Data Scientist - Análise estatística preliminar

Soraia Pereira[a], Tiago Marques[a,b]

[a] CEAUL e FCUL, Universidade de Lisboa
[b] CREEM, University of St Andrews, e Dept de Biologia Animal, FCUL

FCUL, 4 de fevereiro de 2020

House Keeping

Recursos do curso disponíveis na pasta

https://tinyurl.com/CEAULGADESCursoRM2

O curso decorre entre as 18:30 e as 22:30

Entre as 20:30 e as 20:45 faremos uma pausa para café.

Introdução

- A análise exploratória de dados é um passo crucial em qualquer análise de dados. O objetivo é a compreensão dos dados, desde a sua estrutura, distribuição e identificação de potenciais associações entre variáveis.
- O procedimento para esta análise depende daquilo que se pretende investigar, não sendo um processo com regras rígidas comuns a todo o tipo de análise. A melhor forma de o fazer é colocar questões de forma a guiar a investigação. Isso irá determinar o foco da nossa atenção e ajudar a decidir quais as ferramentas mais adequadas para responder a essas questões.

Introdução

Grolemund e Wickham descrevem a análise exploratória de dados como um processo iterativo com os seguintes passos:

1. Colocar questões sobre os dados
2. Procurar respostas usando visualização, manipulação, modelação
3. Refinar questões com base na nova informação, e colocar novas questões (voltar ao passo 1)

Qualquer análise de dados depende naturalmente do tipo de variável/variáveis que se pretende analisar.

Tipos de variáveis

As variáveis podem ser classificadas em dois grandes grupos: quantitativas e qualitativas:

- Quantitativas: podem ser medidas numa escala quantitativa. Podem ser discretas ou contínuas.
  - Discretas: podem assumir valores num conjunto finito ou infinito numerável. Exemplo: número de gatos, número de bactérias, número de desempregados.
  - Contínuas: podem assumir valores num intervalo do conjunto $\Re$. Exemplo: tempo, altura, taxa de desemprego.
- Qualitativas: são definidas em categorias. Podem ser nominais ou ordinais.
  - Nominais: não existe ordenação entre as categorias. Exemplo: sexo, região, estado do mercado de trabalho.
  - Ordinais: existe ordenação entre as categorias. Exemplo: nível de educação, grupo etário, posição obtida num recrutamento.

Visualização gráfica

- Um dos packages mais elegantes e versáteis para visualização gráfica no R é o `ggplot2`. Este package é um dos membros do `tidyverse`, um package um conjunto de packages úteis em qualquer análise de dados.
- Para instalação do `tidyverse`:
  `install.packages("tidyverse")`
- Como em qualquer package, a instalação é necessária apenas uma vez, mas o package deve ser carregado sempre que se inicia nova sessão do R.
  `library(tidyverse)`

Ilustração com dados do package gapminder

```
library(gapminder)
```

```
> head(gapminder)
# A tibble: 6 x 6
  country     continent  year lifeExp      pop gdpPercap
  <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
1 Afghanistan Asia       1952    28.8  8425333      779.
2 Afghanistan Asia       1957    30.3  9240934      821.
3 Afghanistan Asia       1962    32.0 10267083      853.
4 Afghanistan Asia       1967    34.0 11537966      836.
5 Afghanistan Asia       1972    36.1 13079460      740.
6 Afghanistan Asia       1977    38.4 14880372      786.
```

Função summary

```
> summary(gapminder)
        country        continent        year         lifeExp          pop
 Afghanistan:  12   Africa  :624   Min.   :1952   Min.   :23.60   Min.   :6.001e+04
 Albania    :  12   Americas:300   1st Qu.:1966   1st Qu.:48.20   1st Qu.:2.794e+06
 Algeria    :  12   Asia    :396   Median :1980   Median :60.71   Median :7.024e+06
 Angola     :  12   Europe  :360   Mean   :1980   Mean   :59.47   Mean   :2.960e+07
 Argentina  :  12   Oceania : 24   3rd Qu.:1993   3rd Qu.:70.85   3rd Qu.:1.959e+07
 Australia  :  12                  Max.   :2007   Max.   :82.60   Max.   :1.319e+09
 (Other)    :1632
   gdpPercap
 Min.   :   241.2
 1st Qu.:  1202.1
 Median :  3531.8
 Mean   :  7215.3
 3rd Qu.:  9325.5
 Max.   :113523.1
```

Função tapply

```
> length(unique(gapminder$country))
[1] 142
> tapply(gapminder$lifeExp,gapminder$continent,mean)
  Africa Americas     Asia   Europe  Oceania
48.86533 64.65874 60.06490 71.90369 74.32621
> gapminder_2007<-filter(gapminder, year=="2007")
> tapply(gapminder_2007$lifeExp,gapminder_2007$continent,mean)
  Africa Americas     Asia   Europe  Oceania
54.80604 73.60812 70.72848 77.64860 80.71950
```

Histograma

```
ggplot(gapminder, aes(lifeExp)) + geom_histogram()
```

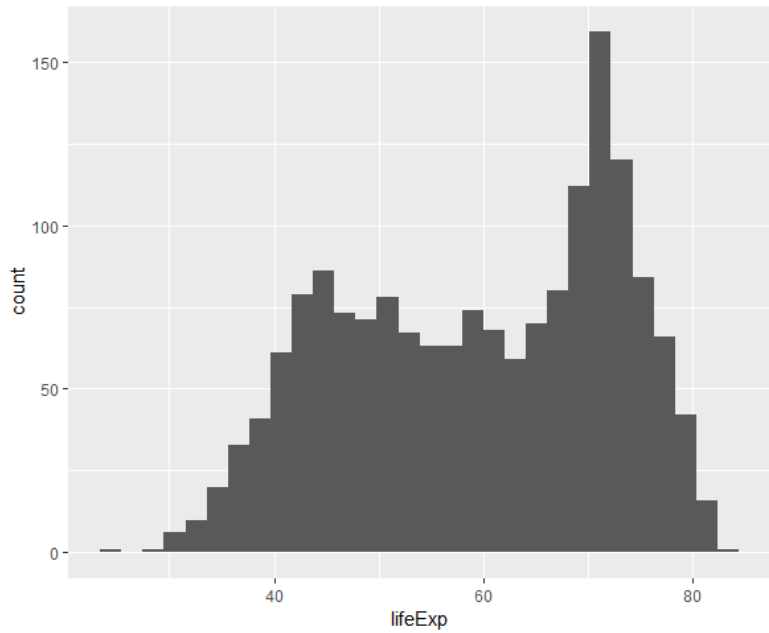
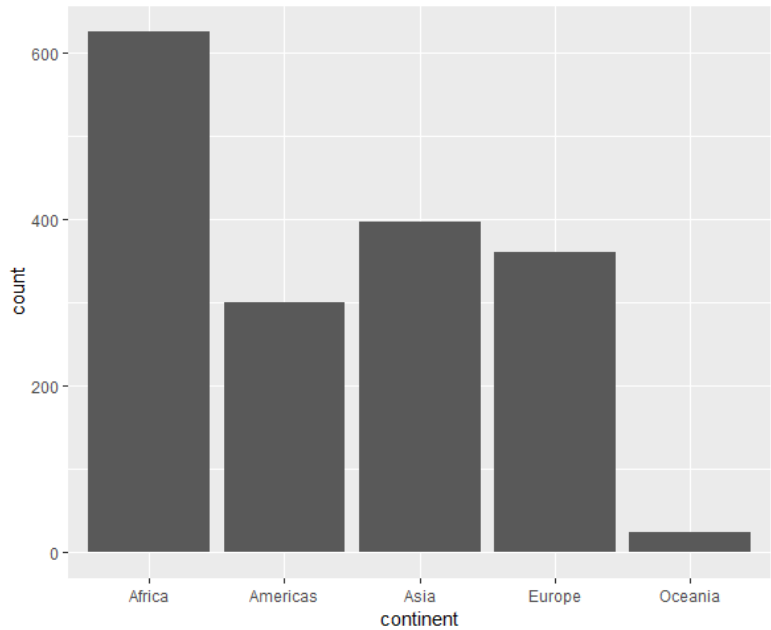


Diagrama de barras

```
ggplot(gapminder, aes(x = continent)) + geom_bar()
```



Boxplot

```
ggplot(gapminder, aes(x="",y=lifeExp)) + geom_boxplot()
```

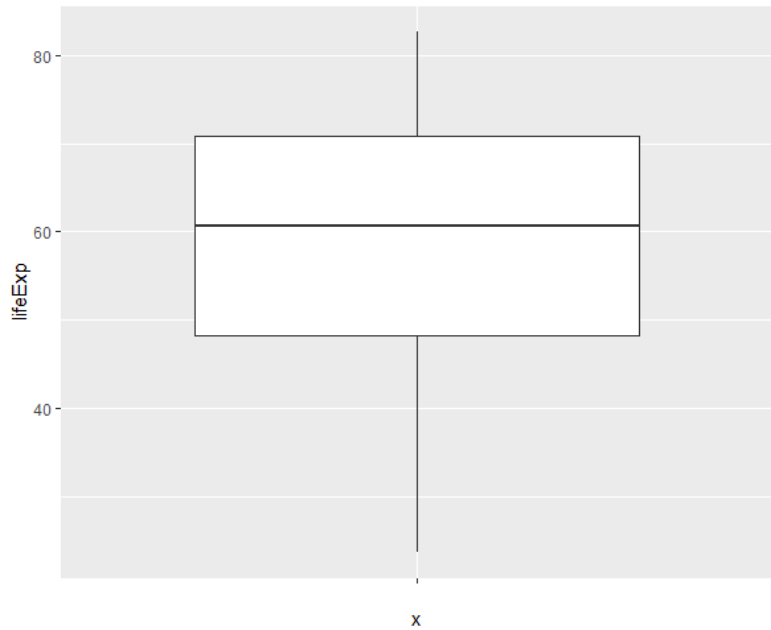
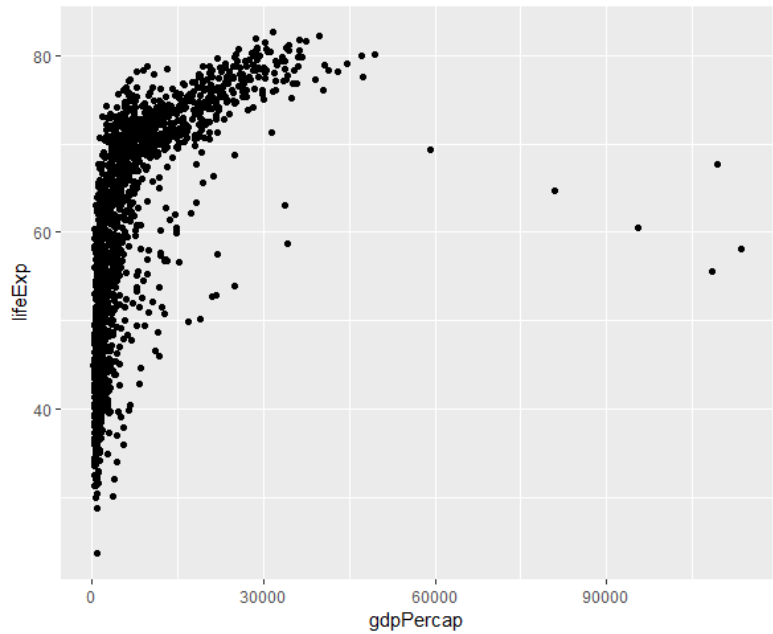


Gráfico de pontos

```
ggplot(data = gapminder, aes(x = gdpPercap, y = lifeExp))
+ geom_point()
```



Heatmap

```
library(broom)

CorMatrix <- broom::tidy(cor(gapminder[, 3:6])) %>%
  rename(Var1 = ".rownames") %>%
  gather(Var2, Cor, -Var1)

ggplot(CorMatrix, aes(Var1, Var2, fill = Cor)) +
  geom_tile()
```