INTELIGÊNCIA ARTIFICIAL

UA02 / LABORATÓRIO # 3

ASSUNTO: EDA - EXPLORATORY DATA ANALYSIS

Materiais de Apoio

Site oficial do Scikit Learn:

scikit-learn.org/

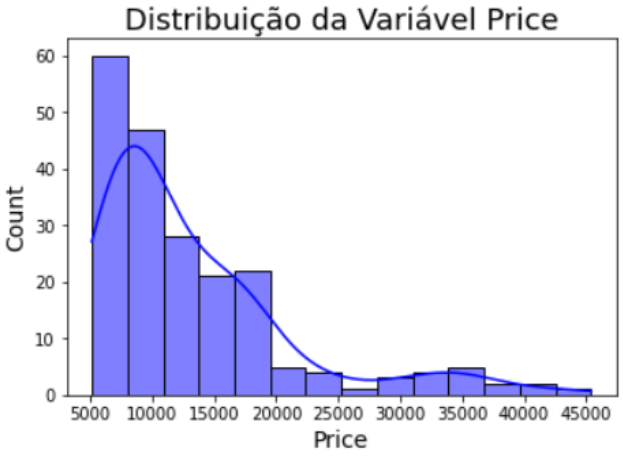
Um bom tutorial de EDA:

https://medium.com/@ugursavci/complete-exploratory-data-analysis-using-python-9f685d67d1e4

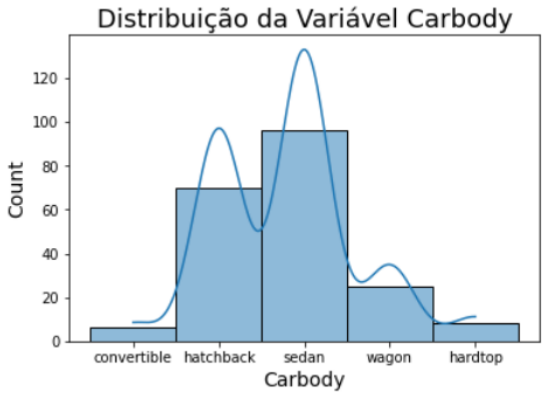
Dataset usado: *CarPrice_Assignment.csv*

```
# Importação de Bibliotecas
import pandas as pd
import seaborn as sns


# Leitura de Dataset - Car Price Assingment, do Kaggle
df = pd.read_csv("CarPrice_Assignment.csv",sep=",")


# Data Understanding - Compreensão dos Dados
df.head()
df.info()
df.describe()
df.describe(include='O')
list(df.carbody.unique())
print(df['carbody'].value_counts())
print(df['fuelsystem'].value_counts())
```

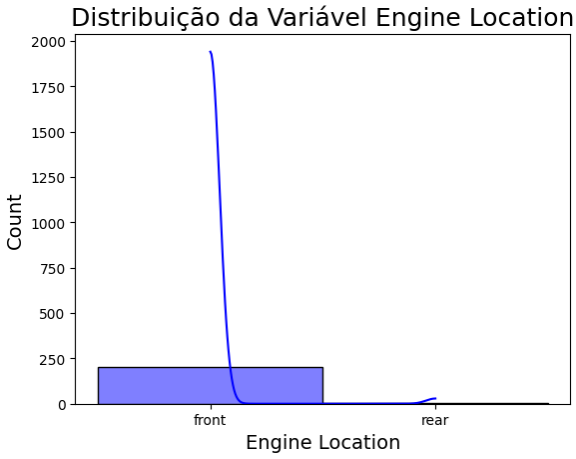
```
# Data Preparation - Preparação dos Dados (Limpeza, por exemplo)

df.isnull().sum() # para totalizar os valores nulos em cada feature
(variável)


df[df.duplicated(keep='first')] # para identificar linhas duplicadas
# df.drop_duplicates(keep='first',inplace=True) - remove linhas
duplicadas, se houver


# Data Visualization

# Distribuição da Variável Price
fig = sns.histplot(data=df, x = 'price', kde=True, color='b')
fig.set_title('Distribuição da Variável Price',size=18)
fig.set_xlabel('Price', size=14)
fig.set_ylabel('Count', size=14)
```



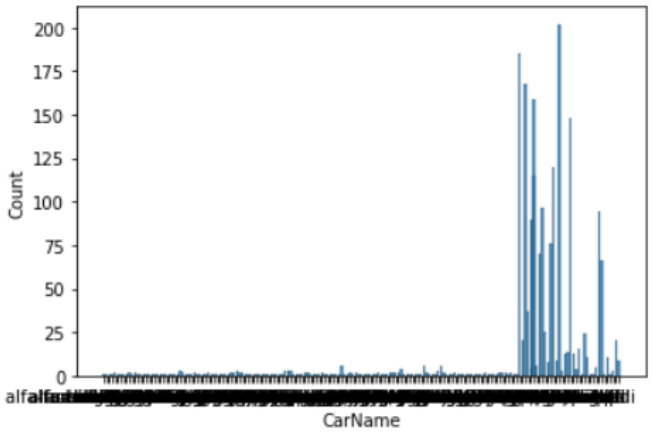

```
# Distribuição da Variável Carbody

fig = sns.histplot(data=df, x = 'carbody', kde=True, color='b')
fig.set_title('Distribuição da Variável Carbody',size=18)
fig.set_xlabel('Carbody', size=14)
fig.set_ylabel('Count', size=14)
```




```
# Distribuição da Variável EngineLocation

fig = sns.histplot(data=df, x = 'enginelocation', kde=True, color='b')
fig.set_title('Distribuição da Variável Engine Location',size=18)
fig.set_xlabel('Engine Location', size=14)
fig.set_ylabel('Count', size=14)
```

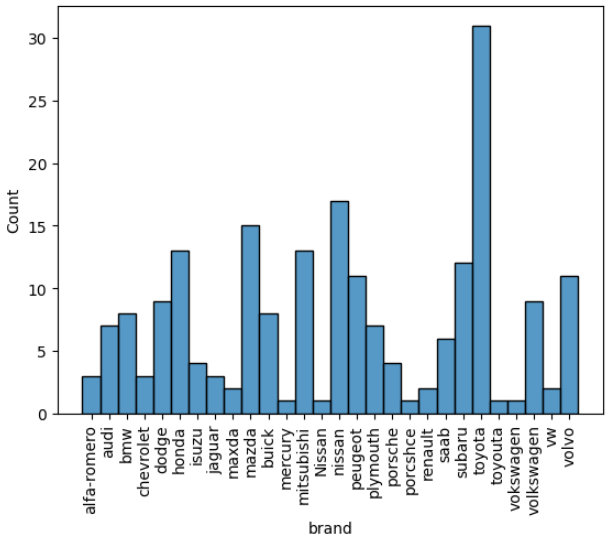



```
# Contagem de Ocorrências de Valores Categóricos - CarName

df_categorical = df.select_dtypes(include = 'object').columns
for i in df_categorical:
    fig = sns.histplot(data=df, x = i, shrink=.8)
```



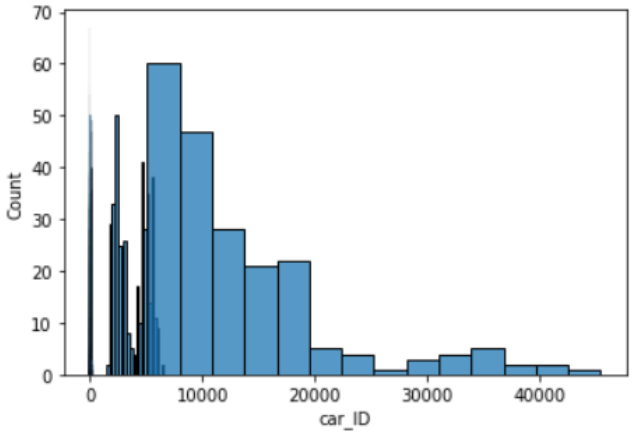
O gráfico bagunçou a exibição com sobreposições de labels do CarName!

Desafio 1: como o gráfico anterior ficou bagunçado, como obter o gráfico abaixo?



Desafio 2: observe que alguns dados estão errados no dataframe, na feature “brand”, tais como “maxda” cuja grafia correta é “mazda”. Portanto, analise os dados de “brand” e corrija, usando comando adequado do Pandas.

```
# Contagem de Ocorrências de Valores Numéricos

df_numerical = df.select_dtypes(exclude = 'object').columns

for i in df_numerical:
    fig = sns.histplot(data=df, x = i)
```

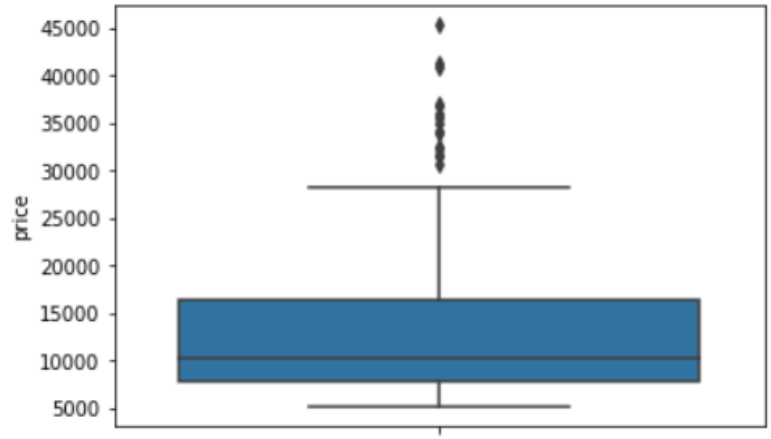


Observe que como campo numérico somente tem o car_ID.

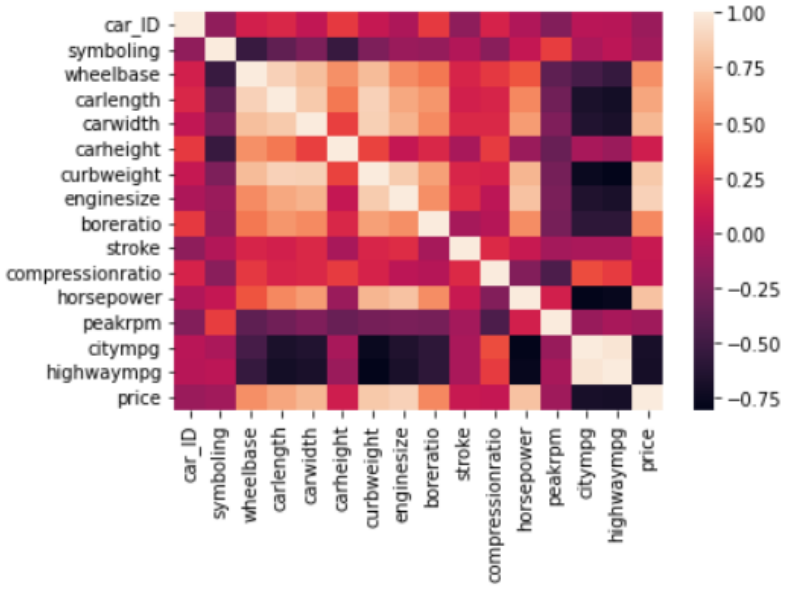
Um ID serve para alguma análise?

```
# Exclusão da Coluna car_ID
df.drop('car_ID', axis = 1, inplace = True)
```

```
# Boxplot do atributo price
fig = sns.boxplot(data=df, y = 'price')
```




```
# Gráfico de Calor (heatmap) de todos os atributos
# Correlação entre atributos - quanto mais clara a cor (mais perto do
valor 1.0), mais alta a correlação
sns.heatmap(df.corr())
```



Problema?

Correção: sns.heatmap(df.select_dtypes(include="number").corr())

Desafio 3: como fazer um gráfico que compare tendências de dois dados numéricos (price e mais outro dado, como cardwidth por exemplo) com alta correlação, positiva ou negativa?

Bom Trabalho!!