

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

*Emanuel Victor A. Coelho 14/0137475, Jonas Prado 13/0117277, Maria Júlia G. R.
Rodrigues 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 (Fernandes et. al., 2018).

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 atribuírem 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* contribui 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; Fernandes et. al., 2018).

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 forma:

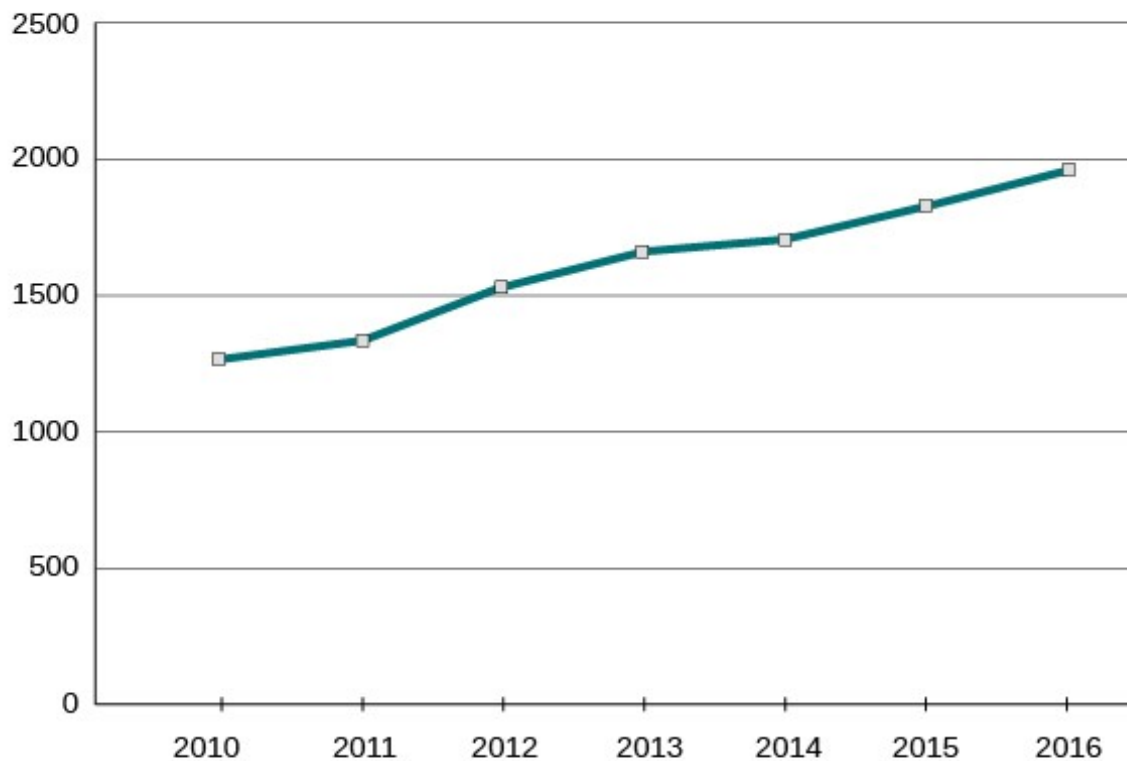
- Colégio de Ciências da Vida - que contempla as grandes áreas de Ciências Agrárias, Ciências Biológicas e Ciências da Saúde.
- Colégio de Ciências Exatas, Tecnológicas e Multidisciplinar - que contempla as grandes áreas de Ciências Exatas e da Terra, Engenharias e Multidisciplinar.
- Colégio de Humanidades - que contempla as grandes áreas de Ciências Humanas, Ciências Sociais Aplicadas e Linguística, 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) (Fernandes et. al., 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).

Produção científica da UnB de 2010 a 2016



Fonte: SciVal - 2018

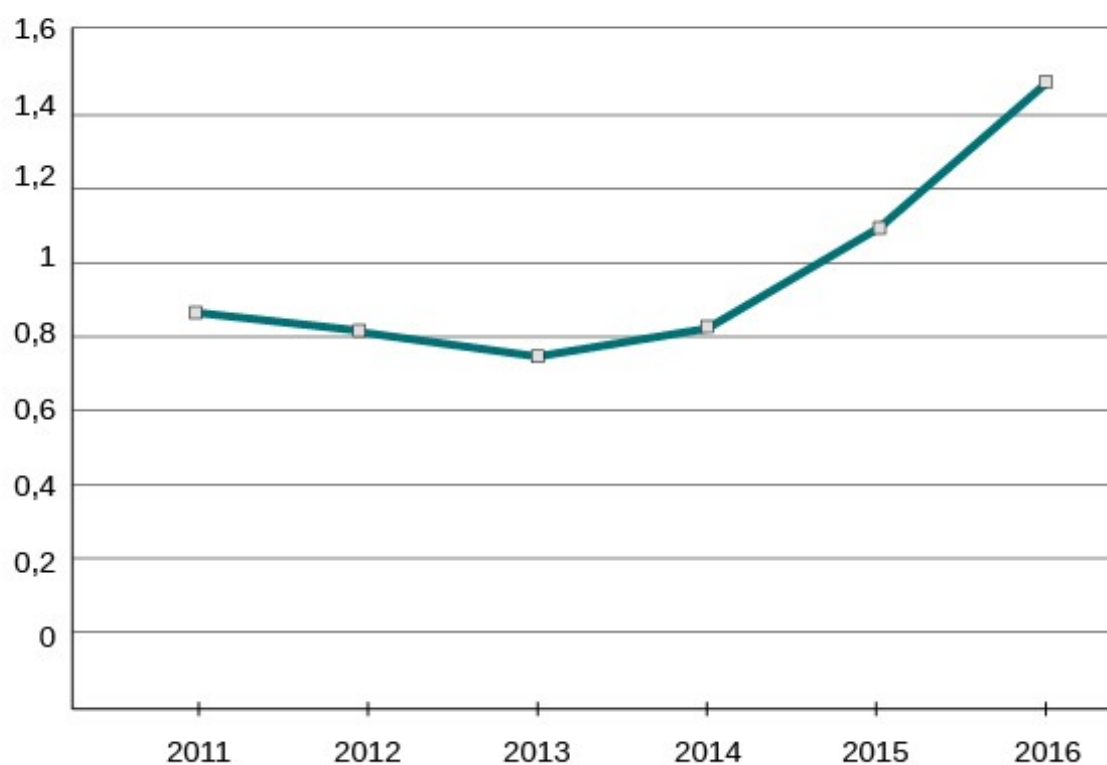
Figure 1: production

Um impacto normalizado de 1,00 significa que a produção teve comportamento similar à média global. Um impacto superior a 1,00 indica maior citação que a média (por exemplo, um impacto de 1,50 indica 50% a mais de citação) enquanto um impacto inferior a 1,00 indica citação inferior à média (SeCom, 2018, Adaptado).

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, houve uma redução branda no número de pesquisadores citados. Entretanto, de 2013 até 2016 houve 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.(SeCom, 2018).

Esse aumento na produção científica 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%) (SeCom, 2018). Com um quadro de professores maior espera-se naturalmente uma maior produção de artigos e outras matérias

Impacto normalizado por citação da produção científica da UnB de 2010 a 2016^[1]



Fonte: SciVal - 2018

^[1] Um impacto normalizado de 1,00 significa que a produção teve comportamento similar à média global. Um impacto superior a 1,00 indica maior citação que a média (por exemplo, um impacto de 1,50 indica 50% a mais de citação) enquanto um impacto inferior a 1,00 indica citação inferior à média (um impacto de 0,91 indica um 9% a menos de citação que a média).

Figure 2: impact

científicos pela UnB. (SeCom, 2018).

Captura dos dados: Definindo bancos de dados para as PPGs a partir de arquivos JSON e CSV

```
library(jsonlite) #Importado para carga dos arquivos JSON para o R
# Perfis
perfil.240 <- fromJSON("240BiologiaAnimal/240profile.json")
perfil.242 <- fromJSON("242BiologiaMicrobiana/242profile.json")
perfil.243 <- fromJSON("243BiologiaMolecular/243profile.json")
perfil.313 <- fromJSON("313PatologiaMolecular/313profile.json")

# Publicações
public.240 <- fromJSON("240BiologiaAnimal/240publication.json")
public.242 <- fromJSON("242BiologiaMicrobiana/242publication.json")
public.243 <- fromJSON("243BiologiaMolecular/243publication.json")
public.313 <- fromJSON("313PatologiaMolecular/313publication.json")

# Orientações
orient.240 <- fromJSON("240BiologiaAnimal/240advise.json")
orient.242 <- fromJSON("242BiologiaMicrobiana/242advise.json")
orient.243 <- fromJSON("243BiologiaMolecular/243advise.json")
orient.313 <- fromJSON("313PatologiaMolecular/313advise.json")

# Grafos
graphl.240 <- fromJSON("240BiologiaAnimal/240graph.json")
graphl.242 <- fromJSON("242BiologiaMicrobiana/242graph.json")
graphl.243 <- fromJSON("243BiologiaMolecular/243graph.json")
graphl.313 <- fromJSON("313PatologiaMolecular/313graph.json")

# Todas PPGs
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 a 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 colaborativas** 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
library(dplyr) #Importado para uso do Operador Pipe
library(tidyr) #Importado para uso da função spread()
library(ggplot2) #Importado para visualizações com ggplot()
library(stringr) #str_to_upper

setwd("~/Repository/DataScience/DS4A-BioAni-BioMic-BioMol-PatMol-master")
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 devem ser 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", "AreaPos"),
                               sep = ";", extra = "drop", fill = "right")
```

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

```
print(length(perfil.240))
```

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

```
ProfileList.240 <- list()
for (i in 1:length(perfil.240)) {
  ProfileList.240 <- rbind(ProfileList.240, perfil.240[[i]]$nome)
}

length(perfil.242)
```

```
## [1] 23
```

```
ProfileList.242 <- list()
for (i in 1:length(perfil.242)) {
  ProfileList.242 <- rbind(ProfileList.242, perfil.242[[i]]$nome)
}

length(perfil.243)
```

```
## [1] 36
```

```

ProfileList.243 <- list()
for (i in 1:length(perfil.243)) {
  ProfileList.243 <- rbind(ProfileList.243, perfil.243[[i]]$nome)
}

```

```
length(perfil.313)
```

```
## [1] 23
```

```

ProfileList.313 <- list()
for (i in 1:length(perfil.313)) {
  ProfileList.313 <- rbind(ProfileList.313, perfil.313[[i]]$nome)
}

```

Análise das listas:

Número de áreas de atuação cumulativas:

```
sum(sapply(perfil.240, function(x) nrow(x$areas_de_atuacao)))
```

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

```
sum(sapply(perfil.242, function(x) nrow(x$areas_de_atuacao)))
```

```
## [1] 95
```

```
sum(sapply(perfil.243, function(x) nrow(x$areas_de_atuacao)))
```

```
## [1] 149
```

```
sum(sapply(perfil.313, function(x) nrow(x$areas_de_atuacao)))
```

```
## [1] 96
```

Numero de areas de atuacao por pessoa

```
table(unlist(sapply(perfil.240, function(x) nrow(x$areas_de_atuacao))))
```

```
##
```

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

```
## 2 3 7 5 2
```

```
table(unlist(sapply(perfil.242, function(x) nrow(x$areas_de_atuacao))))
```

```
##
```

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

```
## 3 6 4 5 5
```

```
table(unlist(sapply(perfil.243, function(x) nrow(x$areas_de_atuacao))))
```

```
##
```

```
## 1 2 3 4 5 6
```

```
## 2 3 6 9 9 7
```

```
table(unlist(sapply(perfil.313, function(x) nrow(x$areas_de_atuacao))))
```

```
##
```

```
## 1 2 3 4 5 6
```

```
## 1 2 3 6 8 3
```

Numero de pessoas por grande area

```
table(unlist(sapply(perfil.240, 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
```

```
table(unlist(sapply(perfil.242, function(x) (x$areas_de_atuacao$grande_area))))
```

```
##
##          CIENCIAS_AGRARIAS          CIENCIAS_BIOLOGICAS
##                9                81
##          CIENCIAS_DA_SAUDE CIENCIAS_EXATAS_E_DA_TERRA
##                4                1
```

```
table(unlist(sapply(perfil.243, function(x) (x$areas_de_atuacao$grande_area))))
```

```
##
##          CIENCIAS_AGRARIAS          CIENCIAS_BIOLOGICAS
##                9                129
##          CIENCIAS_DA_SAUDE CIENCIAS_EXATAS_E_DA_TERRA
##                4                3
##          CIENCIAS_HUMANAS          ENGENHARIAS
##                3                1
```

```
table(unlist(sapply(perfil.313, function(x) (x$areas_de_atuacao$grande_area))))
```

```
##
##          CIENCIAS_AGRARIAS          CIENCIAS_BIOLOGICAS
##                2                77
##          CIENCIAS_DA_SAUDE CIENCIAS_EXATAS_E_DA_TERRA
##                11                4
##          CIENCIAS_HUMANAS          ENGENHARIAS
##                1                1
```

Numero de pessoas que produziram os especificos tipos de producao

```
table(unlist(sapply(perfil.240, function(x) names(x$producao_bibliografica))))
```

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



```
##
##
table(unlist(sapply(perfil.242, function(x) names(x$producao_bibliografica))))
```

```
##
##          ARTIGO_ACEITO
##                2
##          CAPITULO_DE_LIVRO
##                21
##  DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA
##                4
##                EVENTO
##                22
##                LIVRO
##                7
##                PERIODICO
##                23
##          TEXTO_EM_JORNAIS
##                3
```

```
table(unlist(sapply(perfil.243, function(x) names(x$producao_bibliografica))))
```

```
##
##          ARTIGO_ACEITO
##                7
##          CAPITULO_DE_LIVRO
##                32
##  DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA
##                4
##                EVENTO
##                30
##                LIVRO
##                12
##                PERIODICO
##                36
##          TEXTO_EM_JORNAIS
##                7
```

```
table(unlist(sapply(perfil.313, function(x) names(x$producao_bibliografica))))
```

```
##
##          ARTIGO_ACEITO
##                6
##          CAPITULO_DE_LIVRO
##                18
##  DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA
##                2
##                EVENTO
##                20
##                LIVRO
##                4
##                PERIODICO
##                23
##          TEXTO_EM_JORNAIS
##                5
```

Numero de publicacoes por tipo

```

sum(sapply(perfil.240, function(x) length(x$producao_bibliografica$ARTIGO_ACEITO$ano)))

## [1] 6
sum(sapply(perfil.240, function(x) length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano)))

## [1] 41
sum(sapply(perfil.240, function(x) length(x$producao_bibliografica$LIVRO$ano)))

## [1] 5
sum(sapply(perfil.240, function(x) length(x$producao_bibliografica$PERIODICO$ano)))

## [1] 681
sum(sapply(perfil.240, function(x) length(x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano)))

## [1] 4
sum(sapply(perfil.242, function(x) length(x$producao_bibliografica$ARTIGO_ACEITO$ano)))

## [1] 2
sum(sapply(perfil.242, function(x) length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano)))

## [1] 78
sum(sapply(perfil.242, function(x) length(x$producao_bibliografica$LIVRO$ano)))

## [1] 7
sum(sapply(perfil.242, function(x) length(x$producao_bibliografica$PERIODICO$ano)))

## [1] 381
sum(sapply(perfil.242, function(x) length(x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano)))

## [1] 10
sum(sapply(perfil.243, function(x) length(x$producao_bibliografica$ARTIGO_ACEITO$ano)))

## [1] 8
sum(sapply(perfil.243, function(x) length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano)))

## [1] 114
sum(sapply(perfil.243, function(x) length(x$producao_bibliografica$LIVRO$ano)))

## [1] 22
sum(sapply(perfil.243, function(x) length(x$producao_bibliografica$PERIODICO$ano)))

## [1] 1185
sum(sapply(perfil.243, function(x) length(x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano)))

## [1] 10
sum(sapply(perfil.313, function(x) length(x$producao_bibliografica$ARTIGO_ACEITO$ano)))

## [1] 10

```

```

sum(sapply(perfil.313, function(x) length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano)))

## [1] 57
sum(sapply(perfil.313, function(x) length(x$producao_bibliografica$LIVRO$ano)))

## [1] 8
sum(sapply(perfil.313, function(x) length(x$producao_bibliografica$PERIODICO$ano)))

## [1] 891
sum(sapply(perfil.313, function(x) length(x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano)))

## [1] 72

```

Número de pessoas por quantitativo de produções por pessoa

```

table(unlist(sapply(perfil.240, function(x) length(x$producao_bibliografica$ARTIGO_ACEITO$ano))))

##
## 0 2
## 16 3
table(unlist(sapply(perfil.240, function(x) length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano))))

##
## 0 1 2 3 4 5 9
## 5 2 7 1 2 1 1
table(unlist(sapply(perfil.240, function(x) length(x$producao_bibliografica$LIVRO$ano))))

##
## 0 1 2
## 15 3 1
table(unlist(sapply(perfil.240, function(x) length(x$producao_bibliografica$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.240, function(x) length(x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano))))

##
## 0 1
## 15 4
table(unlist(sapply(perfil.242, function(x) length(x$producao_bibliografica$ARTIGO_ACEITO$ano))))

##
## 0 1
## 21 2
table(unlist(sapply(perfil.242, function(x) length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano))))

##
## 0 1 2 3 4 5 11 13
## 2 8 1 4 4 1 2 1
table(unlist(sapply(perfil.242, function(x) length(x$producao_bibliografica$LIVRO$ano))))

```

```
##
## 0 1
## 16 7
table(unlist(sapply(perfil.242, function(x) length(x$producao_bibliografica$PERIODICO$ano))))

##
## 3 6 7 8 9 10 11 13 14 16 19 27 38 44 52
## 1 3 1 1 3 2 1 3 1 1 1 1 2 1 1
table(unlist(sapply(perfil.242, function(x) length(x$producao_bibliografica$TEXTO_EM_JORNAIS$ano))))

##
## 0 1 3 6
## 20 1 1 1
table(unlist(sapply(perfil.243, function(x) length(x$producao_bibliografica$ARTIGO_ACEITO$ano))))

##
## 0 1 2
## 29 6 1
table(unlist(sapply(perfil.243, function(x) length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano))))

##
## 0 1 2 3 4 5 7 8 11
## 4 10 4 6 5 2 1 1 3
table(unlist(sapply(perfil.243, function(x) length(x$producao_bibliografica$LIVRO$ano))))

##
## 0 1 2 3 4
## 24 7 1 3 1
table(unlist(sapply(perfil.243, function(x) length(x$producao_bibliografica$PERIODICO$ano))))

##
## 9 14 15 16 17 19 20 21 22 23 25 26 27 29 31 32 34 36
## 1 1 1 1 1 2 1 2 1 3 1 1 2 2 1 1 1 1
## 38 40 44 45 47 49 63 77 104
## 3 1 2 1 1 1 1 1 1
table(unlist(sapply(perfil.243, function(x) length(x$producao_bibliografica$TEXTO_EM_JORNAIS$ano))))

##
## 0 1 2 3
## 29 5 1 1
table(unlist(sapply(perfil.313, function(x) length(x$producao_bibliografica$ARTIGO_ACEITO$ano))))

##
## 0 1 2 4
## 17 4 1 1
table(unlist(sapply(perfil.313, function(x) length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano))))

##
## 0 1 2 3 4 5 18
## 5 7 2 5 2 1 1
```

```

table(unlist(sapply(perfil.313, function(x) length(x$producao_bibliografica$LIVRO$ano))))

##
## 0 1 2 4
## 19 2 1 1

table(unlist(sapply(perfil.313, function(x) length(x$producao_bibliografica$PERIODICO$ano))))

##
## 9 14 16 18 19 20 21 22 23 25 26 29 31 34 44 45 49 51
## 1 1 2 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1
## 88 233
## 1 1

table(unlist(sapply(perfil.313, function(x) length(x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano))))

##
## 0 1 2 67
## 18 3 1 1

Número de produções por ano

table(unlist(sapply(perfil.240, function(x) (x$producao_bibliografica$ARTIGO_ACEITO$ano))))

##
## 2016 2017
## 1 5

table(unlist(sapply(perfil.240, function(x) (x$producao_bibliografica$CAPITULO_DE_LIVRO$ano))))

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

table(unlist(sapply(perfil.240, function(x) (x$producao_bibliografica$LIVRO$ano))))

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

table(unlist(sapply(perfil.240, function(x) (x$producao_bibliografica$PERIODICO$ano))))

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

table(unlist(sapply(perfil.240, function(x) (x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano))))

##
## 2012 2013 2014 2016
## 1 1 1 1

table(unlist(sapply(perfil.242, function(x) (x$producao_bibliografica$ARTIGO_ACEITO$ano))))

##
## 2013 2016
## 1 1

table(unlist(sapply(perfil.242, function(x) (x$producao_bibliografica$CAPITULO_DE_LIVRO$ano))))

##

```

```

## 2010 2011 2012 2013 2014 2015 2016 2017
##      2      5      8     10     27     10     11      5
table(unlist(sapply(perfil.242, function(x) (x$producao_bibliografica$LIVRO$ano))))

##
## 2010 2012 2013 2015 2016
##      1      1      2      1      2
table(unlist(sapply(perfil.242, function(x) (x$producao_bibliografica$PERIODICO$ano))))

##
## 2010 2011 2012 2013 2014 2015 2016 2017
##     37     41     41     45     63     47     54     53
table(unlist(sapply(perfil.242, function(x) (x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano))))

##
## 2010 2011 2012 2013 2015
##      2      1      3      2      2
table(unlist(sapply(perfil.243, function(x) (x$producao_bibliografica$ARTIGO_ACEITO$ano))))

##
## 2013 2016 2017
##      1      2      5
table(unlist(sapply(perfil.243, function(x) (x$producao_bibliografica$CAPITULO_DE_LIVRO$ano))))

##
## 2010 2011 2012 2013 2014 2015 2016 2017
##      8     10     11     10     28     15     17     15
table(unlist(sapply(perfil.243, function(x) (x$producao_bibliografica$LIVRO$ano))))

##
## 2010 2011 2012 2013 2014 2015 2016 2017
##      3      3      1      2      2      4      5      2
table(unlist(sapply(perfil.243, function(x) (x$producao_bibliografica$PERIODICO$ano))))

##
## 2010 2011 2012 2013 2014 2015 2016 2017
##    140    165    126    160    139    159    158    138
table(unlist(sapply(perfil.243, function(x) (x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano))))

##
## 2010 2011 2012 2014 2015 2016
##      2      2      2      1      2      1
table(unlist(sapply(perfil.313, function(x) length(x$producao_bibliografica$ARTIGO_ACEITO$ano))))

##
##      0      1      2      4
##    17      4      1      1
table(unlist(sapply(perfil.313, function(x) length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano))))

##
##      0      1      2      3      4      5    18

```

```
## 5 7 2 5 2 1 1
table(unlist(sapply(perfil.313, function(x) length(x$producao_bibliografica$LIVRO$ano))))

##
## 0 1 2 4
## 19 2 1 1
table(unlist(sapply(perfil.313, function(x) length(x$producao_bibliografica$PERIODICO$ano))))

##
## 9 14 16 18 19 20 21 22 23 25 26 29 31 34 44 45 49 51
## 1 1 2 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1
## 88 233
## 1 1
table(unlist(sapply(perfil.313, function(x) length(x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano))))

##
## 0 1 2 67
## 18 3 1 1
```

Número de pessoas que realizaram diferentes tipos de orientações

```
length(unlist(sapply(perfil.240, function(x) names(x$orientacoes_academicas))))

## [1] 102
length(unlist(sapply(perfil.242, function(x) names(x$orientacoes_academicas))))

## [1] 108
length(unlist(sapply(perfil.243, function(x) names(x$orientacoes_academicas))))

## [1] 204
length(unlist(sapply(perfil.313, function(x) names(x$orientacoes_academicas))))

## [1] 133
```

Número de pessoas por tipo de orientação

```
table(unlist(sapply(perfil.240, 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
```

```
table(unlist(sapply(perfil.242, function(x) names(x$orientacoes_academicas))))
```

```
##
##          ORIENTACAO_CONCLUIDA_DOUTORADO
##                               13
##          ORIENTACAO_CONCLUIDA_MESTRADO
##                               21
##          ORIENTACAO_CONCLUIDA_POS_DOUTORADO
##                               6
##          ORIENTACAO_EM_ANDAMENTO_DOUTORADO
##                               18
##          ORIENTACAO_EM_ANDAMENTO_GRADUACAO
##                               1
## ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA
##                               15
##          ORIENTACAO_EM_ANDAMENTO_MESTRADO
##                               14
##          OUTRAS_ORIENTACOES_CONCLUIDAS
##                               20
```

```
table(unlist(sapply(perfil.243, function(x) names(x$orientacoes_academicas))))
```

```
##
##          ORIENTACAO_CONCLUIDA_DOUTORADO
##                               33
##          ORIENTACAO_CONCLUIDA_MESTRADO
##                               36
##          ORIENTACAO_CONCLUIDA_POS_DOUTORADO
##                               24
##          ORIENTACAO_EM_ANDAMENTO_DOUTORADO
##                               35
## ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA
##                               18
##          ORIENTACAO_EM_ANDAMENTO_MESTRADO
##                               24
##          OUTRAS_ORIENTACOES_CONCLUIDAS
##                               34
```

```
table(unlist(sapply(perfil.313, function(x) names(x$orientacoes_academicas))))
```

```
##
##          ORIENTACAO_CONCLUIDA_DOUTORADO
##                               22
##          ORIENTACAO_CONCLUIDA_MESTRADO
##                               23
##          ORIENTACAO_CONCLUIDA_POS_DOUTORADO
##                               15
##          ORIENTACAO_EM_ANDAMENTO_DOUTORADO
##                               22
## ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA
##                               13
##          ORIENTACAO_EM_ANDAMENTO_MESTRADO
##                               15
##          OUTRAS_ORIENTACOES_CONCLUIDAS
##                               23
```


Número de orientações concluídas

```
sum(sapply(perfil.240, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano)))

## [1] 136
sum(sapply(perfil.240, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano)))

## [1] 92
sum(sapply(perfil.240, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano)))

## [1] 39
sum(sapply(perfil.242, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano)))

## [1] 130
sum(sapply(perfil.242, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano)))

## [1] 45
sum(sapply(perfil.242, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano)))

## [1] 27
sum(sapply(perfil.243, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano)))

## [1] 275
sum(sapply(perfil.243, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano)))

## [1] 199
sum(sapply(perfil.243, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano)))

## [1] 107
sum(sapply(perfil.313, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano)))

## [1] 242
sum(sapply(perfil.313, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano)))

## [1] 152
sum(sapply(perfil.313, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano)))

## [1] 79
```

Número de pessoas por quantitativo de orientações por pessoa

```
table(unlist(sapply(perfil.240, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano))))

##
##  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.240, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano))))

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

```

table(unlist(sapply(perfil.240, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DO
##
## 0 2 3 5 6 7 14
## 12 2 1 1 1 1 1
table(unlist(sapply(perfil.242, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRA
##
## 0 1 2 3 4 5 6 7 8 9 10 11
## 2 1 3 1 2 2 2 2 2 1 4 1
table(unlist(sapply(perfil.242, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTOR
##
## 0 1 2 3 5 6 7 8
## 10 4 3 1 1 2 1 1
table(unlist(sapply(perfil.242, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DO
##
## 0 1 2 4 17
## 17 2 2 1 1
table(unlist(sapply(perfil.243, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRA
##
## 1 3 4 5 6 7 8 9 10 11 12 14 17
## 2 2 4 2 4 5 2 3 5 2 3 1 1
table(unlist(sapply(perfil.243, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTOR
##
## 0 2 3 4 5 6 7 8 11 17 21
## 3 5 6 2 3 4 5 5 1 1 1
table(unlist(sapply(perfil.243, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DO
##
## 0 1 2 3 4 5 6 7 8 21
## 12 3 6 2 4 3 2 2 1 1
table(unlist(sapply(perfil.313, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRA
##
## 4 5 6 7 8 9 10 11 12 14 27 41
## 3 2 3 1 1 3 3 2 1 2 1 1
table(unlist(sapply(perfil.313, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTOR
##
## 0 2 3 4 5 6 7 8 11 14 22
## 1 3 4 2 3 2 2 2 1 1 2
table(unlist(sapply(perfil.313, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DO
##
## 0 1 2 3 4 5 7 8 14
## 8 1 3 1 4 2 1 1 2

```

Número de orientações por ano

```

table(unlist(sapply(perfil.240, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano),
##
## 2010 2011 2012 2013 2014 2015 2016 2017
## 17 14 21 26 20 13 19 6
table(unlist(sapply(perfil.240, function(x) (x$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.240, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano),
##
## 2010 2011 2012 2013 2014 2015 2016
## 3 4 9 5 9 3 6
table(unlist(sapply(perfil.242, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano),
##
## 2010 2011 2012 2013 2014 2015 2016 2017
## 9 16 17 17 31 18 9 13
table(unlist(sapply(perfil.242, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano),
##
## 2010 2012 2013 2014 2015 2016 2017
## 4 8 13 5 2 9 4
table(unlist(sapply(perfil.242, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano),
##
## 2010 2011 2013 2014 2015 2016
## 11 1 3 6 2 4
table(unlist(sapply(perfil.243, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano),
##
## 2010 2011 2012 2013 2014 2015 2016 2017
## 32 38 39 45 46 28 36 11
table(unlist(sapply(perfil.243, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano),
##
## 2010 2011 2012 2013 2014 2015 2016 2017
## 24 17 33 27 26 22 26 24
table(unlist(sapply(perfil.243, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano),
##
## 2010 2011 2012 2013 2014 2015 2016 2017
## 11 8 13 18 18 12 17 10
table(unlist(sapply(perfil.313, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano),
##
## 2010 2011 2012 2013 2014 2015 2016 2017
## 33 27 34 35 42 24 20 27

```

```

table(unlist(sapply(perfil.313, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano)),
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##   14   14   30   22   20   15   21   16
table(unlist(sapply(perfil.313, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano)),
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##   11    8    9   13   12    8   12    6

```

Análise dos dataframes:

Extrai perfis dos professores:

```

perfil.240.df.professores <- extrai.perfis(perfil.240)

perfil.242.df.professores <- extrai.perfis(perfil.242)

perfil.243.df.professores <- extrai.perfis(perfil.243)

perfil.313.df.professores <- extrai.perfis(perfil.313)

```

Extrai produção bibliográfica de todos os professores

```

perfil.240.df.publicacoes <- extrai.producoes(perfil.240) %>%
  select(tipo_producao, everything()) %>% arrange(tipo_producao)
perfil.240.df.publicacoes$tipo_producao[grepl("DEMAIS", perfil.240.df.publicacoes$tipo_producao)] <- "OUTROS"

perfil.242.df.publicacoes <- extrai.producoes(perfil.242) %>%
  select(tipo_producao, everything()) %>% arrange(tipo_producao)

perfil.243.df.publicacoes <- extrai.producoes(perfil.243) %>%
  select(tipo_producao, everything()) %>% arrange(tipo_producao)

perfil.313.df.publicacoes <- extrai.producoes(perfil.313) %>%
  select(tipo_producao, everything()) %>% arrange(tipo_producao)

```

Extrai orientações

```

perfil.240.df.orientacoes <- extrai.orientacoes(perfil.240) %>%
  select(id_lattes_orientadores, natureza, ano, orientacao, everything()) %>%
  mutate(situacao = ifelse(grepl("CONCLUIDA", orientacao), "Concluída", "Em andamento")) %>%
  mutate(Natureza = case_when(grepl("MESTRADO", str_to_upper(natureza)) ~ "Mestrado",
                              grepl("PÓS-DOUTORADO", str_to_upper(natureza)) ~ "Pós-doutorado",
                              grepl("DOUTORADO", str_to_upper(natureza)) ~ "Doutorado",
                              grepl("INICIACAO", str_to_upper(natureza)) ~ "Iniciação Científica",
                              grepl("INICIAÇÃO", str_to_upper(natureza)) ~ "Iniciação Científica",
                              TRUE ~ "Outras naturezas"))

perfil.242.df.orientacoes <- extrai.orientacoes(perfil.242) %>%
  select(id_lattes_orientadores, natureza, ano, orientacao, everything())

perfil.243.df.orientacoes <- extrai.orientacoes(perfil.243) %>%
  select(id_lattes_orientadores, natureza, ano, orientacao, everything())

```

```

perfil.313.df.orientacoes <- extrai.orientacoes(perfil.313) %>%
  select(id_lattes_orientadores, natureza, ano, orientacao, everything())

```

Extrai áreas de atuação

```

perfil.240.df.areas.de.atuacao <- extrai.areas.atuacao(perfil.240) %>%
  select(idLattes, everything())

perfil.242.df.areas.de.atuacao <- extrai.areas.atuacao(perfil.242) %>%
  select(idLattes, everything())

perfil.243.df.areas.de.atuacao <- extrai.areas.atuacao(perfil.243) %>%
  select(idLattes, everything())

perfil.313.df.areas.de.atuacao <- extrai.areas.atuacao(perfil.313) %>%
  select(idLattes, everything())

```

Cria arquivo com dados quantitativos para análise

```

perfil.240.df <- data.frame()
perfil.240.df <- perfil.240.df.professores %>%
  select(idLattes, nome, resumo_cv, senioridade) %>%
  left_join(
    perfil.240.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.240.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.240.df.areas.de.atuacao %>%
      select(area, idLattes) %>%
      group_by(idLattes) %>%
      summarise(num_areas = n_distinct(area)),
    by = "idLattes") %>%
  mutate(num_areas = factor(num_areas))

perfil.240.df[is.na(perfil.240.df)] <- 0 #Remove NAs do dataframe

perfil.240.df.areas.de.atuacao <- perfil.240.df.areas.de.atuacao %>%
  left_join(perfil.240.df, by = "idLattes") %>%
  rowwise() %>% #realizar sum() corretamente
  mutate(orientacoes_concluidas = sum(ORIENTACAO_CONCLUIDA_DOUTORADO,
                                       ORIENTACAO_CONCLUIDA_POS_DOUTORADO, ORIENTACAO_CONCLUIDA_MESTRADO,
                                       OUTRAS_ORIENTACOES_CONCLUIDAS, na.rm = TRUE)) %>%

```

```
mutate(publicacoes = sum(CAPITULO_DE_LIVRO, EVENTO, PERIODICO,
  LIVRO, TEXTO_EM_JORNAIS, OUTRAS_PRODUCOES, na.rm = TRUE)) %>%
  select(idLattes, grande_area, area, sub_area, especialidade, orientacoes_concluidas, publicacoes)
class(perfil.240.df.areas.de.atuacao) <- c("tbl_df", "data.frame") #desfazer rowwise

glimpse(perfil.240.df)
```

```
## Observations: 19
## Variables: 15
## $ idLattes          <chr> "1612292306950738", "186941...
## $ nome              <chr> "Renato Caparroz", "Monica ...
## $ resumo_cv         <chr> "Formado em Zootecnia pela ...
## $ senioridade      <chr> "9", "8", "9", "9", "9", "9...
## $ ORIENTACAO_CONCLUIDA_DOUTORADO <dbl> 6, 1, 6, 7, 0, 1, 1, 2, 5, ...
## $ ORIENTACAO_CONCLUIDA_MESTRADO <int> 8, 6, 10, 1, 9, 5, 2, 5, 13...
## $ ORIENTACAO_CONCLUIDA_POS_DOUTORADO <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 2, ...
## $ OUTRAS_ORIENTACOES_CONCLUIDAS <dbl> 18, 9, 17, 0, 18, 9, 2, 6, ...
## $ CAPITULO_DE_LIVRO <dbl> 2, 0, 4, 9, 1, 4, 2, 0, 0, ...
## $ EVENTO           <dbl> 19, 14, 12, 0, 11, 10, 4, 3...
## $ LIVRO             <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, ...
## $ OUTRAS_PRODUCOES <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ PERIODICO         <int> 13, 10, 21, 68, 18, 44, 26,...
## $ TEXTO_EM_JORNAIS <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 1, ...
## $ num_areas        <fct> 2, 3, 1, 1, 4, 2, 3, 1, 3, ...
```

```
perfil.242.df <- data.frame()
perfil.242.df <- perfil.242.df.professores %>%
  select(idLattes, nome, resumo_cv, senioridade) %>%
  left_join(
    perfil.242.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.242.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.242.df.areas.de.atuacao %>%
      select(area, idLattes) %>%
      group_by(idLattes) %>%
      summarise(n_distinct(area)),
    by = "idLattes")

glimpse(perfil.242.df)
```

```
## Observations: 23
```

```
## Variables: 15
## $ idLattes <chr> "0098666202802214", "10...
## $ nome <chr> "Napoleão Fonseca Valad...
## $ resumo_cv <chr> "Possui graduação em Fa...
## $ senioridade <chr> "9", "9", "7", "8", "9"...
## $ ORIENTACAO_CONCLUIDA_DOUTORADO <int> NA, 2, 1, NA, 5, 3, 8, ...
## $ ORIENTACAO_CONCLUIDA_MESTRADO <int> NA, 10, 7, 3, 6, 9, 8, ...
## $ ORIENTACAO_CONCLUIDA_POS_DOUTORADO <int> NA, NA, NA, NA, 2, 2, 4...
## $ OUTRAS_ORIENTACOES_CONCLUIDAS <int> NA, 16, 14, 26, NA, 8, ...
## $ CAPITULO_DE_LIVRO <int> 3, 4, 1, 4, 5, 3, 1, 4,...
## $ DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA <int> NA, NA, NA, 2, NA, NA, ...
## $ EVENTO <int> 9, 25, 58, 29, 54, 3, 4...
## $ LIVRO <int> NA, NA, 1, 1, 1, NA, NA...
## $ PERIODICO <int> 13, 14, 10, 3, 19, 16, ...
## $ TEXTO_EM_JORNAIS <int> NA, NA, NA, NA, NA, 1, ...
## $ `n_distinct(area)` <int> 4, 1, 2, 1, 3, 3, 3, 2,...
```

```
perfil.243.df <- data.frame()
perfil.243.df <- perfil.243.df.professores %>%
  select(idLattes, nome, resumo_cv, senioridade) %>%
  left_join(
    perfil.243.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.243.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.243.df.areas.de.atuacao %>%
      select(area, idLattes) %>%
      group_by(idLattes) %>%
      summarise(n_distinct(area)),
    by = "idLattes")
glimpse(perfil.243.df)
```

```
## Observations: 36
## Variables: 15
## $ idLattes <chr> "0576885615997048", "05...
## $ nome <chr> "Dario Grattapaglia", "...
## $ resumo_cv <chr> "Engenheiro florestal p...
## $ senioridade <chr> "9", "9", "9", "9", "9"...
## $ ORIENTACAO_CONCLUIDA_DOUTORADO <int> 7, 5, 6, 2, 2, 7, 2, 6,...
## $ ORIENTACAO_CONCLUIDA_MESTRADO <int> 4, 9, 3, 1, 10, 14, 12,...
## $ ORIENTACAO_CONCLUIDA_POS_DOUTORADO <int> 2, 1, 2, NA, NA, 8, NA,...
## $ OUTRAS_ORIENTACOES_CONCLUIDAS <int> NA, 6, 3, 20, 16, 3, 15...
## $ CAPITULO_DE_LIVRO <int> 2, 4, 7, 1, 4, 3, 8, 1,...
```

```
## $ DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA <int> 2, NA, NA, NA, NA, 2, N...
## $ EVENTO <int> 55, 16, 37, 8, 25, 24, ...
## $ LIVRO <int> 3, 1, NA, 3, NA, NA, 1,...
## $ PERIODICO <int> 63, 20, 32, 38, 14, 49,...
## $ TEXTO_EM_JORNAIS <int> NA, NA, NA, NA, NA, 1, ...
## $ `n_distinct(area)` <int> 2, 2, 4, 2, 1, 3, 3, 3,...
```

```
perfil.313.df <- data.frame()
perfil.313.df <- perfil.313.df.professores %>%
  select(idLattes, nome, resumo_cv, senioridade) %>%
  left_join(
    perfil.313.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.313.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.313.df.areas.de.atuacao %>%
      select(area, idLattes) %>%
      group_by(idLattes) %>%
      summarise(n_distinct(area)),
    by = "idLattes")

glimpse(perfil.313.df)
```

```
## Observations: 23
## Variables: 15
## $ idLattes <chr> "0080556850353422", "05...
## $ nome <chr> "Jaime Martins de Santa...
## $ resumo_cv <chr> "Jaime Martins Santana ...
## $ senioridade <chr> "9", "9", "9", "9", "9"...
## $ ORIENTACAO_CONCLUIDA_DOUTORADO <int> 6, 5, 7, 6, 3, 7, 3, 3,...
## $ ORIENTACAO_CONCLUIDA_MESTRADO <int> 6, 9, 14, 10, 4, 12, 9,...
## $ ORIENTACAO_CONCLUIDA_POS_DOUTORADO <int> NA, 1, 8, 5, 2, 5, 2, 4...
## $ OUTRAS_ORIENTACOES_CONCLUIDAS <int> 8, 6, 3, 5, 16, 5, 8, 9...
## $ CAPITULO_DE_LIVRO <int> NA, 4, 3, 1, 1, 3, 3, 4...
## $ DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA <int> NA, NA, 2, NA, NA, NA, ...
## $ EVENTO <int> NA, 16, 24, 14, 40, 13,...
## $ LIVRO <int> NA, 1, NA, NA, NA, 1, N...
## $ PERIODICO <int> 29, 20, 49, 29, 21, 22,...
## $ TEXTO_EM_JORNAIS <int> NA, NA, 1, NA, NA, NA, ...
## $ `n_distinct(area)` <int> 2, 2, 3, 3, 3, 3, 3, 4,...
```


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.240$PERIODICO, function(x) length(x$natureza)))
```

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

```
sum(sapply(public.242$PERIODICO, function(x) length(x$natureza)))
```

```
## [1] 335
```

```
sum(sapply(public.243$PERIODICO, function(x) length(x$natureza)))
```

```
## [1] 977
```

```
sum(sapply(public.313$PERIODICO, function(x) length(x$natureza)))
```

```
## [1] 736
```

Anos analisados

```
names(public.240$PERIODICO)
```

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

```
names(public.242$PERIODICO)
```

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

```
names(public.243$PERIODICO)
```

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

```
names(public.313$PERIODICO)
```

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

20 revistas mais publicadas

```
head(sort(table(as.data.frame(unlist
  (sapply(public.240$PERIODICO, function(x) unlist(x$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
```

```
##
##          Journal of Nanoscience and Nanotechnology (Print)
##
##          4
```

```
head(sort(table(as.data.frame(unlist
(sapply(public.242$PERIODICO, function(x) unlist(x$periodico)))
)), decreasing = TRUE),20)
```

```
##
##          Archives of Virology
##          13
##          Plant Disease
##          9
##          Plos One
##          8
##          Virus Genes
##          8
##          Microbial Ecology
##          6
##          BMC Genomics
##          5
##          Mycologia
##          5
##          Mycological Progress
##          5
##          Mycotaxon
##          5
##          Biogeochemistry (Dordrecht)
##          4
##          Frontiers in Microbiology (Online)
##          4
##          Fungal Genetics and Biology (Print)
##          4
##          Global Change Biology (Print)
##          4
##          Microbiologia in Foco
##          4
##          Applied Microbiology and Biotechnology
##          3
##          BIOENERG RES
##          3
##          Biotechnology for Biofuels
##          3
##          Brazilian Journal of Microbiology (Impresso)
##          3
##          Genetics and Molecular Research
##          3
##          Journal of Industrial Microbiology & Biotechnology
##          3
```

```
head(sort(table(as.data.frame(unlist
(sapply(public.243$PERIODICO, function(x) unlist(x$periodico)))
)), decreasing = TRUE),20)
```

```
##
```

```

##                               Plos One
##                               49
##                               BMC Genomics
##                               20
##                               BMC Proceedings
##                               20
##                               Archives of Virology
##                               19
##                               Toxicon (Oxford)
##                               12
##                               Frontiers in Microbiology (Online)
##                               11
##                               Genetics and Molecular Research
##                               10
## PLoS Neglected Tropical Diseases (Online)
##                               10
##                               Protein and Peptide Letters
##                               10
##                               Scientific Reports
##                               10
##                               Virus Genes
##                               10
##                               Plant Disease
##                               8
##                               Tree Genetics & Genomes (Print)
##                               8
##                               BMC Plant Biology (Online)
##                               7
##                               Journal of Proteome Research (Print)
##                               7
##                               Proteomics (Weinheim. Print)
##                               7
##                               Acta Horticulturae
##                               6
## Applied Microbiology and Biotechnology
##                               6
##                               Journal of Biomedical Nanotechnology
##                               6
##                               Microbial Ecology
##                               6

```

```

head(sort(table(as.data.frame(unlist
  (sapply(public.313$PERIODICO, function(x) unlist(x$periodico)))
)), decreasing = TRUE),20)

```

```

##
##                               Plos One
##                               44
##                               Frontiers in Microbiology (Online)
##                               22
##                               Peptides (New York, N.Y. 1980)
##                               16
##                               PLoS Neglected Tropical Diseases (Online)
##                               12
##                               BMC Genomics

```

```

##                                     11
##                               Scientific Reports
##                                     11
##                               Protein and Peptide Letters
##                                     10
##                               Biopolymers (New York. Print)
##                                     9
##                               Journal of Proteome Research (Print)
##                                     9
##                               Journal of Biomedical Nanotechnology
##                                     8
##                               Antimicrobial Agents and Chemotherapy (Print)
##                                     7
##                               Biochimica et Biophysica Acta. Biomembranes
##                                     6
##                               BMC Microbiology (Online)
##                                     6
##                               Current Protein and Peptide Science
##                                     6
##                               Experimental Parasitology
##                                     6
##                               Fungal Genetics and Biology (Print)
##                                     6
##                               Journal of Proteomics
##                                     6
##                               Medical Mycology (Oxford. Print)
##                                     6
##                               International Journal of Nanomedicine (Online)
##                                     5
## Revista da Sociedade Brasileira de Medicina Tropical (Impresso)
##                                     5

```

Análise dos dados no formato dataframe

```

public.240.periodico.df <- pub.ls2df(public.240, 1) #artigos
public.240.livros.df <- pub.ls2df(public.240, 2) #livros
public.240.eventos.df <- pub.ls2df(public.240, 5) #eventos

public.242.periodico.df <- pub.ls2df(public.242, 1) #artigos
public.242.livros.df <- pub.ls2df(public.242, 2) #livros
public.242.eventos.df <- pub.ls2df(public.242, 5) #eventos

public.243.periodico.df <- pub.ls2df(public.243, 1) #artigos
public.243.livros.df <- pub.ls2df(public.243, 2) #livros
public.243.eventos.df <- pub.ls2df(public.243, 5) #eventos

public.313.periodico.df <- pub.ls2df(public.313, 1) #artigos
public.313.livros.df <- pub.ls2df(public.313, 2) #livros
public.313.eventos.df <- pub.ls2df(public.313, 5) #eventos

#Dataframe específico para análise de especialidades
especialidade.240.orient <- public.240.eventos.df %>%
  filter (classificacao == "INTERNACIONAL") %>%
  select (`autores-endogeno`) %>%
  separate_rows(`autores-endogeno`, sep = ";") %>%

```

```

distinct() %>%
arrange(`autores-endogeno`) %>% mutate(internacional = "Sim") %>%
right_join(perfil.240.df, by = c("autores-endogeno" = "idLattes")) %>%
left_join(perfil.240.df.areas.de.atuacao, by = c("autores-endogeno" = "idLattes")) %>%
select(`autores-endogeno`, internacional, especialidade, orientacoes_concluidas) %>% distinct() %>%
rename(idLattes = `autores-endogeno`)

especialidade.240.orient[is.na(especialidade.240.orient)] <- 0 #Remove NAs do dataframe
especialidade.240.orient$especialidade <- sub("^$", "Não informada", especialidade.240.orient$especialidade)
especialidade.240.orient$internacional <- sub(0, "Não", especialidade.240.orient$internacional)

#Dataframe especifico para análise de áreas de atuação
perfil.240.areas <- perfil.240.df.areas.de.atuacao %>%
  select(-sub_area, -especialidade) %>%
  distinct() %>%
  group_by(idLattes, orientacoes_concluidas, publicacoes) %>%
  summarise(num_areas = n()) %>% ungroup()
perfil.240.areas$num_areas <- as.factor(perfil.240.areas$num_areas)

```

Publicações por ano

```
table(public.240.periodico.df$ano)
```

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

```

```
table(public.242.periodico.df$ano)
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##    35    39    36    34    50    44    48    49

```

```
table(public.243.periodico.df$ano)
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##   123   136   101   126   116   127   130   118

```

```
table(public.313.periodico.df$ano)
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##    54    73    85   105   100   113   109    97

```

20 revistas mais publicadas (análise com dataframe)

```
head(sort(table(public.240.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

```

```
head(sort(table(public.242.periodico.df$periodico), decreasing = TRUE), 20)
```

```

##
##                               Archives of Virology
##                               13
##                               Plant Disease
##                               9
##                               Plos One
##                               8
##                               Virus Genes
##                               8
##                               Microbial Ecology
##                               6
##                               BMC Genomics
##                               5
##                               Mycologia
##                               5
##                               Mycological Progress
##                               5
##                               Mycotaxon
##                               5
##       Biogeochemistry (Dordrecht)

```

```

##                                     4
##           Frontiers in Microbiology (Online)
##                                     4
##           Fungal Genetics and Biology (Print)
##                                     4
##           Global Change Biology (Print)
##                                     4
##           Microbiologia in Foco
##                                     4
##           Applied Microbiology and Biotechnology
##                                     3
##           BIOENERG RES
##                                     3
##           Biotechnology for Biofuels
##                                     3
##           Brazilian Journal of Microbiology (Impresso)
##                                     3
##           Genetics and Molecular Research
##                                     3
##           Journal of Industrial Microbiology & Biotechnology
##                                     3

```

```
head(sort(table(public.243.periodico.df$periodico), decreasing = TRUE), 20)
```

```

##
##           Plos One
##           49
##           BMC Genomics
##           20
##           BMC Proceedings
##           20
##           Archives of Virology
##           19
##           Toxicon (Oxford)
##           12
##           Frontiers in Microbiology (Online)
##           11
##           Genetics and Molecular Research
##           10
## PLoS Neglected Tropical Diseases (Online)
##           10
##           Protein and Peptide Letters
##           10
##           Scientific Reports
##           10
##           Virus Genes
##           10
##           Plant Disease
##           8
##           Tree Genetics & Genomes (Print)
##           8
##           BMC Plant Biology (Online)
##           7
##           Journal of Proteome Research (Print)
##           7

```



```

##          Proteomics (Weinheim. Print)
##                                7
##          Acta Horticulturae
##                                6
##    Applied Microbiology and Biotechnology
##                                6
##          Journal of Biomedical Nanotechnology
##                                6
##          Microbial Ecology
##                                6
head(sort(table(public.313.periodico.df$periodico), decreasing = TRUE), 20)

##
##                                Plos One
##                                44
##          Frontiers in Microbiology (Online)
##                                22
##          Peptides (New York, N.Y. 1980)
##                                16
##          PLoS Neglected Tropical Diseases (Online)
##                                12
##                                BMC Genomics
##                                11
##          Scientific Reports
##                                11
##          Protein and Peptide Letters
##                                10
##          Biopolymers (New York. Print)
##                                9
##          Journal of Proteome Research (Print)
##                                9
##          Journal of Biomedical Nanotechnology
##                                8
##          Antimicrobial Agents and Chemotherapy (Print)
##                                7
##          Biochimica et Biophysica Acta. Biomembranes
##                                6
##          BMC Microbiology (Online)
##                                6
##          Current Protein and Peptide Science
##                                6
##          Experimental Parasitology
##                                6
##          Fungal Genetics and Biology (Print)
##                                6
##          Journal of Proteomics
##                                6
##          Medical Mycology (Oxford. Print)
##                                6
##          International Journal of Nanomedicine (Online)
##                                5
##    Revista da Sociedade Brasileira de Medicina Tropical (Impresso)
##                                5

```

Número de Orientações de Mestrado e Doutorado

```
sum(sapply(orient.240$ORIENTACAO_CONCLUIDA_DOUTORADO, function(x) length(x$natureza))) +  
sum(sapply(orient.240$ORIENTACAO_CONCLUIDA_MESTRADO, function(x) length(x$natureza)))
```

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

```
sum(sapply(orient.242$ORIENTACAO_CONCLUIDA_DOUTORADO, function(x) length(x$natureza))) +  
sum(sapply(orient.242$ORIENTACAO_CONCLUIDA_MESTRADO, function(x) length(x$natureza)))
```

```
## [1] 175
```

```
sum(sapply(orient.243$ORIENTACAO_CONCLUIDA_DOUTORADO, function(x) length(x$natureza))) +  
sum(sapply(orient.243$ORIENTACAO_CONCLUIDA_MESTRADO, function(x) length(x$natureza)))
```

```
## [1] 460
```

```
sum(sapply(orient.313$ORIENTACAO_CONCLUIDA_DOUTORADO, function(x) length(x$natureza))) +  
sum(sapply(orient.313$ORIENTACAO_CONCLUIDA_MESTRADO, function(x) length(x$natureza)))
```

```
## [1] 386
```

Construção de dataframe de orientações

```
orient.240.posdoutorado.df <- ori.ls2df(orient.240, 6) #pos-Doutorado concluido  
orient.240.doutorado.df <- ori.ls2df(orient.240, 7) #Doutorado concluido  
orient.240.mestrado.df <- ori.ls2df(orient.240, 8) #Mestrado concluido  
orient.240.df <- rbind(rbind(orient.240.posdoutorado.df, orient.240.doutorado.df), orient.240.mestrado.df)
```

```
bolsas.240.df <- orient.240.df %>% filter(bolsa == "SIM") %>%  
  group_by(natureza, ano) %>%  
  summarise(num_bolsas = n()) %>%  
  inner_join((orient.240.df %>%  
    group_by(natureza, ano) %>%  
    summarise(num_orients = n())), by = c("natureza", "ano")) %>%  
  mutate(ratio = num_bolsas/num_orients)
```

```
orient.242.posdoutorado.df <- ori.ls2df(orient.242, 6) #pos-Doutorado concluido  
orient.242.doutorado.df <- ori.ls2df(orient.242, 7) #Doutorado concluido  
orient.242.mestrado.df <- ori.ls2df(orient.242, 8) #Mestrado concluido  
orient.242.df <- rbind(rbind(orient.242.posdoutorado.df, orient.242.doutorado.df), orient.242.mestrado.df)
```

```
orient.243.posdoutorado.df <- ori.ls2df(orient.243, 6) #pos-Doutorado concluido  
orient.243.doutorado.df <- ori.ls2df(orient.243, 7) #Doutorado concluido  
orient.243.mestrado.df <- ori.ls2df(orient.243, 8) #Mestrado concluido  
orient.243.df <- rbind(rbind(orient.243.posdoutorado.df, orient.243.doutorado.df), orient.243.mestrado.df)
```

```
orient.313.posdoutorado.df <- ori.ls2df(orient.313, 6) #pos-Doutorado concluido  
orient.313.doutorado.df <- ori.ls2df(orient.313, 7) #Doutorado concluido  
orient.313.mestrado.df <- ori.ls2df(orient.313, 8) #Mestrado concluido  
orient.313.df <- rbind(rbind(orient.313.posdoutorado.df, orient.313.doutorado.df), orient.313.mestrado.df)
```

Construção de Grafo de Colaborações

```

g.240 <- g.ls2ig(graphl.240)
df.240 <- as.data.frame(V(g.240)$name); colnames(df.240) <- "Idlattes"
df.240 <- left_join(df.240, df.prog, by = c("Idlattes" = "idLattes"))

g.242 <- g.ls2ig(graphl.242)
df.242 <- as.data.frame(V(g.242)$name); colnames(df.242) <- "Idlattes"
df.242 <- left_join(df.242, df.prog, by = c("Idlattes" = "idLattes"))

g.243 <- g.ls2ig(graphl.243)
df.243 <- as.data.frame(V(g.243)$name); colnames(df.243) <- "Idlattes"
df.243 <- left_join(df.243, df.prog, by = c("Idlattes" = "idLattes"))

g.313 <- g.ls2ig(graphl.313)
df.313 <- as.data.frame(V(g.313)$name); colnames(df.313) <- "Idlattes"
df.313 <- left_join(df.313, df.prog, by = c("Idlattes" = "idLattes"))

```

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.240 <- df.240 %>% group_by(Idlattes) %>%
  slice(1L)
V(g.240)$programa <- df.240$Programa
V(g.240)$orient_dout <- perfil.240.df$ORIENTACAO_CONCLUIDA_DOUTORADO
V(g.240)$orient_mest <- perfil.240.df$ORIENTACAO_CONCLUIDA_MESTRADO
V(g.240)$publicacao <- perfil.240.df$PERIODICO
V(g.240)$eventos <- perfil.240.df$EVENTO

df.242 <- df.242 %>% group_by(Idlattes) %>%
  slice(1L)
V(g.242)$programa <- df.242$Programa
V(g.242)$orient_dout <- perfil.242.df$ORIENTACAO_CONCLUIDA_DOUTORADO
V(g.242)$orient_mest <- perfil.242.df$ORIENTACAO_CONCLUIDA_MESTRADO
V(g.242)$publicacao <- perfil.242.df$PERIODICO
V(g.242)$eventos <- perfil.242.df$EVENTO

df.243 <- df.243 %>% group_by(Idlattes) %>%
  slice(1L)
V(g.243)$programa <- df.243$Programa
V(g.243)$orient_dout <- perfil.243.df$ORIENTACAO_CONCLUIDA_DOUTORADO
V(g.243)$orient_mest <- perfil.243.df$ORIENTACAO_CONCLUIDA_MESTRADO
V(g.243)$publicacao <- perfil.243.df$PERIODICO
V(g.243)$eventos <- perfil.243.df$EVENTO

df.313 <- df.313 %>% group_by(Idlattes) %>%
  slice(1L)
V(g.313)$programa <- df.313$Programa
V(g.313)$orient_dout <- perfil.313.df$ORIENTACAO_CONCLUIDA_DOUTORADO
V(g.313)$orient_mest <- perfil.313.df$ORIENTACAO_CONCLUIDA_MESTRADO
V(g.313)$publicacao <- perfil.313.df$PERIODICO
V(g.313)$eventos <- perfil.313.df$EVENTO

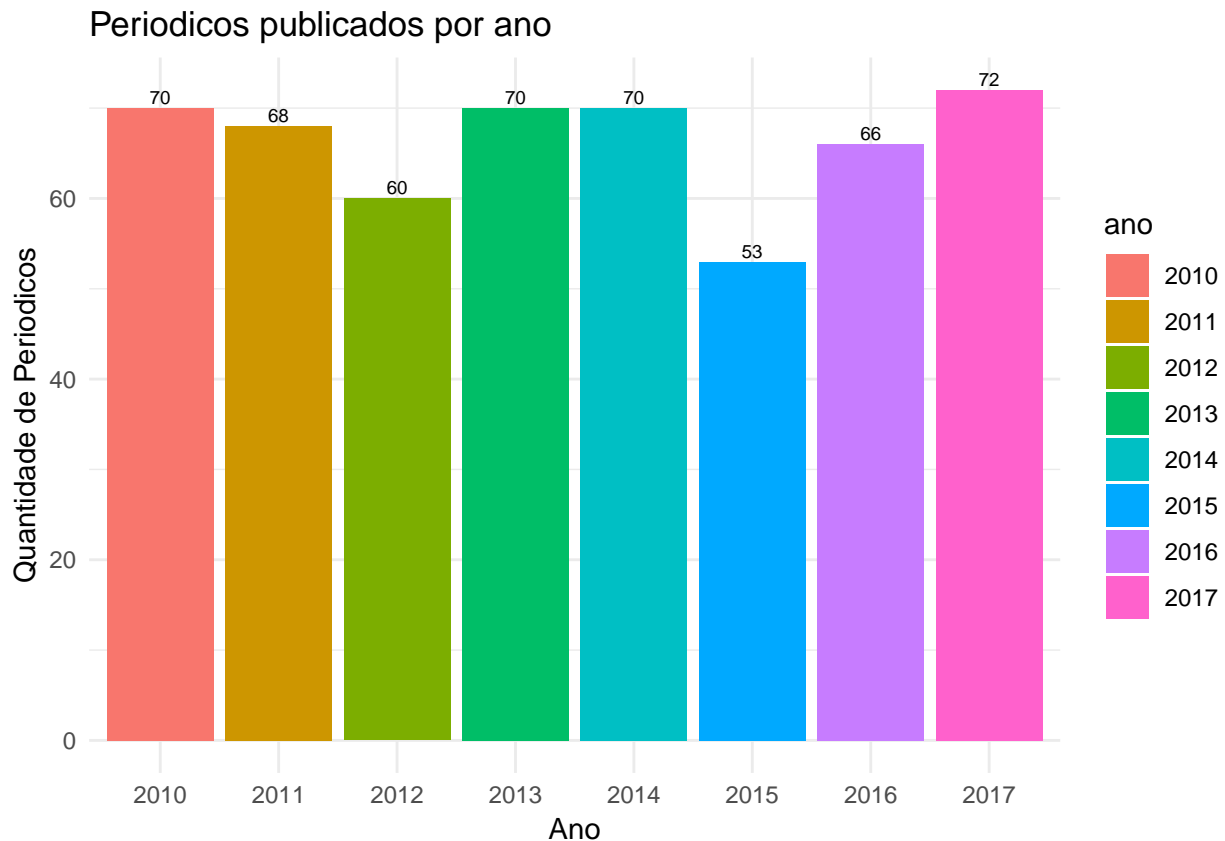
```

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.

Gráfico de barras; periodicos por ano

```
public.240.periodico.df %>%  
  group_by(ano) %>%  
  summarise(Quantidade = n()) %>%  
  ggplot(aes(ano, Quantidade, fill = ano)) +  
  geom_bar(position = "stack", stat = "identity") +  
  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")
```

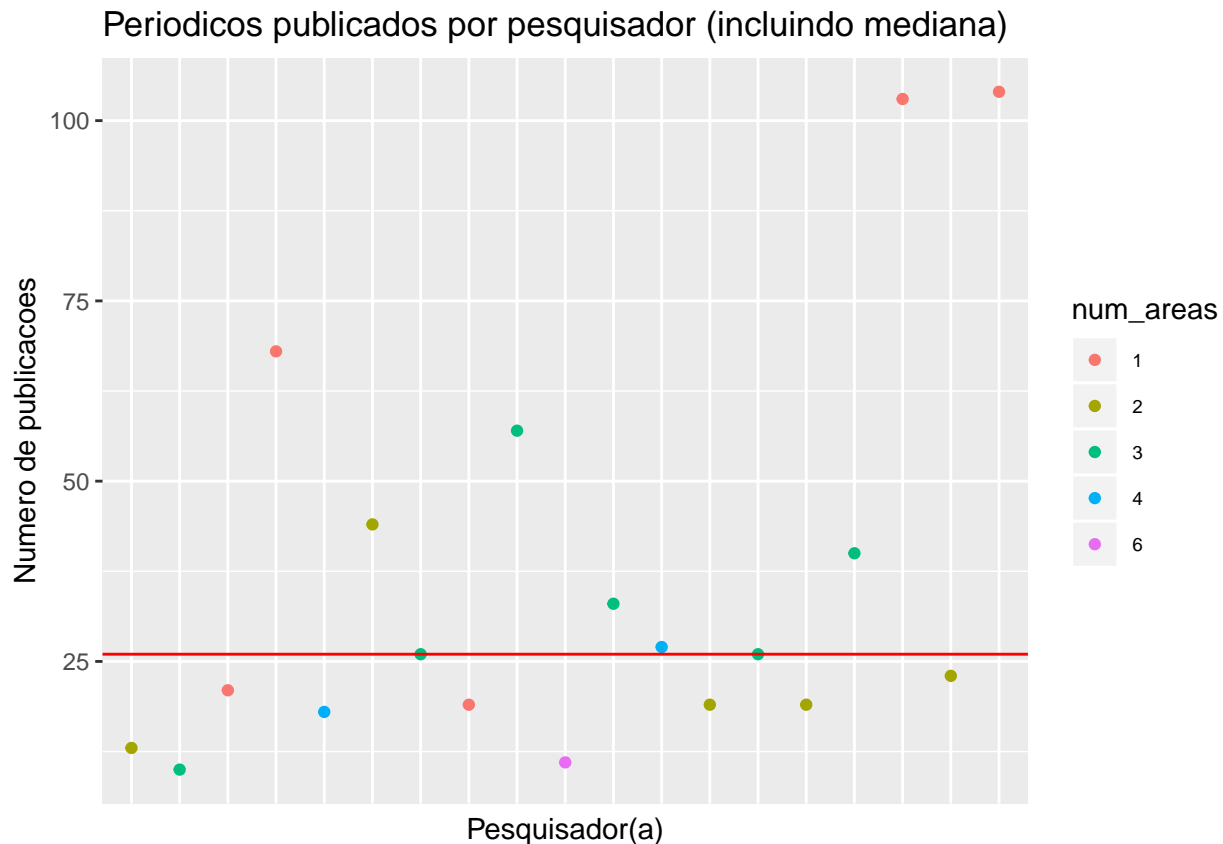


Sob uma perspectiva de publicações em periódicos, o Programa de Pós-Graduação de Biologia Animal mantém-se estável com flutuações normais, mas apresentou uma queda considerável de 2014 para 2015, diminuindo o número de publicações em cerca de 18% mas recuperando em anos posteriores.

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

```
perfil.240.df %>%  
  ggplot(aes(idLattes, PERIODICO, color = num_areas)) +  
  geom_point() +
```

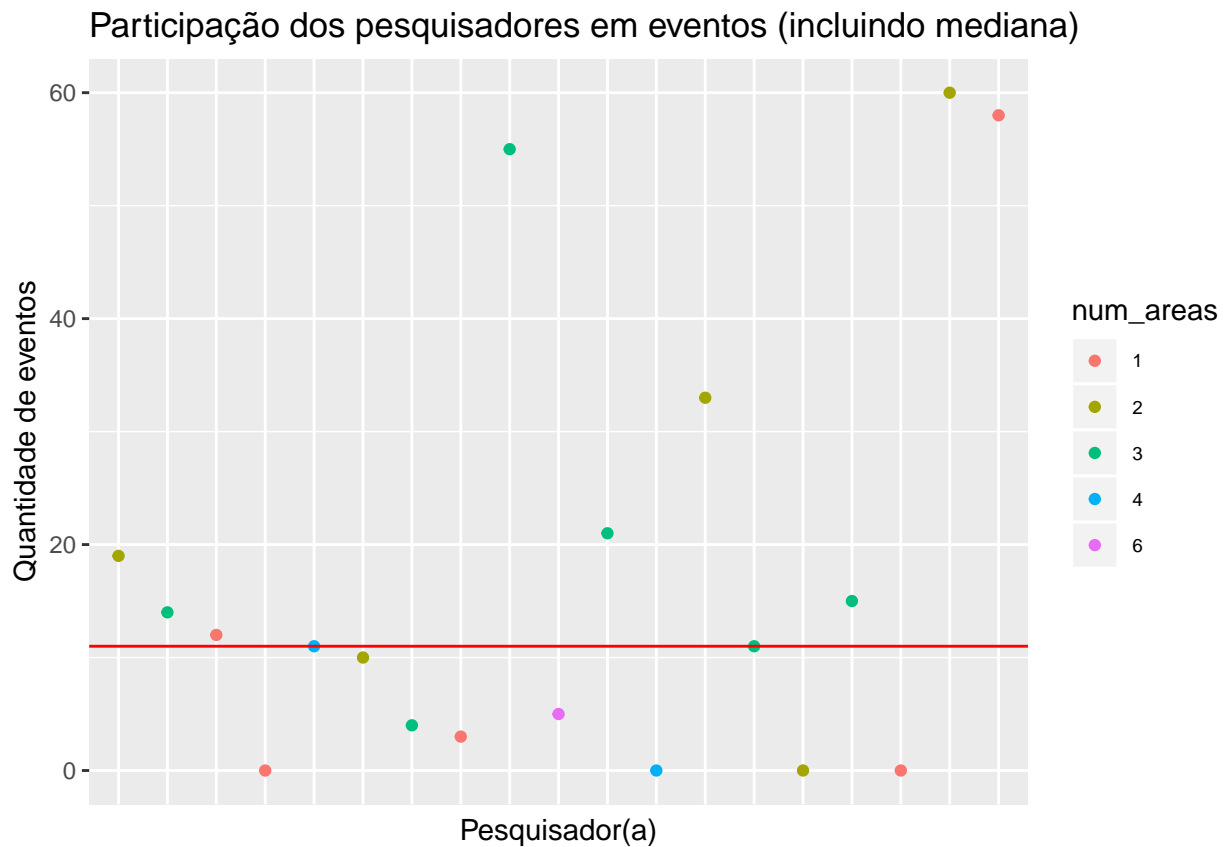
```
ggtitle("Periodicos publicados por pesquisador (incluindo mediana)") +
theme(legend.position="right",legend.text=element_text(size=7)) +
guides(fill=guide_legend(nrow=5, byrow=TRUE, title.position = "top")) +
labs(x="Pesquisador(a)",y="Numero de publicacoes") +
theme(axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
geom_hline(yintercept = sum(perfil.240.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.

Número de areas de pesquisa e quantidade de eventos por pesquisador

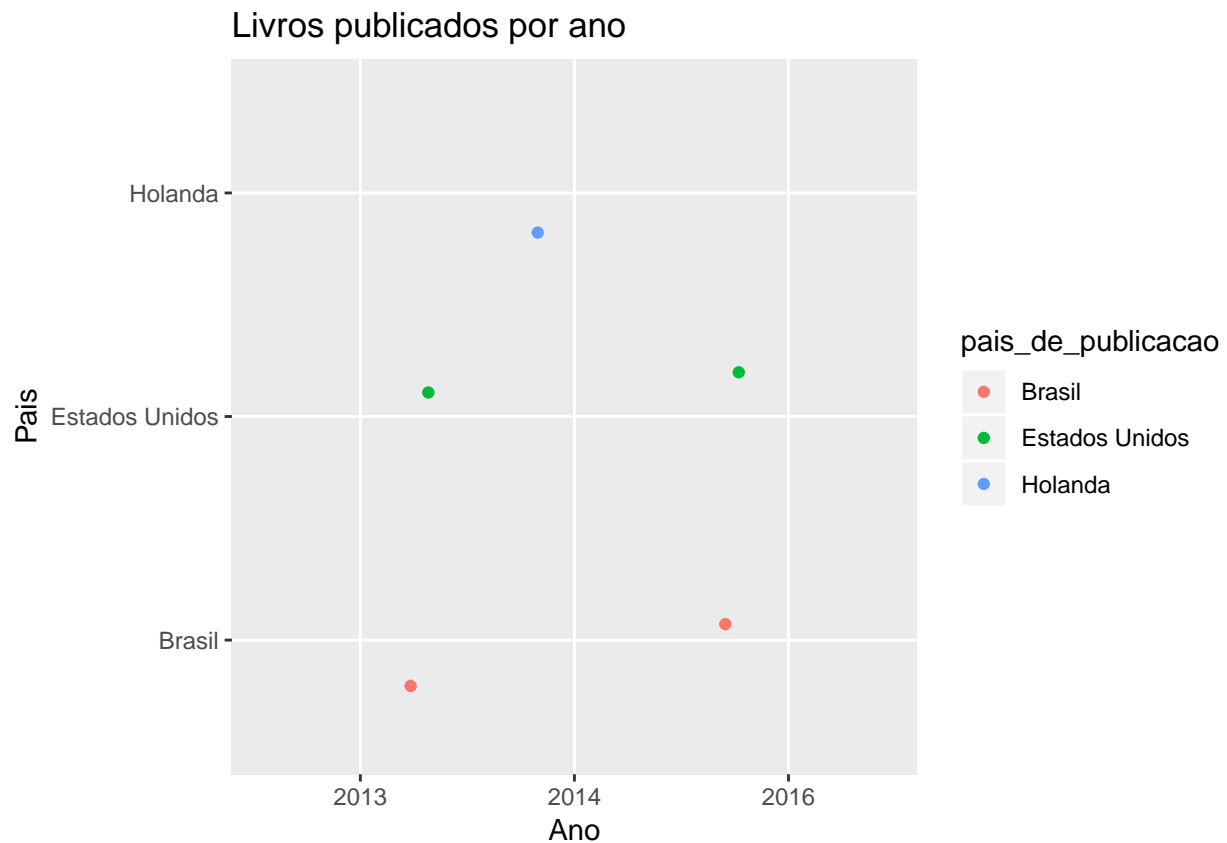
```
perfil.240.df %>%
ggplot(aes(idLattes,EVENTO, color = num_areas)) +
geom_point() +
ggtitle("Participação dos pesquisadores em eventos (incluindo mediana)") +
theme(legend.position="right",legend.text=element_text(size=7)) +
guides(fill=guide_legend(nrow=5, byrow=TRUE, title.position = "top")) +
labs(x="Pesquisador(a)",y="Quantidade de eventos") +
theme(axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
geom_hline(yintercept = sum(perfil.240.df %>% summarize(x = median(EVENTO))), color = "red")
```



AAAAAAAAA

Publicação de livros por pais/ano

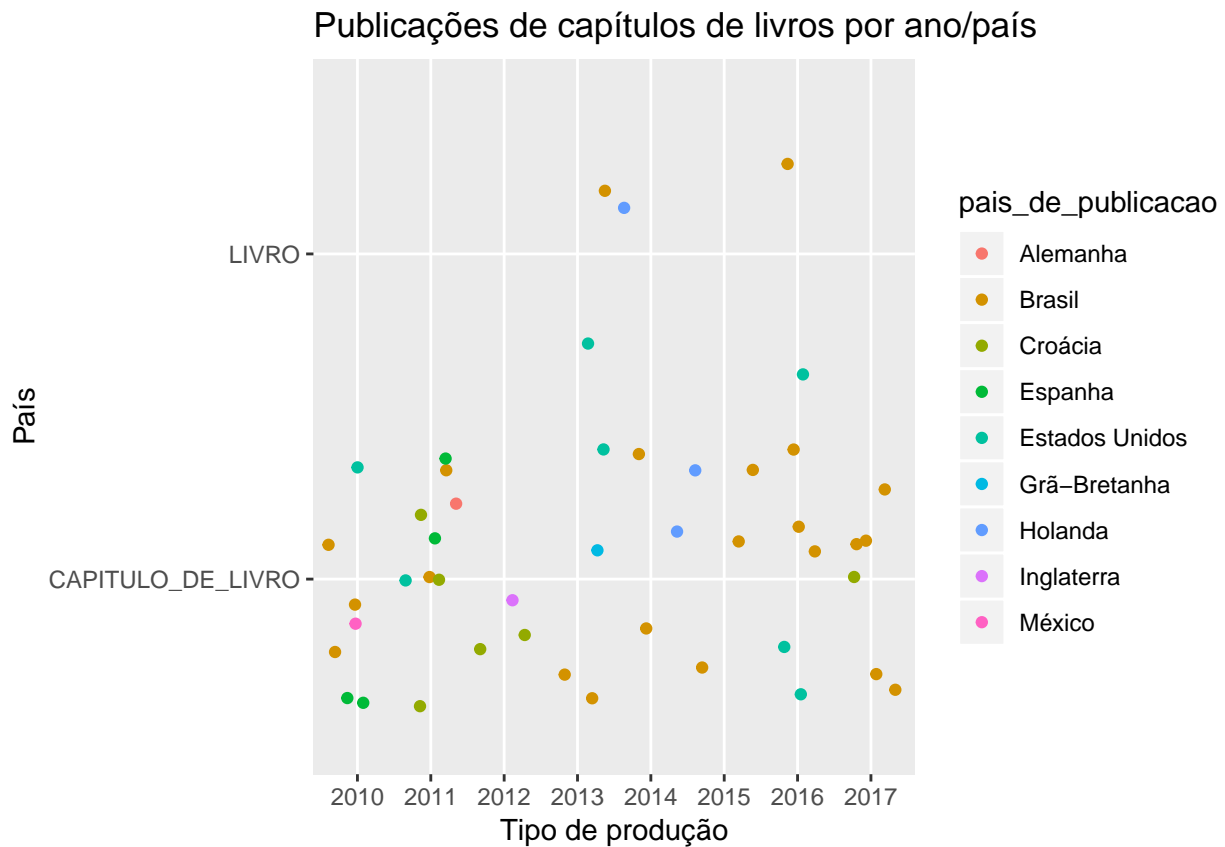
```
public.240.livros.df %>%
  group_by(ano, pais_de_publicacao) %>%
  ggplot(aes(x=ano, y=pais_de_publicacao, color= pais_de_publicacao)) +
  ggtitle("Livros publicados por ano") +
  xlab("Ano") + ylab("Pais") + geom_jitter()
```



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 livros publicados para o PPG em Biologia Animal

Publicações de capítulos de livros por ano/país

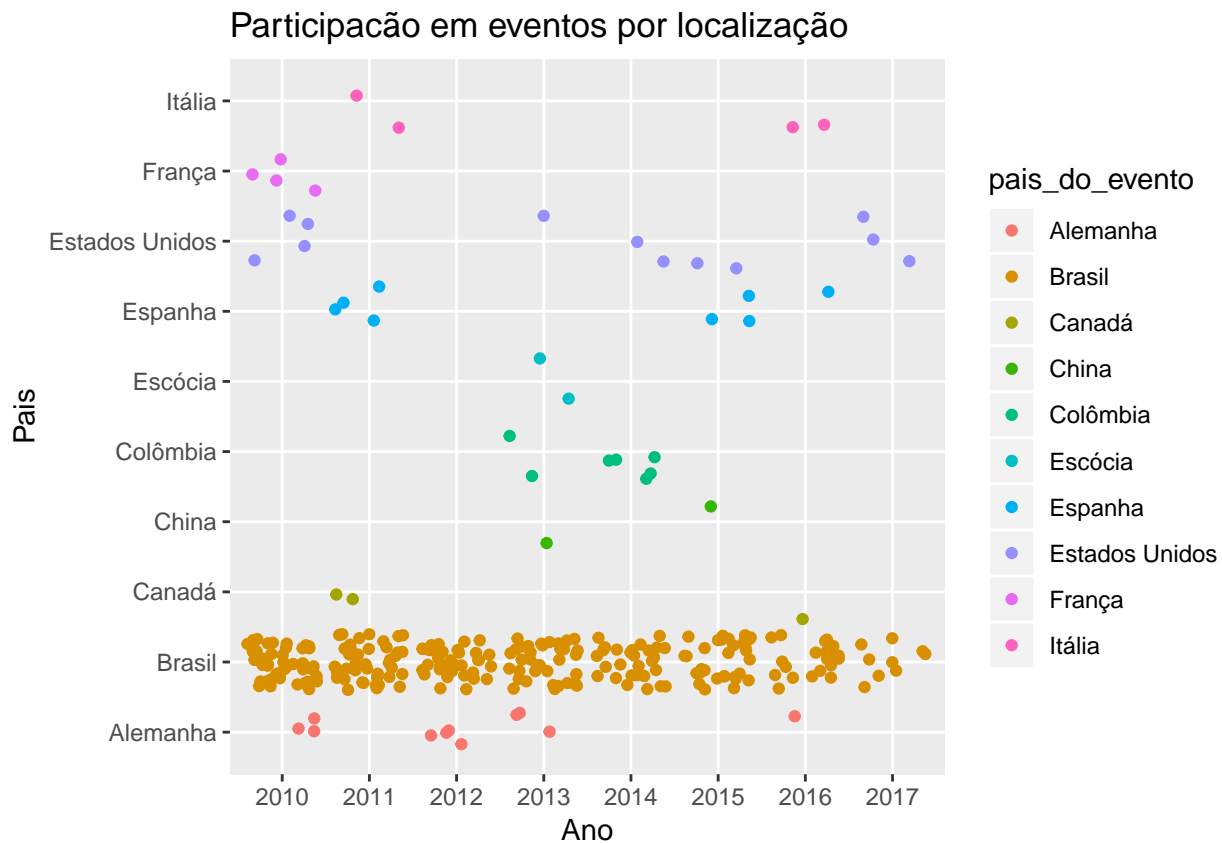
```
perfil.240.df.publicacoes %>%
  filter((tipo_producao %in% c('LIVRO', 'CAPITULO_DE_LIVRO'))) %>%
  group_by(tipo_producao, pais_de_publicacao) %>%
  ggplot(aes(ano, tipo_producao, col = pais_de_publicacao)) +
  geom_jitter() + ggtitle("Publicações de capítulos de livros por ano/país") +
  labs(x = 'Tipo de produção', y = 'País')
```



AAAAAA

Eventos nacionais e internacionais

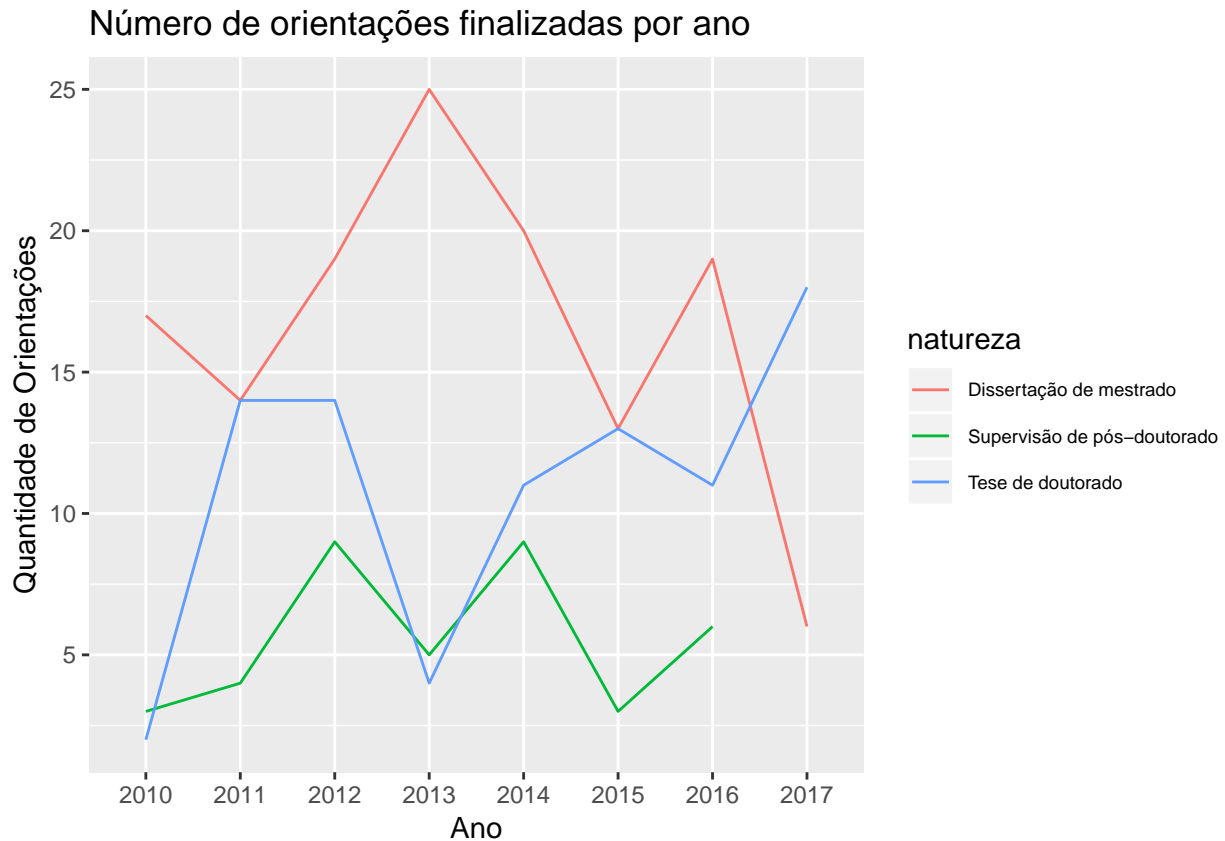
```
public.240.eventos.df %>%
  filter(pais_do_evento %in%
    c(names(head(sort(table(public.240.eventos.df$pais_do_evento)
      , decreasing = TRUE), 10)))) %>%
  group_by(ano_do_trabalho,pais_do_evento) %>%
  ggplot(aes(x=ano_do_trabalho,y=pais_do_evento, color= pais_do_evento)) +
  ggtitle("Participação em eventos por localização") +
  xlab("Ano") + ylab("Pais") + 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

```
orient.240.df %>%
  group_by(ano, natureza) %>% summarise(total = n()) %>%
  ggplot(aes(x=ano,y=total,group=natureza,color=natureza)) +
  geom_line() +
  ggtitle("Número de orientações finalizadas 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 Orientações")
```

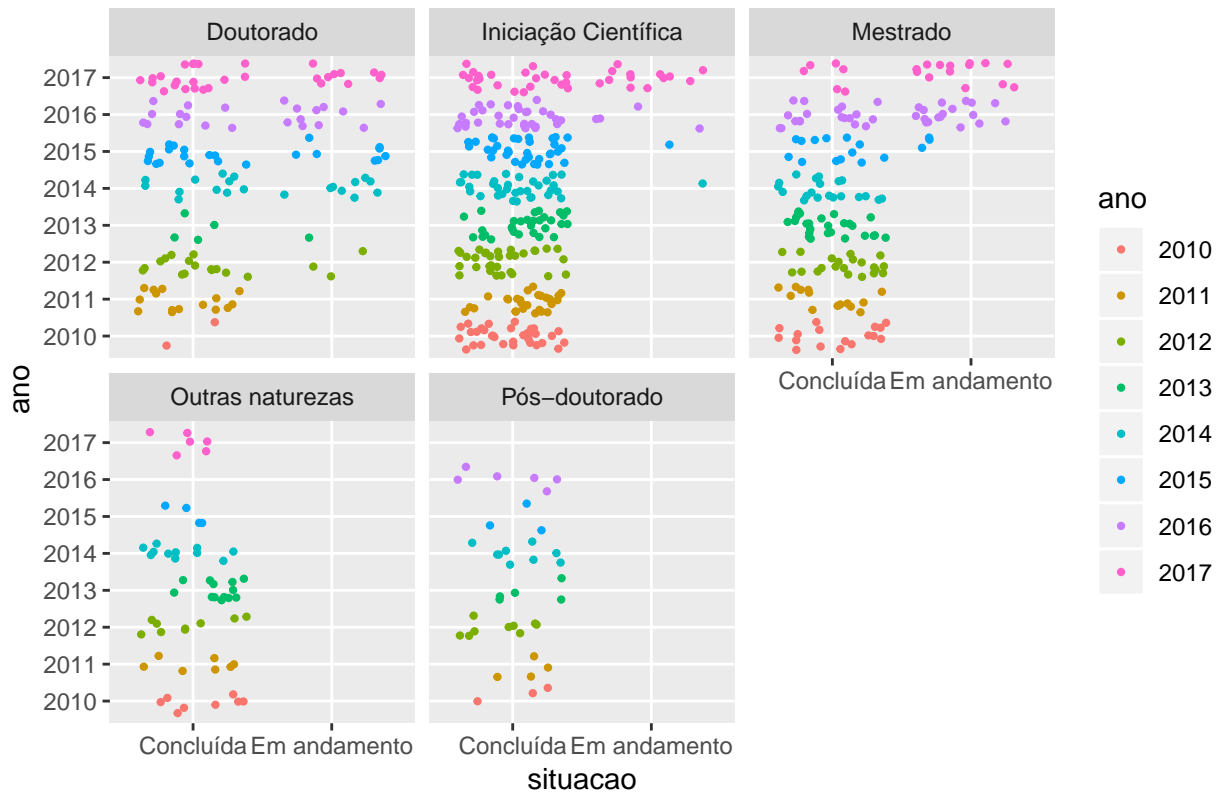


Observando a evolução do número de orientações completas ao longo dos anos, percebe-se que o Programa de Pós-Graduação cresceu consideravelmente na natureza de mestrado após 2010, com uma regressão em 2016. Entretanto, não há um comportamento linear na evolução do número de orientações finalizadas. 2014 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.

Evolução temporal de orientações em andamento e concluídas

```
perfil.240.df.orientacoes %>% group_by(ano, situacao) %>%
  ggplot(aes(x=situacao,y=ano,color=ano)) +
  geom_jitter(size = 0.8) + facet_wrap(. ~ Natureza) +
  ggtitle("Disposição de orientações por ano")
```

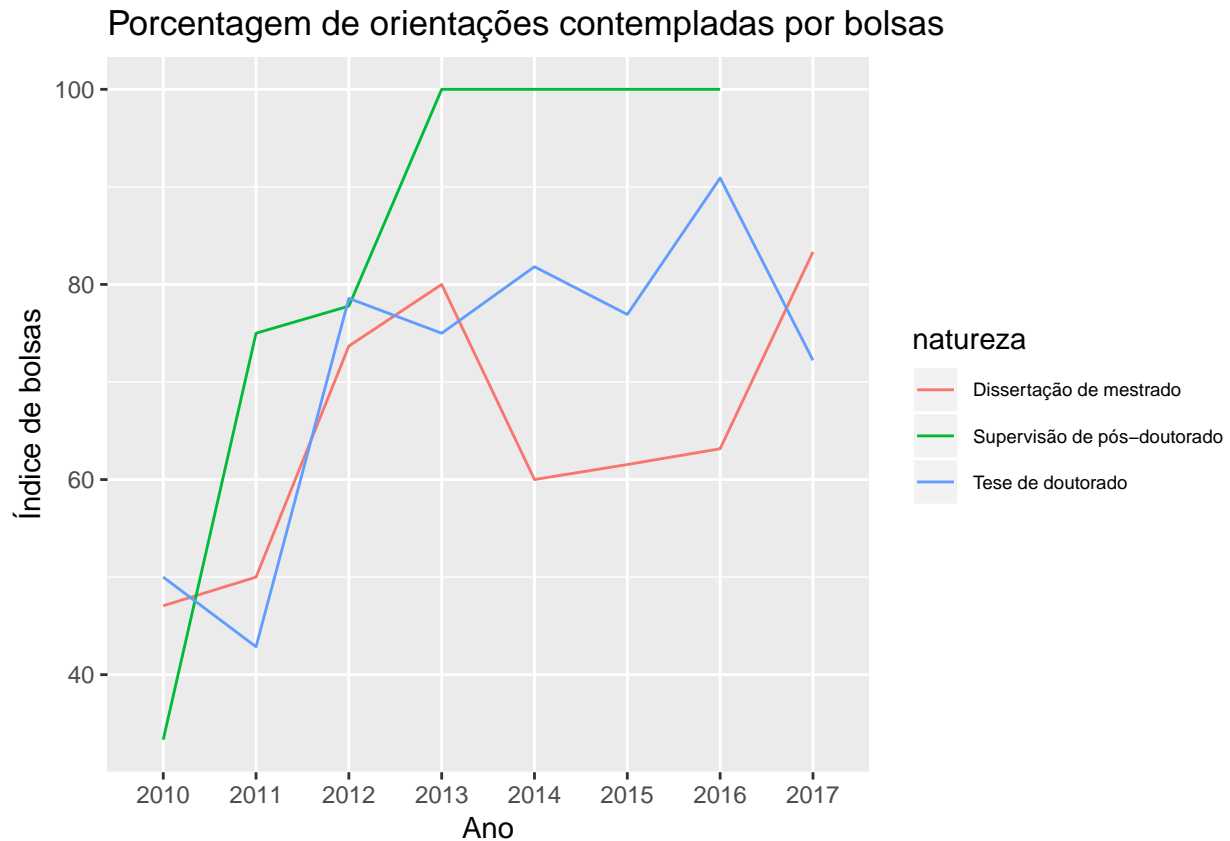
Disposição de orientações por ano



AAAAAAAAA

Índice de bolsas entre naturezas e evolução temporal

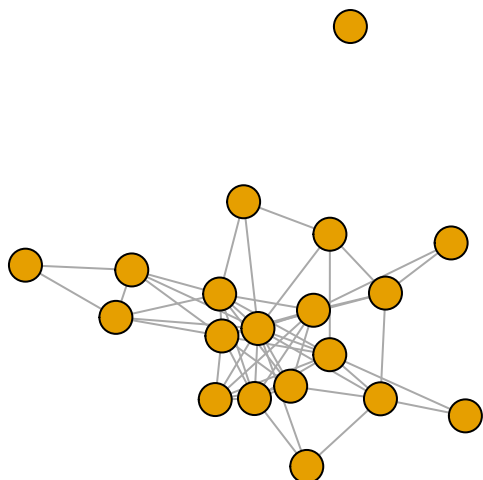
```
bolsas.240.df %>%
  ggplot(aes(x=ano,y=ratio*100,color=natureza, group=natureza)) +
  geom_line() +
  ggtitle("Porcentagem de orientações contempladas por bolsas") +
  theme(legend.position="right",legend.text=element_text(size=7)) +
  guides(fill=guide_legend(nrow=5, byrow=TRUE, title.position = "top")) +
  labs(x="Ano",y="Índice de bolsas")
```



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

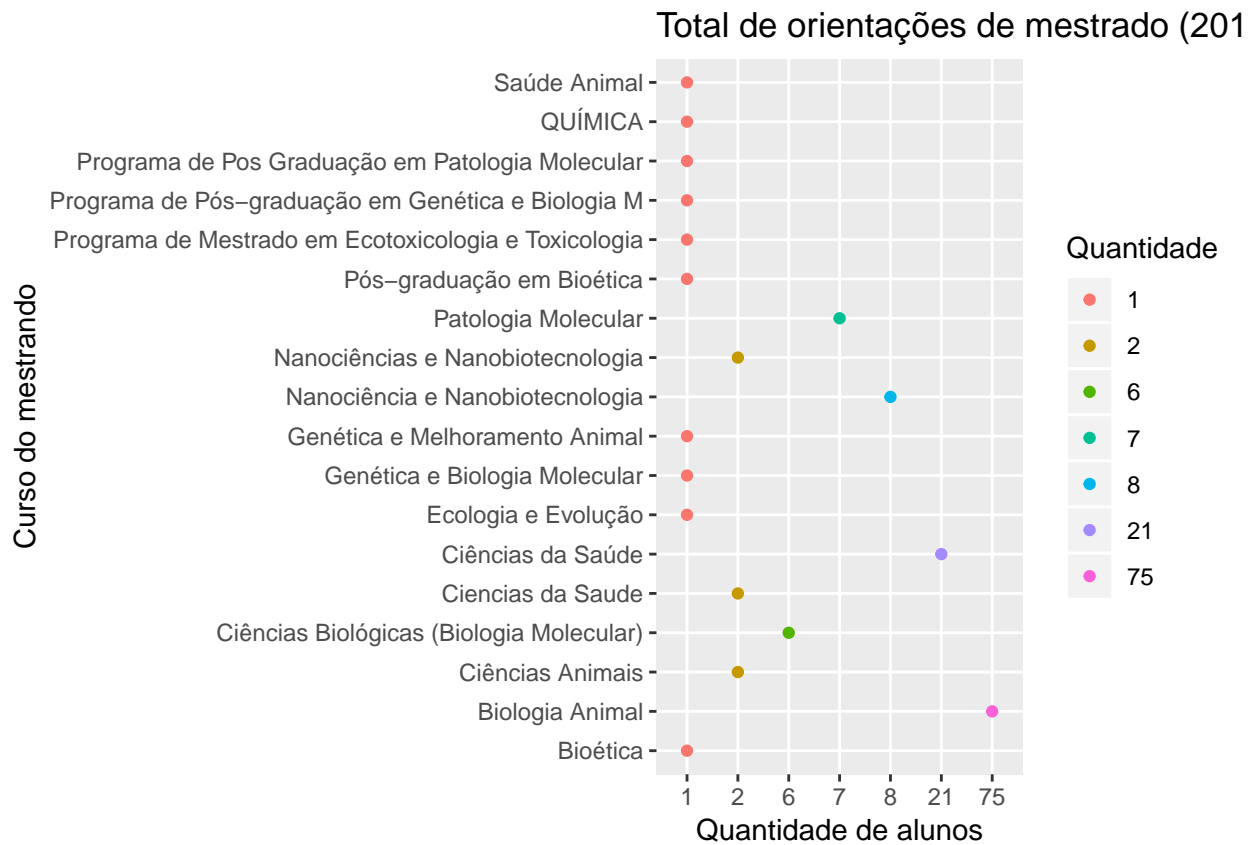
```
plot(g.240, vertex.label = NA)
```



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 exceções tais como os vértices mais afastados e o vértice único que não tem nenhuma aresta com adjacentes.

Cursos de Mestrado e total de alunos no período (2011-2017)

```
orient.240.mestrado.df %>%
  group_by(curso) %>%
  summarise(Quantidade = n()) %>% mutate(Quantidade = factor(Quantidade)) %>%
  ggplot(aes(x=Quantidade, y = curso, color = Quantidade)) +
  geom_point() +
  ggtitle('Total de orientações de mestrado (2011-2017)') +
  labs(x='Quantidade de alunos', y='Curso do mestrando')
```



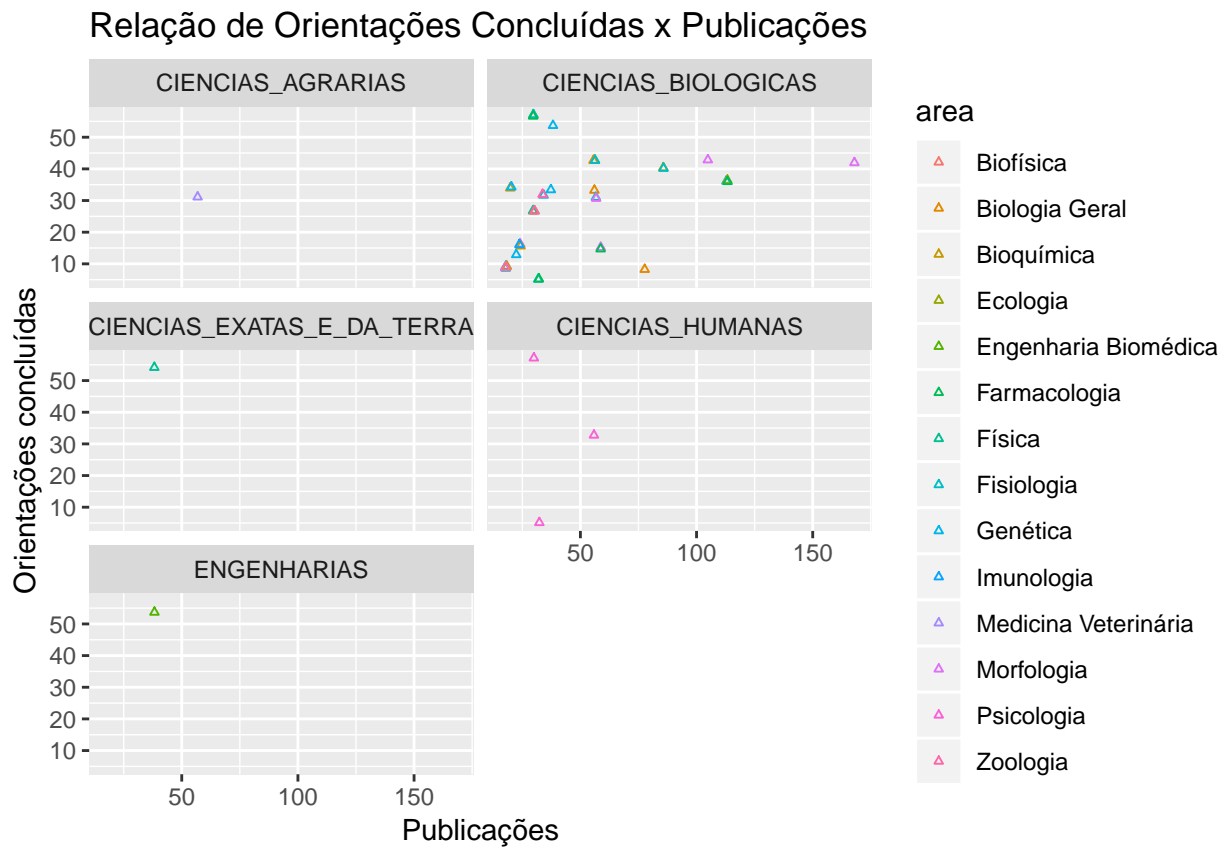
AAAAAAA

Relação de Orientações Concluídas x Publicações por área:

```

perfil.240.df.areas.de.atuacao %>%
  select(-sub_area, -especialidade) %>%
  distinct() %>%
  ggplot(aes(publicacoes, orientacoes_concluidas, color = area)) +
  geom_jitter(shape = 2, size = .8) +
  ggtitle('Relação de Orientações Concluídas x Publicações') +
  labs(x='Publicações',y='Orientações concluídas') + facet_wrap(grande_area ~ ., ncol = 2)

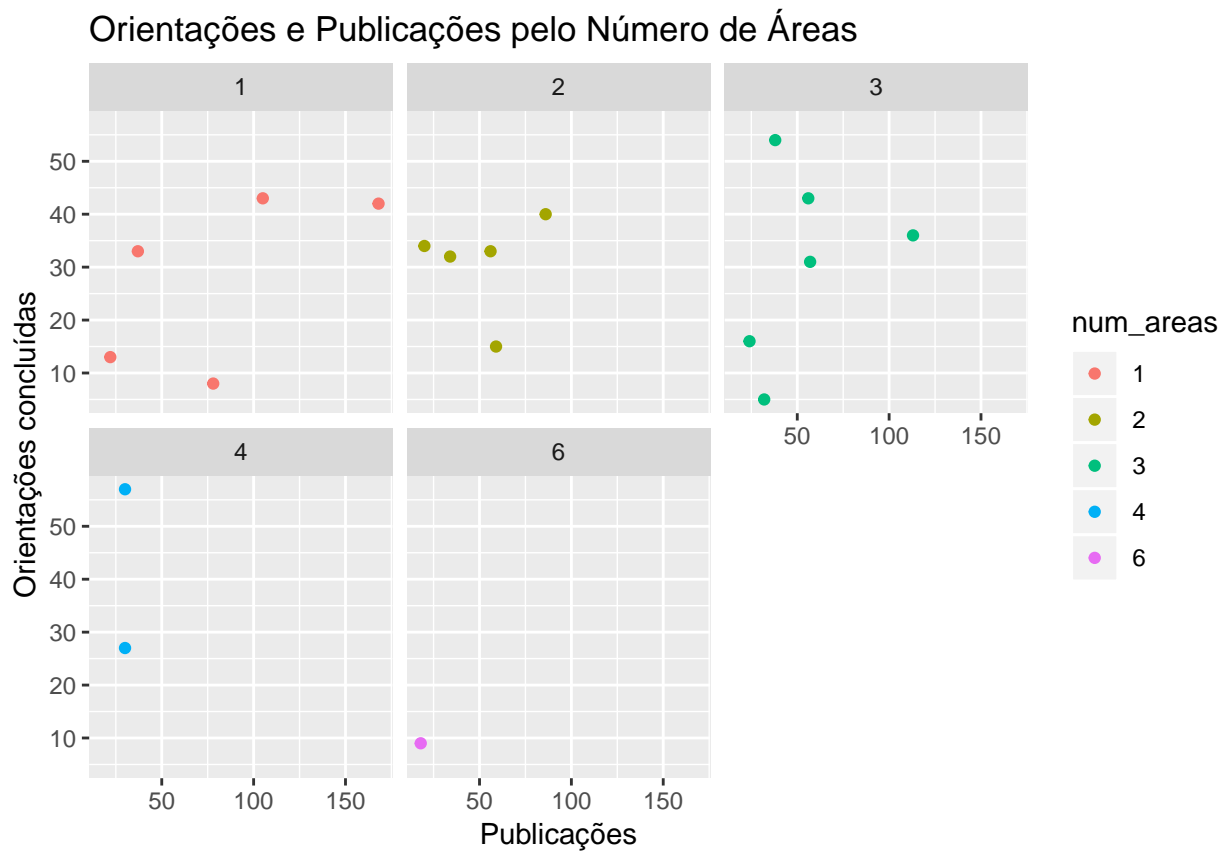
```



AAAAAAA Lembrando que um pesquisador possui mais de uma grande area, o pesquisador se repete entre diferentes graficos

Quantidade de Orientações e Publicações pelo Número de Áreas

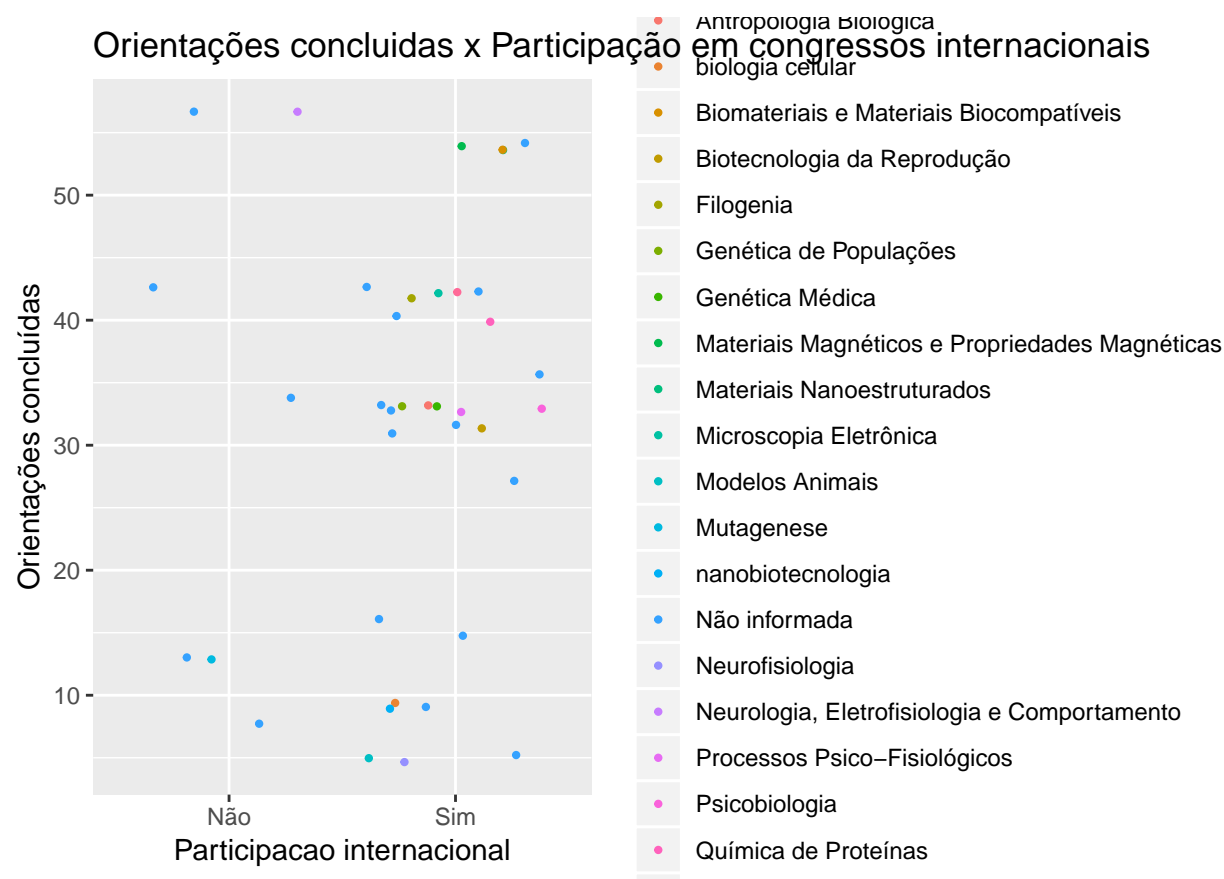
```
perfil.240.areas %>%
  ggplot(aes(publicacoes, orientacoes_concluidas, color = num_areas)) +
  geom_point() +
  ggtitle('Orientações e Publicações pelo Número de Áreas') +
  labs(x='Publicações',y='Orientações concluídas') + facet_wrap(. ~ num_areas)
```



AAAAA Quem trabalha em mais áreas diferentes publica/orienta mais?

Presença em congressos internacionais por número de orientações:

```
especialidade.240.orient %>%
  ggplot(aes(internacional, orientacoes_concluidas, color = especialidade)) +
  geom_jitter(size = .8) +
  ggtitle('Orientações concluídas x Participação em congressos internacionais') +
  labs(x='Participacao internacional',y='Orientações concluídas')
```

AAAA

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. Disponível em: <https://www.the-modeling-agency.com/>, visitado em Novembro, 2019.

Fernandes, Jorge H C, Ricardo Barros Sampaio, e João Ribas de Moura. “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

Secretaria de Comunicação da UnB (SeCom). 2018. “Impacto da produção científica da UnB aumenta mais de 100% em seis anos.” Universidade de Brasília, Junho de 2018, Disponível em <https://www.noticias.unb.br/>, visitado em Novembro de 2018.

Azevedo, Mário Luiz Neves de, João Ferreira de Oliveira, e Afrânio Mendes Catani. “O Sistema Nacional de Pós-Graduação (SNPG) e o Plano Nacional de Educação (PNE 2014-2024): regulação, avaliação e financiamento”. Revista Brasileira de Política e Administração da Educação 32, nº 3 (2016). <http://dx.doi.org/10.21573/vol32n32016.68576>.

Leite, Fernando César Lima. “Considerações básicas sobre a Avaliação do Sistema Nacional de PósGraduação”. Comunicação Pessoal (slides). Universidade de Brasília, abril de 2018.

CAPES. “Documentos de Área”. CAPES.gov.br. Acessado 12 de junho de 2018. Disponível em: <http://avaliacaoquadrienal.capes.gov.br/>, visitado em Dezembro de 2018.

Pátaro, Cristina Saitê de Oliveira, e Frank Antonio Mezzomo. “Sistema Nacional de Pós-Graduação no Brasil: estrutura, resultados e desafios para política de Estado - Lívio Amaral”. Revista Educação e Linguagens 2, nº 3 (julho de 2013): 11–17.

Fernandes, Jorge H C, e Ricardo Barros Sampaio. “DataScienceForAll”. Zotero, 13 de junho de 2018