hours, minutes, and seconds.
2. The `sapply()` function was applied to calculate the total duration in minutes by converting hours to minutes and adding the seconds as fractional minutes.
3. A new column, `Duration_Minutes`, was created to store the calculated values.

```
duration_parts <-
strsplit(as.character(ev_data$`Charging.Time..hh.mm.ss.`), ":")
duration_in_minutes <- sapply(duration_parts, function(x) {
  as.numeric(x[1]) * 60 + as.numeric(x[2]) + as.numeric(x[3]) / 60
})
ev_data$Duration_Minutes <- duration_in_minutes
```

This conversion was essential for performing descriptive statistics and correlation analysis later in the project.

4. Statistical Analysis

Statistical analysis was carried out in R to explore key descriptive statistics and distributions of the charging duration. The `mean()`, `median()`, `var()`, and `sd()` functions were used to compute central tendency and variation measures. Additionally, the `moments` package was employed to calculate skewness and kurtosis, which provide insights into the distribution shape.

4.1. Central Tendency

- **Mean Charging Duration**: 123.14 minutes
- **Median Charging Duration**: 111.75 minutes.

```r
mean_duration <- mean(ev_data$Duration_Minutes, na.rm = TRUE)
median_duration <- median(ev_data$Duration_Minutes, na.rm = TRUE)
```

The **mean duration** of charging sessions was 123.14 minutes, which indicates that on average, most vehicles take just over two hours to charge. The **median** value of 111.75 minutes suggests that half of the sessions lasted less than approximately 112 minutes.

4.2. Variation

- **Variance**: 6176.17 minutes^2
- **Standard Deviation**: 78.59 minutes

```r
variance_duration <- var(ev_data$Duration_Minutes, na.rm = TRUE)
sd_duration <- sd(ev_data$Duration_Minutes, na.rm = TRUE)
```

The variance and standard deviation measure how much the charging durations vary from the average.

The **standard deviation** of 78.59 minutes reveal significant variability in charging times, indicating that some sessions take substantially longer than others.

4.3. Skewness and Kurtosis

- **Skewness**: 1.10 (positive skew)
- **Kurtosis**: 2.46

Skewness indicates asymmetry in the data distribution, while kurtosis indicates how heavy or light the tails are compared to a normal distribution.

```
skew_duration <- skew(ev_data$Duration_Minutes)
kurtosis_duration <- kurtosi(ev_data$Duration_Minutes)
```

The **skewness** value of 1.10 indicates that the distribution of charging durations is positively skewed, meaning there are some unusually long charging sessions, pulling the mean higher than the median.

The **kurtosis** value of 2.46 suggests a slight peak compared to a normal distribution, indicating that many charging durations cluster around the mean, but there are a few outliers with longer durations.

---

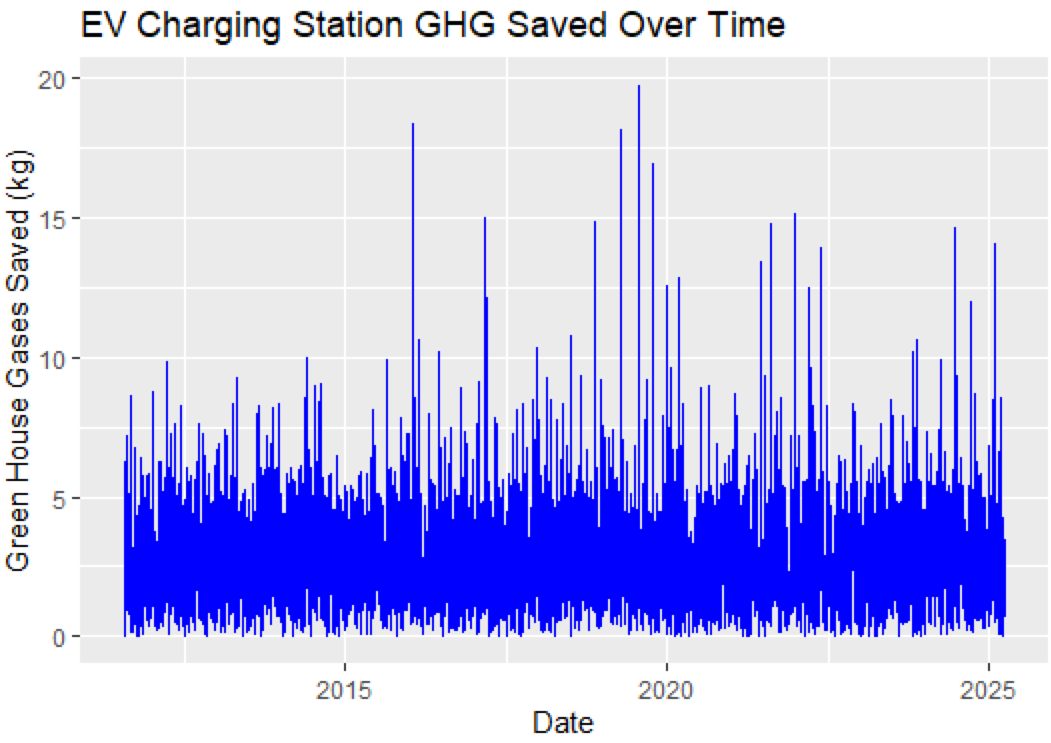
5. Data Visualization

Visualizations help to better understand trends and relationships in the data. Here, various plots were generated using the `ggplot2` package in R.

5.1. GHG Savings Over Time

A time series plot was created to visualize the trend of GHG savings over time, helping identify any seasonal or long-term changes in the data.

```
ggplot(ev_data, aes(x = `End.Date`, y = GHG.Savings..kg.)) +
  geom_line(color = "blue") +
  labs(title = "EV Charging Station GHG Saved Over Time", x = "Date", y =
"Green House Gases Saved (kg)")
```

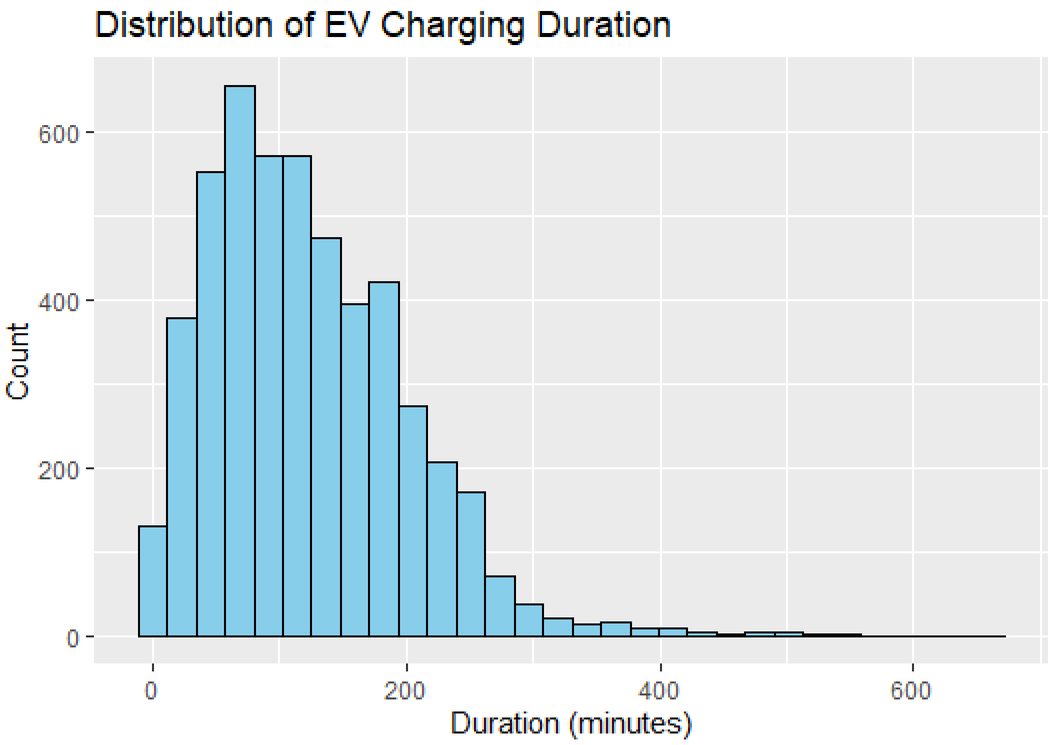


This line plot revealed fluctuations in GHG savings, which could indicate periods of high or low charging activity.

5.2. Distribution of Charging Duration

A histogram was plotted to display the distribution of charging durations across all sessions. This visualization helped to confirm the skewness and revealed the most common duration ranges.

```
ggplot(ev_data, aes(x = Duration_Minutes)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(title = "Distribution of EV Charging Duration", x = "Duration
(minutes)", y = "Count")
```



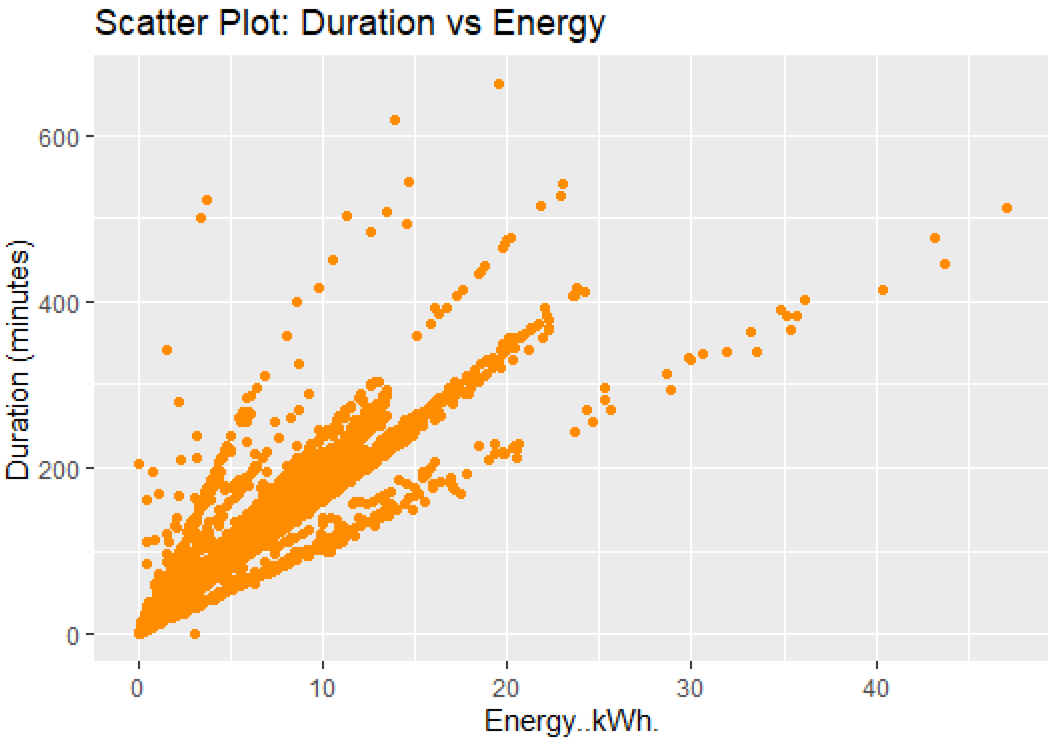
The histogram showed a clear right-skew, with most sessions lasting between 50 and 150 minutes.

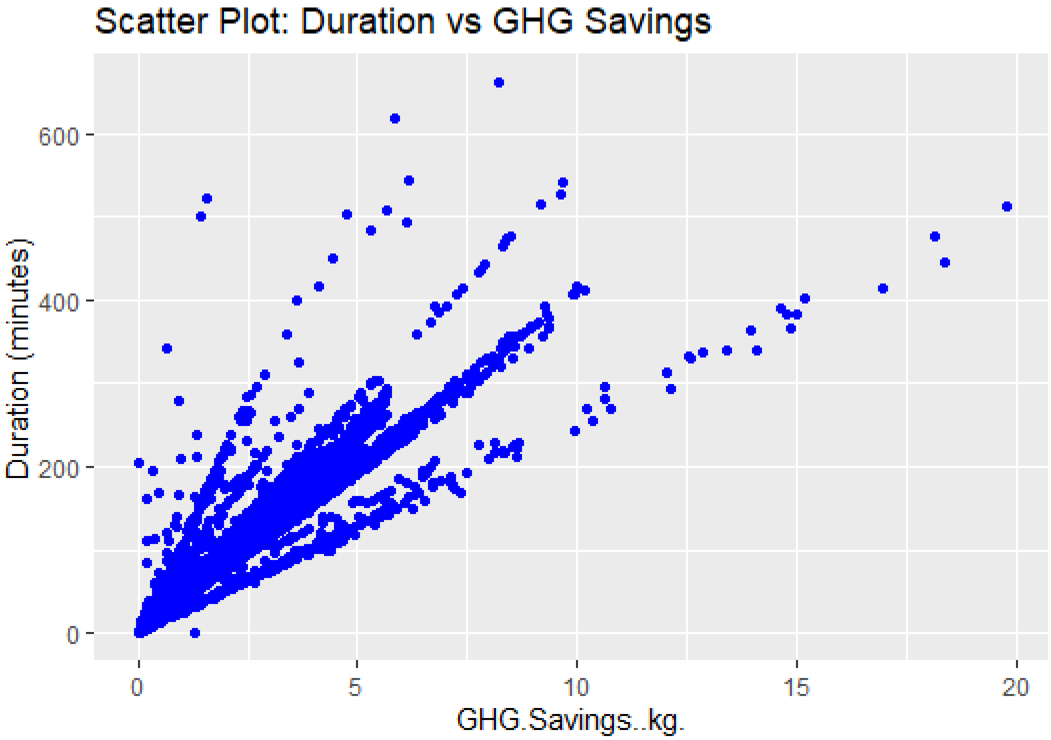
5.3. Scatter Plots: Duration vs GHG Savings and Energy

Scatter plots were used to visualize the relationships between charging duration and two key variables: GHG savings and energy consumption. These plots provided an initial indication of positive correlations, later quantified using statistical methods.

```
ggplot(ev_data, aes(x = `GHG.Savings..kg.`, y = Duration_Minutes)) +
  geom_point(color = "blue") +
  labs(title = "Scatter Plot: Duration vs GHG Savings", x =
"GHG.Savings..kg.", y = "Duration (minutes)")

ggplot(ev_data, aes(x = `Energy..kWh.`, y = Duration_Minutes)) +
  geom_point(color = "darkorange") +
  labs(title = "Scatter Plot: Duration vs Energy", x = "Energy..kWh.", y =
"Duration (minutes)")
```


Scatter Plot: Duration vs Energy
Duration (minutes)
600
400
200
0
0
10
20
30
40
Energy..kWh.


Scatter Plot: Duration vs GHG Savings
Duration (minutes)
600
400
200
0
0
5
10
15
20
GHG.Savings..kg.

6. Correlation Analysis

Correlation analysis was performed to assess the strength of linear relationships between charging duration and other variables like GHG savings and energy consumption. The `cor()` function was used for this purpose:

- **Correlation between Duration and GHG Savings**: 0.89
- **Correlation between Duration and Energy**: 0.89

```
cor_duration_ghg <- cor(ev_data$Duration_Minutes,
ev_data$`GHG.Savings..kg.`, use = "complete.obs")
cor_duration_energy <- cor(ev_data$Duration_Minutes,
ev_data$`Energy..kWh.`, use = "complete.obs")
cor_duration_total <- cor(ev_data$Duration_Minutes,
ev_data$Total_Duration_Minutes, use = "complete.obs")
```

Both correlations are very high (0.89), indicating a strong positive linear relationship between charging duration and both GHG savings and energy consumption. This suggests that as the duration of a charging session increases, the environmental benefits (in terms of GHG savings) and energy consumed also rise substantially.

---

7. Regression Analysis

Linear regression models were built to predict charging duration based on GHG savings and energy consumption. Both simple and multiple regression models were used.

R's `lm()` function was employed for fitting linear regression models.

7.1. Simple Regression

Simple linear regression models were created to explore how charging duration depends on GHG savings and energy consumption. The `summary()` function provided insights into the model's performance, such as the R-squared value, which indicates how much variance in the duration is explained by the independent variables:

```
simple_regression_ghg <- lm(Duration_Minutes ~ `GHG.Savings..kg.`, data =
ev_data)
summary(simple_regression_ghg)
```

The simple regression confirms that GHG savings are a significant predictor of charging duration. Higher GHG savings are associated with longer charging sessions.

```
Call:
lm(formula = Duration_Minutes ~ GHG.Savings..kg., data = ev_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-251.35  -14.60   -5.17    8.83  442.93 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       20.5423     0.8837   23.25   <2e-16 ***
GHG.Savings..kg.  36.7620     0.2613  140.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.29 on 4997 degrees of freedom
Multiple R-squared:  0.7984,	Adjusted R-squared:  0.7984 
F-statistic: 1.979e+04 on 1 and 4997 DF,  p-value: < 2.2e-16
```

7.2. Multiple Regression

A multiple regression model was created to predict charging duration based on both GHG savings and energy consumption:

```
multiple_regression <- lm(Duration_Minutes ~ `GHG.Savings..kg.` +
`Energy..kWh.`, data = ev_data)
summary(multiple_regression)
```

The multiple regression model provided even better predictive power, with a higher R-squared value, suggesting that the combined influence of both GHG savings and energy consumption can effectively predict the duration.

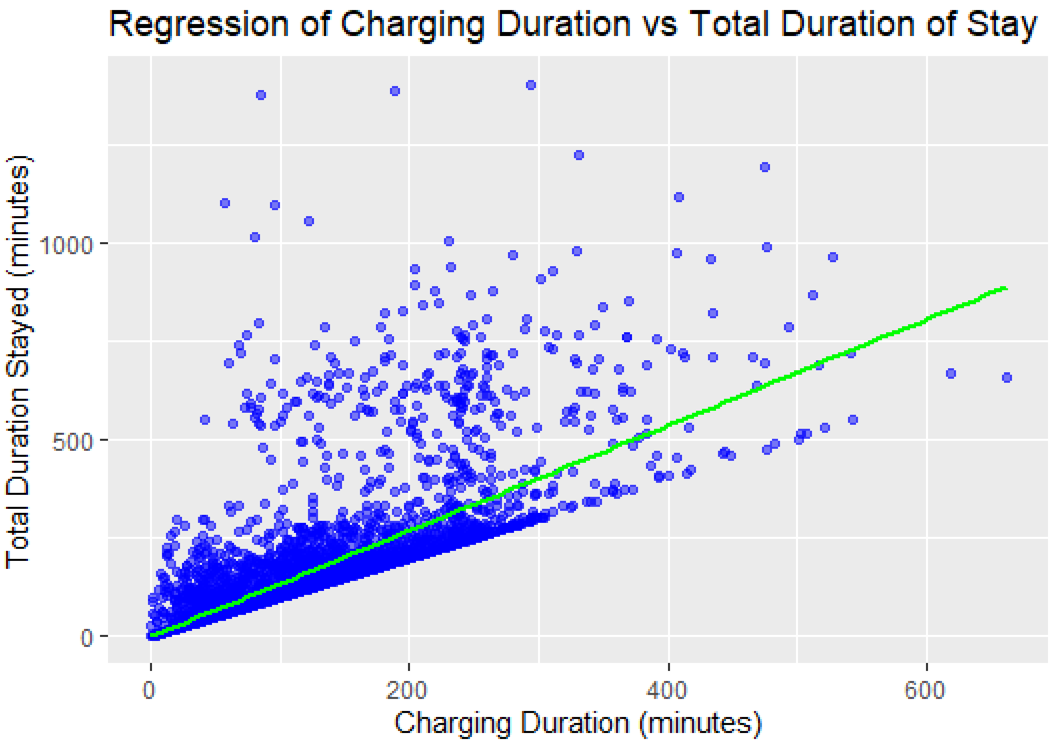
```
Call:
lm(formula = Duration_Minutes ~ GHG.Savings..kg. + Energy..kWh., 
    data = ev_data)
Residuals:
    Min      1Q  Median      3Q     Max 
-250.78  -14.56   -5.19    8.83  442.78 
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        20.5379     0.8838  23.239   <2e-16 ***
GHG.Savings..kg. 1325.5917  1728.0580   0.767    0.443    
```

```
Energy..kWh.     -541.3088   725.7848  -0.746    0.456
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 35.29 on 4996 degrees of freedom
Multiple R-squared:  0.7984,     Adjusted R-squared:  0.7984
F-statistic:  9896 on 2 and 4996 DF,  p-value: < 2.2e-16
```

7.3. Regression of Total Duration Stayed vs Charging Time

The regression analysis between **Total Duration** (the entire time a vehicle stays at the charging station) and **Charge Duration** (the time actively spent charging) provides valuable insights into user behavior, station utilization, and the efficiency of the charging infrastructure. Here's how this specific analysis can be beneficial.

```
ggplot(ev_data, aes(x = Duration_Minutes, y = Total_Duration_Minutes)) +
  geom_point(color = "blue", alpha = 0.5) +
  labs(title = "Regresseion of Charging Duration vs Total Duration of
Stay",
       x = "Charging Duration (minutes)", y = "Total Duration Stayed
(minutes)") +
  geom_smooth(method = "lm", se = FALSE, color = "green")
```



8. T-Test and ANOVA

8.1. T-Test

A t-test was conducted to compare charging durations between two stations: "PALO ALTO CA / BRYANT #1" and "PALO ALTO CA / HAMILTON #1."

- **P-value**: 0.049
- The p-value of 0.049 indicates a statistically significant difference between the two stations' charging durations at the 5% significance level.

```
t_test_result <- t.test(station1_duration, station2_duration)
```

```
	Welch Two Sample t-test

data:  station1_duration and station2_duration
t = -1.9664, df = 1853.9, p-value = 0.0494
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.51548418  -0.01916231
sample estimates:
mean of x mean of y 
 123.2860  130.5534 
```

8.2. ANOVA

An ANOVA (Analysis of Variance) was performed to compare the charging durations across all stations in the dataset.

- **P-value**: 3.06e-09
- The extremely low p-value suggests that there are significant differences in the charging durations across the different stations.

```
anova_result <- aov(Duration_Minutes ~ `Station.Name`, data = ev_data)
```

```
> summary(anova_result)
               Df   Sum Sq Mean Sq F value   Pr(>F)    
Station.Name    4   280192   70048   11.44 3.06e-09 ***
Residuals    4994 30588325    6125                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

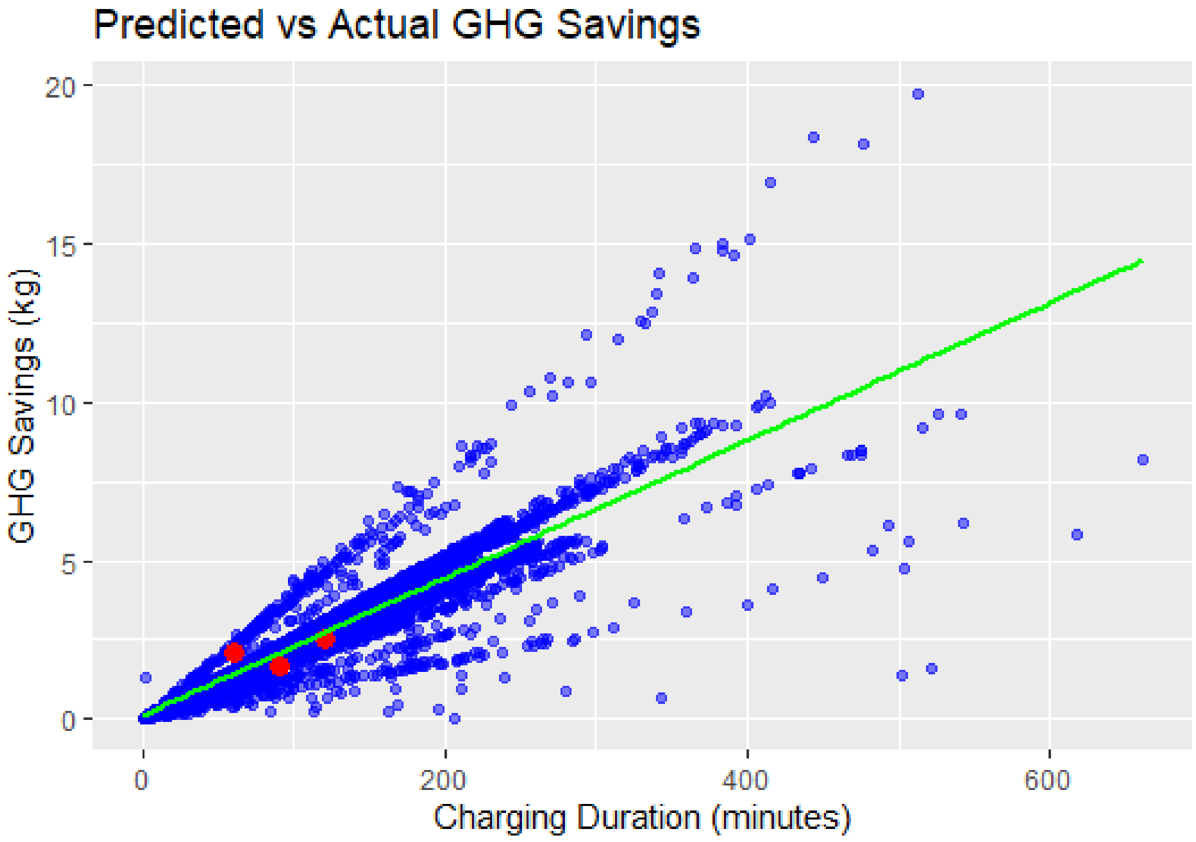
9. Predictions

Finally, the `predict()` function was used to generate predictions for GHG savings based on hypothetical charging durations and energy consumption values. This step demonstrated how regression models can be used to make data-driven predictions and informed decisions.

```
predicted_ghg_savings <- predict(ghg_regression, newdata =
new_data_for_ghg)
```

The predicted values were then compared with actual data using plots.

```
ggplot(ev_data, aes(x = Duration_Minutes, y = `GHG.Savings..kg.`)) +
  geom_point(color = "blue", alpha = 0.5) +  # Actual data points
  geom_point(data = new_data_for_ghg, aes(x = Duration_Minutes, y =
predicted_ghg_savings),
             color = "red", size = 3) +  # Predicted points
  labs(title = "Predicted vs Actual GHG Savings",
       x = "Charging Duration (minutes)", y = "GHG Savings (kg)") +
  geom_smooth(method = "lm", se = FALSE, color = "green")
```



10. Conclusion

The analysis revealed several important insights into EV charging patterns:

- Most charging sessions last between 50 and 150 minutes, with a small number of outliers.
- There is a strong positive correlation between charging duration, GHG savings, and energy consumption, highlighting the environmental benefits of longer charging sessions.
- Significant differences in charging durations exist between different stations, suggesting that station characteristics may influence usage patterns.
- The regression models developed are effective in predicting charging duration based on GHG savings and energy consumption, offering valuable predictive power for future usage planning.

R played a critical role throughout the project, from data preprocessing and visualization to conducting statistical analysis and building predictive models. This analysis provides insights that could help improve EV infrastructure management, optimize station usage, and contribute to environmental sustainability efforts.