Treinamento Bookdown

Robson Wilson Silva Pessoa, Ícaro Bernardes e Daniela Almeida

2020-08-24

Contents


```
knitr::opts_chunk$set(error = TRUE)
```

Chapter 1

Pré-requisitos

Este documento foi elaborado a partir da estrutura mínima obtida pelo *template* do *Bookdown* disponível no ambiente do *Rstudio*.

Este é um *sample* da escrita em **Markdown**. É possível utilizar qualquer recurso que suportado pelo *Markdown* do *Pandoc*, como a equação

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

.

A instalação do **bookdown** package pode instalado pelo CRAN ou Github. A seguir apresentamos uma sequência de passos de instalação pelo modo gráfico. Se você tem familiaridade pule a sequência de figuras e instale utilizando os comandos na aba *Console*, caso contrário siga os seguintes passos:

1. Primeiro é necessário abrir o Rstudio,

2. Selecoine a aba de instalação *Packages* e clique em **install**,

3. No ambiente de busca da interface instalação pesquise por *bookdown*, selecione o pacote e clique em *install*,

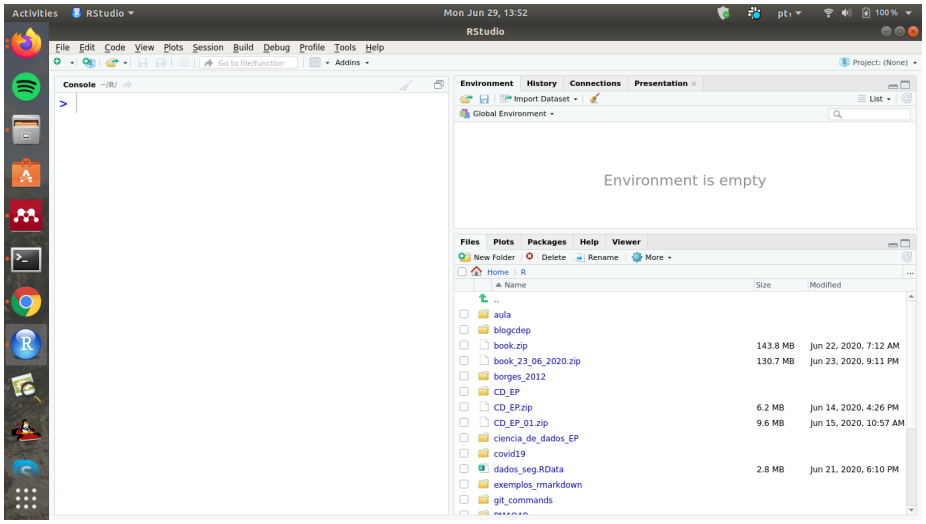


Figure 1.1: Opção de instalação pelo modo gráfico

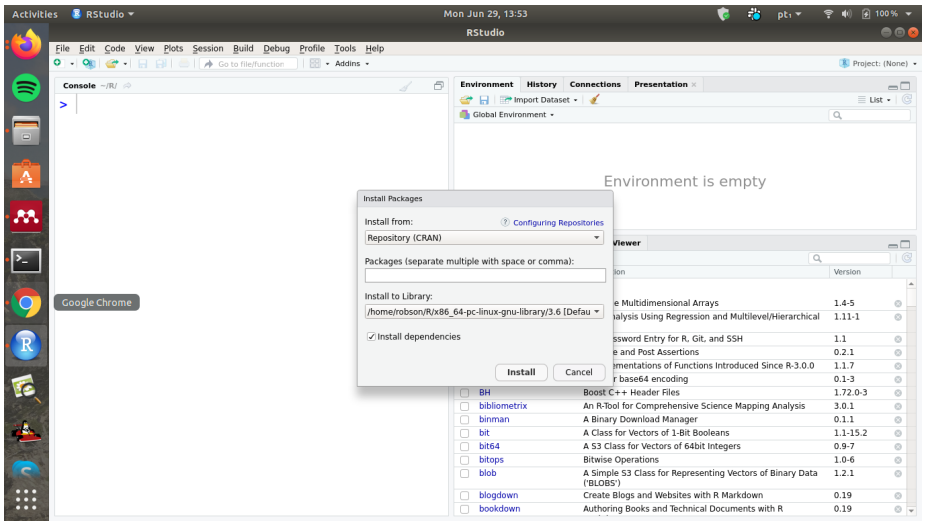


Figure 1.2: Selecoinar aba de instalação *Packages* - *Click em install*

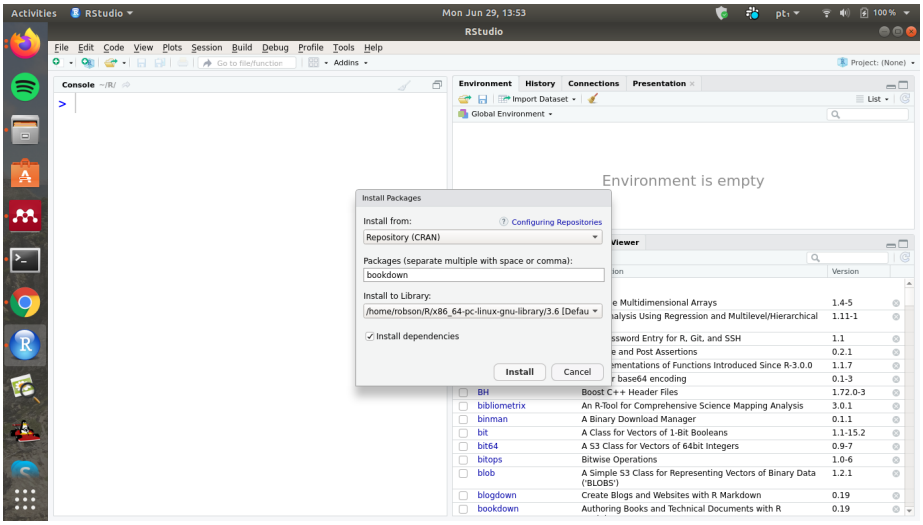
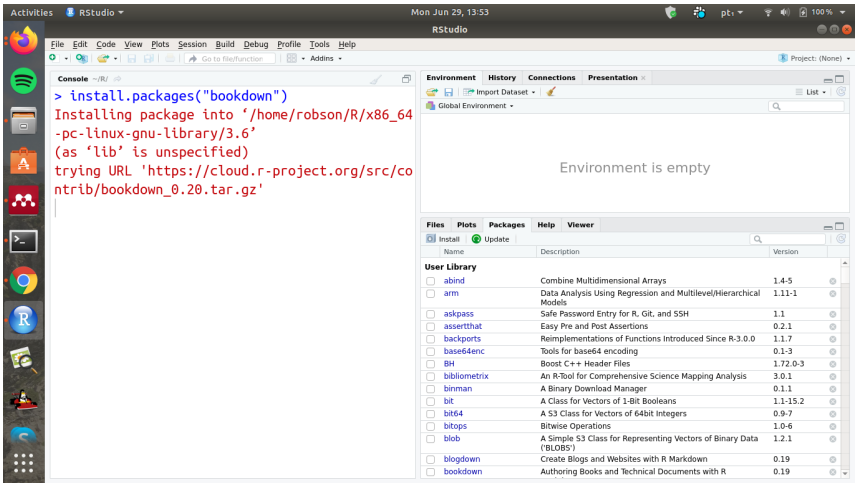


Figure 1.3: *Pesquisar por bookdown e clicar em install*



4. Finalmente, o código será instalado,

5. Ainda é necessário carregá-lo na seção de uso, novamente na aba *Packages* pesquise por *bookdown* e selecione o pacote, o que será suficiente para carregá-lo,

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
library("bookdown")
```

Deve-se lembrar que para cada arquivo *.Rmd* só pode ter um capítulo sendo definido pelo primeiro nível por **#**.

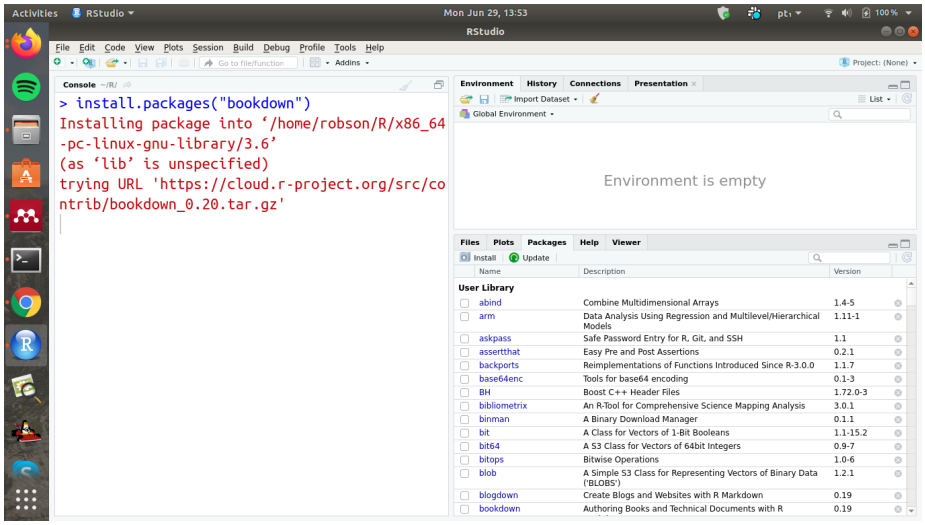


Figure 1.4: Carregar a biblioteca bookdown na aba Package

Para compilar este exemplo para PDF, é necessário o pacote XeLaTeX. É recomendável instalar o TinyTeX (que inclui o XeLaTeX): https://yihui.org/tinytex/.

Nos próximos capítulos serão apresentados outros detalhes sobre instalação e configuração.

Chapter 2

Introdução

A primeira palavra que devemos pensar ao encarar um curso de ferramentas de escrita de textos é *oportunidade*.

Quando pensamos em texto simples e rápidos, podemos naturalmente usar ferramentas como WYSIWYG (*What You See Is What You Get*) como **LibreOffice Writer** ou **Microsof Office Word**. Entretanto, trabalhar com textos longos, como relatórios, trabalhos de conclusão de curso (TCC), dissertações ou teses pode exigir recursos mais avançados como LaTeX.

2.1 Criação do projeto do livro

Faremos uso mais uma vez de recursos gráficos da interface do *Rstudio* para a criação do projeto do livro. Este material foi prepara utilizando a estrutura mínima disponibilizada pelo *template* do pacote *bookdown*. As etapas a seguir serão o suficiente para entender a criação e uso desse *template*:

1. Após a instalação e carregamento da biblioteca *bookdown*, podemos utilizar o *template*, primeiro devemos clicar no canto superior direito em *projetos* como:
2. Em seguida, selecionar *New Directory*:
3. Selecionar *Book Project with bookdown*
4. Em seguida o template com a versão mínima será disponibilizado por meio de uma pasta com o nome escolhido na etapa anterior.

Na lista apresentada acima são identificados arquivos com as seguintes extensões:

- **.Rmd**
- **.bib**

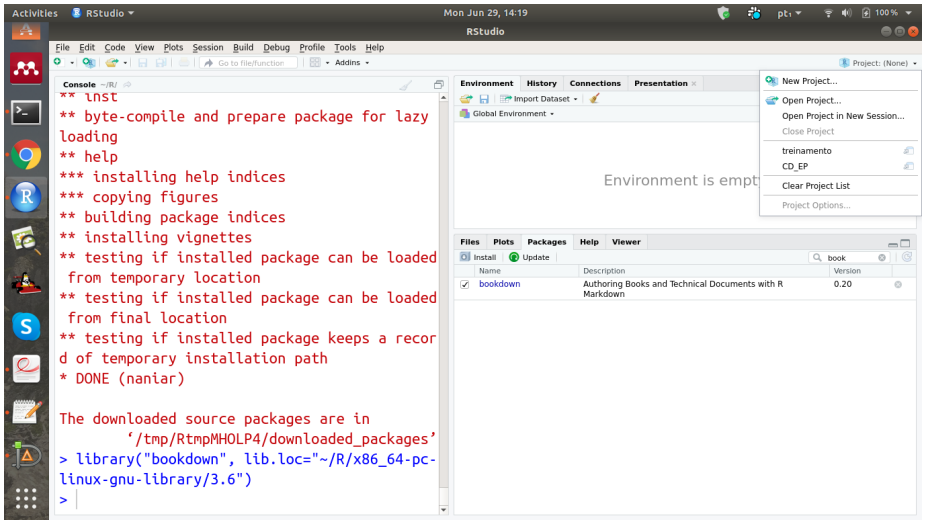


Figure 2.1: Criação de um novo projeto de livro

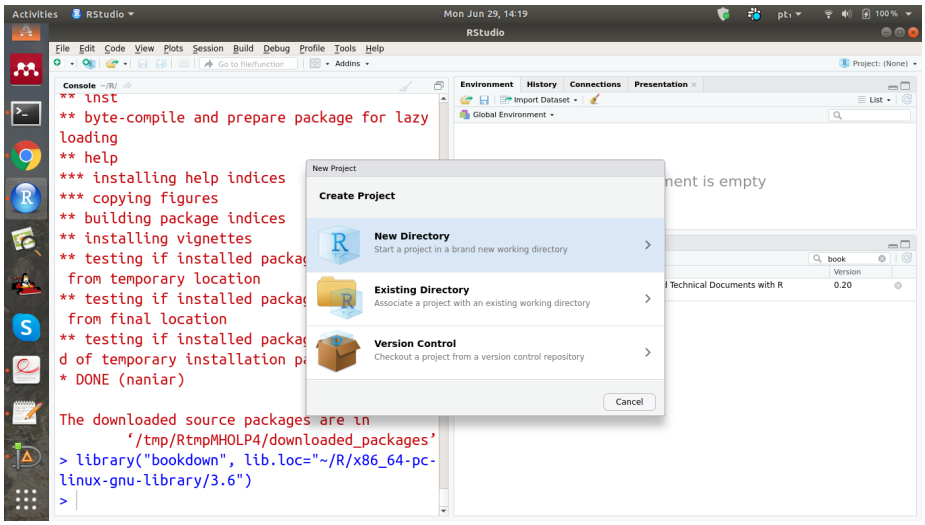


Figure 2.2: Selecionar a criação de um novo diretório

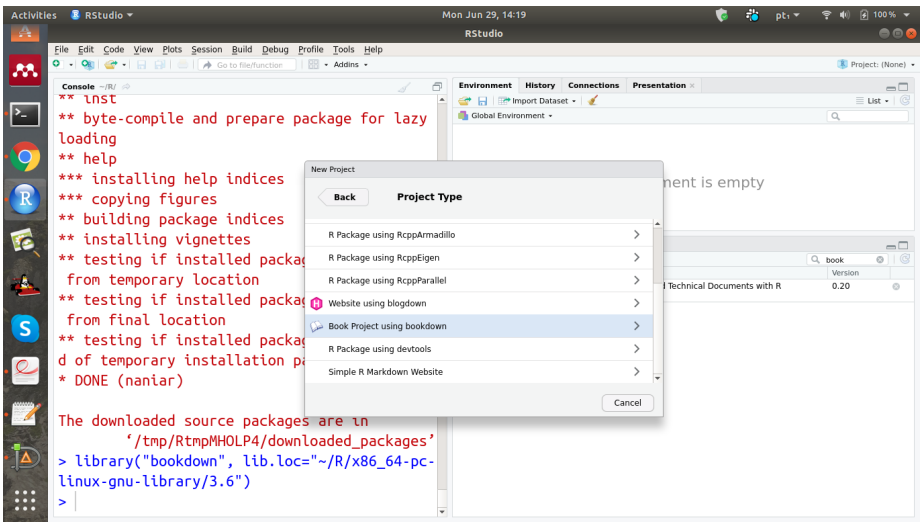


Figure 2.3: Seleção da opção de projeto de livro com pacote bookdown

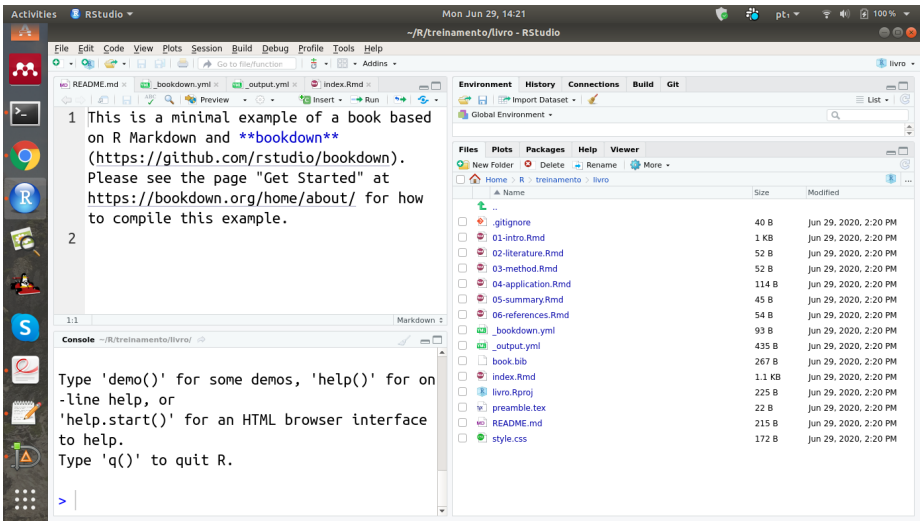


Figure 2.4: Seleção da opção de projeto de livro com pacote bookdown

- **.yml**
- **.tex**
- **.css**

Aqueles arquivos cuja a extensão é *.Rmd* são utilizados para a escrita dos conteúdos do livro em R Markdown. Entretanto, entre eles há um especial, *index.Rmd*, que constrói a página principal, por meio de comandos **yml**. Algumas configurações são reservadas em dois arquivos com extensão **.yml**. Sendo o arquivo _bookdown.yml para configurações gerais que serão úteis para quaisquer tipo de documento de saída, por exemplo, a definição se o título de cada capítulo será chamado de **Chapter** ou **Capítulo** , especialmente para este treinamento fizemos esta alteração. Enquanto que para o arquivo *_output.yml* são apresentadas configurações especiais para cada tipo de saída, como *bookdown::gitbook*, *bookdown::pdf_book:* ou *bookdown::epub_book:*. Especialmente para o caso do *gitbook* é necessário a existência do arquivo *style.css* para algumas configurações. Já o arquivo *book.bib* é uma estrutura especial do pacote *bibtex* do LaTeX e contem as informações de artigos que serão citados.

Chapter 3

Visualização e Ciência de dados

O capítulo 2 apresenta a tabela como uma forma poderosa para estruturar e visualizar informações. No entanto, quando trabalhamos com enormes tabelas com uma imensa quantidade de linhas e colunas se torna difícil interpretar suas informações, não importa o quão organizadas elas estejam. Às vezes, é muito mais fácil interpretar essas informações através dos gráficos, conteúdo que será explorado no decorrer deste capítulo.

A construção e visualização gráfica é de extrema importância na área de ciência de dados, pois é a partir de um bom gráfico que podemos extrair ideias, hipóteses e um melhor entendimento a respeito de um tema ou uma pergunta. A importância desse tipo de análise pode ser expressa por um ditado popular bastante conhecido: "Uma imagem vale mais que mil palavras".

3.1 Objeto de Estudo do capítulo

Para compreender a importância da análise gráfica e como utiliza-la corretamente, iremos analisar o perfil dos estudantes de Salvador que realizaram a prova do Exame Nacional do Ensino Médio (ENEM) no período de 2015 até 2019.

Porém, antes de qualquer coisa: O que é um **Perfil**? Esse termo é muito usado na estatística para **descrever determinado processo ou objeto de estudo, buscando entender características e padrões que o representa**. Para este caso em específico, vamos estudar os estudantes da cidade de Salvador utilizando os microdados do ENEM, publicados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), disponível ao público através deste link de acesso.

Como o termo **perfil** pode ser bem vasto e diversas características podem ser extraídas, é necessário concentrar essa análise em perguntas mais específicas para nortear o estudo. No decorrer deste capítulo, será explorado graficamente as seguintes questões:

- A quantidade de estudantes que realizam o ENEM aumentou de 2015 para 2019 na capital bahiana?
- Como é a distribuição de estudantes em Salvador por cor/raça? Conseguimos identificar algum padrão para esses valores?
- Na era da informação, como está o acesso dos estudantes a internet em suas residências?
- O acesso a computadores pessoais é algo comum para os estudantes de Salvador?
- O tipo de escola (pública ou privada) pode influenciar nas notas dos estudantes neste exame?

A compreensão desses dados é de suma importância para compreender melhor o perfil dos estudantes de Salvador que possuem o ENEM como uma oportunidade de acesso, as vezes única, ao ensino superior no Brasil.

3.2 Gráfico de barra

O **Gráfico de barras** é uma forma bastante comum e versátil de visualização na área de ciência de dados. Ele pode ser utilizado tanto com variáveis categóricas quanto numéricas para expressar grandezas. A Figura abaixo apresenta uma de suas utilizações: demonstrar grandezas numéricas.

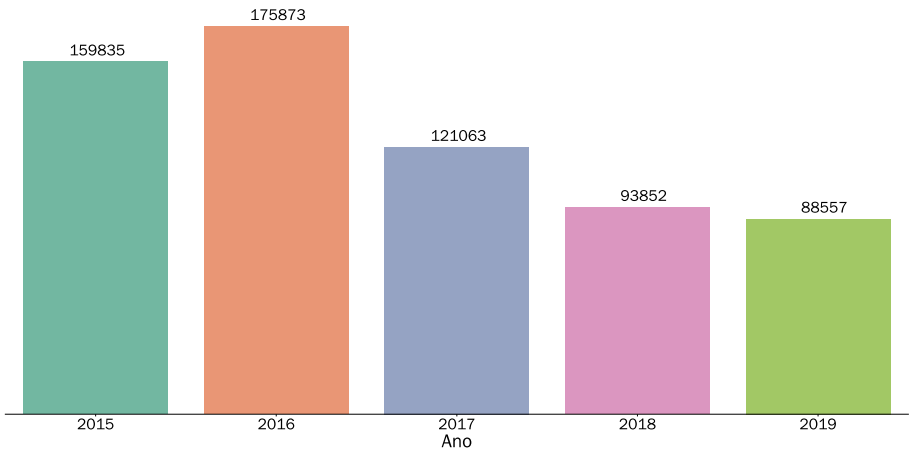


Figure 3.1: Quantidade de estudantes que realizaram o ENEM na capital bahiana

Na Figura 3.1 é apresentado a quantidade de estudantes que realizaram o ENEM de 2015 até 2019 na capital bahiana. É possível notar uma queda drástica na participação de estudantes entre os períodos de 2016 até 2019. Apesar de simples e direto, a análise desse mesmo resultado através de uma tabela pode se mostrar confusa:

Ano	Número de estudantes em Salvador
2015	159835
2016	175873
2017	121063
2018	93852
2019	88557

Note que ao visualizar a Tabela, nenhuma informação visual é passada para destacar os anos com mais ou menos participantes. Além disso, ela contém as mesmas informações demonstradas na Figura 3.1, porém com uma diferença: através da visualização gráfica fica muito mais claro a queda de inscrições no ENEM de 2016 até 2019.

O gráfico de barras apresenta uma característica muito importante relacionado ao tamanho das barras: elas crescem proporcionalmente de acordo as grandezas que elas se referem, ou seja, quanto maior o valor maior será sua barra. Comumente essas barras apresentam a mesma largura neste tipo de gráfico.

É através da Figura 3.1 que podemos responder a primeira pergunta:**A quantidade de estudantes que realizam o ENEM aumentou de 2015 para 2019 na capital bahiana?** E a resposta é não. Apesar do número de estudantes crescer de 2015 para 2016, ocorre uma queda drástica até 2019, chegando a diminuir pela metade o número de inscrições em Salvador.

Essa resposta pode te levar a um questionamento mais profundo: O que realmente motivou essa queda?. Infelizmente encontrar a resposta para este questionamento não é trivial, requer pesquisas mais específicas a cerca do tema, o que foge do escopo deste capítulo. Todavia, é interessante refletir como a partir de um gráfico simples, podemos alcançar perguntas ainda mais complexas.

Agora que respondemos a primeira questão, podemos perceber que a pergunta "**Como é a distribuição de estudantes em Salvador por cor/raça? Conseguimos identificar algum padrão para esses valores?**" está bastante relacionada ao seu resultado. Inicialmente para entender essa relação, precisamos entender o que seria essa distribuição de raças no questionário no ENEM. Trata-se de uma pergunta que busca entender como o estudante se classifica em relação a sua cor. Essa pergunta possui 7 respostas possíveis:

- Não declarado
- Pardo

- Preta
- Branco
- Amarelo
- Indígena
- Opção de não apresentar tal informação

Como foi explicado no Capítulo 2, esse questionamento pode ser definido como uma variável categórica. Ele está bastante relacionada com a primeira questão, pois a quantidade de total de estudantes pode alterar essa distribuição, aumentando ou diminuindo a depender das categorias. Como tivemos uma diferença tão grande entre o número de inscritos em 2016 e 2019 demonstrado na Figura 3.1, uma análise mais aprofundada nesses dois anos podem trazer resultados interessantes para responder nosso segundo questionamento:

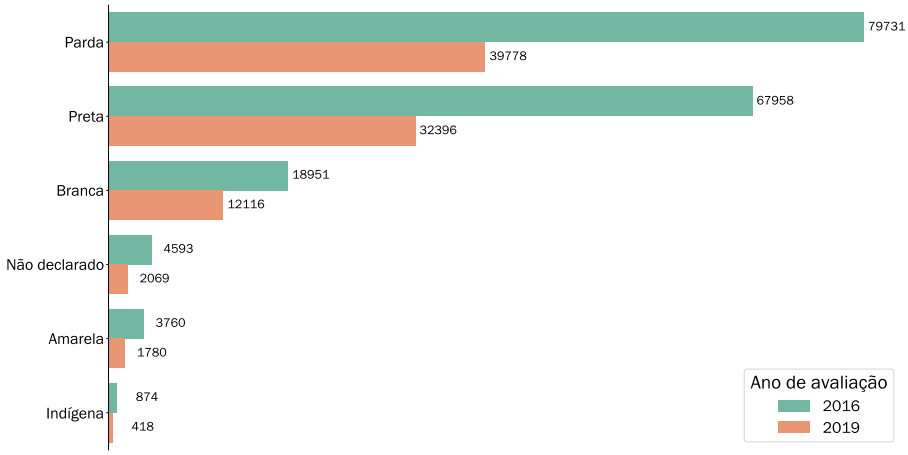


Figure 3.2: Distinção de estudantes por cor/raça da cidade de Salvador para os anos de 2016 e 2019

Através da Figura 3.2, é apresentado os valores absolutos da quantidade de estudantes que realizaram o ENEM em cada ano identificados pela sua raça. Note que a grande queda encontrada na Figura 3.1 se reflete neste gráfico também: Em comparação a 2016, todas as categorias apresentaram valores menores. Por exemplo, a quantidade pessoas pardas que realizaram o ENEM caiu quase pela metade, assim como pessoas auto-declaradas como preta. Além disso, podemos notar uma baixíssima quantidade de pessoas indígenas/amarela que realizam este exame e que em sua grande maioria, os estudantes da capital bahiana se declaram como pardos e negros. Essa situação já era esperada e reflete uma realidade já conhecida: Segundo o Instituto Brasileiro de Estatística e Geografia (IBGE), em uma pesquisa realizada em 2017, Salvador é considerada a capital

mais preta do brasil, onde 8 em cada 10 moradores se autodeclaravam de cor preta ou parda.

Note que a Figura 3.2 demonstra também a principal função do gráfico de barras: dimensionar variáveis categóricas de acordo a frequência de suas categorias. **Frequência** para uma variável categórica pode ser definido com a quantidade de vezes que ela é representada, podendo ser dividida em dois tipos: absoluta e relativa.

A frequência absoluta se trata da representação da quantidade de vezes que cada categoria ocorre. Este tipo de frequência é trabalhada na Figura 3.2, onde apresentamos a quantidade de estudantes por cor/raça que realizaram o ENEM nos anos de 2016 e 2019. Ainda na Figura 3.2, conseguimos notar que para todas as categorias apresentaram uma queda na quantidade de estudantes que realizaram em 2016 para 2019, mas e se quisermos comparar este valores ainda utilizando um gráfico de barras, seria possível?

Uma boa forma para comparar essas frequências absolutas distintas seria através do segundo tipo de frequência apresentada anteriormente: a frequência relativa. A frequência relativa é definida como uma proporção entre o valor que você quer estimar e o valor máximo esperado. Podemos formular este conceito da seguinte forma:

$$Frequncia\ Relativa(\%) = 100 * \binom{Valor\ para\ comparar}{Valor\ mximo}$$

Note que não foi mencionado o valor 100 presente na fórmula. Ele é apresentado para tornar o resultado da frequência relativa em porcentagem. Para compreender melhor este conceito apresentado, vamos continuar respondendo a segunda questão utilizando agora este novo aprendizado:

A Figura 3.3 pode ser vista como uma extensão da Figura 3.2, utilizando a frequência relativa para apresentar uma informação implícita: a proporção dos estudantes que fizeram o ENEM em 2019 em comparação a quantidade que realizaram o exame em 2016. Transcrevendo a fórmula da frequência relativa apresentada anteriormente, temos:

$$Frequncia\ Relativa(\%) = 100 * \binom{Quantidade\ de\ estudantes\ \ realizaram\ o\ ENEM\ em\ 2019}{Quantidade\ de\ estudantes\ \ realizaram\ o\ ENEM\ em\ 2016}$$

Como nos é apresentado uma proporção, podemos ler o gráfico de barras apresentado na Figura 3.3 como sendo **a quantidade de estudantes que fizeram a prova em 2019 em relação a quantidade que realizaram a prova em 2016**. Podemos identificar, por exemplo, que com exceção dos estudantes autodeclarados de cor branca todas as outras cores apresentaram uma proporção de aproximadamente 50%, ou seja, o número de estudantes pardos, pretos, amarelos, indígenas e não declarados caíram pela metade em comparação ao ano de 2016. Esta informação intensifica ainda mais o resultado apresentado na Figura 3.1.

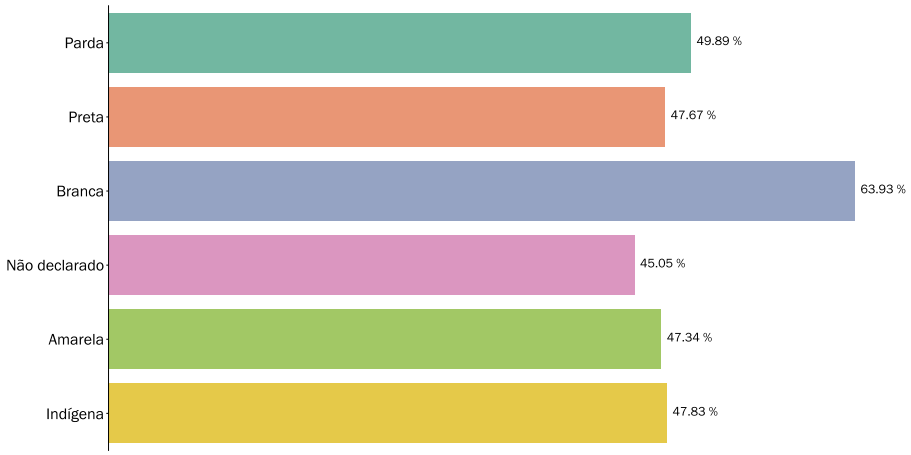


Figure 3.3: Comparação entre os estudantes de Salvador por cor/raça para 2016 e 2019

Através da análise do gráfico de barras conseguimos avaliar dois questionamentos de uma só vez! Porém para analisar como esses resultados ocorreram de 2016 até 2019 ao invés de dois anos separados, qual seria o melhor tipo de gráfico? Iremos explora-lo na próxima seção deste capítulo.

3.3 Séries temporais

3.4 Gráfico de pizza

3.5 Gráficos de Dispersão

3.6 Histograma

3.7 Aprimorando suas visualizações

3.8 Indo Além ...

Chapter 4

Literatura e bibtex

O foco deste capítulo está numa das principais potencialidades do LaTeX utilizas pelo R Markdown a capacidade de citar os documentos adequadamente organizados num arquivo *.bib*, especialmente neste exemplo aproveitamos o arquivo gerado pela estrutura mínima *bookdown*.

Para esta etapa aproveitaremos como exemplo os artigos do projeto organizados na plataforma Mendeley, seguindo os seguintes passos:

1. Abrir Medeley Desktop;
2. Abrir pasta do grupo nomeada por CDnaEP;
3. Selecionar os artigos que pretende citar no seu documento;
4. Clicar com o botão direito do mouse, selecionar *Copy as* e em seguida *BibTex entry*.
5. Abrir o arquivo *book.bib* e colar os metadados dos artigos na última linha do arquivo depois da última chave **}**.
6. Em seguida, salve e feche o arquivo *book.bib*.

As figuras a seguir estão de acordo com a sequência acima apresentada:

Uma informação importante para quem ainda não é familiarizado com LaTeX é o fato da primeira informação dos metados de um artigo dentro do arquivo *.bib* ser a *label*, a informação que será usada para citações ao longo do documento.

Nas subseções a seguir mostramos como citar alguns desses trabalhos.

4.1 Exemplo de citação simples

Vamos considerar a citação do artigo (Partanen et al., 2016).

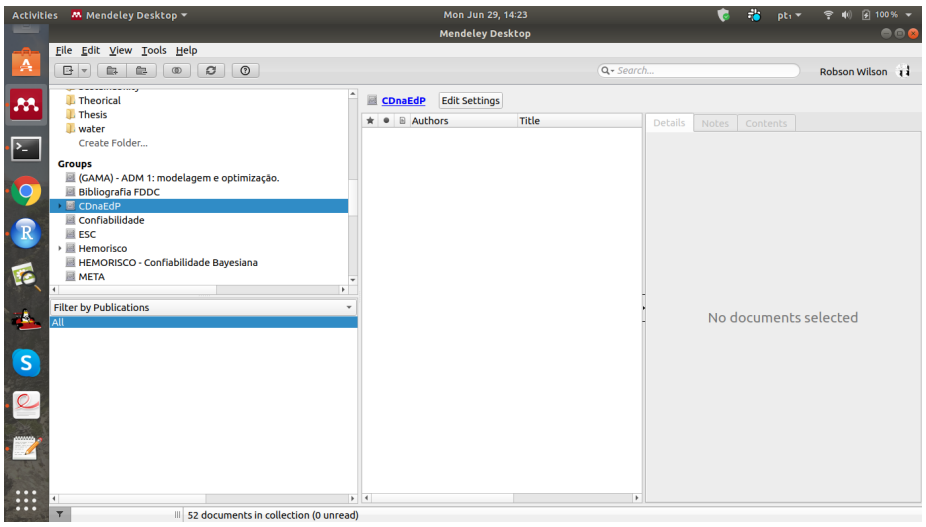


Figure 4.1: Abrir Medeley Desktop e selecionar pasta CDna EP

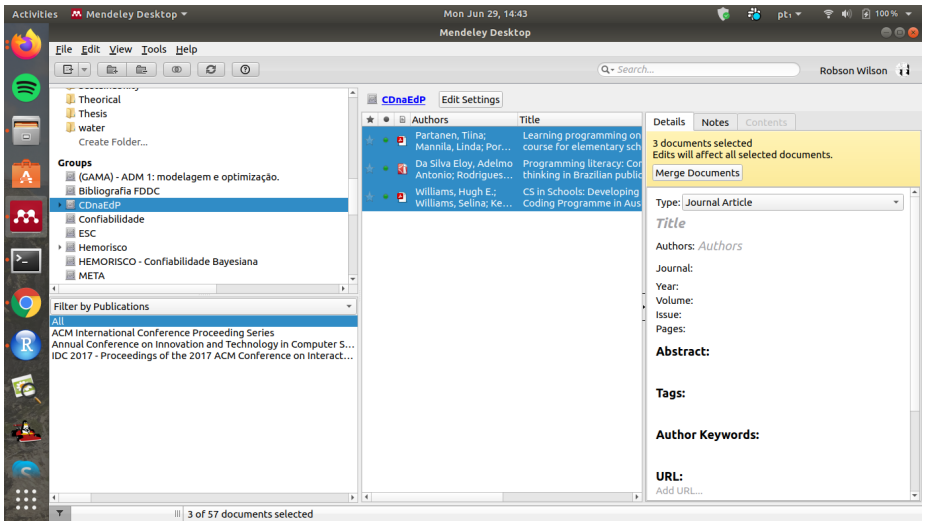


Figure 4.2: Selecionar artigos para citação

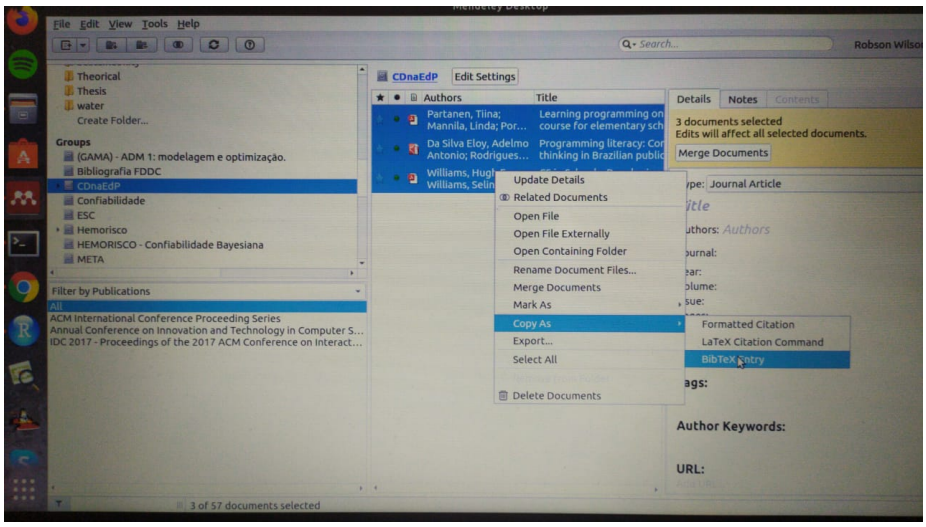


Figure 4.3: Copiar metadados no formato de entrada do bibtex

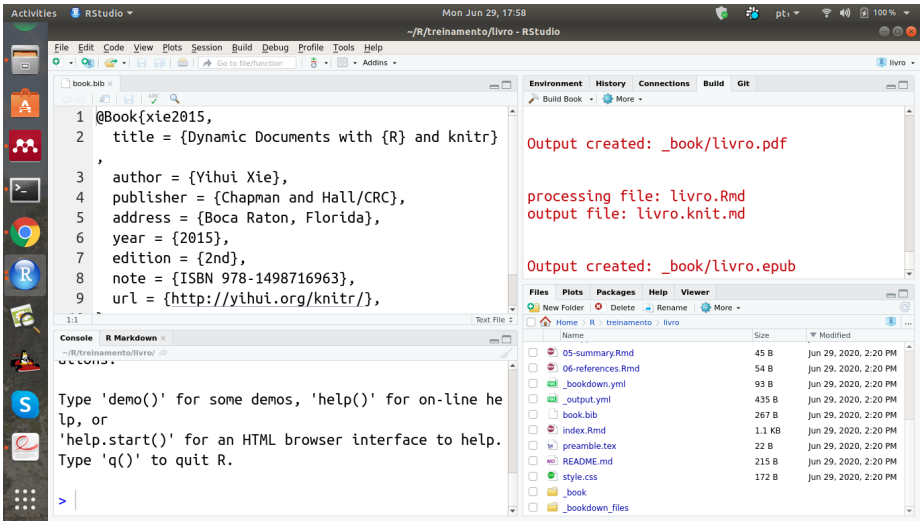


Figure 4.4: Identificação da *label* do livro do ((Xie, 2015))

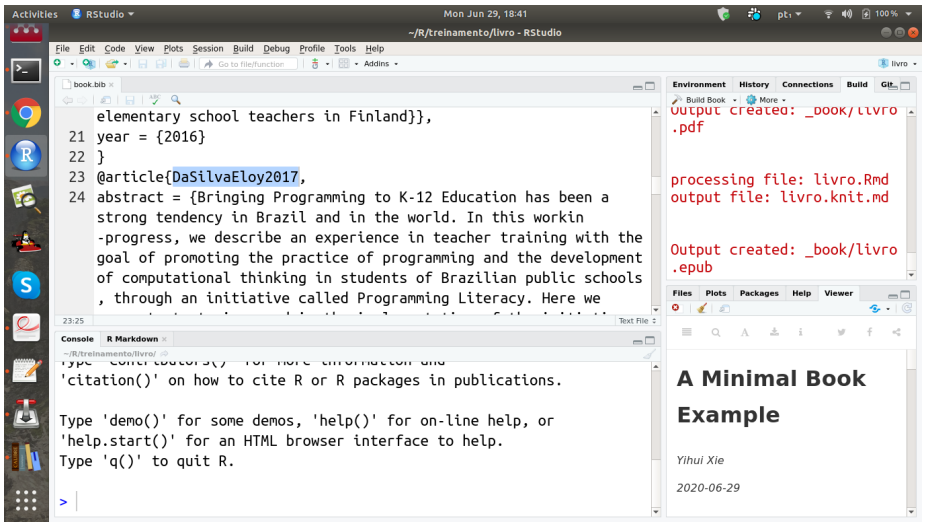


Figure 4.5: Identificação da *label* de cada artigo ((Da Silva Eloy et al., 2017))

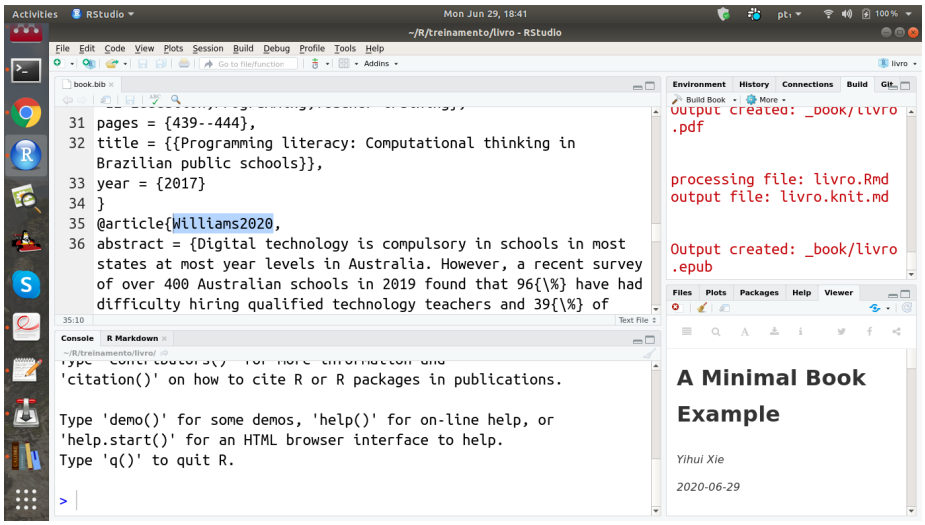


Figure 4.6: Identificação da *label* de cada artigo ((Williams et al., 2020))

4.2 Citações de dois ou mais artigos

Agora incluiremos mais duas citações (Da Silva Eloy et al., 2017, Williams et al. (2020)).

Chapter 5

Aplicações

Neste capítulo apresentaremos alguns exemplos de aplicações de R.

5.1 Exemplo 1

5.1.1 Carregamento de dados

```
##############################################################################################
#1-Carregamento de dados
#1.1-Dados do Covid19
# referencia(22-06-2020) - (https://data.brasil.io/dataset/covid19/_meta/list.html)

library(readr)
```

```
## Error in library(readr): there is no package called 'readr'
```

```
caso <- read_csv("data/caso.csv")
```

```
## Error in read_csv("data/caso.csv"): não foi possível encontrar a função "read_csv"
```

5.2 Análise de dados

5.2.1 Análise Exploratória

```
## Error in library(readr): there is no package called 'readr'
```

```
## Error in library("tidyverse"): there is no package called 'tidyverse'
```

```
## Error in library("tidyr"): there is no package called 'tidyr'
```

```
## Error in library(ggplot2): there is no package called 'ggplot2'
```

```
## Error in read_delim("data/HIST_PAINEL_COVIDBR_21jun2020.csv", ";", escape_double = F
## Error in as.Date(caso_MS$data, "%m/%d/%Y"): objeto 'caso_MS' não encontrado
## Error in eval(lhs, parent, parent): objeto 'caso_MS' não encontrado
## Warning in min(x): nenhum argumento não faltante para min; retornando Inf
## Warning in max(x): nenhum argumento não faltante para max; retornando -Inf
## Warning in min(x): nenhum argumento não faltante para min; retornando Inf
## Warning in max(x): nenhum argumento não faltante para max; retornando -Inf
## Error in plot.window(...): valores finitos são necessários para 'xlim'
```

```
## Error in eval(lhs, parent, parent): objeto 'caso_MS' não encontrado
## Warning in min(x): nenhum argumento não faltante para min; retornando Inf
## Warning in min(x): nenhum argumento não faltante para max; retornando -Inf
## Warning in min(x): nenhum argumento não faltante para min; retornando Inf
## Warning in max(x): nenhum argumento não faltante para max; retornando -Inf
## Error in plot.window(...): valores finitos são necessários para 'xlim'
## Error in UseMethod("weekdays"): método não aplicável para 'weekdays' aplicado a um o
## Error in ggplot(caso_MS_BR, aes(x = data, y = quantidade, fill = dayweek)): não foi
```

Chapter 6

Recomendações finais

A principal sugestão para o contexto desse projeto é que haja um controle para que duas pessoas não trabalhem no mesmo capítulo durante o mesmo período num diretório de github, essa prática pode ampliar demais o trabalho de quem gerencia as pastas do github. Embora haja os recursos de Brunch, se duas ou mais pessoas trabalham num mesmo cápitulo pode se tornar um pouco confuso o merge de capítulos.

6.1 O site principal do Bookdown

\url(https://bookdown.org/)

6.2 Reportagem

\url(https://medium.com/@diegousaiuk/how-i-used-hugo-and-blogdown-to-set-up-my-own-website-e32e2eddbf81)

Treinar Tidyverse

Após o treino dos recursos do tidyverse e especialmente o ggplot apresentados por Ícaro Bernardes, por favor, explorem neste ambinte a inclusão dos exercícios. Procurar pasta de capacitação disponível no github \url(https://github.com/cienciadedadosnaep) . Como recomendado pelo facilitador Ícaro Bernardes, acessar os documentos *Cheat Sheet* no site \url(https://rstudio.com/resources/cheatsheets/).

Tidyverse

- Tidyverse

Ggplot

- ggplot

Bibliography

Da Silva Eloy, A. A., Rodrigues, A., Martins, Q., Pazinato, A. M., De Fatima Polesi Lukjanenko, M. S., and De Deus Lopes, R. (2017). Programming literacy: Computational thinking in Brazilian public schools. *IDC 2017 - Proceedings of the 2017 ACM Conference on Interaction Design and Children*, pages 439–444.

Partanen, T., Mannila, L., and Poranen, T. (2016). Learning programming online: A Racket-course for elementary school teachers in Finland. *ACM International Conference Proceeding Series*, pages 178–179.

Williams, H. E., Williams, S., and Kendall, K. (2020). CS in Schools: Developing a sustainable Coding Programme in Australian Schools. *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, pages 321–327.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.