PRO3200 – EPUSP
Professora Celma de Oliveira Ribeiro
Assistente: Pedro Gerber Machado & Monitor: Vinícius Castanho

Estudo de Caso 0: Uso do R

A disciplina PRO3200-1º Semestre de 2023 utilizará a linguagem **R**. **R** é um ambiente computacional e uma linguagem de programação que vem progressivamente se especializando em manipulação, análise e visualização gráfica de dados. Sugerem-se as seguintes referências para uma abordagem inicial:

https://cran.r-project.org/doc/contrib/Landeiro-Introducao.pdf

https://rpubs.com/EstatBasica

https://cran.r-project.org/web/packages/HSAUR/vignettes/

http://www.r-tutor.com/elementary-statistics

Sugere-se o uso do R-Studio como ambiente para implementação dos códigos.

https://rstudio.com/

Este estudo de caso contém algumas indicações sobre comandos básicos e tem o intuito de familiarizar os inscritos no curso com o ambiente de programação.

A disciplina **não tem** como objetivo o ensino da linguagem R! O objetivo é desenvolver o ferramental de **estatística** com o suporte dessa linguagem. Porém todos os exercícios **devem ser feitos em R** para que haja organização tanto em sala de aula quanto na entrega dos trabalhos. Adicionalmente, tentaremos apresentar o máximo de informações quanto ao R nos estudos de casos que vocês fizerem, então caso você perca o prazo de um estudo de caso, além de perder nota, você pode também perder informações importantes para o próximo case, então sempre tente fazer os cases em dia, se não der, faça o estudo de caso do mesmo jeito. Não deixe de estudar todos eles.

Assume-se que os alunos tenham tido contato prévio com linguagens de programação o que viabiliza o uso da plataforma R.

No que se refere à parte conceitual da Estatística abordada nesse estudo de caso e nos próximos, sugerem-se as seguintes referências:

*Notas de aula (pdf. material das professoras Celma e Linda) -> Disponiível e-disciplinas

*DEVORE, J.L.; Probabilidade e Estatística para Engenharia e Ciências, Editora Thomson

*Outros livros empregados nas turmas de PRO3200 do segundo semestre

PRO3200 – EPUSP
Professora Celma de Oliveira Ribeiro
Assistente: Pedro Gerber Machado & Monitor: Vinícius Castanho

Neste estudo de caso, utilizaremos a base de dados IMDB.rds para aplicar nossas primeiras funções no R com o intuito de manipular a base de dados e deixa-la apta a uma análise exploratória de dados. Esse dataframe possui 15 colunas e 3713 linhas.

1. Importação de dados.

Em um primeiro momento, é interessante que se crie uma pasta diretório onde serão salvas as análises deste caso. Sugere-se que a base de dados esteja dentro dessa pasta diretório.

Para importar a planilha no ambiente do RStudio, utilizaremos o comando read_rds( ). Quaisquer dúvidas sobre a função podem ser sanadas utilizando o "help" do ambiente.

Após o comando, algo desse tipo deve aparecer em seu RStudio:

	titulo	ano	diretor	duracao	cor	generos
1	AvatarÂ	2009	James Cameron	178	Color	Action|Adventure|Fantasy|Sci-Fi
2	Pirates of the Caribbean: At World's EndÂ	2007	Gore Verbinski	169	Color	Action|Adventure|Fantasy
3	The Dark Knight RisesÂ	2012	Christopher Nolan	164	Color	Action|Thriller
4	John CarterÂ	2012	Andrew Stanton	132	Color	Action|Adventure|Sci-Fi
5	Spider-Man 3Â	2007	Sam Raimi	156	Color	Action|Adventure|Romance
6	TangledÂ	2010	Nathan Greno	100	Color	Adventure|Animation|Comedy|Far
7	Avengers: Age of UltronÂ	2015	Joss Whedon	141	Color	Action|Adventure|Sci-Fi
8	Batman v Superman: Dawn of JusticeÂ	2016	Zack Snyder	183	Color	Action|Adventure|Sci-Fi
9	Superman ReturnsÂ	2006	Bryan Singer	169	Color	Action|Adventure|Sci-Fi
10	Pirates of the Caribbean: Dead Man's ChestÂ	2006	Gore Verbinski	151	Color	Action|Adventure|Fantasy

2. Manipulação de dados.

Algumas linhas do nosso df possui valores NA, que representam ausência de informação. Por vezes, pode ser interessante eliminar tais linhas. Parai isso, vamos utilizar a seguinte função:

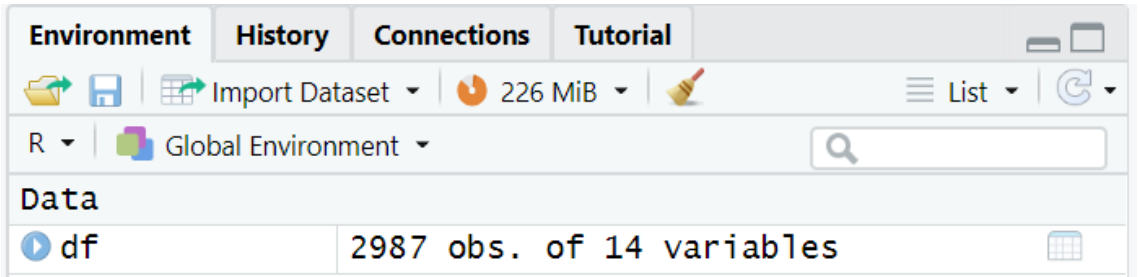
```
df <- na.omit(df)
```

Perceba que ao digitar "df$" numa linha disponível, algumas opções de colunas aparecem. Essa dinâmica facilita a seleção da coluna que desejamos manipular.

A coluna "país" não é do nosso interesse, visto que possui a mesma resposta atribuída a todas as linhas. Podemos, então, excluir essa coluna que não nos tem valor:

```
df$pais = NULL
```

Você pode confirmar as modificações no seu environmente:



Pode ser do nosso interesse uma variável que não está apresentada no banco de dados, mas que a partir dele pode ser construída. Suponha que queremos analisar a idade de cada produção. Para isso, devemos criar uma nova coluna e atribuir a cada linha o valor numérico da idade do filme.

E não para por aí. Suponha que ainda classificaremos um filme com mais de 50 anos de idade como “Velho” ao passo que menos são classificados como “Novo”. As estruturas condicionais de for( ) e if( ) são, então, utilizadas para a construção dessa atualização.

```
# Construindo colunas.

df$idade = 2023 - df$ano
df$status = "Novo/Velho"

for (i in 1:3713){
  if(df$idade[i]>=50){
    df$status[i]="Velho"
  }
  else{
    df$status[i]="Novo"
  }
}
```

Podemos tabelar as quantidades de “Velho” e “Novo” via função table( ):

```
Novo Velho
2946    41
```

3. Entrega no Moodle.

Os cases devem ser enviados no e-disciplinas em um arquivo .pdf com o script do R anexo ao final do próprio PDF, de forma a possibilitar o Ctrl c, Ctrl v do mesmo para efeitos de correção.

O prazo de entrega é domingo, 26/03, às 23h59.