Aditya Sumbaraju

Bellevue University

[DSC540-T302 Data Preparation](https://cyberactive.bellevue.edu/webapps/blackboard/execute/courseMain?course_id=_500220_1)

Professor Williams

Apr 11, 2021

**Identifying the Dataset**

We know there is a vital necessity to increase Car sales' productivity because of the anticipated need to lower selling price, as industrial systems are increasingly entering into and operating within a new, more competitive international market. We also recognize the everyday necessity to embrace automation through data analytics to boost productivity, which would be crucial to support transformative and sustainable change in the industry. Sometimes, data is scattered in different locations that need to be consolidated together to be effectively extracted. More effectively, data is stored in databases where it exists in relational forms, and different relations are connected by having similar relationships. However, data can also be joined together when it doesn't exist in a database on the columns that are common in the different datasets.

In this project, the datasets of interest were the car datasets which all are available at kaggle.com. Precisely three different datasets have been used in which the relationship between them exists by the car type, year, fuel type columns. It means that the dataset can be joined together by the standard columns and explored together. Due to the varying characteristics of the car's year, fuel type was more consistent and didn't have much variation and will be used to join the datasets.

In subsequent milestones of the project, I would be leveraging the option to pull the data from Kaggle API/ carroya.com and apply data wrangling techniques that I have learned throughout the course. And as a part of data visualization, I would be using matplotlib and ggplot2.

**Three data sources and its relationship:**

The first dataset was car price which is available at <https://www.kaggle.com/hellbuoy/car-price-prediction> and had 26 variables. The variables in the dataset were both continuous and categorical.

**Data Dictionary:**

|  |  |  |
| --- | --- | --- |
|  | **Variable** | **Description** |
| 1 | **Car\_ID** | Unique id of each observation (Interger) |
| 2 | **Symboling** | Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical) |
| 3 | **carCompany** | Name of car company (Categorical) |
| 4 | **fueltype** | Car fuel type i.e gas or diesel (Categorical) |
| 5 | **aspiration** | Aspiration used in a car (Categorical) |
| 6 | **doornumber** | Number of doors in a car (Categorical) |
| 7 | **carbody** | body of Car (Categorical) |
| 8 | **drivewheel** | type of drive wheel (Categorical) |
| 9 | **enginelocation** | Location of car engine (Categorical) |
| 10 | **wheelbase** | Weelbase of Car (Numeric) |
| 11 | **carlength** | Length of Car (Numeric) |
| 12 | **carwidth** | Width of Car (Numeric) |
| 13 | **carheight** | height of Car (Numeric) |
| 14 | **curbweight** | The weight of a car without occupants or baggage. (Numeric) |
| 15 | **enginetype** | Type of engine. (Categorical) |
| 16 | **cylindernumber** | cylinder placed in the Car (Categorical) |
| 17 | **enginesize** | Size of Car (Numeric) |
| 18 | **fuelsystem** | Fuel system of Car (Categorical) |
| 19 | **boreratio** | Boreratio of car (Numeric) |
| 20 | **stroke** | Stroke or volume inside the engine (Numeric) |
| 21 | **compressionratio** | compression ratio of Car (Numeric) |
| 22 | **horsepower** | Horsepower (Numeric) |
| 23 | **peakrpm** | car peak rpm (Numeric) |
| 24 | **citympg** | Mileage in city (Numeric) |
| 25 | **highwaympg** | Mileage on highway (Numeric) |
| 26 | **price(Dependent variable)** | Price of Car (Numeric) |

**Data Source 2: API Data Source:**

**Code snippet to pull dataset from Kaggle API:**

import kaggle

from kaggle.api.kaggle\_api\_extended import KaggleApi

api.dataset\_download\_file('avikasliwal/used-cars-price-prediction', file\_name='test-data.csv')

The second dataset was US car price which is available at <https://www.kaggle.com/avikasliwal/used-cars-price-prediction> and had 13 columns. Similarly, the dataset had both numerical and categorical variables.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Name | Car Make |
| Location | Car Location |
| Year | Car Make Year |
| Kilometers\_Driven | Total Mileage |
| Fuel\_Type | Categorizing Car w.r.t its fuel type |
| Transmission | Auto or manual |
| Owner\_Type | to identify Car is pre-owned or new |
| Mileage | Average Mileage |
| Engine | Engine Capacity |
| Power | Horse power |
| Seats | Number of seats |
| New\_Price | Car Price |

**Data Source 3:** **HTML source from carroya.com website**

URL: <https://www.carroya.com/buscar/vehiculos/t4e0.do#paginaActual=4>

I would be scraping the url to pull the used car details

|  |  |
| --- | --- |
| Variable | Description |
| Name | Car Brand |
| Year | Make Year |
| Mileage | Total Mileage |
| Price | Current car value |

To pull car basic info I would be using

HTML class details:

item.find\_all(**'div'**, attrs={**'class'**: **'car-ad-info'**}))

* Car name: find\_all(**"h2"**, attrs = {**"class"** : **"car-ad-name"**})
* year: find\_all(**"h3"**, attrs = {**"class"** : **"car-ad-year"**})
* price: find\_all(**"div"**, attrs = {**"class"** : **"car-ad-price"**})
* Mileage: find\_all(**"table"**, attrs = {**"class"** : **"used-specs-table"**})

**Data Relationship:** The grain of the dataset is identified as Car Make Year, Model and Fuel Type.