Kaggle dataset "Car Features and MSRP" was compiled by Cooper Union. It combines vehicle features and prices data from The Edmunds Vehicle API with vehicle popularity data obtained using web scraping of Twitter based on the number of times a vehicle brand was mentioned in tweets. The dataset contains 11,915 car samples that were sold in the USA between 1990 to 2017 with features and prices of the cars.

The Edmunds Vehicle API is a proprietary dataset that is currently closed to public and is only available to limited strategic partners of Edmunds. It contains various vehicle specifications and prices of new and used cars. Car samples were obtained from Edmunds API by Cooper Union and their dataset that is used in our project is distributed freely on Kaggle. Vehicle popularity data was obtained by Cooper Union using streaming API that obtained current tweets related to each vehicle brand. The streaming API was run for 4 hours that obtained 54,000 tweets. Using this method, Cooper Union was able to rank popularity of each vehicle brand.

We have chosen this dataset to develop an algorithm to correctly predict price of a car. Furthermore, we are able to effectively analyze vehicle features in comparison to vehicle prices and their popularity. The effects of branding can be analyzed and see how the prices can change relative to the make of the car. It is therefore possible to see whether a car is overpriced relative to its features and find the degree of overpricing of a car using price projection algorithm.

Furthermore, most important features in a car can be analyzed using this dataset to see which features have the most effect on a vehicle prices and popularity of a vehicle. We can further outline the correlation amongst vehicle features and see how the correlation varies across vehicle brands. Furthermore, using a timeline analysis, the ever-changing features of the vehicles, as well as their gas consumption efficiency and horsepower can be analyzed through time and investigate how these changes have affected vehicle prices through time.

The dataset has 16 columns as follows:

# Make Vehicle Make

# Model Vehicle Model

# Year Vehicle Year (Marketing)

# Engine Fuel Type Engine Fuel Type (Unleaded, Diesel, Electric,

# Engine HP Engine Horse Power

# Engine Cylinders No. of Engine Cylinders (Usually 4 to Cylinders but would be 0 for electric cars

# Transmission Type Transmission Type (Automatic, Manual, Automated

Manual or Direct Drive)

# Driven\_Wheels Driven Wheels (Front Wheel Drive, Rear Wheel Drive,

Four Wheel Drive or All Wheel Drive)

# Number of Doors Number of Doors (2, 4 or 3 doors (Minivans))

# Market Category Market Category (Such as Luxury, Performance, Etc.)

# Vehicle Size Size of Vehicle (Compact, Midsize or Large)

# Vehicle Style Type of Vehicle (Such as Convertible, Coupe, etc.)

# highway MPG Highway Miles Per Gallon

# city mpg City Miles Per Gallon

# Popularity Popularity of the Vehicle (Per Twitter)

# MSRP Manufacturer Suggested Retail Price

**Further Explanation:**

***Miles Per Gallon:*** How far a vehicle can travel with one gallon of fuel.

***MSRP:*** Manufacturer's suggested retail price is the recommended price for the sale of a vehicle at the outlets. In the dataset MSRP of new cars sold in 2017 refers to the manufacturer’s suggested price, while for used vehicles produced before 2017, Edmunds API was used to obtain the market value.

The dataset has a high quality and low number of missing values as seen in the table below. Aside from “Market Category”, the rest of the missing values can be easily filled with a simple algorithm such as the mean value. As for “Market Category”, due to its large number, a new field of denoting their unknown category can be defined.

