Computer Science Department

Mathematical Foundations of Analytics

CS660

CRN – 22921

Spring 2025

Project EDA

Submitted by:
Ran Roffe
Nipun Navadia

Purpose of Project

The goal of this project is to create a Notebook (Jupyter, Colab) and become familiar with the Pandas and Matplotlib/Seaborn/Plotly libraries.

Questions to be Answered:

1. Which car manufacturers are the most commonly used for EVs in Washington?
2. What are the highest and lowest electric ranges in this dataset, and which car makers and models do they correspond to?
3. Is the maximum electric range value unique? If not, which cars share this range?
4. Is the minimum electric range value unique? If not, which cars share this range?
5. How does the electric range vary between car makers and between models ?
6. Which are the top 5 cities adopting EVs?
7. How does the EV adoption rate vary among car makers over the years?
8. Is there a correlation between the electric range and the city of an EV?
9. Which county has the greatest variety of EV car models?

Details of Dataset

Electric Vehicle Population Data in Washington State, USA (Dataset)
Publisher: data.wa.gov
Maintainer: Department of Licensing
Source URL: https://data.wa.gov/api/views/f6w7-q2d2/rows.csv?accessType=DOWNLOAD
Category: Transportation
Public Access Leve: Public

This dataset shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department of Licensing (DOL).

This CSV file contains **200,048** rows, each representing an electric vehicle.

The dataset includes the following columns: VIN, County, City, State, Postal Code, Model Year, Make, Model, Electric Vehicle Type, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Electric Range, Base MSRP, Legislative District, DOL Vehicle ID, Vehicle Location, Electric Utility, and 2020 Census Tract.

Data Analysis Process and Methodology

As part of analysis we went through below steps:

1. Understanding the data Context:
    a. We analysed the requirements/problem statement related to end result.
    b. The source of the data was checked and thoroughly understood.
2. Import Libraries and Load data
    a. The required Libraries which we require throughout the project to achieve the goal have been imported at this step.
    b. We loaded the data using Pandas as data frame.
3. Initial Inspection
    a. We checked number of rows and columns (i.e. shape of the data frame).
    b. Used methods like info() and describe() to get an overview.
    c. We checked the Missing Values in the data frame.
4. Data Cleaning
    a. Handling missing data:
        i. We created a function to check the common rows having missing values with common columns and remove them.
        ii. We used KNN imputation to impute the data into data frame column “Legislative District”.
    b. Handling Outliers:
        i. We used box plots to visualise the outliers.
        ii. We checked for the outliers using zscore with threshold of 3
        iii. Handled the outliers:
            1. Still there are outliers, as the distribution have changed. Practices to be made for future research:
            2. 1. Replace with the mean/median of the same column - repeat the process until there are no more outliers.
            3. 2. Remove the outlier values entirely - won't necessarily help in this case.
            4. 3. Replace with a statistically related value - say we look for high correlation with the target variable, we can impute with a more relevant value instead of mean.
    c. Correct Data Types
    d. Handling Duplicates:
5. Univariate Analysis
    a. Reviewed measures of central tendency (mean, median) and dispersion (standard deviation, variance).
    b. Visualize Distributions: Use histograms, box plots, and bar charts to understand the distribution of individual variables.
6. Bivariate Analysis
    a. Checked correlation using Heatmap
    b. Crosstab analysis has been done to check relation between pair of categorial variables.
    c. Visualize Relationships:
        i. Scatter Plots: For numerical variables.
        ii. Box Plots and Violin Plots: To compare distributions across categories.
        iii. Heatmaps: To visualize correlations
7. Feature Engineering:
    a. New Features were created at this step:

    b. Showed Vehicle Locations on Map using Longitudes and Latitudes.
    c. We scaled and normalised the data at this point.
    d. Visualisation of the Normalised data has been done below are the finding:
    e. We performed One hot encoding.
8. PCA has been performed.
9. Key Insights documentation, next steps and further research.

Note: The project is available on Github. Please refer to below URL-
https://github.com/nipunnavadia/Fundamental-Mathematics-for-Data-Science-EDA-Project/tree/main/Project1

Key Insights

Below are the findings:

1. As per the analysis we found that the data was mostly filled and had almost0.3% missing data which shows the data collection was done in very organised manner.
2. Since the dataset is from Washington, we see the data from other states which needs to be removed in case we don't have future aims with the datasets to study out of the scope of Washington.
3. We see the columns like "Electric Range" and "Base MSRP" shows mostly 0's which interprets that they might be missing values.
4. We did not find any relations in numerical features.
5. As per the analysis of Outliers we see the data is not normally distributed.
6. Several numerical features had to be normalised.
7. Most of the numerical columns are skewed.
8. Vehicle locations were not in standard form but they were seen normally distributed.
9. Tesla has the highest electric vehicles in the Washington.
10. Most of the cars were seen from 2023.
11. Most of the cars are purely electric in comparison to hybrid cars.
12. Most of the data we have is from King county of Washington.

Future Work

We plan for the further analysis on the data set as mentioned below:

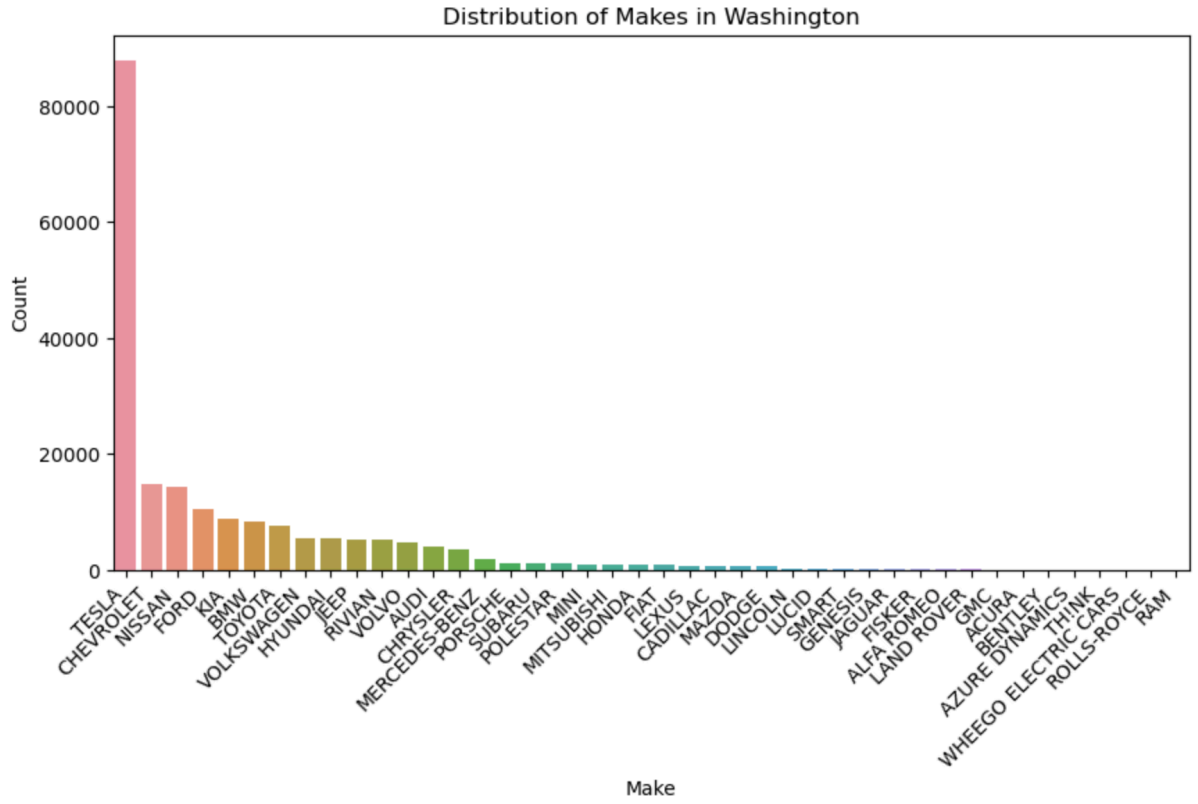
1. The data can be either specific for Washington considering the highly biased towards that or we need data from other states.
2. Data needs to be collected correctly since we see 0's in the crucial features like Base MSRP.
3. Further analysis needs to be done from location perspective to get reduce the number of features.
4. Target variable needs to be defined. As the data can be more meaningful with the target variable.
5. More features can be added with more specific context.
6. Groups (bins) of the features can be created based on the frequence of the categories.
7. Based on the desired aims (targeted experimental goal) we can further analyse the data: if the requirement need to drive the analysis towards the supervised analysis we can study the corelation with the dependent variable. However for unsupervised goals we can initiate the use of the model.
8. Next action will to split the data in train, test and validate sets.
9. Further analysis needs to be done to find strong correlations between the features.
10. Outliers of categorical features need to be defined with more context.

Question 1

Which car manufacturers are the most commonly used for EVs in Washington?

Answer

The most commonly used EV manufacturers in Washington are **Tesla, Nissan, and Chevrolet**, based on the highest number of registered EVs.



Questions 2

What are the highest and lowest electric ranges in this dataset, and which car makers and models do they correspond to?

Answer

```
<=====Minimum Electric Range: 0=====>

Make: TESLA
['MODEL 3' 'MODEL Y' 'MODEL X' 'MODEL S' 'CYBERTRUCK' 'ROADSTER']

Make: RIVIAN
['R1T' 'R1S' 'EDV']

Make: AUDI
['E-TRON' 'Q8' 'Q4' 'E-TRON GT' 'E-TRON SPORTBACK' 'RS E-TRON GT' 'SQ8']

Make: NISSAN
['LEAF' 'ARIYA']

Make: FORD
['MUSTANG MACH-E' 'F-150' 'TRANSIT']

Make: HYUNDAI
['IONIQ 5' 'KONA ELECTRIC' 'IONIQ 6' 'IONIQ' 'IONIQ 5 N']

Make: KIA
['EV6' 'EV9' 'NIRO' 'SOUL EV']

Make: CHEVROLET
['BOLT EUV' 'BOLT EV' 'BLAZER EV' 'SILVERADO EV' 'EQUINOX EV']

Make: BMW
['I4' 'IX' 'I5' 'I7' 'I3']

Make: MERCEDES-BENZ
['EQE-CLASS SUV' 'EQS-CLASS SUV' 'EQS-CLASS SEDAN' 'EQE-CLASS SEDAN'
 'EQB-CLASS' 'ESPRINTER']

Make: VOLKSWAGEN
['ID.4']

Make: CADILLAC
['LYRIQ']

Make: TOYOTA
['BZ4X']

Make: SUBARU
['SOLTERRA']

Make: VOLVO
['XC40' 'C40']

Make: POLESTAR
['PS2']

Make: MINI
```

```
['HARDTOP']

Make: JAGUAR
['I-PACE']

Make: PORSCHE
['TAYCAN']

Make: GENESIS
['GV60' 'GV70' 'G80']

Make: LUCID
['AIR']

Make: LEXUS
['RZ']

Make: FISKER
['OCEAN']

Make: HONDA
['PROLOGUE' 'CR-V']

Make: GMC
['HUMMER EV PICKUP']

Make: MAZDA
['MX-30']

Make: ACURA
['ZDX']

Make: FIAT
['500E']

Make: ROLLS-ROYCE
['SPECTRE']

Make: RAM
['PROMASTER 3500']
<=====Maximum Electric Range 337=====>

Make: TESLA
['MODEL S']
```

Question 3.

Is the maximum electric range value unique? If not, which cars share this range?

Answer

```
<=====Maximum Electric Range 337=====>

Make: TESLA
['MODEL S']
```

Question 4

Is the minimum electric range value unique? If not, which cars share this range?

Answer

```
<=====Minimum Electric Range: 0=====>

Make: TESLA
['MODEL 3' 'MODEL Y' 'MODEL X' 'MODEL S' 'CYBERTRUCK' 'ROADSTER']

Make: RIVIAN
['R1T' 'R1S' 'EDV']

Make: AUDI
['E-TRON' 'Q8' 'Q4' 'E-TRON GT' 'E-TRON SPORTBACK' 'RS E-TRON GT' 'SQ8']

Make: NISSAN
['LEAF' 'ARIYA']

Make: FORD
['MUSTANG MACH-E' 'F-150' 'TRANSIT']

Make: HYUNDAI
['IONIQ 5' 'KONA ELECTRIC' 'IONIQ 6' 'IONIQ' 'IONIQ 5 N']

Make: KIA
['EV6' 'EV9' 'NIRO' 'SOUL EV']

Make: CHEVROLET
['BOLT EUV' 'BOLT EV' 'BLAZER EV' 'SILVERADO EV' 'EQUINOX EV']

Make: BMW
['I4' 'IX' 'I5' 'I7' 'I3']

Make: MERCEDES-BENZ
['EQE-CLASS SUV' 'EQS-CLASS SUV' 'EQS-CLASS SEDAN' 'EQE-CLASS SEDAN'
 'EQB-CLASS' 'ESPRINTER']

Make: VOLKSWAGEN
['ID.4']

Make: CADILLAC
['LYRIQ']

Make: TOYOTA
['BZ4X']
```

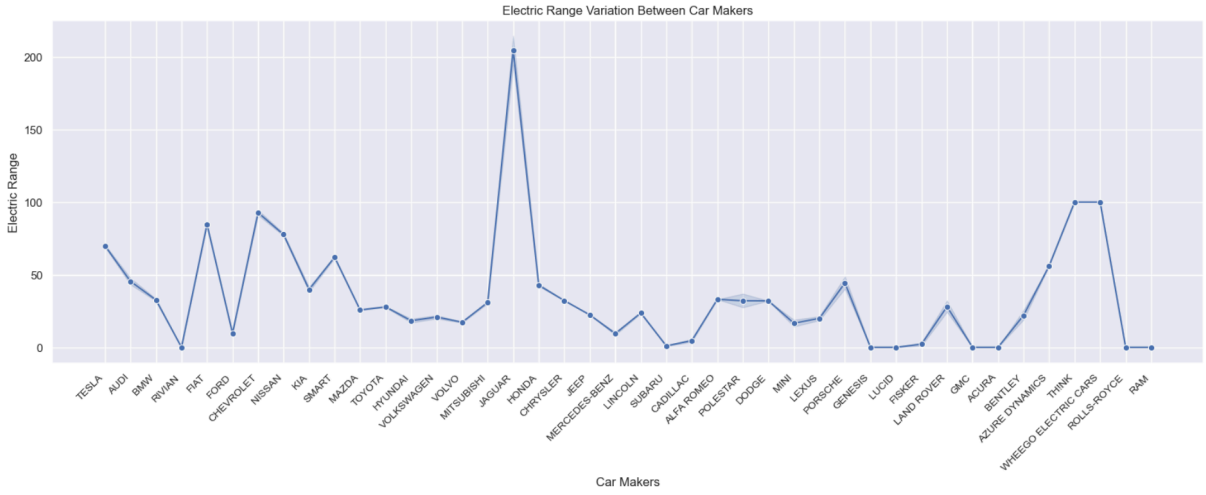
```
Make: SUBARU
['SOLTERRA']

Make: VOLVO
['XC40' 'C40']

Make: POLESTAR
['PS2']

Make: MINI
['HARDTOP']

Make: JAGUAR
['I-PACE']

Make: PORSCHE
['TAYCAN']

Make: GENESIS
['GV60' 'GV70' 'G80']

Make: LUCID
['AIR']

Make: LEXUS
['RZ']

Make: FISKER
['OCEAN']

Make: HONDA
['PROLOGUE' 'CR-V']

Make: GMC
['HUMMER EV PICKUP']

Make: MAZDA
['MX-30']

Make: ACURA
['ZDX']

Make: FIAT
['500E']

Make: ROLLS-ROYCE
['SPECTRE']

Make: RAM
['PROMASTER 3500']
```

Question 5

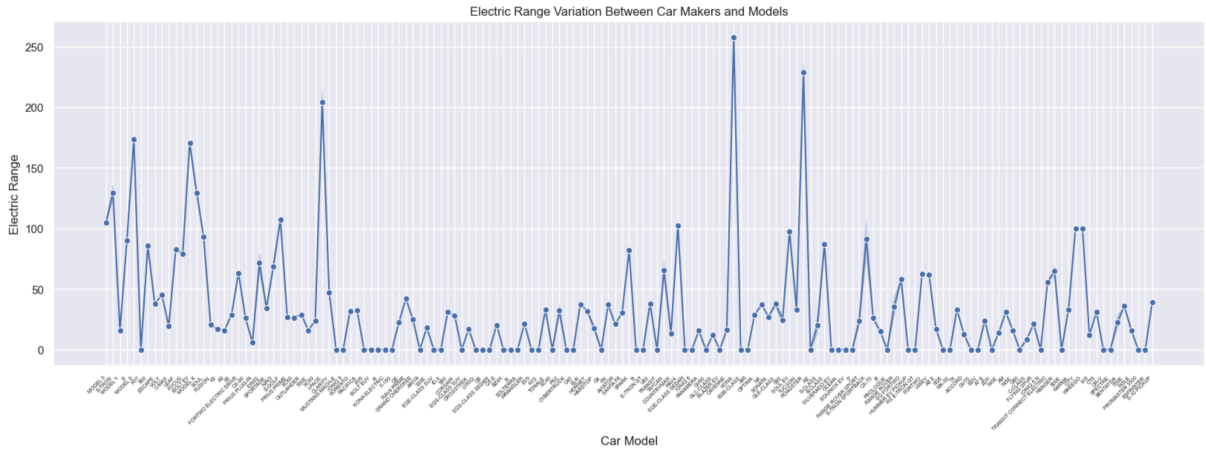
How does the electric range vary between car makers and between models ?

Answer

Map to showcase electric range vary between car makers.



Map to showcase electric range vary between car models



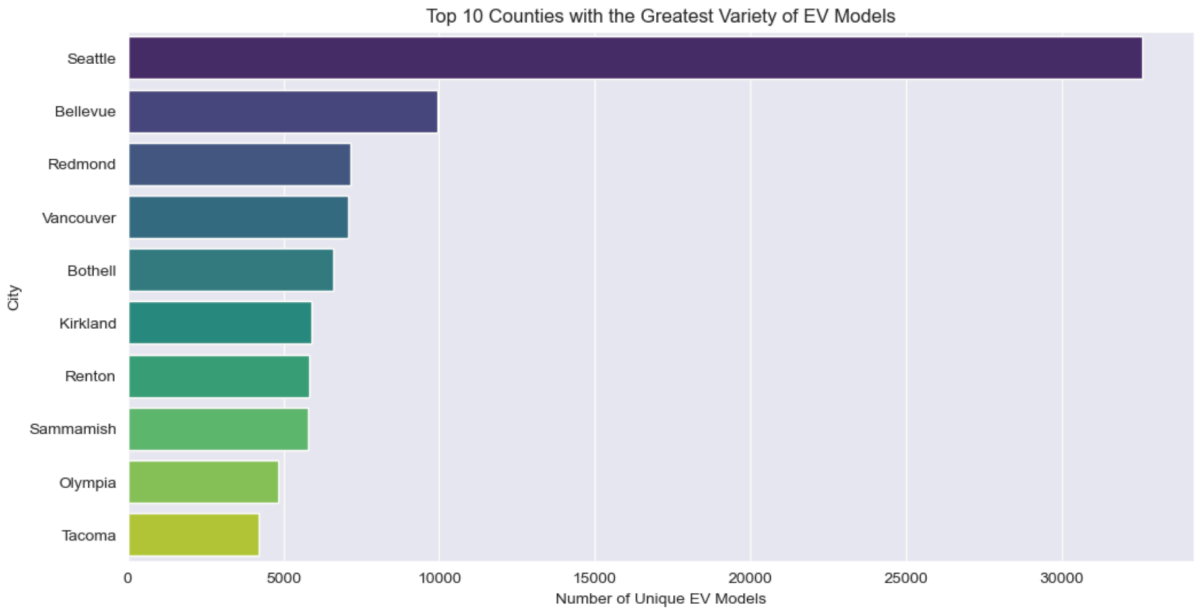
Question 6

Which are the top 5 cities adopting EVs?

Answer

Top 5 Counties with greatest variety of EV are:

1. Seattle
2. Bellevue
3. Redmond
4. Vancouver
5. Bothell



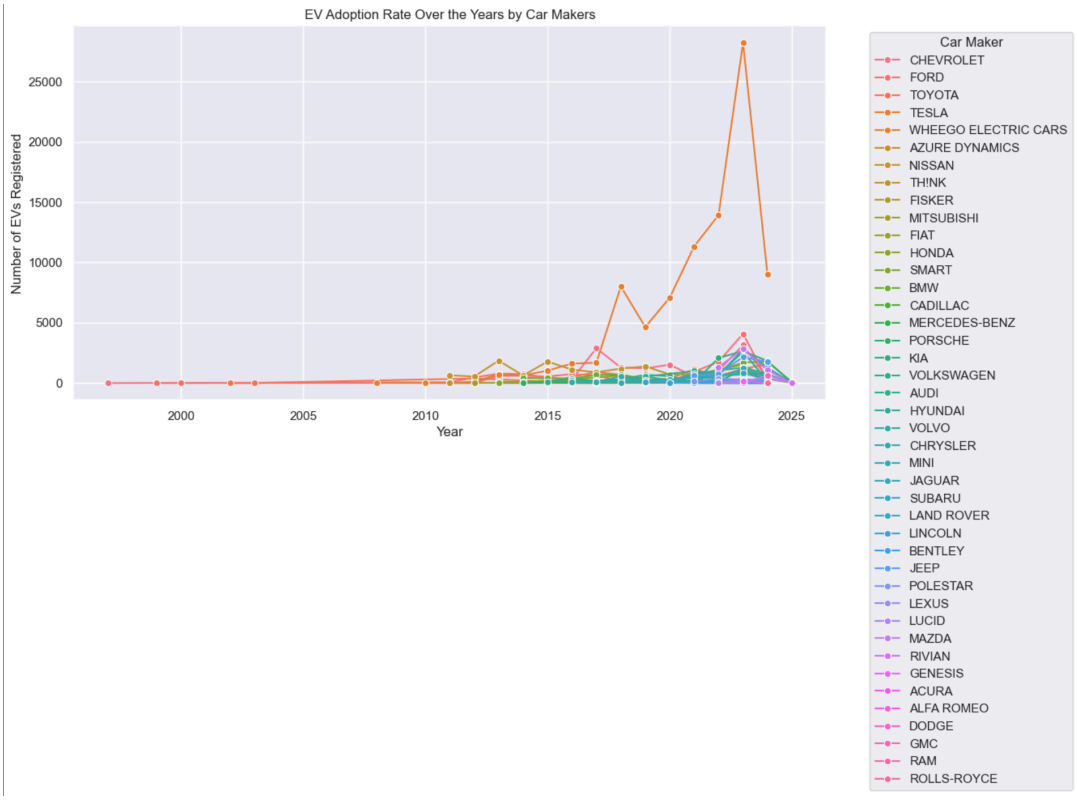
Question 7

How does the EV adoption rate vary among car makers over the years?

Answer

The EV adoption rate has varied across car manufacturers over the years.

1. **Tesla's Rapid Growth**: Tesla has experienced significant growth in EV registrations, particularly after 2015, peaking in the early 2020s before declining.
2. **Steady Increase for Other Brands**: Companies like Chevrolet, Nissan, and BMW have shown a more gradual rise in EV adoption compared to Tesla.
3. **Market Expansion**: The presence of numerous brands in the legend indicates that many manufacturers have entered the EV market, though their adoption rates remain lower than Tesla's.
4. **Post-2020 Surge**: A noticeable rise in EV registrations among several brands suggests growing industry-wide adoption.
5. **Fluctuations in Growth**: Some manufacturers display inconsistent adoption trends, possibly due to production challenges, changing policies, or shifts in consumer demand.



Question 8

Is there a correlation between the electric range and the city of an EV?

Answer

Here we are using ANOVA (Analysis of Variance) to answer this question.
ANOVA (Analysis of Variance) is a statistical test used to determine if there are significant differences between the means of two or more independent groups. It is commonly used when analyzing the relationship between a categorical independent variable and a numerical dependent variable.

How ANOVA Works
It compares the variance within each group to the variance between groups.
If the between-group variance is significantly larger than the within-group variance, it suggests that at least one group mean is different from the others.
ANOVA produces an F-statistic and a p-value:
F-statistic: Measures the ratio of between-group variance to within-group variance.
p-value: If $p < 0.05$, we reject the null hypothesis, meaning that at least one group mean is significantly different.

```
ANOVA Results:
F-statistic: 2.89, p-value: 0.0000
There is a statistically significant relationship between city and electric range.
```

Question 9

Which county has the greatest variety of EV car models?

Answer

King has the greatest EV car models.