Introdução à Ciência de Dados com Python

Análise Exploratória de Dados

Residência em TIC 09
Programa Prioritário da Lei de Informática

MCTI FUTURO

INSTITUTO FEDERAL
Maranhão

FUNETEC - PB
FUNDAÇÃO DE APOIO AO IFPB

Softex

GOVERNO FEDERAL
BRASIL
UNIÃO E RECONSTRUÇÃO

Agenda

Conceitos gerais sobre Análise Exploratória de Dados

Exemplo passo à passo

Conceitos sobre Análise Exploratória de Dados

O que é análise exploratória de dados?

AED (EDA, em inglês) é a atividade de extrair informações e padrões dos dados utilizando técnicas de estatística e visualizações de dados

- Etapa menos automatizada que outras etapas
- Mais intuitiva e próxima do negócio
- Potencial de auxiliar em tomada de decisões

Visão geral

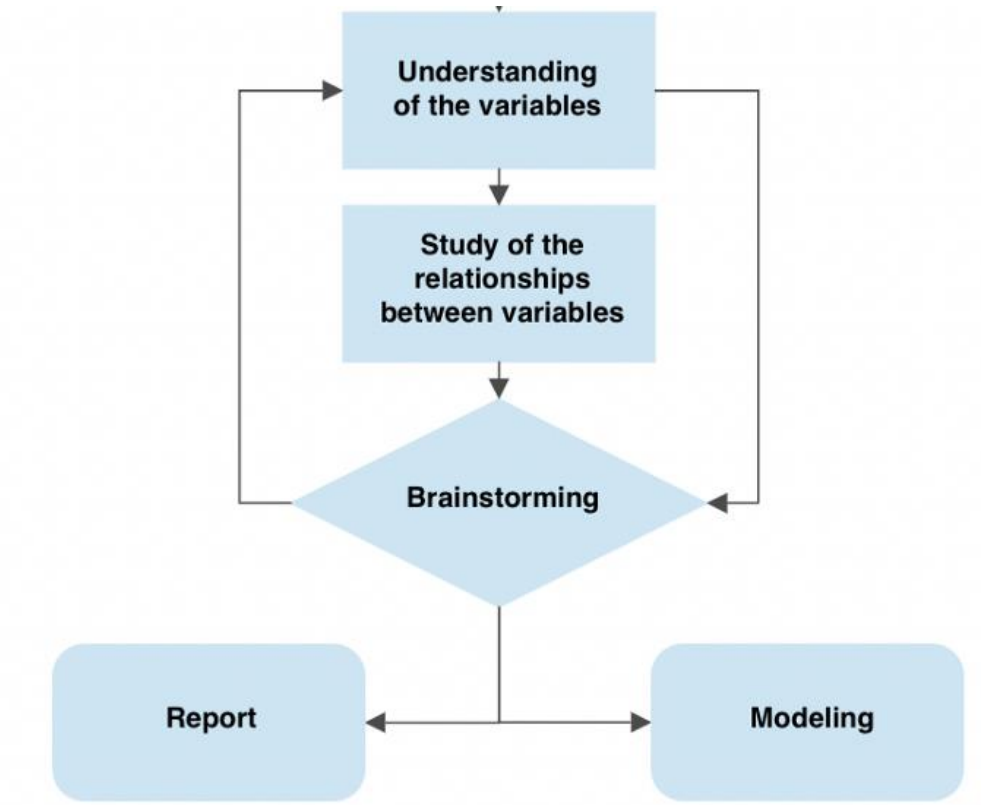
Análise univariada

- Quais são os atributos e tipos?
- Quais as distribuições?
- Há necessidade de transformações?

Análise bivariada

- Há correlações entre atributos?
- Relações lineares ou não lineares?
- Há necessidade de novas colunas?

Análise multivariada e teste de hipóteses



Análise Univariada

INSTITUTO FEDERAL
Maranhão

Softex

Quais são os atributos e tipos?**.info(), .dtypes**

Há variáveis categóricas

- Nominais ou Ordinais
- Quais os valores únicos? **.unique(), .value_count(), .bar()**
- Há desbalanceamento nos valores

Quais as distribuições? **.describe(), .hist(), sns.histplot()**

- Avançado: qqplot, teste de normalidade, etc

Há necessidade de transformações?

Há coluna sem utilidade?

- Mesmo valor ou muitos nulos

Análise Bivariada

Há correlações entre atributos?

.corr(), heatmap()

Relações lineares ou não lineares?

.scatter(), boxplot()

Há necessidade de novas colunas?

dummies? dia? mês? etc.

get_dummies(), OneHotEncoder(),

Há colunas redundantes?

Ex: peso, altura e BMI

Análise Multivariada

INSTITUTO FEDERAL
Maranhão

Softex

Qual a relação da variável alvo (se houver) com pares de outras variáveis?

- **.pairplot()**
- **.scatterplot()** com hue definido pela variável alvo
- **.boxplot()** com hue definido pela variável alvo

Quais relações entre outras variáveis?

- Use cores e tamanho para analisar até 4 variáveis ao mesmo tempo
- Comece pelas mais correlacionadas entre si
- **.boxplot(), catplot(), violinplot()**

Análise Multivariada (avançado)

INSTITUTO FEDERAL
Maranhão

Softex

Levante hipóteses

Quais possíveis influencias de uma variável, ou grupo, sobre a variável alvo?

Faça testes estatísticos

T-Test

ANOVA

Questões em geral

Tente levantar outras perguntas relevantes

- Como a resposta vai ser útil para o negócio?
- Como responder a pergunta com os dados?
- É uma pergunta recorrente?
- É uma pergunta de análise descritiva ou de predição? report, dashboard

Converse com especialista no domínio para encontrar perguntas relevantes

Produtos

Escreva relatório com os principais achados

- Distribuições: normais, lognormais, etc
- Variáveis redundantes ou com baixa qualidade devido a nulos e erros
- Padrões, tendências e correlações encontradas
- Indique ações ou decisões que podem ser tomadas

Código + Dados

- Notebooks, scripts, dashboards, datasets processados

Exemplo

Estudo de Caso

INSTITUTO FEDERAL
Maranhão

Softex

Dataset da House Prices Kaggle Competition

- 79 atributos sobre as características de casas em Ames, Iowa, EUA.
- Os dados estão em dois conjuntos, um de treino e outro de teste.
- Há 1460 instancias (tuplas) em cada um deles.
- Há um total de 81 atributos, dos quais 43 são categóricos.
- Há ainda um campo de identificação e o valor da venda (SalesPrice)

Dependências e Leitura dos Dados

INSTITUTO FEDERAL
Maranhão

Softex

Imports

```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import scipy.stats as st
from sklearn import ensemble, tree, linear_model
import missingno as msno
```

Carregar os dados

```
train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
```

Descrição básica (numéricos)

INSTITUTO FEDERAL
Maranhão

Softex

```
train.describe()
```

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond
count	1460.000000	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000
mean	730.500000	56.897260	70.049958	10516.828082	6.099315	5.575342
std	421.610009	42.300571	24.284752	9981.264932	1.382997	1.112799
min	1.000000	20.000000	21.000000	1300.000000	1.000000	1.000000
25%	365.750000	20.000000	59.000000	7553.500000	5.000000	5.000000
50%	730.500000	50.000000	69.000000	9478.500000	6.000000	5.000000
75%	1095.250000	70.000000	80.000000	11601.500000	7.000000	6.000000
max	1460.000000	190.000000	313.000000	215245.000000	10.000000	9.000000

8 rows × 38 columns

Descrição básica (não numéricos)

INSTITUTO FEDERAL Maranhão

Softex

Codificar? Ord? One Hot?

```
train.describe(exclude='number')
```

nulos?

	MSZoning	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood
count	1460	1460	91	1460	1460	1460	1460	1460	1460
unique	5	2	2	4	4	2	5	3	25
top	RL	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	NAmes
freq	1151	1454	50	925	1311	1459	1052	1382	225

4 rows × 43 columns

Descrição básica

INSTITUTO FEDERAL
Maranhão

Softex

Use head(), tail() e sample(), .shape

```
train.shape , test.shape
```

```
((1460, 81), (1459, 80))
```

```
train.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea
0	1	60	RL	65.0	8450
1	2	20	RL	80.0	9600
2	3	60	RL	68.0	11250
3	4	70	RL	60.0	9550
4	5	60	RL		

5 rows × 81 columns

```
train.tail()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea
1455	1456	60	RL	62.0	7917
1456	1457	20	RL	85.0	13175
1457	1458	70	RL	66.0	9042
1458	1459	20	RL	68.0	9717
1459	1460	20	RL	75.0	9937

5 rows × 81 columns

```
train.sample(n=10)
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley
30	31	70	C (all)	50.0	8500	Pave	Pave
833	834	20	RL	100.0	10004	Pave	NaN
1411	1412	50	RL	80.0	9600	Pave	NaN
9	10	190	RL	50.0	7420	Pave	NaN
1136	1137	50	RL	80.0	9600	Pave	NaN
1185	1186	50	RL	60.0	9738	Pave	NaN
1319	1320	20	RL	75.0	10215	Pave	NaN
1044	1045	20	RL	80.0	9600	Pave	NaN
229	230	120	RL	43.0	3182	Pave	NaN
193	194	160	RM	24.0	2522	Pave	NaN

10 rows × 81 columns

Separe colunas numéricas e categóricas

```
numeric_features = train.select_dtypes(include=['number'])

numeric_features.columns
```

```
Index(['Id', 'MSSubClass', 'LotFrontage', 'LotArea', 'OverallQual',
       'OverallCond', 'YearBuilt', 'YearRemodAdd', 'MasVnrArea', 'BsmtFinSF1',
       'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF',
       'LowQualFinSF', 'GrLivAr
       'HalfBath', 'BedroomAbv(
       'Fireplaces', 'GarageYrB
       'OpenPorchSF', 'Enclose(
       'MiscVal', 'MoSold', 'Yr
      dtype='object')
```

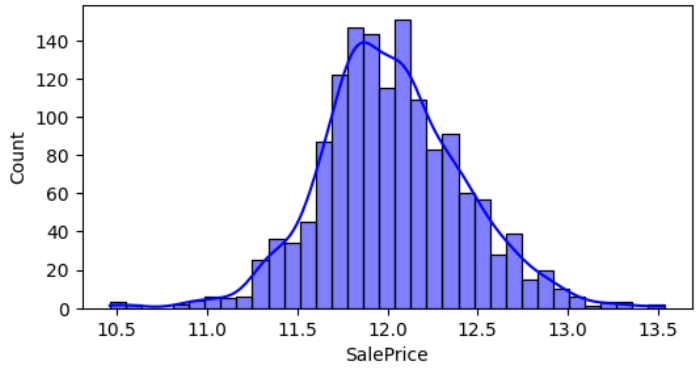
```
categorical_features = train.select_dtypes(include=['object'])

categorical_features.columns
```

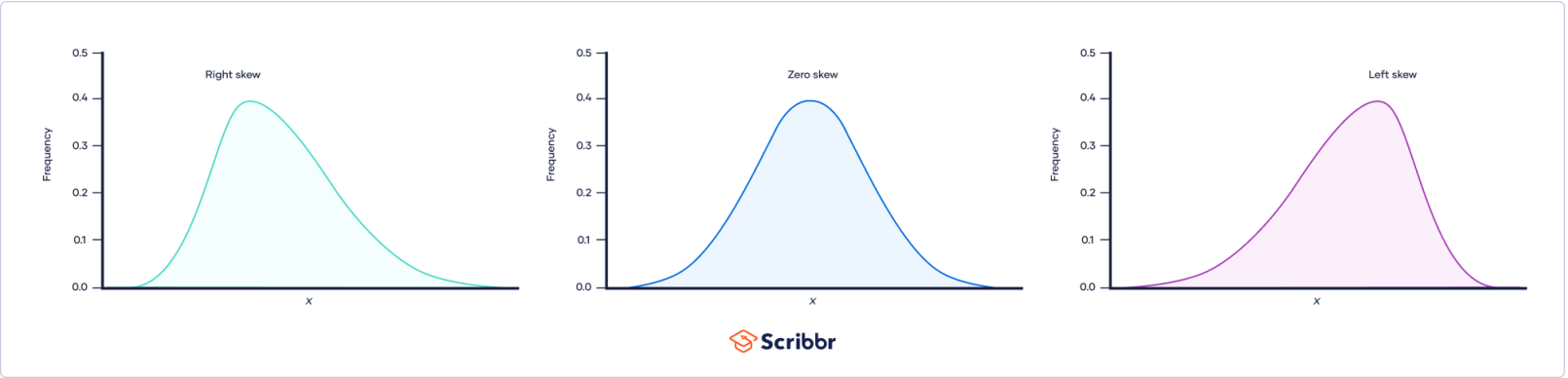
```
Index(['MSZoning', 'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities',
       'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'Condition2',
       'BldgType', 'HouseStyle', 'RoofStyle', 'RoofMatl', 'Exterior1st',
       'Exterior2nd', 'MasVnrType', 'ExterQual', 'ExterCond', 'Foundation',
       'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2',
       'Heating', 'HeatingQC', 'CentralAir', 'Electrical', 'KitchenQual',
       'Functional', 'FireplaceQu', 'GarageType', 'GarageFinish', 'GarageQual',
       'GarageCond', 'PavedDrive', 'PoolQC', 'Fence', 'MiscFeature',
       'SaleType', 'SaleCondition'],
      dtype='object')
```

Análise univariada

Descrição de cada atributo (numérico)

- Histograma
- Assimetria (skew): normal = 0
- Curtose (kurt): achatamento normal = 3
- Boxplots individuais



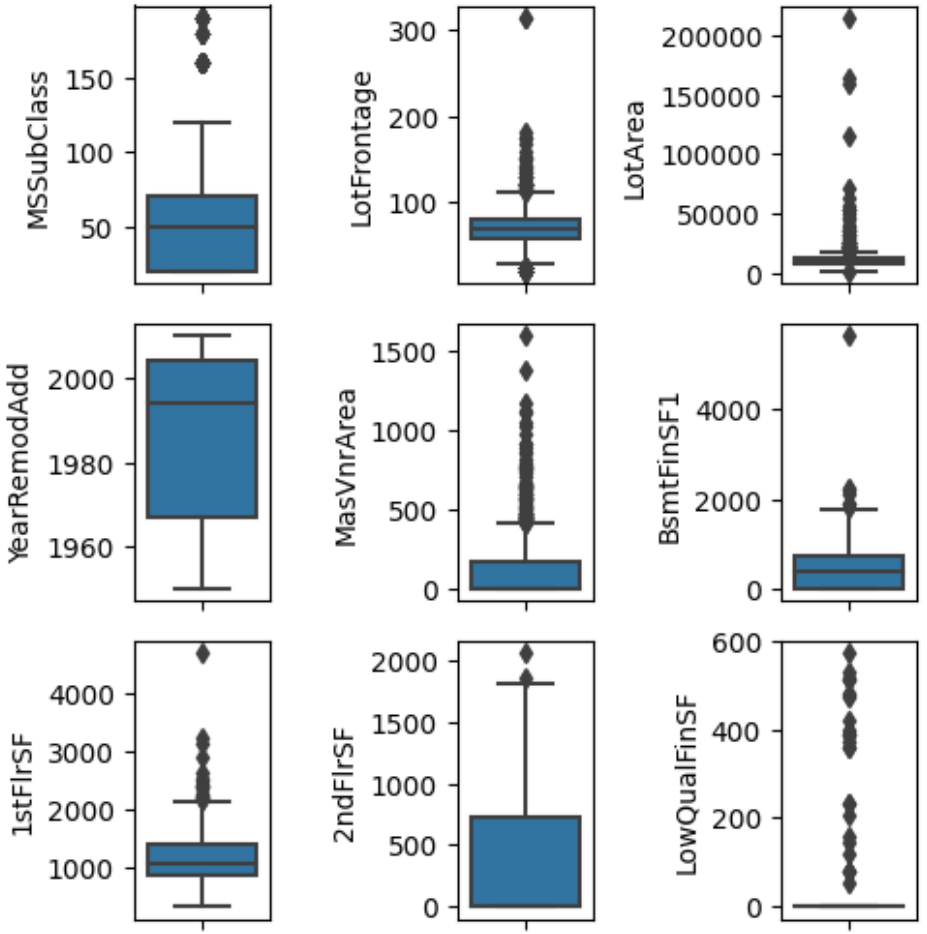


Análise univariada

Exemplo com boxplots

```
nf_cols = numeric_features.columns
```

```
fig,axes=plt.subplots(nrows=6,
                      ncols=6,
                      figsize=(10,10))
axes = axes.flatten() # ou .ravel()
for i,col in enumerate(nf_cols[1:-1]):
    if (i<len(axes)):
        sns.boxplot(data=train,y=col,ax=axes[i])
plt.tight_layout()
```



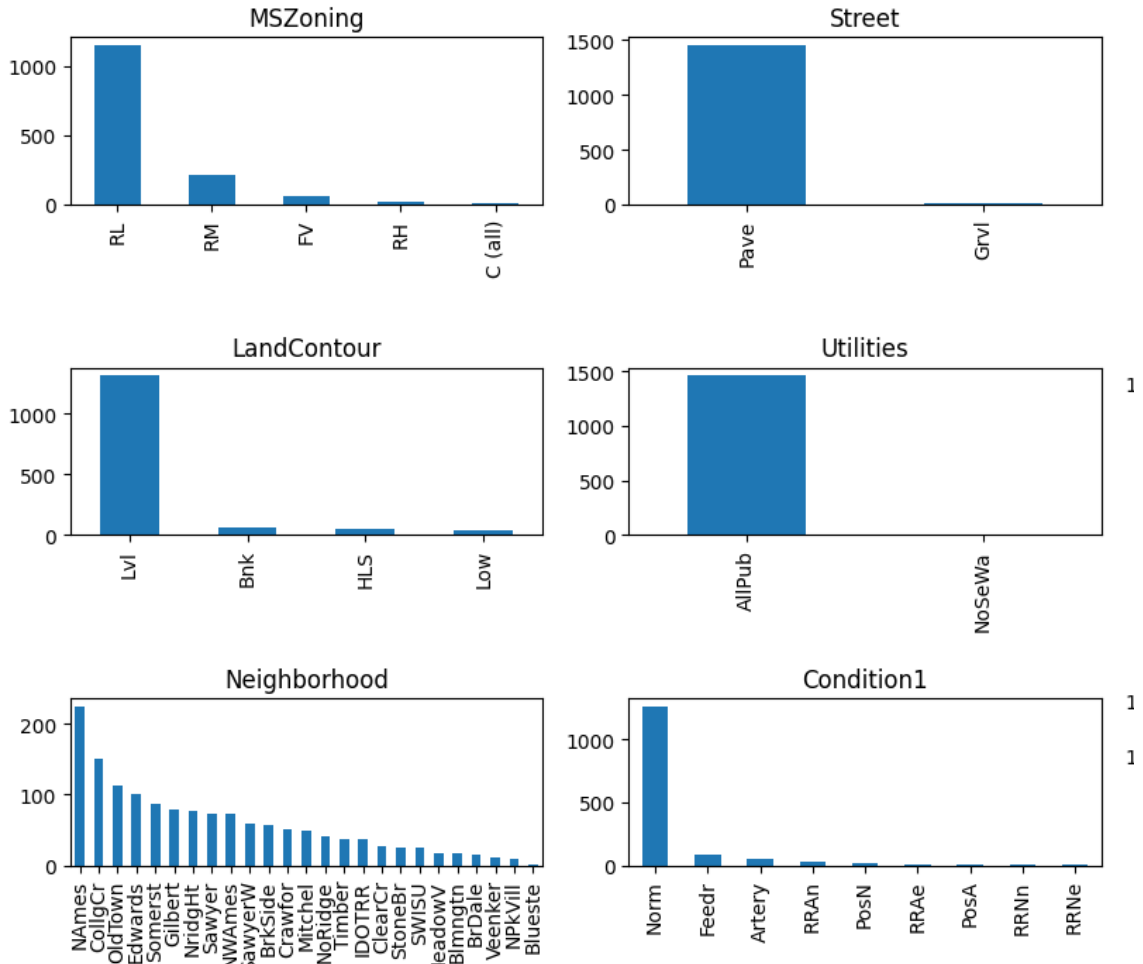
Análise univariada

INSTITUTO FEDERAL
Maranhão

Softex

Atributos categóricos

```
rows, cols = 6,4
fig,axes = plt.subplots(nrows=rows,
                        ncols=cols,
                        figsize=(16,14))
axes=axes.flatten()
for i,c in enumerate(cat_cols[:cols*rows]):
    train_cat_df.iloc[:,i].value_counts().\
                    plot.bar(title=c,ax=axes[i])
plt.tight_layout()
```



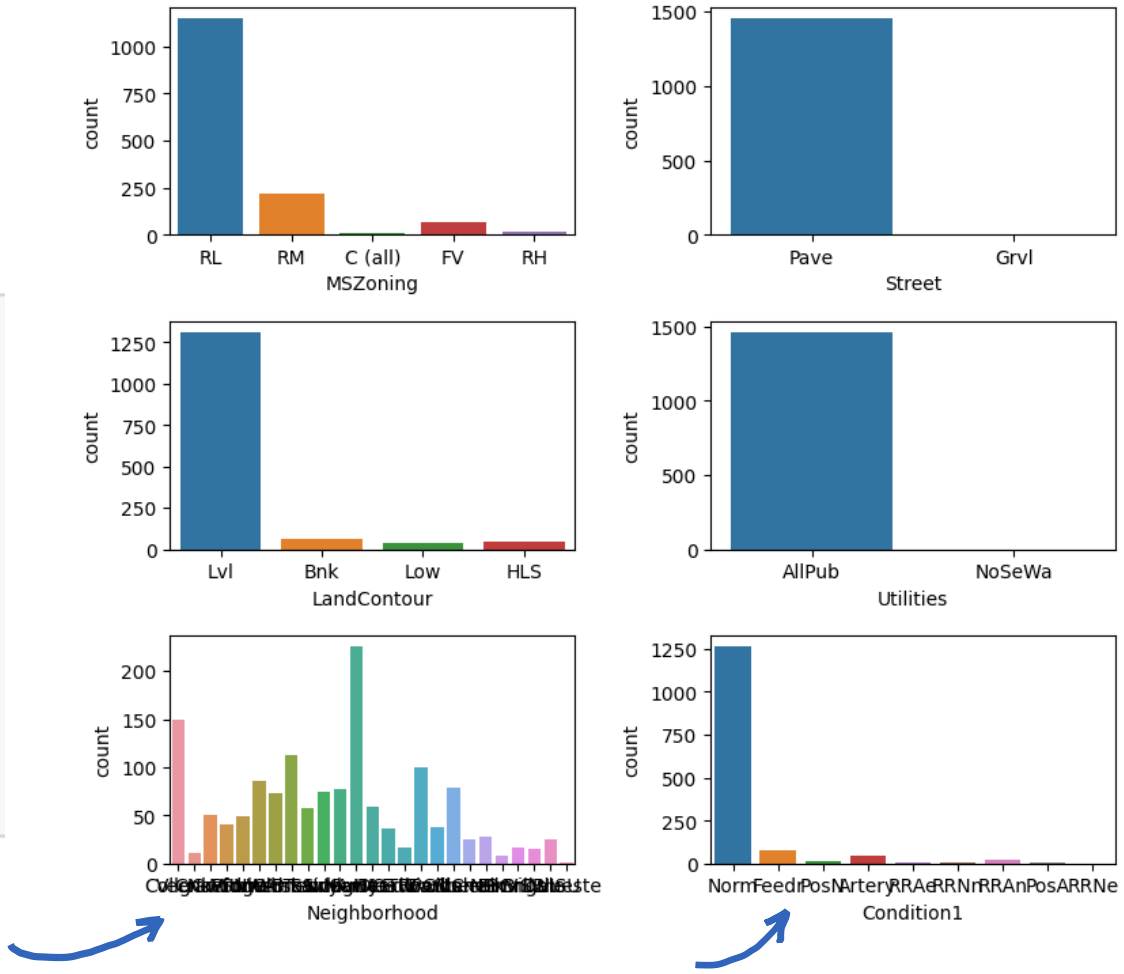
Análise univariada

INSTITUTO FEDERAL
Maranhão

Softex

Atributos categóricos com seaborn countplot

```
rows, cols = 6,4
fig,axes = plt.subplots(nrows=rows,
                        ncols=cols,
                        figsize=(16,14))
axes=axes.flatten()
for i,c in enumerate(cat_cols[:cols*rows]):
    sns.countplot(data=train_cat_df,
                  x=c,
                  ax=axes[i])
plt.tight_layout()
```



Análise bivariada

Há correlação entre os atributos?

Há atributos redundantes?

Há relações lineares ou não-lineares?

Análise bivariada

INSTITUTO FEDERAL
Maranhão

Softex

Correlações

```
correlation = train_num_df.corr()
print(correlation['SalePrice'].sort_values(ascending = False))
```

Forte > 0.7

```
SalePrice        1.000000
OverallQual      0.790982
GrLivArea        0.708624
GarageCars       0.640409
GarageArea       0.623431
TotalBsmtSF      0.613581
1stFlrSF         0.605852
FullBath         0.560664
TotRmsAbvGrd     0.533723
YearBuilt        0.522897
YearRemodAdd     0.507101
GarageYrBlt      0.486362
MasVnrArea       0.477493
```

Positiva

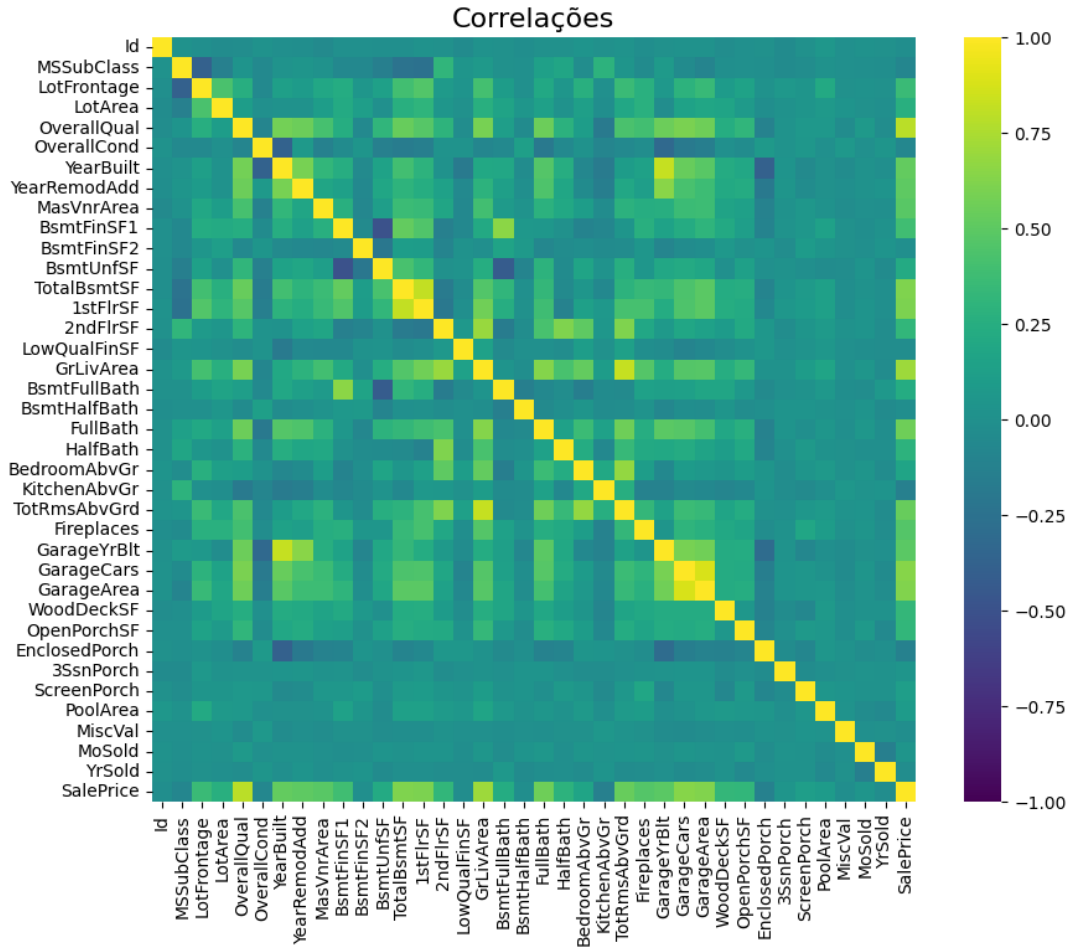
```
BsmtHalfBath    -0.016844
MiscVal         -0.021190
Id              -0.021917
LowQualFinSF    -0.025606
YrSold          -0.028923
OverallCond     -0.077856
MSSubClass      -0.084284
EnclosedPorch   -0.128578
KitchenAbvGr    -0.135907
Name: SalePrice, dtype: float64
```

Negativa

Análise bivariada

Correlações

```
f , ax = plt.subplots(figsize = (10,8))
plt.title('Correlações',y=1,size=16)
sns.heatmap(correlation,
            square = True,
            vmin=-1.0,
            vmax=1.0,
            cmap='viridis')
plt.show()
```



Análise bivariada

Suspeita de multicolinearidade nas correlações muito altas >90

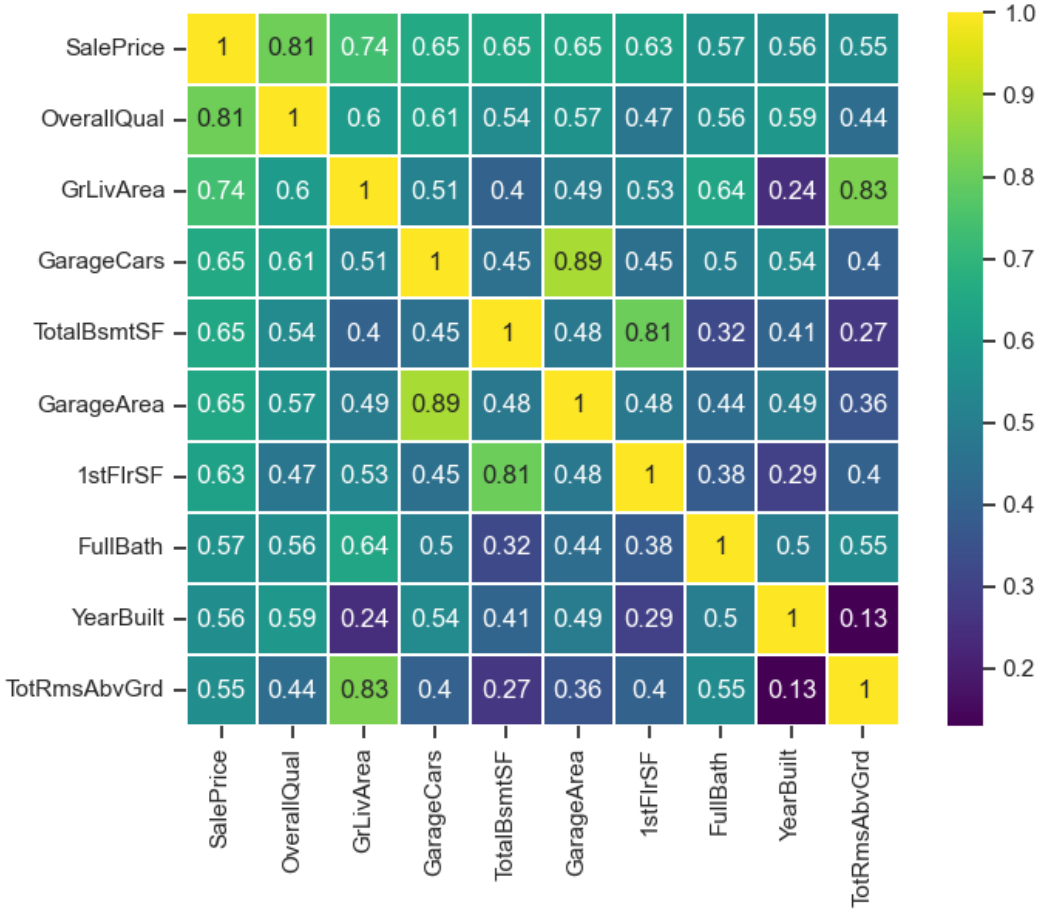
- Colunas 'TotalBsmtSF' e '1stFlrSF' são suspeitas
- Também as colunas Garage_* são suspeitas

Uma delas pode ser excluída na etapa de modelagem

Análise bivariada

Zoom nos atributos que tem maiores correlações com preço

- Se o analista está interessado em saber o que afeta o preço.
- O mesmo serve para qualquer outra coluna

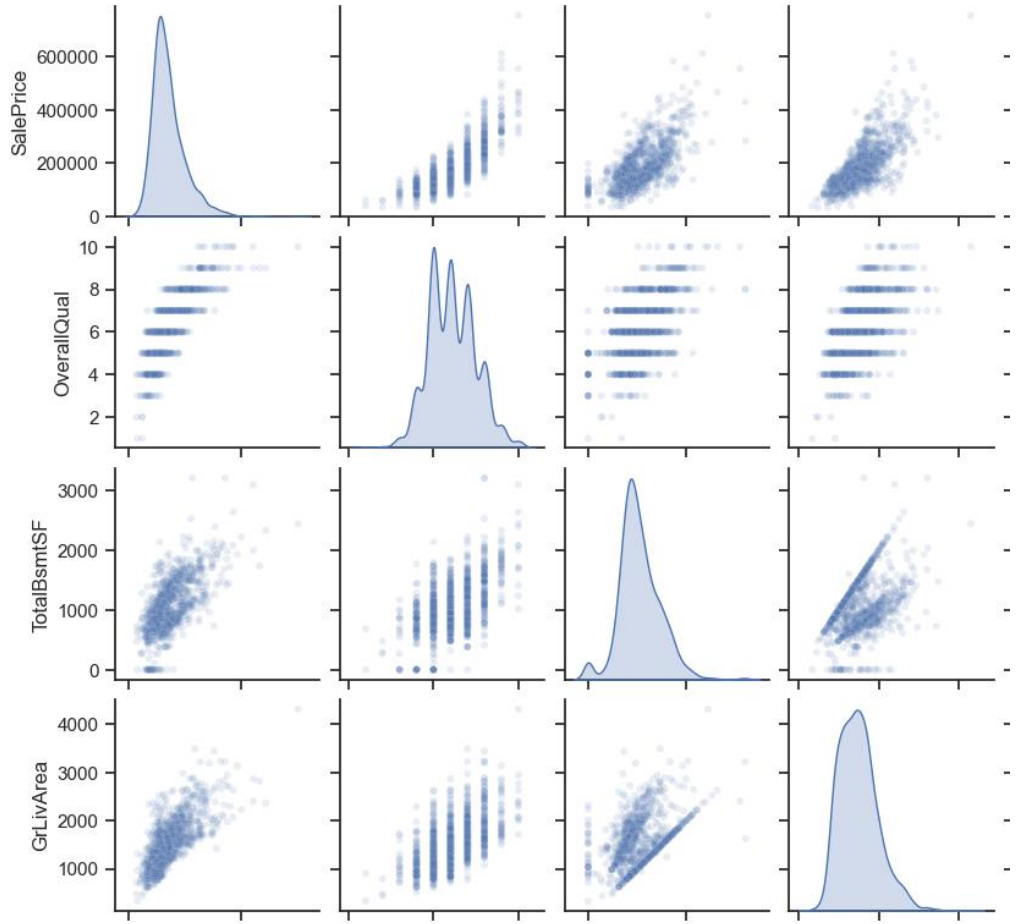


Análise bivariada

Pairplot

SalePrice' e 'YearBuilt' podem tem uma relação não linear, talvez exponencial.

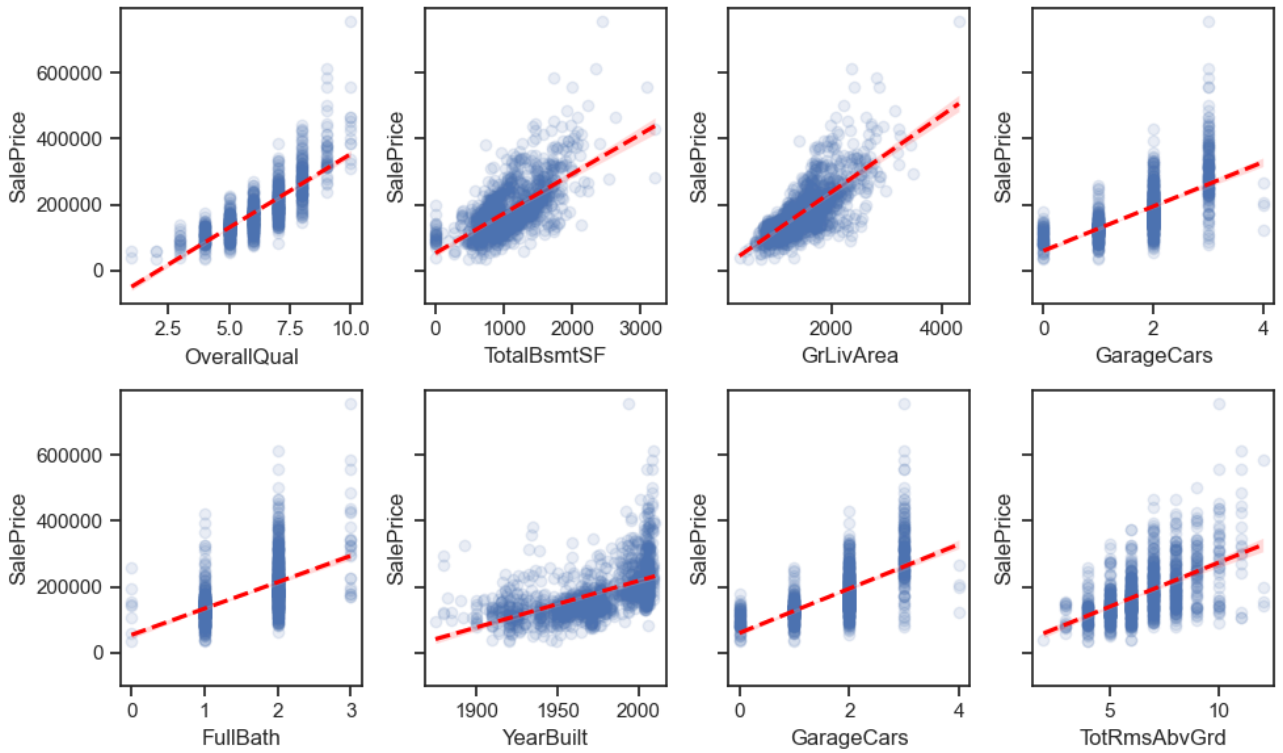
Os preços estão aumentando mais rapidamente em relação aos anos anteriores.



Análise bivariada

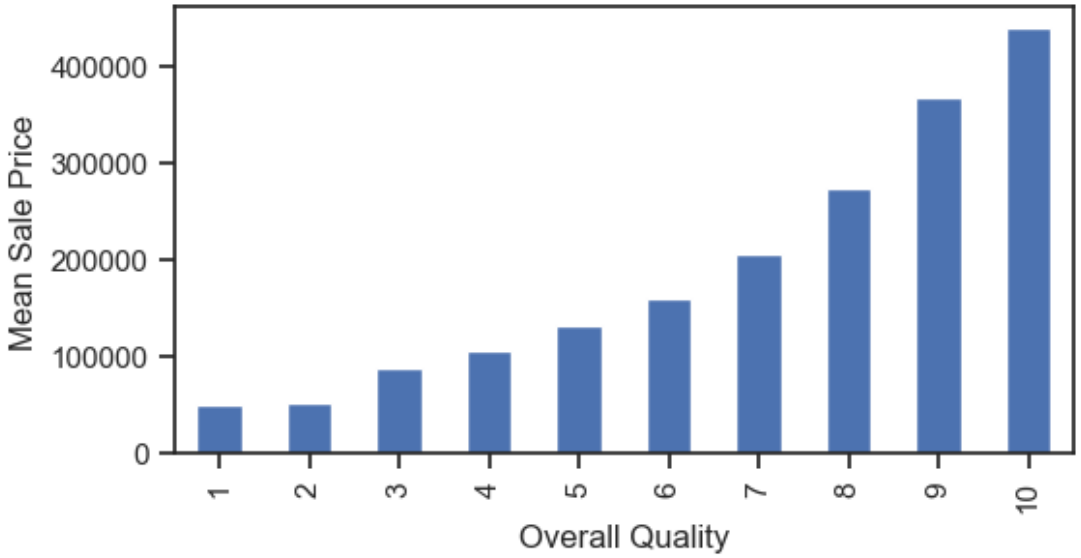
Scatter com linha de regressão

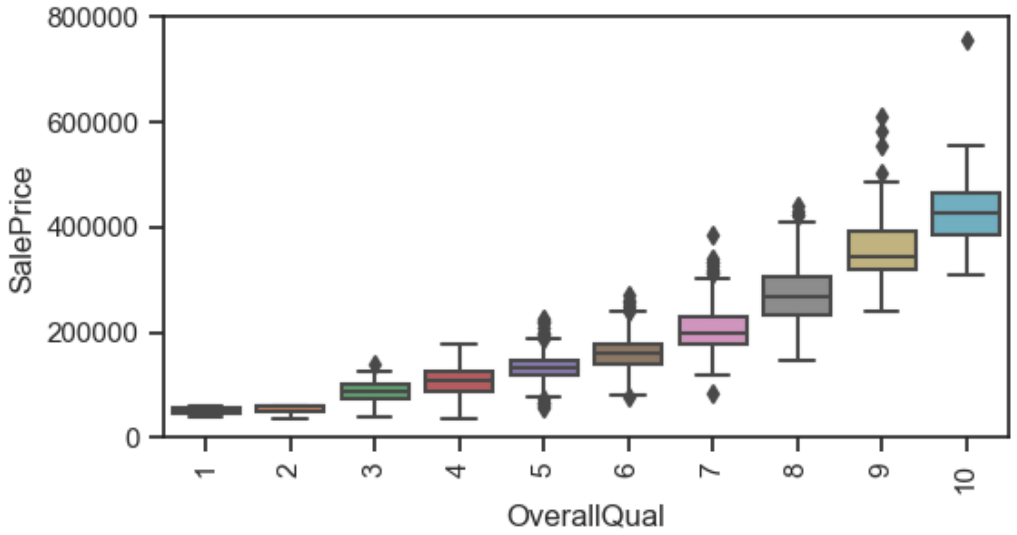
Aqui se pode identificar se há relação linear, não linear ou mesmo não há relação por inspeção visual



Análise bivariada

SalePrice parece ser influenciado pelo OverallQual



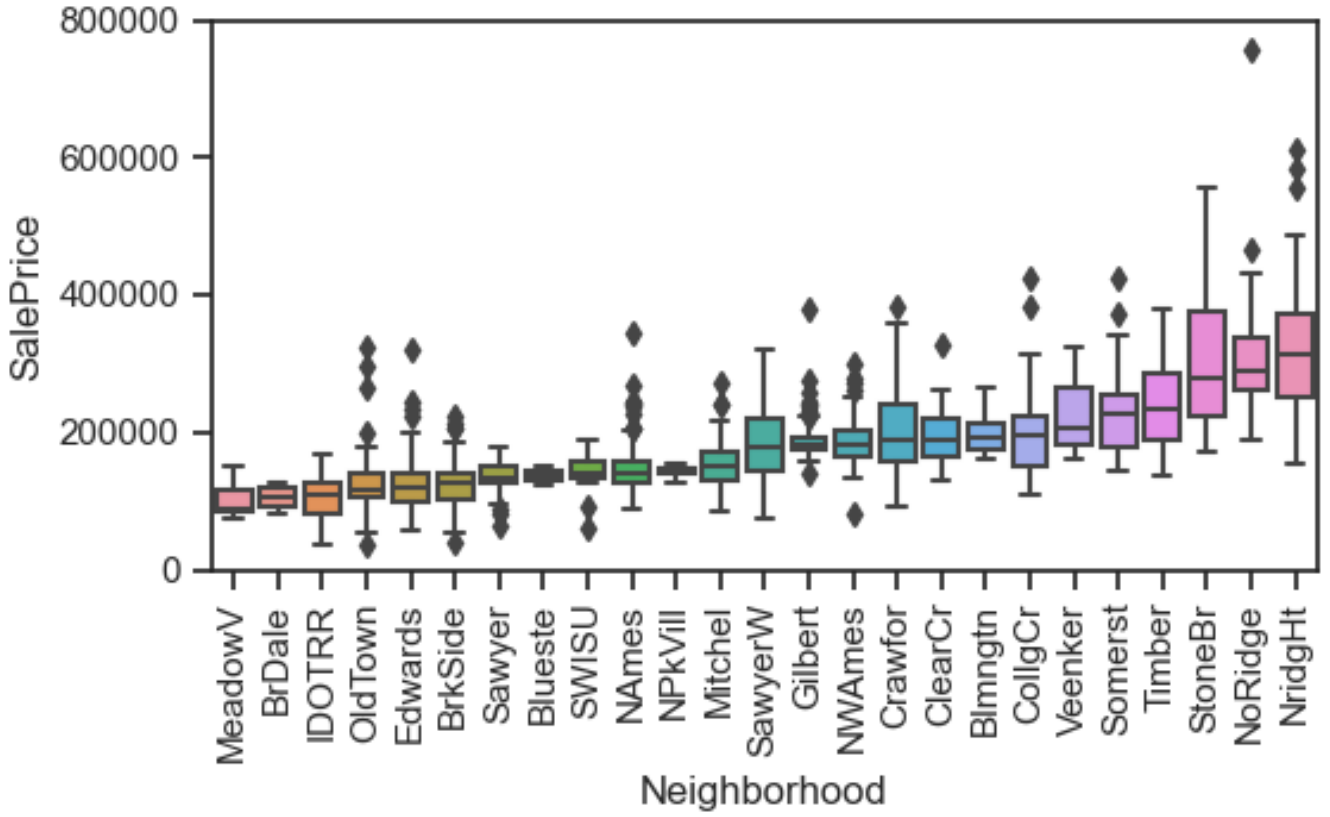


Análise bivariada

INSTITUTO FEDERAL
Maranhão

Softex

SalePrice em relação aos vários bairros



Análise bivariada

INSTITUTO FEDERAL
Maranhão

Softex

Comentários:

OverallQual', 'GrLivArea' e 'TotalBsmtSF‘ parecem influenciar fortemente 'SalePrice'.

'GarageCars' and 'GarageArea' são fortemente correlacionadas entre si. Recomenda-se manter 'GarageCars' pela correlação com 'SalePrice'

'TotalBsmtSF' e '1stFloor' são fortemente correlacionadas. Recomenda-se manter 'TotalBsmtSF'

'YearBuilt' parece ser fracamente correlacionado com 'SalePrice'.

Análise Multivariada

INSTITUTO FEDERAL
Maranhão

Softex

Para esta etapa pode-se utilizar

- Algoritmo de regressão linear, regressão logística ou mesmo árvores de decisão
- Analisar o peso de cada atributo nestes algoritmos e levantar hipóteses sobre sua importância para variável alvo

Próximas etapas

Testes de hipótese

Testar as várias hipóteses levantadas sobre quais atributos mais influenciam o SalePrice

Ex: T-test, ANOVA, etc.

Aprendizagem de máquina

SalePrice talvez possa ser predito por um algoritmo de regressão.

Conclusão

Resumo

Análise Exploratória de Dados é a etapa principal de compreensão dos dados

Nessa aula tivemos uma visão geral de como aplicar todos os conhecimentos de manipulação e visualização para encontrar padrões e ajudar a tomada de decisão

Os produtos dessa etapa incluem: relatório com os principais achados, notebooks, scripts, figuras, e dados transformados

A próxima etapa é a modelagem de dados com técnicas de aprendizagem de máquina ou criação de produtos de tais como dashboards coorporativos

Referências

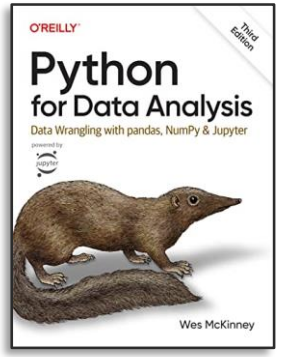
Livros

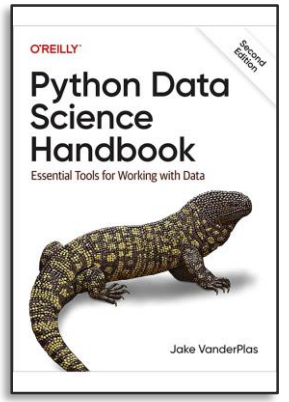
McKINNEY, Wes. "Python for Data Analysis". O'Reilly, 2017.

VANDERPLASS, Jake. "Python Data Science Handbook". O'Reilly, 2016.

HARRISON, Matt. “Machine Learning: guia de referencia rápida”. Novatec, 2020.

CHEN, Daniel Y. “Analise de Dados com Python e Pandas”. Novatec, 2018.





Links

https://towardsdatascience.com/exploratory-data-analysis-in-python-a-step-by-step-process-d0dfa6bf94ee