FIAP

MBA+

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

FIAP

MBA+

PROGRAMANDO IA COM R

Prof. Elthon Manhas de Freitas
elthon@alumini.usp.br /
profelthon.freitas@fiap.com.br
2021

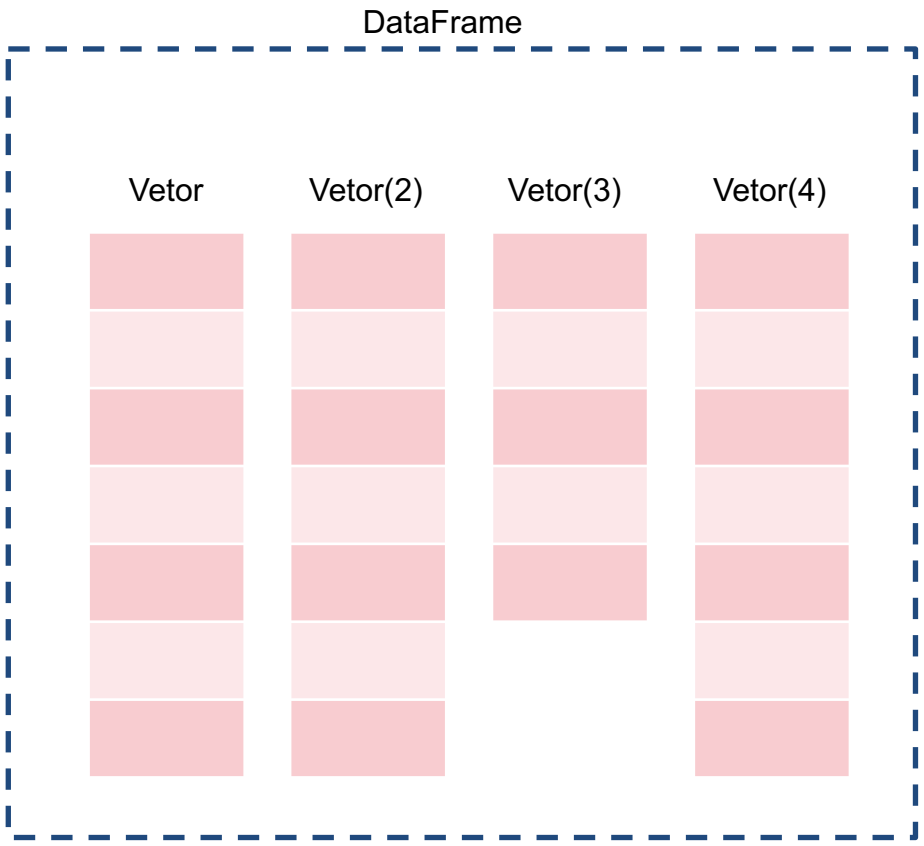
Revisão da última aula

- O que vimos na aula passada?

Manipulação “simples” de Data.Frame

Estrutura de dados : Data Frame

- A estrutura de dados mais usada em análise de dados.
- Equivalem às planilhas do Excel!



Acessamos aos vetores das colunas de um data frame pelo símbolo $

```
• mtcars
• class(mtcars)
```

Data Frame: Funções especiais

- Primeiros registros
- Últimos
- Resumo estatístico
- Visualização em Janela
- Pivoteamento

- `head()`
- `tail()`
- `summary()`
- `View()`
- `aggregate()`

```
aggregate(mtcars,
          by=list(mtcars$cyl),
          FUN=mean)
```

Manipulação básica de um dataset

Usando o R em substituição a uma planilha Excel

FIAP

Parte 1, Usando apenas as instruções vista em sala, você é capaz de:

- Carregar o dataset BrFlights2.Rdata
  https://storage.googleapis.com/ds-publico/BrFlights2.RData
  - Trata-se de todos os voos comerciais brasileiros de 2016 e 2017 (2.542.519x21)

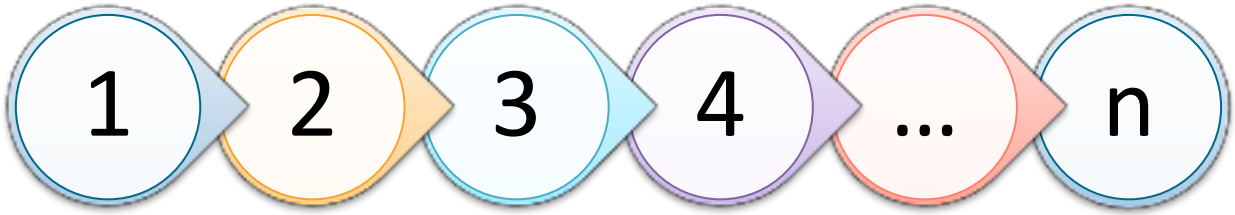
- Crie coluna com o atraso da partida e outra com o atraso da chegada.
- Crie coluna com distância euclidiana entre origem e destino.
- Crie coluna com tempo de viagem real.

- Como ver o primeiro quartil, média, mediana, etc. da coluna com o atraso na partida?
- Como ver o resumo da tabela toda?

Parte 2, veja o help da função agregate e tente resolver:

- Qual companhia aérea com maior atraso médio?
- Qual estado de origem com maior atraso médio?
- Qual a relação média entre distância percorrida e tempo de voô?
  - É possível identificar a companhia mais rápida?

Seqüencias



Sequências

- Dados sequenciais podem ser usados rapidamente das seguintes formas:
  - : (isso mesmo, dois pontos)
  - função seq (e suas variações)
    - seq.int
    - seq_along
    - seq_len

```
1:9
80:32
4:-2
3:3
1:0
```

presente no help:

```
seq(0, 1, length.out = 11)
seq(stats::rnorm(20))
seq(1, 9, by = 2)
seq(1, 9, by = pi)     # stays below 'end'
seq(1, 6, by = 3)
seq(1.575, 5.125, by = 0.05)
seq(17) # same as 1:17, or even better
seq_len(17)
```

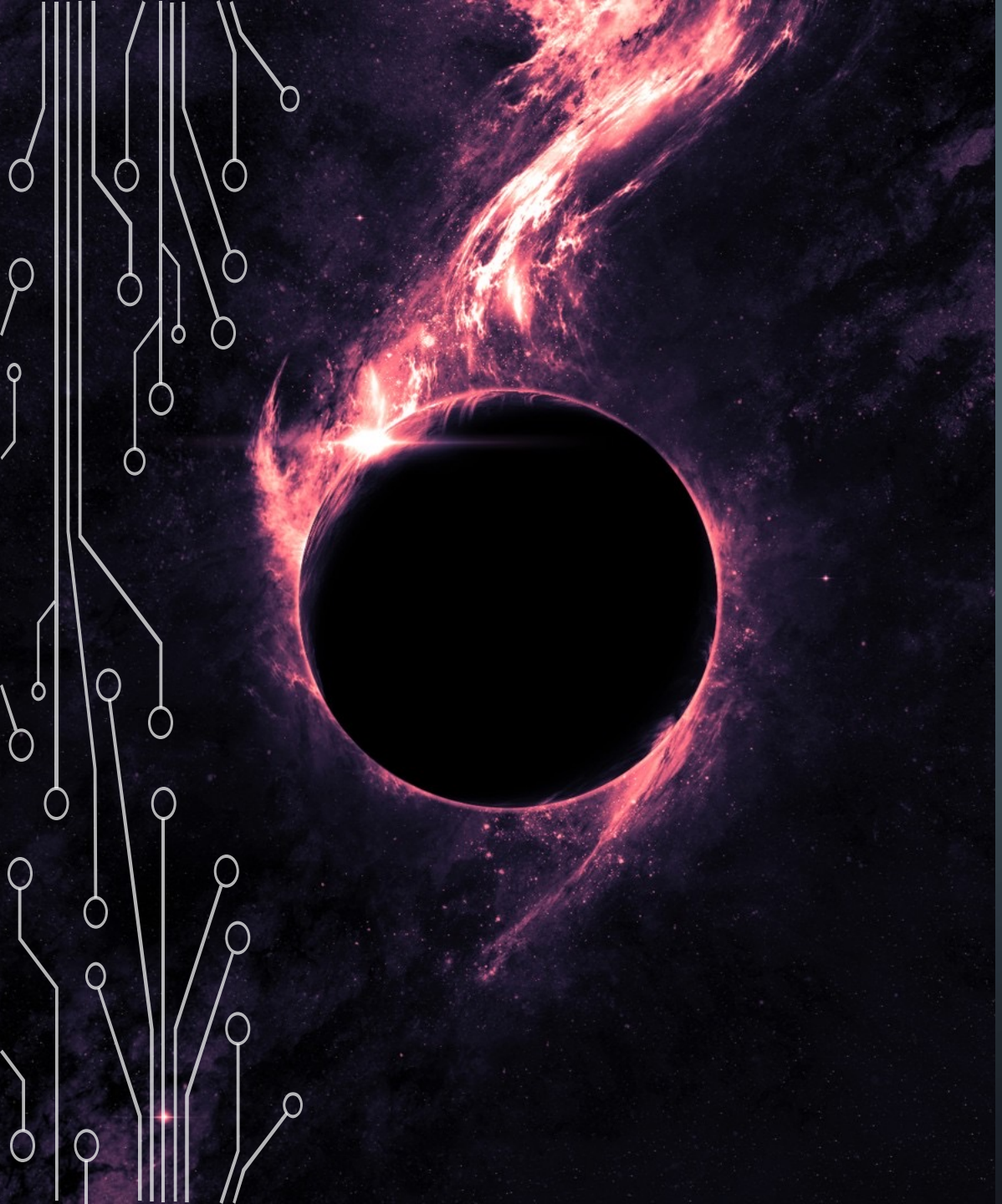
Replicação

FIAP

- Comando que “repete” um valor n vezes:
  - `rep(0, times = 40)`
  - `rep(c(0, 1, 2), times = 10)`
  - `rep(c(0, 1, 2), each = 10)`

Exercícios com seqüencias

- Crie uma variável `my_seq` com 30 valores entre 5 e 10
- Veja o help da função `:`
- Qual a diferença das instruções
  - `pi:10`
  - `10:pi`
- Como consultar o tamanho do vetor `my_seq`
- Como fazer uma sequencia que acompanhe o tamanho do vetor `my_seq`?



VOID

MISSING VALUES:

`NA, NaN e NULL`

SPECIAL VALUE:

`Inf`

OUTROS TIPOS DE DADOS (DATA E SEQUENCIAS)

Valores especiais

FIAP

- NA
  - Missing genérico, para representar Not Available
- NaN (Not a Number)
  - Resultados de contas que causam erros aritméticos
  - Exemplo: $\frac{0}{0}$ ou $\sqrt{-4}$
- Inf e -Inf
  - Números infinitos
  - Exemplo: $\frac{19}{0}$
- NULL
  - Null representa o vazio, o nada!
  - Até NA tem um tipo, mas NULL não!
  - Não é possível colocar NULL em um vetor

Exercícios sobre Special Values

FIAP

- Criar 5 variáveis. Cada uma com um tipo dos valores especiais (+Inf e –Inf)
- Tentar criar um vetor com as 5 variáveis.
  - Foi possível?

- Verificar os valores das variáveis e vetor através das funções
  - is.na()
  - is.nan()
  - is.null()
  - is.infinite()
  - is.finite()

Filtrando dados das estruturas – Subsets simples

Filtrando dados das estruturas (subset)

FIAP

- Os dados são filtrados em diversos tipos de estruturas através do uso de colchetes.

```
• mtcars
• mtcars[2,5]
• mtcars[2:4,5]
• mtcars[2:4,5:7]
• mtcars[8:9, c(1,2,4)]
• mtcars[2:4, c("mpg", "wt")]
```

- Nas matrizes, a primeira dimensão é a linha e a segunda é a coluna;
- Pode-se filtrar pelo número ou pelo nome, quando a dimensão for nomeada, ou por valor lógico;
- Ao sub-conjunto de dados se dá o nome de “subset”

Sub-sets simples: Vetores

FIAP

- Vetores, por terem apenas uma dimensão, possuem o subset mais simples de se obter

- Uso dos colchetes simples:

```
vetor[filtro] -> resultado
```

(subset)

nome do vetor (set)

Diversos modos de aplicar filtros (subscript)

R-Markdown: Subsets

Sub-sets simples: Matrizes

FIAP

- Similar ao subset dos vetores, porém com 2 dimensões

- Uso dos colchetes simples:

  - `matriz[linha,coluna] -> resultado`

R-Markdown:
Subsets

Sub-sets simples: Listas

FIAP

- Para listas, há mais opções de filtros:
  - `lista[filtro] -> sub-lista`
  - `lista[[filtro]] -> elemento`
  - `lista$elemento -> elemento`

R-Markdown:
Subsets

Exercício

Filtrar o dataset mtcars

Buscar todas as colunas dos registros cujo “mpg” seja maior ou igual a 15.

Obter do dataset BrFlights todos os voôs da companhia “AZUL”

Armazenar em outro data frame

Desafio: Armazenar em ordem cronológica de partida.

Data

Data e Hora

Data e Hora

FIAP

- Data é representada pela classe Date
  - São armazenados os dias a partir de 01-01-1970

- Data-e-Hora pode ser representada por duas classes:
  - POSIXct
    - Número de segundos a partir de 01-01-1970
  - POSIXlt
    - Lista de ano, mês, dia, hora, minuto, segundo, etc.

Experimento com Data e Hora

- Façam o seguinte experiento, analizem e comentem o resultado:

```
dia_texto <- "28/09/2017 T 18:51:30"
dia_date  <- as.Date(dia_texto,format="%d/%m/%Y T %H:%M:%S",tz="America/Sao_Paulo")
dia.time1 <- as.POSIXct(dia_texto,format="%d/%m/%Y T %H:%M:%S",tz="America/Sao_Paulo")
dia.time2 <- as.POSIXlt(dia_texto,format="%d/%m/%Y T %H:%M:%S",tz="America/Sao_Paulo")

dia_date
dia.time1
dia.time2
unclass(dia_date)
unclass(dia.time1)
unclass(dia.time2)

dia.time1$year
dia.time2$year
```

- O que a função “unclass’ faz?
- Qual a diferença entre os objetos dia.time1 e dia.time2?

Funções especiais com data e hora

- Para obter a data e hora atual

```
• Sys.time()
• Sys.Date()
• Sys.timezone()
```

- Converter um texto em data

```
• as.Date()
• as.POSIXct()
• as.POSIXlt()
• strptime("Janeiro 10, 2012 10:40", "%B %d, %Y %H:%M")
```

Pacotes famosos para manipulação de data/hora

- hms
  - Exibe diferenças temporais em HH:mm:ss

```
BrF$Partida.Atraso3 <- as_hms(BrF$Partida.Atraso )
```

- lubridate
  - Pacote muito utilizado para manipulação de dados (provavelmente o mais famosos também)
  - Principais funcionalidades:
    - Conversão de datas com facilitador de formatos
    - Extração de componentes da data
    - Intervalos de datas (dia, hora, segundos, etc)
    - Tratamento de fuso horário
    - Operações com datas (somar dias, minutos, etc)

Data e Hora : `lubridate`

- Conversão de datas com facilitador de formatos

```
library(lubridate)

ymd("20110604")
mdy("06-04-2011")
dmy("04/06/2011")
ymd_hms("2018-04-23T19:02:13")
dmy_hms("23/04/2018 19:03:14")
```

- Extração de componentes da data
- Intervalos de datas
  - dia, hora, segundos, etc
- Tratamento de fuso horário
- Operações com datas
  - somar dias, minutos, etc

Abrir Rmarkdown:

Aula 02 - Lubridate

Exercícios Lubridate

- Calcular a Black Friday de 2021:

  O Black Friday acontece 1 dia depois do dia de ação de graças americano, que é celebrado na 4ª quinta-feira de novembro. (o seguinte roteiro pode ajudar

  - Crie uma variável representando 1-nov-2021
  - Verifique que dia da semana é 1-nov
  - Adicione a quantidade necessária para chegar na 1ª quinta feira e armazene em uma segunda variável
  - Adicione 3 semanas e some 1 dia.
- Testar para 2018, 2019*, 2020 e 2021

2018 – 23/11
2019 – 29/11
2020 – 27/11
2021 – 26/11

Trabalho em grupo

Trabalho em grupo

- Escolha dos Cases :

https://sites.google.com/usp.br/programando-ia-com-r

Exercícios individuais de revisão (Aprenda R no R)

FIAP

MBA+