Ciência dos Dados. Projeto Fase A - № 1

Ricardo Sousa e Manuel Pita, Universidade Lusófona. November 20, 2022

Introdução

O objetivo deste projeto prático é implementar os métodos de análise de dados e testes estatísticos estudados na cadeira através do estudo de um conjunto de dados real sobre um assunto atual e que interessa a todas as pessoas do mundo, a pandemia da Covid-19.

Juntamente com este enunciado encontrarás dois ficheiros: o primeiro é `covid-data.csv` que será a base do trabalho que irás desenvolver, individualmente ou em grupo, e o segundo é `covid.ipynb`, um caderno Jupyter que deverás utilizar como base para o teu projeto. Não deves alterar a estrutura do caderno, mas sim preencher as células com o código, funções e texto *markdown* pedidos nas perguntas deste enunciado.

O primeiro passo que deves seguir é entender o conjunto de dados. Este passo é essencial para que possas formar uma ideia clara sobre se os dados são ou não uma estrutura retangular, no caso positivo, como é que cada fila está identificada, se os dados numa determinada coluna são consistentes com o correspondente título e outros aspetos.

A primeira seção do caderno Jupyter está destinada unicamente a importar bibliotecas ou dados, cada uma destas componentes numa célula separada. No resto deste enunciado, quando encontres a etiqueta `TO-DO` no início dum parágrafo significa que o mesmo contém instruções que deverás implementar no teu código. No correspondente caderno Jupyter encontrarás texto que te permitirá identificar facilmente onde deves responder a cada item `TO-DO`.

Pré-processamento

`TO-DO P1:` deverás importar a biblioteca `pandas` e logo o conjunto de dados, garantindo que o índice do *dataframe* seja a coluna zero, a qual identifica cada uma das filas com um único número desde o zero até o tamanho do conjunto de dados menos um. Utiliza o nome `df` para o *dataframe*.

`TO-DO P2:` documentar as colunas do *dataframe*. Os nomes das colunas são auto-explicativos. Nota que os valores de algumas das colunas são variáveis mas outros são constantes, dependendo da localidade.

`TO-DO P3:` verifica que o tipo de dados da coluna `date` não é um `datetime`. Chama a função da biblioteca `pandas` que permite converter uma coluna para que a mesma seja do tipo `datetime`.

Análise de dados

`TO-DO A1:` Explora o conteúdo da coluna *location*. Os valores que observas nesta coluna são todos da mesma categoria (por exemplo países, ou continentes)?

`TO-DO A2:` Será que termos valores que não são da mesma categoria poderá causar problemas na análise dos dados?

`TO-DO A3:` Vamos focar unicamente no subconjunto dos dados correspondente aos seguintes países: Áustria, França, Espanha e Portugal. Cria um *dataframe* `dfx` que contém estes dados.

`TO-DO A4:` Escreve uma função `get_basic_plots(country, var, opt)` que recebe o nome do país, e o nome duma coluna, e guarda as seguintes visualizações no diretório onde se encontra o caderno Jupyter:

1. `country-var-histogram.png`
2. `country-var-boxplot.png`

onde `country` deve ser o nome do país dado como argumento quando executamos a função e `var` o nome da variável. Neste caso só vamos considerar como válidas as variáveis `new_cases` e `new_deaths`. A variável `opt` será utilizada mais tarde na segunda versão desta função.

`TO-DO A5:` Escreve uma função `remove_outliers(df,col)` que recebe um dataframe `df` e devolve o mesmo *dataframe* mas com os outliers da coluna `col` eliminados utilizando a técnica do $1.5 \times$ IQR.

`TO-DO A6:` Editar a função `get_basic_plots(country, var, opt)` para que quando `opt=False` os gráficos sejam produzidos sem eliminação de *outliers*, e caso contrário (quando `opt=True`) os *outliers* na variável `var` sejam eliminados antes de gerar os gráficos. No caso em que removemos os *outliers*, adicionar os carateres `-wo` no nome dos ficheiros antes do `.png` para diferenciar os dois casos possíveis.

`TO-DO A7:` Produz os gráficos, com e sem *outliers*, para os quatro países foco do nosso estudo, mas só para os primeiros seis meses da pandemia. Considerando os histogramas, (a) será que temos alguma suspeita de que alguma das variáveis para algum dos países poderá ser Normal? e (b) observa os *outliers* nos *boxplots* e reflete sobre a razão pela qual "normalmente" removemos os *outliers* das nossas amostras. Será que neste conjunto de dados devemos eliminar? Ainda não estás 100% no nível de conhecimento necessário para responder a esta pergunta, por isso o que deves fazer, para já, com os recursos disponíveis, é gerar os gráficos todos e os valores das medianas e IQR para os cenários sem e com remoção dos *outliers*. Analiticamente, observas muita diferença nos resultados? Discute este assunto com o professor da teórica (o assunto é: será que sempre queremos eliminar *outliers*? mas pensa

sobre o assunto criticamente antes).

`TO-DO A8`: Tudo o que temos feito até agora é útil para a fase da análise univariada de dados. Quando queremos passar a fase bi-variada deparamo-nos com um problema: os números de casos de um país para outro não são diretamente comparáveis porque a população de cada um dos países é diferente. Aumenta o teu dataframe para incluir colunas `new_cases_ht` e `new_deaths_ht` que representem os valores das variáveis para novos casos e número de mortes por 100.000 habitantes (o `ht` significa *hundred thousand*). Utiliza todos os recursos que implementaste até agora para pensar critica e analiticamente sobre a situação nos quatro países durante os primeiros seis meses da pandemia. Documenta todas as tuas observações no caderno Jupyter, utiliza *markdown* para estruturar o teu texto.