Lista 3 - EDA

João Paulo Roberto Delucca - 2016108546

Mateus Cadar - 2014070592

Pedro Dourado - 2012020288

Victor Bráulio Moreira Roberto - 2014046497

Leitura

In [3]:

```
import numpy as np
import matplotlib.pyplot as plt

import pandas as pd
import seaborn as sns

%matplotlib inline
```

In [4]:

```
from sklearn.datasets import load_boston
boston_dataset = load_boston()
```

In [5]:

```
boston = pd.DataFrame(boston_dataset.data, columns=boston_dataset.feature_names)
boston.head()
```

Out[5]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	L
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	

In [6]:

```
boston['CMEDV'] = boston_dataset.target
```

In [7]:

```
boston.isnull().sum()
```

Out[7]:

```
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
AGE        0
DIS        0
RAD        0
TAX        0
PTRATIO    0
B          0
LSTAT      0
CMEDV      0
dtype: int64
```

Análise Exploratória

Algumas perguntas foram formuladas para obtermos um entendimento maior do dataset. Dessa forma, utilizamos histogramas, boxplots, scatterplots e matriz de correlação para responder a tais questões. Escolhemos concentrar todos os insights juntos para que a análise fique mais integrada, mas os gráficos utilizados se encontram logo abaixo.

Como é a distribuição do preço das casas? Ele varia entre quais valores?

> A partir da análise exploratória dos dados, é possível identificar que a variável resposta CMEDV segue uma distribuição normal, mas com alguns outliers. O preço mediano das residências ocupadas por região vai de 5 mil a 50 mil dólares. O preço médio dos preços medianos é 22 mil e 500 dólares.

Quais as variáveis que mais influenciam no preço das casas?

> As variáveis mais correlacionadas ao preço da casa são RM (número médio de quartos) e LSTAT (% população de baixa renda). Outro ponto importante é o fato de as variáveis mais importantes, RM e LSTAT, possuirem um comportamento não linear, o que deve ser considerado em um posterior ajuste do modelo.

Como a taxa de criminalidade afeta o preço das casas?

> Existe uma correlação negativa entre a criminalidade e o preço das casas, ou seja, quanto maior a criminalidade, menor o preço da casa. Entretanto, exitem muitas observações em que a taxa de criminalidade é proxima de 0, portanto a criminalidade explica bem o preço apenas quando ela é maior do que 0.

Outras informações relevantes encontradas:

1) A informação de DIS (distância dos centros de emprego) provavelmente está contida em outras 3 variáveis, que também estão interrelacionadas: AGE (unidades construídas antes de 1940), NOX (concentração de óxido nítrico) e INDUS (quantidade de indústrias próximas)

2) RAD (acessibilidade a rodovias) e TAX (imposto predial) têm correlação muito forte

In [8]:

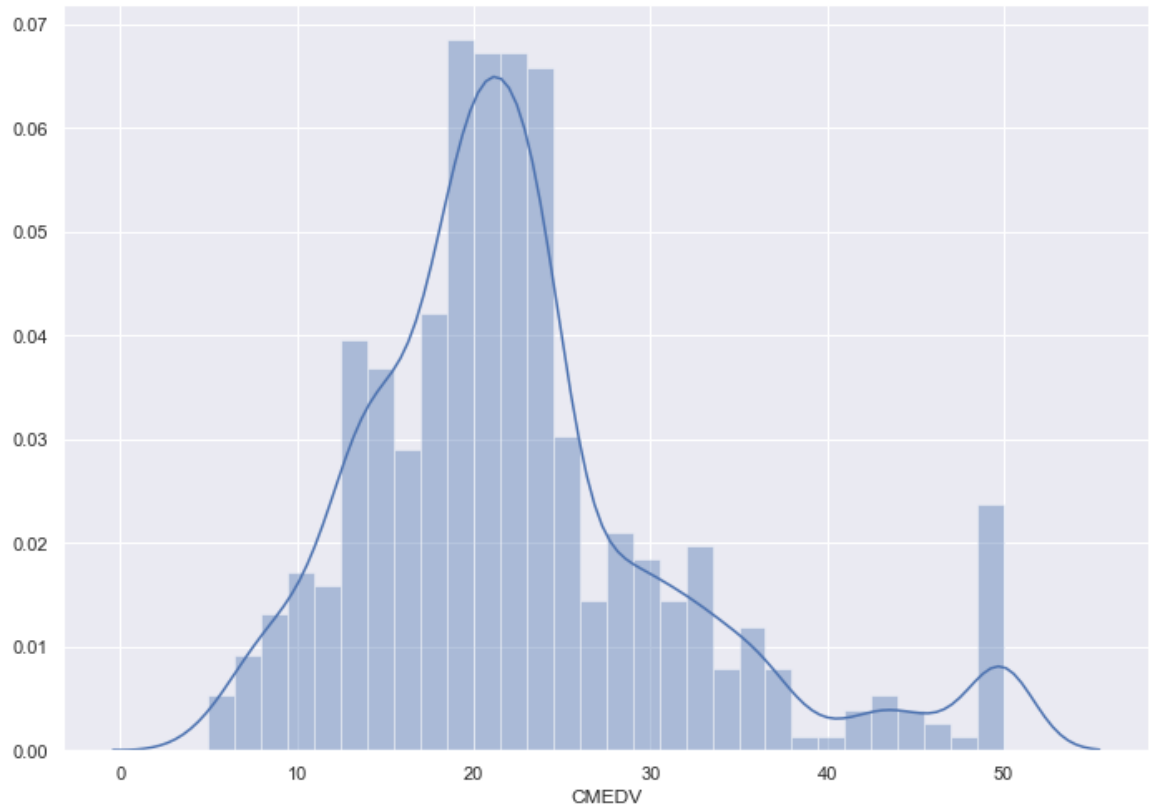
```
boston['CMEDV'].describe()
```

Out[8]:

```
count    506.000000
mean      22.532806
std        9.197104
min        5.000000
25%       17.025000
50%       21.200000
75%       25.000000
max       50.000000
Name: CMEDV, dtype: float64
```

In [9]:

```
sns.set(rc={'figure.figsize':(11.7,8.27)})
sns.distplot(boston['CMEDV'], bins=30)
plt.show()
```



In [10]:

```
from scipy.stats import shapiro
stat, p = shapiro(boston)
print('Statistics=%.3f, p=%.3f' % (stat, p))
```

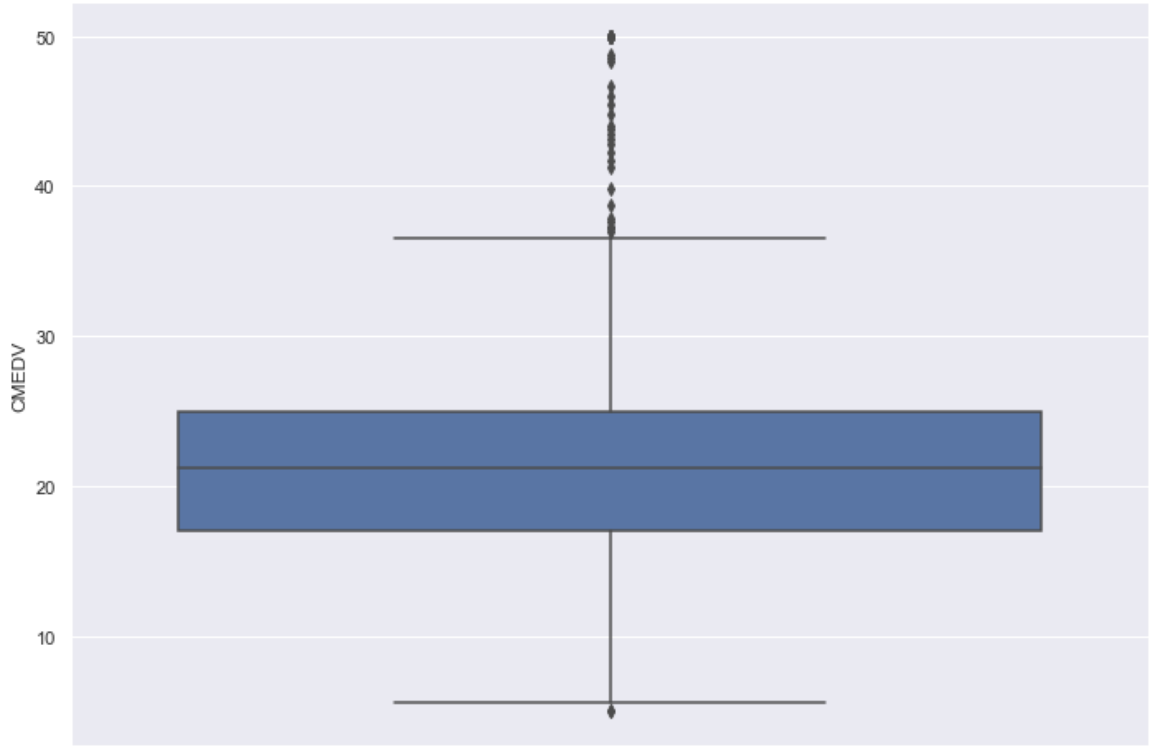
```
Statistics=0.519, p=0.000

/Users/victorbmr/anaconda3/lib/python3.7/site-packages/scipy/stats/m
orestats.py:1660: UserWarning: p-value may not be accurate for N > 5
000.
  warnings.warn("p-value may not be accurate for N > 5000.")
```

In [8]:

```
sns.set(rc={'figure.figsize':(11.7,8.27)})
sns.boxplot(boston['CMEDV'], orient="v")
```

Out[8]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a19e1c5f8>
```

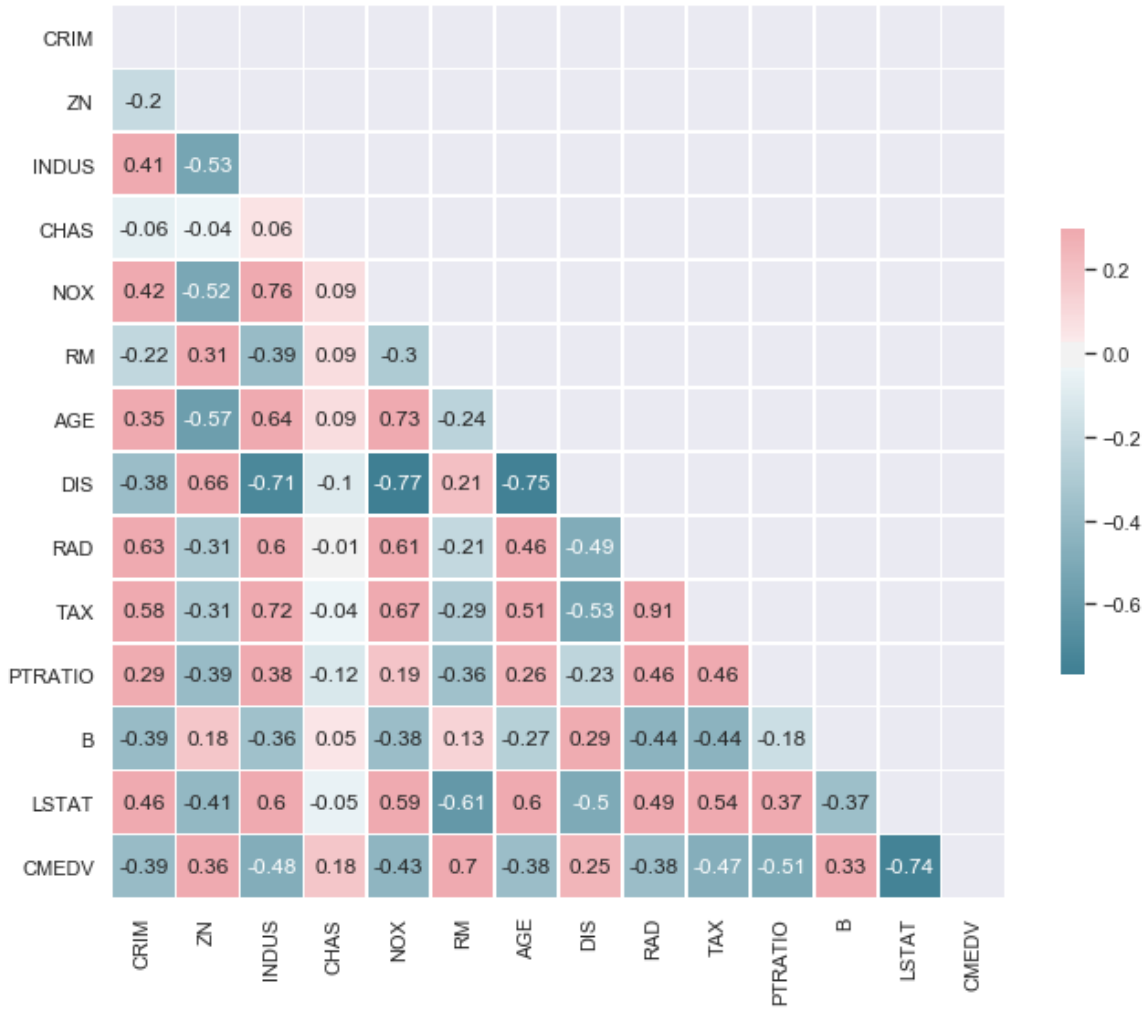


In [9]:

```
corr = boston.corr().round(2)

mask = np.zeros_like(corr, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True

f, ax = plt.subplots(figsize=(11, 9))

cmap = sns.diverging_palette(220, 10, as_cmap=True)

sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5}, annot=True)
```

Out[9]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1a203f28>
```

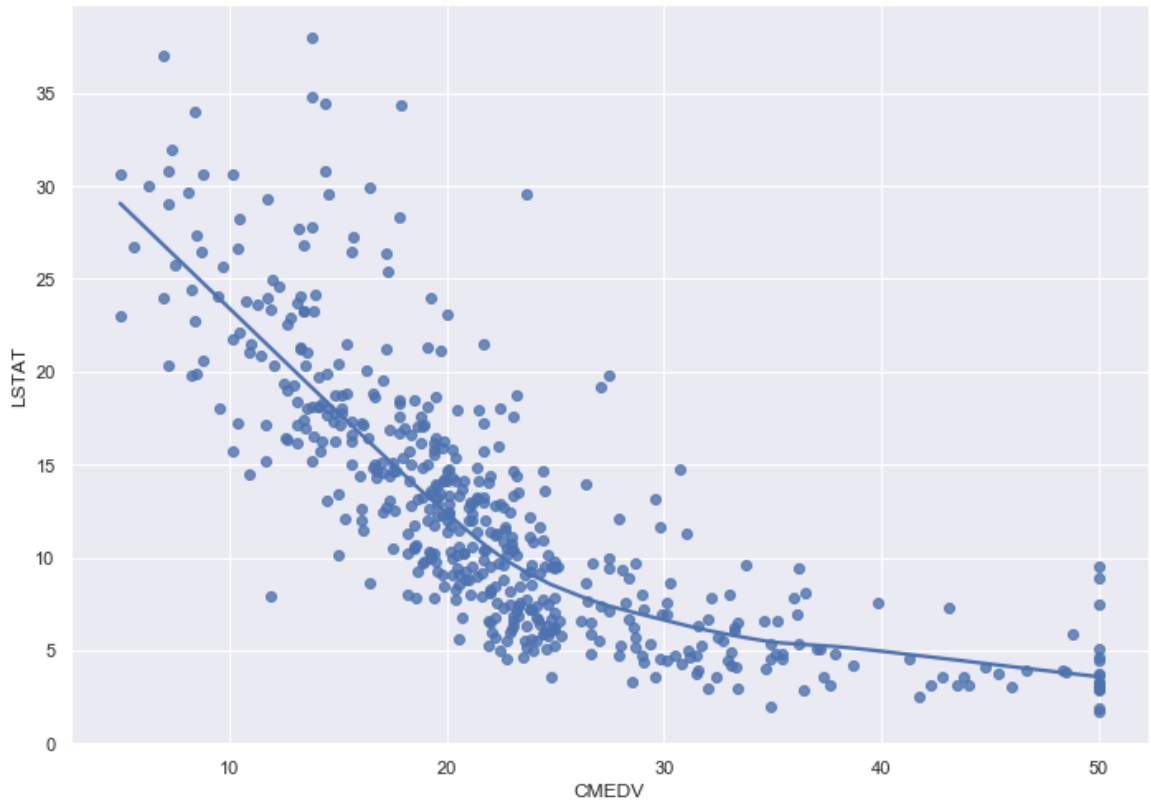


In [18]:

```
sns.regplot(x="CMEDV", y="LSTAT", lowess=True, data=boston)
```

Out[18]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1b3cbb38>
```

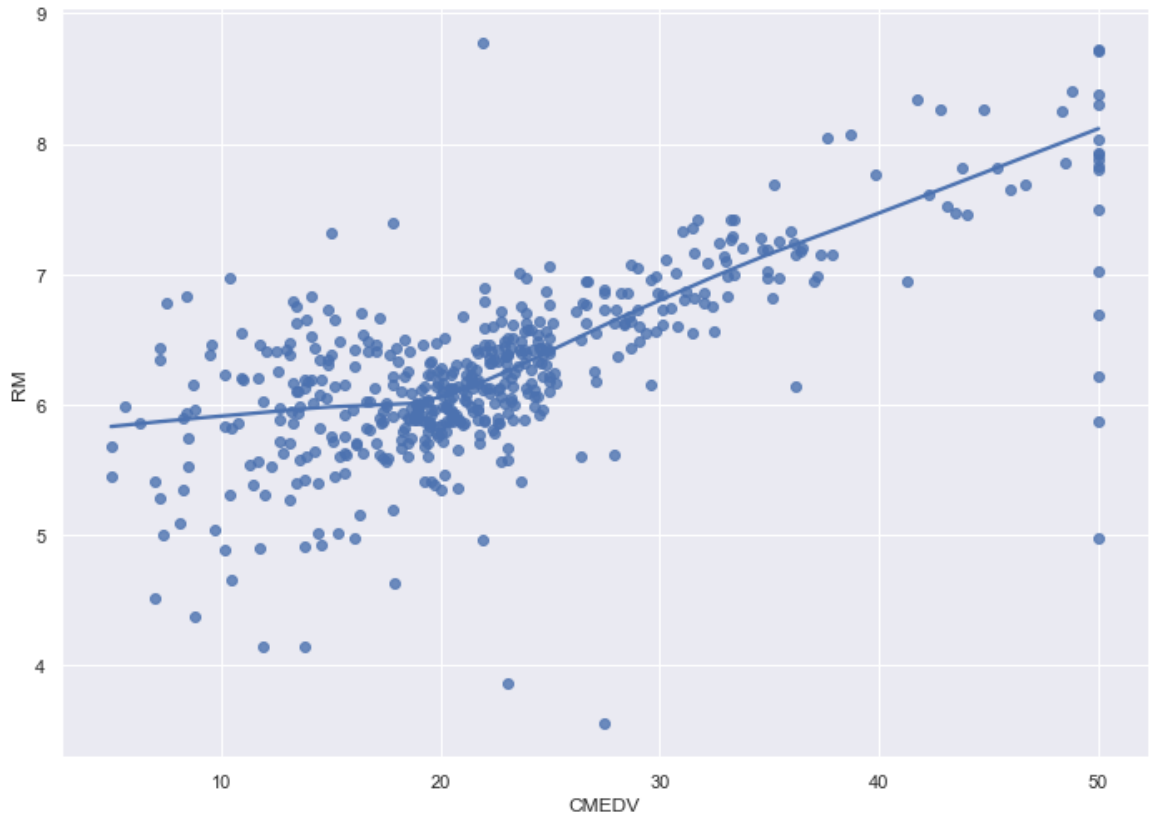


In [32]:

```
sns.regplot(x="CMEDV", y="RM", lowess=True, data=boston)
```

Out[32]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1ca49c88>
```

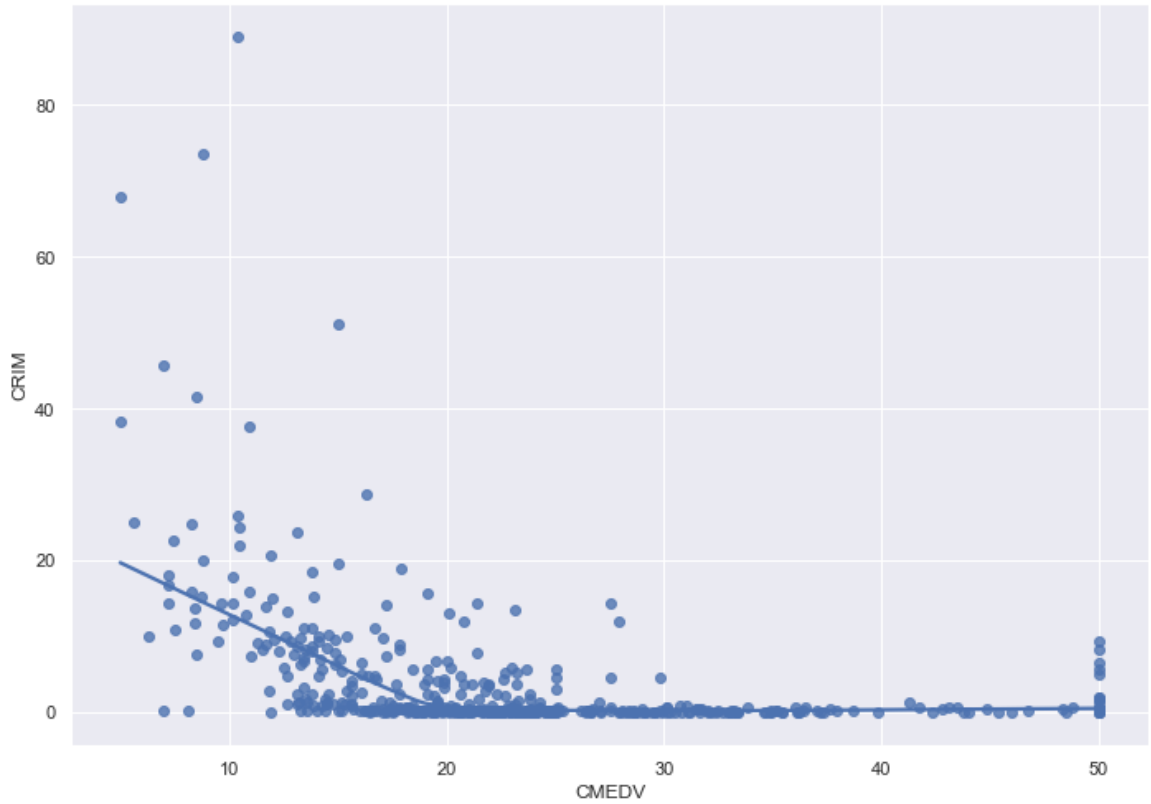


In [12]:

```
sns.regplot(x="CMEDV", y="CRIM", lowess=True, data=boston)
```

Out[12]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a14f543c8>
```



In [34]:

```
sns.scatterplot(x="CMEDV", y="RM",hue="TAX", data=boston)
```

Out[34]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1cc3c9b0>
```