MDA 620 Capstone Project: Electric Vehicles Analysis

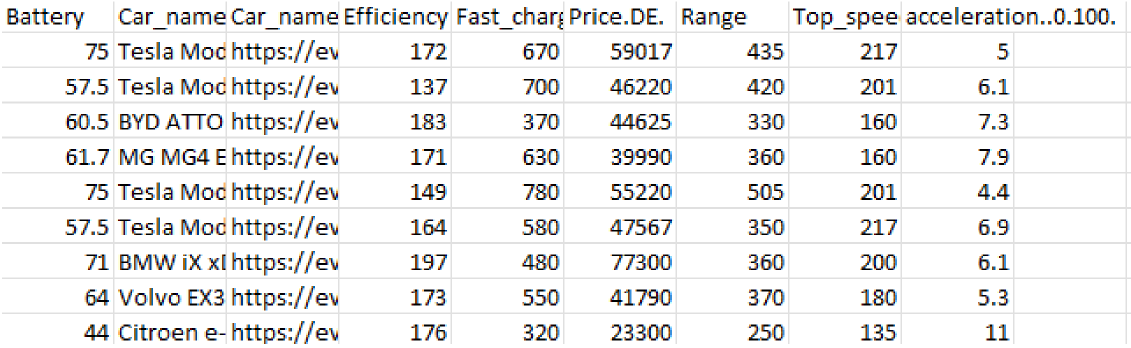
By:Eoin Gawronski

Table of Contents

- Dataset/Background
- Problem Scenario/Business Issue
- Objectives/Goals
- Data Exploration/Visualization
- Methodology/Model Building
- Model Selection
- Conclusion/Recommendations

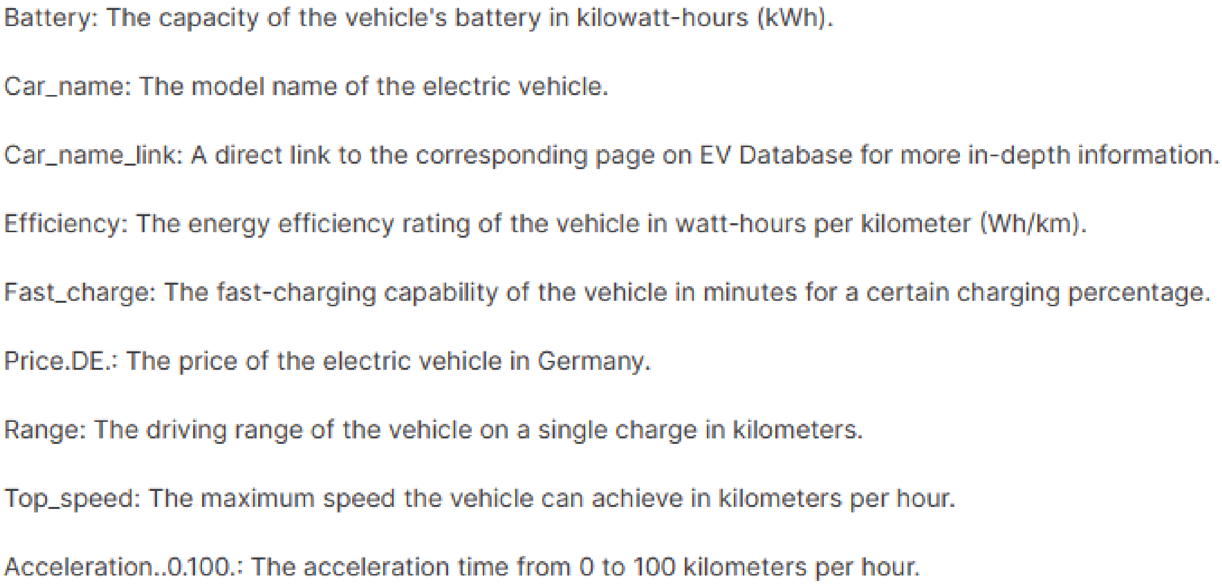
Dataset/Background

The dataset I chose to analyze was an electric vehicle dataset. This dataset was important to me because I wanted to learn more about electric vehicles as it's the future of cars. Ev's are relevant in the present, and it's important to understand pricing. The "Electric Vehicle Specifications and Prices" dataset, sourced from EV Database, offered detailed information on electric vehicles' specifications, pricing, and performance metrics. The dataset contained 360 observations and 9 different variables that helped me better understand prices and the effects of the features.



Battery	Car_name	Car_name	Efficiency	Fast_charg	Price.DE.	Range	Top_spee	acceleration..0.100.
75	Tesla Moc	https://ev	172	670	59017	435	217	5
57.5	Tesla Moc	https://ev	137	700	46220	420	201	6.1
60.5	BYD ATTO	https://ev	183	370	44625	330	160	7.3
61.7	MG MG4 E	https://ev	171	630	39990	360	160	7.9
75	Tesla Moc	https://ev	149	780	55220	505	201	4.4
57.5	Tesla Moc	https://ev	164	580	47567	350	217	6.9
71	BMW iX xI	https://ev	197	480	77300	360	200	6.1
64	Volvo EX3	https://ev	173	550	41790	370	180	5.3
44	Citroen e-	https://ev	176	320	23300	250	135	11

Below is a description of each column variable directly from the Kaggle dataset data field:



Battery: The capacity of the vehicle's battery in kilowatt-hours (kWh).

Car_name: The model name of the electric vehicle.

Car_name_link: A direct link to the corresponding page on EV Database for more in-depth information.

Efficiency: The energy efficiency rating of the vehicle in watt-hours per kilometer (Wh/km).

Fast_charge: The fast-charging capability of the vehicle in minutes for a certain charging percentage.

Price.DE.: The price of the electric vehicle in Germany.

Range: The driving range of the vehicle on a single charge in kilometers.

Top_speed: The maximum speed the vehicle can achieve in kilometers per hour.

Acceleration..0.100.: The acceleration time from 0 to 100 kilometers per hour.

Problem Scenario/Business Issue

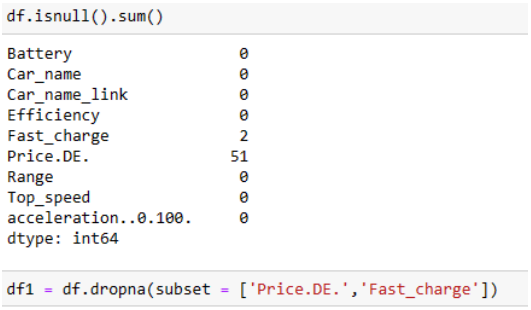
The main business issue within this analysis is trying to understand which features are the most important in determining prices of electric vehicles. It's important to understand which features affect the prices and fluctuation between prices. Ultimately, it's important to understand the relationship between car prices and its features to maximize market penetration and competitiveness.

Objectives/Goals

The main objectives of this project include looking at how prices of Electric vehicles are affected by range, top speed, efficiency, and other factors. The goal is to look further into the different Electric vehicles and look at the price trends based on different factors and features. The main question I'll try to answer is what features affect the prices of Electric vehicles the most. This analysis will be done through data exploration, data cleaning, and visualizations.

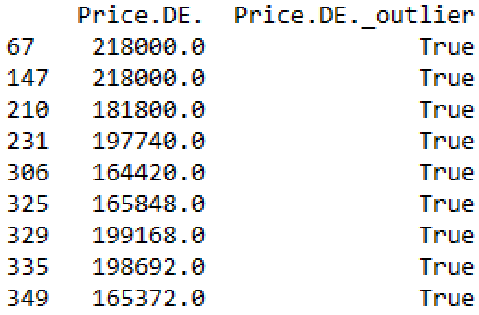
Data Exploration/Visualizations

Some basic data exploration I started with was reading the csv file of my dataset, and naming it df. This allowed me to convert my excel into a csv and make it accessible for data analysis. The first method applied was checking for missing values and dropping them to clean the data.



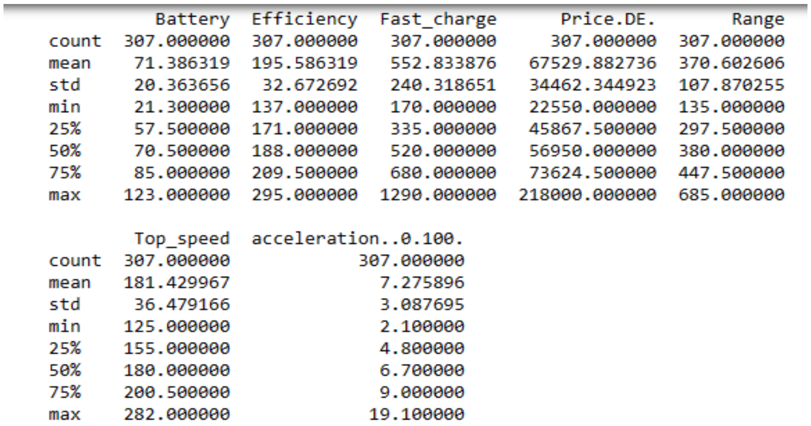
```
df.isnull().sum()

Battery                0
Car_name               0
Car_name_link          0
Efficiency             0
Fast_charge            2
Price.DE.             51
Range                  0
Top_speed              0
acceleration..0.100.   0
dtype: int64

df1 = df.dropna(subset = ['Price.DE.','Fast_charge'])
```

This was followed by checking the dataset for outliers with the set car price of 160,000. Any value greater than 160,000 was considered an outlier and was marked by the statement ‘TRUE’.



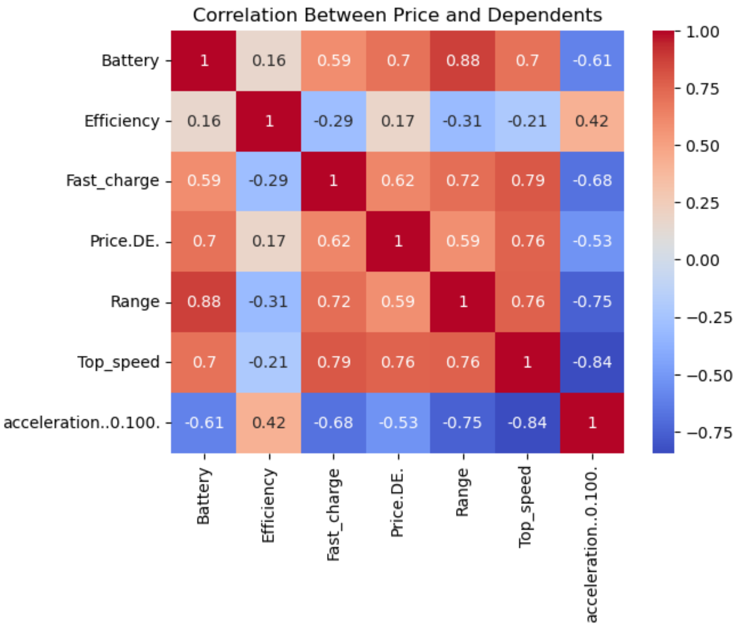
```
     Price.DE.  Price.DE._outlier
67    218000.0               True
147   218000.0               True
210   181800.0               True
231   197740.0               True
306   164420.0               True
325   165848.0               True
329   199168.0               True
335   198692.0               True
349   165372.0               True
```

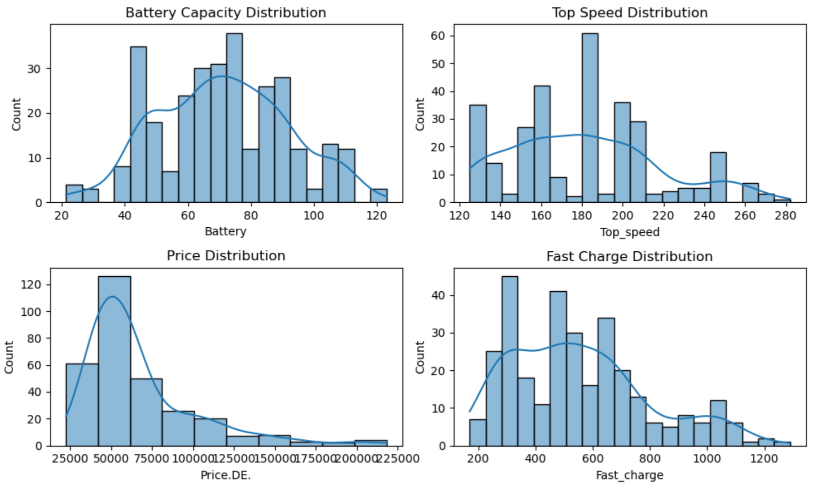
After cleaning the data, the describe and info function in python displayed basic statistics of each individual feature.



```
          Battery  Efficiency  Fast_charge      Price.DE.       Range
count  307.000000  307.000000   307.000000     307.000000  307.000000
mean    71.386319  195.586319   552.833876   67529.882736  370.602606
std     20.363656   32.672692   240.318651   34462.344923  107.870255
min     21.300000  137.000000   170.000000   22550.000000  135.000000
25%     57.500000  171.000000   335.000000   45867.500000  297.500000
50%     70.500000  188.000000   520.000000   56950.000000  380.000000
75%     85.000000  209.500000   680.000000   73624.500000  447.500000
max    123.000000  295.000000  1290.000000  218000.000000  685.000000

        Top_speed  acceleration..0.100.
count  307.000000            307.000000
mean   181.429967              7.275896
std     36.479166              3.087695
min    125.000000              2.100000
25%    155.000000              4.800000
50%    180.000000              6.700000
75%    200.500000              9.000000
max    282.000000             19.100000
```

These functions are important to begin analysis of a dataset. These allowed me to look deeper into analysis by looking at important visualizations.





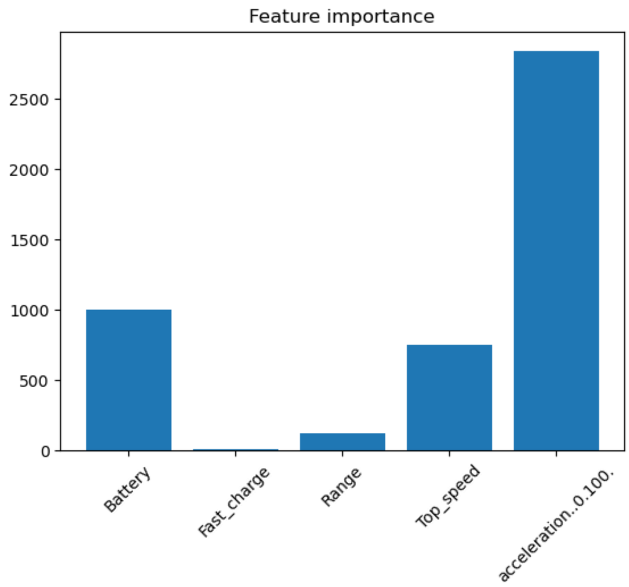
The left correlation matrix heatmap shows a clear image of the relationship between different variables regarding electric vehicles. Based on the heatmap, it's clear that there are high correlations between various features. The correlations between Price and its dependents are listed below:

- Price and Battery (0.7)
- Price and Fast Charge (0.72)
- Price and Range (0.59)
- Price and Top Speed (0.76)
- Price and Acceleration (-0.53)

The right plots show distribution plots that show each feature and its distribution, with a normal distribution curve. The Price Distribution plot is skewed to the right because the mean is around 50,000 which shows that there are some outliers towards the right end of the plot.

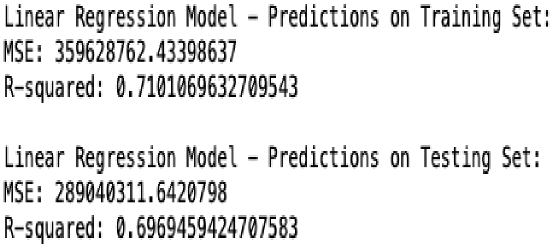
This shows a feature importance model which shows 5 features from the dataset.

This shows a lasso regression plot of the feature importance for the Electric Vehicles. The acceleration from 0-100 mph has the highest feature importance, followed by the Battery and Top speed that also have a bit of importance, and the fast charge feature has little to no importance.



Model Building

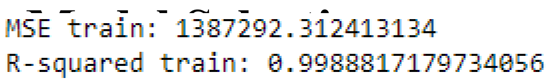
Two models used during this analysis were Linear Regression and a combination of Lasso/XG boost. The results for the linear regression are shown below:



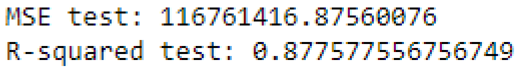
```
Linear Regression Model - Predictions on Training Set:
MSE: 359628762.43398637
R-squared: 0.7101069632709543

Linear Regression Model - Predictions on Testing Set:
MSE: 289040311.6420798
R-squared: 0.6969459424707583
```

These results show very high MSE for both and R-squared that are relatively high. These MSE's are very high potentially because of the wide range of car prices from 22,500 EU to 218,000 EU. This shows that the linear regression model may not be the best fit for this dataset. The Lasso/ XG boost model was used to improve predictions and reduce error. Lasso helps prevent overfitting in the dataset and XGboost can produce more accurate models. The MSE for the training and testing dataset were 138,792 and 116,761,416 respectively. Although these are still high, they are significantly less than MSEs of linear regression. This indicates more accurate predictions . Also, the R-squared for the training and testing dataset were

0.99 and 0.87 which is very high. In conclusion, improvement in the numbers indicates this model is a better fit for the data.

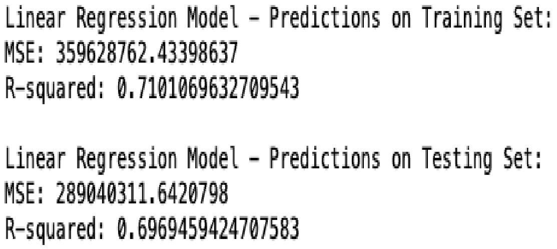


```
MSE train: 1387292.312413134
R-squared train: 0.9988817179734056
```

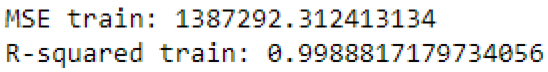


```
MSE test: 116761416.87560076
R-squared test: 0.877577556756749
```

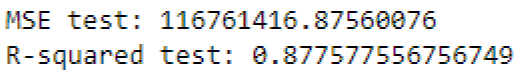
After building some models in python, the best model was the combination of the Lasso/XG model. This model showed improved R-Squared data and better MSE then the linear regression model. The small improvement of these numbers show that this model best fits the EV dataset. Below is a side by side of the two testing methods for a clearer image:



```
Linear Regression Model - Predictions on Training Set:
MSE: 359628762.43398637
R-squared: 0.7101069632709543

Linear Regression Model - Predictions on Testing Set:
MSE: 289040311.6420798
R-squared: 0.6969459424707583
```



```
MSE train: 1387292.312413134
R-squared train: 0.9988817179734056
```



```
MSE test: 116761416.87560076
R-squared test: 0.877577556756749
```

Tools and Techniques

- Data cleaning
  - addressed missing values and outliers in your dataset
- Data analysis
  - used statistical summaries and visualizations (like histograms and scatter plots) to understand the distribution and relationships of features
- Correlation Analysis
  - Correlation Matrix
- Regression Analysis
  - Implemented the Lasso/XGboost regression models

Conclusions/Recommendations

Some takeaways I had after completing the project was being able to see the best variables that affect price in Electric cars. Top speed and fast charge were the two most influential factors that affect EV prices alone. I was also able to analyze how other features affect each other not solely in comparison to price. This was shown by the high correlations between several factors, also known as multicollinearity which impacted the linear regression model. The use of more complex regression models showed that EV pricing is complex and could be due to multiple factors at once. The MSE is extremely high for both but is likely caused by the wide range in prices of EVs. In conclusion, Deeper analysis could be done to see these patterns, and also more sophisticated regression models/model refinement. My recommendation would be to implement a Feature-based pricing strategy and try to segment to a target group of consumers. The Feature-based pricing strategy is important because you can analyze which features affect price and choose features that are useful. By doing this, I think we should optimize key features influencing price and customer preference in EV models, consider tiered pricing models, and offer basic, mid-range, and premium models to cater to different market segments

References

- **Electric Vehicle Specifications and Prices (kaggle.com)**
- Excel File
- Jupyter Notebook