Análise da Produção Científica e Acadêmica da Universidade de Brasília - Relatório Parcial sobre Pós-Graduação em Ciências Biológicas

Emanuel Victor Araújo 14/0137475, Jonas Prado 13/0117277 , Maria Júlia Gonçalves 14/0153713

18/11/2018

Resumo

Este documento apresenta o relatório final da disciplina Tópicos Avançados em Computadores - Turma D - 2018.2, do Departamento de Ciência da Computação da Universidade de Brasília, que trata da análise da produção científica e acadêmica na Universidade de Brasília, na área de ciências biológicas. Para isso, foi aplicada o modelo metodológico de mineração de dados denominado CRISP-DM e através dele foram dispostas diversas informações sobre a evolução destes programas de pós-graduação entre os anos de 2010 e 2017.

Introdução

Este trabalho visa, no contexto da pós-graduação do sentido amplo e restrito do Brasil, buscar relações e comparações entre os Programas de Pós-Graduação em Biologia Molecular, Biologia Animal, Biologia Microbiana e Patologia Molecular da Universidade de Brasília para sintetizar o panorama e o paradigma da pós-graduação nestes programas.

CRISP-DM:

A metodologia para desenvolvimento do relatório é baseada no modelo de mineração de dados denominado CRISP-DM (Chapman et al., 2000, Mariscal et al., 2010). Este modelo é caracterizado por um projeto dividido em seis fases, as quais serão tratadas ao longo das seções seguintes.

- Fase 1) Entendimento do negócio: Busca definir um problema de mineração de dados com base nos objetivos e necessidades do projeto.
- Fase 2) Entendimento dos dados: Consiste na coleta e descrição dos dados, incluindo uma análise de qualidade e quantidade dos mesmos - necessários para avaliar a viabilidade do projeto.
- Fase 3) Preparação dos dados: Consiste na estruturação dos dados capturados, incluindo uma fase de limpeza de dados indesejados ou inconsistentes. Esta fase tem como objetivo produzir datasets prontos para análises estatísticas e produção de gráficos.
- Fase 4) Modelagem: Envolve a construção e revisão de modelos estatísticos de interesse para o projeto, além de testes para descarte dos modelos produzidos que não atingirem o grau desejado de confiabilidade.
- Fase 5) Avaliação: Consiste na avaliação dos resultados e do processo como um todo, revisando se todas as questões relevantes para o projeto foram abordadas adequadamente.
- Fase 6) Implementação: Consiste no planejamento e implementação dos entregáveis finais desenvolvidos pelo projeto, incluindo monitoramento e manutenção dos mesmos.

CRISP-DM Fase 1 - Entendimento do Negócio

O que é o Sistema Nacional de Pós-Graduação?

A produção do conhecimento científico, no Brasil, é predominantemente efetuada por meio do Sistema Nacional de Pós-Graduação - SNPG, e mais fortemente relacionada com a formação de doutores nesse sistema (Pátaro e Mezzomo, 2013), por meio de cursos de pós-graduação *strictu sensu*.

Fernandes e Sampaio (2017) já indicaram que a ciência é reconhecidamente um elemento essencial para o desenvolvimento social e econômico de qualquer nação. Assim sendo, faz-se mister aprimorar o SNPG como forma de promoção desse crescimento, visando maximizar o retorno decorrente do emprego dos recursos nele aplicados. A promoção do crescimento do SNPG se dá predominantemente por meio de avaliações regulares de seus programas de pós-graduação, sob responsabilidade da CAPES, que realiza a cada quatro anos um complexo (Leite, 2018, p. 13) e custoso processo de coleta de dados, análise e deliberação sobre as pós-graduações *strictu sensu*, em coerência com o estabelecido no Plano Nacional de Pós-Graduação (PNPG) 2012-2020 (CAPES, 2010) e nos diversos documentos que definem os critérios de organização da pós-graduação em cada área do conhecimento (CAPES, 2018). Leite (2018) faz uma apresentação geral de como se organizam e são avaliadas as pós-graduações no Brasil.

O Plano Nacional de Pós-Graduação (PNPG), por outro lado, define diretrizes estratégicas para desenvolvimento da pós-graduação brasileira, que deve abordar prioritariamente grandes temas de interesse nacional, tais como a redução das assimetrias de desenvolvimento entre as regiões do Brasil, a formação de professores para a educação básica, a formação de recursos humanos para as empresas, a resposta aos grandes desafios brasileiros sobre Água, Energia, Transporte, Controle de Fronteiras, Agronegócio, Amazônia, Amazônia Azul (Mar), Saúde, Defesa, Programa Espacial, além de Justiça, Segurança Pública, Criminologia e Desequilíbrio Regional. O PNPG também traça as diretrizes para financiamento da pós-graduação e sua internacionalização, apresentando conclusões e recomendações.

As avaliações do SNPG, ao atribuirem mensurações de desempenho às diversas pós-graduações que dele fazem parte, geram incentivos e penalidades aos programas, tendo em vista a limitada disponibilidade de recursos para investimento em bolsas, taxas de bancada etc. Embora o sistema seja altamente sofisticado ele é também altamente criticado (Azevedo et al., 2016), sobretudo porque há percalços na busca por um equilíbrio entre as diferentes concepções de finalidade da ciência. Se de um lado a promoção do conhecimento gerado predominantemente nas ditas ciências *hard* constribui para criar fluxos econômicos mais intensos, isso não significa que essa promoção possa ocorrer em detrimento da menor promoção na geração de conhecimento sobre problemas sociais, predominantemente gerado nas ditas ciências *soft*, especialmente das áreas de humanidades, sob pena de ampliação de desigualdades (Azevedo et al., 2016). Esta disciplina propõe que uma maior agilidade e a utilização de critérios mais objetivos na avaliação dos programas poderá facilitar a melhoria do sistema.

Os Colégios, Grandes áreas e áreas da Pós-Graduação Brasileira

A partir de 2018, as diversas áreas da pós-graduação brasileira foram organizadas na forma de colégios, grandes áreas e áreas. Os colégios foram separados da seguinte formar:

- Colégio de Ciências da Vida - que contempla as grandes áreas de Ciências Agrárias, Ciências Biológicas e Ciências da Saúde.
- Colégio de Ciências Exatas, Tecnológicas e Multidisciplinar - que contempla as grandes áreas de Ciencias Exatas e da Terra, Engenharias e Multidisciplinar.
- Colégio de Humanidades - que contempla as grandes áreas de Ciencias Humanas, Ciencias Sociais Aplicadas e Linguistica, Letras e Artes.

Cada um desses colégios, grandes áreas e áreas de conhecimento possuem dinâmicas próprias, e, portanto, não há um modelo universal que se aplique a todas. Existem aspectos comuns, mas também grandes peculiaridades,

descritas parcialmente nos correspondentes documentos de área disponíveis em CAPES (2018).

A UnB dentro do Sistema Nacional de Pós-Graduação

A produção científica da Universidade de Brasília despontou como uma das mais relevantes do país. Segundo o Plano de Internacionalização da Universidade de Brasília 2018-2022, a produção científica da UnB se expandiu consideravelmente período de 2011 a 2016, assim como a relevância das mesma (SciVal, 2018).

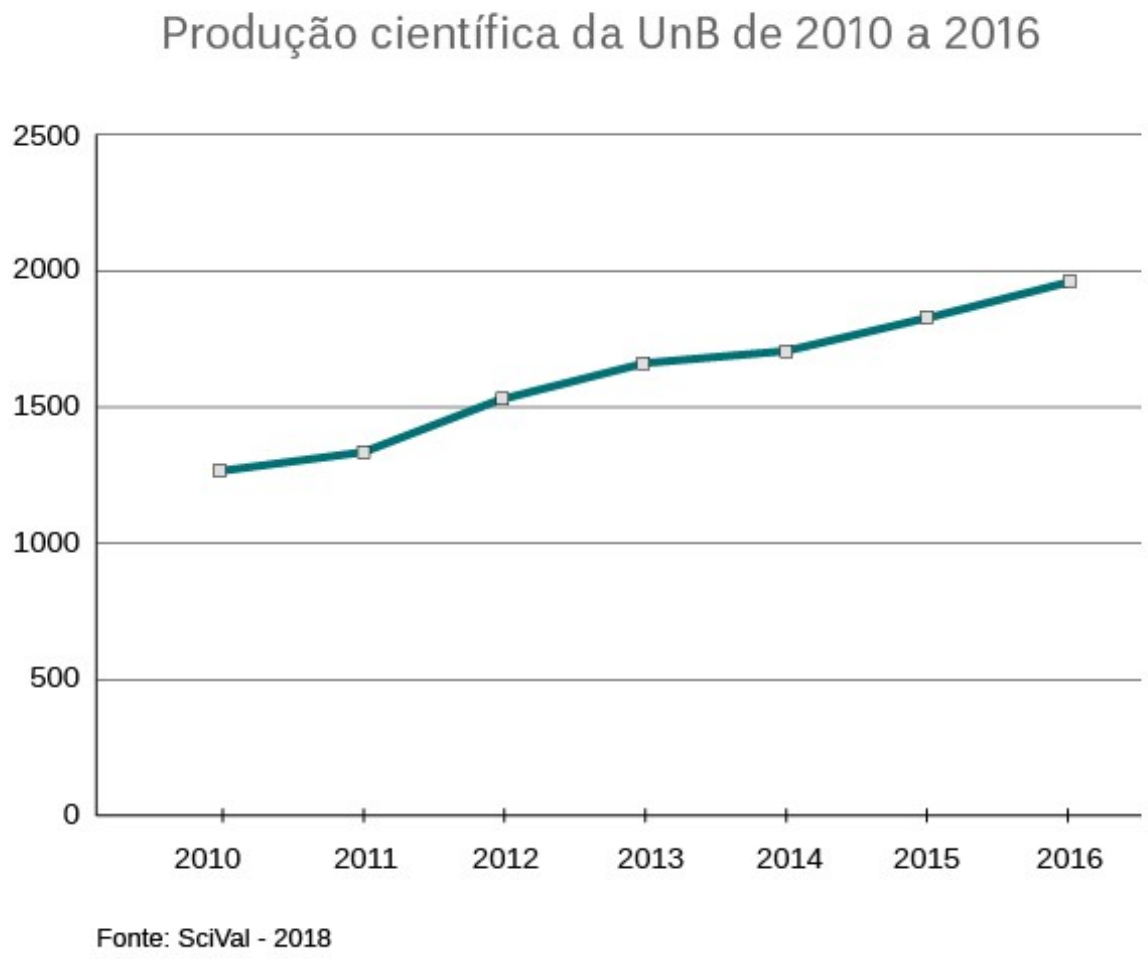


Figure 1: production

Em 2011, o impacto normalizado por citação foi, de acordo com o gráfico, de aproximadamente 0.85, o que indica que os pesquisadores da UnB foram 15% menos citados do que a média global. Similarmente em 2013, ouve uma redução branda no número de pesquisadores citados. Entretanto, de 2013 até 2016 ouve um aumento considerável nesse impacto normalizado - os pesquisadores da UnB atingiram índice de impacto por citação de 1.45 em 2016, sendo 45% mais citados que a média global. Esse aumento na produção cientifica está diretamente associada ao aumento no número de docentes da Universidade. Entre os anos de 2009 e 2013, 577 professores ingressaram no quadro permanente da instituição, com um aumento de 32%. Entre 2013 e 2016, este aumento foi de 185 docentes (8%) [fonte]. Com um quadro de professores maior espera-se naturalmente uma maior produção de artigos e outros matérias científicos pela UnB.

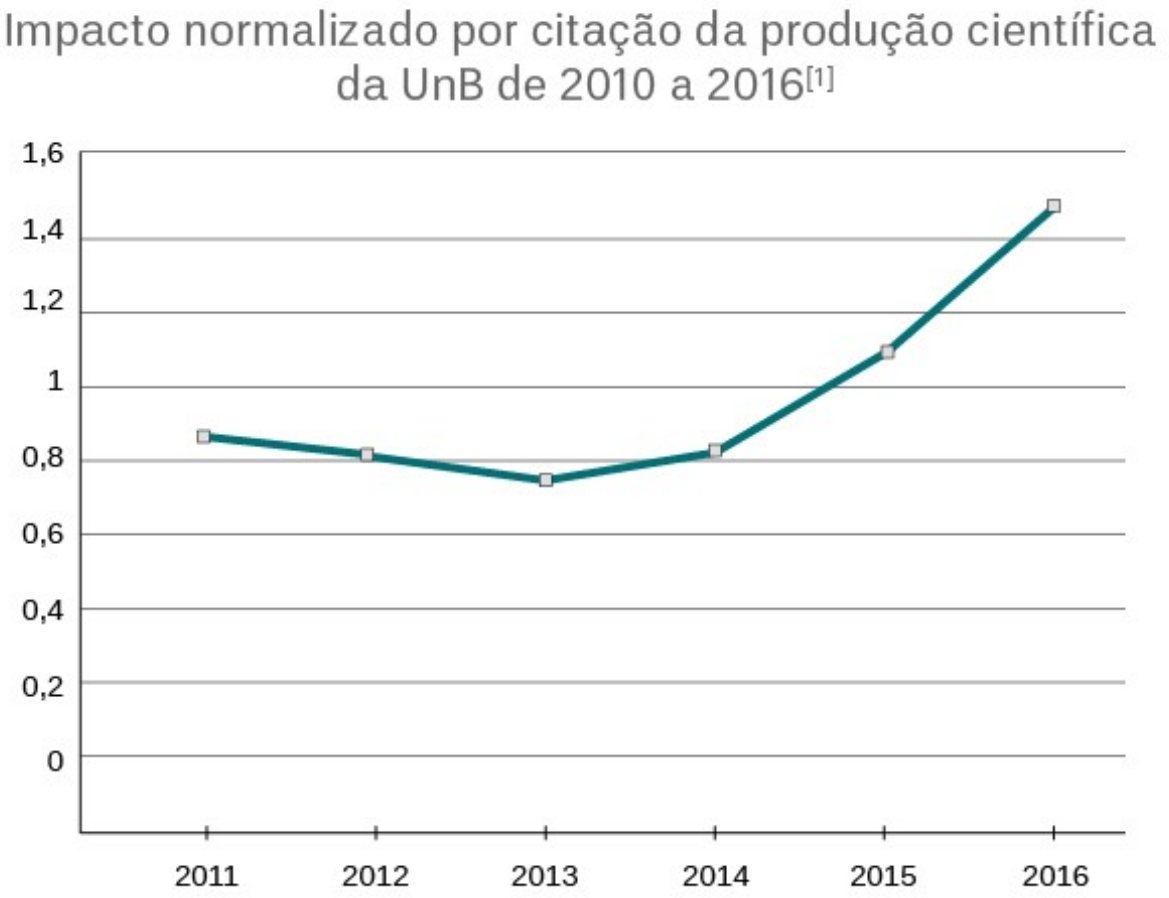


Figure 2: impact

Captura dos dados - Definir JSONS

```
library(jsonlite) #Importado para carga dos arquivos JSON para o R
perfil <- fromJSON("240BiologiaAnimal/240profile.json")
public <- fromJSON("240BiologiaAnimal/240publication.json")
orient <- fromJSON("240BiologiaAnimal/240advise.json")
graphl <- fromJSON("240BiologiaAnimal/240graph.json")
df.prog <- read.table("UnBPosGeral/prof_prog.csv", sep = ",",
                      colClasses = "character", encoding = "UTF-8", header = TRUE)
```

CRISP-DM Fase 2 - Entendimento dos Dados

A segunda parte do CRISP-DM consiste no entendimento dos dados. Para realizar análises significativas com os **datasets** disponíveis, é essencial ter um bom entendimento sobre a forma que estão organizados.

Os arquivos utilizados são provenientes da plataforma Elattes e compilam informações sobre aa produção científica dos professores do **Programa de Pós-graduação em Biologia Animal** no período de 2010 a 2017.

Os **datasets** que serão trabalhados consistem em: perfil profissional; orientações de mestrado e doutorado realizadas; produções bibliográficas e redes de colaboração entre os pesquisadores.

Arquivos Analisados

Os arquivos com informações sobre os pesquisadores do Programa de **Biologia Animal**:

- **240BiologiaAnimal/240profile.json**: apresenta dados sobre o **perfil** de todos os pesquisadores.
- **240BiologiaAnimal/240publication.json**: apresenta dados sobre as **publicações** e **produções bibliográficas** geradas por todos os pesquisadores.
- **240BiologiaAnimal/240advise.json**: apresenta dados sobre **orientações de mestrado e doutorado** feitas por todos os pesquisadores.
- **240BiologiaAnimal/240graph.json**: apresenta dados sobre **produções bibliográficas coolaborativas** feitas entre os pesquisadores.

Análise estrutural dos dados

Para continuar com as análises, as seguintes bibliotecas são selecionadas:

```
#library(tidyverse) #Importado para manipulação de tibbles
library(listviewer) #Importado para análise dos arquivos JSON
library(igraph) #Importado para manipulação de grafo
```

```
##
## Attaching package: 'igraph'

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```
library(dplyr) #Importado para uso do Operador Pipe
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:igraph':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr) #Importado para uso da função spread()
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:igraph':
##
##     crossing
```

```
library(ggplot2) #Importado para visualizações com ggplot()

setwd("~/Repository/DataScience/DS4A-BioAni-BioMic-BioMol-PatMol")
source("elattes.ls2df.R")
```

Após a importação e definição das bibliotecas utilizadas, podemos utilizar funções de tais pacotes para análise sistêmica dos dados, mas para isto, estes deves sem primeiramente descritos por meio de funções de descrição do pacote `dplyr` que facilita a visualização e manipulação dos dados:

SEPARACAO DOS CAMPOS DE DF.PROG

```
df.prog <- df.prog %>% separate(idLattes.Docente.Categoria.Grande.Area.Area.de.Avaliacao.Codigo.AreaPos
                      c("idLattes", "Docente", "Categoria", "GrandeArea", "AreaDeAvaliacao", "Codigo", "
                      sep = ";", extra = "drop", fill = "right")
```

Número de docentes na base e a construção das listas dos mesmos:

```
length(perfil)
```

```
## [1] 19
```

```
ProfileList <- list()
for (i in 1:length(perfil)) {
  ProfileList <- rbind(ProfileList, perfil[[i]]$nome)
}
```

Análise das listas:

Número de áreas de atuação cumulativas:

```
sum(sapply(perfil, function(x) nrow(x$areas_de_atuacao)))
```

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

Numero de areas de atuacao por pessoa

```
table(unlist(sapply(perfil, function(x) nrow(x$areas_de_atuacao))))
```

```
##
## 2 3 4 5 6
## 2 3 7 5 2
```

Numero de pessoas por grande area

```
table(unlist(sapply(perfil, function(x) (x$areas_de_atuacao$grande_area))))
```

```
##
##          CIENCIAS_AGRARIAS        CIENCIAS_BIOLOGICAS
##                          2                         68
## CIENCIAS_EXATAS_E_DA_TERRA           CIENCIAS_HUMANAS
##                          2                          5
##                ENGENHARIAS
##                          1
```

Numero de pessoas que produziram os especificos tipos de producao

```
table(unlist(sapply(perfil, function(x) names(x$producao_bibiografica))))
```

```
##
##                          ARTIGO_ACEITO
##                                      3
##                      CAPITULO_DE_LIVRO
##                                     14
## DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA
##                                      1
##                                 EVENTO
##                                     15
##                                  LIVRO
##                                      4
##                               PERIODICO
##                                     19
##                         TEXTO_EM_JORNAIS
##                                      4
```

Numero de publicacoes por tipo

```
sum(sapply(perfil, function(x) length(x$producao_bibiografica$ARTIGO_ACEITO$ano)))
```

```
## [1] 6
```

```
sum(sapply(perfil, function(x) length(x$producao_bibiografica$CAPITULO_DE_LIVRO$ano)))
```

```
## [1] 41
```

```
sum(sapply(perfil, function(x) length(x$producao_bibiografica$LIVRO$ano)))
```

```
## [1] 5
```

```
sum(sapply(perfil, function(x) length(x$producao_bibiografica$PERIODICO$ano)))
```

```
## [1] 681
```

```
sum(sapply(perfil, function(x) length(x$producao_bibiografica$TEXTO_EM_JORNAIS$ano)))
```

```
## [1] 4
```

Número de pessoas por quantitativo de produções por pessoa

```
table(unlist(sapply(perfil, function(x) length(x$producao_bibiografica$ARTIGO_ACEITO$ano))))
```

```
##
##  0  2
## 16  3
```

```
table(unlist(sapply(perfil, function(x) length(x$producao_bibiografica$CAPITULO_DE_LIVRO$ano))))
```

```
##
## 0 1 2 3 4 5 9
## 5 2 7 1 2 1 1
```

```
table(unlist(sapply(perfil, function(x) length(x$producao_bibiografica$LIVRO$ano))))
```

```
##
##  0  1  2
## 15  3  1
```

```
table(unlist(sapply(perfil, function(x) length(x$producao_bibiografica$PERIODICO$ano))))
```

```
##
##  10  11  13  18  19  21  23  26  27  33  40  44  57  68 103 104
##   1   1   1   1   3   1   1   2   1   1   1   1   1   1   1   1
```

```
table(unlist(sapply(perfil, function(x) length(x$producao_bibiografica$TEXTO_EM_JORNAIS$ano))))
```

```
##
##  0  1
## 15  4
```

Número de produções por ano

```
table(unlist(sapply(perfil, function(x) (x$producao_bibiografica$ARTIGO_ACEITO$ano))))
```

```
##
## 2016 2017
##    1    5
```

```
table(unlist(sapply(perfil, function(x) (x$producao_bibiografica$CAPITULO_DE_LIVRO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##    7    9    3    4    3    4    5    6
```

```
table(unlist(sapply(perfil, function(x) (x$producao_bibiografica$LIVRO$ano))))
```

```
##
## 2013 2014 2016
##    2    1    2
```

```
table(unlist(sapply(perfil, function(x) (x$producao_bibiografica$PERIODICO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##   81   81   81   90   90   76   91   91
```

```
table(unlist(sapply(perfil, function(x) (x$producao_bibiografica$TEXTO_EM_JORNAIS$ano))))
```

```
##
## 2012 2013 2014 2016
```

```
##    1    1    1    1
```

Número de pessoas que realizaram diferentes tipos de orientações

```
length(unlist(sapply(perfil, function(x) names(x$orientacoes_academicas))))
```

```
## [1] 102
```

Número de pessoas por tipo de orientação

```
table(unlist(sapply(perfil, function(x) names(x$orientacoes_academicas))))
```

```
##
##               ORIENTACAO_CONCLUIDA_DOUTORADO
##                                           18
##                ORIENTACAO_CONCLUIDA_MESTRADO
##                                           19
##           ORIENTACAO_CONCLUIDA_POS_DOUTORADO
##                                            7
##            ORIENTACAO_EM_ANDAMENTO_DOUTORADO
##                                           16
## ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA
##                                           10
##             ORIENTACAO_EM_ANDAMENTO_MESTRADO
##                                           14
##                OUTRAS_ORIENTACOES_CONCLUIDAS
##                                           18
```

Número de orientacões concluídas

```
sum(sapply(perfil, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano)))
```

```
## [1] 136
```

```
sum(sapply(perfil, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano)))
```

```
## [1] 92
```

```
sum(sapply(perfil, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano)))
```

```
## [1] 39
```

Número de pessoas por quantitativo de orientacões por pessoa

```
table(unlist(sapply(perfil, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$an
```

```
##
##  1  2  5  6  7  8  9 10 11 12 13
##  1  2  3  2  2  3  1  1  2  1  1
```

```
table(unlist(sapply(perfil, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$a
```

```
##
##  0  1  2  3  5  6  7  8 10 13
##  1  4  2  2  1  2  2  3  1  1
```

```
table(unlist(sapply(perfil, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORA
```

```
##
##  0  2  3  5  6  7 14
## 12  2  1  1  1  1  1
```

Número de orientações por ano

```
table(unlist(sapply(perfil, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##   17   14   21   26   20   13   19    6
```

```
table(unlist(sapply(perfil, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##    2   15   15    4   11   15   11   19
```

```
table(unlist(sapply(perfil, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano
```

```
##
## 2010 2011 2012 2013 2014 2015 2016
##    3    4    9    5    9    3    6
```

Análise dos dataframes:

Extrai perfis dos professores:

```
perfil.df.professores <- extrai.perfis(perfil)
```

Extrai produção bibliográfica de todos os professores

```
perfil.df.publicacoes <- extrai.producoes(perfil) %>%
  select(tipo_producao, everything()) %>% arrange(tipo_producao)
```

Extrai orientações

```
perfil.df.orientacoes <- extrai.orientacoes(perfil) %>%
  select(id_lattes_orientadores, natureza, ano, orientacao, everything())
```

Extrai áreas de atuação

```
perfil.df.areas.de.atuacao <- extrai.areas.atuacao(perfil) %>%
  select(idLattes, everything())
```

Cria arquivo com dados quantitativos para análise

```
perfil.df <- data.frame()
perfil.df <- perfil.df.professores %>%
  select(idLattes, nome, resumo_cv, senioridade) %>%
  left_join(
    perfil.df.orientacoes %>%
      select(orientacao, idLattes) %>%
      filter(!grepl("EM_ANDAMENTO", orientacao)) %>%
      group_by(idLattes) %>%
      count(orientacao) %>%
      spread(key = orientacao, value = n),
    by = "idLattes") %>%
  left_join(
    perfil.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(
    perfil.df.areas.de.atuacao %>%
      select(area, idLattes) %>%
      group_by(idLattes) %>%
      summarise(n_distinct(area)),
    by = "idLattes")

glimpse(perfil.df)
```

```
## Observations: 19
## Variables: 15
## $ idLattes                                <chr> "1612292306950738", "18...
## $ nome                                    <chr> "Renato Caparroz", "Mon...
## $ resumo_cv                               <chr> "Formado em Zootecnia p...
## $ senioridade                             <chr> "9", "8", "9", "9", "9"...
## $ ORIENTACAO_CONCLUIDA_DOUTORADO          <int> 6, 1, 6, 7, NA, 1, 1, 2...
## $ ORIENTACAO_CONCLUIDA_MESTRADO           <int> 8, 6, 10, 1, 9, 5, 2, 5...
## $ ORIENTACAO_CONCLUIDA_POS_DOUTORADO      <int> NA, NA, NA, NA, NA, NA,...
## $ OUTRAS_ORIENTACOES_CONCLUIDAS           <int> 18, 9, 17, NA, 18, 9, 2...
## $ CAPITULO_DE_LIVRO                       <int> 2, NA, 4, 9, 1, 4, 2, N...
## $ DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA  <int> NA, NA, NA, NA, NA, NA,...
## $ EVENTO                                  <int> 19, 14, 12, NA, 11, 10,...
## $ LIVRO                                   <int> NA, NA, NA, 1, NA, NA, ...
## $ PERIODICO                               <int> 13, 10, 21, 68, 18, 44,...
## $ TEXTO_EM_JORNAIS                        <int> NA, NA, NA, NA, NA, 1, ...
## $ `n_distinct(area)`                      <int> 2, 3, 1, 1, 4, 2, 3, 1,...
```

CRISP-DM Fase 3 - Preparação dos Dados

A terceira fase do CRISP é conhecida por ser a parte de preparação dos dados. Tal fase possui como característica a execução de atividades para construir o conjunto final de dados a partir dos dados brutos iniciais. Pode-se separar essa etapa em cinco momentos que serão descritos nessa seção.

- Seleção dos dados
- Limpeza dos dados
- Construção dos dados
- Integração dos dados
- Formatação dos dados

Na etapa de seleção dos dados a entrada é o conjunto de dados bruto e nela ocorre a decisão dos dados a serem usados para análise. Os critérios incluem relevância para as metas de mineração de dados, qualidade e restrições técnicas, como limites no volume de dados ou tipos de dados. Então vem a fase da limpeza que recebe a seleção de dados úteis efetuada anteriormente e é efetuado um aumento na qualidade dos dados para o nível exigido pelas técnicas de análise selecionadas. Aqui pode haver o uso de técnicas mais elaboradas, como a estimativa de dados ausentes por modelagem e inserção de padrões adequados.

O terceiro passo é a construção dos dados. Essa tarefa inclui operações de preparação de dados construtivos, como a produção de atributos derivados, novos registros ou valores transformados para atributos existentes. A penúltima atividade é a integração dos dados. Este é o momento no qual as informações são combinadas de

vários bancos de dados, tabelas ou registros para criar novos registros ou valores. Por fim, ocorre a tarefa de formatação dos dados, que é a realização de modificações na estrutura dos dados de forma que as operações planejadas possam ser efetuadas de forma conveniente.

Para tornar a análise mais fácil de ser feita e até mesmo para possibilitar a realização de comparações ao final, os mesmos procedimentos foram realizados para os três programas de pós-graduação. Além disso, é importante ressaltar que as variáveis e estruturas montadas foram nomeados de forma mnemônica permitindo a distinção de diferentes programas e aspectos, como orientações, publicações, entre outros.

Análise dos dados no formato lista

Número de Publicacoes em periódicos

```
sum(sapply(public$PERIODICO, function(x) length(x$natureza)))
```

```
## [1] 529
```

Anos analisados

```
names(public$PERIODICO)
```

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

20 revistas mais publicadas

```
head(sort(table(as.data.frame(unlist
    (sapply(public$PERIODICO, function(x) unlist(x$periodico)))
  )), decreasing = TRUE),20)
```

```
##
##                                                Plos One
##                                                      18
##                                        Toxicon (Oxford)
##                                                      14
##                    Genetics and Molecular Biology (Impresso)
##                                                      11
##                         Journal of Biomedical Nanotechnology
##                                                       8
##                              Genetics and Molecular Research
##                                                       7
##               International Journal of Nanomedicine (Online)
##                                                       7
##                                  Protein and Peptide Letters
##                                                       7
## Revista de la Universidad Industrial de Santander. Salud
##                                                       7
##                                Journal of Nanobiotechnology
##                                                       6
##                                         Scientific Reports
##                                                       6
##                                            Theriogenology
##                                                       6
##                                 Behavioural Brain Research
##                                                       5
##                         Frontiers in Behavioral Neuroscience
##                                                       5
##                                   Frontiers in Pharmacology
```

```
##                                                    5
##                                  Genética na Escola
##                                                    5
##                Journal of Nanomedicine & Nanotechnology
##                                                    5
##                                           Nanomedicine
##                                                    5
##                                Small Ruminant Research
##                                                    5
##   Bulletin of Environmental Contamination and Toxicology
##                                                    4
##        Journal of Nanoscience and Nanotechnology (Print)
##                                                    4
```

Análise dos dados no formato dataframe

```
public.periodico.df <- pub.ls2df(public, 1) #artigos
public.livros.df <- pub.ls2df(public, 2) #livros
public.eventos.df <- pub.ls2df(public, 5) #eventos
```

Publicações por ano

```
table(public.periodico.df$ano)
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##   70   68   60   70   70   53   66   72
```

20 revistas mais publicadas (análise com dataframe)

```
head(sort(table(public.periodico.df$periodico), decreasing = TRUE), 20)
```

```
##
##                                          Plos One
##                                                18
##                                   Toxicon (Oxford)
##                                                14
##               Genetics and Molecular Biology (Impresso)
##                                                11
##                    Journal of Biomedical Nanotechnology
##                                                 8
##                         Genetics and Molecular Research
##                                                 7
##           International Journal of Nanomedicine (Online)
##                                                 7
##                             Protein and Peptide Letters
##                                                 7
## Revista de la Universidad Industrial de Santander. Salud
##                                                 7
##                            Journal of Nanobiotechnology
##                                                 6
##                                   Scientific Reports
##                                                 6
##                                      Theriogenology
##                                                 6
##                              Behavioural Brain Research
##                                                 5
```

```
##                       Frontiers in Behavioral Neuroscience
##                                                          5
##                                  Frontiers in Pharmacology
##                                                          5
##                                       Genética na Escola
##                                                          5
##                   Journal of Nanomedicine & Nanotechnology
##                                                          5
##                                                Nanomedicine
##                                                          5
##                                    Small Ruminant Research
##                                                          5
##   Bulletin of Environmental Contamination and Toxicology
##                                                          4
##        Journal of Nanoscience and Nanotechnology (Print)
##                                                          4
```

Número de Orientações de Mestrado e Doutorado

```
sum(sapply(orient$ORIENTACAO_CONCLUIDA_DOUTORADO, function(x) length(x$natureza))) +
  sum(sapply(orient$ORIENTACAO_CONCLUIDA_MESTRADO, function(x) length(x$natureza)))
```

```
## [1] 220
```

Construição de dataframe de orientações

```
orient.posdoutorado.df <- ori.ls2df(orient, 6) #pos-Doutorado concluido
orient.doutorado.df <- ori.ls2df(orient, 7) #Doutorado concluido
orient.mestrado.df <- ori.ls2df(orient, 8) #Mestrado concluido

orient.df <- rbind(rbind(orient.posdoutorado.df, orient.doutorado.df), orient.mestrado.df)
```

Construção de Grafo de Colaborações

```
g <- g.ls2ig(graphl)
df <- as.data.frame(V(g)$name); colnames(df) <- "Idlattes"
df <- left_join(df, df.prog, by = c("Idlattes" = "idLattes")) #
```

```
## Warning: Column `Idlattes`/`idLattes` joining factor and character vector,
## coercing into character vector
```

Apenas para fins de análise inicial, foram retiradas do grafo as observacoes de pesquisadores duplicados quando inclusos em mais de um programa, vide código abaixo.

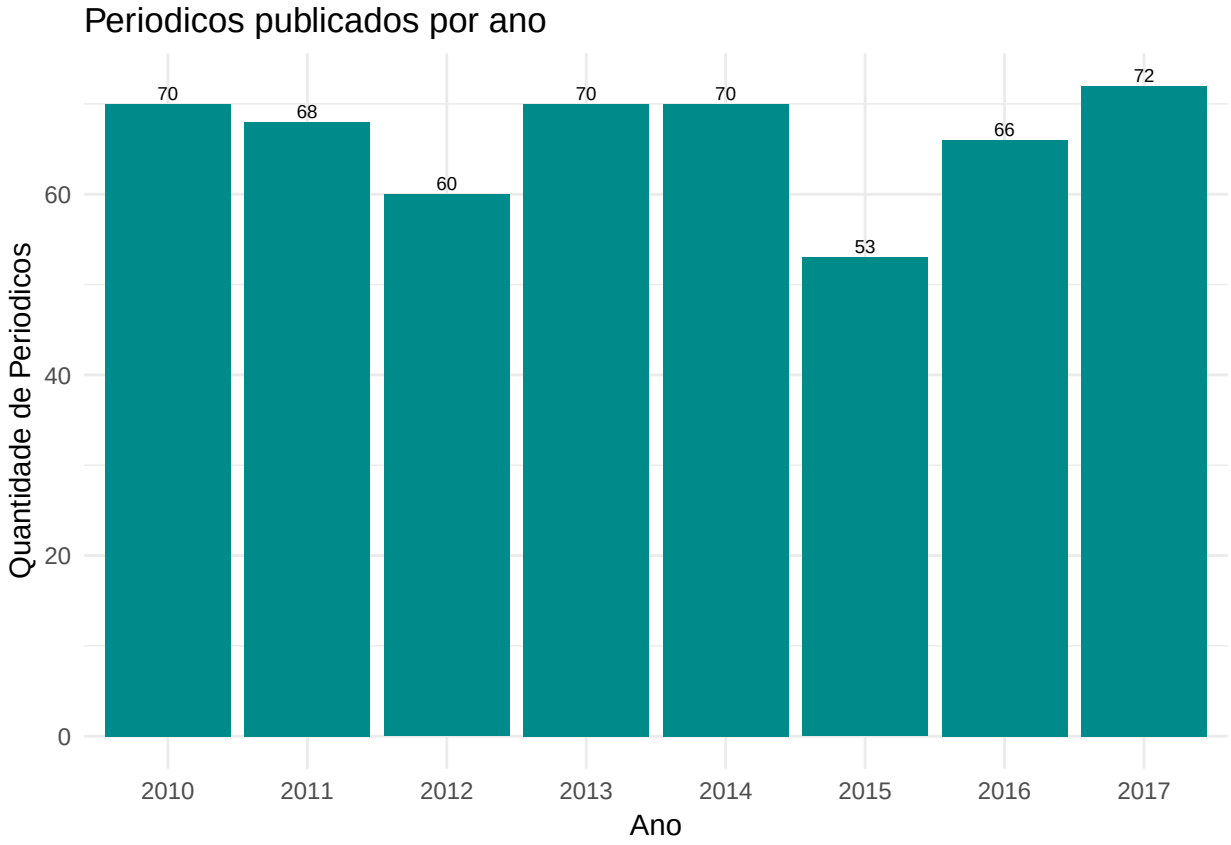
```
df <- df %>% group_by(Idlattes) %>%
  slice(1L)
V(g)$programa <- df$Programa
V(g)$orient_dout <- perfil.df$ORIENTACAO_CONCLUIDA_DOUTORADO
V(g)$orient_mest <- perfil.df$ORIENTACAO_CONCLUIDA_MESTRADO
V(g)$publicacao <- perfil.df$PERIODICO
V(g)$eventos <- perfil.df$EVENTO
```

CRISP-DM Fases 4 a 6 - Resultados e visualizações:

Foram escolhidos alguns resultados, em relação aos dados encontrados durante o processo, para serem plotados em gráficos.

Grafico de barras; periodicos por ano

```
public.periodico.df %>%
  group_by(ano) %>%
  summarise(Quantidade = n()) %>%
  ggplot(aes(x = ano, y = Quantidade)) +
  geom_bar(position = "stack",stat = "identity", fill = "darkcyan")+
  ggtitle("Periodicos publicados por ano") +
  geom_text(aes(label=Quantidade), vjust=-0.3, size=2.5)+
  theme_minimal() + labs(x="Ano",y="Quantidade de Periodicos")
```

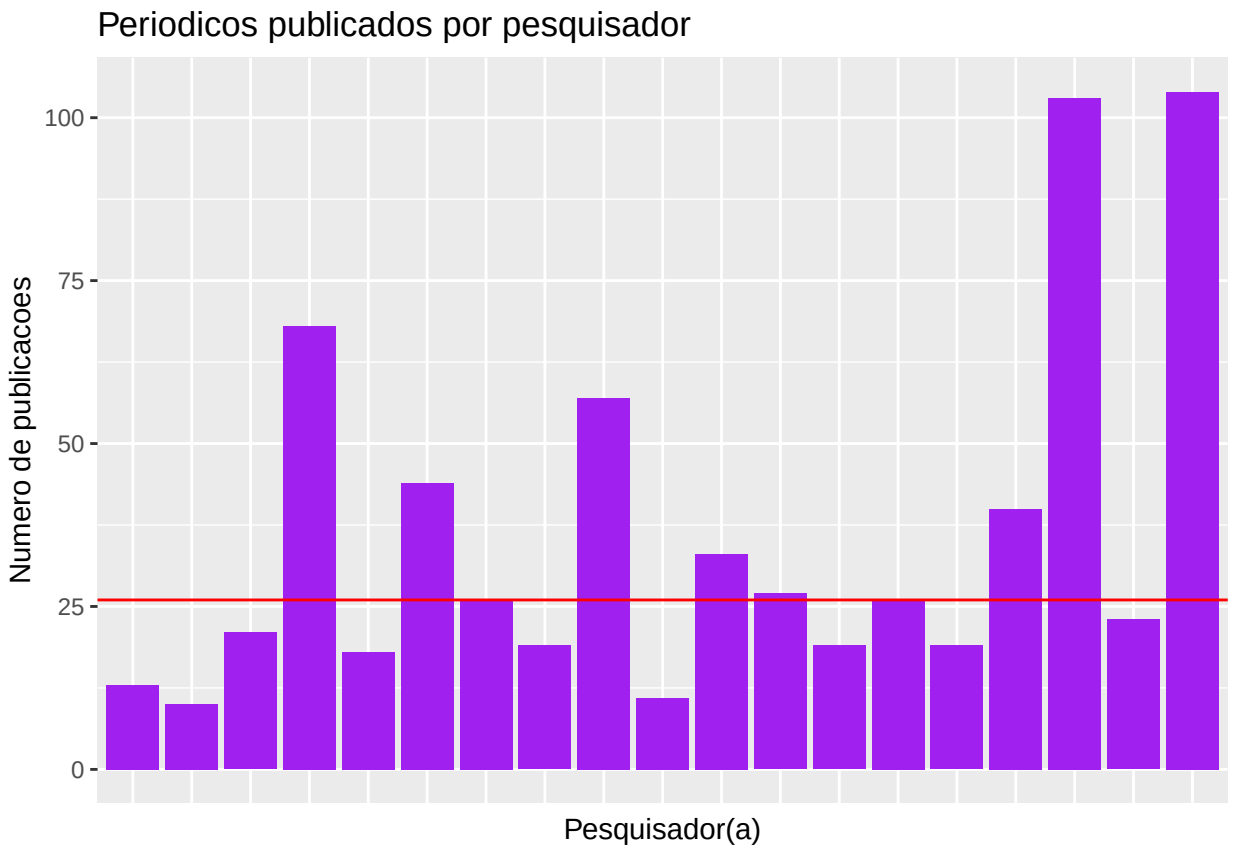


Sob uma perspectiva de publicações em periódicos, o Programa de Pós-Graduação de Biologia Animal mantém se estável com flutuações normais, mas apresentou uma queda considerável de 2014 para 2015, diminuindo o número de publicações em cerca de 18% mas recuperando em anos posteriores.

Quantidade de periodicos publicados por professor(a) entre 2010 e 2017

```
perfil.df %>%
  ggplot(aes(idLattes,PERIODICO)) +
  geom_col(fill = "purple") +
```

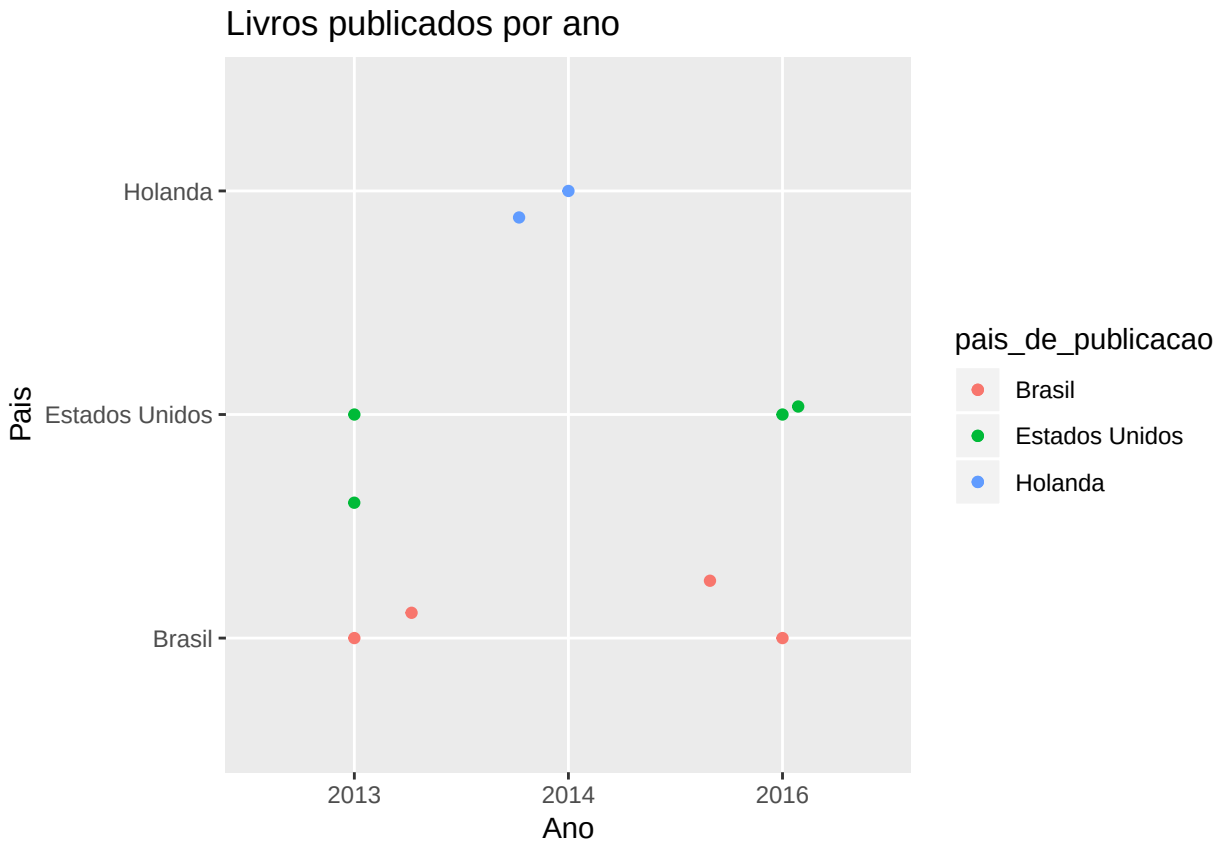
```
  ggtitle("Periodicos publicados por pesquisador") +
  theme(legend.position="right",legend.text=element_text(size=7)) +
  guides(fill=guide_legend(nrow=5, byrow=TRUE, title.position = "top")) +
  labs(x="Pesquisador(a)",y="Numero de publicacoes") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
  geom_hline(yintercept = sum(perfil.df %>% summarize(x = median(PERIODICO))), color = "red")
```



O gráfico acima demonstra o total de publicações para cada pesquisador, demonstrando relativa homogeneidade entre a quantidade de pesquisadores em média, porém há aqueles que deslocam a média para cima por estarem fora do desvio padrão acima ou abaixo, sendo que há destaque para dois pesquisadores que tem um número de publicação que corresponde cerca de quatro vezes o valor da média.

Publicação de livros por pais/ano

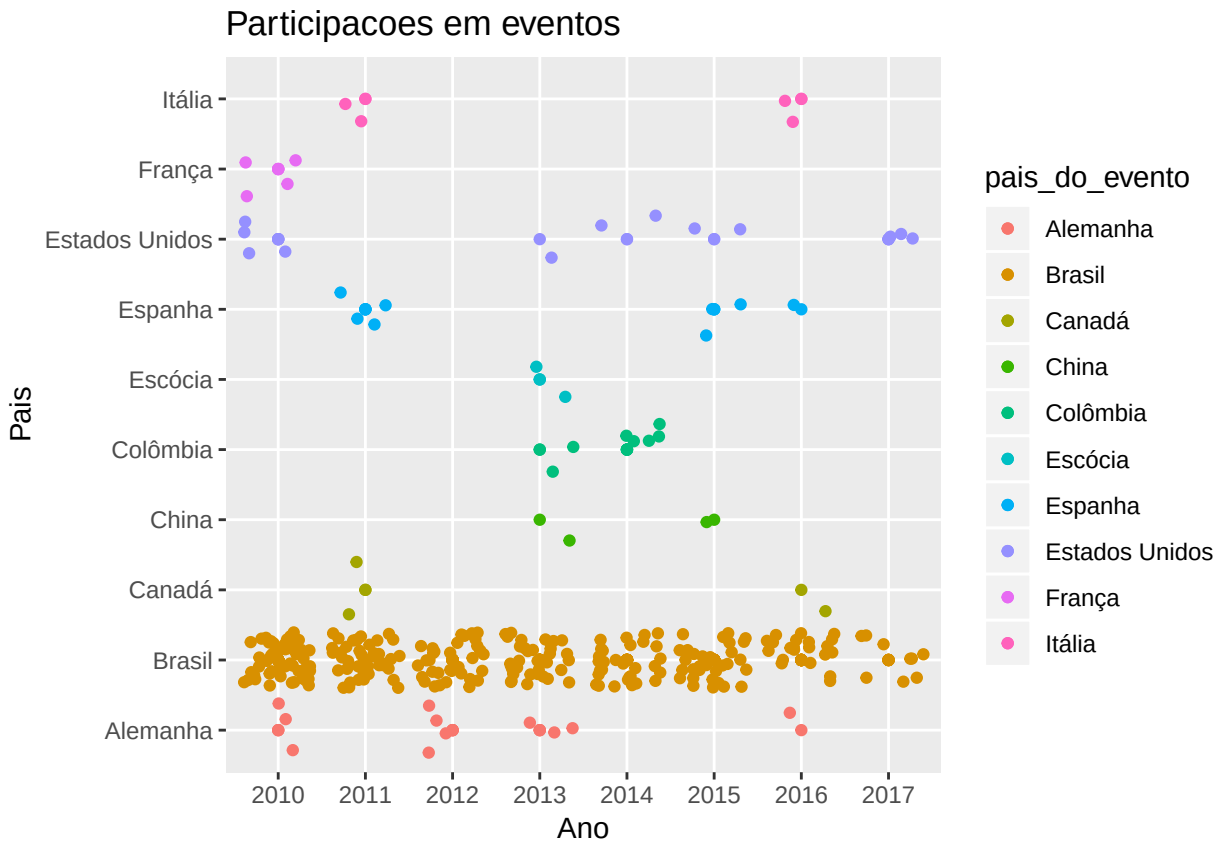
```
public.livros.df %>%
  group_by(ano,pais_de_publicacao) %>%
  ggplot(aes(x=ano,y=pais_de_publicacao, color= pais_de_publicacao)) +
  ggtitle("Livros publicados por ano") +
  xlab("Ano") + ylab("Pais") + geom_point() + geom_jitter()
```



Denota-se homogeneidade na distribuição de livros publicados para cada país de publicação, onde os estados unidos se equiparam ao Brasil quanto a número de lívros publicados para o PPG em Biologia Animal

Eventos nacionais e internacionais

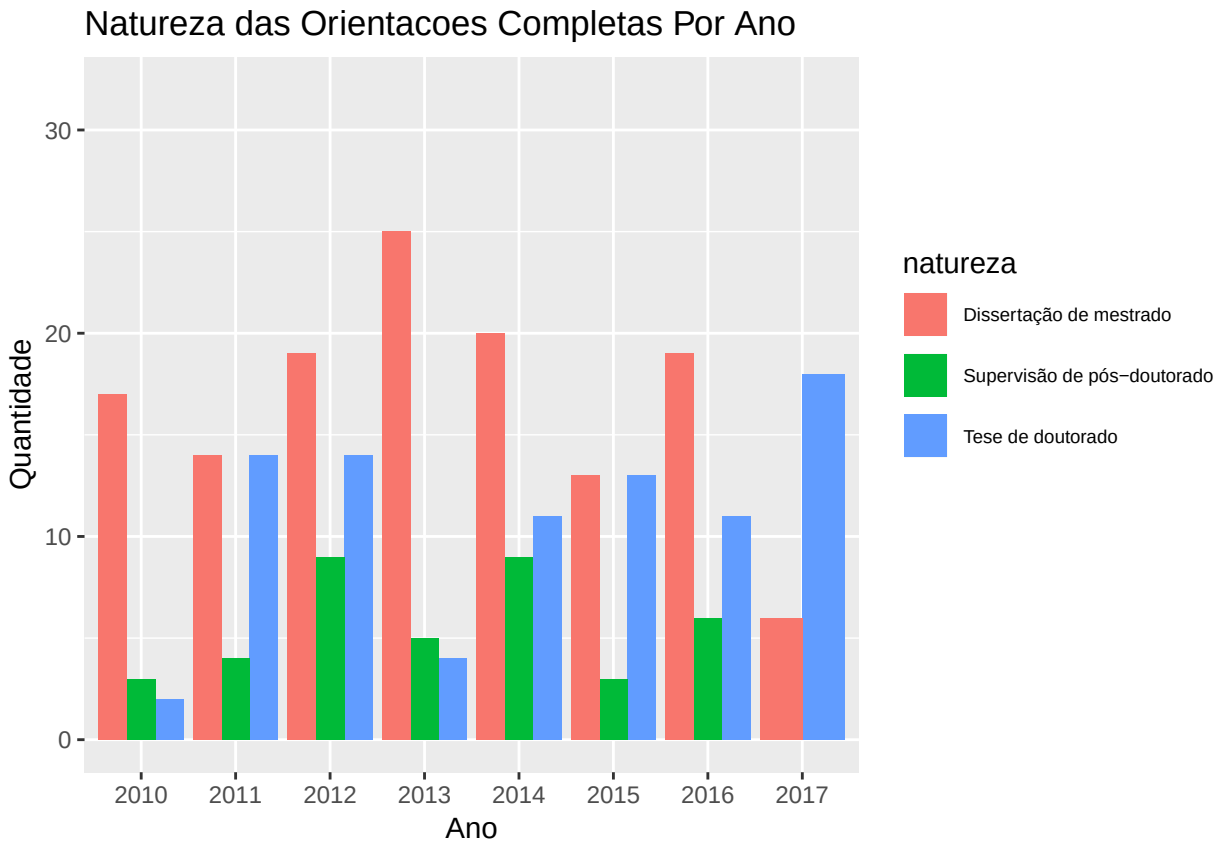
```
public.eventos.df %>%
  filter(pais_do_evento %in%
           c(names(head(sort(table(public.eventos.df$pais_do_evento)
                             , decreasing = TRUE), 10)))) %>%
  group_by(ano_do_trabalho,pais_do_evento) %>%
  ggplot(aes(x=ano_do_trabalho,y=pais_do_evento, color= pais_do_evento)) +
  ggtitle("Participacoes em eventos") +
  xlab("Ano") + ylab("Pais") + geom_point() + geom_jitter()
```



Considerando eventos, o Programa de Biologia Animal teve comparecimento maior em eventos no Brasil que em países estrangeiros. Ainda assim, o gráfico mostra que o programa esteve presente em eventos de sede internacional em todos os anos registrados na base de dados, principalmente nos Estados Unidos.

Orientacoes completas por ano e natureza

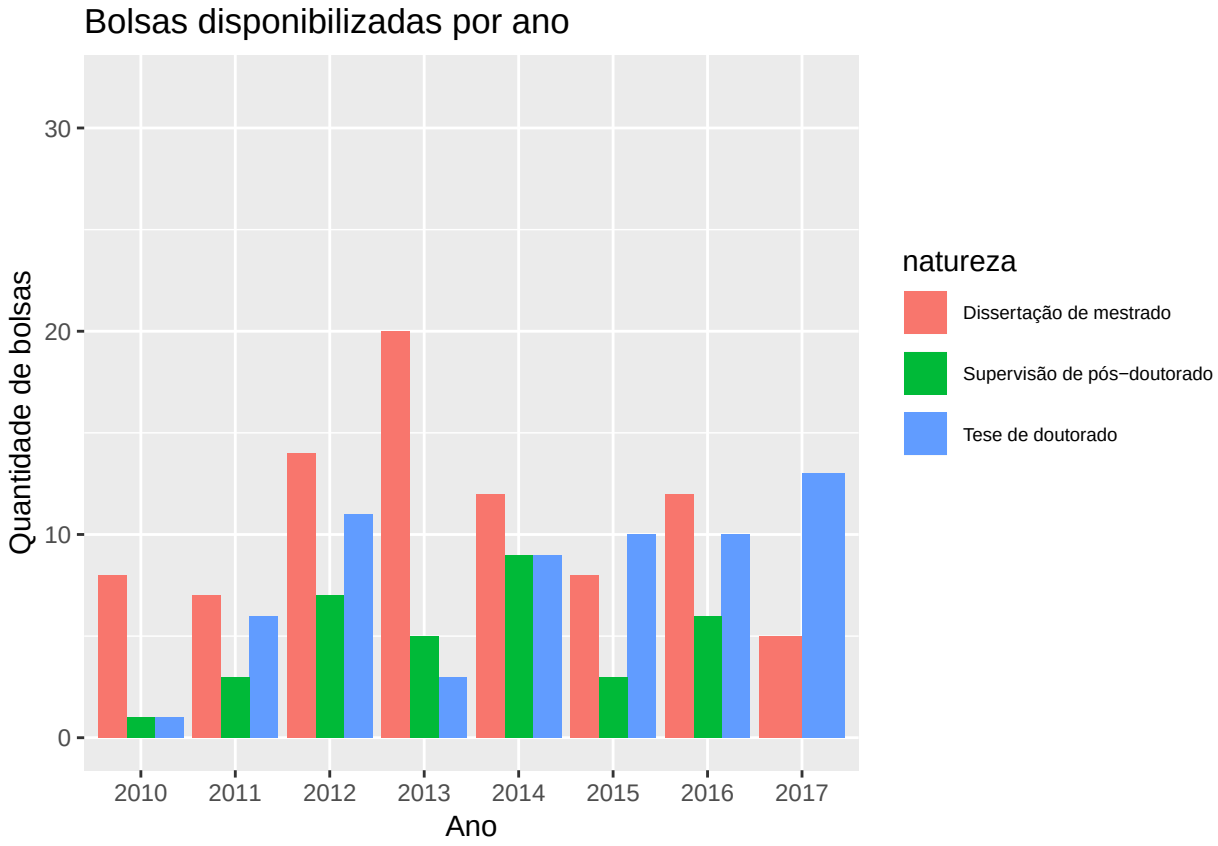
```
ggplot(orient.df,aes(ano,fill=natureza)) +
  geom_bar(stat = "count", position="dodge") +
  ggtitle("Natureza das Orientacoes Completas Por Ano") +
  theme(legend.position="right",legend.text=element_text(size=7)) +
  guides(fill=guide_legend(nrow=5, byrow=TRUE, title.position = "top")) +
  labs(x="Ano",y="Quantidade") + scale_y_continuous(limits = c(0, 32))
```



Observando a evolução do número de orientações completas ao longo dos anos, percebe-se que o Programa de Pós-Graduação cresceu consideravelmente na natureza de mestrado após 2010, com uma regressão em 2016. Entretanto, não há um comportamento linear na evolução do número de orientações finalizadas. 2014 despontou no número de dissertações de mestrado, mas este comportamento não se repetiu posteriormente. Enquanto as orientações de mestrado do Programa parecem bem estabelecidas, as supervisões de pós-doutorado ainda parecem incipientes, tendo em vista que não despontaram em nenhum dos anos pesquisados e chegaram a zero registros em 2017.

Bolsas distribuidas por ano

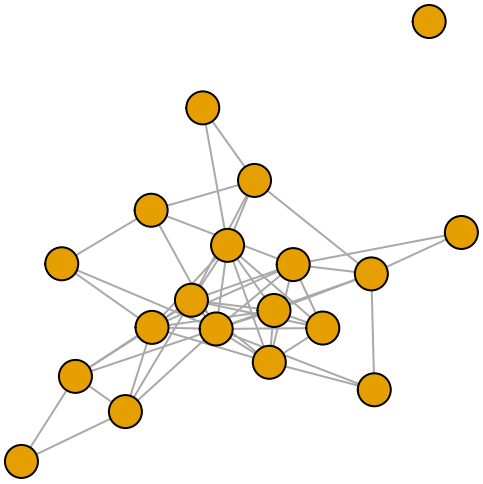
```
orient.df %>% filter(bolsa == "SIM") %>%
ggplot(aes(ano,fill=natureza)) +
  geom_bar(stat = "count", position = "dodge") +
  ggtitle("Bolsas disponibilizadas por ano") +
  theme(legend.position="right",legend.text=element_text(size=7)) +
  guides(fill=guide_legend(nrow=5, byrow=TRUE, title.position = "top")) +
  labs(x="Ano",y="Quantidade de bolsas") + scale_y_continuous(limits = c(0, 32))
```



Comparando os gráficos de orientações completas e de bolsas, é possível perceber que o número de bolsas oferecidas para o Programa acompanhou o total de orientações de maneira satisfatória ao longo dos anos. Durante todo o período, todas as naturezas apresentaram um índice de pelo menos 50% de bolsas, chegando a 100% em alguns casos. Além disso, as teses de pós-doutorado se mostram a natureza de pesquisa melhor contemplada pelas agências financiadoras, visto que apenas uma das observações não recebeu bolsa. Por fim, é possível observar que o pico no número de orientações de mestrado registrado em 2014 não foi tão expressivo em número de bolsas, caracterizando um ano com elevado número de alunos não bolsistas.

Grafo de proximidade entre pesquisadores do Programa de Pós-Graduação

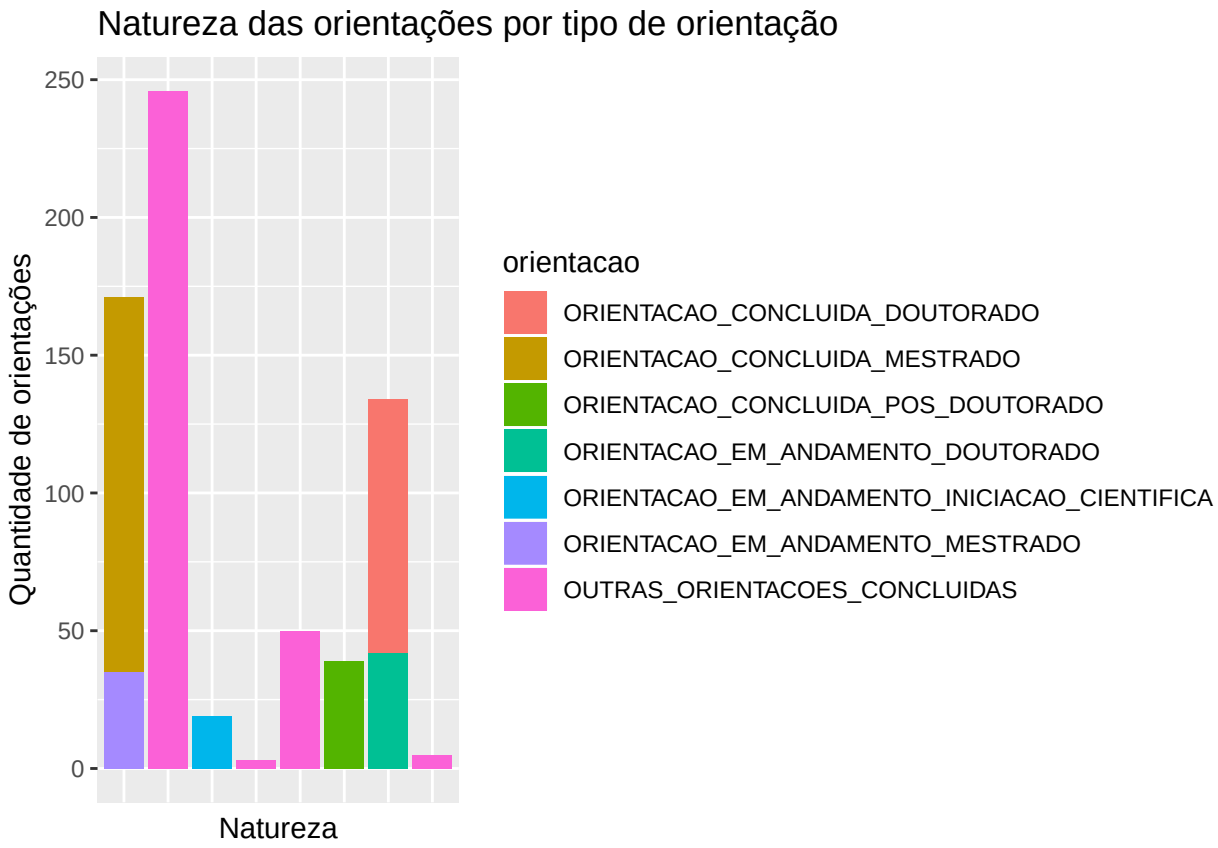
```
plot(g, vertex.label = NA)
```



O grafo acima representa os pesquisadores do Programa de Pós-Graduação em seus vértices e a existência de cooperação entre eles em suas arestas. É possível observar um grau profundo de cooperação entre a maioria dos pesquisadores, que, publicam trabalhos em grande cooperatividade, gerando um complexo emaranhado de arestas para cada vértice. Existem excessões tais como os vértices mais afastados e o vértice único que não tem nenhuma aresta com adjacentes.

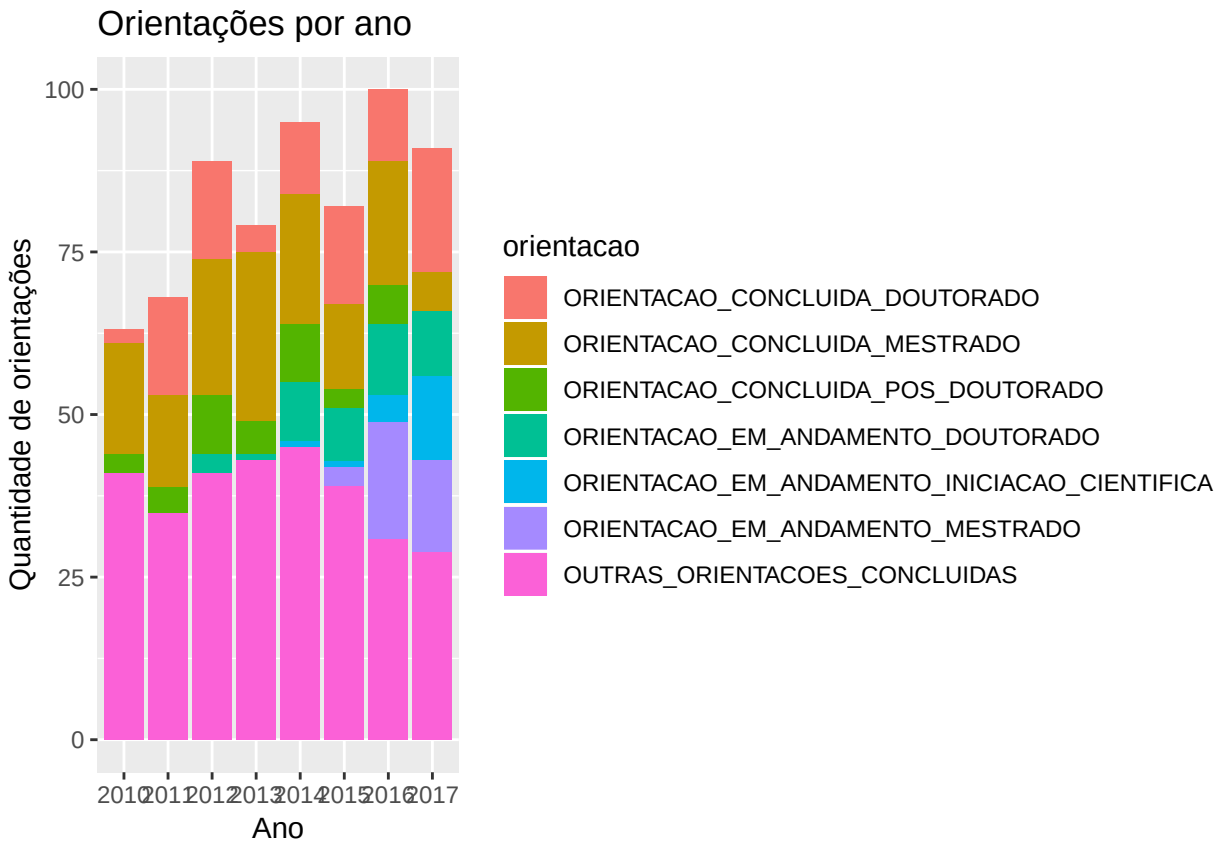
Natureza das orientações por tipo de orientação

```
ggplot(perfil.df.orientacoes, aes(natureza,fill=orientacao)) +
  geom_bar(stat = 'count') +
  theme(legend.position = 'right') +
  ggtitle('Natureza das orientações por tipo de orientação') +
  labs(x='Natureza',y='Quantidade de orientações') +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



Orientações por ano:

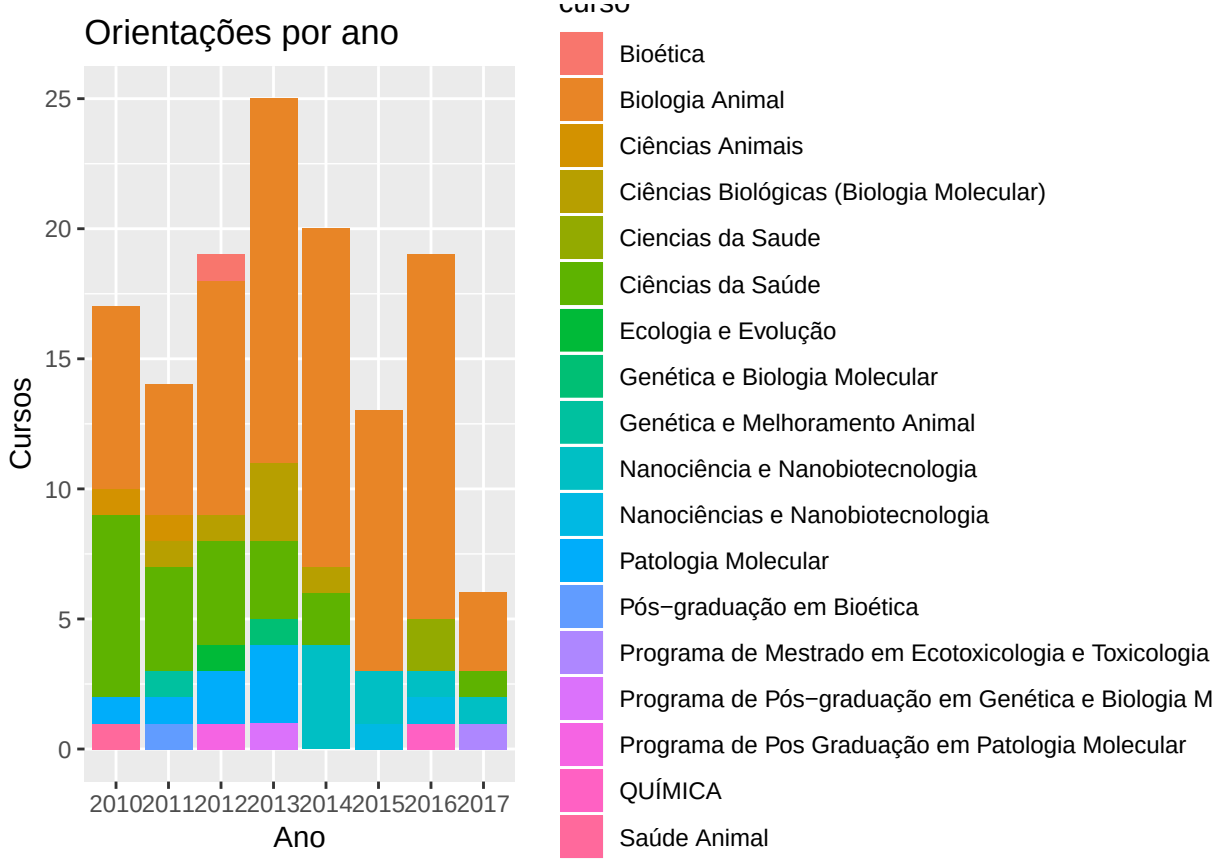
```
ggplot(perfil.df.orientacoes, aes(ano, fill=orientacao)) +
  geom_bar(stat = 'count') +
  ggtitle('Orientações por ano') +
  theme(legend.position = 'right') +
  labs(x='Ano',y='Quantidade de orientações')
```



A partir do número de orientações por ano e pela natureza destas, percebe-se que a quantidade de orientações de pós graduações no sentido restrito era pequena em proporção a outras orientações conduzidas por PPG por ano, até sua volta em 2015, onde o número de pós graduações no sentido restrito foram maiores que outras orientações.

Mestrados e cursos que mais ocorrem por ano

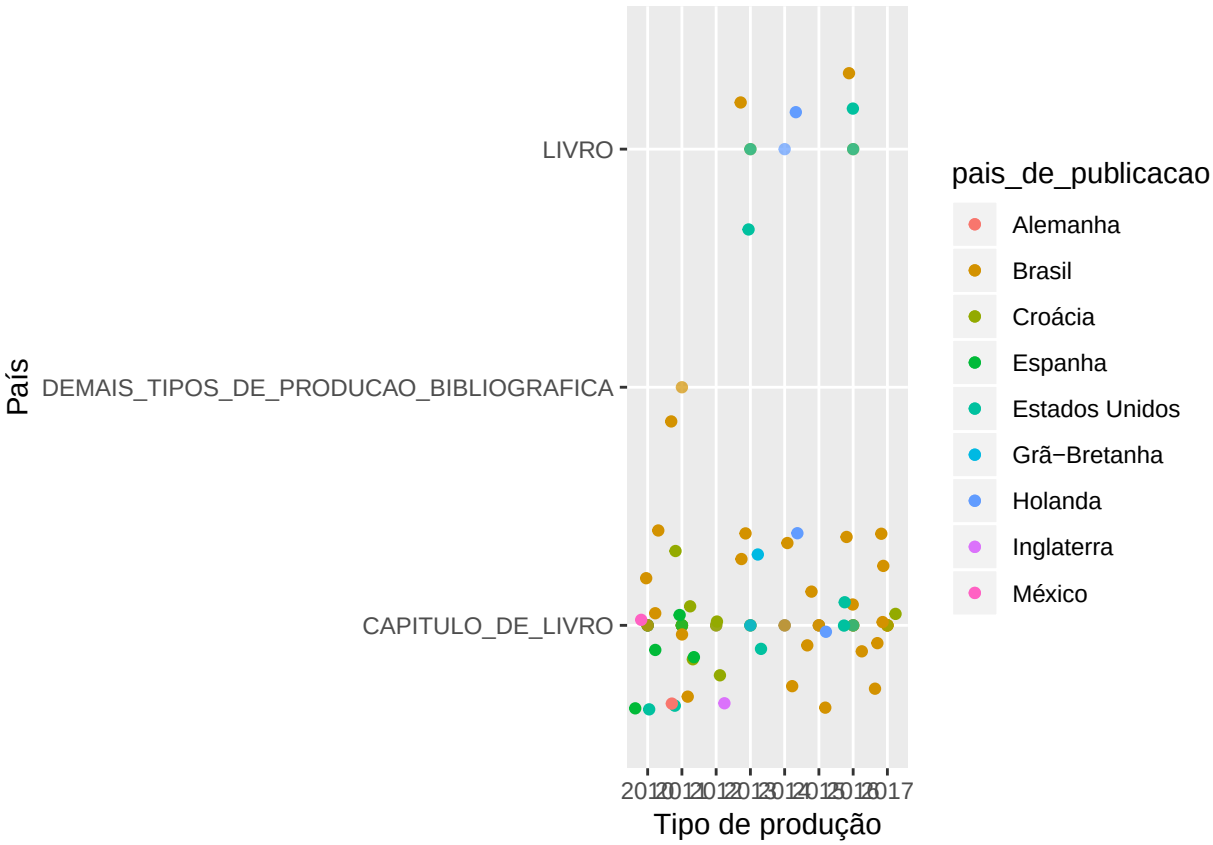
```
ggplot(orient.mestrado.df, aes(ano, fill=curso)) +
  geom_bar(stat = 'count') +
  ggtitle('Orientações por ano') +
  theme(legend.position = 'right') +
  labs(x='Ano',y='Cursos')
```



Dentre os mestrados nos programas de pós graduação estudados, pode se perceber o grande domínio em volume da pós graduação em Biologia Animal, que na maioria dos anos corresponde a quase metade das orientações de pós graduação por ano dentre os PPG estudados.

Publicações em países:

```
perfil.df.publicacoes %>%
  filter(!(tipo_producao %in% c('EVENTO','TEXTO_EM_JORNAIS','PERIODICO','ARTIGO_ACEITO'))) %>%
  group_by(tipo_producao,pais_de_publicacao) %>%
  ggplot(aes(ano,tipo_producao,col=pais_de_publicacao)) +
  geom_point(alpha = 0.7) + geom_jitter() +
  labs(x='Tipo de produção',y='País')
```



Observa-se deficiência de dados quanto ao país de publicação para periódicos, textos em jornais e artigos aceitos para os PPG analizados. Ainda assim, este gráfico demonstra bem a heterogeniedade das publicações de livros e/ou capítu- los no Brasil e demais países.

Bibliografia

- *SciVal Metrics Guidebook*. ELSEVIER, 2014.
- Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, e Rüdiger Wirth. "CRISP-DM 1.0: Step-by-Step Data Mining Guide". USA: CRISP-DM Consortium, 2000. https://www.the-modeling-agency.com/crisp-dm.pdf.
- Fernandes, Jorge H C, Ricardo Barros Sampaio, João Ribas de Moura e Jerônimo Avelar Filho. "Ciência de Dados para Todos (Data Science For All) - 2018.1 - Análise da Produção Científica e Acadêmica da Universidade de Brasília - Modelo de Relatório Final da Disciplina - Departamento de Ciência da Computação da UnB". Disciplina 116297 - Tópicos Avançados em Computadores, turma D, do semestre 2018.1, do Departamento de Ciência da Computação do Instituto de Ciências Exatas da Universidade de Brasília, 13 de junho de 2018.