Inferential Statistics

Tesla Battery Degradation

The work described in this report is contained in the tesla-battery-degradation.ipynb notebook of the same repository.

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 a majority of data points - above 90% of remaining battery capacity. The graphical representation of these relationships is shown in Figure 1.

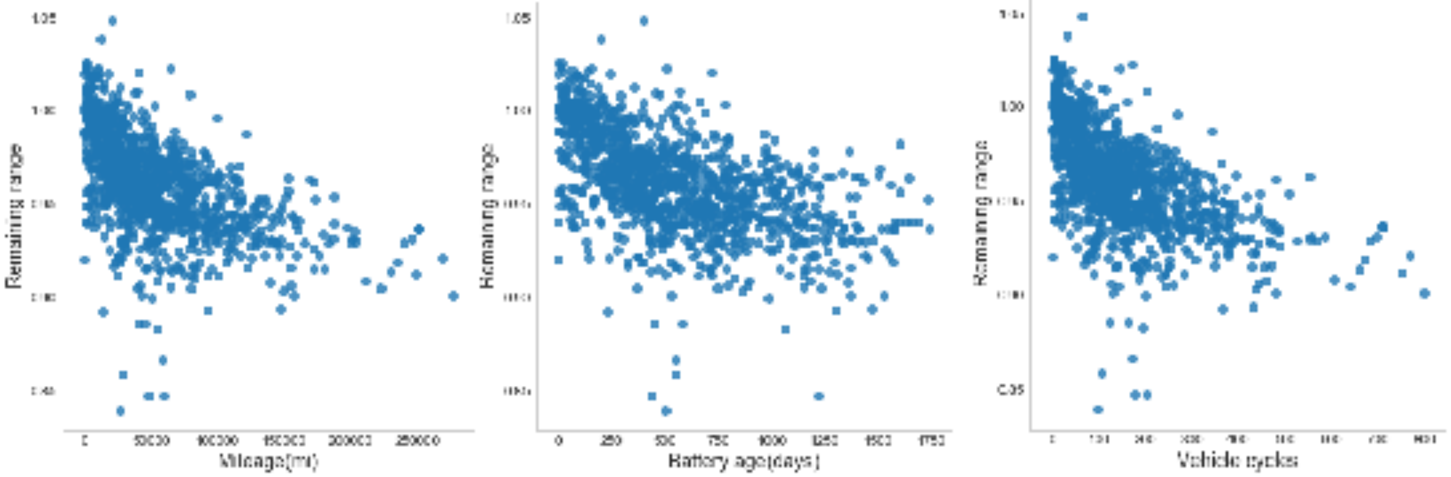

Figure 1: Relationships between the dependent variable (remaining range) and independent variables (mileage, battery age, and vehicle cycles).

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. 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.

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. 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.

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%.

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.

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.

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

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.

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 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.

Conclusion

In conclusion, using Pearson correlation coefficient, 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.

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