Tesla Battery Degradation

Summary

Degradation of lithium-ion batteries represents a challenging problem for electric car companies that concentrate on delivering high-quality cars with high driving range and low cost of charging. Even though Li-ion batteries became the future of the battery market, they are still not impervious to decrease in performance that is caused by cycling, elevated temperatures, and aging. Tesla represents one of the most prominent disruptors of the electric car market as they deliver a variety of models with different battery packs and performance specifications. Every Tesla driver is provided with a warranty for battery failure which does not cover degradation. Degradation is equivalent to the decrease in battery capacity over time which leads to the decrease in range (miles) that a Tesla can drive before re-charging. My goal is to develop a model that can predict the expected battery degradation on the basis of the Tesla Model S Survey taken by drivers in Asia, Europe, Canada, and USA. This information is useful for future owners of Tesla as well as other businesses that are interested in improving the performance of electric cars.

Data Wrangling

Dataset description

The data set was downloaded from Google Drive in a form of MaxRange Tesla Battery Survey.xlsx file. The goal was to use python to import and inspect the data for separate spreadsheets from different locations which include Asia & Europe, USA, Canada, and UK; isolate relevant variables; determine the correlation between the target variable (remaining range in %) and other variables, organize the data frames for different locations into one data frame, and resolve any missing, invalid, or corrupted rows.

Import packages, .xlsx file, and inspect the data

After the raw data set was imported from the Google Drive, the initial inspection of the Excel file established that there were 4 spreadsheets of interest from different locations. Each location was imported separately by parsing each spreadsheet into a data frame.
Each location data frame was cleaned of empty rows which contained only NaN values. All the separate data frames were merged into one master data frame. The resulting master data frame `tesla survey` contains 1156 observations and 54 columns which represent answers provided by Tesla drivers and related to characteristics of their cars.

Organize and re-name columns

Before merging the location data frames, the column names were changed because the original header was too descriptive. All 54 columns were renamed to merge the data frames. The last five columns with names that include `book_keeping…` were eliminated because they had 0 non-null values. Columns which contain information about chart data used by the owner of the file were eliminated too. Some of the columns represent additional optional information which users could provide if they had any recent trips with overnight charging. Majority of the users did not answer these questions (92.5%). Therefore, these columns were eliminated. As a result, we have 36 columns which contain required, calculated, and optional information about Tesla battery performance.

Some of the columns for Asia & Europe as well as Canada had inconsistent units for mileage, mileage per day, range for new cars, range at 100% charge, and range after correction because in the original data frame, SI units were used. These columns were transformed to miles and miles/day from km and km/day. Additionally, the units for lifetime average energy consumption were also different for countries with SI units (Wh/km as compared to Wh/mi used in USA and UK). Therefore, lifetime average energy consumption column for Asia & Europe as well as Canada was converted to Wh/km.

Evaluate NULL values

Because we identified the empty rows in each location data frame before merging, we did not have to perform any extensive investigation of the master data frame. There are 4 observations (rows) that have no information about manufacturing date which is why the mileage per day was not calculated for them. It is also impossible to obtain any information about vehicle cycles and other age-related characteristics for these rows. Therefore, they were removed.

Additionally, vehicle cycles column contained 13 NaN observations. After investigating these rows, it was determined that users simply chose not to provide any information for lifetime average energy consumption which is used for calculating the number of vehicle cycles. In the original Excel file, the owner proposed to drivers who did not know the lifetime average energy consumptions of their vehicles to substitute it with an average for their respective location. However, some drivers still did not use the average to fill it in. That is why the missing lifetime average energy consumption values were substituted with the averages of this column for each location. Then, the missing vehicle cycles values were filled in using the following formula:

$$Cycles = \frac{Mileage \cdot Lifetime\ avg\ energy\ consumption + vampire\ loss \cdot battery\ life}{0.5(1 + remaining\ range) \cdot original\ capacity}$$

The resulting data frame contained 1152 observations and 36 columns.

Data manipulations with columns

As some features of the data frame for Tesla survey represent categorical variables, it would be useful to find any groups which are of a smaller size than the rest to merge them with bigger groups. For example, location column had only 10 observations for drivers residing in UK. Asia & Europe category seemed to be broad enough to be able to include UK.

Therefore, all the observations containing UK as location were re-named to Asia and Europe, including UK.

The columns describing charge frequency also had some small groups which could be included in the bigger groups for statistical analysis. For the full charge frequency, ‘B) twice a week’ group contained only 8 observations. It was merged with ‘C) weekly’, and the new group was re-named to become ‘B) once or twice a week’. For the empty charge frequency, ‘B) twice a week’ group had only 1 observation while ‘C) weekly’ group had 6 observations. These groups were added to ‘D) twice a month’ group, and a new group called ‘D) one to four times a month’ was created. For the daily charge level, only 6 observations were attributed to the daily charge level of 50%. Therefore, it was decided to add these observations to the 60% group and re-name it to ‘<= 60%’ for accuracy of representation.

Inferential Statistics

Relationships to investigate from initial data exploration

In the process of initial data exploration, it was determined that some features, including the total mileage of a car, battery age, and vehicle cycles are negatively correlated with the remaining range. The remaining range stayed quite high for the majority of data points above 90% of remaining battery capacity. The graphical representation of these relationships is shown in Figure 1.

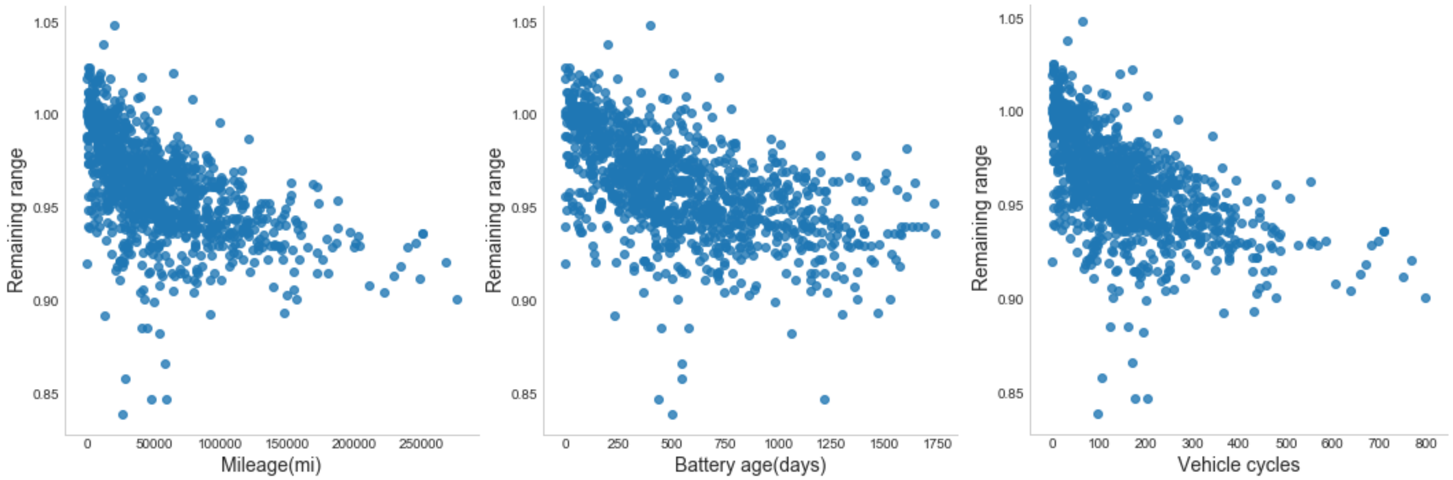


Figure 1: Relationships between the dependent variable (remaining range) and independent variables (mileage, battery age, and vehicle cycles).

Optional survey questions related to car charge frequency and other battery characteristics as well as location of the drivers were also investigated. It was determined that majority of the drivers were from ‘Asia & Europe’ region. Most of them supercharged at the Tesla charging station twice a month (Figure 2). They almost never had a fully charged battery while driving (only a few times a year), and once or twice a year, they tend to completely run out of battery charge. Most of survey participants had 90% daily charge level.

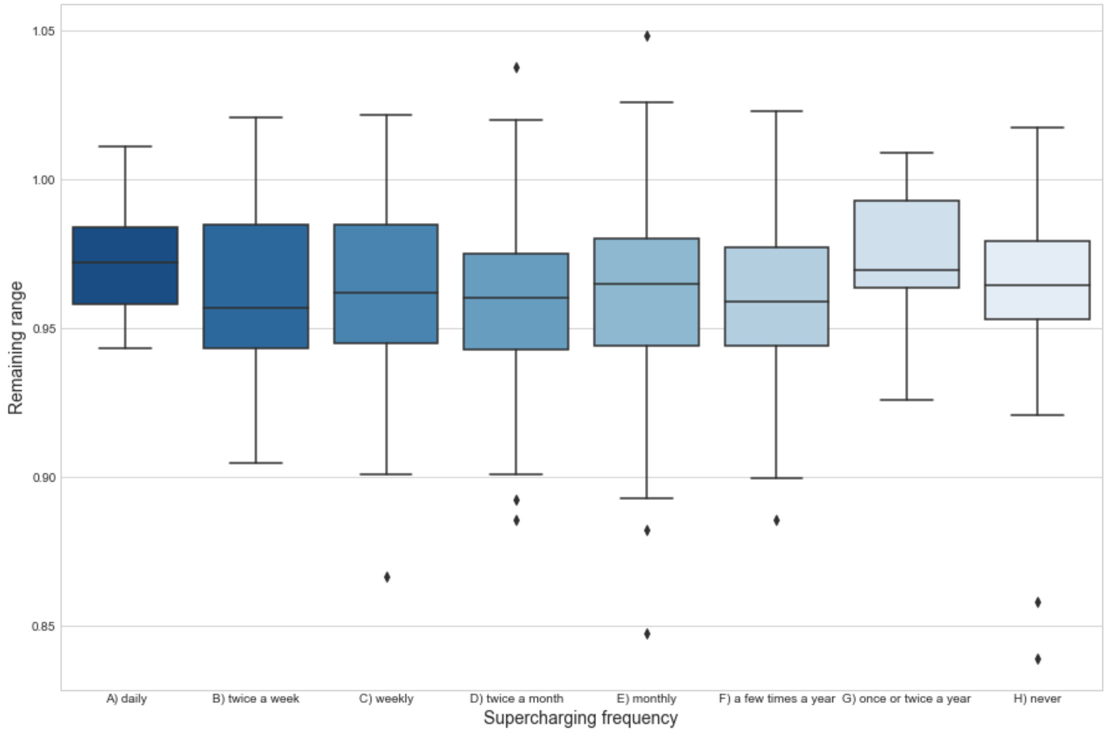


Figure 2: Relationship between the supercharging frequency and remaining range.

The remaining range median for different locations was approximately the same which means that geographical region did not have any effect on the remaining range (Figure 3). Frequency of supercharging groups had approximately the same medians for remaining range. The relationship between remaining range and fully charged battery frequency had a negative trend with the lowest remaining range for drivers who had 100% charge on a daily basis (Figure 4).

The daily charge value also had some influence on the remaining range - the lowest remaining range median was obtained for a group of drivers who had an average daily charge level around 100% (Figure 5).

The goal is to determine the statistical significance of correlations between the remaining range and other features. Because optional features describing battery use are represented by categorical variables, we can use analysis of variance for comparing the means of three or more groups.

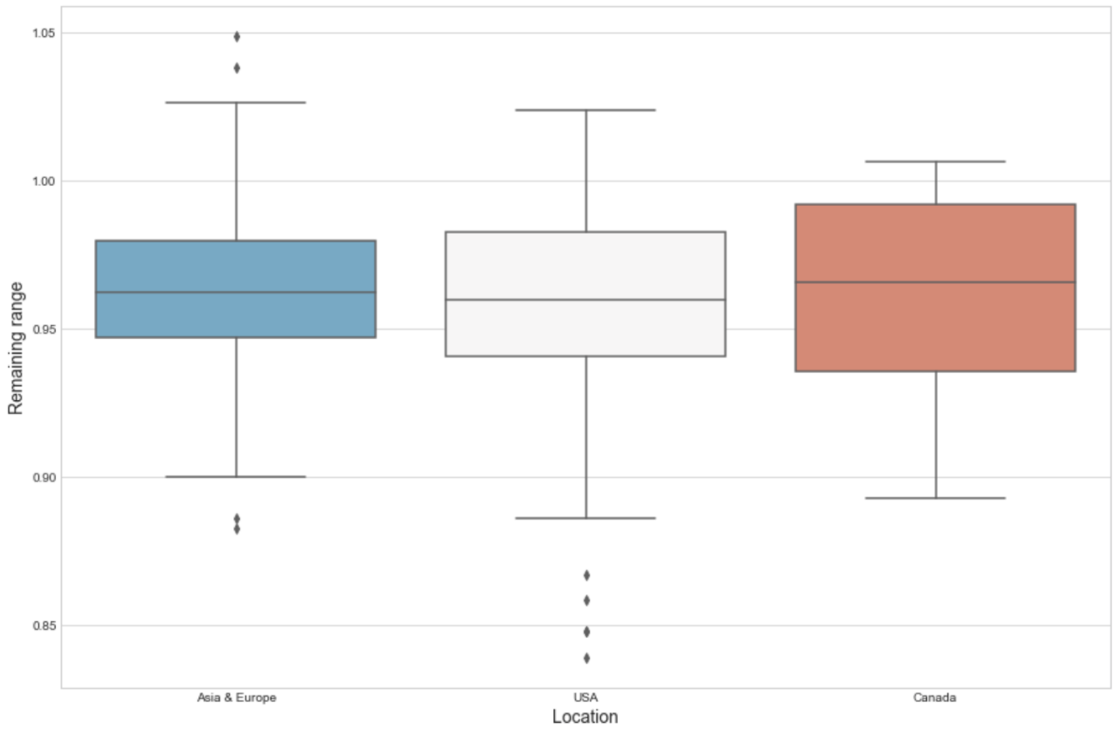


Figure 3: Relationship between the location and remaining range.

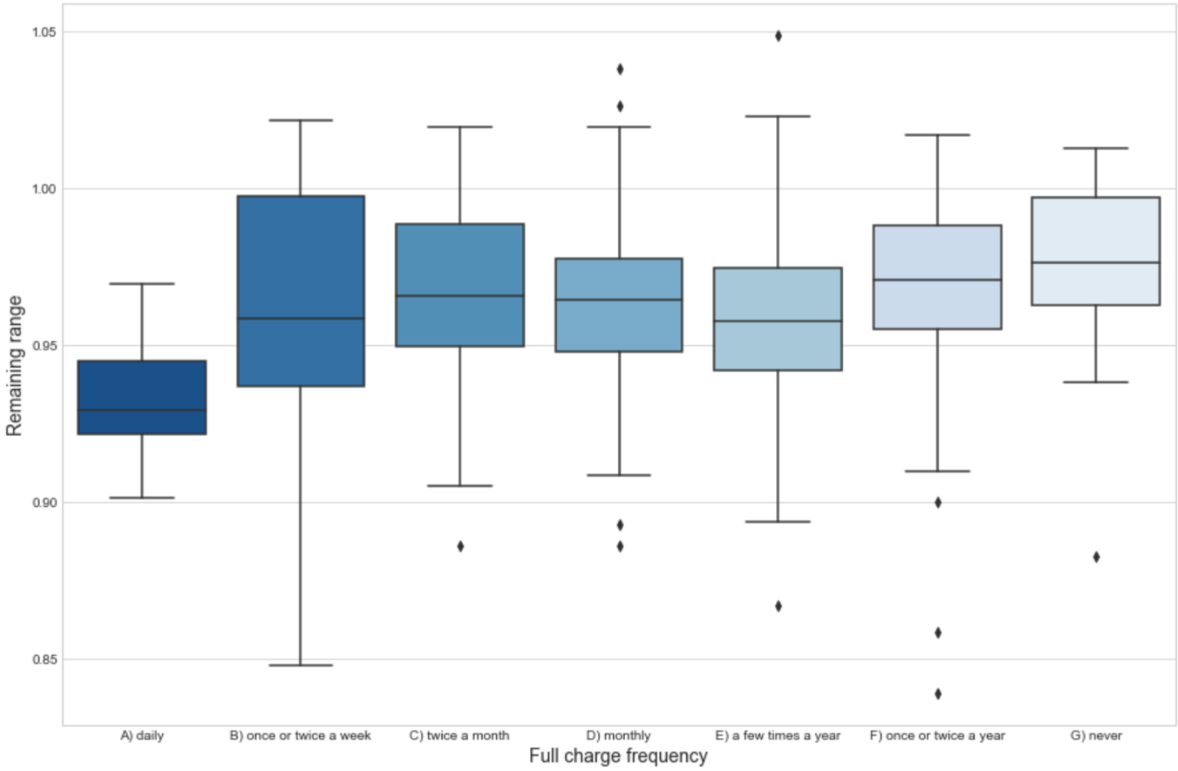


Figure 4: Relationship between the full charge frequency and remaining range.

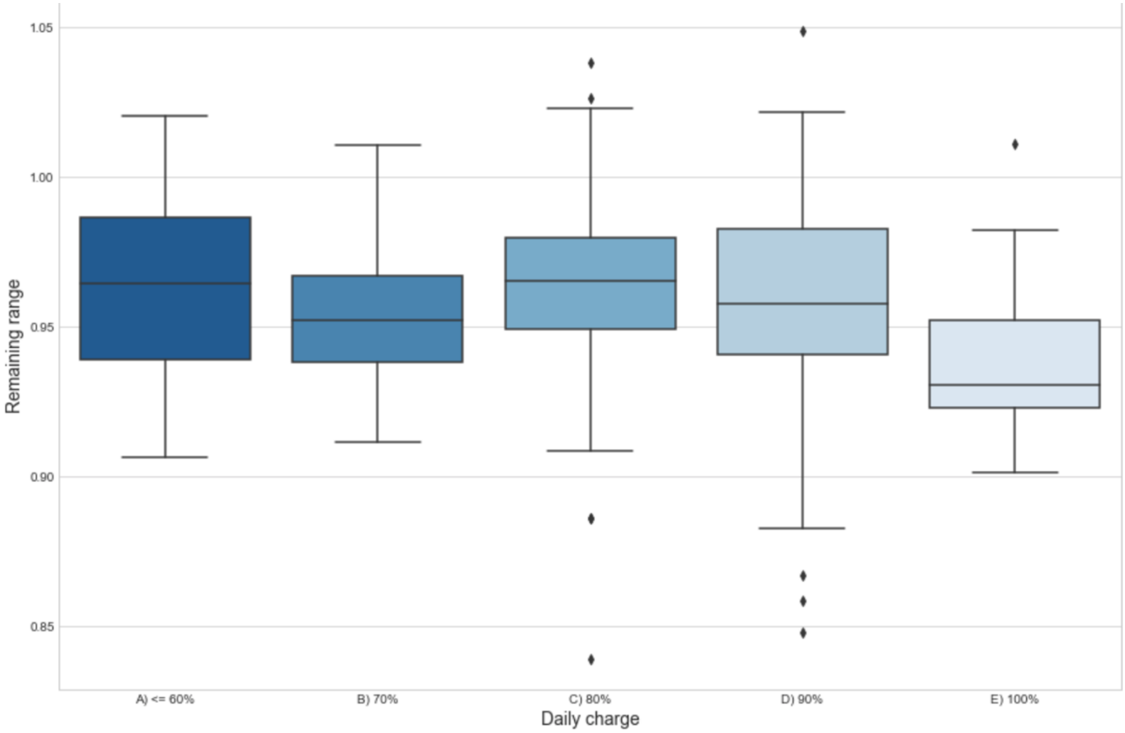


Figure 5: Relationship between the daily charge level and remaining range.

Analysis of correlations

The goal of this section is to determine how statistically significant the correlations between remaining range and independent features (mileage, battery age, and vehicle cycles) are. First, we set up an appropriate hypothesis test for each correlation.

In our case, it would be useful to determine the Pearson correlation coefficient between each feature and remaining range as it would measure the strength of a linear association between two variables. The value of the Pearson correlation ranges from -1 to 1 with 0 denoting the absence of correlation between two variables. According to the preliminary scatter plot, we could expect that the calculated Pearson correlation coefficient is negative as the high values of remaining range are associated with low car mileage, small battery age, and few vehicle cycles. The results obtained for each analysis of correlation are given in Table 1.

From Table 1, we can conclude that we can reject the null hypothesis for each feature analysis with 95% confidence. Therefore, there is some correlation between the remaining range and independent car features which include mileage, battery age, and vehicle cycles. All correlations are negative which was expected from preliminary data exploration.

Table 1: Pearson correlation coefficients for remaining range and independent features (mileage, battery age, and vehicle cycles).

Feature name	Hypothesis	Pearson correlation coefficient	p-value < 0.05
Mileage (mi)	**Ho:** There is no correlation between the remaining range and car mileage **Ha:** There is a correlation between the remaining range and car mileage	-0.564	TRUE
Battery age (days)	**Ho:** There is no correlation between the remaining range and battery age **Ha:** There is a correlation between the remaining range and battery age	-0.558	TRUE
Vehicle cycles	**Ho:** There is no correlation between the remaining range and vehicle cycles **Ha:** There is a correlation between the remaining range and vehicle cycles	-0.583	TRUE

Analysis of variance (ANOVA)

Analysis of variance is used for comparing the ratio of systematic variance to unsystematic variance in a data set. We are primarily interested in variance due to groups which the optional categorical features consist of. The ratio obtained as a result of this comparison is called F-ratio. A one-way ANOVA can be seen as a regression model with a single categorical predictor. The goal is to determine if the remaining range means of groups are the same or if at least two groups differ from each other in the mean remaining range value. ANOVA is quite robust, i.e. it can deal with some deviation in distributions of groups. However, it is important to make sure that the groups with smaller number of samples do not have higher standard deviation than groups with bigger sample size. For each ANOVA test, the standard deviations of different groups were compared. In all tests, the standard deviations were found to be approximately the same. The results of ANOVA, including F-ratio and p-value, can be found in Table 2 for different categorical features of the data set.

According to Table 2, we can conclude that we cannot reject the null hypothesis that the mean remaining range for all locations is the same. Therefore, we can conclude that location has no influence on battery degradation. Additionally, we cannot reject the null hypothesis that the mean remaining range for all supercharging frequencies is the same. Therefore, supercharging frequency has no influence on battery degradation. However, for the rest of categorical features, we can conclude that the null hypothesis can be rejected and two or more groups have different means for remaining range. For 100% charge frequency, drivers who had 100% charge on a daily basis had a much smaller mean remaining range. For empty charge frequency, drivers who never had 0 charge battery level had a higher mean remaining range. For daily charge level, drivers who had an average of 100% charge had a much smaller mean remaining range.

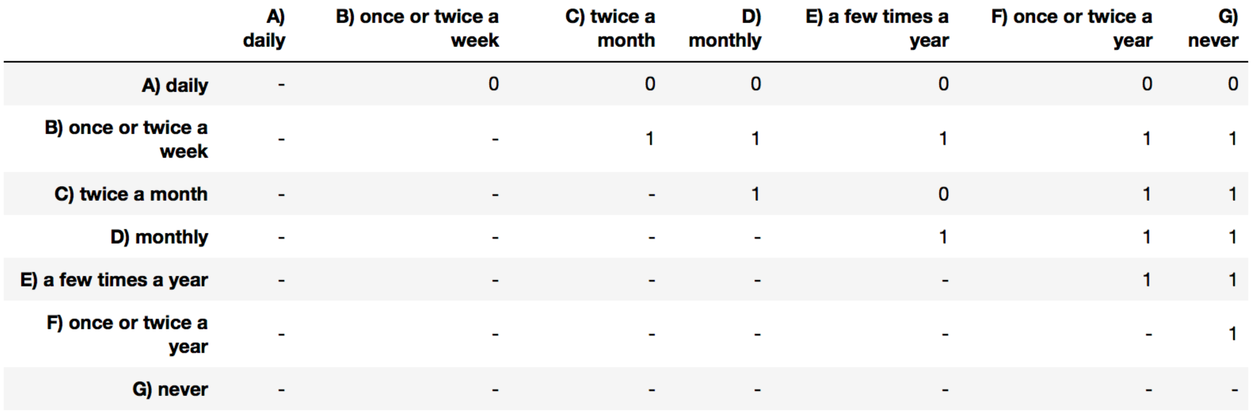
Pairwise t-tests adjusted for Bonferroni correlation

After the categorical features which affect the remaining range were determined, each feature was tested using pairwise t-testing adjusted for Bonferroni correction to eliminate the possibility of making Type I error (rejection of a true null hypothesis). The results are shown in Tables 3, 4, and 5 where 0 denotes that the null hypothesis (means of two groups are equal) was rejected.

Table 2: ANOVA results for different categorical features of the Tesla survey with respect to remaining range.

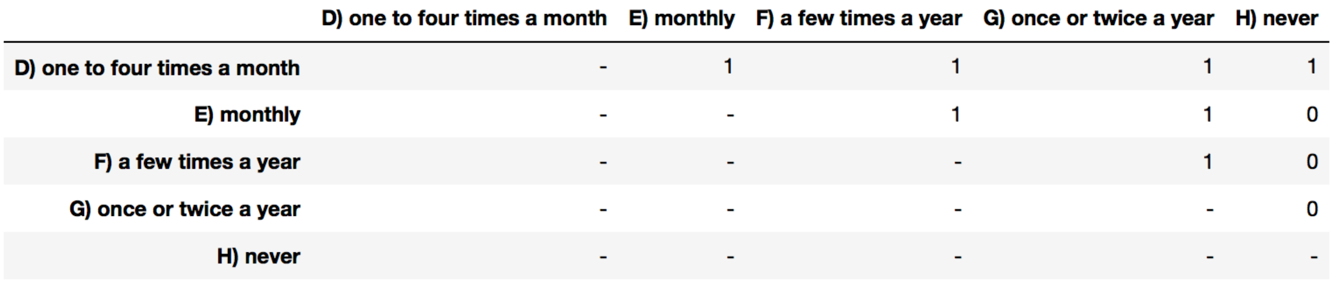
Feature name	Groups	Hypothesis	F-ratio	p-value
Location	1. Asia & Europe 2. USA 3. Canada	Ho: The mean remaining range for all locations is the same Ha: Two or more means for remaining range are different from others	2.29	0.101
Supercharging frequency	1. daily 2. twice a week 3. weekly 4. twice a month 5. monthly 6. a few times a year 7. once or twice a year 8. never	Ho: The mean remaining range for all supercharging frequencies is the same Ha: Two or more means for remaining range are different from others	1.55	0.145
100% charge frequency	1. daily 2. once or twice a week 3. twice a month 4. monthly 5. a few times a year 6. once or twice a year 7. never	Ho: The mean remaining range for all 100% charge frequencies is the same Ha: Two or more means for remaining range are different from others	9.49	~0
Empty charge frequency	1. one to four times a month 2. monthly 3. a few times a year 4. once or twice a year 5. never	Ho: The mean remaining range for all empty charge frequencies is the same Ha: Two or more means for remaining range are different from others	15	~0
Daily charge level	1. <= 60% 2. 70% 3. 80% 4. 90% 5. 100%	Ho: The mean remaining range for all daily charge levels is the same Ha: Two or more means for remaining range are different from others	8.78	~0

Table 3: Multiple pairwise t-test results for full charge frequency groups adjusted for α = 0.00238



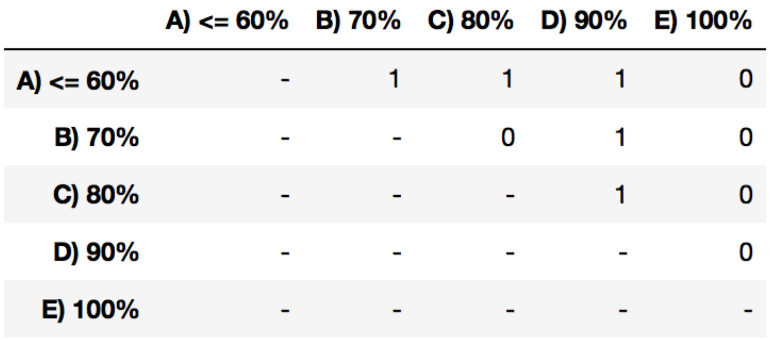
	A) daily	B) once or twice a week	C) twice a month	D) monthly	E) a few times a year	F) once or twice a year	G) never
A) daily	-	0	0	0	0	0	0
B) once or twice a week	-	-	1	1	1	1	1
C) twice a month	-	-	-	1	0	1	1
D) monthly	-	-	-	-	1	1	1
E) a few times a year	-	-	-	-	-	1	1
F) once or twice a year	-	-	-	-	-	-	1
G) never	-	-	-	-	-	-	-

Table 4: Multiple pairwise t-test results for empty charge frequency groups adjusted for α = 0.005



	D) one to four times a month	E) monthly	F) a few times a year	G) once or twice a year	H) never
D) one to four times a month	-	1	1	1	1
E) monthly	-	-	1	1	0
F) a few times a year	-	-	-	1	0
G) once or twice a year	-	-	-	-	0
H) never	-	-	-	-	-

Table 5: Multiple pairwise t-test results for daily charge levels groups adjusted for α = 0.005



	A) <= 60%	B) 70%	C) 80%	D) 90%	E) 100%
A) <= 60%	-	1	1	1	0
B) 70%	-	-	0	1	0
C) 80%	-	-	-	1	0
D) 90%	-	-	-	-	0
E) 100%	-	-	-	-	-

From Table 3, we can conclude that the presence of full charge on a daily basis had the most significant deviation from the mean remaining range in comparison to other frequencies of fully charged battery. Therefore, having a fully charged battery on a daily basis contributes to battery degradation. Additionally, we have rejected the null hypothesis for most of the cases (3/4) in which cars never had empty charge (Table 4). We can conclude that never discharging the battery fully is beneficial for remaining range and could also contribute to decrease in the rate of the battery degradation. According to the results obtained from Table 5, we have rejected the null hypothesis for all the cases in which one of daily charge values was 100%. We can conclude that daily charge level of 100% contributes to decrease in the remaining range. Also, 80% daily charge level seems to be more beneficial to the remaining range than 70% daily charge level.

Machine Learning Analysis

Pre-processing conversion of categorical variables

After performing pairwise t-testing on optional variables that represent full charge frequency, empty charge frequency, and daily charge level, it was determined that only one category represented a statistically significant result for each variable. Therefore, it was decided to convert these variables from categorical to boolean type. Full charge frequency values which were not equal to ‘A) daily’ were converted to 0, and ‘A) daily’ values were converted to 1. Empty charge frequency values which were not equal to ‘H) never’ were converted to 0, and ‘H) never’ observations were changed to 1. Daily charge level values which were not equal to ‘E) 100%’ were converted to 0, and ‘E) 100%’ observations were changed to 1.

Imputing missing boolean variables

Full charge frequency as well as empty charge frequency columns had 96 values missing after data cleaning while daily charge frequency column had 253 NaN values. It was decided to impute the missing values before performing the prediction using machine learning. The missing values were found using logistic regression in Sklearn. Mileage, battery age, vehicle cycles, and range capacity (kW) were used to predict the missing values in each column. The resulting prediction was substituted back into the original data frame instead of missing values. As a result, full charge frequency, empty charge frequency, and daily charge level had 1152 observations in each column which is consistent with the non-null observations in all other columns of importance.

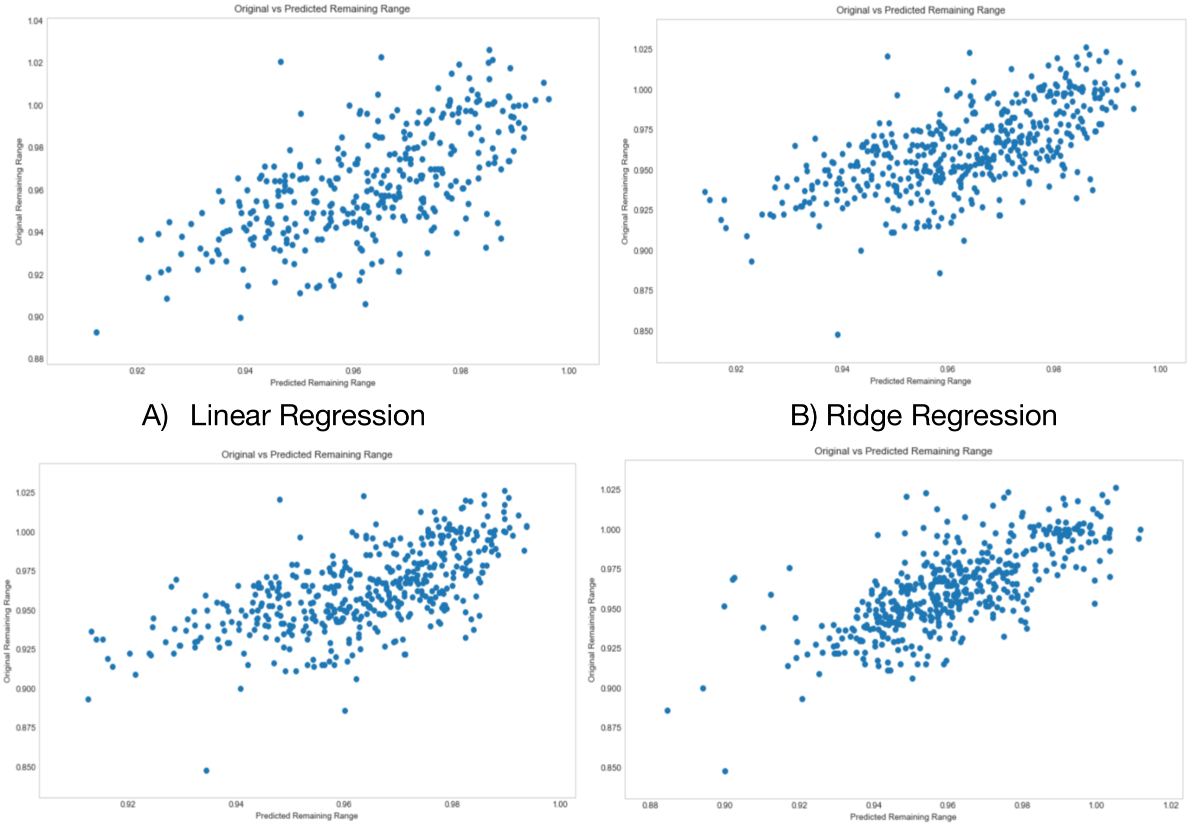
Additionally, to make the prediction for the remaining range better, optional columns which are based on frequency were used to create helper columns represented as the frequency column multiplied by the battery age to amplify the frequency result with respect to time. Therefore, three more columns denoted as ‘full_charge_with_time’, ‘empty_charge_with_time’, and ‘daily_charge_with_time’ were added to the master data frame.

Regression analysis

Using Statsmodels for regression with vehicle cycles, battery age, mileage, full charge frequency, empty charge frequency, daily charge, full charge with time, empty charge with time, daily charge with time, and range when new to predict remaining range resulted into R^2 of 0.573. The results for regression models built using Sklearn are shown in Table 6. The graphs depicting original and predicted range for each regression model are shown in Figure 6.

Table 6: Regression results for predicting remaining range (in %).

Regression	R^2	RMSE
Linear Regression	0.396	0.02
Elastic Net	0.4	0.02
Ridge Regression	0.4	0.02
Random Forest	0.5	0.018



A) Linear Regression

B) Ridge Regression

C) Elastic Net

D) Random Forest

Figure 6: Original remaining range vs Predicted remaining range for different types of regression in Sklearn.

According to results obtained for estimated regression coefficients, optional frequency columns, including full charge frequency, empty charge frequency, and daily charge, had the highest regression coefficients while mileage and battery age were ineffective at predicting the remaining range. As expected, Random Forest regression produced the best results while Linear regression modelling had the lowest R^2.

Additionally, the remaining range in miles was predicted and showed much better modelling results as depicted in Figure 7. The range of a new battery was a good predictor of the remaining range in miles. However, optional frequency variables demonstrated the highest correlation coefficients for linear regression. While remaining range in miles had mostly a normal distribution for its predicted results, remaining range in % had a distribution skewed towards an upper boundary of approximately 97% which demonstrates that majority of the Tesla drivers did not experience any significant battery degradation (see Figure 8).

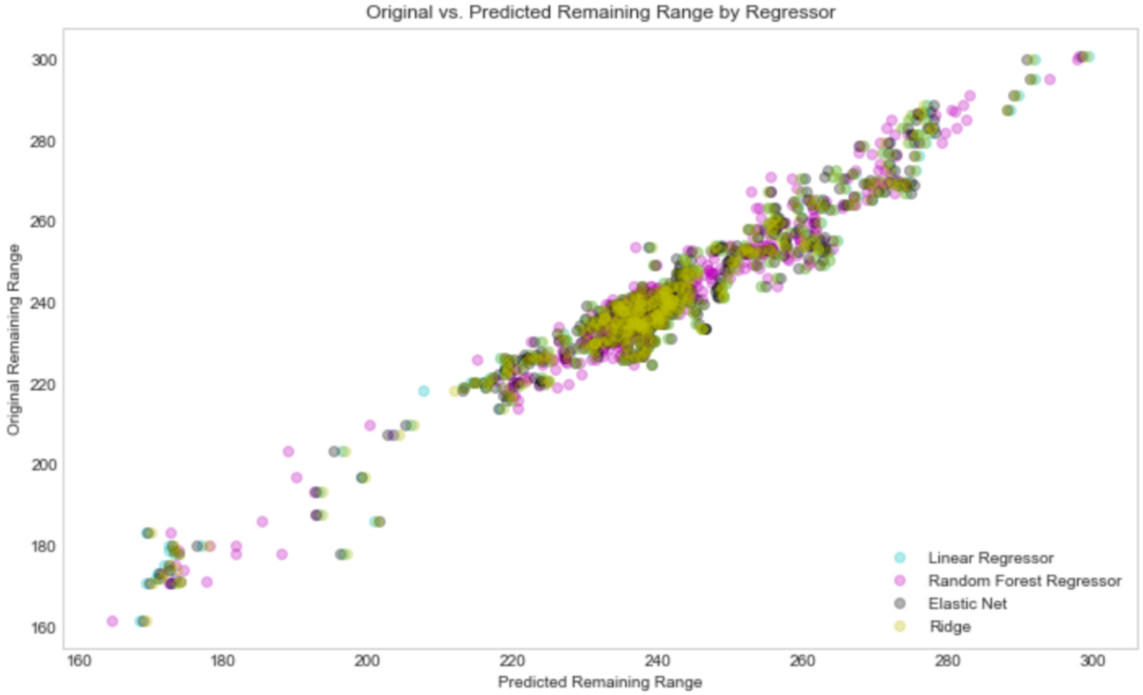
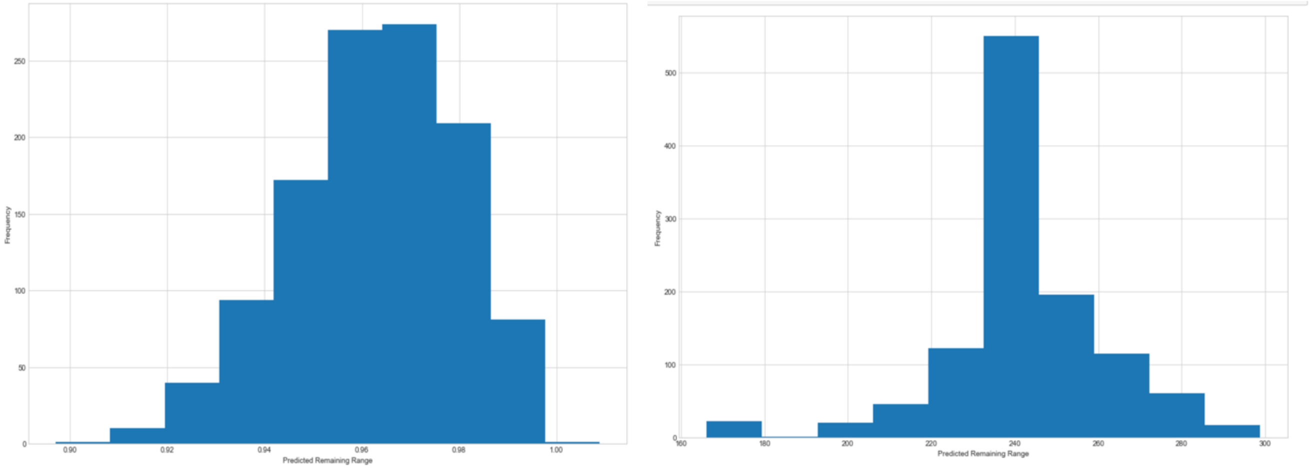


Figure 7: Regression results for predicting remaining range in miles.

Predicting the remaining range in miles resulted in obtaining the highest R^2 of 0.95 as compared to the highest R^2 of 0.5 for remaining range in %. We can conclude that to produce a more robust model for remaining range in terms of percentage of the original battery range, we need to obtain more variables including the average battery temperature and recommend the optional frequency columns to be a part of the main survey as they had the highest regression coefficients as compared to the rest of the variables used in prediction. Additionally, more observations are required to obtain a more consistent result in terms of statistical analysis. Optional categories (full charge frequency, empty charge frequency, and daily charge level) had a lot of frequency categories, but very little data for some of them which prompted us to unite them. This might have affected the resulting analysis of statistical significance as well as the imputation of the optional columns using logistic regression.



A) Remaining range in %

B) Remaining range in mi

Figure 8: Remaining range prediction distribution in different units.

After performing the regression analysis on Tesla Model S 85 observations separately, much more accurate results were obtained as shown in Table 7. The best prediction was calculated using Random Forest regression with the highest R^2 of 0.61 and the lowest root mean squared error of 0.016. This result was obtained using 438 observations for Model S 85. Therefore, it is recommended to build the predictive models for remaining range based on different car models as the type of model affects the predictive powers of the algorithm. Each model has different original ranges as well as performance efficiency (as denoted by P and D in the model name). Tesla models also range in horsepower ratings which also affects model performance demonstrating that model type cannot be equated to battery pack size. In the future analysis, more observations should be obtained for each model type to analyze regression more accurately.

Table 7: Regression results for predicting remaining range (in %) of Tesla Model S 85.

Regression	**R^2**	**RMSE**
Elastic Net	0.421	0.019
Ridge Regression	0.447	0.018
Linear Regression	0.453	0.018
Random Forest	0.61	0.016

Conclusion

In conclusion, we proved that remaining range is negatively correlated with three independent Tesla features - car mileage, battery age, and vehicle cycles. Additionally, we determined that location and supercharging frequency have no influence on remaining range while 100% charge frequency, empty charge frequency, and daily charge level have some effect on remaining range. We have investigated the pairwise relationships between different groups for 100% charge frequency, empty charge frequency, and daily charge level. It was determined that not charging the battery fully and never fully discharging it helped to decrease the rate of battery degradation. The developed regression model for Tesla Model S 85 was designed to predict the remaining car range in % of the original battery range. The best regression was achieved by using Random Forest which resulted in R^2 of 0.61 and a root-mean-squared-error of 0.016. Full charge frequency demonstrated the highest negative regression coefficient. It was also established that regression analysis was the most effective when performed separately for different models as model type affected the battery degradation results. It is recommended to obtain more observations for each model type and include battery charge characteristics as a part of the main survey for further analysis.