chapter_01

April 1, 2022

1 0.0. Imports

1.1 0.1. Julia & Python Imports

[78]:
```
using CSV;
using PyCall;
using PyPlot;
using Printf;
using DataFrames, FreqTables
using HypothesisTests

include("../scripts/utils_1.jl")
pd = pyimport("pandas")
np  = pyimport("numpy");
sns = pyimport("seaborn");
```

[2]:
```
wrg = pyimport("warnings");
wrg.filterwarnings("ignore")
```

2 1.0. Capítulo 1

2.1 1.1. Correlação

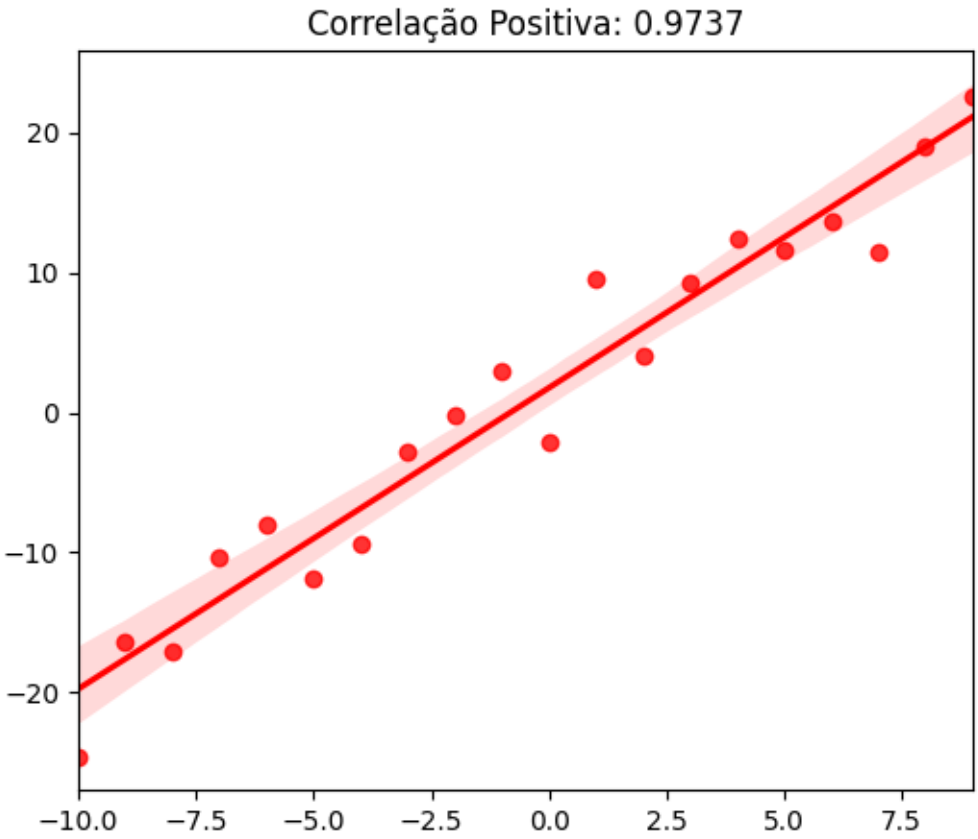
2.1.1 1.1.1. R de Pearson

O coeficiente de correlação de pearson é muitas vezes o primeiro coeficiente estudado ou abordado em livros. São ditos os dados que são positivamente correlacionados quando os valores de **x** acompanham os valores de **y** e negativamente correlacionados se os valores altos de **x** acompanharem os valores baixos de **y**.**Causalidade** a variável **x** é a causa da variável **y**, logo por exemplo a correlação entre número de vendas e clientes é positiva, mas não quer dizer que quantos mais clientes existem mais vendas eu tenha.Ex: O número de consumo de **margarina** e o número de **divórcios** em Maine.

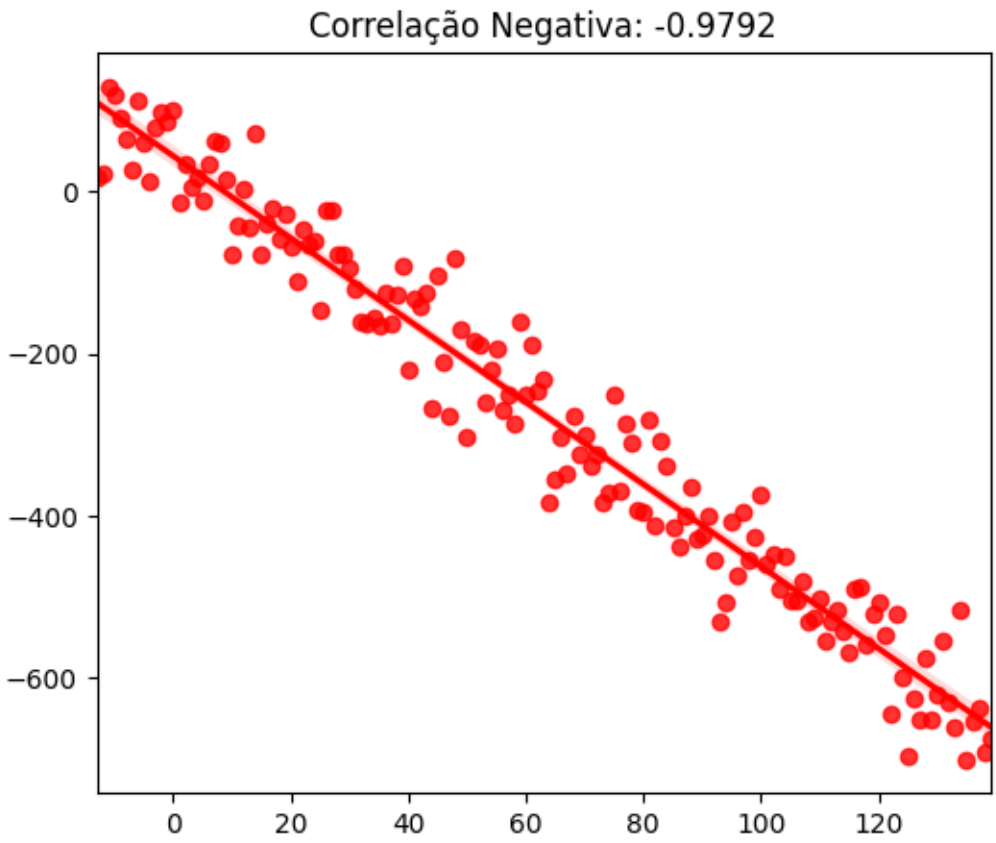
Fórmula do coeficiente de pearson.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y))}{\sqrt{(n\sum(x^2) - (\sum x)^2)\ (n\sum(y^2) - (\sum y)^2)}}$$

[191]:
```
plot_linear(2, 2, 3, -10, 10);
```



[8]:
```
plot_linear(-5, 40, 50, -13, 140);
```

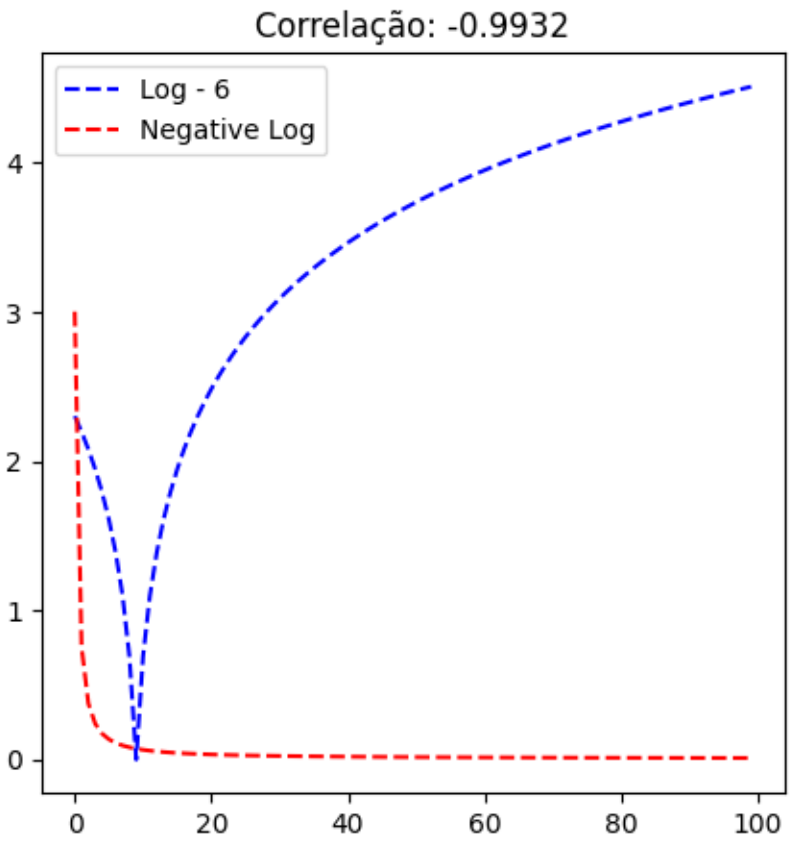


2.1.2 1.1.2. Rho de spearman

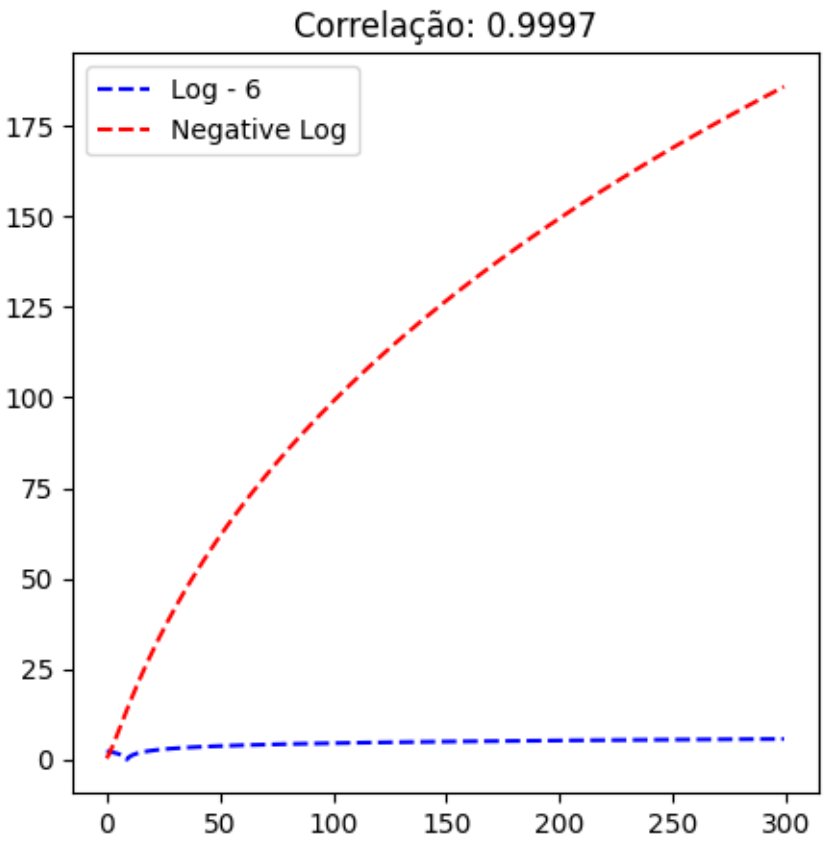
Robusto contra outliers e calculado em relação ao ranqueamento ou ordens dos dados, também mede relações lineares e não lineares.

$$r_s = 1 - \frac{6 \sum d^2}{n^3 - n}$$

[96]:
```
spearman_plot( 100, -3 );
```



[97]:
```
spearman_plot( 300, 3 );
```



2.1.3 1.1.4. V de Cramér

O V de cramér basicamente serve para calcular a correlação entre variaveis categoricas partindo da frequência.

Existe a versão corrigida da fórmula de cramér que esta abaixo, k e r são as dimensões da matriz.

$$V = \sqrt{\frac{\varphi^2 \ ou \ X^2/n}{min(k-1, r-1)}}$$

$$\varphi^2 = max(0, \varphi^2 - \frac{(k-1)-(r-1)}{n-1}$$

$$cor \ k = k - \frac{(k-1)^2}{n-1}$$

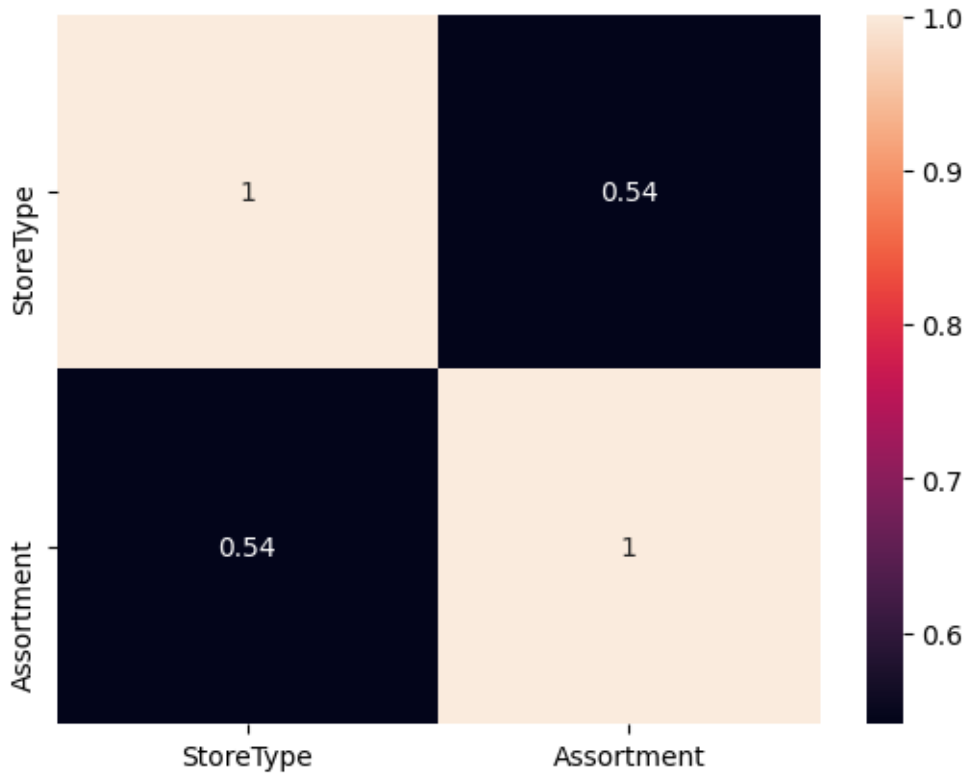
$$cor \ r = r - \frac{(r-1)^2}{n-1}$$

[75]:
```
df = DataFrame(CSV.File("../data/store.csv"))
df = df[:, 2:3]

df2 = cramer_simple_corr(df)
df2 = pd.DataFrame( Array(df2), columns=names(df2), index=names(df2) )
```

[75]:
```
PyObject             StoreType  Assortment
StoreType    1.001348    0.540680
Assortment   0.540680    1.000898
```

[76]:
```
sns.heatmap( df2, annot=true );
```

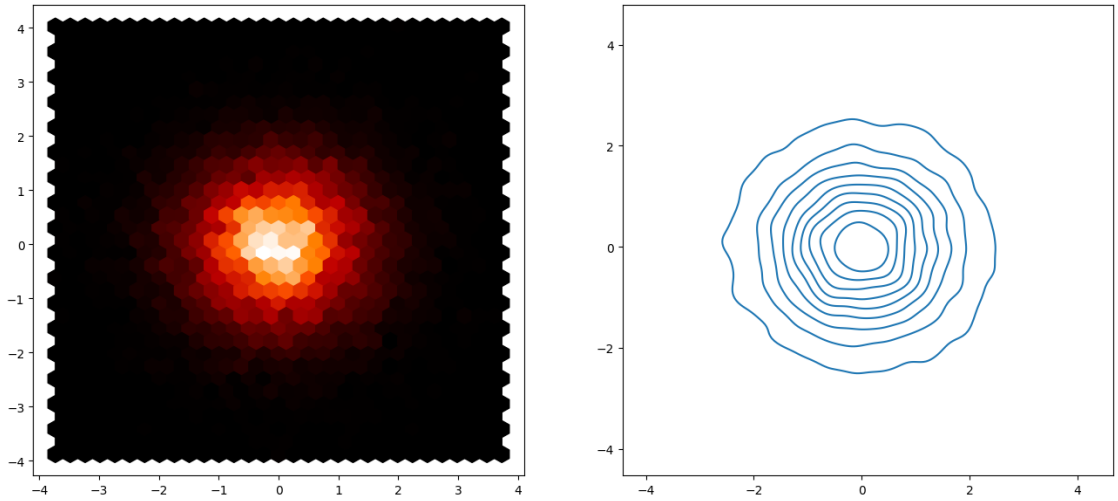


Não existe uma grande correlação entre o tipo de loja e seu sortimento.

2.2 1.2. Dois Gráficos de Densidade

Hexagonal Binning relaciona as duas variaveis aleatorias normais em hexágonos, mesma coisa que o Histograma.Kernel Density Estimate, Análogo análogo ao Hexagonal, porem em densidades com curvas.

[160]:
```
plot_density( 20000, 20000 )
plt.savefig("Density.png")
```



3 x.0. Referências

PETER BRUCE & ANDREW BRUCE **Estatística prática para cientistas de dados: 50 conceitos essenciais**. Link: https://www.amazon.com.br/Estat%C3%ADstica-Pr%C3%A1tica-Para-Cientistas-Dados/dp/855080603X DAVID MATOS **8 Conceitos Estatísticos Fundamentais Para Data Science**. Link: https://www.cienciaedados.com/8-conceitos-estatisticos-fundamentais-para-data-science/IGOR SOARES **Correlação não implica em Causalidade**.Link: https://medium.com/@felipemaiapolo/correla%C3%A7%C3%A3o-n%C3%A3o-implica-em-causalidade-8459179ad1bc.annahaensch **Número de Casos de Divórcio em Maine**Link: https://blogs.ams.org/blogonmathblogs/2017/04/10/divorce-and-margarine/