

***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

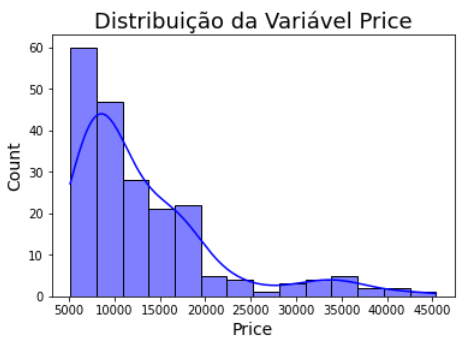
# Distribuição da Variável Price

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)

****

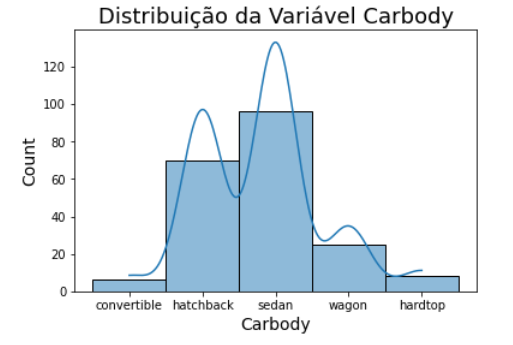
# Distribuição da Variável Carbody

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)



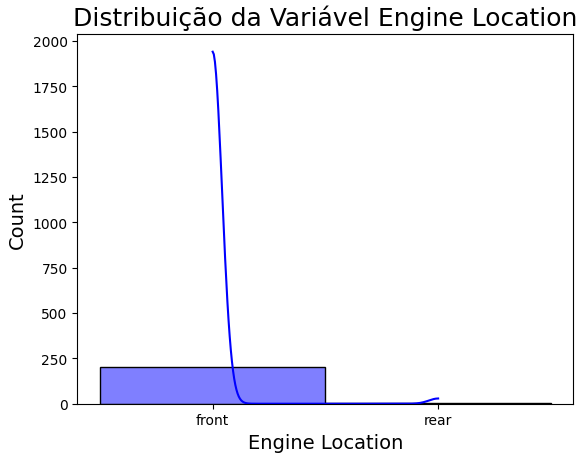
# Distribuição da Variável EngineLocation

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)

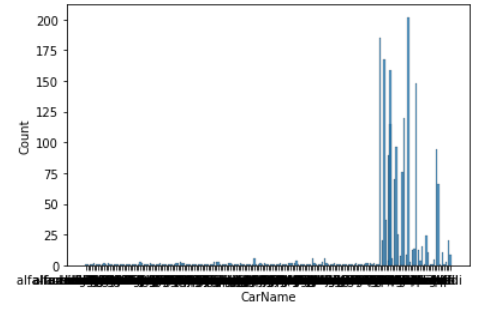


# Contagem de Ocorrências de Valores Categóricos - CarName

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?**

Gráfico, Gráfico de barras

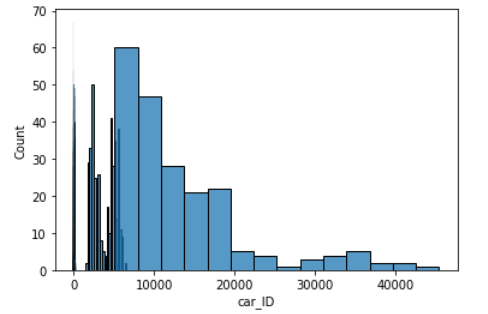
Descrição gerada automaticamente

# 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.**

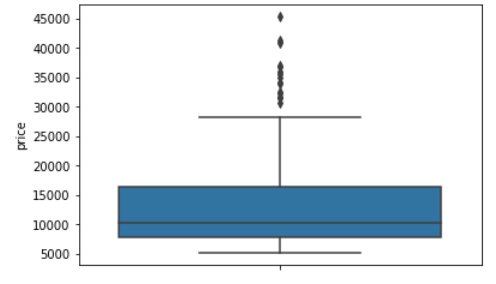
**Um ID serve para alguma análise?**

# Exclusão da Coluna car\_ID

df.drop('car\_ID', axis = 1, inplace = True)

# Boxplot do atributo price

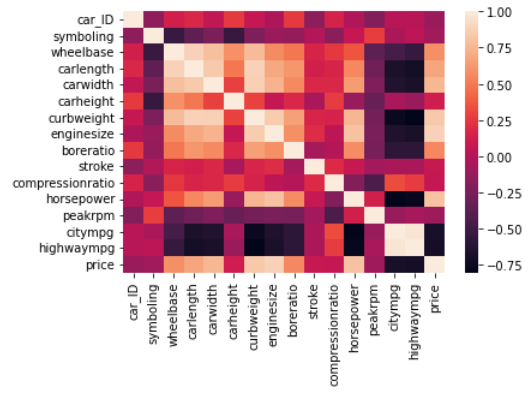
fig = sns.boxplot(data=df, y = 'price')



# 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())



**Desafio 2: 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!!**