Ciência de Dados para Todos (Data Science For All) - 2018.2 - Análise da Produção Científica e Acadêmica da Universidade de Brasília - Engenharia Biomédica

Marcos Vinicius Prescendo Tonin, Lucas Nascimento, Carlos Aragão

20/09/2018

Introdução

O presente trabalho tem por finalidade entender e correlacionar dados que possam algum significado mais profundo do que aparentam ter, assim com os dados da pós-graduação em engenharia biomédica em mãos busca-se encontrar principais fatores, professores mais envolvidos, temas mais relevantes e afins.

Metodologia

Para um melhor resultado do trabalho buscou-se seguir e adaptar-se a metodologia CRISP-DM, para isso baseou-se no ciclo de projeto usado pelo CRISP-DM.

O ciclo é basicamente definido pelas seguintes fases:

- **Entendimento do negócio** : primeiramente deve-se entender o que se busca encontrar, haja visto que não faz sentido fazer uma análise de dados sem saber o que se busca.
- **Entendimento dos dados** : Busca-se entender os dados de forma mais superficial primeiramente.
- **Preparação de dados** : Então limpa-se o dado e prepara o dado para que se possa facilitar seu processamento.
- **Modelagem** : Faz a modelagem do dado.
- **Avaliação** : Avalia-se o resultado da modelagem.
- **Implantação** : Faz uso dos dados.

As fases não são independentes entre si, mas possuem certa comunicação, depedendo das fases, além de não ser estritamente sequencial, sendo melhor visualizado na figura abaixo.

Por esta imagem consegue-se perceber que o entendimento do negócio e do dado pode muitas vezes ser alternada, indo e voltando.

Delimitações iniciais

Domínio de Aplicação do projeto

O domínio de aplicação do projeto é produção científica ou produção acadêmica de um subgrupo de pesquisadores vinculados à Universidade de Brasília, sendo vinculado ao subtema **engenharia biomédica**, tais arquivos foram pegos na plataforma elattes.

Problema abordado

Problema abordado tem por finalidade obter dados de forma descritiva, quantitativa e de modelagem computacional ou estatástica, que permitam caracterizar como, porque e também para que ocorre a produção científica e acadêmica na Área de engenharia biomédica.

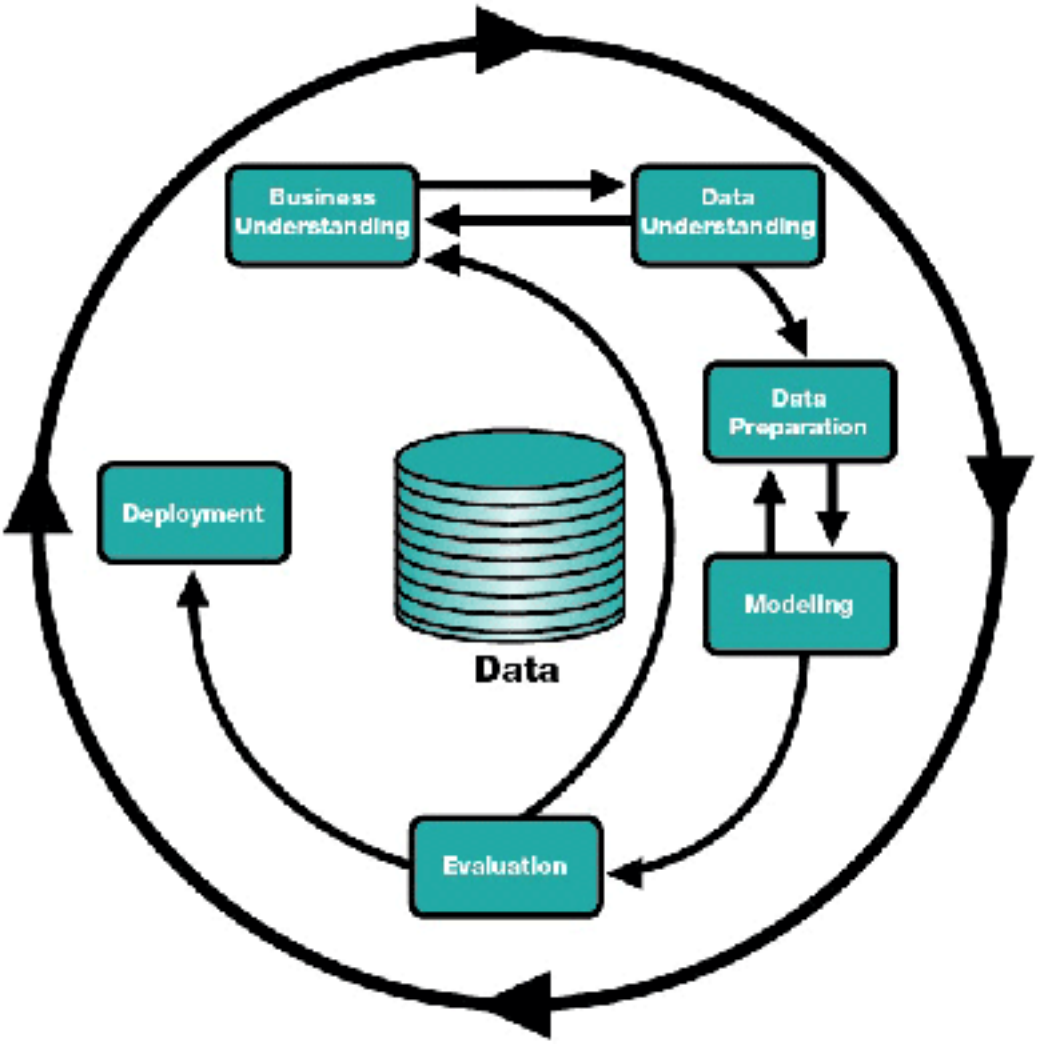


Figure 1: Ciclo do CRISP-DM que será usado como base

CRISP-DM Fase 1 - Entendimento do Negócio

A Universidade de Brasília é uma das grandes universidades federais, sendo responsável pela produção de artigos visando artigos acadêmicos, este presente trabalho de como programas :

Programas de Pós-Graduação Grupo 15

Ids	Área de Avaliação	Nome do Programa	Link Capes	ME	DO	MP
53001010059P8	ENGENHARIAS IV	ENGENHARIA ELÉTRICA	ENGENHARIA ELÉTRICA (53001010059P8)	-	-	3
53001010080P7	ENGENHARIAS IV	ENGENHARIA ELÉTRICA	ENGENHARIA ELÉTRICA (53001010080P7)	4	4	-
53001010081P3	ENGENHARIAS IV	ENGENHARIA DE SISTEMAS ELETRÔNICOS E DE AUTOMAÇÃO	ENGENHARIA DE SISTEMAS ELETRÔNICOS E DE AUTOMAÇÃO (53001010081P3)	4	4	-
53001010083P6	ENGENHARIAS IV	ENGENHARIA BIOMÉDICA	ENGENHARIA BIOMÉDICA (53001010083P6)	3	-	-

Sendo o enfoque principal os dados referente a engenharia biomédica, que por sua vez tem sua pesquisa muitas voltada a novas soluções na área de biomédica.

Avaliação das Circunstancias

Este trabalho terá o limite que os próprios dados gerado pela plataforma elattes tem, já que se os dados fornecidos delimitariam o escopo do projeto e até pode-se ir, ou melhor, até onde os dados permitem uma análise fidedgina.

CRISP-DM Fase 2 - Entendimento dos Dados

CRISP-DM Fase.Atividade 2.1 - Coleta inicial dos dados

Todos os arquivos com dados iniciais a seguir apresentados foram fornecidos pelos professores responsáveis pela disciplina, através da plataforma elattes. Os dados foram gerados no mês de setembro de 2018, data em que o presente grupo começou a realizar o trabalho e compilam informações entre os anos de 2010 e 2017, das áreas expostas na seção anterior. Os arquivos estão no formato JSON, sendo todos fornecidos pelos docentes responsável por esta disciplina.

Perfil profissional dos docentes vinculados às pós-graduações

```
json.perfil <- "dados-2018-2/engenharia-biomedica/279.profile.json"
file.info(json.perfil)
```

```
##                                                    size isdir mode
## dados-2018-2/engenharia-biomedica/279.profile.json 770014 FALSE  666
##                                                                mtime
## dados-2018-2/engenharia-biomedica/279.profile.json 2018-09-22 19:33:32
##                                                                ctime
## dados-2018-2/engenharia-biomedica/279.profile.json 2018-09-22 19:33:32
##                                                                atime exe
## dados-2018-2/engenharia-biomedica/279.profile.json 2018-09-24 19:51:00  no
```

O arquivo dados-2018-2/engenharia-biomedica/279.profile.json apresenta dados sobre o perfil de todos os docentes vinculados a programas de pós-graduação, em engenharia biomédia, da UnB, entre 2010 e 2017.

Orientações de mestrado e doutorado realizadas pelos docentes vinculados às pós-graduações

```
json.advise <- "dados-2018-2/engenharia-biomedica/279.advise.json"
file.info(json.advise)
```

```
##                                                    size isdir mode
## dados-2018-2/engenharia-biomedica/279.advise.json 361237 FALSE  666
##                                                                mtime
## dados-2018-2/engenharia-biomedica/279.advise.json 2018-09-22 19:33:32
##                                                                ctime
## dados-2018-2/engenharia-biomedica/279.advise.json 2018-09-22 19:33:32
##                                                                atime exe
## dados-2018-2/engenharia-biomedica/279.advise.json 2018-09-24 19:51:01  no
```

O arquivo dados-2018-2/engenharia-biomedica/279.advise.json apresenta dados sobre o orientações de mestrado e doutorado feitas por todos os docentes vinculados a programas de pós-graduação em engenharia biomédica, da UnB, entre 2010 e 2017.

Produção bibliográfica gerada pelos docentes vinculados Ã s pós-graduações

```
json.producao.bibliografica <- "dados-2018-2/engenharia-biomedica/279.publication.json"
file.info(json.producao.bibliografica)
```

```
##                                                      size isdir mode
## dados-2018-2/engenharia-biomedica/279.publication.json 294169 FALSE  666
##                                                                  mtime
## dados-2018-2/engenharia-biomedica/279.publication.json 2018-09-22 19:33:32
##                                                                  ctime
## dados-2018-2/engenharia-biomedica/279.publication.json 2018-09-22 19:33:32
##                                                                  atime
## dados-2018-2/engenharia-biomedica/279.publication.json 2018-09-24 19:51:01
##                                                                 exe
## dados-2018-2/engenharia-biomedica/279.publication.json  no
```

O arquivo dados-2018-2/engenharia-biomedica/279.publication.json apresenta dados sobre a produção bibliográfica gerada por todos os docentes vinculados a programas de pós-graduação, em engenharia biomédica, da UnB, entre 2010 e 2017.

ID's dos docentes participantes e o que contêm o arquivo list.json

```
json.list <- "dados-2018-2/engenharia-biomedica/279.list.json"
file.info(json.list)
```

```
##                                                 size isdir mode
## dados-2018-2/engenharia-biomedica/279.list.json  943 FALSE  666
##                                                              mtime
## dados-2018-2/engenharia-biomedica/279.list.json 2018-09-22 19:33:32
##                                                              ctime
## dados-2018-2/engenharia-biomedica/279.list.json 2018-09-22 19:33:32
##                                                              atime exe
## dados-2018-2/engenharia-biomedica/279.list.json 2018-09-24 19:51:01  no
```

O arquivo dados-2018-2/engenharia-biomedica/279.list.json apresenta o id de todos docentes vinculados a programas de pós-graduação, em engenharia biomédica, da UnB, entre 2010 e 2017.Porém tal arquivo se mostra inútil, haja visto que só contém o id e que as outras variáveis como **"nome"** sempre tem seu valor igual a **""**.

Redes de colaboração entre docentes

```
json.graph<-'dados-2018-2/engenharia-biomedica/279.graph.json'
file.info(json.graph)
```

```
##                                                    size isdir mode
## dados-2018-2/engenharia-biomedica/279.graph.json 3901 FALSE  666
##                                                               mtime
## dados-2018-2/engenharia-biomedica/279.graph.json 2018-09-22 19:33:32
##                                                               ctime
## dados-2018-2/engenharia-biomedica/279.graph.json 2018-09-22 19:33:32
##                                                               atime exe
## dados-2018-2/engenharia-biomedica/279.graph.json 2018-09-24 19:51:01  no
```

O arquivo dados-2018-2/engenharia-biomedica/279.graph.json apresenta redes de colaboração na co-autoria de artigos científicos, feitas entre os docentes vinculados a programas de pós-graduação da UnB, entre 2010 e 2017.

CRISP-DM Fase.Atividade 2.2 - Descrição dos Dados

Para ler e manipular inicialmente esses dados, serão usadas primordialmente as bibliotecas seguintes

```
library(jsonlite)
library(listviewer)
library(readxl)
library(readr)
library(readtext)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v purrr   0.2.5     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------ tidy
## x dplyr::filter()  masks stats::filter()
## x purrr::flatten() masks jsonlite::flatten()
## x dplyr::lag()     masks stats::lag()
```

```
library(stringr)
```

Com estas bibliotecas seremos capazes de responder e determinar qual o volume de dados, a estrutura dos dados (tipos), codificações usadas, etc..

Descrição dos dados do perfil

```
unb.prof <- fromJSON("dados-2018-2/engenharia-biomedica/279.profile.json")
```

A quantidade de docentes sob análise é apresentada a seguir.

```
length(unb.prof)
```

```
## [1] 15
```

Para um melhor entendimento é importante saber como os dados estão dispostos para algum número de ocorrência.

```
### Usando glimpse
glimpse(unb.prof[[1]], width = 30)
```

```
## List of 7
##  $ nome                  : chr "Leandro Xavier Cardoso"
##  $ resumo_cv             : chr "Possui graduação em Física Bacharelado e Licenciatura pela Universida
##  $ areas_de_atuacao      :'data.frame':  1 obs. of  4 variables:
##   ..$ grande_area  : chr ""
##   ..$ area         : chr ""
##   ..$ sub_area     : chr ""
##   ..$ especialidade: chr ""
##  $ endereco_profissional :List of 8
##   ..$ instituicao: chr "UnB - UNIVERSIDADE DE BRASÃ?LIA - FGA - FACULDADE GAMA"
##   ..$ orgao      : chr ""
##   ..$ unidade    : chr ""
##   ..$ DDD        : chr "61"
##   ..$ telefone   : chr "31078903"
##   ..$ bairro     : chr "Setor Leste (Gama)"
##   ..$ cep        : chr "72444240"
##   ..$ cidade     : chr "Brasília"
##  $ producao_bibiografica :List of 2
##   ..$ EVENTO   :'data.frame':    3 obs. of  11 variables:
##   .. ..$ natureza        : chr [1:3] "COMPLETO" "COMPLETO" "COMPLETO"
##   .. ..$ titulo          : chr [1:3] "As Tic's na Educação: Mudnça ou Modernização" "A Educação no Co
##   .. ..$ nome_do_evento  : chr [1:3] "IV Colóquio Internacional Educação e Contemporaneidade" "IV Col
##   .. ..$ ano_do_trabalho : chr [1:3] "2010" "2010" "2014"
##   .. ..$ pais_do_evento  : chr [1:3] "Brasil" "Brasil" "Brasil"
##   .. ..$ cidade_do_evento: chr [1:3] "São Cristovão" "São Cristovão" "Uberlândia"
##   .. ..$ doi             : chr [1:3] "" "" ""
##   .. ..$ classificacao   : chr [1:3] "INTERNACIONAL" "INTERNACIONAL" "NACIONAL"
##   .. ..$ paginas         : chr [1:3] " - " " - " "2636 - 2638"
##   .. ..$ autores         :List of 3
##   .. ..$ autores-endogeno:List of 3
##   ..$ PERIODICO:'data.frame':    10 obs. of  10 variables:
##   .. ..$ natureza        : chr [1:10] "COMPLETO" "COMPLETO" "COMPLETO" "COMPLETO" ...
##   .. ..$ titulo          : chr [1:10] "Thermoluminescent dose reconstruction using quartz extracted f
##   .. ..$ periodico       : chr [1:10] "Journal of Physics. Conference Series (Online)" "Scientia Ple
##   .. ..$ ano             : chr [1:10] "2010" "2011" "2011" "2013" ...
##   .. ..$ volume          : chr [1:10] "249" "7" "80" "477" ...
##   .. ..$ issn            : chr [1:10] "17426596" "18082793" "20103778" "17426588" ...
##   .. ..$ paginas         : chr [1:10] "012031 - " "014101 - " "285 - 290" "012011 - " ...
##   .. ..$ doi             : chr [1:10] "10.1088/1742-6596/249/1/012031" "" "" "10.1088/1742-6596/477/1
##   .. ..$ autores         :List of 10
##   .. ..$ autores-endogeno:List of 10
##  $ orientacoes_academicas:List of 2
##   ..$ ORIENTACAO_EM_ANDAMENTO_MESTRADO:'data.frame': 5 obs. of  13 variables:
##   .. ..$ natureza                   : chr [1:5] "Dissertação de mestrado" "Dissertação de mestrado"
##   .. ..$ titulo                     : chr [1:5] "Utilização de dosímetros termoluminescentes comercia
```

```
##   .. ..$ ano                        : chr [1:5] "2014" "2014" "2015" "2015" ...
##   .. ..$ id_lattes_aluno            : chr [1:5] "" "" "" "" ...
##   .. ..$ nome_aluno                 : chr [1:5] "Rafael Assunção Gomes de Souza" "Marcelo Oppermann"
##   .. ..$ instituicao                : chr [1:5] "Universidade de Brasília" "Universidade de Brasília
##   .. ..$ curso                      : chr [1:5] "Engenharia Biomédica" "Engenharia Biomédica" "Engenh
##   .. ..$ codigo_do_curso            : chr [1:5] "60059672" "90000006" "90000006" "60059672" ...
##   .. ..$ bolsa                      : chr [1:5] "NAO" "NAO" "NAO" "NAO" ...
##   .. ..$ agencia_financiadora       : chr [1:5] "" "" "" "" ...
##   .. ..$ codigo_agencia_financiadora: chr [1:5] "" "" "" "" ...
##   .. ..$ nome_orientadores          :List of 5
##   .. ..$ id_lattes_orientadores     :List of 5
##   ..$ OUTRAS_ORIENTACOES_CONCLUIDAS   :'data.frame': 5 obs. of  13 variables:
##   .. ..$ natureza                   : chr [1:5] "MONOGRAFIA_DE_CONCLUSAO_DE_CURSO_APERFEICOAMENTO_E_
##   .. ..$ titulo                     : chr [1:5] "Insuficiência na Aprendizagem de Matemática do 9º a
##   .. ..$ ano                        : chr [1:5] "2011" "2011" "2011" "2011" ...
##   .. ..$ id_lattes_aluno            : chr [1:5] "" "" "" "" ...
##   .. ..$ nome_aluno                 : chr [1:5] "Décio Luiz Alves Barreto e outros" "Izaque dos Sant
##   .. ..$ instituicao                : chr [1:5] "Faculdade Serigy" "Faculdade Serigy" "Faculdade Ser
##   .. ..$ curso                      : chr [1:5] "Fundamentos e Métodos do ensino da Matemática" "Fun
##   .. ..$ codigo_do_curso            : chr [1:5] "90000005" "90000005" "90000005" "90000005" ...
##   .. ..$ bolsa                      : chr [1:5] "NAO" "NAO" "NAO" "NAO" ...
##   .. ..$ agencia_financiadora       : chr [1:5] "" "" "" "" ...
##   .. ..$ codigo_agencia_financiadora: chr [1:5] "" "" "" "" ...
##   .. ..$ nome_orientadores          :List of 5
##   .. ..$ id_lattes_orientadores     :List of 5
##  $ senioridade           : chr "9"
```

Podemos inferir que:

- Que o professor não é da área da engenharia elétrica, por formação, mas acabou por aderir a subárea engenharia biomédica.
- Não é nativo da UnB, sendo formado no UFG.
- Atualmente trabalha na UnB do Gama.
- Sua senioridade é de 9.

Potencial de utilização dos dados do perfil dos docentes

Descrição dos dados de orientações

```
unb.adv <- fromJSON("dados-2018-2/engenharia-biomedica/279.advise.json")
# Mostrando as listas presentes neste arquivo.
names(unb.adv)
```

```
## [1] "ORIENTACAO_EM_ANDAMENTO_DE_POS_DOUTORADO"
## [2] "ORIENTACAO_EM_ANDAMENTO_DOUTORADO"
## [3] "ORIENTACAO_EM_ANDAMENTO_MESTRADO"
## [4] "ORIENTACAO_EM_ANDAMENTO_GRADUACAO"
## [5] "ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA"
## [6] "ORIENTACAO_CONCLUIDA_POS_DOUTORADO"
## [7] "ORIENTACAO_CONCLUIDA_DOUTORADO"
## [8] "ORIENTACAO_CONCLUIDA_MESTRADO"
## [9] "OUTRAS_ORIENTACOES_CONCLUIDAS"
```

```
# Explorando um nível de detalhe de Orientações de doutorados concluídas
names(unb.adv$ORIENTACAO_CONCLUIDA_DOUTORADO)
```

```
## [1] "2010" "2011" "2012" "2013" "2014" "2015" "2016" "2017"
```

```
############## DOUTORADO 2017
#Buscando cursos que mais produziram doutorados.
head(sort(table(unb.adv$ORIENTACAO_CONCLUIDA_DOUTORADO$`2017`$curso), decreasing = TRUE), 10)
```

```
##
## ENGENHARIA DE SISTEMAS ELETRÔNICOS E DE AUTOMAÇÃO
##                                                 1
##                 Pós-graduação em Ciências Médicas
##                                                 1
```

```
#Sabendo suas instituições
head(sort(table(unb.adv$ORIENTACAO_CONCLUIDA_DOUTORADO$`2017`$instituicao), decreasing = TRUE), 10)
```

```
##
## Universidade de BrasÃlia - Campus Darcy Ribeiro
##                                               1
##                         Universidade de Brasília
##                                               1
```

```
#Sabendo seus orientadores
data_orienM <- capture.output(str(unb.adv$ORIENTACAO_CONCLUIDA_DOUTORADO$`2017`$nome_orientadores))
unique(data_orienM)
```

```
## [1] "List of 2"
## [2] " $ : chr \"Marcus Vinícius Chaffim Costa\""
## [3] " $ : chr \"Marilia Miranda Forte Gomes\""
```

```
############## MESTRADO 2017
#Buscando cursos que mais produziram mestrados.
head(sort(table(unb.adv$ORIENTACAO_CONCLUIDA_MESTRADO$`2017`$curso), decreasing = TRUE), 10)
```

```
##
##             Engenharia Biomédica Mestrado em Engenharia Biomédica
##                                9                                 4
##              ENGENHARIA ELÉTRICA                           Física
##                                1                                 1
##                       PPDSCI/CEAM
##                                1
```

```
head(sort(table(unb.adv$ORIENTACAO_CONCLUIDA_MESTRADO$`2017`$instituicao), decreasing = TRUE), 10)
```

```
##
##                         Faculdade do Gama da UnB
##                                                6
##                         Universidade de Brasília
##                                                5
##    Universidade de Brasília - Faculdade UnB-Gama
##                                                3
##                         Faculdade UNB Gama - FGA
##                                                1
## Universidade de BrasÃlia - Campus Darcy Ribeiro
##                                                1
```

```
#Sabendo suas instituições
head(sort(table(unb.adv$ORIENTACAO_CONCLUIDA_MESTRADO$`2017`$instituicao), decreasing = TRUE), 10)
```

```
##
```

```
##                          Faculdade do Gama da UnB
##                                                 6
##                          Universidade de Brasília
##                                                 5
##    Universidade de Brasília - Faculdade UnB-Gama
##                                                 3
##                          Faculdade UNB Gama - FGA
##                                                 1
## Universidade de BrasÃlia - Campus Darcy Ribeiro
##                                                 1
```

```
#Sabendo seus orientadores
data_orien <- capture.output(str(unb.adv$ORIENTACAO_CONCLUIDA_MESTRADO$`2017`$nome_orientadores))
unique(data_orien)
```

```
## [1] "List of 16"
## [2] " $ : chr \"Georges Daniel Amvame Nze\""
## [3] " $ : chr \"Jose Felicio da Silva\""
## [4] " $ : chr \"Lourdes Mattos Brasil\""
## [5] " $ : chr \"Marcelino Monteiro de Andrade\""
## [6] " $ : chr \"Marilia Miranda Forte Gomes\""
## [7] " $ : chr \"Ronni Geraldo Gomes de Amorim\""
```

Como se pode perceber apenas u, professor que orientou doutorado concluido em 2017 (“Marilia Miranda Forte Gomes”) também fez parte dos professores que orientaram no mestrado:

- “Georges Daniel Amvame Nze”
- “Jose Felicio da Silva”
- “Lourdes Mattos Brasil”
- “Marcelino Monteiro de Andrade”
- “Marilia Miranda Forte Gomes”
- “Ronni Geraldo Gomes de Amorim”

Sendo o professor “Marcus Vinícius Chaffim Costa” responsável apenas por orientação de doutorado em 2017.

Descrição dos dados list.json

```
unb.list <- fromJSON("dados-2018-2/engenharia-biomedica/279.list.json")
#analisando a quantidade de elementos presente em list.json
length(unb.list$fiocruz$id)
```

```
## [1] 15
```

```
# Mostrando alguns id, verificar que eles são diferentes
d_listID <- unb.list$fiocruz$id
unique(d_listID)
```

```
##  [1] "1141716826787805" "0535100751136568" "5330755818114960"
##  [4] "5810353896294133" "2957228356035337" "0201204222182378"
##  [7] "9190489069187153" "4739013535126469" "4839052902231824"
## [10] "1524924375222848" "9169095482512290" "4086384842130773"
## [13] "5928104758017036" "1154673226500318" "7294738832905991"
```

```
#Mostrando nome sempre igual ""
d_listNO <- unb.list$fiocruz$nome
unique(d_listNO)
```

```
## [1] ""
```

```
#Mostrando periodo sempre igual 2010-2017
d_listDT <- unb.list$fiocruz$periodo
unique(d_listDT)
```

```
## [[1]]
## [1] "2010" "2017"
```

Feito a análise, percebe-se que apenas o campo id que muda, tanto **"nome"** (obtendo sempre o valor "") e **"periodo"**(obtendo sempre o valor["2010","2017"]), por isso considera-se este arquivo JSON como inútil, haja visto que não dá para obter quaisquer dado plaúsivel, somente os id.

Descrição dos dados de produção bibliográfica

```
unb.pub <- fromJSON("dados-2018-2/engenharia-biomedica/279.publication.json")
# Verificando os tipos de produções que existe.
names(unb.pub)
```

```
## [1] "PERIODICO"
## [2] "LIVRO"
## [3] "CAPITULO_DE_LIVRO"
## [4] "TEXTO_EM_JORNAIS"
## [5] "EVENTO"
## [6] "ARTIGO_ACEITO"
## [7] "DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA"
```

```
#Analisando o qual tipo de informação se tem em periódicos no ano 2012.
names(unb.pub$PERIODICO$`2012`)
```

```
##  [1] "natureza"         "titulo"           "periodico"
##  [4] "ano"              "volume"           "issn"
##  [7] "paginas"          "doi"              "autores"
## [10] "autores-endogeno"
```

```
#Nomes dos periódicos em que tiveram mais publicações na área de engenharia biomédica.
head(sort(table(unb.pub$PERIODICO$`2017`$periodico), decreasing = TRUE), 10)
```

```
##
## International Journal of Biosensors & Bioelectronics
##                                                    2
##      Revista Brasileira de Ensino de Física (Online)
##                                                    2
##                                  ACTA PAUL DE ENFERM
##                                                    1
##                                ADOLESCÊNCIA & SAÚDE
##                                                    1
##                           Advanced Materials Letters
##                                                    1
##                        Advances in High Energy Physics
##                                                    1
##                      AFRICAN JOURNAL OF BIOTECHNOLOGY
##                                                    1
##                                     ANNALS OF PHYSICS
##                                                    1
##                          ARTEFACTUM (RIO DE JANEIRO)
##                                                    1
```

```
##                             CADERNO DE FÍSICA DA UEFS
##                                                     1
```

```
#Nomes dos periódicos em que tiveram mais publicações de artigos aceitos na área de engenharia biomédic
head(sort(table(unb.pub$ARTIGO_ACEITO$`2017`$periodico), decreasing = TRUE), 10)
```

```
##
##                       Acta Paulista de Enfermagem
##                                                 1
##                        Advanced Materials Letters
##                                                 1
##               Ciência e Saúde Coletiva (Impresso)
##                                                 1
## IEEE Journal of Biomedical and Health Informatics
##                                                 1
##                    IEEE Latin America Transactions
##                                                 1
##         JOURNAL OF NANOSCIENCE AND NANOTECHNOLOGY
##                                                 1
```

```
#Nomes dos autores que produziram um tipo de produção que não estava contemplada .
head(sort(table(unb.pub$DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA$`2011`$autores), decreasing = TRUE), 10)
```

```
##
##
##
## GOMES, Marília Miranda Forte;GOMES, MARÍLIA MIRANDA FORTE;GOMES, MARÍLIA M F;GOMES, Marília Miranda
##
##
##
##
##
##
##
##
##
```

Descrição dos dados de redes de colaboração

```
unb.graph <- fromJSON("dados-2018-2/engenharia-biomedica/279.graph.json")
# Suas variáveis
names(unb.graph)
```

```
## [1] "label" "nodes" "links"
```

```
# Quantidade de nós
length(unb.graph$nodes$id)
```

```
## [1] 15
```

```
# Quantidade de links de fonte
length(unb.graph$links$source)
```

```
## [1] 28
```

```
# Quantidade de links de chegada
length(unb.graph$links$target)
```

```
## [1] 28
```

```
#exemplos de pesos da aresta
str(unb.graph$links$weigth)
```

```
##  chr [1:28] "1" "3" "9" "1" "2" "3" "1" "1" "5" "15" "1" "3" "9" "1" ...
```

CRISP-DM Fase.Atividade 2.3 - Análise exploratória dos dados

Arquivo Profile

```
# Total de áreas de atuação de todos profissionais
sum(sapply(unb.prof, function(x) nrow(x$areas_de_atuacao)))
```

```
## [1] 57
```

```
# Número de pessoas por grande area
table(unlist(sapply(unb.prof, function(x) (x$areas_de_atuacao$grande_area))))
```

```
##
##                                  CIENCIAS_BIOLOGICAS
##                            1                        4
##          CIENCIAS_DA_SAUDE CIENCIAS_EXATAS_E_DA_TERRA
##                            7                        9
## CIENCIAS_SOCIAIS_APLICADAS                ENGENHARIAS
##                            1                       34
##                       OUTROS
##                            1
```

```
# Número de pessoas que produziram os tipos de produção específico
table(unlist(sapply(unb.prof, function(x) names(x$producao_bibiografica))))
```

```
##
##                           ARTIGO_ACEITO
##                                       7
##                       CAPITULO_DE_LIVRO
##                                       9
## DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA
##                                       1
##                                  EVENTO
##                                      14
##                                   LIVRO
##                                       3
##                                PERIODICO
##                                      15
##                          TEXTO_EM_JORNAIS
##                                       1
```

```
# Número de publicações por tipo
#####ARTIGO ACEITO################
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$ARTIGO_ACEITO$ano)))
```

```
## [1] 7
```

```
# Número de pessoas por quantitativo de produções por pessoa 0 = 1; 1 = 2...
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$ARTIGO_ACEITO$ano))))
```

```
##
```

```
## 0 1
## 8 7
```

```
#####CAPITULO DE LIVRO###############
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$CAPITULO_DE_LIVRO$ano)))
```

```
## [1] 47
```

```
# Número de pessoas por quantitativo de produções por pessoa 0 = 1; 1 = 2...
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$CAPITULO_DE_LIVRO$ano))))
```

```
##
##  0  1  3  5  9 17
##  6  4  1  1  2  1
```

```
#####LIVRO###############
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$LIVRO$ano)))
```

```
## [1] 3
```

```
# Número de pessoas por quantitativo de produções por pessoa 0 = 1; 1 = 2...
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$LIVRO$ano))))
```

```
##
##  0  1
## 12  3
```

```
#####PERIÓDICO###############
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$PERIODICO$ano)))
```

```
## [1] 172
```

```
# Número de pessoas por quantitativo de produções por pessoa 0 = 1; 1 = 2...
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$PERIODICO$ano))))
```

```
##
##  1  2  3  4  6  8  9 10 11 13 16 22 29 36
##  1  2  1  1  1  1  1  1  1  1  1  1  1  1
```

```
#####TEXTO EM JORNAIS###############
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$TEXTO_EM_JORNAIS$ano)))
```

```
## [1] 1
```

```
# Número de pessoas por quantitativo de produções por pessoa 0 = 1; 1 = 2...
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$TEXTO_EM_JORNAIS$ano))))
```

```
##
##  0  1
## 14  1
```

```
#####DEMAIS TIPOS###############
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA$a
```

```
## [1] 1
```

```
# Número de pessoas por quantitativo de produções por pessoa 0 = 1; 1 = 2...
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$DEMAIS_TIPOS_DE_PRODUCAO$ano))
```

```
##
##  0  1
## 14  1
```

```r
# Número de produções por ano
table(unlist(sapply(unb.prof, function(x) (x$producao_bibiografica$ARTIGO_ACEITO$ano))))
```

```
##
## 2015 2017
##    1    6
```

```r
table(unlist(sapply(unb.prof, function(x) (x$producao_bibiografica$CAPITULO_DE_LIVRO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##    1    3    8    2    2   21    5    5
```

```r
table(unlist(sapply(unb.prof, function(x) (x$producao_bibiografica$LIVRO$ano))))
```

```
##
## 2015 2016 2017
##    1    1    1
```

```r
table(unlist(sapply(unb.prof, function(x) (x$producao_bibiografica$PERIODICO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##   14    9   10   29   18   39   21   32
```

```r
table(unlist(sapply(unb.prof, function(x) (x$producao_bibiografica$TEXTO_EM_JORNAIS$ano))))
```

```
##
## 2010
##    1
```

```r
# Número de pessoas que realizaram diferentes tipos de orientações
length(unlist(sapply(unb.prof, function(x) names(x$orientacoes_academicas))))
```

```
## [1] 60
```

```r
# Número de pessoas por tipo de orientação
table(unlist(sapply(unb.prof, function(x) names(x$orientacoes_academicas))))
```

```
##
##                ORIENTACAO_CONCLUIDA_DOUTORADO
##                                             6
##                 ORIENTACAO_CONCLUIDA_MESTRADO
##                                            13
##            ORIENTACAO_CONCLUIDA_POS_DOUTORADO
##                                             1
##             ORIENTACAO_EM_ANDAMENTO_DOUTORADO
##                                             2
##             ORIENTACAO_EM_ANDAMENTO_GRADUACAO
##                                             3
## ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA
##                                             7
##              ORIENTACAO_EM_ANDAMENTO_MESTRADO
##                                            13
##                 OUTRAS_ORIENTACOES_CONCLUIDAS
##                                            15
```

```r
#Número de orientações concluidas
sum(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano)))
```

```
## [1] 109
```

```
sum(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano)))
```

```
## [1] 14
```

```
sum(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano)
```

```
## [1] 2
```

```
# Número de pessoas por quantitativo de orientações por pessoa 0 = 1; 1 = 2...
table(unlist(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$
```

```
##
##  0  1  3  6  7  9 11 18 19
##  2  1  2  1  3  3  1  1  1
```

```
table(unlist(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO
```

```
##
## 0 1 2 3 4
## 9 2 1 2 1
```

```
table(unlist(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTO
```

```
##
##  0  2
## 14  1
```

```
# Número de orientações por ano
table(unlist(sapply(unb.prof, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##    7    4   10   12   23   24   13   16
```

```
table(unlist(sapply(unb.prof, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano)
```

```
##
## 2011 2012 2013 2014 2015 2017
##    2    2    6    1    1    2
```

```
table(unlist(sapply(unb.prof, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$a
```

```
##
## 2014 2017
##    1    1
```

Arquivo Publicação

```
#Criando um data-frame com todos os anos
unb.pub.df <- data.frame()
for (i in 1:length(unb.pub[[1]]))
  unb.pub.df <- rbind(unb.pub.df, unb.pub$PERIODICO[[i]])
glimpse(unb.pub.df)
```

```
## Observations: 147
## Variables: 10
## $ natureza           <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
```

```
## $ titulo             <chr> "Prótese para controle de fluxo esofagiano ...
## $ periodico          <chr> "Revista Brasileira de Engenharia Biomédica...
## $ ano                <chr> "2010", "2010", "2010", "2010", "2010", "20...
## $ volume             <chr> "26", "249", "20", "11", "14", "9", "53", "...
## $ issn               <chr> "15173151", "17426596", "10546618", "151942...
## $ paginas            <chr> "49 - 54", "012031 - ", "192 - 200", "23 - ...
## $ doi                <chr> "", "10.1088/1742-6596/249/1/012031", "10.1...
## $ autores            <list> [<"Rosa, S. S. R. F.", "da Rocha, A F", "B...
## $ `autores-endogeno` <list> ["1141716826787805", "0201204222182378", "...
```

```
# Limpando o data-frame de listas
unb.pub.df$autores <- gsub("\",\"|\", \"", "; ", unb.pub.df$autores)
unb.pub.df$autores <- gsub("\"|c\\(|\\)", "", unb.pub.df$autores)
unb.pub.df$`autores-endogeno` <- gsub(",", ";", unb.pub.df$`autores-endogeno`)
unb.pub.df$`autores-endogeno` <- gsub("\"|c\\(|\\)", "", unb.pub.df$`autores-endogeno`)
glimpse(unb.pub.df)
```

```
## Observations: 147
## Variables: 10
## $ natureza           <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo             <chr> "Prótese para controle de fluxo esofagiano ...
## $ periodico          <chr> "Revista Brasileira de Engenharia Biomédica...
## $ ano                <chr> "2010", "2010", "2010", "2010", "2010", "20...
## $ volume             <chr> "26", "249", "20", "11", "14", "9", "53", "...
## $ issn               <chr> "15173151", "17426596", "10546618", "151942...
## $ paginas            <chr> "49 - 54", "012031 - ", "192 - 200", "23 - ...
## $ doi                <chr> "", "10.1088/1742-6596/249/1/012031", "10.1...
## $ autores            <chr> "Rosa, S. S. R. F.; da Rocha, A F; Brasil, ...
## $ `autores-endogeno` <chr> "1141716826787805", "0201204222182378", "91...
```

```
### ARQUIVO PROCESSADO E LIMPO
## Publicações por natureza - todas foram completas
table(unb.pub.df$natureza)
```

```
##
## COMPLETO
##      147
```

```
## Publicações 2010 até 2017
table(unb.pub.df$ano)
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##   12    8    9   27   14   31   20   26
```

```
## Publicações por periódico, mostrando os top-5 na área de biomédica
d <- table(unb.pub.df$periodico)
head(sort(d,decreasing = TRUE),n=5)
```

```
##
##         Revista Brasileira de Engenharia Biomédica (Impresso)
##                                                             7
## Global Journal of Engineering Science and Research Management
##                                                             6
##                        Biomedical Engineering Online (Online)
##                                                             4
##               Revista Brasileira de Ensino de Física (Online)
```

```
##                                                          4
##           Revista Brasileira de Inovação Tecnologica em Saúde
##                                                          4
```

```
## Mostrando os autores que tiveram mais publicação
### Neste caso é importante mostra que como a não uniformidade de como se escreve acaba por tornar este
head(sort(table(toupper(unlist(strsplit(unb.pub.df$autores,";")))),decreasing = TRUE),n=10)
```

```
##
##                          AMORIM, R. G. G.
##                                        17
##                          AMORIM, R. G. G.
##                                        16
##                       MARÃES, V. R. F. S.
##                                        14
##                             BRASIL, L. M.
##                                         7
##                       MARÃES, V. R. F. S.
##                                         7
##                              SILVA, W. B.
##                                         7
##              GOMES, MARÍLIA MIRANDA FORTE
##                                         7
## ROSA, SUÉLIA DE SIQUEIRA RODRIGUES FLEURY
##                                         7
##                            DA ROCHA, A. F.
##                                         6
##              GOMES, MARÍLIA MIRANDA FORTE
##                                         6
```

```
## Mostrando os autores-endogeno que tiveram mais publicação
head(sort(table(unlist(unb.pub.df$`autores-endogeno`)),decreasing = TRUE),n=10)
```

```
##
## 1154673226500318 4086384842130773 1141716826787805 7294738832905991
##               31               29               13               12
## 9169095482512290 4839052902231824 9190489069187153 0535100751136568
##               10                9                9                4
## 4739013535126469 0201204222182378
##                4                3
```

Arquivo Orientação

```
#Orientação
#Visualizar a estrutura do json no painel Viewer
#jsonedit(unb.adv)
#Reunir todos os anos e orientações concluidas em um mesmo data-frame
unb.adv.tipo.df <- data.frame(); unb.adv.df <- data.frame()
for (i in 1:length(unb.adv[[1]]))
  unb.adv.tipo.df <- rbind(unb.adv.tipo.df, unb.adv$ORIENTACAO_CONCLUIDA_POS_DOUTORADO[[i]])
unb.adv.df <- rbind(unb.adv.df, unb.adv.tipo.df); unb.adv.tipo.df <- data.frame()
for (i in 1:length(unb.adv[[1]]))
  unb.adv.tipo.df <- rbind(unb.adv.tipo.df, unb.adv$ORIENTACAO_CONCLUIDA_DOUTORADO[[i]])
unb.adv.df <- rbind(unb.adv.df, unb.adv.tipo.df); unb.adv.tipo.df <- data.frame()
for (i in 1:length(unb.adv[[1]]))
```

```
  unb.adv.tipo.df <- rbind(unb.adv.tipo.df, unb.adv$ORIENTACAO_CONCLUIDA_MESTRADO[[i]])
unb.adv.df <- rbind(unb.adv.df, unb.adv.tipo.df)
glimpse(unb.adv.df)
```

```
## Observations: 124
## Variables: 13
## $ natureza                    <chr> "Supervisão de pós-doutorado", "Su...
## $ titulo                      <chr> "", "", "Influência da Eletroestim...
## $ ano                         <chr> "2014", "2017", "2011", "2011", "2...
## $ id_lattes_aluno             <chr> "", "", "7129464687368571", "92327...
## $ nome_aluno                  <chr> "Leandro Xavier Cardoso", "Glécia ...
## $ instituicao                 <chr> "Universidade de Brasília - Faculd...
## $ curso                       <chr> "", "", "Ciências Médicas", "ENGEN...
## $ codigo_do_curso             <chr> "", "", "60021152", "60057840", "6...
## $ bolsa                       <chr> "SIM", "SIM", "SIM", "NAO", "NAO",...
## $ agencia_financiadora        <chr> "Coordenação de Aperfeiçoamento de...
## $ codigo_agencia_financiadora <chr> "045000000000", "045000000000", "0...
## $ nome_orientadores           <list> ["Lourdes Mattos Brasil", "Lourde...
## $ id_lattes_orientadores      <list> ["9190489069187153", "91904890691...
```

```
#Transformar as colunas de listas em caracteres eliminando c("")
unb.adv.df$nome_orientadores <- gsub("\"|c\\(|\\)", "", unb.adv.df$nome_orientadores)
unb.adv.df$id_lattes_orientadores <- gsub("\"|c\\(|\\)", "", unb.adv.df$id_lattes_orientadores)
#Separar as colunas com dois orientadores
unb.adv.df <- separate(unb.adv.df, nome_orientadores, into = c("ori1", "ori2"), sep = ",")
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 123 rows [1,
## 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, ...].
```

```
# Contando quem tem apenas 1 orientador
sum(is.na(unb.adv.df$ori2))
```

```
## [1] 123
```

```
unb.adv.df <- separate(unb.adv.df, id_lattes_orientadores, into = c("idLattes1", "idLattes2"), sep = ",
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 123 rows [1,
## 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, ...].
```

```
# Contando quem tem apenas 1 orientador
sum(is.na(unb.adv.df$idLattes2))
```

```
## [1] 123
```

```
#Numero de orientacoes por ano
table(unb.adv.df$ano)
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##    7    6   12   17   25   25   13   19
```

```
#Numero de instituições
sort(table(unb.adv.df$instituicao),decreasing = TRUE)
```

```
##
##                                          Universidade de Brasília
##                                                                82
##                          Universidade de Brasília - Faculdade UnB-Gama
##                                                                21
```

```
##                                                 Faculdade do Gama da UnB
##                                                                        7
##                                                 Faculdade UNB Gama - FGA
##                                                                        7
##                         Universidade de BrasÃlia - Campus Darcy Ribeiro
##                                                                        3
## Centro de Desenvolvimento e Planejamento Regional - Cedeplar/FACE/UFMG
##                                                                        1
##                                                   Faculdade do Gama/ UnB
##                                                                        1
##                           Universidade do Estado do Rio Grande do Norte
##                                                                        1
##                                       Universidade Nacional de Brasília
##                                                                        1
```

```
#Cursos que estão envolvido na pós-graduação de engenharia biomédica
cursos_d <- sort(table(unb.adv.df$curso), decreasing = TRUE)
cursos_d
```

```
##
##                                Engenharia Biomédica
##                                                  41
##                    Mestrado em Engenharia Biomédica
##                                                  26
## Programa de Pós-Graduação em Engenharia Biomédica
##                                                   7
##                                 Engenharia Elétrica
##                                                   6
##                                             Física
##                                                   6
##                                    Ciências Médicas
##                                                   5
## ENGENHARIA DE SISTEMAS ELETRÔNICOS E DE AUTOMAÇÃO
##                                                   5
##               Pós-Graduação em Engenharia Biomédica
##                                                   5
##                                 ENGENHARIA ELÉTRICA
##                                                   3
##
##                                                   2
##                               ciências de materiais
##                                                   2
##                                          PPDSCI/CEAM
##                                                   2
##                               Sistemas Mecatrônicos
##                                                   2
##               Ciência da Computação - Uern - Ufersa
##                                                   1
##                                  Ciências da Saúde
##                                                   1
##                     Ciências e Tecnologias em Saúde
##                                                   1
##                                  Ciências Mecânicas
##                                                   1
##                                         Demografia
```

```
##                                             1
##                                 Educação Física
##                                             1
##                                     Informática
##                                             1
##                         Medicina (Clínica Médica)
##                                             1
##                   Nanociência e Nanobiotecnologia
##                                             1
##                 Pós-graduação em Ciências Médicas
##                                             1
##                                     Psicologia
##                                             1
##                       Psicologia Clínica e Cultura
##                                             1
```

```
cursos_d5<-head(cursos_d,5)
table(unb.adv.df$codigo_do_curso)
```

```
##
##          51500027 51500132 51500140 51500248 60009322 60018704 60021152
##        2        6        1        6        1        1        2        5
## 60021160 60021179 60027002 60027894 60045850 60057831 60057840 60059672
##        1        1        1        1        1        3        5       33
## 60059753 60471972 90000003 90000013 90000019 90000022 90000023 90000024
##        1        1        5        1        1        1        7        9
## 90000027 90000035 90000036 90000046 90000052
##        1        1        1       19        6
```

```
#Quantidas de natureza dos trabalhos orientado
table(unb.adv.df$natureza)
```

```
##
##     Dissertação de mestrado Supervisão de pós-doutorado
##                         109                           2
##           Tese de doutorado
##                          13
```

```
#Tabela com nome de professor e numero de orientacoes
head(sort(table(rbind(unb.adv.df$ori1, unb.adv.df$ori2)), decreasing = TRUE), 20)
```

```
##
##                     Lourdes Mattos Brasil
##                                       24
##  Suélia de Siqueira Rodrigues Fleury Rosa
##                                       20
##                   Adson Ferreira da Rocha
##                                       13
##                     Jose Felicio da Silva
##                                       11
##               Marilia Miranda Forte Gomes
##                                       11
## Cristiano Jacques Miosso Rodrigues Mendes
##                                       10
##             Marcelino Monteiro de Andrade
##                                        7
```

```
##             Ronni Geraldo Gomes de Amorim
##                                         7
##     Vera Regina Fernandes da Silva Marães
##                                         7
##                 Georges Daniel Amvame Nze
##                                         6
##           Marcella Lemos Brettas Carneiro
##                                         3
##             Sergio Ricardo Menezes Mateus
##                                         3
##  Suélia de Siqueira Rodrigues Fleury Rosa
##                                         1
##                     Fabiano Araujo Soares
##                                         1
##             Marcus Vinícius Chaffim Costa
##                                         1
```

```
# Tabela com nome dos alunos que mais foram orientados
head(sort(table(toupper(unb.adv.df$nome_aluno)),decreasing = TRUE),n=10)
```

```
##
##   AMILTON DOS REIS CAPISTRANO        ANTONIO DOMINGUES NETO
##                             3                             2
##      CAMILA CADENA DE ALMEIDA  CRISTINA AKEMI SHIMODA UECHI
##                             2                             2
## LEINA ADRIANA BARBOSA PIMENTA              LUIZ ALBER LEMOS
##                             2                             2
##       MARIA DO CARMO DOS REIS           ROBERTO AGUIAR LIMA
##                             2                             2
##             ROOZBEH TAHMASEBI         SIMONE BEZERRA FRANCO
##                             2                             2
```

```
# Quantidade de alunos que não são bolsistas
sum(unb.adv.df$bolsa == 'NAO')
```

```
## [1] 78
```

```
# Quantidade de alunos que são bolsistas
sum(unb.adv.df$bolsa == 'SIM')
```

```
## [1] 46
```

```
# Reparar que a maioria não tem agência financiadora por trás e há dois alunos que tem bolsa mas que nã
table(unb.adv.df$agencia_financiadora)
```

```
##
##
##                                                          80
##               Centro de Apoio ao Desenvolvimento Tecnologico
##                                                           1
## Conselho Nacional de Desenvolvimento Científico e Tecnológico
##                                                          10
##   Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
##                                                          32
##             Fundação de Apoio à Pesquisa do Distrito Federal
##                                                           1
```

CRISP-DM Fase.Atividade 2.4 - Verificação da qualidade dos dados.

Para o levantamento de informação os dados em geral são preciso, como por exemplo saber quais os professores mais envolvido em orientações, quantos artigos foram concluidos, informações por ano. Mas houve dificuldade quando o dado que se queria analisar dependia de como a pessoa que cadastrou escreveu, como por exemplo os autores nos artigos, haja visto que não foi encontrado um padrão, dificultou qualquer análise.

CRISP-DM Fase 3 - Preparação dos Dados

CRISP-DM Fase.Atividade 3.1 - Seleção dos dados.

Para a principal utilização dos dados foi definidos os **data frames** unb.adv.df, que neste caso contêm alguns informações de qual a natureza da pesquisa produzida, qual aluno produziu, seus respectivos orientadores, o ano e etc. Também foi definido os dados importantes do arquivo profile.json, gerando o **data frame** unb.prof.df, que já está limpo, com as principais colunas definidas.

CRISP-DM Fase.Atividade 3.2 - Limpeza dos dados

Esta etapa se mistura com a próxima pelo fato de limpar e preparar o dado, ou seja, já construir.

CRISP-DM Fase.Atividade 3.3 - Construção dos dados

Construindo e limpando alguns dados de interesse.

```
# Funcoes auxiliares que serão usados
# estas funções foram disponibilizada pelo professor da disciplina
# converte as colunas de um dataframe tipo lista em tipo character
cv_tplista2tpchar <- function( df  ) {
  for( variavel in names(df)) {
    if (class(df[[variavel]]) == "list" ) {
      df[[variavel]] <- lapply(df[[variavel]] ,   function(x)   lista2texto( x  ) )
      df[[variavel]] <- as.character( df[[variavel]] )
    }
  }
  return(df)
}
###


# converte o conteudo de lista em array de characters
lista2texto <- function( lista  ) {
  if(is.null(lista)) {
    return ( NULL )
  }
  saida <- ""
  for( j in 1:length(lista)) {
    for( i in 1:length(lista[[j]]) ) {
      elemento <- lista[[j]][i]
      if( !is.null(elemento)) {
        if( i == length(lista[[j]]) & j == length(lista)  ) {
          # se for o ultimo elemento nao coloque o ponto e virgula no final
          saida <- paste0( saida , elemento  )
        } else {
          # enquanto nao for o ultimo coloque ; separando os elementos concatenados
          saida <- paste0( saida , elemento , sep = " ; ")
```

```
        }
      }
    }
  }
  return( saida )
}

# Converte producao elattes separada por anos em um unico dataframe
converte_producao2dataframe<- function( lista_producao ) {
  df_saida <- NULL

  for( ano in names(lista_producao)) {
    df_saida <- rbind(df_saida , lista_producao[[ano]])
  }

  # converte tipo lista em array de character
  df_saida <- cv_tplista2tpchar(df_saida)
  return(df_saida)


}

#concatena dois dataframes com  colunas diferentes
concatenadf <- function( df1, df2) {
  #cria colunas de df1 que faltam em df2
  for( coluna in names(df1 ) ) {
    if( !is.element(coluna, names(df2) )) {
      df2[coluna] <- NA
    }
  }

  #cria colunas de df2 que faltam em df1
  for( coluna in names(df2 ) ) {

    if( !is.element(coluna, names(df1) )) {
      df1[coluna] <- NA
    }
  }


  #faz o rbind dos dois dataframes
  df_final <- rbind(df1 , df2)
  return(df_final)

}

# Extracao dos perfis dos professores

extrai_1perfil <- function( professor ) {
  idLattes <- names(professor)
  nome <- professor[[idLattes]]$nome
  resumo_cv <- professor[[idLattes]]$resumo_cv
  endereco_profissional <- professor[[idLattes]]$endereco_profissional #list
```

```
  instituicao <- endereco_profissional$instituicao
  orgao <- endereco_profissional$orgao
  unidade <- endereco_profissional$unidade
  DDD <- endereco_profissional$DDD
  telefone <- endereco_profissional$telefone
  bairro <- endereco_profissional$bairro
  cep <- endereco_profissional$cep
  cidade <- endereco_profissional$cidade
  senioridade <- professor[[idLattes]]$senioridade
  df_1perfil <- data.frame( idLattes , nome, resumo_cv ,instituicao ,
                            orgao, unidade , DDD, telefone, bairro,cep,cidade , senioridade,
                            stringsAsFactors = FALSE)

  return(df_1perfil)
}

extrai_perfis <- function(jsonProfessores) {
  df_saida <- data.frame()
  for( i in 1:length(jsonProfessores)) {
    jsonProfessor <- jsonProfessores[i]
    df_professor <- extrai_1perfil(jsonProfessor)
    if( nrow(df_saida) > 0 ) {
      df_saida <- rbind(df_saida , df_professor)
    } else {
      df_saida <- df_professor
    }
  }

  return(df_saida)
}

# Extracao da producao bibliografica dos professores

extrai_1producao <- function(professor) {
  idLattes <- names(professor)
  df_1producao <<- NULL
  producao_bibliografica <- professor[[idLattes]]$producao_bibiografica  #list
  for( tipo_producao in names(producao_bibliografica)) {
    df_temporario <- cv_tplista2tpchar ( producao_bibliografica[[tipo_producao]])
    df_temporario$tipo_producao <-  tipo_producao
    df_temporario$idLattes <-  idLattes
    df_1producao <- concatenadf( df_1producao , df_temporario  )
  }
  return(df_1producao)
}

extrai_producoes <- function( jsonProfessores) {
  df_saida <- data.frame()
  for( i in 1:length(jsonProfessores)) {
    jsonProfessor <- jsonProfessores[i]
    df_producao <- extrai_1producao(jsonProfessor)
    if( nrow(df_saida) > 0 ) {
      df_saida <- concatenadf(df_saida , df_producao)
```

```
    } else {
      df_saida <- df_producao
    }
  }
  df_saida <- df_saida %>% filter( !is.na(tipo_producao))
  return(df_saida)
}

# Extracao das orientacoes dos professores

extrai_1orientacao <- function(professor) {
  idLattes <- names(professor)
  df_1orientacao <- NULL
  orientacoes_academicas  <- professor[[idLattes]]$orientacoes_academicas  #list
  for( orientacao in names(orientacoes_academicas )) {
    df_temporario <- cv_tplista2tpchar ( orientacoes_academicas[[orientacao]])
    df_temporario$orientacao <-  orientacao
    df_temporario$idLattes <-  idLattes
    df_1orientacao <- concatenadf( df_1orientacao , df_temporario  )
  }
  return(df_1orientacao)
}

extrai_orientacoes <- function(jsonProfessores) {
  df_saida <- data.frame()
  for( i in 1:length(jsonProfessores)) {
    jsonProfessor <- jsonProfessores[i]
    df_orientacao <- extrai_1orientacao(jsonProfessor)
    if( nrow(df_saida) > 0 ) {
      df_saida <- concatenadf(df_saida , df_orientacao)
    } else {
      df_saida <- df_orientacao
    }
  }
  df_saida <- df_saida %>% filter(!is.na(idLattes))
  return(df_saida)
}

# Extracao das areas de atuacao dos professores

extrai_1area_de_atuacao <- function(professor){
  idLattes <- names(professor)
  df_1area <-  professor[[idLattes]]$areas_de_atuacao
  df_1area$idLattes <- idLattes
  return(df_1area)
}

extrai_areas_atuacao <- function(jsonProfessores){
  df_saida <- data.frame()
  for( i in 1:length(jsonProfessores)) {
    jsonProfessor <- jsonProfessores[i]
    df_area_atuacao <- extrai_1area_de_atuacao(jsonProfessor)
    if( nrow(df_saida) > 0 ) {
```

```
      df_saida <- concatenadf(df_saida , df_area_atuacao)
    } else {
      df_saida <- df_area_atuacao
    }
  }
  df_saida <- df_saida %>% filter( !is.na(idLattes))
  return(df_saida)
}
############################ Inicio ################################
#### Começo da preparação dos dados referente a profile

unb.prof.json <- read_file("dados-2018-2/engenharia-biomedica/279.profile.json")
unb.prof.df.capes <- read.csv("dados-2018-2/PesqPosCapes.csv",
                              sep = ";", header = TRUE, colClasses = "character")
unb.prof <- fromJSON(unb.prof.json)
length(unb.prof)
```

```
## [1] 15
```

```
# extrai perfis dos professores
unb.prof.df.professores <- extrai_perfis(unb.prof)

# extrai producao bibliografica de todos os professores
unb.prof.df.publicacoes <- extrai_producoes(unb.prof)

#extrai orientacoes
unb.prof.df.orientacoes <- extrai_orientacoes(unb.prof)

#extrai areas de atuacao
unb.prof.df.areas.de.atuacao <- extrai_areas_atuacao(unb.prof)

#salva os daframes
save(unb.prof.df.professores, unb.prof.df.publicacoes,
     unb.prof.df.orientacoes, unb.prof.df.areas.de.atuacao, file = "dataframes.Rda")

#cria arquivo para análise
unb.prof.df <- data.frame()
unb.prof.df <- unb.prof.df.professores %>%
  select(idLattes, nome, resumo_cv, senioridade) %>%
  left_join(
    unb.prof.df.orientacoes %>%
      select(orientacao, idLattes) %>%
      filter(!grepl("EM_ANDAMENTO", orientacao)) %>%
      group_by(idLattes) %>%
      count(orientacao) %>%
      spread(key = orientacao, value = n),
    by = "idLattes") %>%
  left_join(
    unb.prof.df.publicacoes %>%
      select(tipo_producao, idLattes) %>%
      filter(!grepl("ARTIGO_ACEITO", tipo_producao)) %>%
      group_by(idLattes) %>%
      count(tipo_producao) %>%
      spread(key = tipo_producao, value = n),
```

```
    by = "idLattes") %>%
  left_join(
    unb.prof.df.areas.de.atuacao %>%
      select(area, idLattes) %>%
      group_by(idLattes) %>%
      summarise(n_distinct(area)),
    by = "idLattes") %>%
  left_join(
    unb.prof.df.capes %>%
      select(AreaPos, idLattes) %>%
      group_by(idLattes) %>%
      summarise(n_distinct(AreaPos)),
    by = "idLattes")

glimpse(unb.prof.df)
```

```
## Observations: 15
## Variables: 16
## $ idLattes                              <chr> "0201204222182378", "05...
## $ nome                                  <chr> "Leandro Xavier Cardoso...
## $ resumo_cv                             <chr> "Possui graduação em Fí...
## $ senioridade                           <chr> "9", "6", "9", "9", "5"...
## $ ORIENTACAO_CONCLUIDA_DOUTORADO        <int> NA, 1, 4, 3, 1, NA, NA,...
## $ ORIENTACAO_CONCLUIDA_MESTRADO         <int> NA, 9, 9, 18, NA, 11, 7...
## $ ORIENTACAO_CONCLUIDA_POS_DOUTORADO    <int> NA, NA, NA, NA, NA, NA,...
## $ OUTRAS_ORIENTACOES_CONCLUIDAS         <int> 5, 2, 6, 148, 48, 20, 1...
## $ CAPITULO_DE_LIVRO                     <int> NA, 1, 9, 9, NA, 1, NA,...
## $ DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA <int> NA, NA, NA, NA, NA, NA,...
## $ EVENTO                                <int> 3, 23, 51, 65, 10, 15, ...
## $ LIVRO                                 <int> NA, NA, NA, 1, NA, NA, ...
## $ PERIODICO                             <int> 10, 8, 22, 36, 1, 2, 29...
## $ TEXTO_EM_JORNAIS                      <int> NA, NA, NA, 1, NA, NA, ...
## $ `n_distinct(area)`                    <int> 1, 3, 2, 3, 2, 3, 2, 1,...
## $ `n_distinct(AreaPos)`                 <int> 1, 1, 3, 1, 1, 1, 2, 1,...
```

```
head(unb.prof.df,3)
```

```
##           idLattes                                      nome
## 1 0201204222182378                    Leandro Xavier Cardoso
## 2 0535100751136568 Cristiano Jacques Miosso Rodrigues Mendes
## 3 1141716826787805                   Adson Ferreira da Rocha
##
## 1
## 2
## 3 Engenheiro Eletricista pela Universidade de Brasília (1988), Mestre em Engenharia Elétrica pela Uni
##   senioridade ORIENTACAO_CONCLUIDA_DOUTORADO ORIENTACAO_CONCLUIDA_MESTRADO
## 1           9                             NA                            NA
## 2           6                              1                             9
## 3           9                              4                             9
##   ORIENTACAO_CONCLUIDA_POS_DOUTORADO OUTRAS_ORIENTACOES_CONCLUIDAS
## 1                                 NA                             5
## 2                                 NA                             2
## 3                                 NA                             6
##   CAPITULO_DE_LIVRO DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA EVENTO LIVRO
```

```
## 1                NA                                      NA      3    NA
## 2                 1                                      NA     23    NA
## 3                 9                                      NA     51    NA
##   PERIODICO TEXTO_EM_JORNAIS n_distinct(area) n_distinct(AreaPos)
## 1        10               NA                1                   1
## 2         8               NA                3                   1
## 3        22               NA                2                   3
```

```
## Mostrando dados processados referente ao arquivo 279.publication.json
glimpse(unb.pub.df)
```

```
## Observations: 147
## Variables: 10
## $ natureza           <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo             <chr> "Prótese para controle de fluxo esofagiano ...
## $ periodico          <chr> "Revista Brasileira de Engenharia Biomédica...
## $ ano                <chr> "2010", "2010", "2010", "2010", "2010", "20...
## $ volume             <chr> "26", "249", "20", "11", "14", "9", "53", "...
## $ issn               <chr> "15173151", "17426596", "10546618", "151942...
## $ paginas            <chr> "49 - 54", "012031 - ", "192 - 200", "23 - ...
## $ doi                <chr> "", "10.1088/1742-6596/249/1/012031", "10.1...
## $ autores            <chr> "Rosa, S. S. R. F.; da Rocha, A F; Brasil, ...
## $ `autores-endogeno` <chr> "1141716826787805", "0201204222182378", "91...
```

```
#Podemos ver que contêm natureza, titulo, periodico , autores ....
head(unb.pub.df,3)
```

```
##   natureza
## 1 COMPLETO
## 2 COMPLETO
## 3 COMPLETO
##                                                                                                                                ti
## 1 Prótese para controle de fluxo esofagiano como nova técnica para o tratamento da obesidade (QUALIS
## 2                    Thermoluminescent dose reconstruction using quartz extracted from unfired buildi
## 3                Breast cancer image assessment using an adaptative network-based fuzzy inference sys
##                                                periodico  ano volume
## 1 Revista Brasileira de Engenharia Biomédica (Impresso) 2010     26
## 2        Journal of Physics. Conference Series (Online) 2010    249
## 3                Pattern Recognition and Image Analysis 2010     20
##       issn   paginas                              doi
## 1 15173151   49 - 54
## 2 17426596 012031 -  10.1088/1742-6596/249/1/012031
## 3 10546618 192 - 200      10.1134/S1054661810020112
##
## 1
## 2 Campos, Simara S; Almeida, Geângela M; CARDOSO, L. X.;Cardoso, Leandro X;CARDOSO, L X;XAVIER CARDOS
## 3                                          Fernandes, F. C.; BRASIL, L. M.;Brasil, L. M.;Brasil, L
##   autores-endogeno
## 1 1141716826787805
## 2 0201204222182378
## 3 9190489069187153
```

```
## Mostrando dados processados referente ao arquivo 279.advise.json
glimpse(unb.adv.df)
```

```
## Observations: 124
```

```
## Variables: 15
## $ natureza                   <chr> "Supervisão de pós-doutorado", "Su...
## $ titulo                     <chr> "", "", "Influência da Eletroestim...
## $ ano                        <chr> "2014", "2017", "2011", "2011", "2...
## $ id_lattes_aluno            <chr> "", "", "7129464687368571", "92327...
## $ nome_aluno                 <chr> "Leandro Xavier Cardoso", "Glécia ...
## $ instituicao                <chr> "Universidade de Brasília - Faculd...
## $ curso                      <chr> "", "", "Ciências Médicas", "ENGEN...
## $ codigo_do_curso            <chr> "", "", "60021152", "60057840", "6...
## $ bolsa                      <chr> "SIM", "SIM", "SIM", "NAO", "NAO",...
## $ agencia_financiadora       <chr> "Coordenação de Aperfeiçoamento de...
## $ codigo_agencia_financiadora <chr> "045000000000", "045000000000", "0...
## $ ori1                       <chr> "Lourdes Mattos Brasil", "Lourdes ...
## $ ori2                       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ idLattes1                  <chr> "9190489069187153", "9190489069187...
## $ idLattes2                  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
```

```
#Podemos ver que contêm natureza, titulo, autores, ano, nome aluno, orientadores ....
head(unb.adv.df,3)
```

```
##                     natureza
## 1 Supervisão de pós-doutorado
## 2 Supervisão de pós-doutorado
## 3           Tese de doutorado
##                                                                                titulo
## 1
## 2
## 3 Influência da Eletroestimulação Neuromuscular de Baixa Freqüência nas Variáveis Eletromiográficas
##    ano  id_lattes_aluno                  nome_aluno
## 1 2014                        Leandro Xavier Cardoso
## 2 2017                 Glécia Virgolino da Silva Luz
## 3 2011 7129464687368571           Kenia Fonseca Pires
##                                   instituicao            curso
## 1 Universidade de Brasília - Faculdade UnB-Gama
## 2 Universidade de Brasília - Faculdade UnB-Gama
## 3                     Universidade de Brasília Ciências Médicas
##   codigo_do_curso bolsa
## 1                  SIM
## 2                  SIM
## 3        60021152  SIM
##                                        agencia_financiadora
## 1   Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
## 2   Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
## 3 Conselho Nacional de Desenvolvimento Científico e Tecnológico
##   codigo_agencia_financiadora                    ori1 ori2
## 1               045000000000   Lourdes Mattos Brasil <NA>
## 2               045000000000   Lourdes Mattos Brasil <NA>
## 3               002200000000 Adson Ferreira da Rocha <NA>
##          idLattes1 idLattes2
## 1 9190489069187153      <NA>
## 2 9190489069187153      <NA>
## 3 1141716826787805      <NA>
```

```
###################### Os arquivos
###### 279.graph.json - usado para mostrar correlações
```

```
#### 279.list.json - Como explicado este arquivo não contem nenhuma informação interessante.
```

CRISP-DM Fase.Atividade 3.4 - Integração dos dados

Neste presente trabalho, não se viu a necessidade de fazer merge entres os datas frames.

CRISP-DM Fase.Atividade 3.5 - Formatação dos dados

As formatações de dados necessária já foram feitas, como por exemplo "orientadores" virou "orie1" e "orie2", fazendo com que cada variável contenha apenas um elemento, e não mais uma lista de orientadores.

CRISP-DM Fase 4 - Modelagem

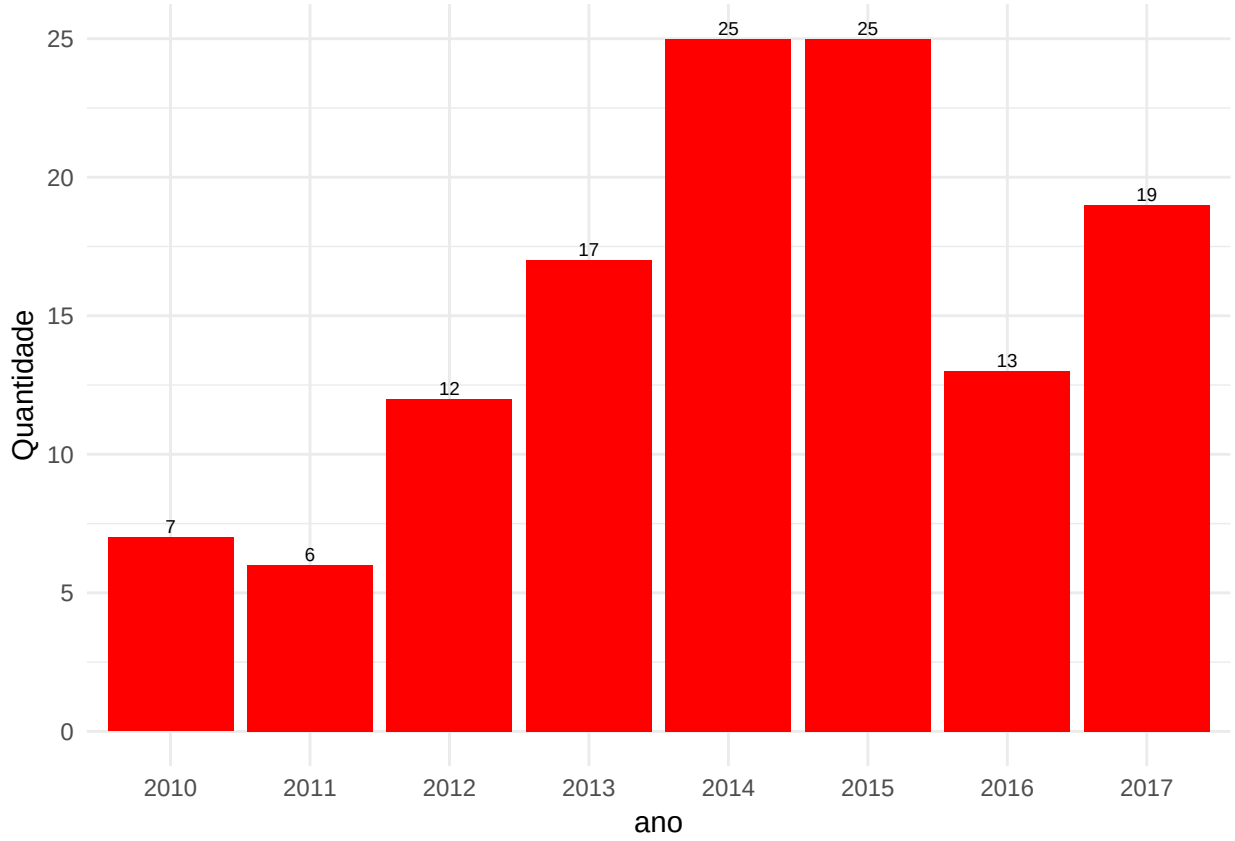
Nesta fase, após ter feito limpeza e preparação dos dados, tem-se que se decidir qual ferramenta usar : computacionais, matemáticas ou estátistica. Esta fase, por enquanto, será omitida.

CRISP-DM Fase 5 - Avaliação

Será exposto em forma de gráficos certos dados que se acharam convenientes e interessante.

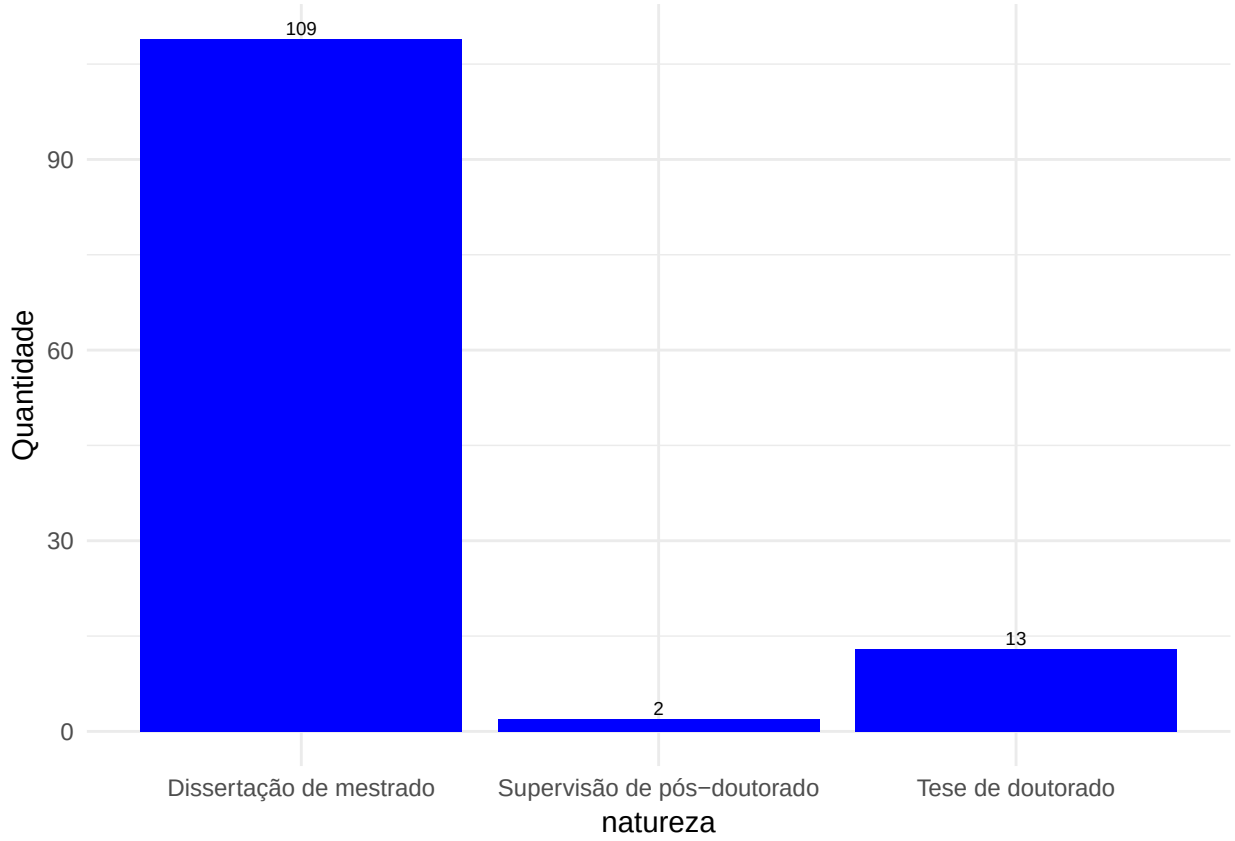
Aqui visa-se ver as quantidades de publicações feitas por anos na área de engenharia de biomédica.

```
unb.adv.df %>%
group_by(ano) %>%
summarise(Quantidade = n()) %>%
ggplot(aes(x = ano, y = Quantidade)) +
geom_bar(position = "stack",stat = "identity", fill = "red")+
geom_text(aes(label=Quantidade), vjust=-0.3, size=2.5)+
theme_minimal()
```



Agora verificando a quantidade separada por tipo de publicação.

```
unb.adv.df %>%
group_by(natureza) %>%
summarise(Quantidade = n()) %>%
ggplot(aes(x = natureza, y = Quantidade)) +
geom_bar(position = "stack",stat = "identity", fill = "blue")+
geom_text(aes(label=Quantidade), vjust=-0.3, size=2.5)+
theme_minimal()
```

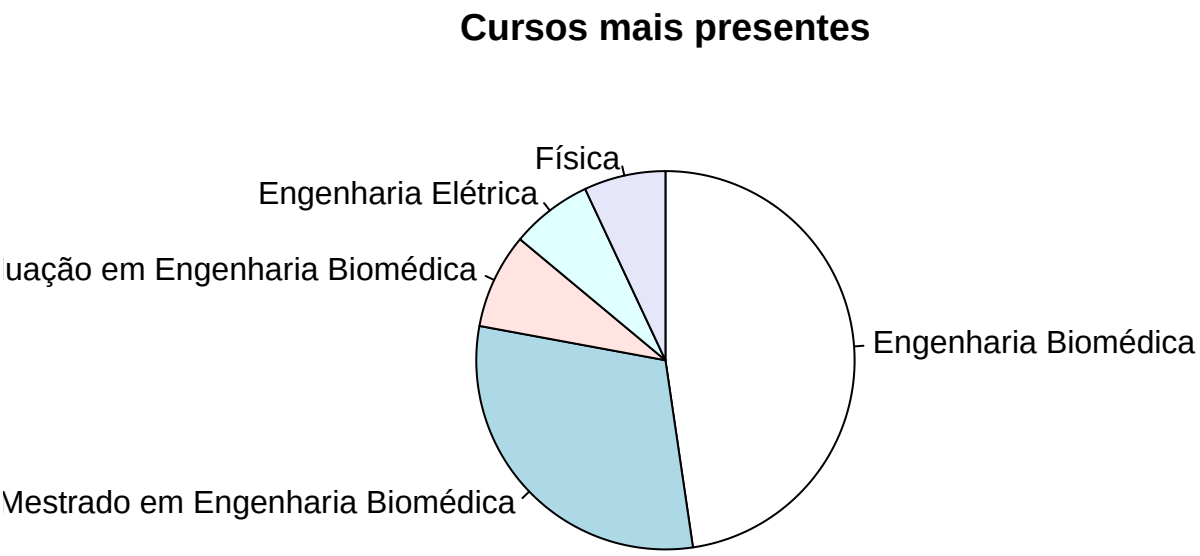


Verificando os cursos mais presentes.

```
cursos_d5
```

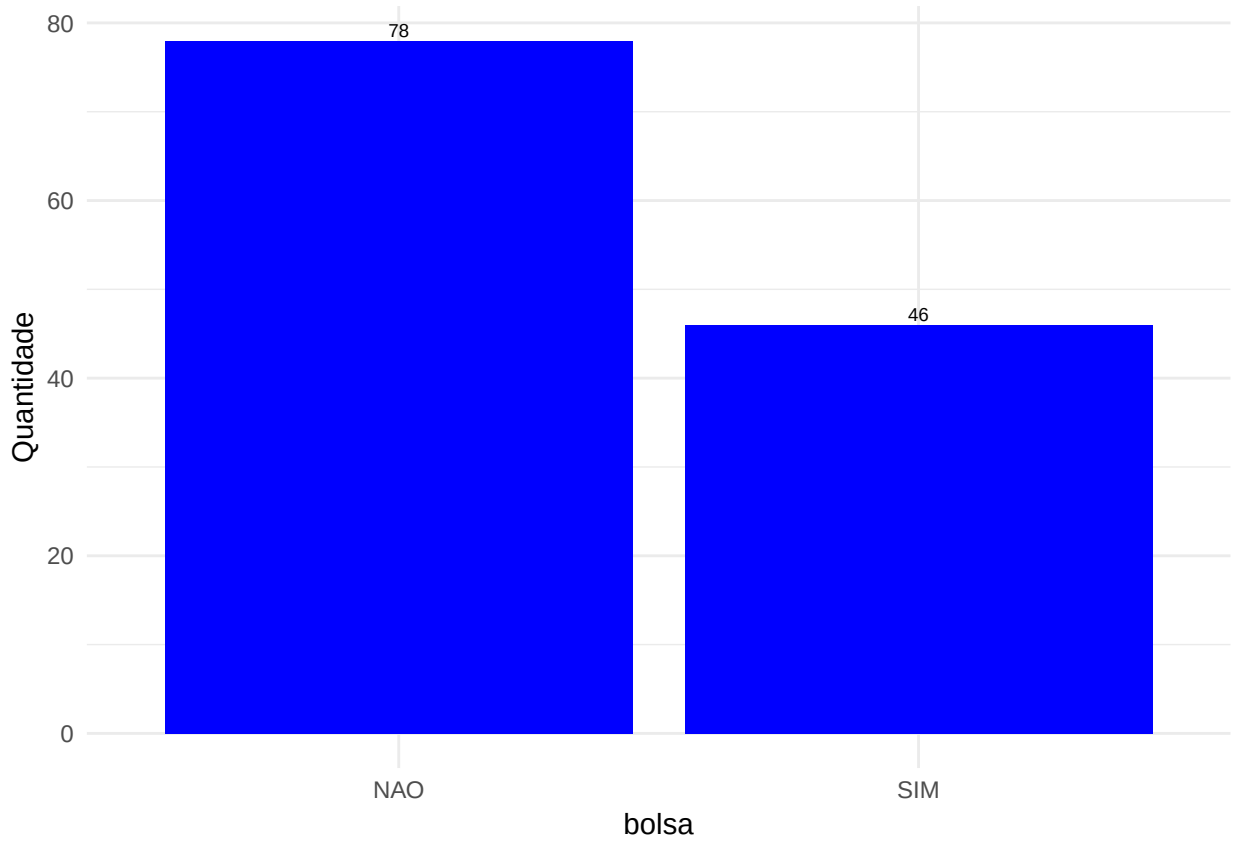
```
##
##                              Engenharia Biomédica
##                                                41
##                  Mestrado em Engenharia Biomédica
##                                                26
## Programa de Pós-Graduação em Engenharia Biomédica
##                                                 7
##                               Engenharia Elétrica
##                                                 6
##                                             Física
##                                                 6
```

```
pie(cursos_d5,clockwise=TRUE,main="Cursos mais presentes")
```



Quantos bolsistas tem ?

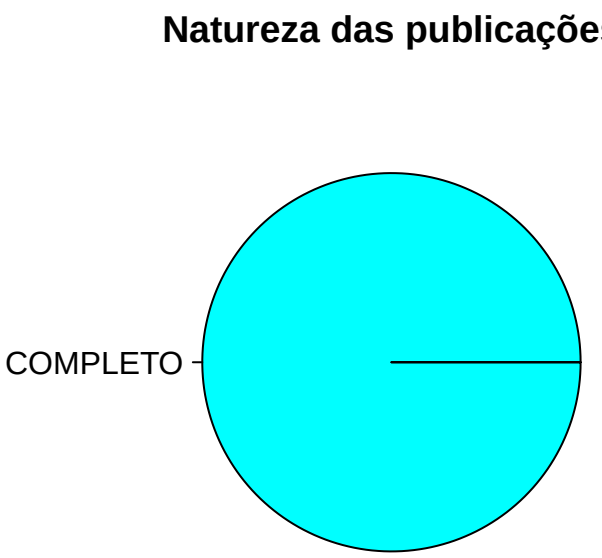
```
unb.adv.df %>%
group_by(bolsa) %>%
summarise(Quantidade = n()) %>%
ggplot(aes(x = bolsa, y = Quantidade)) +
geom_bar(position = "stack",stat = "identity", fill = "blue")+
geom_text(aes(label=Quantidade), vjust=-0.3, size=2.5)+
theme_minimal()
```



Assim, estes gráficos permitem verificar alguns aspectos importantes sobre as orientações.

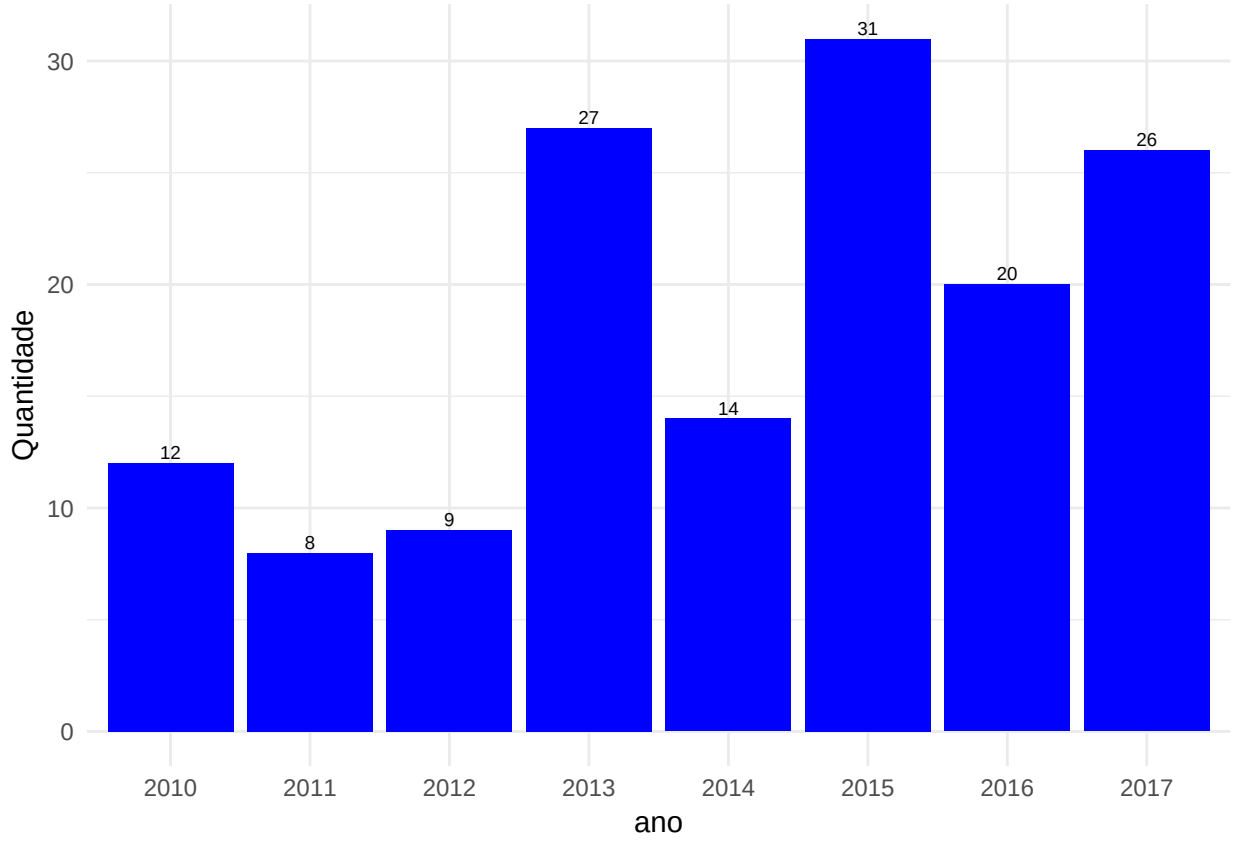
Verificando, agora, o arquivo 279.publication.json, contendo informações sobre as publicações.

```
pie(table(unb.pub.df$natureza),main="Natureza das publicações",col = "cyan")
```



Separando por ano e depois por periódico.

```
unb.pub.df %>%
group_by(ano) %>%
summarise(Quantidade = n()) %>%
ggplot(aes(x = ano, y = Quantidade)) +
geom_bar(position = "stack",stat = "identity", fill = "blue")+
geom_text(aes(label=Quantidade), vjust=-0.3, size=2.5)+
theme_minimal()
```




```
d <- table(unb.pub.df$periodico)
head(sort(d,decreasing = TRUE),n=5)
```

```
##
##         Revista Brasileira de Engenharia Biomédica (Impresso)
##                                                             7
## Global Journal of Engineering Science and Research Management
##                                                             6
##                         Biomedical Engineering Online (Online)
##                                                             4
##                Revista Brasileira de Ensino de Física (Online)
##                                                             4
##           Revista Brasileira de Inovação Tecnologica em Saúde
##                                                             4
```

Mostrando os autores que mais participaram de publicações.

```
head(sort(table(toupper(unlist(strsplit(unb.pub.df$autores,";")))),decreasing = TRUE),n=10)
```

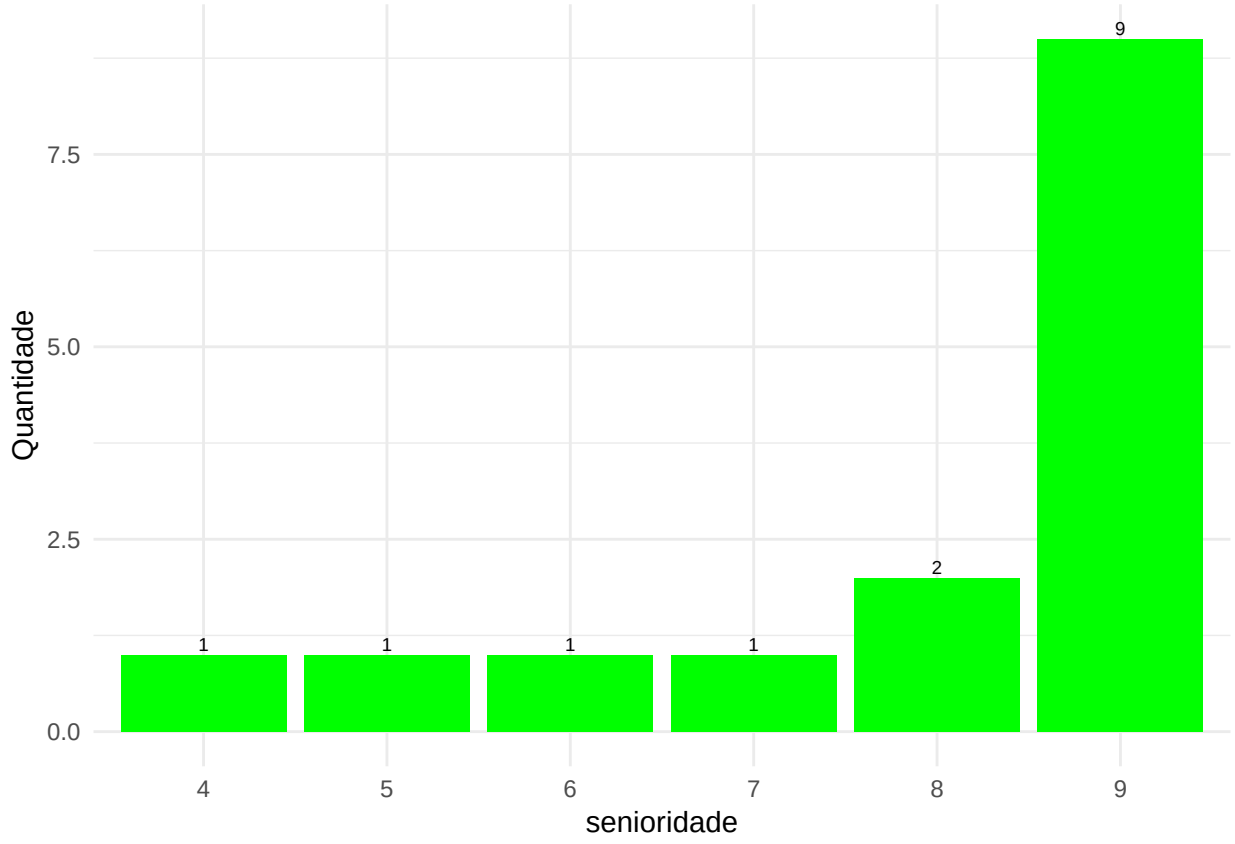
```
##
##                           AMORIM, R. G. G.
##                                         17
##                           AMORIM, R. G. G.
##                                         16
##                         MARÃES, V. R. F. S.
##                                         14
##                              BRASIL, L. M.
##                                          7
```

```
##                       MARÃES, V. R. F. S.
##                                         7
##                                SILVA, W. B.
##                                         7
##                GOMES, MARÍLIA MIRANDA FORTE
##                                         7
## ROSA, SUÉLIA DE SIQUEIRA RODRIGUES FLEURY
##                                         7
##                              DA ROCHA, A. F.
##                                         6
##                GOMES, MARÍLIA MIRANDA FORTE
##                                         6
```

Assim, vemos que há algumas incoerência neste dado acima, isto se dá pela falta de uniformidade.

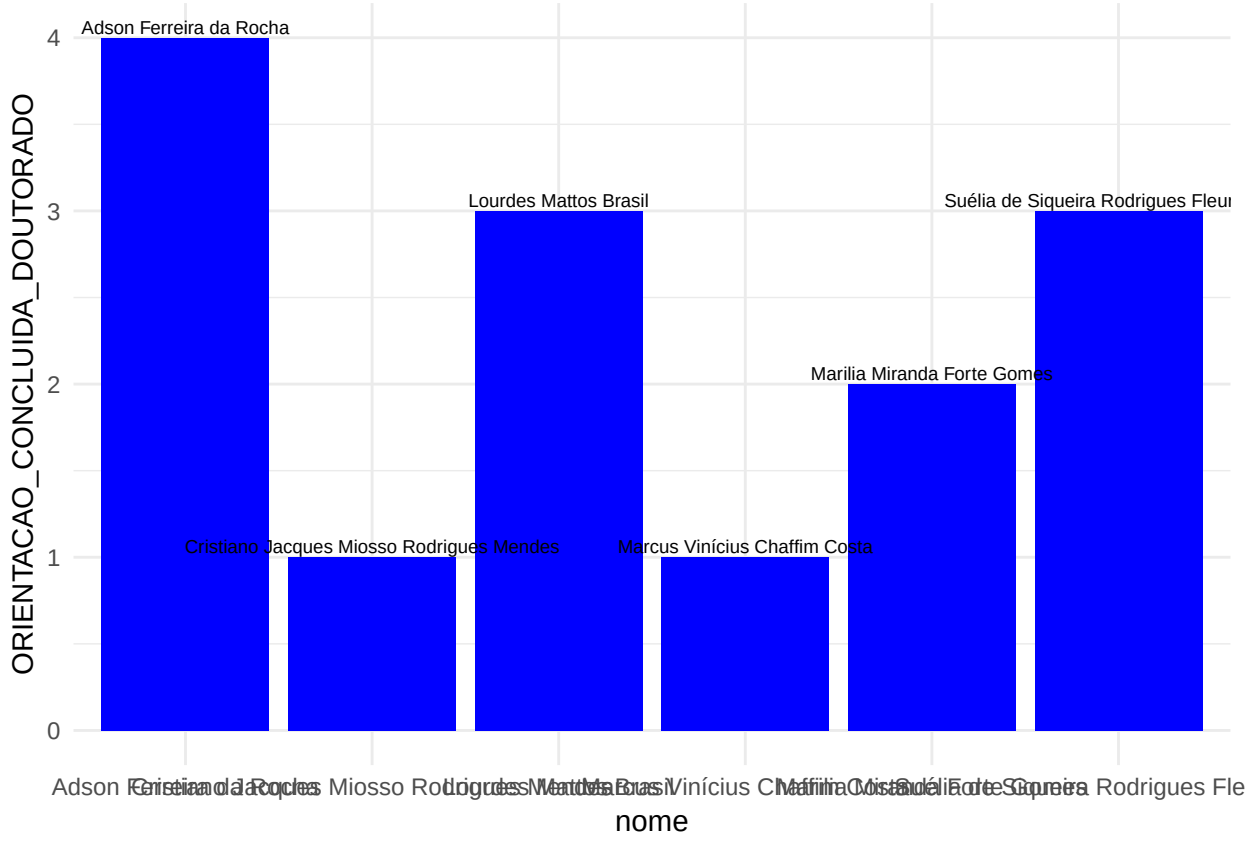
Analisando o arquivo 279.profile.json.

```
unb.prof.df %>%
group_by(senioridade) %>%
summarise(Quantidade = n()) %>%
ggplot(aes(x = senioridade, y = Quantidade)) +
geom_bar(position = "stack",stat = "identity", fill = "green")+
geom_text(aes(label=Quantidade), vjust=-0.3, size=2.5)+
theme_minimal()
```



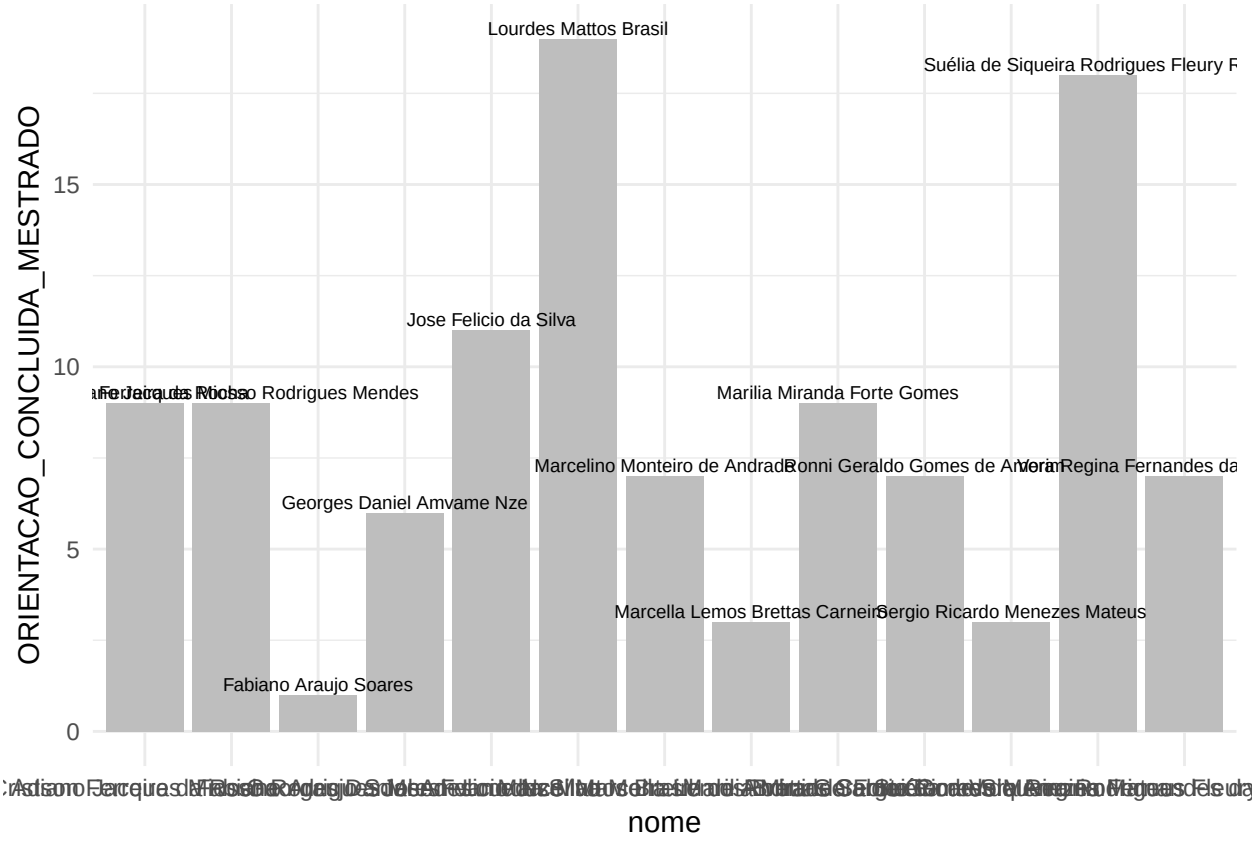
Analisando quandidade de orientações concluídas no doutorado por professor.

```r
unb.prof.df %>%
group_by(nome) %>%
filter(!is.na(ORIENTACAO_CONCLUIDA_DOUTORADO ))%>%
ggplot(aes(x = nome, y = ORIENTACAO_CONCLUIDA_DOUTORADO)) +
geom_bar(position = "stack",stat = "identity", fill = "blue")+
geom_text(aes(label=nome), vjust=-0.3, size=2.5)+
theme_minimal()
```



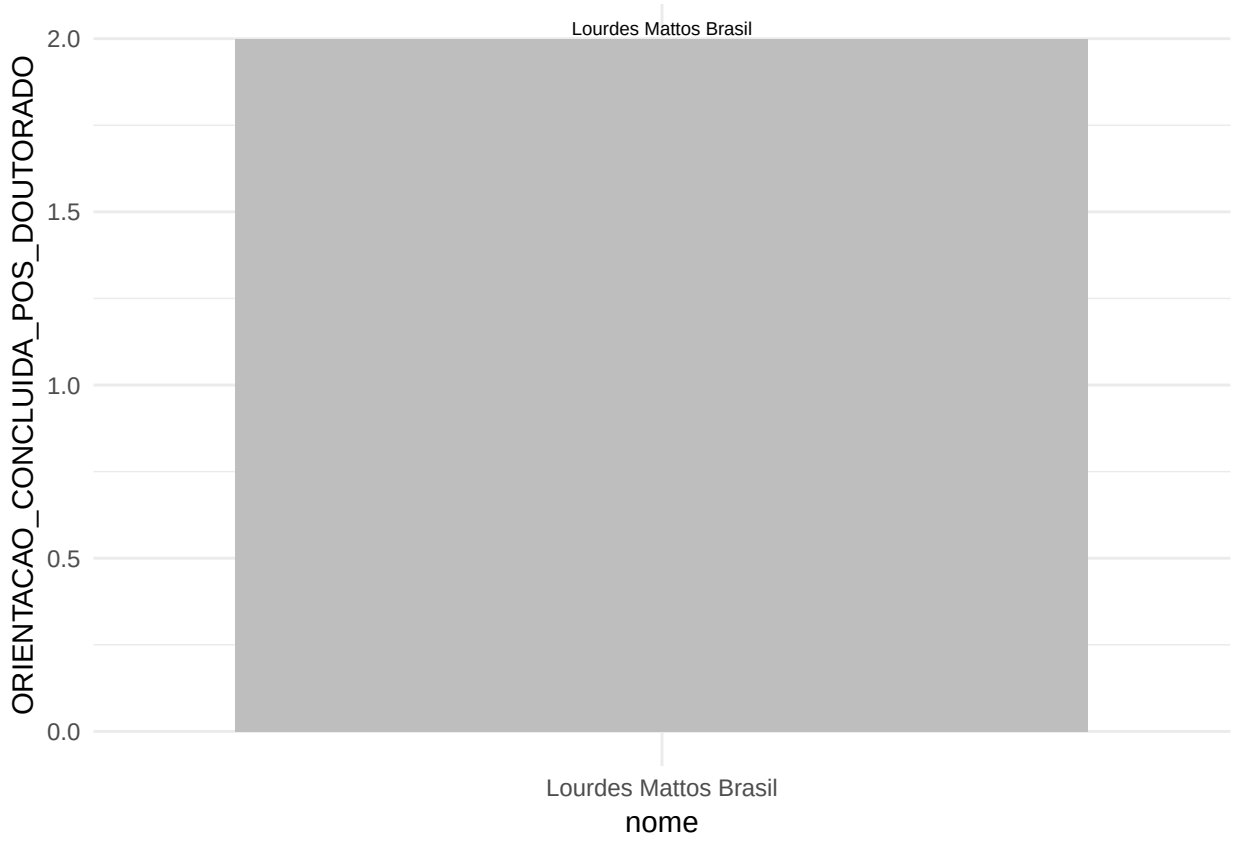
Analisando quandidade de orientações concluídas no mestrado por professor.

```r
unb.prof.df %>%
group_by(nome) %>%
filter(!is.na(ORIENTACAO_CONCLUIDA_MESTRADO ))%>%
ggplot(aes(x = nome, y = ORIENTACAO_CONCLUIDA_MESTRADO)) +
geom_bar(position = "stack",stat = "identity", fill = "gray")+
geom_text(aes(label=nome), vjust=-0.3, size=2.5)+
theme_minimal()
```



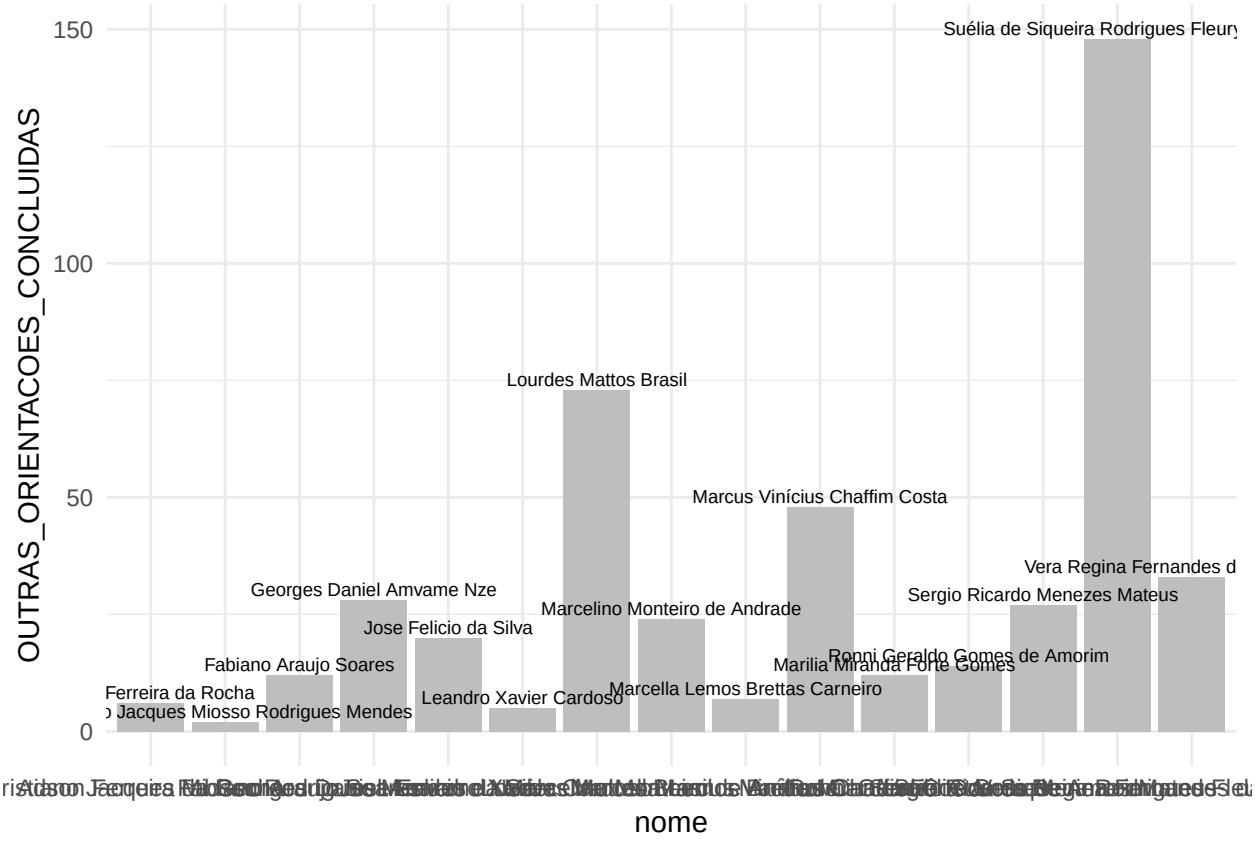
Analisando quandidade de orientações concluídas no pós doutorado por professor.

```
unb.prof.df %>%
group_by(nome) %>%
filter(!is.na(ORIENTACAO_CONCLUIDA_POS_DOUTORADO))%>%
ggplot(aes(x = nome, y = ORIENTACAO_CONCLUIDA_POS_DOUTORADO)) +
geom_bar(position = "stack",stat = "identity", fill = "gray")+
geom_text(aes(label=nome), vjust=-0.3, size=2.5)+
theme_minimal()
```



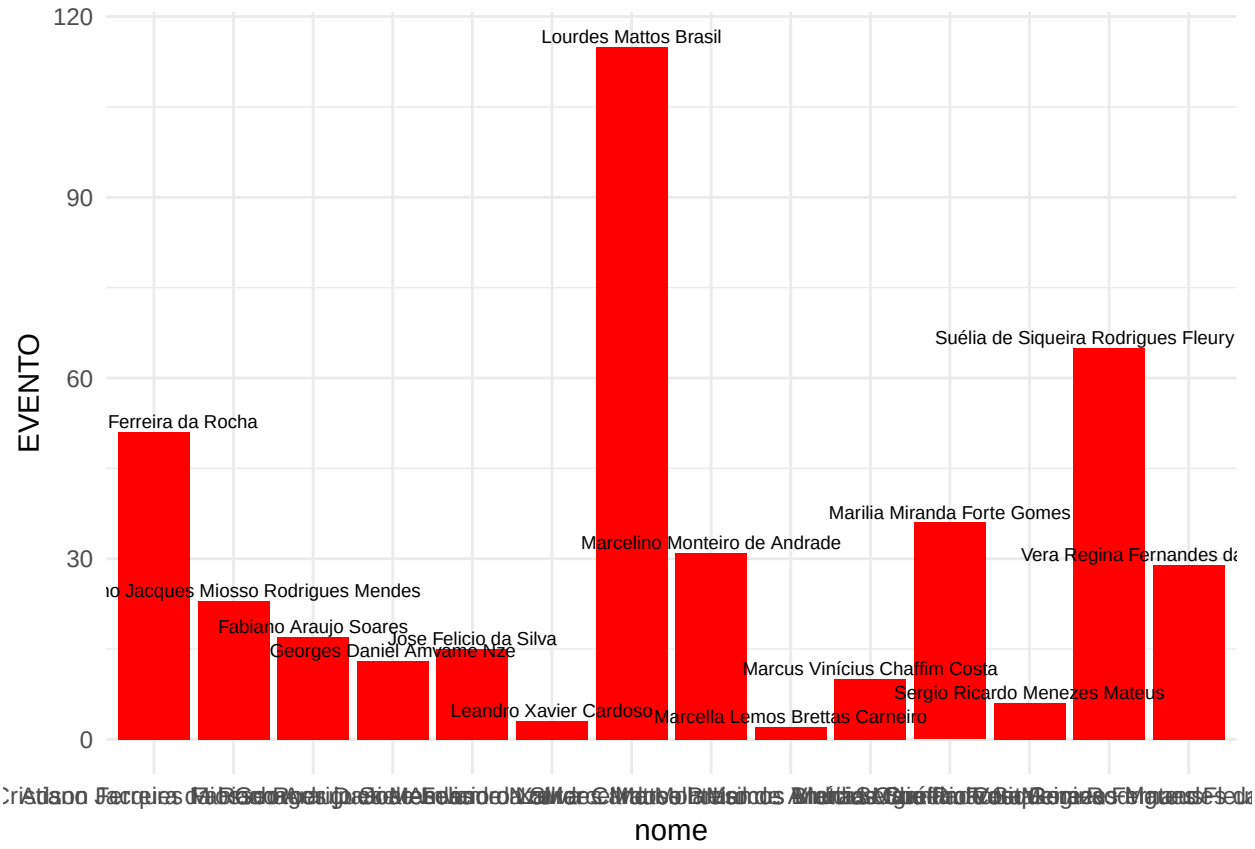
Analisando quandidade de orientações concluídas em outros tipos por professor.

```
unb.prof.df %>%
group_by(nome) %>%
filter(!is.na(OUTRAS_ORIENTACOES_CONCLUIDAS))%>%
ggplot(aes(x = nome, y = OUTRAS_ORIENTACOES_CONCLUIDAS)) +
geom_bar(position = "stack",stat = "identity", fill = "gray")+
geom_text(aes(label=nome), vjust=-0.3, size=2.5)+
theme_minimal()
```



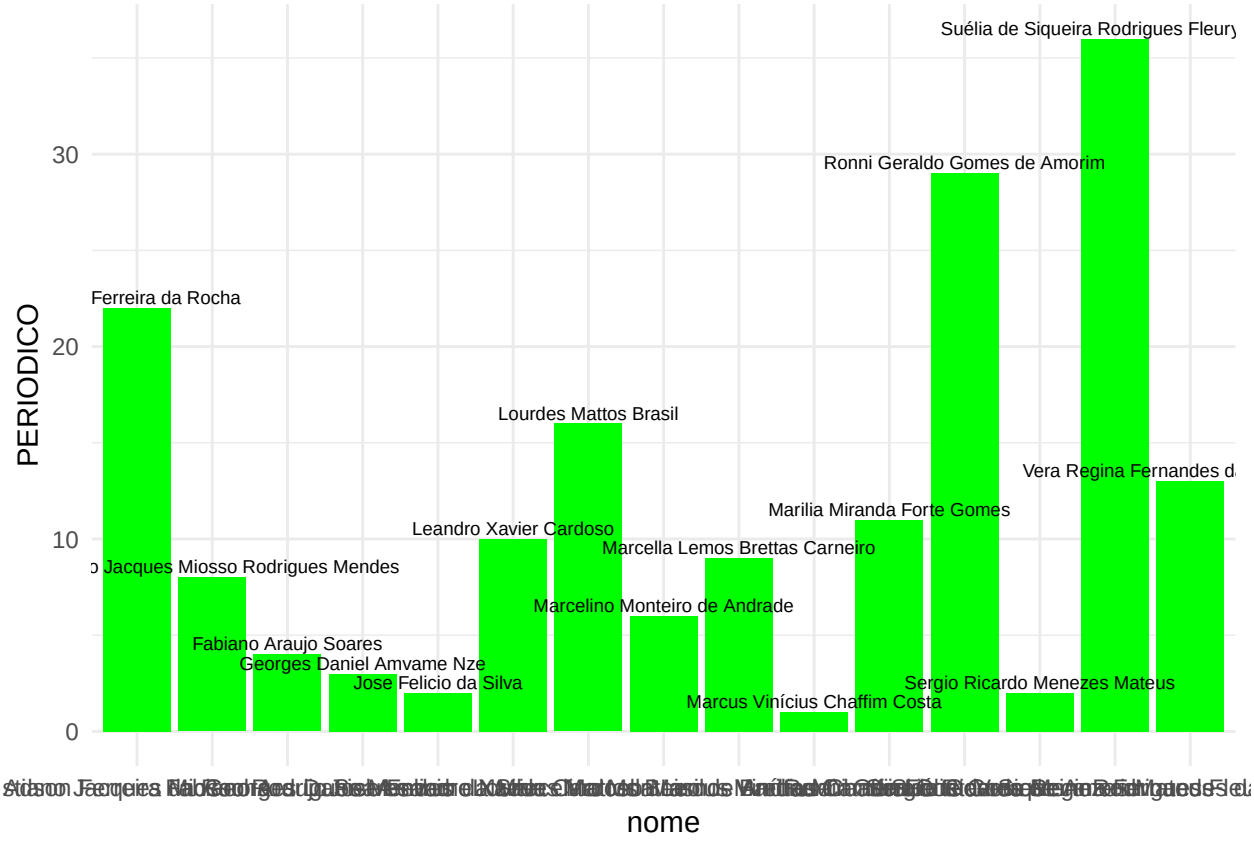
Analisando quandidade de eventos ido por professor.

```
unb.prof.df %>%
group_by(nome) %>%
filter(!is.na(EVENTO))%>%
ggplot(aes(x = nome, y = EVENTO)) +
geom_bar(position = "stack",stat = "identity", fill = "red")+
geom_text(aes(label=nome), vjust=-0.3, size=2.5)+
theme_minimal()
```



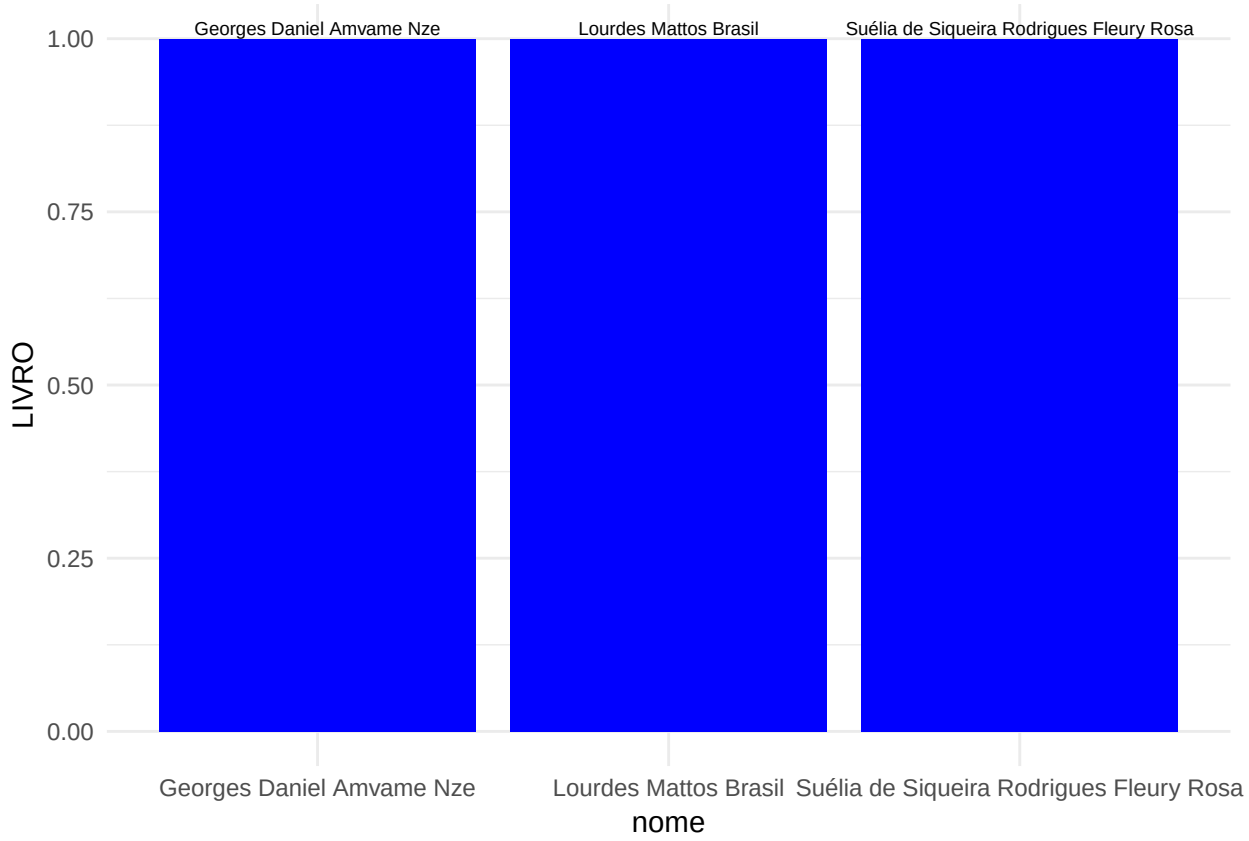
Analisando quandidade de periódicos por professor.

```
unb.prof.df %>%
group_by(nome) %>%
filter(!is.na(PERIODICO))%>%
ggplot(aes(x = nome, y = PERIODICO)) +
geom_bar(position = "stack",stat = "identity", fill = "green")+
geom_text(aes(label=nome), vjust=-0.3, size=2.5)+
theme_minimal()
```



Analisando quandidade de livro por professor.

```
unb.prof.df %>%
group_by(nome) %>%
filter(!is.na(LIVRO))%>%
ggplot(aes(x = nome, y = LIVRO)) +
geom_bar(position = "stack",stat = "identity", fill = "blue")+
geom_text(aes(label=nome), vjust=-0.3, size=2.5)+
theme_minimal()
```

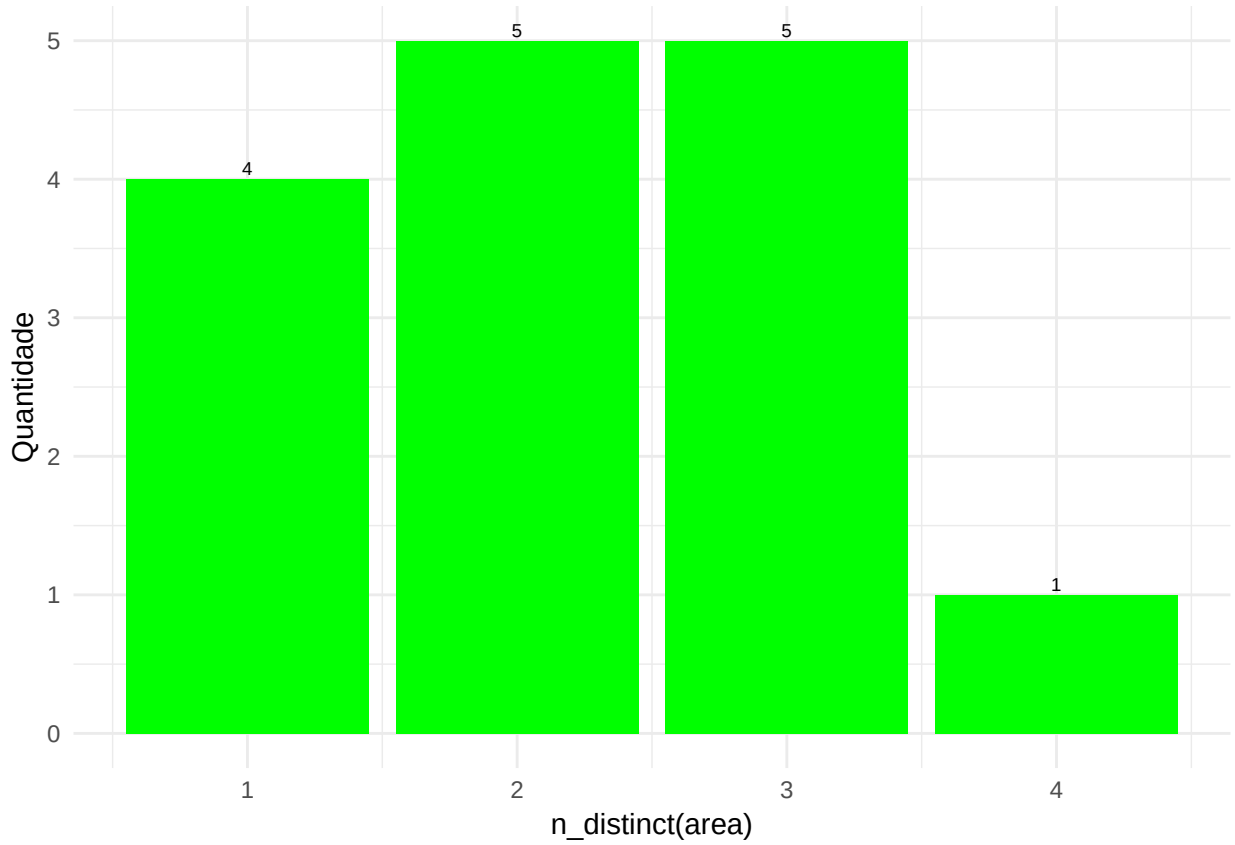


Analisando quandidade de texto em jornais por professor.

```
unb.prof.df %>%
group_by(nome) %>%
filter(!is.na(TEXTO_EM_JORNAIS))%>%
ggplot(aes(x = nome, y = TEXTO_EM_JORNAIS)) +
geom_bar(position = "stack",stat = "identity", fill = "black")+
geom_text(aes(label=nome), vjust=-0.3, size=2.5)+
theme_minimal()
```

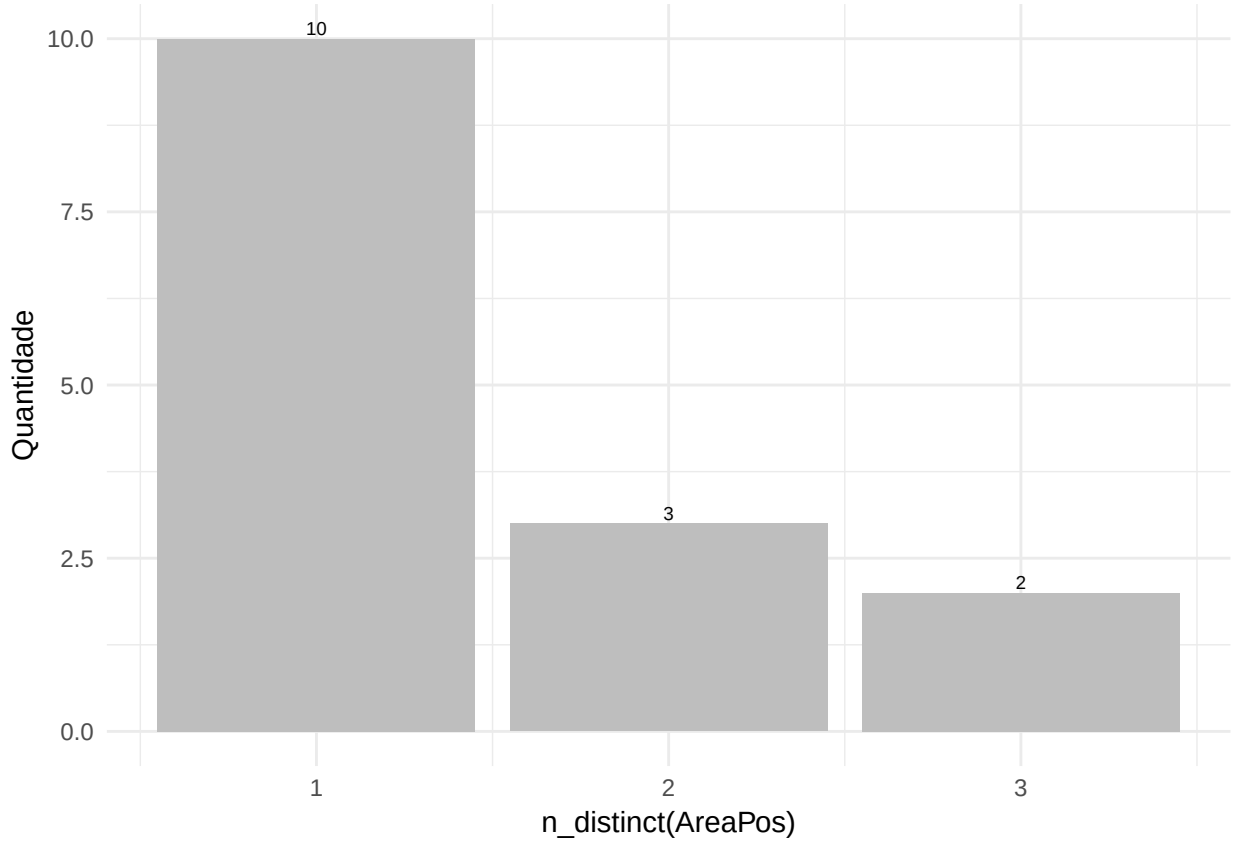
Analisando quantidade de professor por major área e posteriormente o área pós-graduação.

```
unb.prof.df %>%
group_by(`n_distinct(area)`) %>%
summarise(Quantidade = n()) %>%
ggplot(aes(x = `n_distinct(area)`, y = Quantidade)) +
geom_bar(position = "stack",stat = "identity", fill = "green")+
geom_text(aes(label=Quantidade), vjust=-0.3, size=2.5)+
theme_minimal()
```



Analisando quantidade de professor por major área e posteriormente o área pós-graduação.

```
unb.prof.df %>%
group_by(`n_distinct(AreaPos)`) %>%
summarise(Quantidade = n()) %>%
ggplot(aes(x = `n_distinct(AreaPos)`, y = Quantidade)) +
geom_bar(position = "stack",stat = "identity", fill = "grey")+
geom_text(aes(label=Quantidade), vjust=-0.3, size=2.5)+
theme_minimal()
```



CRISP-DM Fase 6 - Implantação (*deployment*)

No caso deste presente trabalho, é basicamente os scripts desenvolvido ao decorrer do mesmo.

Conclusão

Ao final deste trabalho, podemos chegar a conclusão que com a "ferramenta" CRISP-DM foi possível realizar uma análise de dados apartir de vários arquivos **JSON**, assim, foi possível perceber como funciona - pelo menos de maneira simplória - a vida de cientista de dados, como se deve preparar os dados, como se deve buscar informações úteis, analisar determinadas informações, plotar gráficos para um efeito mais visual, ver a dificuldade que falta de padronização causa, ainda mais quando se tem os mais diversos arquivos para analisar das mais diversas fontes.

O trabalho usou como base o arquivo disponibilizado pelo professor, no qual continha as explicações de todas as fase do CRISP-DM, seguindo o modelo passado pelo professor, tentou-se ir descorrendo sobre o script em R que se vinha fazendo e correlacionando com alguma fase do CRISP-DM, mas tal divisão não pode ser considerada totalmente fidedigna, haja visto que muitas vezes há mistura de fases e não apenas fases totalmente isoladas, sem influência de outras, procurou-se explicar de maneira simples como se encaixava cada fase.

Por fim, podemos chegar a alguns resultados valorosos ,como por exemplo : as revistas que tem mais publicações advindas da pós-graduação em engenharia biómedica, mas nem tudo são flores, algumas dificuldades vieram em decorrência da falta de uniformidade, porém podemos aprender bastante com este trabalho e nos interar de como começar a fazer análise de dados.

Referências