**Electric Vehicle Average Charge Energy and Seasonality Hypothesis Testing**

Group 16 Analysis

Statistics for Data Science - University of Waterloo

December 12, 2022

[**1. Background**](#_dzs0ilwmjiy6) **3**

[**2. Objectives**](#_7kz9e66dva9r) **3**

[3. Approach](#_9dszkvgo3oki) 3

[**4. EDA (Data Preparation, Cleaning and Pipeline)**](#_rzaigzblsy4l) **4**

[4.1 Data Source](#_5u8kktl96nm0) 4

[4.2 Data Quality](#_txn9k972wm1v) 4

[4.3 Data Preparation and Data Cleaning](#_6xvflbz47omy) 5

[**4.4 Data Exploration**](#_i9z165seh6b5) **5**

[**5. Analysis/Model**](#_j112ikire9jd) **5**

[**6. Conclusions**](#_khka6s7p5w32) **5**

# Background

A team led by public policy professor Omar Asensio used a field experiment to collect data on 3,395 electric vehicle charging sessions between November 2014 and October 2015. The resulting [dataset](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NFPQLW)  contains charging sessions from 85 EV drivers with repeat usage at 105 charging stations across 25 sites at a workplace charging program in the United States. It indicates the date and length of each charging session, total energy used, cost, and more.

# Objectives

The objective of this analysis is to determine if there is a correlation between seasonality and the average charging energy per day.

To do this we will perform a hypothesis test on the following null and alternative hypotheses:

* Ho (Null Hypothesis): The average charging energy per day has no correlation with the season.
* Ha (Alternative Hypothesis): The average charging energy per day is correlated with the season.

Performing this hypothesis test will provide information about the consumer behavior of energy by EV drivers, and may assist EV charging outlets and electric utilities in tailoring their services to meet consumer demand.

Seasons are defined as follows:

* Spring - March 1 - May 31
* Summer - June 1 - August 31
* Fall - September 1 - November 30
* Winter - December 1 - February 28 (29th if a leap year)

## Approach

Working with the available data we created an expanded data set, adding a seasonal component to correlate this to the dependent variable, total kWh charged.

In order to look for trends in the analysis first, we will use various data visualizations to understand trends over time and by season, and use correlation coefficients and interpret the results before setting up a model.

# Exploratory Data Analysis

## 4.1 Data Source

For this analysis, the [Electric Vehicle Charging Dataset](https://www.kaggle.com/datasets/michaelbryantds/electric-vehicle-charging-dataset) was used from [www.kaggle.com](http://www.kaggle.com), an online community of data scientists and machine learning practitioners. This dataset contains a collection of EV charging data from a workplace charging program

## 4.2 Data Quality

The below is a heatmap to show the missing data.

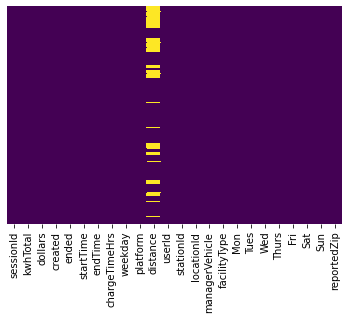


Figure 1: Heatmap of data

Distance is the only data point with missing values. According to the study details, <https://github.com/asensio-lab/workplace-charging-experiment/>, distance column is the distance from a user's home to the charging location, expressed in miles except where the user did not report address. In our study, distance from the EV driver’s home to the charging station was considered irrelevant, the column was ignored.

## 4.3 Data Preparation and Data Cleaning

In order to ease our use of the dataset, we performed the following cleaning actions:

* Modified the date/time columns data to fit a timestamp type. For example “0014-11-18 15:40:26” became: “2014-11-18 15:40:26”. This has been done specifically for the “created” and “ended” columns.
* Created the “year”, “month” and “season” columns based on the “created” timestamp column.
* Dropped the “distance” column as it has a high percentage of null values and it is irrelevant to our hypothesis.

## 4.4 Data Exploration

After data cleaning, we used different tools to explore the data.

### 4.4.1 Scatter Plot of kWh Totals by Station and Season

Using a scatter plot, we viewed the total kWh by station per season. Depending on the geographic location of the charging station, seasonal effects on EV charging may be more or less pronounced. Unfortunately, the data set did not include geographical information, so we were unable to determine exact locations of each charging station.

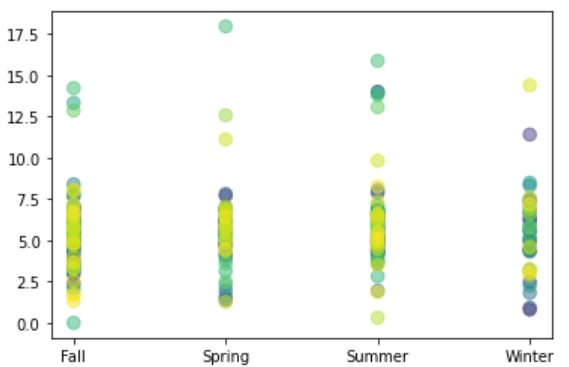


Figure 2: Scatter Plot of kWh Totals

### 4.4.2 Histogram of Charging Time and kWh by Season

From a plot of charge time mean by season, we see that winter has the highest charge time mean, followed by summer and fall, whereas spring has the lowest charge time mean.

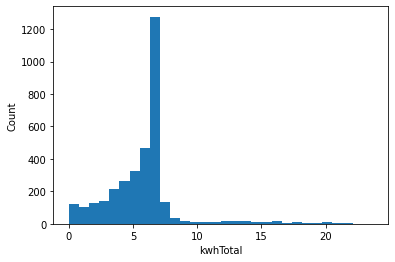
From a plot of total kWh by season, summer has the highest total kWh, followed by spring and fall, whereas winter has the lowest total kWh.

# 

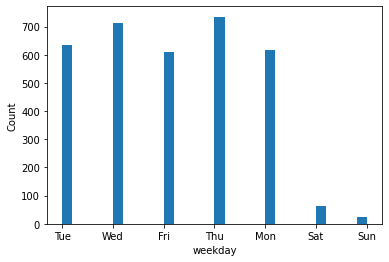
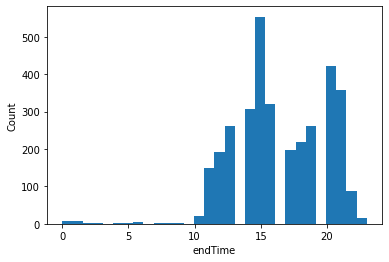
Figure 3: Histograms of Mean Charging Time and kWh Totals

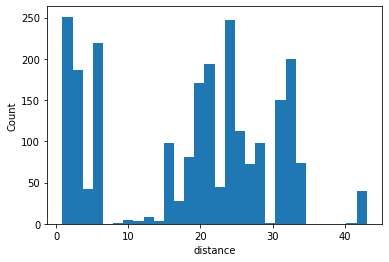
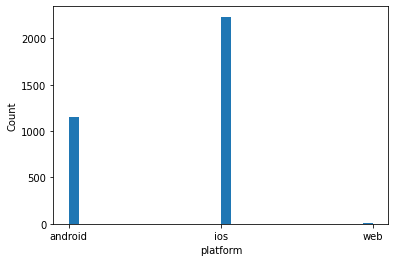
### 4.4.3 Data Set Histograms

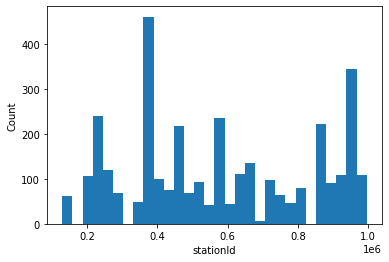
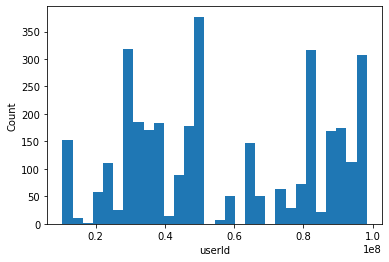
Next, we created histogram plots of all data points to determine if the data contained any significant outliers.

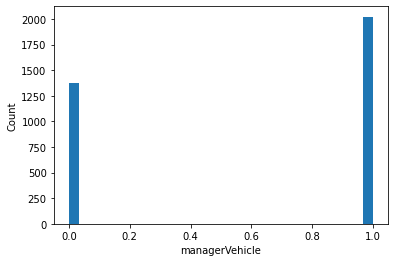
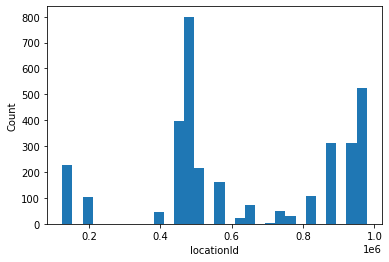


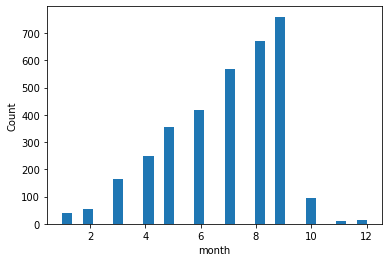
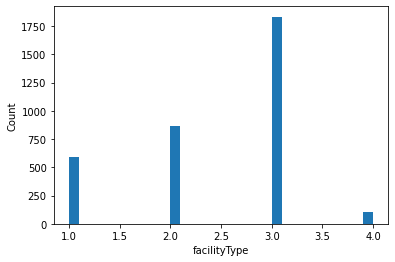
# 











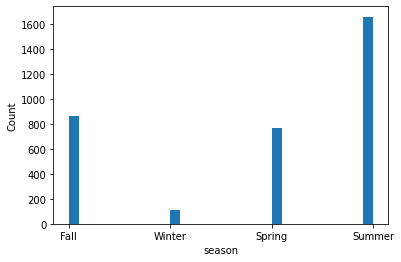


Figure 4: Histograms of Data Set

### 4.4.4 Correlation Matrix

We used a correlation matrix to see the correlation between different variables. A correlation coefficient tells us the relationship between two variables. A positive value means that as one variable increases, so does the other, and a negative value means that as one variable increases the other decreases.

When examining the results for the dependent variable, charging energy, most variables have small degrees of positive correlations with charge energy (kWh Total). The strongest relationship exists between charging energy and charging time, which may be expected as more charging energy could be added to an electric vehicle over time.

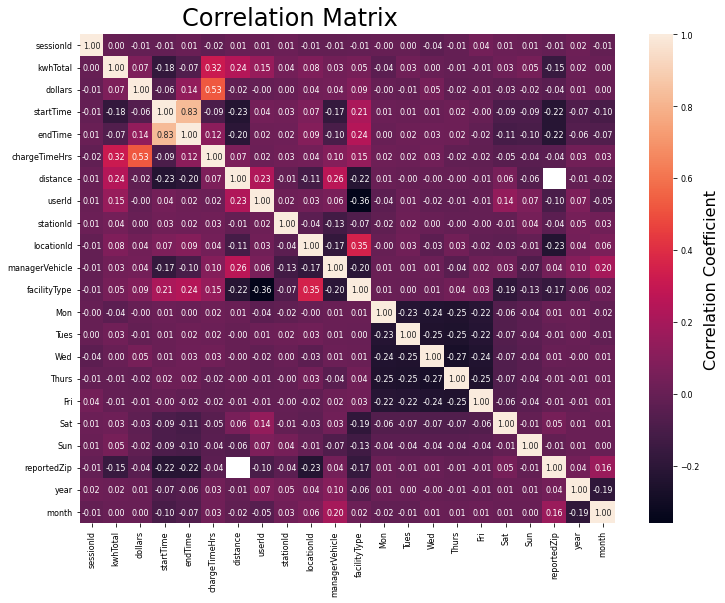


Figure 5: Correlation Matrix of Data

### 4.4.5 Scatterplot

Scatterplots of the variables were created to visualize the correlations identified in the correlation matrix.

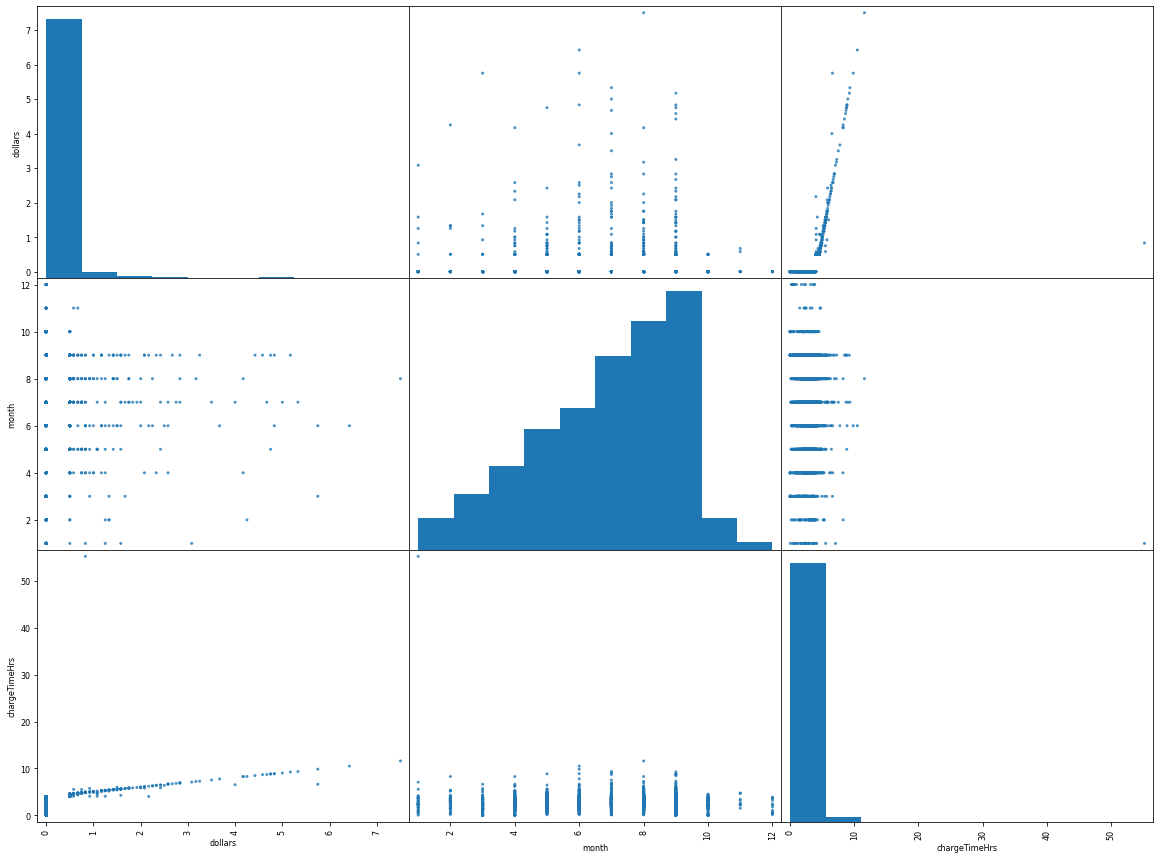


Figure 6: Scatterplots of Data Set

# Analysis/Model

We deployed an OLS regression model to the cleaned data set. The OLS regression results without transformation had an R2 value of 0.70. Further transformation was applied, a square root transformation resulted in an R2 value of 0.811.

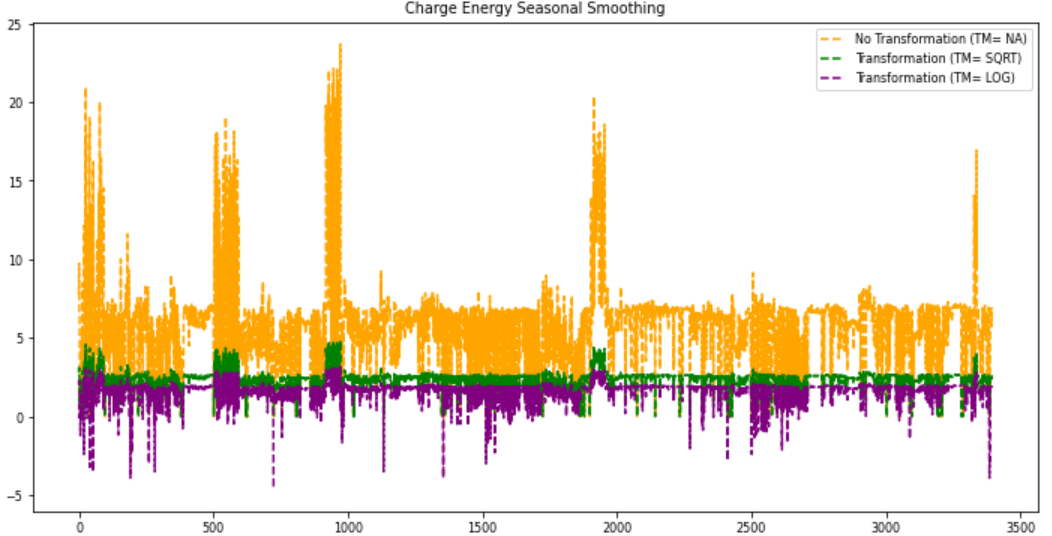


Figure 7: Charge Energy Smoothing by Transformation

# Conclusions

Using this data set, we reached a conclusion that the average charge energy per day is correlated with the season. The OLS regression model indicates that the R2 value is 0.811, which suggests the correlation is strong. As such we fail to reject the null hypothesis for this analysis.

For a future analysis, it is recommended reviewing the geographic location of each charging station as the seasonal effects may vary based on the location.