



## 4. R markdown

Tags	In progress
Módulo	<a href="https://www.coursera.org/learn/data-scientists-tools/home/module/4">https://www.coursera.org/learn/data-scientists-tools/home/module/4</a>

### ▼ Instalação

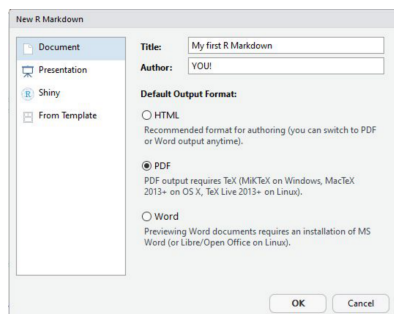
instale o pacote no R

```
install.packages("rmarkdown")  
library(rmarkdown)
```

e mano é isso

### ▼ Arquivos no formato R markdown

No Rstudio, abra um arquivo novo no formato .Rmd

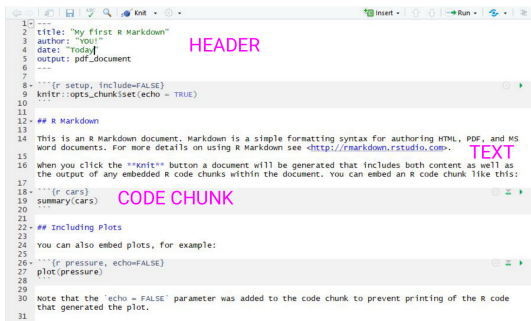


essa janelinha vai abrir

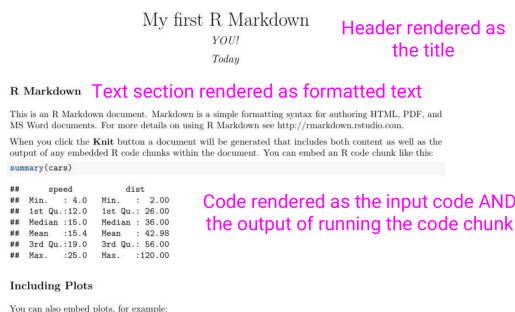
### R Markdown

Markdown é uma linguagem de texto que combina metadados, texto e snippets de código para gerar documentos.

<https://vimeo.com/178485416>

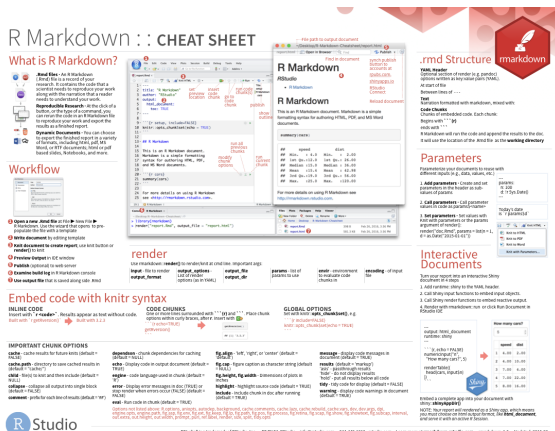


esse é o modelo padrão



Knit é a tricotagem do código para documento

Para dar conta de lembrar os comandos, existem algumas planilhas



ou simplesmente abrir no Rstudio

File > Help > Cheatsheets > R Markdown Cheat Sheet

## Perguntas da ciência de dados

De maneira geral, existem 6 tipos de perguntas em análise de dados

## ▼ Descritiva

análises descritivas tem como objetivo descrever e sumarizar conjuntos de dados.

Geralmente é a primeira etapa a ser feita em analisar um conjunto de dados novos.

Esse tipo de análise tem como objetivos entender as amostras, não é usado pra gerar conclusões. Descrever os dados é diferente de fazer interpretações.

## ▼ Exploratória

O objetivo é explorar os dados e encontrar relação entre as variáveis. Entretanto, essa relação não implica causalção, a análise exploratória não identifica causalidade.

Por isso a análise exploratória pode ser usada para inferir hipóteses ou delimitar novas questões, mas não responde questões, principalmente em como conjuntos de dados estão relacionados com outros. Análises exploratórias mostram que relações existem, mas não as suas causas.

## ▼ Inferencial

Análises inferenciais usam uma parcela relativamente pequena de dados para inferir algo maior, como por exemplo uma população.

Análises inferenciais geralmente são o objetivo de modelos estatísticos, onde um pequeno grupo de informação é usado para extrapolar e generalizar esta informação para grupos maiores

## ▼ Preditiva

enquanto as análises inferenciais funcionam no tempo, análises preditivas funcionam no espaço. Dados pretéritos são utilizados para fazer predições sobre dados futuros.

Nota-se que assim como as análises exploratórias, análises preditivas não exploram as causalções da relações entre

dados, mas apenas capitaliza em cima das relações observadas.

### ▼ Causal

análises causais descrevem a causa e efeito entre variáveis, manipulando uma parte dos dados para observar o comportamento de outros.

geralmente são complicadas de serem feitas, principalmente por que é difícil definir a validade das premissas em dados observados.

outro ponto é que os dados são analisados de maneira agregada, ou seja, descreve características para uma população geral, mas indivíduos entre os dados não seguem necessariamente as mesmas características.

### ▼ Mecanista

Similar á causal, as análises mecanicas, ou de mecanincas, ou de mecanismos???? observam mudanças específicas em variáveis através da manipulação específica de outras variáveis.

Em dados biológicos, geralmente há muitas forças agindo para que estas análises sejam práticas, sendo utilizadas majoritariamente em pesquisas físicas.

## Desenho experimental

Um mal desenho experimental pode levar à uma coleta de dados que não responde a pergunta proposta ou leva a conclusões e inferências errôneas.

*Por exemplo, meu desenho experimental para contar Pbio não padronizou o volume analisado e portanto no meu mestrado os dados de fluxo e abundância estão errados.*

Para contornar isso, desenhos experimentais devem levar alguns conceitos

## Compartilhando dados

dropa lá no gitHub man

<https://github.com/jtleek/datasharing.git>

## p-hacking!

caralho, chato pra caralho...

p-valor é a probabilidade de que os resultados do experimento foram observados por acaso. Ou seja, descreve o comportamento da variabilidade dos dados. Quanto maior o p-valor, maior a probabilidade de observações não determinadas pelas variáveis analisadas.

em conta:

### ▼ variável independente (fator)

é a variável manipulada pelo experimento, não depende das outras variáveis para variar, geralmente representada no eixo x.

*no caso dos experimento com armadilhas de sedimento, essa variável é o tempo, ou características oceanográficas (T, SST, Chl)*

### ▼ variável dependente

é a variável, ou o grupo de variáveis que mudam em resposta à variável independente. Geralmente representadas no eixo y.

*Fluxo de massa, fluxo de bSi, fluxo de Pbio, mas aqui também pode ser forçantes oceanográficas em relação ao tempo*

### ▼ hipótese

a hipótese é a síntese da pergunta a ser respondida. Ou melhor, a proposta de relação entre as variáveis que estão sendo investigadas. Hipótese nula ( $H_0$ ) é quando esta relação não é observada com as análises realizadas.

*No caso da minha dissertação é uma pergunta exploratória: Como o fluxo de Pbio varia com o fluxo de massa. A hipótese nula: O fluxo de pBio não varia com o fluxo de massa.*

### ▼ número amostral

é o número de amostras determinadas durante o experimento. Existem modos de estimar este número ideal.

What is a p-value? (Updated and extended version)

An introduction to the concept of the p-value, in the context of one-sample Z tests for the population mean. Much of the underlying logic holds for other tests as


 <https://www.youtube.com/watch?v=UsU-02Z1rAs>

quando o p-valor é  $<0.05$  quer dizer que existe até 5% de chances de dados por acaso. Ou seja, as observações são significantes.

Entretanto, se os testes forem refeitos diversas vezes, por exemplo 20, uma das vezes (5%) irá apresentar significância.

Em big data, p-hacking explora esse conceito para achar padrões dentro dos conjuntos de dados que possuem alto valor de significância devido ao grande volume de dados. Podem levar a interpretações falsa, criando uma tendência de observar apenas o que quer observar.

FiveThirtyEight

 <https://projects.fivethirtyeight.com/p-hacking/>

18 amostras +  
3 amostras +  
2 amostras

#### ▼ **variável cofundadora**

é a variável (ou o grupo de variáveis) fora do experimento que influenciam a relação entre variáveis dependentes e independentes. Quando é feito o desenho experimental deve-se leva-las em consideração.

*Por exemplo, as relações ecológicas, origens de massa de água, perturbações de fluxo, etc..... . . . .*

#### ▼ **controle**

controle é um recorte no desenho experimental que contorna os dados de variáveis cofundadoras, por exemplo recortes de idade, gênero ou etnia.

#### ▼ **grupo de controle vs grupo de tratamento**

é um método de analisar o efeito de uma variável estabelecendo um grupo onde essa variável é fixa, ou inexistente, e um grupo de tratamento, comparando os resultados. Outro modo de aplicar o conceito é com testes cegos, onde uma parcela da população amostral recebe um placebo.

#### ▼ **aleatorizada**

Quando as variáveis cofundadoras não são conhecidas, amostras aleatórias de um determinado grupo amostral irão "balancear" a variabilidade resultante das cofundadoras

#### ▼ **réplica**

replicar as observações de dados descreve a variabilidade intrínseca dos dados e reduzir os efeitos de variáveis cofundadoras;

## **Datalhão! (big data)**

ou *dataralho* (do português: “dados para um caralho”), são conjuntos de dados que seguem o padrão de Volume, Velocidade, Variedade. Com características emergentes como capacidade de armazenamento e ferramentas de análise, existe uma popularidade crescente do dataralho.

A partir da obtenção destes dados de diversas fontes, eles devem ser organizados em tabelas coesas para serem analisados, isso é estruturar os dados.