

IMD0033 - Probabilidade

Aula 20 - Correlação e covariância

Ivanovitch Silva
Maio, 2018



Agenda

- Correlação e covariância
- Coeficiente de correlação

Atualizar o repositório

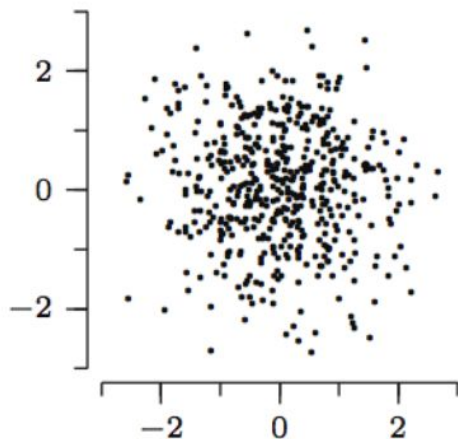
```
git clone https://github.com/ivanovitchm/imd0033_2018_1.git
```

Ou

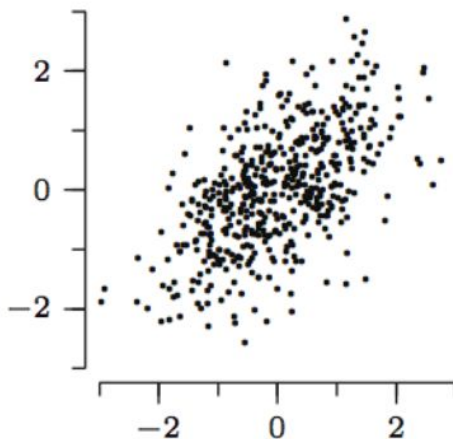
```
git pull
```

Tipos de correlação e intensidade

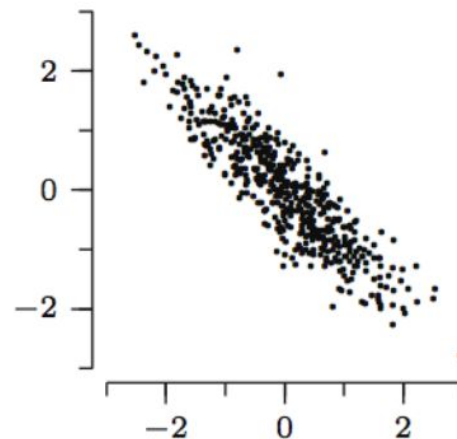
Gráficos de Dispersão



Não correlacionado

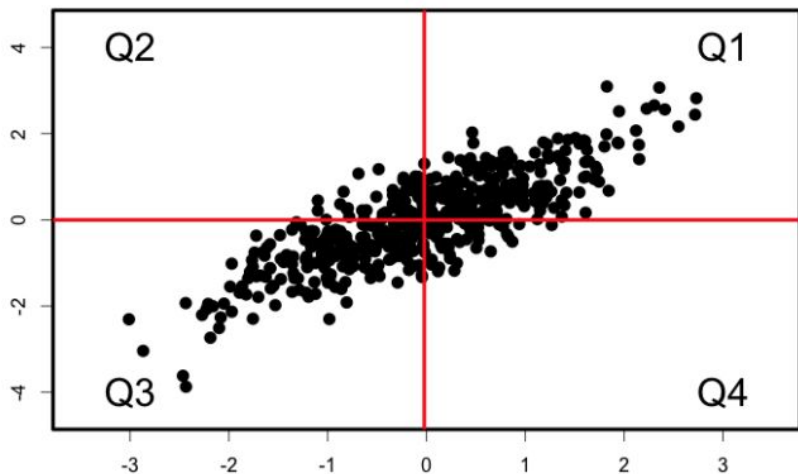


Correlação positiva, fraca



Correlação negativa, forte

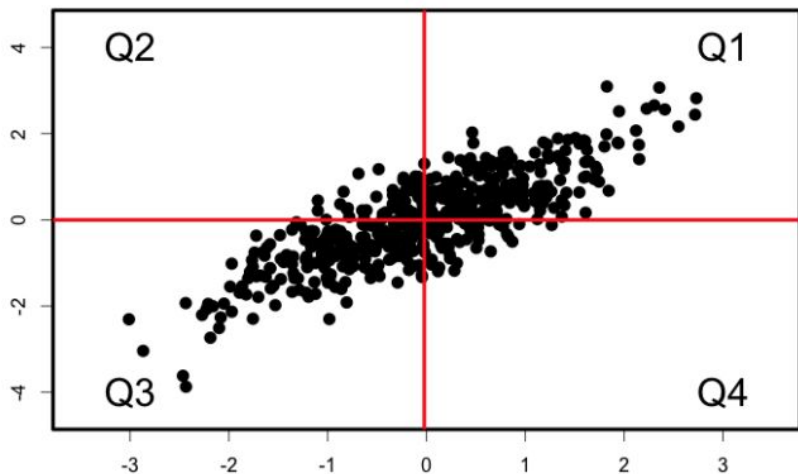
Analizando os gráficos de dispersão



É possível incluir no gráfico de dispersão as retas vertical e horizontal que passam, respectivamente, pelas médias amostrais \bar{x} e \bar{y}

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{e} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Analizando os gráficos de dispersão



$(x_i - \bar{x})$ Desvio da média para a variável x

$(y_i - \bar{y})$ Desvio da média para a variável y

$(x_i - \bar{x})(y_i - \bar{y})$ Produto dos desvios

Como se comporta o produto dos desvios para Q1, Q2, Q3, Q4?

Covariância

Covariância é uma medida que reflete a forma como duas variáveis variam conjuntamente.

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{n}$$

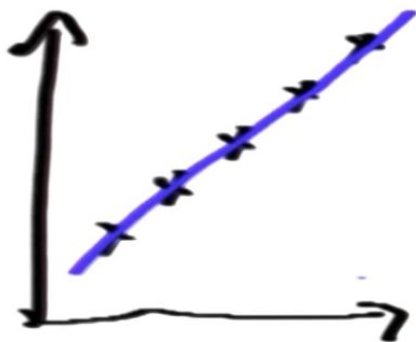
Coeficiente de correlação

A covariância, no entanto, não fornece uma medida da intensidade da relação, já que depende das unidades em que as variáveis são expressas. Uma maneira de contornar este problema é através da padronização dos dados

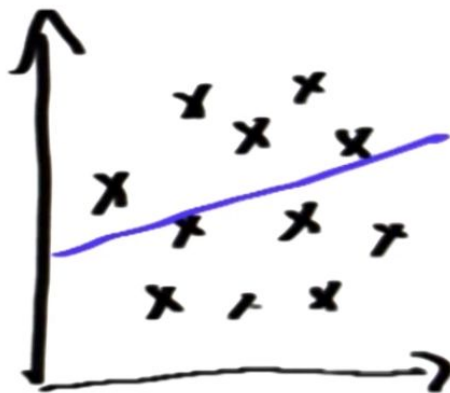
$$\frac{cov(\mathbf{x}, \mathbf{y})}{\sigma_x \sigma_y}$$

Coeficiente de correlação

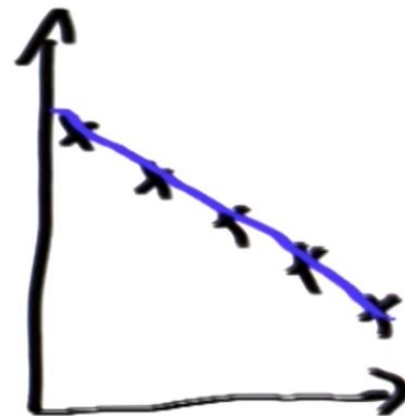
$\in [-1 \dots 1]$



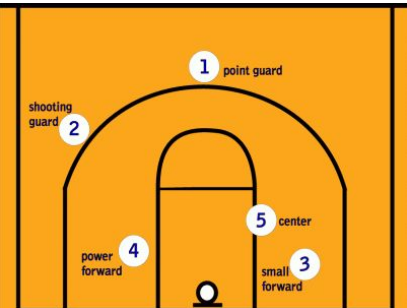
$\gamma = 1$



0



-1



Estudo de Caso

	player	pos	age	bref_team_id	g	gs	mp	fg	fga	fg.	...	drb	trb	ast	stl	blk	tov	pf	pts	season	sea
0	Quincy Acy	SF	23	TOT	63	0	847	66	141	0.468	...	144	216	28	23	26	30	122	171	2013-2014	
1	Steven Adams	C	20	OKC	81	20	1197	93	185	0.503	...	190	332	43	40	57	71	203	265	2013-2014	
2	Jeff Adrien	PF	27	TOT	53	12	961	143	275	0.520	...	204	306	38	24	36	39	108	362	2013-2014	
3	Arron Afflalo	SG	28	ORL	73	73	2552	464	1011	0.459	...	230	262	248	35	3	146	136	1330	2013-2014	
4	Alexis Ajinca	C	25	NOP	56	30	951	136	249	0.546	...	183	277	40	23	46	63	187	328	2013-2014	

g - número de jogos

gs - jogos como titular

mp - minutos jogados/partida

fg - lançamentos feitos

fga - tentativas de lançamentos

fg. - eficiência

drb - rebotes defensivos

trb - total de rebotes

ast - assistência por jogo

stl - roubadas de bola

pf - faltas pessoais

pts - pontos

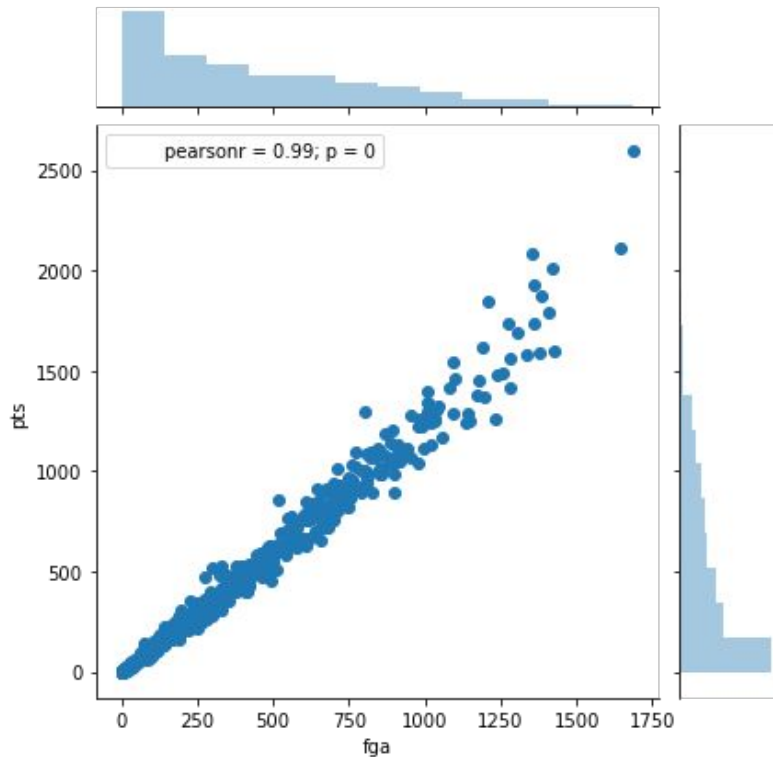
<http://stats.nba.com/help/glossary/>

Pearsonr

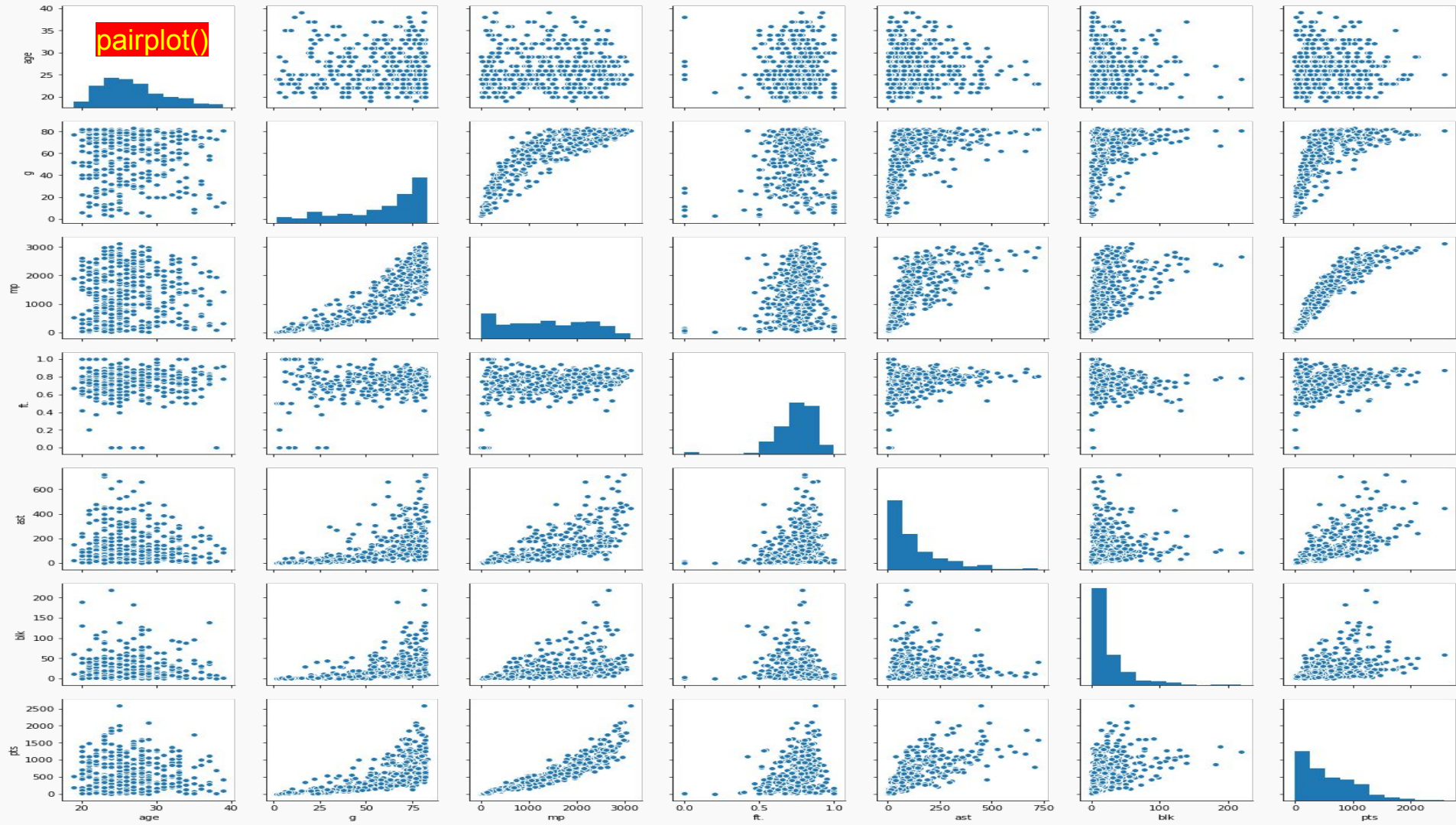
```
from scipy.stats.stats import pearsonr
```

```
# The pearsonr function will find the correlation between two columns of data.  
# It returns the r value and the p value. We'll learn more about p values later on.  
r, p_value = pearsonr(nba["fga"], nba["pts"])
```

Seaborn Jointplot

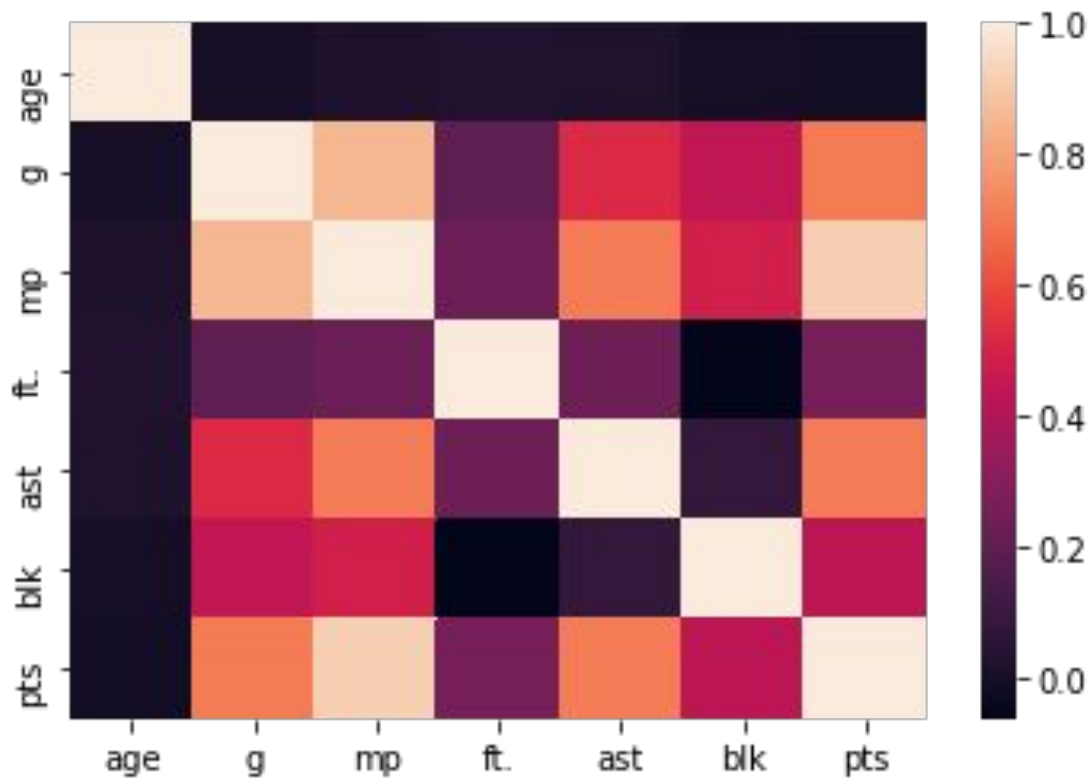


```
# Draw a plot of two variables  
# with bivariate and univariate graphs.  
sns.jointplot(x=nba["fga"], y=nba["pts"])
```

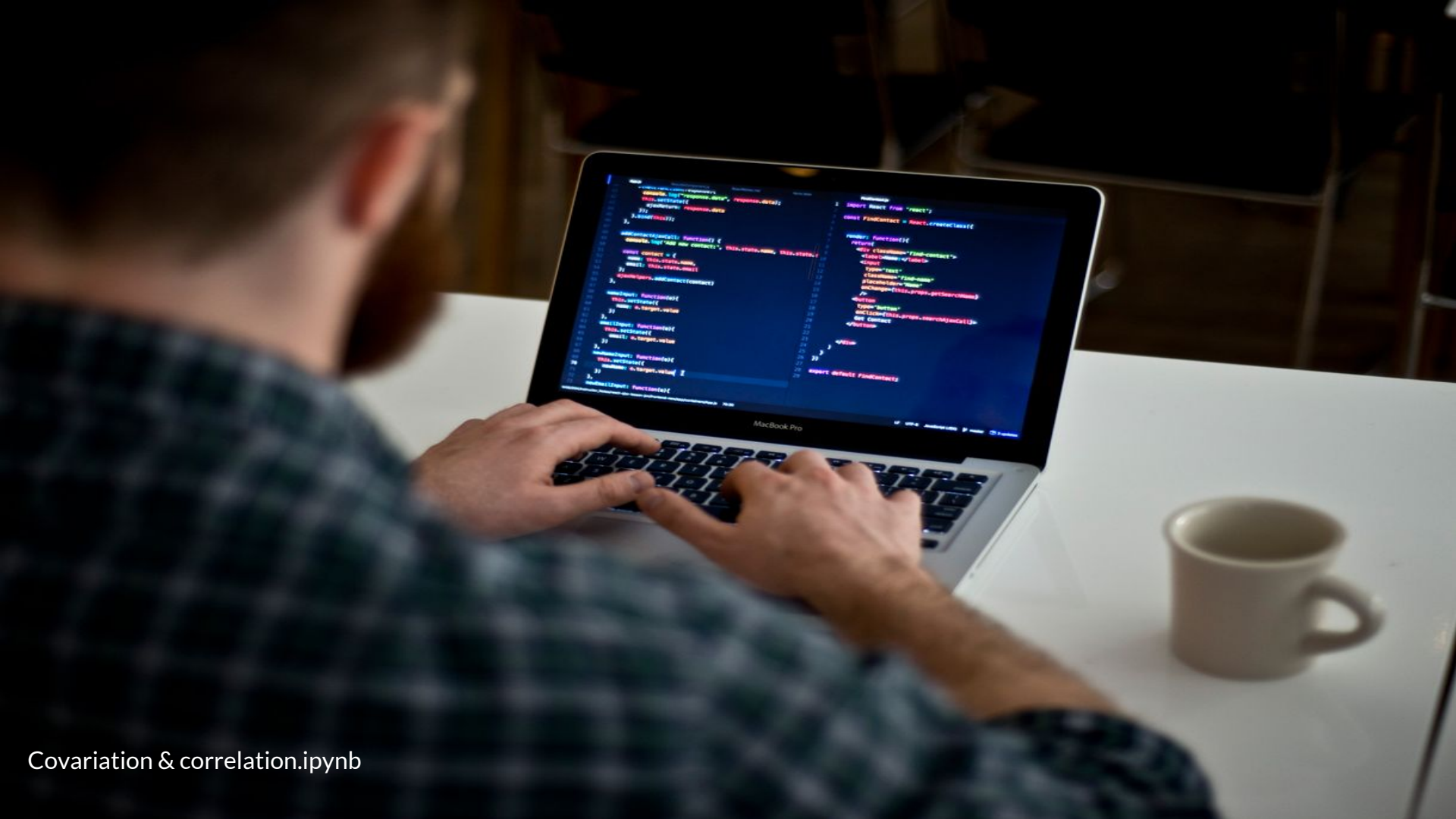


```
columns = ['age', 'g', 'mp', 'ft.', 'ast', 'blk', 'pts']  
nba[columns].corr()
```

	age	g	mp	ft.	ast	blk	pts
age	1.000000	0.003149	0.019843	0.033372	0.026157	0.001864	-0.007520
g	0.003149	1.000000	0.855091	0.198547	0.520201	0.444877	0.708630
mp	0.019843	0.855091	1.000000	0.232772	0.711095	0.489242	0.920194
ft.	0.033372	0.198547	0.232772	1.000000	0.235162	-0.060122	0.258744
ast	0.026157	0.520201	0.711095	0.235162	1.000000	0.083110	0.710765
blk	0.001864	0.444877	0.489242	-0.060122	0.083110	1.000000	0.432895
pts	-0.007520	0.708630	0.920194	0.258744	0.710765	0.432895	1.000000



```
sns.heatmap(nba[columns].corr())
```



<http://globoesporte.globo.com/cartola-fc/>

<https://github.com/henriquepgomide/caRtola>