**US cars Data Analysis**

US Cars data was scraped from AUCTION EXPORT.com. This dataset included Information about 28 brands of clean and used vehicles for sale in US. Twelve features were assembled for each car in the dataset.

**Overview of dataset.**

The dataset contains 2499 observations and 12 Feature variables. Let’s take a further look at our dataset

|  |  |  |
| --- | --- | --- |
| Feature Variable | Data Type | Feature Description |
| Price | Integer | Sale price of vehicle in ad. |
| Year | Integer | Vehicle registration year |
| Brand | String | Car Brand |
| Model | String | Car Model |
| Color | String | Vehicle color |
| State/city | String | Location in which car is sold |
| Mileage | Float | Miles traveled by vehicle |
| VIN | Alphanumeric String | Vehicle identification number |
| Title Status | String | Describes whether the vehicle was clean vehicle or salvage insured |
| Lot | Integer | Lot number is combined with a serial number to form the Vehicle Identification Number. |
| Condition | String | Time left for the insurance |

**Exploratory data analysis**

On checking the dataset, it was found that there are no missing values.

**Variable treatment**

Now let’s take a look at how each variable is treated for the final analysis.

**Numerical variables**: - As checked the ‘Price’ and ‘mileage’ variables are already numerical variables also there are no missing variables hence there isn’t much that is needed to alter these variables.

**Categorical variables**: - we have a few categorical variables, we will see how we will treat each of them.

**Color**: - The color variable shows which color the vehicle is. On removing all the duplicate values we get around 49 unique values of color i.e. in our dataset there 49 kinds of colors. When checked further it was found that there were few colors which have a count of 1 which means that in whole data set there is only one single instance of such color. Further it was found that there were few colors which were the derivatives of the basic VIBGYOR, Thus the colors which were the derivatives of the Primary colors were clubbed into the Primary ones. For example glacier white is a derivative of white color, hence clubbed into white color, similarly ***ruby red*** and ***royal crimson metallic tinted clearcoat*** are derivatives of ***red color*** hence clubbed into **red color**.

Once this is done the again after removing the duplicate values we get 25 count of unique colors. The new set of colors has very less categories. Further this new set of colors was superimposed on the original color variable in order to get a new data with less variability.

**Condition**: - As per the original description of dataset in Kaggle the clear description is not present, however for the analysis it was taken as the time left for insurance. When we check it further we get that the time is either in “X hours left” or “X Days left”. There were few instances where it was “X minutes left”. Hence each of the observation is converted into the X minutes using excel CONVERT() Function.

**VIN**: - VEHICLE IDENTIFICATION NUMBER is an important variable which can provide us lots of insights, Let us try to understand what exactly VIN.

A vehicle identification number (VIN) is a unique code, including a serial number, used by the automotive industry to identify individual motor vehicles, towed vehicles, motorcycles, scooters and mopeds, as defined in ISO 3779 (content and structure) and ISO 4030 (location and attachment).

As per ISO 3779 Standards VIN has 3 Sections.

1. World manufacturer identifier
2. Vehicle descriptor section
3. Vehicle identifier section

**World manufacturer identifier**

The first three characters uniquely identify the manufacturer of the vehicle using the world manufacturer identifier or WMI code. The first three characters uniquely identify the manufacturer of the vehicle using the world manufacturer identifier or WMI code. The first character of the WMI is the region in which the manufacturer is located.

**Vehicle descriptor section**

The fourth to ninth positions in the VIN are the vehicle descriptor section or VDS. This is used, according to local regulations, to identify the vehicle type, and may include information on the automobile platform used, the model, and the body style. Each manufacturer has a unique system for using this field.

### Vehicle identifier section

The 10th to 17th positions are used as the 'vehicle identifier section' (VIS). This is used by the manufacturer to identify the individual vehicle in question. This may include information on options installed or engine and transmission choices, but often is a simple sequential number. For our analyses we will only take WIM section of the identification as the info from the other sections of the VIN isn’t available on public domain and usually differs from company to company.

**World manufacturer identifier:-**

We can use WMI to extract the manufacturing location of our vehicle. We used the below chart to extract the manufacturing location of our vehicles.

