Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Hyperiondev Copyright © 2018 Hyperion Development. All rights reserved.

Introduction

Auto-mobile data set consists of three types of entities:

1. The specification of a vehicle based on a number of various characteristics (make, fuel-type, aspiration, number-of-doors, body-style etc.).

2. Its insurance risk rating.

3. Its normalized losses in use as compared to other vehicle.

A risk factor is assigned on the vehicles linked to its price. +3 depicts that the vehicle is risky while 3 the auto is less risky. Lastly the average loss payment per insured vehicle year. This value is normalized for all vehicles within a particular size classification.

DATA CLEANING

The Dataset was loaded in a data frame for manipulation. Upon displaying the data it was clearly visible that the dataset contained special character in more than one of the columns in the data set, this would create problems for the pandas library as it does not know how to handle this type of character within its fields. The following methods and visualizations were used during data cleaning: (please refer to the jupyter notebook called **automobile.ipynb**

1. automobile = pd.read_csv('automobile.txt').
2. automobile.info().
3. automobile.head(5).
4. automobile.describe().
5. automobile.isnull().sum()
6. Replace inappropriate values.

The first Step was to read the data set and load it to a data frame, secondly get all the data set info (records, column names, number of entries , data types etc.) to better understand the data set and what values it contains.

Thirdly display the first five records contained in the dataset to see if there any missing or inconsistent values found and in which fields. Fourthly describe the data set to get the statistics of the data.

Last two steps I check the number of missing values per column, in the case of this particular dataset there was no null values only a special character was found in a few columns. I replaced the special character with appropriate values for each field in the dataset to ensure consistency of data within my dataset to ensure that I perform a realistic EDA.

Hyperiondev Copyright © 2018 Hyperion Development. All rights reserved.

MISSING DATA

Missing data was found is the following columns:

- normalized_losses
- num-of-doors
- bore
- stroke
- horsepower
- peak-rpm
- price

All fields that contained no number of doors on the vehicle were dropped, because it does not make sense for a vehicle not to have any number of doors and an average cannot be used.

All other fields an average of the column was computed and used to replace the missing values. This was easy to compute and will not affect the dataset greatly as opposed to dropping the fields all together.

DATA STORIES AND VISUALIZATIONS

Let us take a look at our dataset in-depth and try and visualize the meaning of our data through visualization. (Refer to jupyter notebook **automobile.ipynb**)

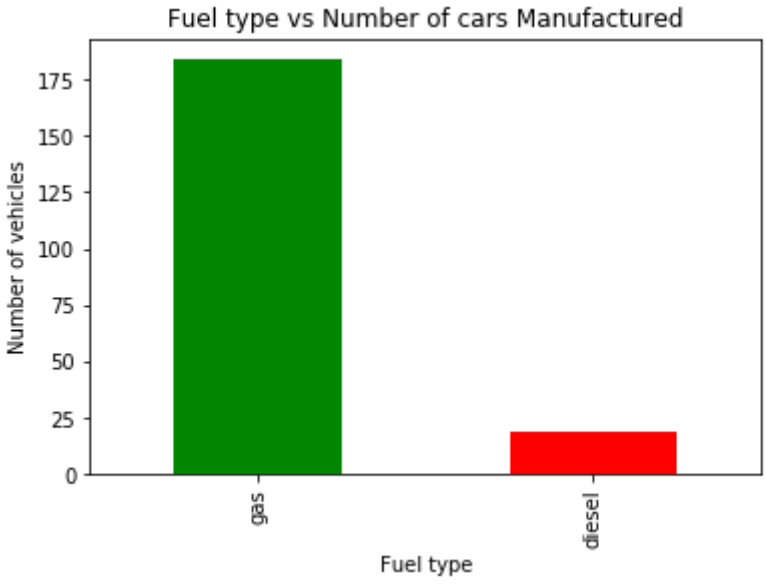

Figure 1. Fuel type versus Number of vehicles manufactured.

Hyperiondev
Copyright © 2018 Hyperion Development. All rights reserved.

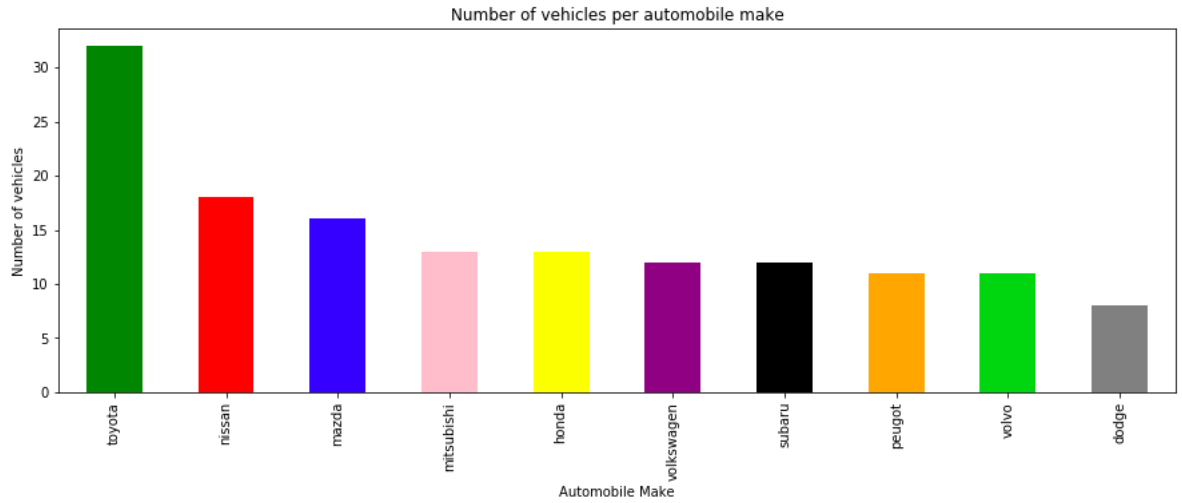

Figure 2. Vehicle make versus number of vehicles manufactured

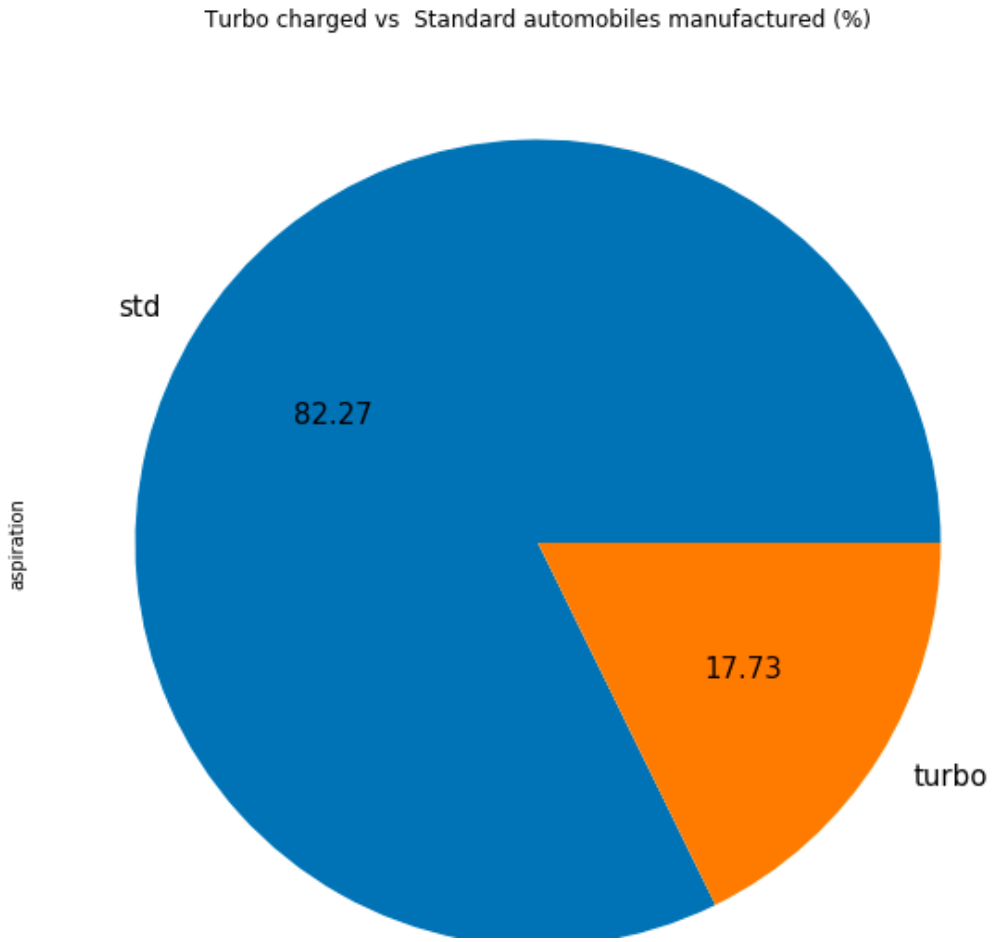

Figure 3. Turbo charged versus Naturally Aspirated engine vehicles produced

Hyperiondev
Copyright © 2018 Hyperion Development. All rights reserved.

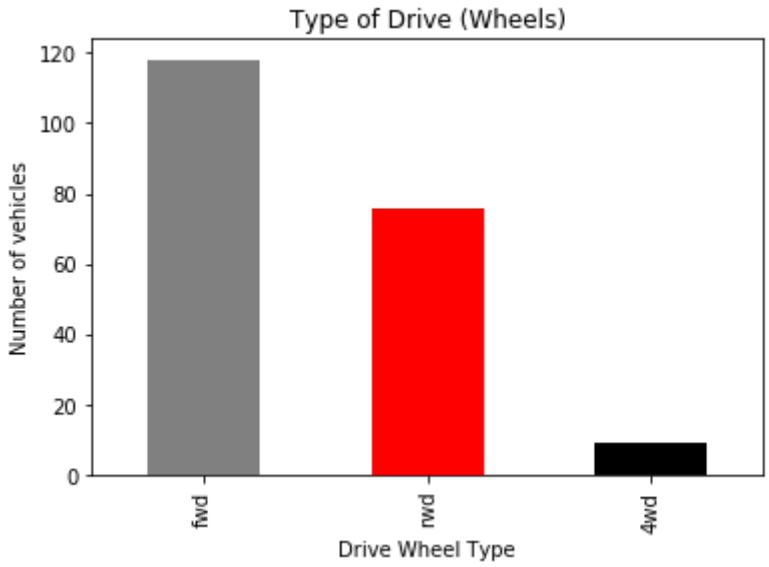

Figure 4. Type of Wheel drive versus number of vehicles produced

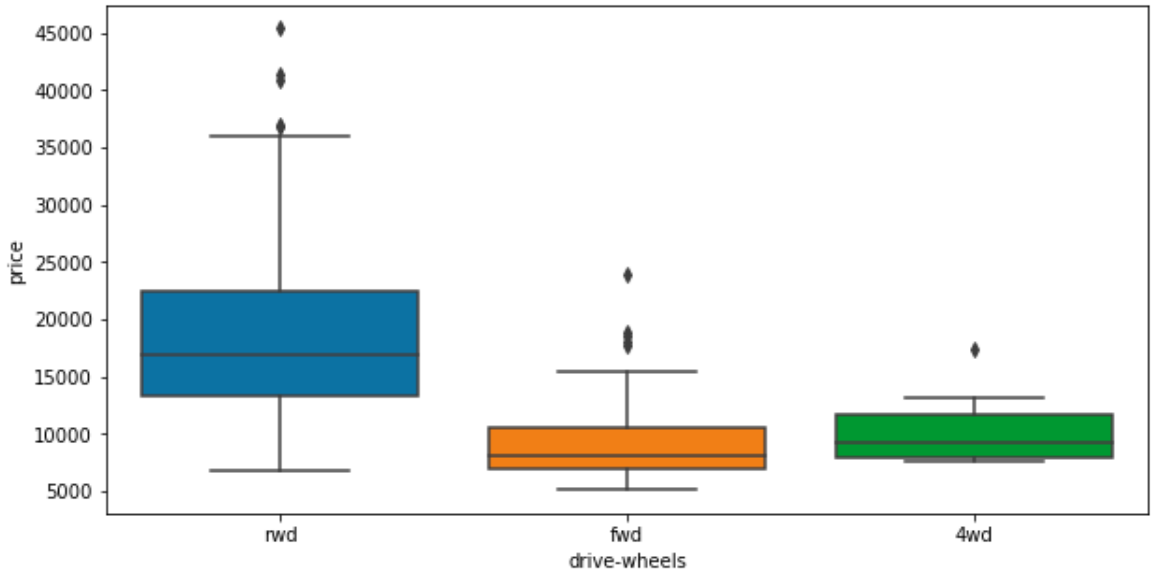

Figure 5. Drive wheel type versus Price of vehicles

Hyperiondev
Copyright © 2018 Hyperion Development. All rights reserved.

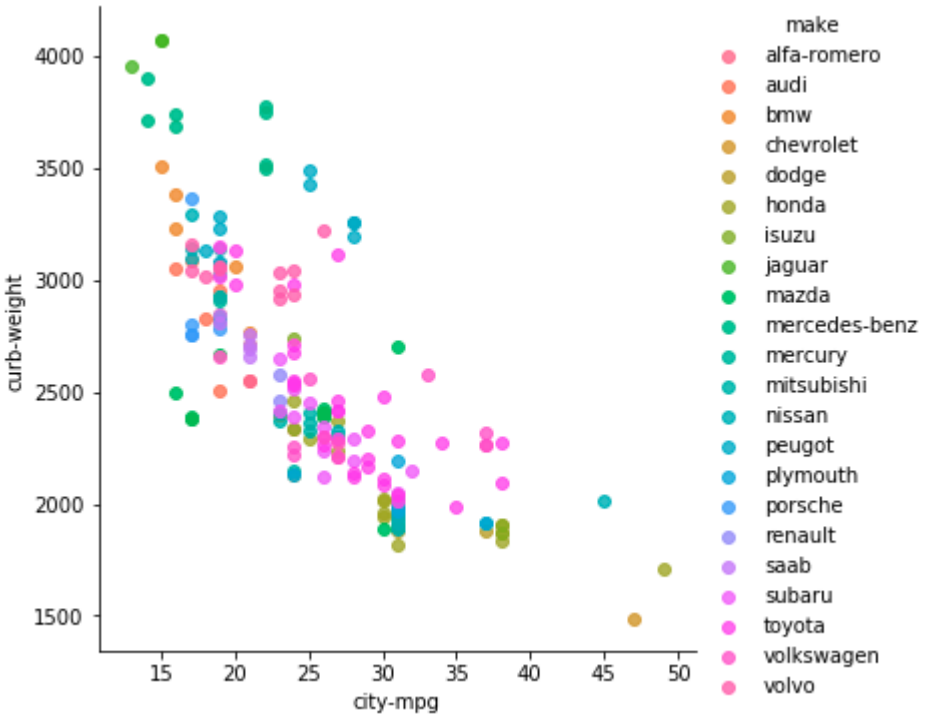

Figure 6. Weight of vehicle versus mileage in the city

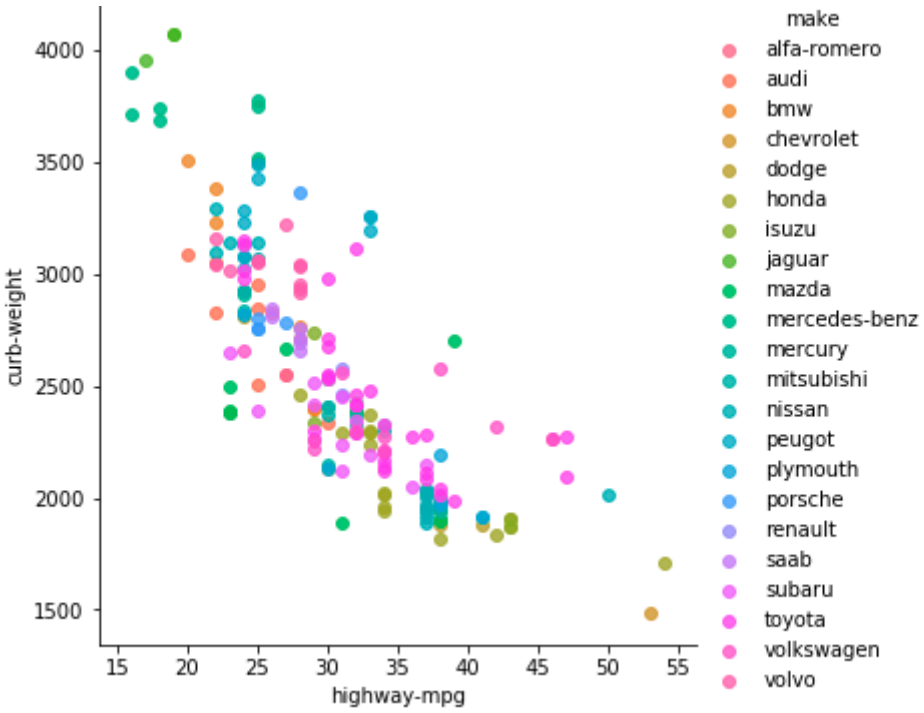

Figure 7. Weight of vehicle versus mileage on the highway

Hyperiondev
Copyright © 2018 Hyperion Development. All rights reserved.

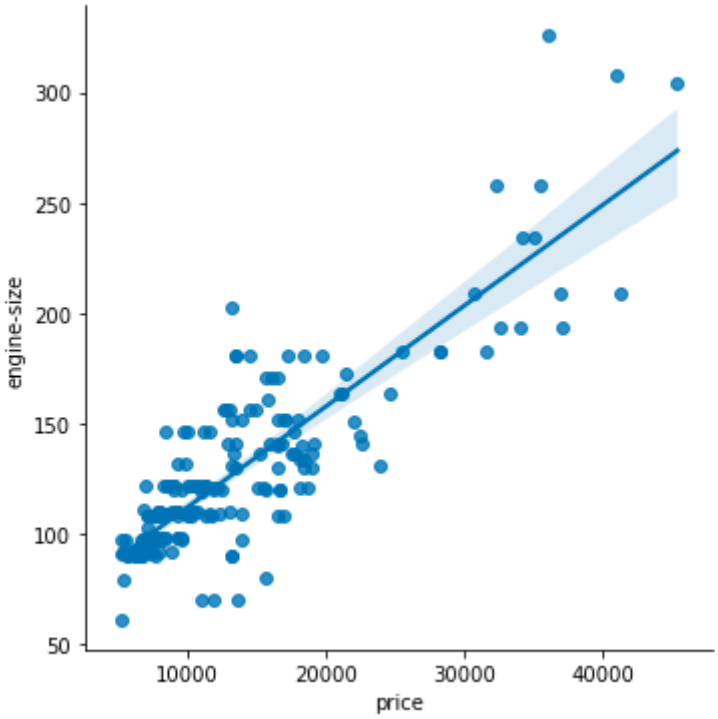

Figure 8. Price versus Engine Capacity/size

From the above different types of visualization and analysis using key features of the automobile dataset it is found that:

1. Most of the vehicles use gas rather than diesel
2. Front wheel drive has most number of vehicles followed by rear wheel drive and four wheel drive.
3. Most cars have naturally aspirated engines versus Turbo Engine
4. Toyota is the make of the car which has most number of vehicles with more than 40% than the 2nd highest Nissan
5. The bigger the engine size the more expensive the vehicle
6. Vehicles with less weight travel more in the city and highway this tells us that more and more people prefer to drive smaller cars.

CONCLUSION

Analysis of the dataset can be concluded to show the following:

- Importance of drive wheels and curb weight
- How the data set are distributed
- Correlation between different fields and how they are related
- Factors affecting Price of the Automobile.
- Mileage based on City and Highway driving for various make and attributes

THIS REPORT WAS WRITTEN BY: Molato Paul Sekgobela

Hyperiondev
Copyright © 2018 Hyperion Development. All rights reserved.