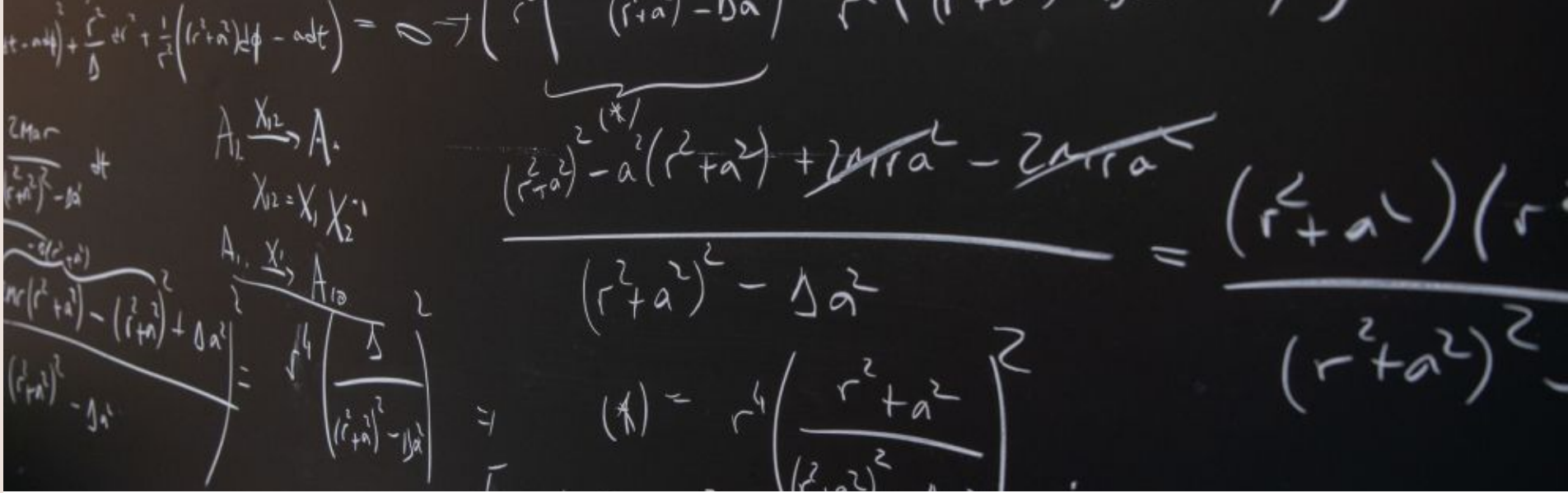
WHAT FUELS THE PRICE?

What factors contribute to the price of a used car?

TABLE OF CONTENTS



ABOUT ME

from Physics to Data Science

Bs in Physics

Started Teaching

2010

Teaching Diploma

Moved to Dubai, taught AP Physics

2012 - 2013

MA Education Management & Leadership

2014

OUR EVOLUTION

2015

Jeddah

High School Physics Teacher,
AP, IB and IGCSE

Curriculum Designer and
Trainer (STEM EDUCATION)

2020

Houston, TX

Tutor (IB Physics)
Caregiver

2021 - 2022

Data Science Certificate

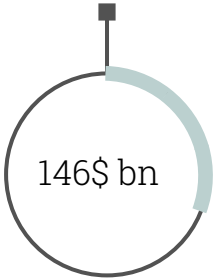
USED CARS MARKET IN THE US

The USED CARS market in the US

Used car sales in any country is a derivative of the new car market. In a market like the US where 70-80% of new vehicles sold are either leased/financed, the new cars make their way to the used car market in a span of 3-4 years, when the lease/finance period gets over.

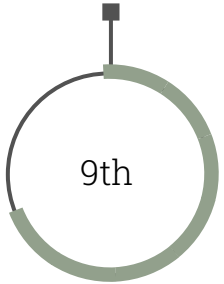
Market Size

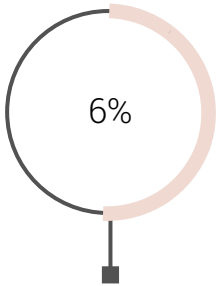
The market size, measured by revenue, of the Used Car Dealers industry is $146.5bn in 2022



Rank in Market

the 90th largest in the US.





GROWTH RATE

The market size of the Used Car Dealers industry is expected to increase 6% in 2022.

LOW SUPPLY
HIGH DEMAND

- The cost of used cars jumped 10 percent month on month in Oct 2021 and was up 21 percent compared with a year earlier.
- The surge in prices is driven by the slowdown in new-car production.
- Due to high Demand, Prices have recently been rising on used cars, not going down over time,.
- Consumers can now find a greater inventory of used vehicles online, and dealers are accelerating digital efforts too.
- There is a shift in sales towards EV cars.

AND HENCE,

WHAT FEATURES OF A CAR

contribute to the increase in its price?

THIS STUDY

Predicting Used cars price given its specification

FOR SELLERS

1. have a better understanding of what makes a car pricier
2. rethink the stock
3. find the strongest purchase potentials

As a Final Project

1. showcase of knowledge
2. Apply ML regression using python
3. Graduate!

Methods

01 **Multiple Linear Regression**

There are several dependent variables

02 **Python**

personal preference

03 **Packages**

Pandas, Numpy, Seaborn, Sklearn

GET TO KNOW
THE DATA

missing data

categorical: dummy code

categorical: dummy code

DROPPED

categorical: dummy code

DROPPED

```
[4]: cars.describe(include='all')
```

[4]:

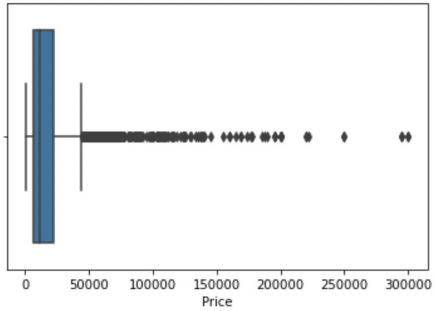
	Brand	Price	Body	Mileage	EngineV	Engine Type	Registration	Year	Model
count	4345	4173.000000	4345	4345.000000	4195.000000	4345	4345	4345.000000	4345
unique	7	NaN	6	NaN	NaN	4	2	NaN	312
top	Volkswagen	NaN	sedan	NaN	NaN	Diesel	yes	NaN	E-Class
freq	936	NaN	1649	NaN	NaN	2019	3947	NaN	199
mean	NaN	19418.746935	NaN	161.237284	2.790734	NaN	NaN	2006.550058	NaN
std	NaN	25584.242620	NaN	105.705797	5.066437	NaN	NaN	6.719097	NaN
min	NaN	600.000000	NaN	0.000000	0.600000	NaN	NaN	1969.000000	NaN
25%	NaN	6999.000000	NaN	86.000000	1.800000	NaN	NaN	2003.000000	NaN
50%	NaN	11500.000000	NaN	155.000000	2.200000	NaN	NaN	2008.000000	NaN
75%	NaN	21700.000000	NaN	230.000000	3.000000	NaN	NaN	2012.000000	NaN
max	NaN	300000.000000	NaN	980.000000	99.990000	NaN	NaN	2016.000000	NaN

Outliers situation
values beyond accepted values

Change the quantitative variables from float to int

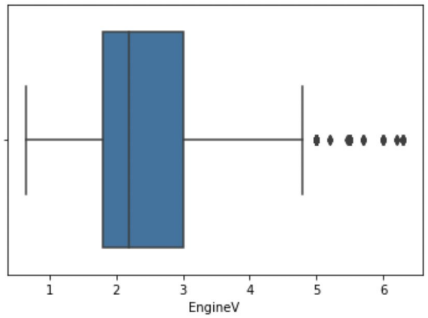
[11]:
```
sns.boxplot(cars2['Price'])
```

[11]: <AxesSubplot:xlabel='Price'>



[25]:
```
sns.boxplot(cars2.EngineV)
```

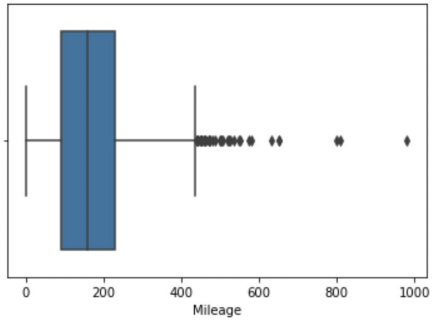
[25]: <AxesSubplot:xlabel='EngineV'>



Boxplot to examine the outliers

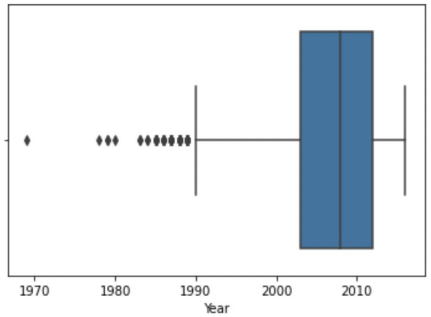
[17]:
```
sns.boxplot(cars2['Mileage'])
```

[17]: <AxesSubplot:xlabel='Mileage'>



[31]:
```
sns.boxplot(cars2.Year)
```

[31]: <AxesSubplot:xlabel='Year'>



Testing ASSUMPTIONS

Multicollinearity

[36]:
```
cars2.corr()
```

[36]:

	Price	Mileage	EngineV	Year
Price	1.000000	-0.554466	0.381024	0.569548
Mileage	-0.554466	1.000000	0.020944	-0.717477
EngineV	0.381024	0.020944	1.000000	-0.029689
Year	0.569548	-0.717477	-0.029689	1.000000

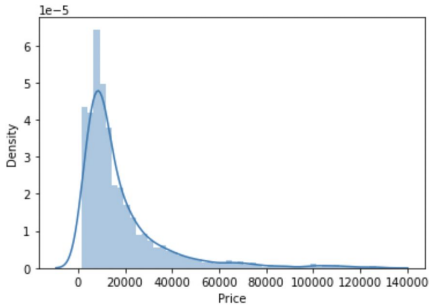
High correlation between Year and Mileage. Drop the Year

Linearity and Normality: Transforming Data

1.Price:

[38]:
```
sns.distplot(cars2['Price'])
```

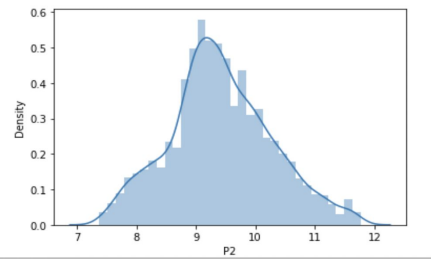
[38]:
```
<AxesSubplot:xlabel='Price', ylabel='Density'>
```



[39]:
```
cars2['P2'] = boxcox(cars2['Price'],0)
```

[40]:
```
sns.distplot(cars2['P2'])
```

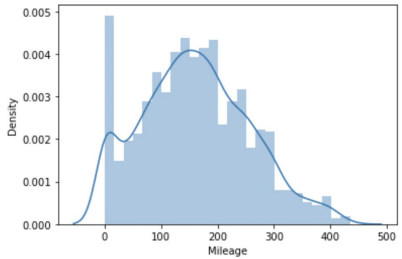
[40]:
```
<AxesSubplot:xlabel='P2', ylabel='Density'>
```



2. Mileage

[41]:
```
sns.distplot(cars2['Mileage'])
```

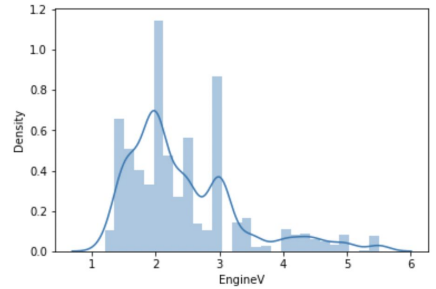
[41]:
```
<AxesSubplot:xlabel='Mileage', ylabel='Density'>
```



3. EngineV:

[42]:
```
sns.distplot(cars2['EngineV'])
```

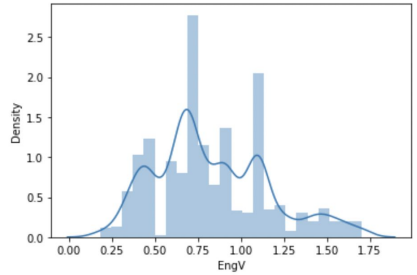
[42]:
```
<AxesSubplot:xlabel='EngineV', ylabel='Density'>
```



[43]:
```
cars2['EngV'] = boxcox(cars2['EngineV'],0)
```

[44]:
```
sns.distplot(cars2['EngV'])
```

[44]:
```
<AxesSubplot:xlabel='EngV', ylabel='Density'>
```



Analysis

Create the Basic Model

```
x = df.drop(['P2'], axis=1)
y = df['P2']
```

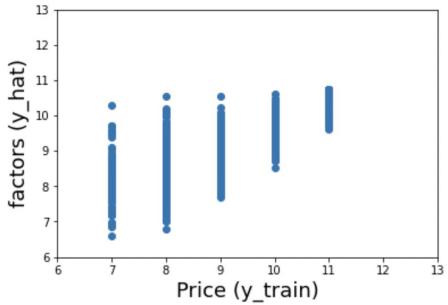
Split and Train

```
x_train, x_test, y_train, y_test = train_test_split (inputs_scaled, y, test_size = 0.2, random_state=365)
```

```
reg = LinearRegression()
reg.fit(x_train, y_train)
```

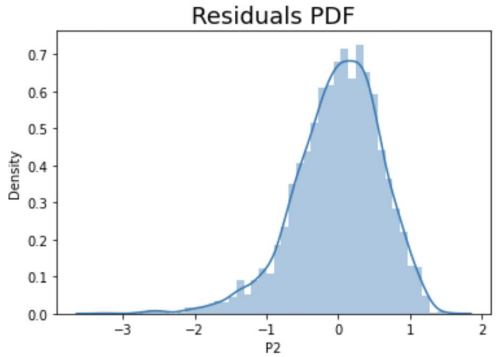
[58]:
```
y_hat = reg.predict(x_train)
```

[59]:
```
plt.scatter(y_train, y_hat)
plt.xlabel('Price (y_train)', size=18)
plt.ylabel('factors (y_hat)', size=18)
plt.xlim(6,13)
plt.ylim(6,13)
plt.show()
```



[60]:
```
sns.distplot(y_train - y_hat)
plt.title("Residuals PDF", size = 18)
```

[60]:
```
Text(0.5, 1.0, 'Residuals PDF')
```



Find the regression Score

Find the Weight of Each Feature

[61]:
```
reg.score(x_train, y_train)
```

[61]: 0.5726634422978878

[62]:
```
reg.intercept_ ## this is the bias
```

[62]: 8.926014446121723

[64]:

	Features	Weights
0	Mileage	-5.194189e-01
1	EngV	1.861117e-01
2	Audi	-3.761139e+11
3	BMW	-4.503778e+11
4	Mercedes-Benz	-4.809212e+11
5	Mitsubishi	-3.281750e+11
6	Renault	-3.833605e+11
7	Toyota	-4.059913e+11
8	Volkswagen	-5.048994e+11
9	crossover	5.622135e+11
10	hatch	3.371466e+11
11	other	4.041481e+11
12	sedan	6.743811e+11
13	vagon	4.107041e+11
14	van	5.077486e+11
15	Diesel	8.338364e+12
16	Gas	5.925950e+12
17	Other	2.661072e+12
18	Petrol	7.990651e+12

Basic Findings

Mileage

Volkswagen and Mercedes

Sedan

Limitations

- this study does not include EV cars though there is a bigger demand for them with time.
- Petrol and gas are used interchangeably in the US, it is not clear why there has been a distinction between the two in the study.

Recommendations

- Whether selling or buying a used car, mileage is the biggest factor determining the car's worth.
- Demand should be studied more deeply as in what brands / types of cars customers prefer before setting up a dealership since some brands / car types are more expensive than others (regardless of other features).

THANKS!

Questions!

CREDITS

This is where you give credit to the ones who are part of this project.
Did you like the resources on this template? Get them for free at our other websites.

- Presentation template by Slidesgo
- Icons by Flaticon
- Images & infographics by Freepik
- Author introduction slide photo created by Freepik
- Text & Image slide photo created by Freepik.com
- Big image slide photo created by Freepik.com