
# Licenciatura de Engenharia Informática

# Ciência de Dados

# Data Analysis
# Regressão Linear

**Executado por:**
nº 2405 Tiago Cardoso

**Orientado por:**
Ricardo Ferreira

**Entregue em:**
09/06/2022

# Índice

# Introdução

O objetivo deste trabalho é analisar e estudar grandes quantidades de informação utilizando a regressão linear para descrever a relação entre duas das variáveis estudadas.

```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
from scipy import stats
def myfunc(x):
  return slope * x + intercept


df = pd.read_csv("Stores.csv")
df.drop(df.columns[[0]], axis = 1, inplace = True)  # remove o id
sns.set()
sns.pairplot(df, diag_kind='kde', kind='reg', height = 3)
plt.show(block = True)
corrmat = df.corr()
plt.subplots(figsize = (5,5))
sns.heatmap(corrmat, vmax=.8, linewidths=0.01, square=True, annot=True, linecolor="white", annot_kws = {'size':12})
plt.show(block = True)
x = df["Store_Area"]
y = df["Items_Available"]
slope, intercept, r, p, std_err = stats.linregress(x, y)
mymodel = list(map(myfunc, x))
plt.scatter(x, y)
plt.plot(x, mymodel, color="r")
plt.show(block = True)

print("B0 =", slope)
print("B1 =", intercept)
print("Erro =", std_err)
```

Figura 1 - Código Python utilizado na análise dos dados no ficheiro “Stores.csv”

1. São importadas as bibliotecas necessárias para a realização do trabalho.
2. O programa lê o ficheiro csv e remove a coluna que contém o id.
3. São criados um pairplot (Figura 2) e um heatmap (Figura 3).
4. Analisa-se o heatmap e seleciona-se as variáveis a utilizar (neste caso selecionou-se a área da loja (x) e a quantidade de artigos disponíveis (y)).
5. É criado um gráfico scatter (Figura 4) com as variáveis selecionadas no ponto 4.
6. É utilizada a função “myfunc” na criação de uma linha estimada de regressão (Figura 4).
7. São apresentados os valores de β0, β1 e ε

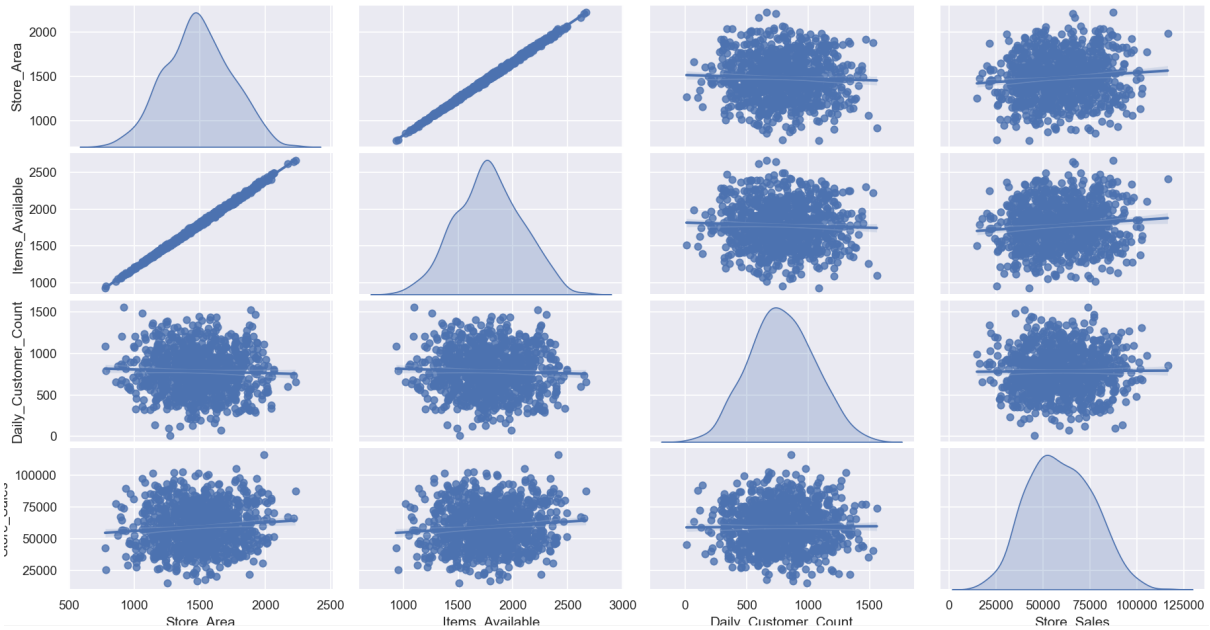


Figura 2 - Gráfico pairplot

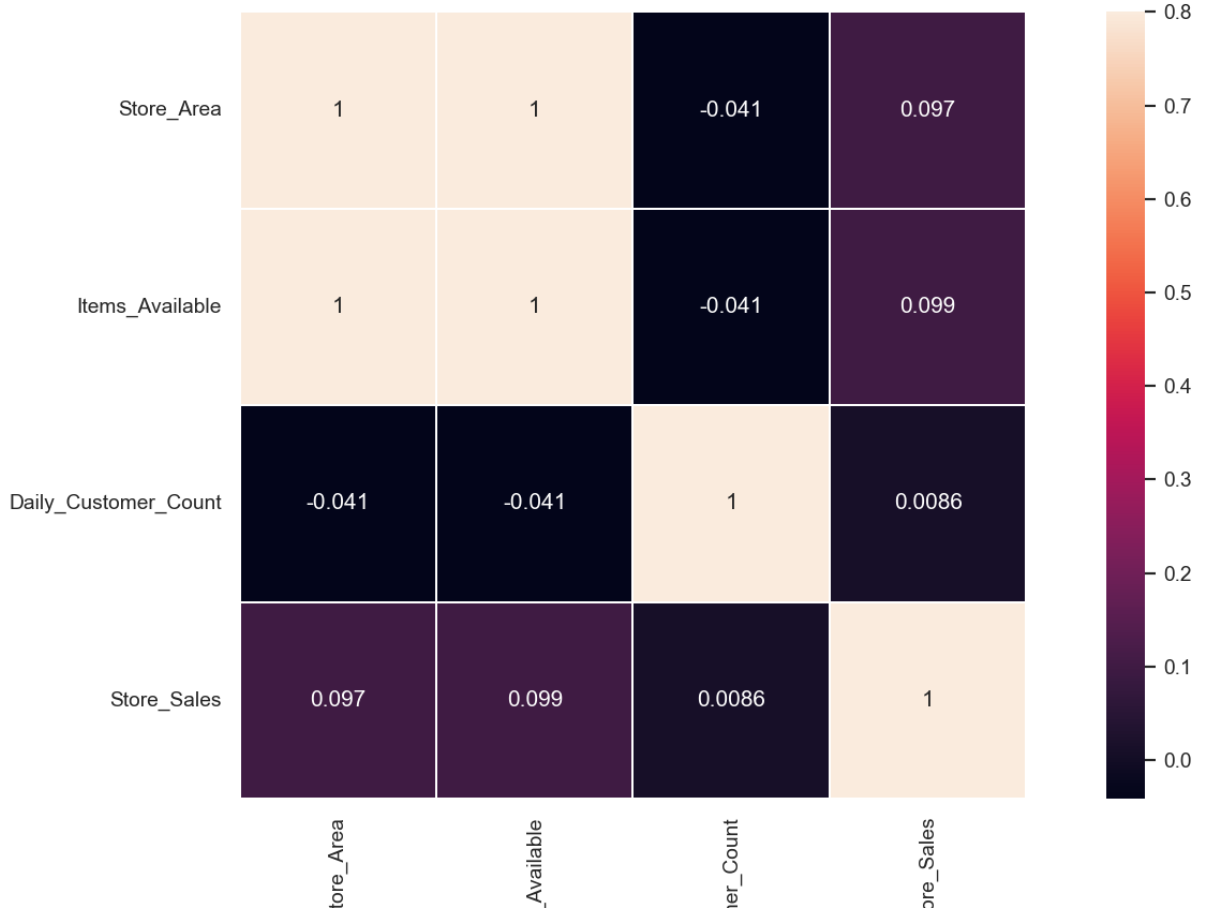


Figura 3 - Gráfico heatmap

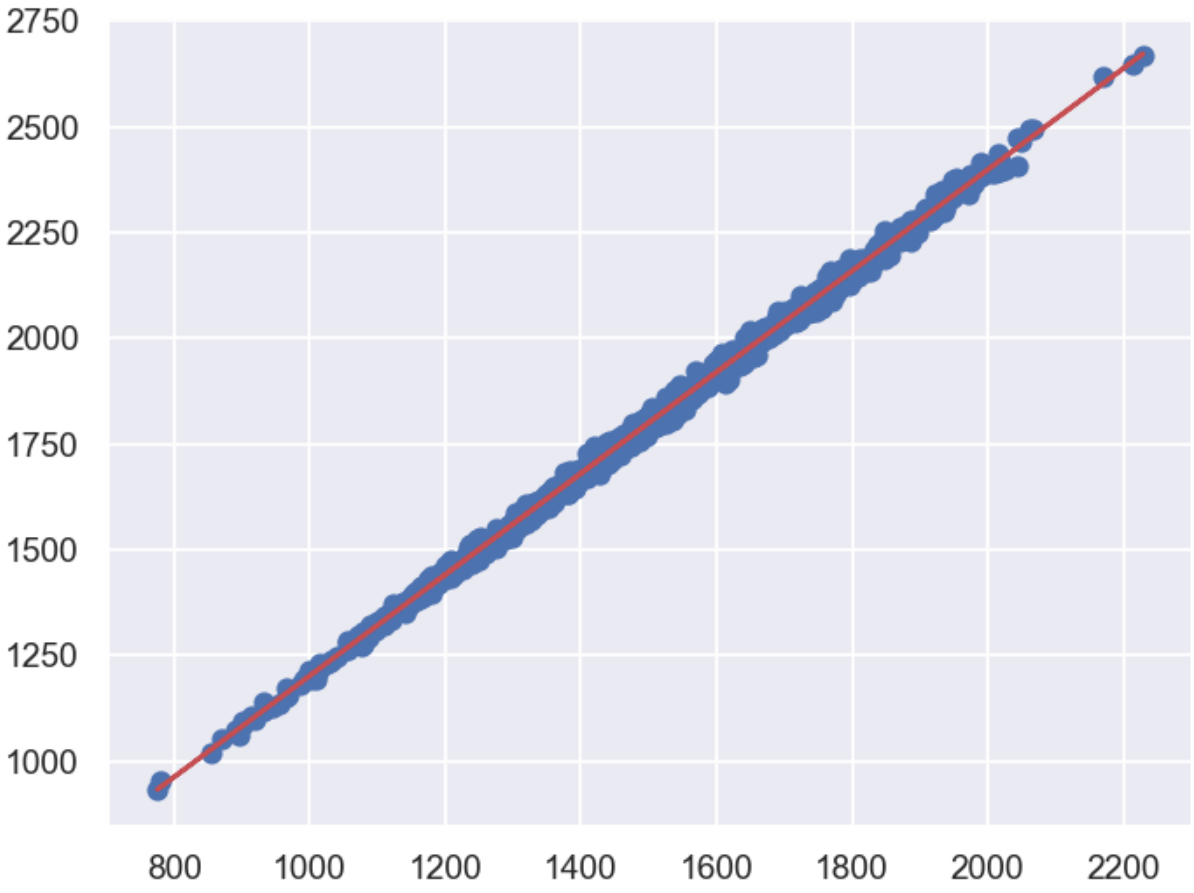


Figura 4 - Gráfico Scatter e o valor esperado de regressão representado por uma linha vermelha

# Conclusão

Este projeto permitiu-nos estudar e analisar uma grande quantidade de dados utilizando o Python, também foi possível estudar a relação entre variáveis e como estas se influenciam umas às outras.

Algumas das dificuldades sentidas na realização deste projeto foi a utilização de novas ferramentas no Python (ex: seaborn e o scipy) que nos permitiram analisar os dados mais facilmente.