Machine-learning analysis for automobile dataset

Article · May 2020

CITATIONS
0

READS
4,039

1 author:

Ahmed Ali
University of the Cumberlands
44 PUBLICATIONS **2** CITATIONS

SEE PROFILE

Ahmed Ali

Ph.D. Student

University of Cumberlands

Automobile Data Analysis

Introduction

The automobile data analysis includes a dataset introduced from the Univerisity of California Irvine Machine Learning Repository UCI and refined from Kaggle. According to UCI (1985), the attributes consist of three different types of entities: (a) the model and specification of an auto, which includes the characteristics, (b) the personal insurance, (c) its normalized losses in use as compared to other cars. The data set source for this model collected from Insurance collision reports, personal insurance, and car models. According to Kaggle, There are 26 data attributes in this model descript the data set model from different angles. The objective of this report is to perform exploratory data analysis to find the primary relationships between features, which include univariate analysis, which includes finding the maximum and minimum, such as the weight, engine size, horsepower, and price.

Moreover, perform a regression to predict the car prices. The first step after determining the attributes is performing the data clean-up to find out the missing data and replace the unwanted data with context data; for example, the UCI indicates the missing attribute donated by "?" NAN will replace with zero. The data downloaded from the source website UCI. Figure 2 shows the data set attributes in which two columns the first one include 26 attributes, while the second columns include the attributes ranges.

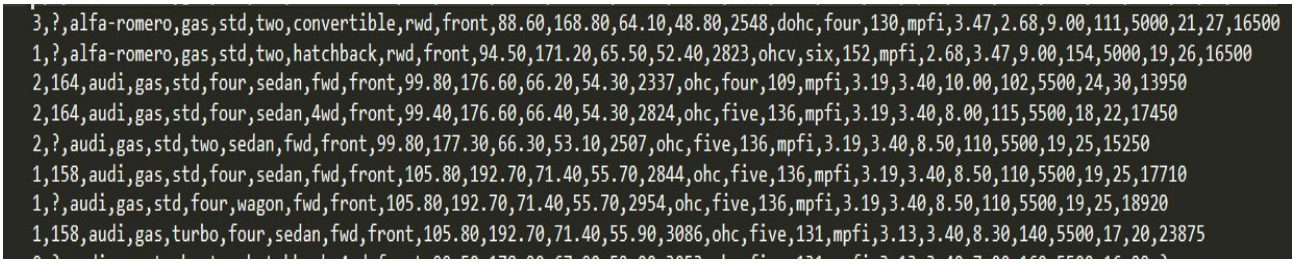

Figure 1 - Sample Orignal data from (https://archive.ics.uci.edu/ml/datasets/Automobile)

```
7. Attribute Information:
     Attribute:                Attribute Range:
     ----------                ------------------------------------------------
  1. symboling:                -3, -2, -1, 0, 1, 2, 3.
  2. normalized-losses:        continuous from 65 to 256.
  3. make:                     alfa-romero, audi, bmw, chevrolet, dodge, honda,
                               isuzu, jaguar, mazda, mercedes-benz, mercury,
                               mitsubishi, nissan, peugot, plymouth, porsche,
                               renault, saab, subaru, toyota, volkswagen, volvo
  4. fuel-type:                diesel, gas.
  5. aspiration:               std, turbo.
  6. num-of-doors:             four, two.
  7. body-style:               hardtop, wagon, sedan, hatchback, convertible.
  8. drive-wheels:             4wd, fwd, rwd.
  9. engine-location:          front, rear.
 10. wheel-base:               continuous from 86.6 120.9.
 11. length:                   continuous from 141.1 to 208.1.
 12. width:                    continuous from 60.3 to 72.3.
 13. height:                   continuous from 47.8 to 59.8.
 14. curb-weight:              continuous from 1488 to 4066.
 15. engine-type:              dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
 16. num-of-cylinders:         eight, five, four, six, three, twelve, two.
 17. engine-size:              continuous from 61 to 326.
 18. fuel-system:              1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
 19. bore:                     continuous from 2.54 to 3.94.
 20. stroke:                   continuous from 2.07 to 4.17.
 21. compression-ratio:        continuous from 7 to 23.
 22. horsepower:               continuous from 48 to 288.
 23. peak-rpm:                 continuous from 4150 to 6600.
 24. city-mpg:                 continuous from 13 to 49.
 25. highway-mpg:              continuous from 16 to 54.
 26. price:                    continuous from 5118 to 45400.

8. Missing Attribute Values: (denoted by "?")
   Attribute #:   Number of instances missing a value:
   2.             41
   6.             2
   19.            4
   20.            4
   22.            2
   23.            2
   26.            4
```

Figure2 Orignal data attaributes from (https://archive.ics.uci.edu/ml/datasets/Automobile)

The data preparation includes constructing one excel sheet to combine the two data set from Figure 1 and Figure 1; the header row includes the 26 columns attributes and the rest of rows imported from the file in Figure 1. Figure 3 shows a snapshot of the final excel sheet after importing the data.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
2	?	alfa-romei	gas	std	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130
3	?	alfa-romei	gas	std	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130
4	?	alfa-romei	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv	six	152
5	164	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc	four	109
6	164	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc	five	136
7	?	audi	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc	five	136
8	158	audi	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc	five	136
9	?	audi	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc	five	136
10	158	audi	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc	five	131
11	?	audi	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52	3053	ohc	five	131
12	192	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108
13	192	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108
14	188	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710	ohc	six	164
15	188	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765	ohc	six	164
16	?	bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3055	ohc	six	164
17	?	bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3230	ohc	six	209
18	?	bmw	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380	ohc	six	209

Figure 3 Data set after importing

Exploratory Data Analysis

In this section, we will start by importing the data into the google Colab exploratory data analysis to find the primary relationship between attributes. The next step is to clean the unwanted data such as "?" with NAN value if it is a character or mean value

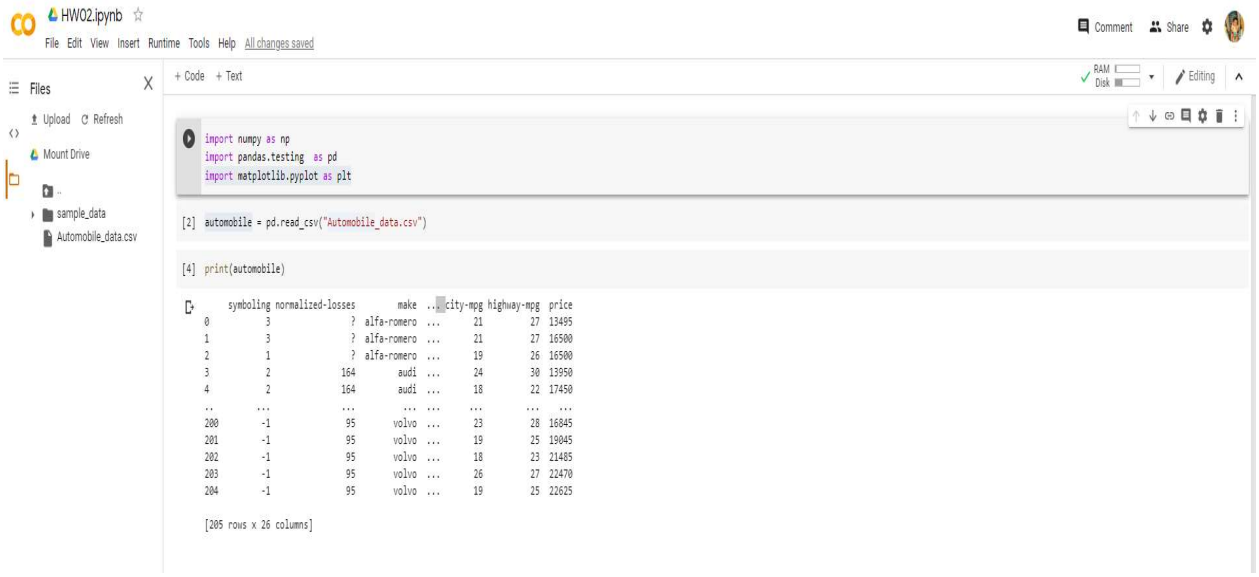

Figure 4 data set imported from the Automobile datasheet

The next step fills the missing numerical data with the mean this done by calculating the mean for every numerical column after excluding the nan-numerical data and then replace the missing data with the mean; figure 4 show snap code of this steps, this steps will be repeated for columns (price, curb-weight, engine-size, horsepower,peak-rpm)

```
temp = automobile[automobile['horsepower']!='?']
horsepower_mean = temp['horsepower'].astype(int).mean()
automobile['horsepower'] = automobile['horsepower'].replace('?',horsepower_mean).astype(int)
```

Figure 4 – Replace the missing data with the mean

```
temp = automobile[automobile['horsepower']!='?']
horsepower_mean = temp['horsepower'].astype(int).mean()
automobile['horsepower'] = automobile['horsepower'].replace('?',horsepower_mean).astype(int)

temp = automobile[automobile['peak-rpm']!='?']
peak_rpm_mean = temp['peak-rpm'].astype(int).mean()
automobile['peak-rpm'] = automobile['peak-rpm'].replace('?',peak_rpm_mean).astype(int)

temp = automobile[automobile['bore']!='?']
bore_mean = temp['bore'].astype(float).mean()
automobile['bore'] = automobile['bore'].replace('?',bore_mean).astype(float)

temp = automobile[automobile['price']!='?']
price_mean = temp['price'].astype(int).mean()
automobile['price'] = automobile['price'].replace('?',price_mean).astype(int)

temp = automobile[automobile['stroke']!='?']
stroke_mean = temp['stroke'].astype(float).mean()
automobile['stroke'] = automobile['stroke'].replace('?',stroke_mean).astype(float)

automobile[['horsepower','price',  'peak-rpm','curb-weight','engine-size']].hist(figsize=(13,10),bins=8,color='Y')
plt.figure(figsize=(13,10))
plt.tight_layout()
plt.show()
```

Figure 5 – Replace the missing data with the mean for performing the Univariate analysis

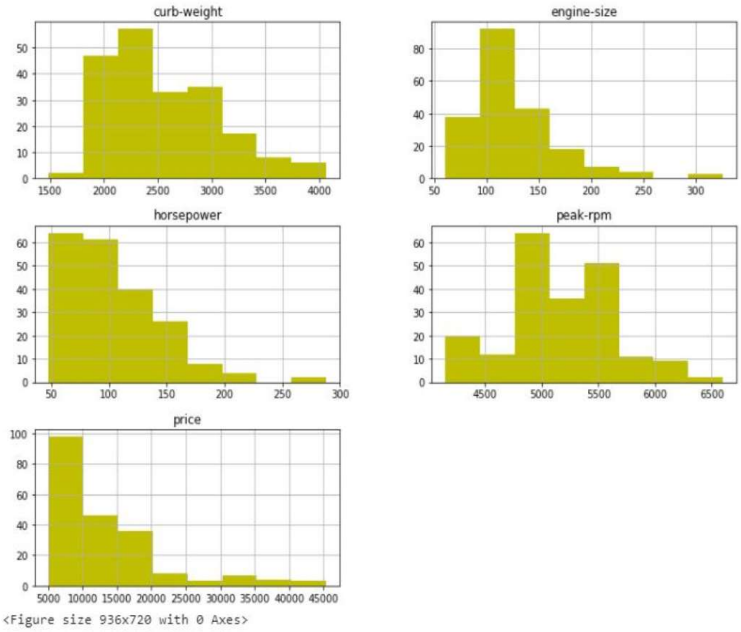

Figure 6 – Univariate analysis

The univariate analysis shows that most car price between 5000 – 8000, for the engine size is in range 60 to 170. Most of the car has a Curb Weight in range 1800 to 3100. Also, the graph shows the peak rpm distributed in the range between 4600 to 5100. The most vehicle has horsepower 50 to 120.

Figure 6 shows the correlation between the different attributes, the heat map ranging from zero to one, and it also goes under zero. For example, car prices highly correlated with engine size and horsepower; another observation is the highway-mpg and city-mpg negatively correlated if they are higher in the price.

Figure 7 shows the linear regression model relation between the price and engine-size variables by using a machine learning library and splitting the dataset into two groups, the training data with 75% and test data 25%. The linear regression shows that the price tends to go hight when the engine size becomes higher (Mediasittich, 2019).

```
import seaborn as sns
corr = automobile.corr()
plt.figure(figsize=(20,9))
a = sns.heatmap(corr, annot=True, fmt='.2f')
```

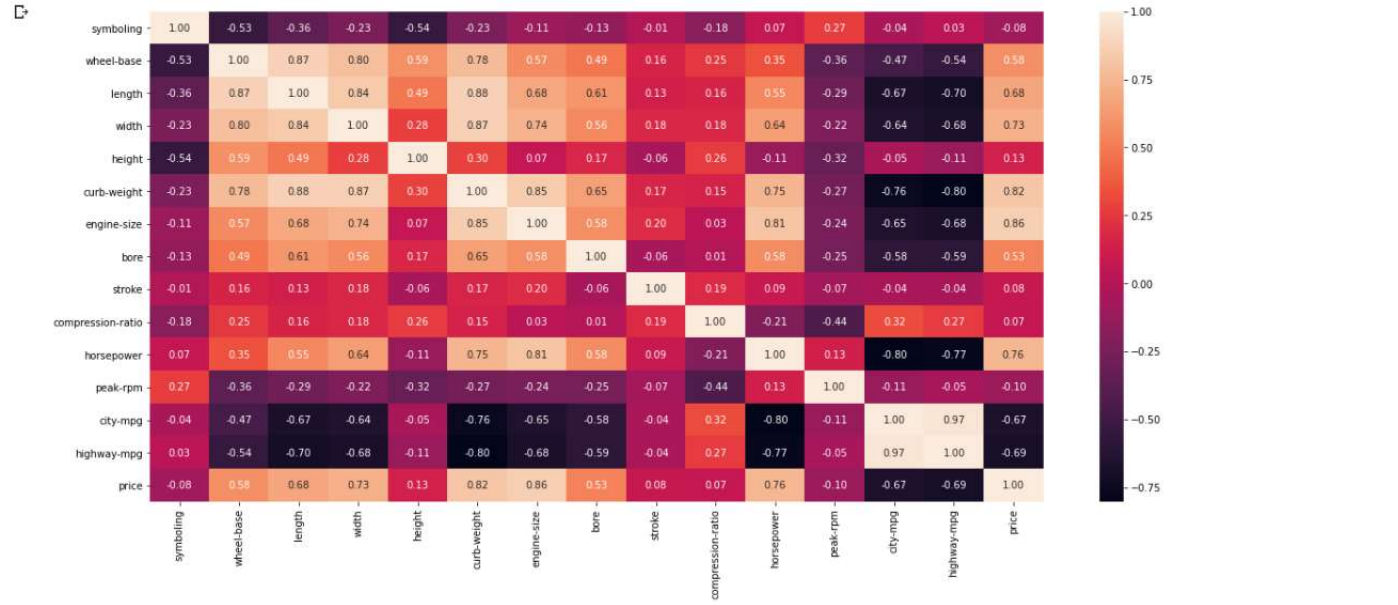

Figure 6 – Heat Maps

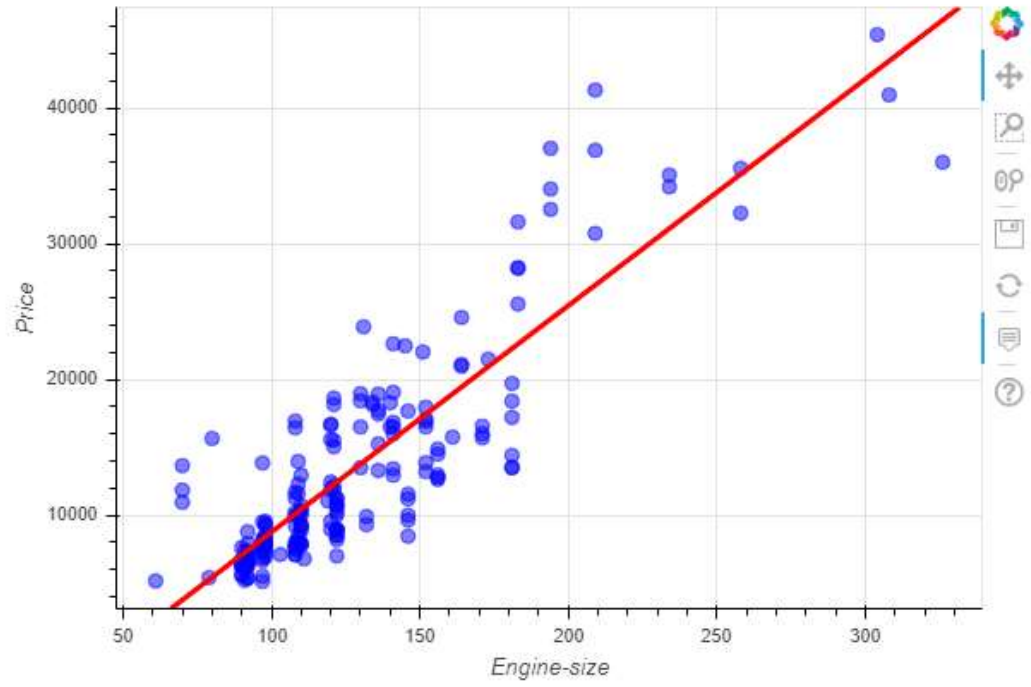

Figure 7 – linear regression

Figure 8– linear regression

References

Mediasittich. (2019). Linear regression for car price prediction. Retrieved from https://www.kaggle.com/mediasittich/linear-regression-for-car-price-prediction

Automobile Data Set. (n.d.). Retrieved from https://archive.ics.uci.edu/ml/datasets/Automobile

Shubhamsinghgharsele. (2018). Exploratory Data Analysis on Automobile Dataset. Retrieved from https://www.kaggle.com/shubhamsinghgharsele/exploratory-data-analysis-on-automobile-dataset