Publish

Amnanazim Follow

Apr 8, 2020 · 4 min read · Listen

Exploratory Analysis

Linear Regression and Correlation

The data consist of multiple features of a vehicle varing from it's length to it's horsepower to it's body style etc. It consist of both categorical and numerical features. The use of the data can be done in two ways

1. Analysis
2. Prediction

In analysis we can use the help of statistics to find relevant insights from the dataset however in prediction , we can use the price column as Y and other column as X['factors affecting price'] to generate price prediction based on its features

Linear Regression is one of the most common used method for analyzing data , it involves best fit line i.e $y=mx+b$ tells the relationship between dependent and independent variable .

Linear regression model is used for Qualitative data (Numerical)

Most of the machine learning algorithm based on Linear Regression Model , used to predict data

Here I will share my piece of code on the Automobile Dataset.

The format of file is csv (comma seperated file)

The file consist of(201 , 29) i.e 201 rows and 29 columns

```
df.shape
```

```
(201, 29)
```

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

1. Import data on Jupyter Notebook using pandas

```
path = "https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DA0101EN/autom
df = pd.read_csv(path)
df.head(5)
```

	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	...	compression-ratio	horsep
0	3	122	alfa-romero	std	two	convertible	rwd	front	88.6	0.811148	...	9.0	
1	3	122	alfa-romero	std	two	convertible	rwd	front	88.6	0.811148	...	9.0	
2	1	122	alfa-romero	std	two	hatchback	rwd	front	94.5	0.822681	...	9.0	
3	2	164	audi	std	four	sedan	fwd	front	99.8	0.848630	...	10.0	
4	2	164	audi	std	four	sedan	4wd	front	99.4	0.848630	...	8.0	

5 rows × 29 columns

The file has 29 columns some of which are numerical and other are categorical in nature .

2. Seperate categorical and numerical columns

```
df_num = df[["wheel-base","length","width","height","curb-weight","engine-size","bore","stroke","compre
df_num.head()
```

	wheel-base	length	width	height	curb-weight	engine-size	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg
0	88.6	0.811148	0.890278	48.8	2548	130	3.47	2.68	9.0	111.0	5000.0	21	27
1	88.6	0.811148	0.890278	48.8	2548	130	3.47	2.68	9.0	111.0	5000.0	21	27
2	94.5	0.822681	0.909722	52.4	2823	152	2.68	3.47	9.0	154.0	5000.0	19	26
3	99.8	0.848630	0.919444	54.3	2337	109	3.19	3.40	10.0	102.0	5500.0	24	30
4	99.4	0.848630	0.922222	54.3	2824	136	3.19	3.40	8.0	115.0	5500.0	18	22

13 columns are numeric in nature

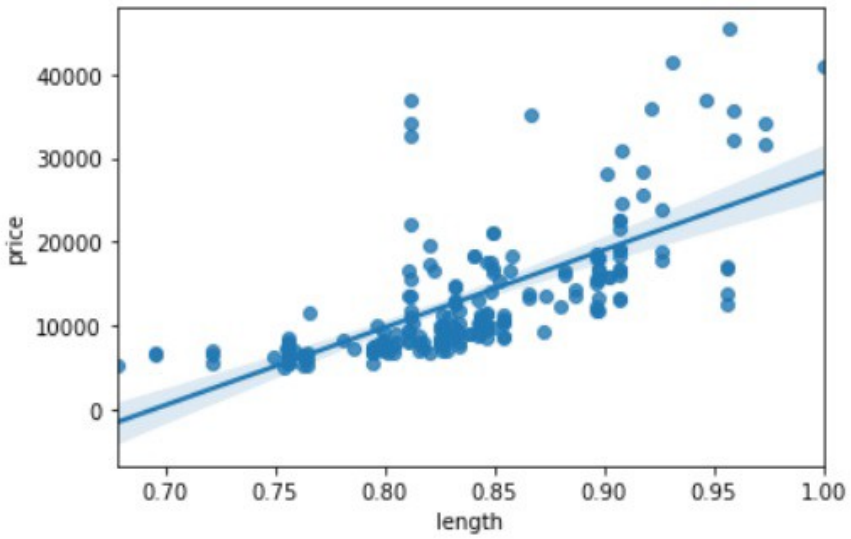
3. Using seaborn library , scatter plot for all the qualitative data. This helps us to quickly understand the trend and factors affecting the price of automobile

dependent

There are mainly three **types of Correlation**

- **Positive correlation** shows strong relationship between 2 variables

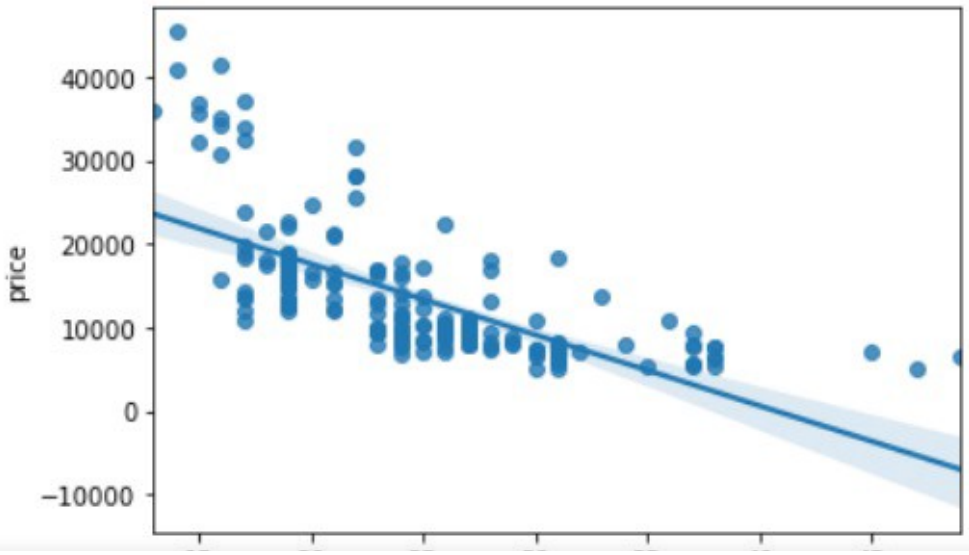
```
In [17]: length = sns.regplot(x = df_num['length'], y='price' , data=df)
```



Price increases with the increase in length

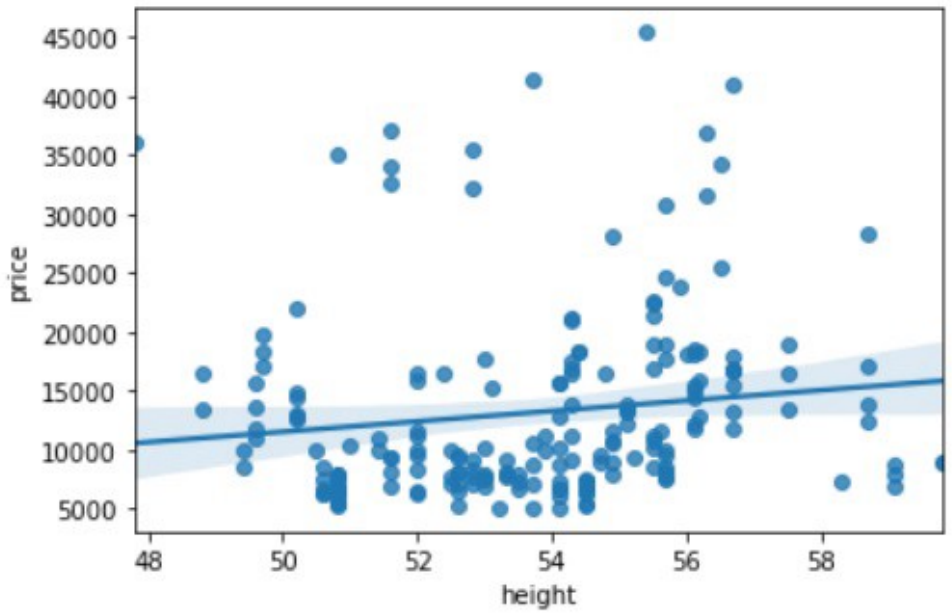
- **Negative Correlation** shows weak relationship between 2 variables

```
city_mpg =sns.regplot(x = df_num['city-mpg'] , y='price' , data=df)
```



- **No Cor**

```
height = sns.regplot(x=df_num['height'] , y='price' , data=df)
```

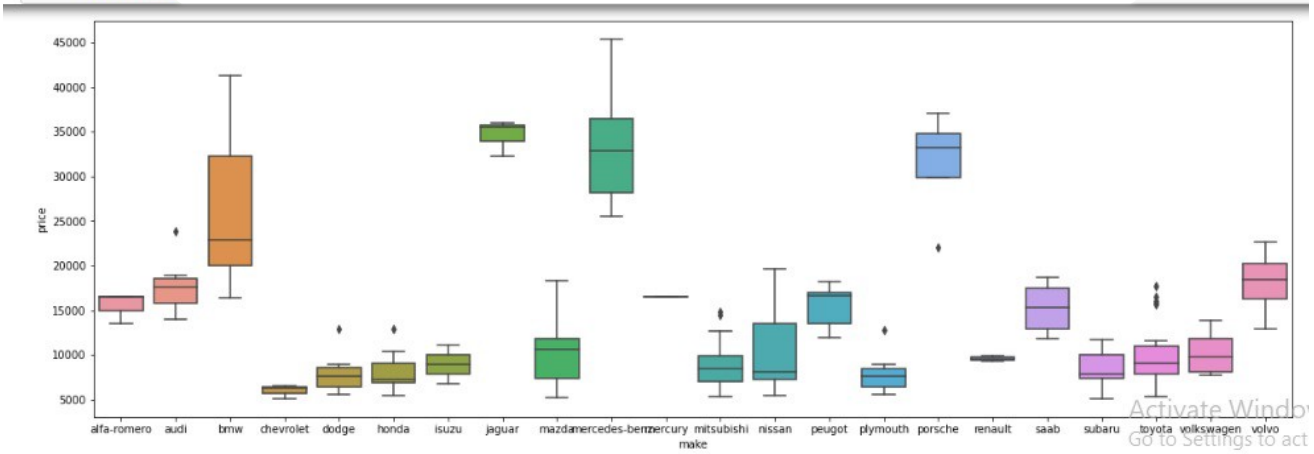


Price doesnot have much increase with the increase in height

There Values ranges from -1 to +1 . Values from +1 (shows strong positive relationship) to -1 (shows strong negative relationship). However values near or close to 0 shows no relationship

4. For categorical column , I have used Box plot and concluded that Drive-wheel and Engine-Location are good predictor of price

```
sns.boxp
count +=
```



For all the Numerical columns , I have used Pearson correlation technique

Pearson correlation measures the strength and direction of a linear relationship between two variables. Pearson correlation test is also known as parametric correlation test because it depends to the distribution of the data.

```
from scipy import stats
```

```
df_rel = df[["wheel-base","length","width","curb-weight","engine-size","bore","horsepower","city-mpg","highway-mpg"]]
```

```
for i in range(len(df_rel.columns)):
    pearson_coef,p_value = stats.pearsonr(df_rel[df_rel.columns[i]],df['price'])
    print("Pearson Co-efficeient for {}: \t\t".format(df_rel.columns[i]),pearson_coef," and P-value : ",p_value)
```

```
Pearson Co-efficeient for wheel-base:           0.584641822265508  and P-value :  8.076488270733218e-20
Pearson Co-efficeient for length:               0.6906283804483639  and P-value :  8.016477466159328e-30
Pearson Co-efficeient for width:                0.7512653440522672  and P-value :  9.20033551048217e-38
Pearson Co-efficeient for curb-weight:          0.8344145257702843  and P-value :  2.189577238894065e-53
Pearson Co-efficeient for engine-size:          0.8723351674455182  and P-value :  9.265491622200232e-64
Pearson Co-efficeient for bore:                 0.5431553832626604  and P-value :  8.049189483935032e-17
Pearson Co-efficeient for horsepower:           0.809574567003656  and P-value :  6.369057428259557e-48
Pearson Co-efficeient for city-mpg:             -0.6865710067844678  and P-value :  2.321132065567641e-29
Pearson Co-efficeient for highway-mpg:          -0.7046922650589529  and P-value :  1.7495471144477352e-31
```

To interpret the result based on the above calculations:

From the above result we can conclude that :

- Wheel-base P-value is statistically significant but linear relationship is moderate
- Length P-value is statistically significant but linear relationship is moderate
- Width P-value is statistically significant h… …p is strong
- Curb-weight P-value is statistically signi… …tionship is strong
- Engine-size P-value is statistically significant but linear relationship is strong
- Bore P-value is statistically significant but linear relationship is moderate
- Horsepower P-value is statistically significant but linear relationship is strong