FIAP

MBA+

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

FIAP

MBA+

PROGRAMANDO IA COM R

Prof. Elthon Manhas de Freitas
elthon@usp.br

2018

Revisão da última aula

- O que vimos na aula passada?

As fronteiras do R
Obtendo dados de arquivos

Obtendo dados com o R

FIAP

- Primeira observação:
  - Qual seu diretório de trabalho?
    - O que fazem mesmo os comandos?
      - `setwd()`
      - `getwd()`
    - Caminho relativo : "`./data`", "`../`"
    - Caminho absoluto: "`/r/fiap/data`"
    - Formato Windows: "`C:\\r\fiap\data`"

Obtendo dados com o R

FIAP

- Boa prática: Separar seus dados de análise dos scripts de análise

```
if(!file.exists('data')){
  dir.create('data')
}
```

- Primeira observação:
  - Qual seu diretório de trabalho?
    - O que fazem mesmo os comandos?
      - `setwd()`
      - `getwd()`
    - Caminho relativo : `"./data", "../"`
    - Caminho absoluto: `"/r/fiap/data"`
    - Formato Windows: `"C:\\r\\fiap\\data"`

Obtendo dados da internet

FIAP

RMarkdown

- Simples download de um arquivo da internet:
  - `download.file()`
    - O modo padrão é texto. Use `mode='wb'` para evitar problemas.

- Comandos auxiliares:
  - Obter o nome "base" do arquivo, ignorando seu caminho:
    - basename()
  - Montar um caminho 'path' para um arquivo:
    - file.path()
- Exemplo:

```
file.url = 'http://www.bcb.gov.br/pec/Indeco/Port/IE1-04.xlsx'
file.local = file.path('./data', basename(file.url))
download.file(url = file.url, destfile = file.local , mode='wb')
```

Exercício - Download

- O que faz cada uma das instruções do exemplo?

```
file.url = 'http://www.bcb.gov.br/pec/Indeco/Port/IE1-04.xlsx'
file.local = file.path('./data', basename(file.url))
download.file(url = file.url, destfile = file.local , mode='wb')
```

- Qual o conteúdo do arquivo baixado?

- Criar uma função que recebe uma url e baixa o arquivo sempre na pasta “./data”.
  - Chamar a função para os seguintes arquivos:

```
https://raw.githubusercontent.com/elthonf/fiap-mba-r/master/data/Copas.csv
https://raw.githubusercontent.com/elthonf/fiap-mba-r/master/data/Copas-Partidas.csv
https://raw.githubusercontent.com/elthonf/fiap-mba-r/master/data/Copas-Jogadores.csv
```

Lendo arquivos locais CSV (Prática)

FIAP

- Para ler um arquivo local, vamos usar o comando
  - `read.table`

- Este recupera um objeto tipo data.frame.

- Copa do mundo
  - Usar read.table para ler o arquivo ‘Copas.csv’
  - Os arquivos Copas-Partidas.csv e Copas-Jogadores.csv podem ser lidos da mesma forma?

- Alternativa: read.csv (o que este comando faz?)
- Avaliando esta função. O que são os ‘...’?

RMarkdown

Lendo arquivos locais tipo Excel

- Biblioteca `xlsx`
  - Necessita do Java e do pacote rJava
  - Funções `read.xlsx` e `read.xlsx2`
  - Lê e escreve em arquivos Excel
- Biblioteca `readxl`
  - Não necessita nenhum pacote, plugin, etc.
  - Apenas lê arquivos Excel (biblioteca mais usada e mantida pela equipe do RStudio)
  - Função `read_excel`
- Biblioteca `openxlsx`
  - Não necessita nenhum pacote, plugin, etc.
  - Lê e escreve

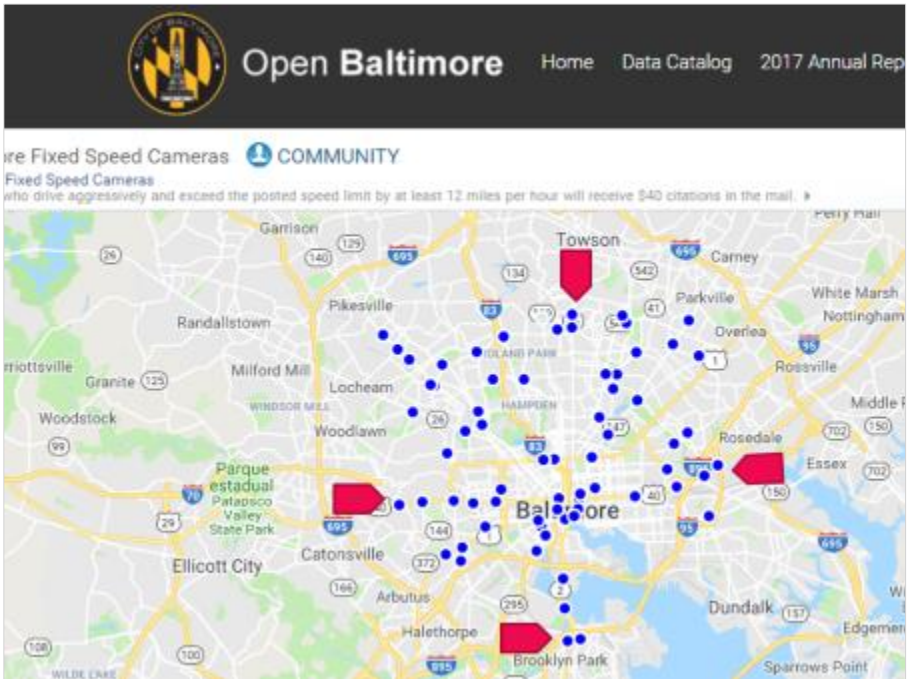
Atividade: Ler Excel

- Para esta atividade, vamos buscar dados públicos atualizados.

https://data.baltimorecity.gov/

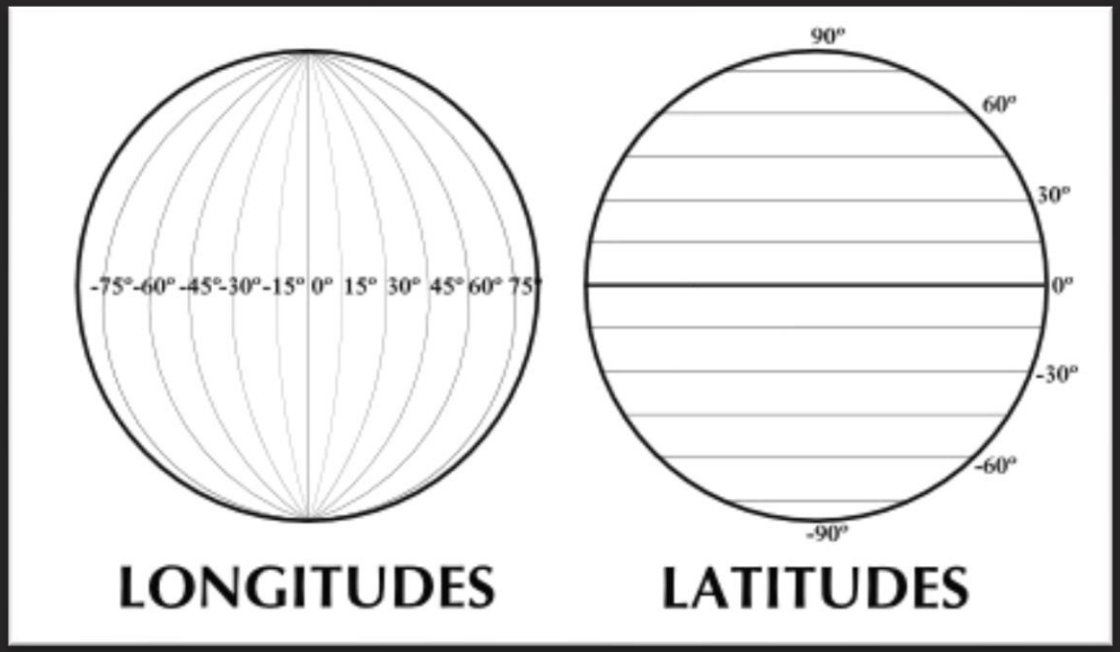
- Baixar o arquivo usando sua função de download:

```
https://github.com/elthonf/fiap-mba-r/raw/master/data/cameras.baltimore.xlsx
```



Atividade: Ler Excel

- Em Baltimore
  - Qual a câmera mais ao Norte?
  - Qual a câmera mais ao Sul?
  - Qual a câmera mais ao Leste?
  - Qual a câmera mais ao Oeste?



Outros tipos de leitura local

- Ainda é possível a leitura de outros tipos de arquivos (dados locais), entre eles
  - XML, JSON, TSV, HTML, PDF, etc.
  - Imagens BMP, JPEG, GIF, etc.

- Um pacote muito versátil para leitura de diversos tipos de arquivos simpes é o
  - readr

Dados de banco de dados (MySQL)

FIAP

- Instalar pacote `RMySQL`

```
dbConn <- dbConnect(MySQL(), user='elthon',
password='xxxxxxx', host='localhost')

resultado <- dbGetQuery(dbConn, 'SELECT * FROM
clientes')

dbDisconnect(dbConn)
```

Outros tipos de leitura remota

FIAP

- Além de Bancos é possível:
  - Trabalhar com dados HDF5
  - Baixar páginas da Web ou fazer um scrapper
  - Chamar API's, incluindo API's de IA
    - Watson da IBM
    - Azure Microsoft
    - GCP da Google
    - etc.
  - Abrir dados de sistemas diversos, como SAS, Matlab/Octave, S, Minitab, etc.
  - Compactar e descompactar arquivos

Manipulação de dados

Manipulação tradicional

- Manipular dados de forma tradicional é possível, porém trabalhoso.
- Até o momento já vimos:
  - Como filtrar dados
  - Criar colunas de um data.frame / matriz
  - Remover colunas
  - Alterar tipo de dados das colunas
  - Alterar informações contidas nas estruturas
  - ...
  - etc

Conhecendo o pacote dplyr

FIAP

DPLYR - Mutate

FIAP

- Adiciona uma coluna ao data.frame

```
tabela <- mutate(tabela,
                 col1 = `formula`,
                 col2 = `formula` )
```

- O pacote dplyr traz um dataset chamado ‘starwars’ para pequenos experimentos
  - Conhecendo o dataset : head(starwars)
  - Criar coluna com índice de massa corpórea:

```
s2 <- mutate(starwars,
             imc = mass / ((height / 100) ^ 2) )
```

DPLYR – Mutate Exercício

- Carregar o dataset BrFlights2 (aula 02)
  - Criar as colunas vistas na aula 2, mas agora com um único comando usando o mutate.

  - Partida.Atraso
  - Chegada.Atraso
  - DistanciaEuc (distância euclidiana)
  - TempoViagem.Real

Pacote magrittr e o pipe

FIAP

- Pipe para substituir argumento “.”
- Pipe usado para substituir o primeiro argumento
- Transformações em série.

Ctrl + Shift + M

DPLYR – Mutate + pipe

FIAP

- Carregar o dataset BrFlights2

- Fazer o mesmo procedimento já realizado:
  - criar as 4 colunas, porém agora com a utilização do pipe.

- Para pensar um pouco:
  - Há um atraso médio de chegadas. Qual é este valor?
  - Crie uma coluna de atraso relativo com a diferença do atraso real para o atraso médio.

DPLYR – Filter (Filtra linhas)

- Utilizado para filtrar de forma mais fácil os dados de um data.frame.

```
starwars %>%
  filter(species == "Droid")
```

- E o que faz este comando?

```
starwars %>%
  filter(species == "Droid") %>%
  View()
```

- Quais são os voôs da Azul no dataset BrFlights2? Atribua à variável Azul e use o pipe.

DPLYR – SELECT

- Utilizado para ordenar e filtrar colunas.

```
starwars %>%
  select(name, ends_with("color"))
```

- Remonte o dataset BrFlights2 para que:
  - Partida.Atraso real fique após partida Partida.Real
  - O mesmo com a chegada
  - TempoViagem.Real fique após Chegada.Atraso

  - Atribua esta busca a um novo dataset: BrFlights3
  - Garanta que não tenha perdido nenhuma coluna

DPLYR – SELECT

- Para filtrar uma seqüencia de colunas, é possível usar os conhecidos “:”

```
starwars %>%
  mutate(name, imc = mass / ((height / 100)  ^ 2)) %>%
  select(name:mass, imc)
```

- O exercício anterior teria sido mais fácil se você soubesse disso?
- Vamos praticar?

DPLYR – SELECT

- Pode ser usado para renomear colunas

- SELECT (col1.nome.novo = col1.nome.antigo)

DPLYR – ARRANGE

- Usado para Reordenar linhas de um data.frame

```
starwars %>%
  arrange(desc(mass))
```

  - Mas e o sort?

- No dataset BrFlights2:
  - Quais os 100 maiores atrasos de chegada de voôs nacionais / regionais?
  - Quantas dias/horas/minutos/segundos foi o maior atraso?

DPLYR – RENAME

- Renomar uma coluna é bem complicado no R
  - É preciso criar uma nova coluna
  - Apagar a antiga
  - Depois copiar as demais
  - Ficar atento à ordem
  - etc.

```
rename(starwars,
       nome = name,
       altura=height,
       massa = mass)
```

- Obs.: O Comando SELECT pode ser usado para o mesmo fim

DPLYR – Joins

FIAP

RMarkdown

DPLYR – Agregações

- group_by
- summarise
- n

```
starwars %>%
  group_by(species) %>%
  summarise(
    j = n())
```

```
starwars %>%
  group_by(species) %>%
  summarise(
    j = n(),
    mass = mean(mass, na.rm = TRUE)
  ) %>%
  filter(j > 1)
```

DPLYR – Agregações

- Lembram do “aggregate” do pacote stats, que usamos no BrFlights2?

- Agora vamos refazer as agregações usando o dplyr:
  - 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?

Prática extra 01 (opcional)

- Quais pacotes mais foram baixados do CRAN há exaramente 1 ano atrás?

`http://cran-logs.rstudio.com/yyyy/yyyy-mm-02.csv.gz`

  - Substituir yyyy, mm e dd pela data desejada

  - É possível abrir este arquivo no Excel?
  - Quantos downloads houveram neste dia?
  - Qual sistema operacional que mais baixou pacotes?
    - linux-gnu = Linux
    - darwinX = Mac
    - mingw32 = Windows
  - Quais os 10 pacotes mais baixados?

Prática extra 02 (opcional)

FIAP

- Arquivo ‘Fifa/fifa game-2.csv’
  - Refere-se às características de todos os jogadores do jogo para vídeo-game ‘Fifa 2018’
  - As posições que os jogadores podem / preferem jogar são as colunas rs:gk (27 posições)
  - Existe uma coluna de avaliação geral (overall). Qual o histograma desta coluna? Parece familiar?
  - Agrupe os jogadores nas posições favoritas e diga quais as principais características dos jogadores das posições.
  - Como fazer o JOIN com o dataset da copa do mundo?
  - O que mais vocês você consegue extrair destes dados?

Exercícios individuais

- Aprenda R no R (exercícios de revisão)

- Portfólio individual

MBA+