

Ciência de Dados para Todos (Data Science For All) - 2019.2 - Análise da Produção Científica e Acadêmica da Universidade de Brasília - Relatório - Departamento de Ciência da Computação da UnB

Luan mendes Gonçalves Freitas - 15/0015585

25/10/2019

Contents

1	Introdução	3
2	Metodologia	3
2.1	O que é ciência?	3
2.2	O que é ciência no Brasil?	3
2.3	O que é CRISP-DM?	4
3	Fase 1 - Entendimento do Negócio	6
3.1	O que é o Sistema Nacional de Pós-Graduação? (Contextualização)	6
3.2	A UnB dentro do Sistema Nacional de Pós-Graduação (Contextualização)	7
4	Fase 2 - Entendimento dos Dados	8
4.1	Coleta inicial dos dados	8
4.2	Descrição dos Dados	8
4.3	Análise exploratória dos dados	15
4.3.1	Arquivo Perfil	15
4.3.2	Arquivo Publicação	21
4.3.3	Arquivo Orientação	22
4.4	Verificação da qualidade dos dados	23
5	Fase 3 - Preparação dos Dados	24
5.1	Seleção dos dados.	24
5.2	Limpeza dos dados	24
5.3	Construção dos dados	24
5.4	Integração dos dados	25
5.5	Formatação dos dados	25

6 Fase 4 - Modelagem	26
6.1 Análise de Redes	26
7 Fase 5 - Avaliação	31
7.0.1 Arquivo Perfil	31
7.0.2 Arquivo Publicação	52
7.0.3 Arquivo Orientação	57
7.1 Revisão do processo	59
8 Fase 6 - Implantação (deployment)	60
9 Conclusão	60
Referências	60

1 Introdução

Este relatório apresenta uma análise da produção científica e acadêmica de dados coletados da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), à respeito de programas de pós-graduação da disciplina Tópicos Avançados em Computadores Data Science For All (Ciência de Dados para Todos) - Turma D - 2019.2, ministrada por Jorge Henrique Cabral Fernandes e Ricardo Barros Sampaio, do Departamento de Ciência da Computação da Universidade de Brasília.

O tema de estudo escolhido para realização da prática de ciência de dados foi o cenário atual da Pós-Graduação Brasileira, para que seus dados sejam processados e estudados com o objetivo de se retirar análises a respeito da qualidade, relevância, e produtividade dos programas de pós-graduação brasileiros.

Órgão Capes Sucupira já coleta informações, para realizar análises e avaliações e ser a base de referência do Sistema Nacional de Pós-Graduação (SNPG). A Plataforma disponibiliza em tempo real e com muito mais transparência as informações, processos e procedimentos que a CAPES realiza no SNPG para toda a comunidade acadêmica. Igualmente, a Plataforma propiciará a parte gerencial-operacional de todos os processos e permitirá maior participação das pró-reitorias e coordenadores de programas de pós-graduação (Capes 2006).

O tema de estudo escolhido tem como objetivo apresentar as análises descritivas, quantitativas e de modelagem realizadas dos programas de pós-graduações da brasileira, seguindo o modelo de metodologia CRISP-DM. O programa escolhido é **Geotecnia (53001010032P2)** da área de conhecimento de Engenharias I (Geotecnia 2013).

2 Metodologia

2.1 O que é ciência?

Ciência é o conhecimento que explica os fenômenos obedecendo a leis que foram verificadas por métodos experimentais. Aristóteles define a ciência como o “conhecimento das causas pelas causas. É o conhecimento demonstrativo”.

A ciência é composta por três componentes: a observação, a experimentação e as leis. Visa a união entre o conhecimento teórico, a prática e a técnica. Não se utiliza de suposições, mas da comprovação após a aplicação do método científico.

Foi o próprio Aristóteles quem definiu que as ciências (no plural) estão relacionadas à maneira de realização do ideal de cientificidade de acordo com os fatos investigados e os métodos empregados (Todamateria 2006).

2.2 O que é ciência no Brasil?

Diz-se que o sistema de ciência e tecnologia do Brasil começa oficialmente com o CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), criado em 1951 para incentivar nosso progresso na área. Claro, já havia antes instituições de destaque: do Observatório Nacional no Rio (ainda do tempo do Império) às universidades federal do Rio de Janeiro (UFRJ) e estadual de São Paulo (USP), fundadas em 1920 e 1934. De qualquer forma, trata-se de uma história muito jovem. Basta lembrar que os Estados Unidos viram sua prestigiosa Universidade Harvard surgir em 1636. Ou seja: estamos correndo atrás, e não faz pouco tempo. A boa notícia é que nos últimos 20 anos a coisa finalmente parece ter engrenado.

Passamos da 21ª para a 13ª posição no ranking mundial de produção científica (veja ao lado) e levantamentos já indicam que o Brasil responde por 2,7% dos trabalhos científicos publicados no mundo (em 1994, era apenas 0,7%). “Se for pensar que o PIB do Brasil [Produto Interno Bruto, soma de todas as riquezas produzidas pelo país em um ano] representa cerca de 2,9% do mundo, você vê que nossa ciência já atingiu

um tamanho proporcional à nossa economia”, diz Marco Antonio Raupp, ministro da Ciência, Tecnologia e Inovação. Em evento recente na Sociedade Brasileira para o Progresso da Ciência (SBPC), Raupp mostrou que os investimentos do governo na área foram de R\$ 1,1 bilhão em 2000 para R\$ 12,7 bilhões em 2013. “Nunca antes na história desse país, como costuma dizer nosso ex-presidente”, brincou o ministro.

Mas nem tudo são flores. Pra começar, a parcela do PIB investida em pesquisa e desenvolvimento, 1,16%, ainda é pequena se comparada com a de nações desenvolvidas como a Alemanha (2,7%) ou EUA (2,8%) - e não vem crescendo expressivamente na última década. Isso deixa muito trabalho bom de fora. “Se a taxa de projetos aprovados no CNPq é de 50% e a taxa de financiados é de 20%, nos sentimos no direito de pleitear mais”, diz Helena Nader, presidente da SBPC.

Além do dinheiro, há entraves puramente burocráticos. Não à toa está entre as prioridades do governo fazer aprovar no Congresso Nacional um Código Nacional de Ciência, Tecnologia e Inovação. Trata-se de um conjunto de leis que deve resolver alguns desses problemas - mas não todos. Até mesmo o ex-jogador e deputado federal Romário de Souza Faria (PSB-RJ) anda enchendo a bola dos cientistas, com um projeto de lei para facilitar as importações, uma das piores vias-crúcis de quem quer fazer pesquisa de ponta no país. São sinais alvissareiros: pouco a pouco, os freios da ciência nacional começam a ser colocados de lado, após décadas de protestos dos meios acadêmicos.

Mas há muito a ser feito: de facilitar o acesso de cientistas a recursos a fazer o governo se mexer quando surgem oportunidades em parcerias científicas internacionais. Neste Dossiê, mostramos algumas das principais dificuldades ainda enfrentadas pelos pesquisadores brasileiros, e como o país se prepara para lidar com elas.

2.3 O que é CRISP-DM?

CRISP-DM (Cross-Industry Standard Process for Data Mining) é um modelo de análise de mineração de dados, feita de forma sistemática, sendo amplamente utilizada por ser flexível, podendo ser aplicada em qualquer negócio, e sua execução não ser dependente de ferramentas (CRISP-DM 2006). As fases que compõem o CRISP-DM são descritas abaixo e exemplificadas na figura 1 :

- Entendimento do negócio: foca em entender o objetivo do projeto a partir de uma perspectiva de negócios, definindo um plano preliminar para atingir os objetivos. Pode ser subdividido em três atividades:
 - Definir objetivos: Descrever seu objetivo principal a partir de uma perspectiva de negócios.
 - Produzir plano de projeto: Descrever o plano para atingir os objetivos de mineração de dados e negócios.
 - Critérios de sucesso nos negócios: Definir os critérios que serão usados para determinar se o projeto foi bem-sucedido do ponto de vista comercial
- Entendimento dos dados: Recolhimento de dados e início de atividades para familiarização com os dados, identificando problemas ou conjuntos interessantes. Pode ser subdividido em cinco atividades:
 - Coleta dos dados: Liste as fontes de dados adquiridas juntamente com seus locais, os métodos usados para adquiri-las e quaisquer problemas encontrados.
 - Descrição dos dados: Descrever os dados que foram adquiridos, incluindo seu formato, sua quantidade (por exemplo, o número de registros e campos em cada tabela)
 - Exploração dos dados: Descrever os resultados de sua exploração de dados, incluindo as primeiras descobertas ou hipóteses iniciais e seu impacto no restante do projeto.
 - Qualidade dos dados: Analisar a qualidade dos dados recolhidos.
- Preparação dos dados: Construção do conjunto de dados final a partir dos dados iniciais. Normalmente ocorre várias vezes no processo. Pode ser subdividido em cinco atividades:
 - Seleção dos dados: Liste os dados a serem incluídos / excluídos e os motivos dessas decisões.
 - Limpeza dos dados: Descrever quais decisões e ações você tomou para solucionar problemas de qualidade de dados.

- Construção dos dados: Descrever operações construtivas de preparação de dados, como a produção de atributos derivados ou novos registros inteiros ou valores transformados para atributos existentes.
- Integração dos dados: Integre fontes e armazene o resultado (novas tabelas e registros).
- Formatação dos dados: Organização e alterações na estrutura de dados para adequação ao método de data mining escolhido.
- Modelagem: Áreas técnicas de modelagem são aplicadas, e seus parâmetros calibrados para otimização. Assim, é comum retornar à Preparação dos Dados durante essa fase. Pode ser dividido em quatro atividades:
 - Seleção das técnicas de modelagem: Escolher e ajustar os parâmetros do algoritmo a ser utilizado.
 - Realização de testes de modelagem: Descrever o plano pretendido para treinar, testar e avaliar os modelos
 - Modelo de construção: Definir parâmetros iniciais e motivos do documento para escolher esses valores.
 - Avaliação do modelo: Avalie o resultado com relação aos critérios de avaliação. Classifique os resultados com relação aos critérios de sucesso e avaliação e selecione os melhores modelos.
- Avaliação: É construído um modelo que parece ter grande qualidade de uma perspectiva de análise de dados. No entanto, é necessário verificar se o modelo atinge os objetivos do negócio. Pode ser dividido em quatro atividades:
 - Avalie os resultados: entenda o resultado da mineração de dados. Verifique o impacto da meta de mineração de dados.
 - Revisão do processo: Resuma a revisão do processo (atividades que foram perdidas ou devem ser repetidas)
 - Determine as próximas etapas: Analise o potencial de implantação de cada resultado. Estimar o potencial de melhoria do processo atual.
 - Decisão: De acordo com os resultados e a revisão do processo, é decidido como proceder para a próxima etapa (recursos e orçamento restantes)
- Implantação: O conhecimento adquirido pelo modelo é organizado e apresentado de uma maneira que o cliente possa utilizar.
 - Planejar a implantação: Identifique possíveis problemas ao: implantar os resultados da mineração de dados
 - Planejar monitoramento e manutenção: O que poderia mudar no ambiente? Como a precisão será monitorada?
 - Produzir um relatório final: Identificar os relatórios necessários (apresentação de slides, resumo de gerenciamento, descobertas detalhadas, explicação de modelos etc.). Quão bem os objetivos iniciais de mineração de dados foram alcançados.
 - Revisão do projeto: Entrevistar as pessoas envolvidas no projeto. Entreviste usuários finais. O que poderia ter sido feito melhor? Eles precisam de suporte adicional? Resuma o feedback e escreva a documentação da experiência

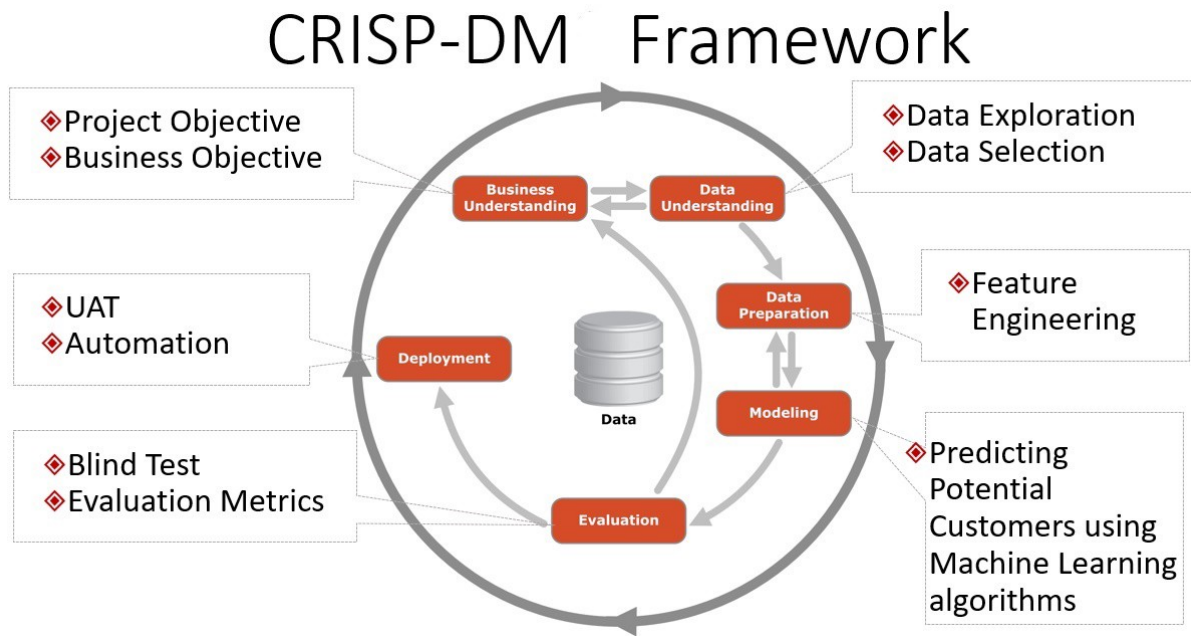


Figure 1: Diagrama da CRISP-DM Bharat (2019)

3 Fase 1 - Entendimento do Negócio

3.1 O que é o Sistema Nacional de Pós-Graduação? (Contextualização)

O Sistema de Avaliação da Pós-graduação foi implantado pela CAPES em 1976 e desde então vem cumprindo papel de fundamental importância para o desenvolvimento da pós-graduação e da pesquisa científica e tecnológica no Brasil. Abrange dois processos conduzidos por comissões de consultores do mais alto nível, vinculados a instituições de ensino das diferentes regiões do país: a Avaliação das Propostas de Cursos Novos e a Avaliação dos Programas de Pós-graduação (Sucupira 2006).

A Avaliação das Propostas de Cursos Novos é parte do rito estabelecido para a admissão de novos programas e cursos como integrantes do Sistema Nacional de Pós-graduação, SNPG. Ao avaliar as propostas de cursos novos, a CAPES verifica a qualidade de tais propostas e se elas atendem ao padrão de qualidade requerido desse nível de formação. Os resultados desse processo são encaminhados ao Conselho Nacional de Educação para fundamentar a deliberação desse órgão sobre o reconhecimento dos novos cursos.

A Avaliação dos Programas de Pós-graduação compreende os processos de Acompanhamento Anual e de Avaliação Trienal do desempenho dos programas e cursos que integram o Sistema Nacional de Pós-graduação, SNPG.

O Acompanhamento Anual é realizado no período compreendido entre os anos de realização das avaliações trienais. Tem por objetivo o estabelecimento de um diálogo entre a CAPES e as instituições promotoras de cursos de mestrado e doutorado com vistas à orientação da atuação dos programas de forma que possam elevar a qualidade de seu desempenho e superar os problemas que eventualmente estejam a enfrentar - se possível antes da Avaliação Trienal subsequente. O Acompanhamento não implica na atribuição de conceitos aos programas, mas apenas na apresentação de um parecer com os comentários considerados pertinentes pela Comissão de Área, e não enseja que seus resultados sejam contestados mediante a apresentação de recursos ou pedidos de reconsideração.

A Avaliação Trienal é realizada ao final de cada triênio, sendo o ano de sua realização estabelecido pela sequência histórica do processo de avaliação da CAPES. Os resultados da avaliação de cada programa

são apresentados na “Ficha de Avaliação” definida pelo CTC, de que constam, no que se refere aos vários quesitos e itens avaliados, os atributos a ele consignados, com os respectivos comentários e justificativas da comissão avaliadora, e, ao final, o conceito correspondente ao seu desempenho no triênio, na escala de 1 a 7 adotada. Tais resultados podem ser contestados pelas instituições de ensino mediante a apresentação de recurso contra a decisão inicial comunicada pela CAPES e, uma vez homologados pelo Ministro da Educação, são válidos até a homologação dos resultados da Avaliação Trienal subsequente. Os resultados da Avaliação Trienal realizada pela CAPES, além de indicarem a qualidade do desempenho e a posição relativa de cada programa no contexto de sua respectiva área, servem de referência para as decisões dos órgãos governamentais de investimento na pesquisa e na pós-graduação e fundamentam as deliberações do Conselho Nacional de Educação sobre quais cursos de mestrado e de doutorado obterão, para vigência no triênio seguinte, a renovação de seu “reconhecimento”.

3.2 A UnB dentro do Sistema Nacional de Pós-Graduação (Contextualização)

Historicamente, o desenvolvimento da ciência na UnB é realizado nas unidades acadêmicas com apoio, acompanhamento e supervisão do Decanato de Pesquisa e Pós-Graduação (DPP). A atuação do decanato promove todas as áreas do conhecimento com o auxílio de diretorias específicas para pesquisa, desenvolvimento institucional e inovação e iniciação científica (UnB 2006).

Para estimular a pesquisa e a inovação e tornar a Universidade de Brasília uma referência na área, foi criado, no início de 2017, o Decanato de Pesquisa e Inovação (DPI). Com a mudança, o DPP passou a se chamar Decanato de Pós-Graduação (DPG). A intenção é que as pró-reitorias somem expertises em seus respectivos setores de atuação e possam conduzir a UnB adiante na produção científica.

Editais próprios e de agências de fomento como CNPq, Capes e FAPDF são responsáveis por financiar parte significativa das pesquisas na Universidade. O DPG anuncia em sua página as publicações internas e as principais chamadas públicas que podem beneficiar a comunidade científica com a concessão de aportes financeiros, equipamentos, bolsas e realização e participação em eventos científicos.

Algumas das principais definições estratégicas para o progresso da ciência na UnB são realizadas no Conselho de Ensino, Pesquisa e Extensão (Cepe), com apoio da Câmara de Pesquisa e Pós-Graduação (CPP). Essas estruturas colegiadas permitem decisões democráticas com participação ativa dos segmentos interessados.

De acordo com a plataforma sucupira, existem 97 programas em funcionamento na Universidade de Brasília no momento. As notas por programa se separam da seguinte maneira:

- 5 programas com nota 7: Antropologia, Desenvolvimento Sustentável, Geologia, Matemática e Sociologia.
- 10 programas com nota 6: Geotecnia, Ciência Política, Ciências Biológicas (Biologia Molecular), Direito e outros.
- 18 programas com nota 5: Administração, Administração (pública), Bioética, Ciências animais, Ciências da informação, Ciências da saúde e outros.
- 46 programas com nota 4: Ensino de Ciências Ambientais em Rede Nacional, Agronegócios, Agronomia, Arquitetura E Urbanismo, Artes, Artes Cênicas, Biologia Animal, Biologia Microbiana e outros.

4 Fase 2 - Entendimento dos Dados

4.1 Coleta inicial dos dados

Os arquivos json para análise foram fornecidos na plataforma unb.elattes da UnB, que disponibiliza de forma acessível informações relevantes dos programas avaliados. Os dados fornecem informações entre os anos de 2010 e 2019 (Fernandes and Sampaio 2011).

Perfil profissional dos docentes vinculados às pós-graduações

```
file.info("Geotecnia/profile.json")
```

O arquivo profile.json apresenta dados sobre o perfil de todos os docentes vinculados a programas de pós-graduação da UnB, entre 2010 e 2019. Esse arquivo foi fornecido pelos docentes responsáveis pela disciplina.

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

```
file.info("Geotecnia/advise.json")
```

O arquivo advise.json apresenta dados sobre as orientações de mestrado e doutorado feitas por todos os docentes vinculados a programas de pós-graduação da UnB, entre 2010 e 2019. Esse arquivo foi fornecido pelos docentes responsáveis pela disciplina.

Produção bibliográfica gerada pelos docentes vinculados às pós-graduações

```
file.info("Geotecnia/publication.json")
```

O arquivo publication.json apresenta dados sobre a produção bibliográfica gerada por todos os docentes vinculados a programas de pós-graduação da UnB, entre 2010 e 2019.

Agrupamento dos docentes conforme áreas de atuação

```
file.info("Geotecnia/graph.json")
```

O arquivo graph.json de Geotecnia apresenta redes de colaboração na co-autoria de artigos científicos, feitas entre os docentes vinculados a programas de pós-graduação da UnB, entre 2010 e 2019.

4.2 Descrição dos Dados

Para ler, manipular, analisar e visualizar estes dados, serão utilizadas as seguintes bibliotecas:

```
library(tidyverse)
library(jsonlite)
library(listviewer)
library(scales)
library(dplyr)
library(readxl)
library(readr)
library(readtext)
library(ggplot2)
library(igraph)
library(knitr)
```


Com essas bibliotecas habilitadas será possível de responder e determinar qual o volume de dados, a estrutura dos dados (tipos), codificações usadas, entre outras atividades importantes para análise dos dados.

Descrição dos dados do perfil

O arquivo profile.json, que contém dados que caracterizam o perfil profissional de todos os docentes do grupo sob análise, podem ser lido por meio do comando seguinte:

```
profile <- fromJSON("Geotecnia/profile.json")
```

A quantidade de docentes de Geotecnia sob análise é apresentada a seguir.

```
length(profile)
```

```
## [1] 12
```

Para apresentar os dados que estão contido nos dados de perfil dos docentes, vamos usar a função glimpse, da biblioteca dplyr, como ilustra o código seguinte, que apresenta os atributos típicos que podem ser obtidos relativamente a um pesquisador específico, o mais antigo docente ainda em exercício na UnB a ter criado seu registro na plataforma unb.elattes.

```
glimpse(profile[[1]], width = 30)
```

```
## List of 7
## $ nome           : chr "André Luís Brasil Cavalcante"
## $ resumo_cv       : chr "André L. Brasil Cavalcante graduou-se em Engenharia Civil pela Univer
## $ areas_de_atuacao : 'data.frame':  4 obs. of  4 variables:
## ..$ grande_area  : chr [1:4] "ENGENHARIAS" "ENGENHARIAS" "ENGENHARIAS" "ENGENHARIAS"
## ..$ area         : chr [1:4] "Engenharia Civil" "Engenharia Civil" "Engenharia Civil" "Engenharia Civil"
## ..$ sub_area     : chr [1:4] "Geotécnica" "Geotécnica" "Geotécnica" "Fenômenos de Transporte"
## ..$ especialidade: chr [1:4] "Geotecnia Ambiental e Mineração" "Remediação de Áreas Contaminadas"
## $ endereco_profissional :List of 8
## ..$ instituicao: chr "Universidade de Brasília"
## ..$ orgao      : chr "ENC/FT/UNB"
## ..$ unidade    : chr ""
## ..$ DDD        : chr "61"
## ..$ telefone   : chr "31071269"
## ..$ bairro     : chr "ASA NORTE"
## ..$ cep        : chr "70910900"
## ..$ cidade     : chr "Brasília"
## $ producao_bibliografica :List of 6
## ..$ ARTIGO_ACEITO           : 'data.frame':  2 obs. of  10 variables:
## .. ..$ natureza            : chr [1:2] "NAO_INFORMADO" "NAO_INFORMADO"
## .. ..$ titulo              : chr [1:2] "The Iota-Delta Function as an Alternative to Boolean Formalism
## .. ..$ periodico           : chr [1:2] "INTERNATIONAL JOURNAL OF FOUNDATIONS OF COMPUTER SCIENCE" "JOU
## .. ..$ ano                 : chr [1:2] "2018" "2018"
## .. ..$ volume              : chr [1:2] "" ""
## .. ..$ issn                : chr [1:2] "01290541" "00220000"
## .. ..$ paginas             : chr [1:2] " - " " - "
## .. ..$ doi                 : chr [1:2] "" ""
## .. ..$ autores             :List of 2
## .. ..$ autores-endogeno :List of 2
## ..$ CAPITULO_DE_LIVRO      : 'data.frame':  5 obs. of  13 variables:
```

```

## ..$ tipo : chr [1:5] "Capítulo de livro publicado" "Capítulo de livro publica
## ..$ titulo_do_capitulo : chr [1:5] "Modelos teóricos de infiltração em meios porosos: equa
## ..$ titulo_do_livro : chr [1:5] "Tópicos sobre infiltração: teoria e prática aplicadas a
## ..$ ano : chr [1:5] "2012" "2013" "2014" "2015" ...
## ..$ doi : chr [1:5] "" "" "10.1201/b17017-190" "" ...
## ..$ pais_de_publicacao : chr [1:5] "Brasil" "França" "Brasil" "Brasil" ...
## ..$ isbn : chr [1:5] "9788560313419" "9782859784775" "9781138001466" "978856
## ..$ nome_da_editora : chr [1:5] "José Camapum de Carvalho, Gilson de Farias Neves Gitir
## ..$ numero_da_edicao_revisao: chr [1:5] "1" "1" "1" "1" ...
## ..$ organizadores : chr [1:5] "José Camapum de Carvalho; Gilson de Farias Neves Gitir
## ..$ paginas : chr [1:5] "249 - 268" "2807 - 2810" "1073 - 1077" "531 - 553" ...
## ..$ autores :List of 5
## ..$ autores-endogeno :List of 5
## ..$ DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA:'data.frame': 1 obs. of 9 variables:
## ..$ natureza : chr "Artigo Completo Publicado no Periódico arXiv"
## ..$ titulo : chr "Continuum versus Discrete: A Physically Interpretable General Rule
## ..$ ano : chr "2012"
## ..$ pais_de_publicacao: chr "Estados Unidos"
## ..$ editora : chr "arXiv"
## ..$ doi : chr ""
## ..$ numero_de_paginas : chr "37"
## ..$ autores :List of 1
## ..$ autores-endogeno :List of 1
## ..$ EVENTO : 'data.frame': 68 obs. of 11 variables:
## ..$ natureza : chr [1:68] "COMPLETO" "RESUMO" "COMPLETO" "COMPLETO" ...
## ..$ titulo : chr [1:68] "Cálculo Probabilístico do Fator de Segurança em Projetos de Ba
## ..$ nome_do_evento : chr [1:68] "XII Congresso Nacional de Geotecnia" "VI Simpósio Inovação no
## ..$ ano_do_trabalho : chr [1:68] "2010" "2010" "2010" "2010" ...
## ..$ pais_do_evento : chr [1:68] "Portugal" "Brasil" "Brasil" "Brasil" ...
## ..$ cidade_do_evento: chr [1:68] "Guimarães" "Brasília" "Gramado, RS" "Gramado, RS" ...
## ..$ doi : chr [1:68] "" "" "" "" ...
## ..$ classificacao : chr [1:68] "NACIONAL" "REGIONAL" "NACIONAL" "NACIONAL" ...
## ..$ paginas : chr [1:68] "1 - 8" "1 - 1" "1 - 8" "1 - 7" ...
## ..$ autores :List of 68
## ..$ autores-endogeno:List of 68
## ..$ LIVRO : 'data.frame': 1 obs. of 13 variables:
## ..$ titulo : chr "Anais do Simpósio de Prática de Engenharia Geotécnica na Regi
## ..$ ano : chr "2017"
## ..$ tipo : chr "LIVRO_ORGANIZADO_OU_EDICAO"
## ..$ natureza : chr "ANAIS"
## ..$ pais_de_publicacao : chr "Brasil"
## ..$ isbn : chr "9788567950051"
## ..$ doi : chr ""
## ..$ nome_da_editora : chr "ABMS"
## ..$ numero_da_edicao_revisao: chr "1"
## ..$ numero_de_paginas : chr "854"
## ..$ numero_de_volumes : chr "1"
## ..$ autores :List of 1
## ..$ autores-endogeno :List of 1
## ..$ PERIODICO : 'data.frame': 31 obs. of 10 variables:
## ..$ natureza : chr [1:31] "COMPLETO" "COMPLETO" "COMPLETO" "COMPLETO" ...
## ..$ titulo : chr [1:31] "Lagrange's Inversion Theorem and Infiltration" "On Modelling (
## ..$ periodico : chr [1:31] "World Academy of Science, Engineering and Technology (Online)
## ..$ ano : chr [1:31] "2012" "2012" "2012" "2012" ...

```

```

## .. ..$ volume      : chr [1:31] "6" "1" "01" "1" ...
## .. ..$ issn         : chr [1:31] "20103778" "22518843" "22518843" "22518843" ...
## .. ..$ paginas      : chr [1:31] "388 - 393" "11 - 16" "11 - 16" "64 - 70" ...
## .. ..$ doi          : chr [1:31] "" "" "" "" ...
## .. ..$ autores      :List of 31
## .. ..$ autores-endogeno:List of 31
## $ orientacoes_academicas:List of 8
## ..$ ORIENTACAO_CONCLUIDA_DOUTORADO      :'data.frame': 5 obs. of 13 variables:
## .. ..$ natureza      : chr [1:5] "Tese de doutorado" "Tese de doutorado" "Tese de dou
## .. ..$ titulo        : chr [1:5] "Concepção e Validação de um Modelo Matemático-digit
## .. ..$ ano           : chr [1:5] "2014" "2014" "2016" "2016" ...
## .. ..$ id_lattes_aluno : chr [1:5] "1853213451758913" "4994189165685449" "5372672441111
## .. ..$ nome_aluno     : chr [1:5] "Luan Carlos de S.M. Ozelim" "Leonardo Ramos da Silv
## .. ..$ instituicao     : chr [1:5] "Universidade de Brasília" "Universidade de Brasília
## .. ..$ curso          : chr [1:5] "Geotecnia" "Geotecnia" "Geotecnia" "Geotecnia" ...
## .. ..$ codigo_do_curso : chr [1:5] "51500329" "51500329" "51500329" "51500329" ...
## .. ..$ bolsa         : chr [1:5] "SIM" "SIM" "SIM" "SIM" ...
## .. ..$ agencia_financiadora : chr [1:5] "Conselho Nacional de Desenvolvimento Científico e T
## .. ..$ codigo_agencia_financiadora: chr [1:5] "002200000000" "002200000000" "045000000000" "002200
## .. ..$ nome_orientadores :List of 5
## .. ..$ id_lattes_orientadores :List of 5
## ..$ ORIENTACAO_CONCLUIDA_MESTRADO      :'data.frame': 15 obs. of 13 variables:
## .. ..$ natureza      : chr [1:15] "Dissertação de mestrado" "Dissertação de mestrado"
## .. ..$ titulo        : chr [1:15] "Modelagem Multidimensional de Transporte de Contam
## .. ..$ ano           : chr [1:15] "2011" "2011" "2013" "2013" ...
## .. ..$ id_lattes_aluno : chr [1:15] "" "3289800404391423" "" "1811065164099319" ...
## .. ..$ nome_aluno     : chr [1:15] "Juan Fernando Díaz-Sánchez" "Renata Conciani" "Orl
## .. ..$ instituicao     : chr [1:15] "Universidade de Brasília" "Universidade de Brasília
## .. ..$ curso          : chr [1:15] "Geotecnia" "Geotecnia" "Geotecnia" "Geotecnia" ...
## .. ..$ codigo_do_curso : chr [1:15] "51500329" "51500329" "51500329" "51500329" ...
## .. ..$ bolsa         : chr [1:15] "SIM" "SIM" "SIM" "SIM" ...
## .. ..$ agencia_financiadora : chr [1:15] "Conselho Nacional de Desenvolvimento Científico e
## .. ..$ codigo_agencia_financiadora: chr [1:15] "002200000000" "002200000000" "002200000000" "002200
## .. ..$ nome_orientadores :List of 15
## .. ..$ id_lattes_orientadores :List of 15
## ..$ ORIENTACAO_CONCLUIDA_POS_DOUTORADO :'data.frame': 1 obs. of 13 variables:
## .. ..$ natureza      : chr "Supervisão de pós-doutorado"
## .. ..$ titulo        : chr ""
## .. ..$ ano           : chr "2017"
## .. ..$ id_lattes_aluno : chr ""
## .. ..$ nome_aluno     : chr "Luan Carlos de Sena Monteiro Ozelim"
## .. ..$ instituicao     : chr "Universidade de Brasília"
## .. ..$ curso          : chr ""
## .. ..$ codigo_do_curso : chr ""
## .. ..$ bolsa         : chr "SIM"
## .. ..$ agencia_financiadora : chr "Conselho Nacional de Desenvolvimento Científico e Tecnolô
## .. ..$ codigo_agencia_financiadora: chr "002200000000"
## .. ..$ nome_orientadores :List of 1
## .. ..$ id_lattes_orientadores :List of 1
## ..$ ORIENTACAO_EM_ANDAMENTO_DOUTORADO :'data.frame': 6 obs. of 13 variables:
## .. ..$ natureza      : chr [1:6] "Tese de doutorado" "Tese de doutorado" "Tese de dou
## .. ..$ titulo        : chr [1:6] "Avaliação do Fluxo de Contaminantes em Sistemas de L
## .. ..$ ano           : chr [1:6] "2014" "2016" "2017" "2017" ...
## .. ..$ id_lattes_aluno : chr [1:6] "" "9492644429017937" "6392971663534518" "" ...

```

```

## ..$ nome_aluno : chr [1:6] "Silvana Fava Marchezini" "Lucas Parreira de Faria B
## ..$ instituicao : chr [1:6] "Universidade de Brasília" "Universidade de Brasília
## ..$ curso : chr [1:6] "Geotecnia" "Geotecnia" "Geotecnia" "Geotecnia" ...
## ..$ codigo_do_curso : chr [1:6] "51500329" "51500329" "51500329" "51500329" ...
## ..$ bolsa : chr [1:6] "SIM" "SIM" "SIM" "SIM" ...
## ..$ agencia_financiadora : chr [1:6] "Conselho Nacional de Desenvolvimento Científico e T
## ..$ codigo_agencia_financiadora: chr [1:6] "002200000000" "002200000000" "002200000000" "002200
## ..$ nome_orientadores :List of 6
## ..$ id_lattes_orientadores :List of 6
## ..$ ORIENTACAO_EM_ANDAMENTO_GRADUACAO :'data.frame': 2 obs. of 13 variables:
## ..$ natureza : chr [1:2] "Trabalho de conclusão de curso de graduação" "Traba
## ..$ titulo : chr [1:2] "Experimento de Fluxo em Meio Poroso Artificial Gera
## ..$ ano : chr [1:2] "2017" "2017"
## ..$ id_lattes_aluno : chr [1:2] "" ""
## ..$ nome_aluno : chr [1:2] "Nicholas Veres Barros" "Dhara Vieira Alcântara"
## ..$ instituicao : chr [1:2] "Universidade de Brasília" "Universidade de Brasília
## ..$ curso : chr [1:2] "Engenharia Civil" "Engenharia Civil"
## ..$ codigo_do_curso : chr [1:2] "60070293" "60070293"
## ..$ bolsa : chr [1:2] "NAO" "NAO"
## ..$ agencia_financiadora : chr [1:2] "" ""
## ..$ codigo_agencia_financiadora: chr [1:2] "" ""
## ..$ nome_orientadores :List of 2
## ..$ id_lattes_orientadores :List of 2
## ..$ ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA:'data.frame': 6 obs. of 13 variables:
## ..$ natureza : chr [1:6] "Iniciação Científica" "Iniciação Científica" "Inici
## ..$ titulo : chr [1:6] "Fluxo de Água em Meio Poroso Artificial Gerado por
## ..$ ano : chr [1:6] "2017" "2017" "2017" "2017" ...
## ..$ id_lattes_aluno : chr [1:6] "" "" "" "" ...
## ..$ nome_aluno : chr [1:6] "Nicholas Veres Barros" "Thália Raelly de Lima Menes
## ..$ instituicao : chr [1:6] "Universidade de Brasília" "Universidade de Brasília
## ..$ curso : chr [1:6] "Engenharia Civil" "Engenharia Civil" "Engenharia Ci
## ..$ codigo_do_curso : chr [1:6] "60070293" "60070293" "60070293" "60070293" ...
## ..$ bolsa : chr [1:6] "SIM" "NAO" "SIM" "NAO" ...
## ..$ agencia_financiadora : chr [1:6] "Fundação de Apoio à Pesquisa do Distrito Federal" "
## ..$ codigo_agencia_financiadora: chr [1:6] "786500000001" "" "002200000000" "" ...
## ..$ nome_orientadores :List of 6
## ..$ id_lattes_orientadores :List of 6
## ..$ ORIENTACAO_EM_ANDAMENTO_MESTRADO :'data.frame': 3 obs. of 13 variables:
## ..$ natureza : chr [1:3] "Dissertação de mestrado" "Dissertação de mestrado"
## ..$ titulo : chr [1:3] "Modelagem fracionária de fluxo de água em meio poroso
## ..$ ano : chr [1:3] "2017" "2018" "2018"
## ..$ id_lattes_aluno : chr [1:3] "" "" ""
## ..$ nome_aluno : chr [1:3] "Pedro Victor Serra Mascarenhas" "Mariana dos Santos
## ..$ instituicao : chr [1:3] "Universidade de Brasília" "Universidade de Brasília
## ..$ curso : chr [1:3] "Geotecnia" "Geotecnia" "Geotecnia"
## ..$ codigo_do_curso : chr [1:3] "51500329" "51500329" "51500329"
## ..$ bolsa : chr [1:3] "SIM" "SIM" "SIM"
## ..$ agencia_financiadora : chr [1:3] "Conselho Nacional de Desenvolvimento Científico e T
## ..$ codigo_agencia_financiadora: chr [1:3] "002200000000" "002200000000" "002200000000"
## ..$ nome_orientadores :List of 3
## ..$ id_lattes_orientadores :List of 3
## ..$ OUTRAS_ORIENTACOES_CONCLUIDAS :'data.frame': 57 obs. of 13 variables:
## ..$ natureza : chr [1:57] "INICIACAO_CIENTIFICA" "INICIACAO_CIENTIFICA" "INIC
## ..$ titulo : chr [1:57] "Ensaio laboratorial em geotecnia e aplicações da

```

```
## .. ..$ ano : chr [1:57] "2010" "2010" "2010" "2010" ...
## .. ..$ id_lattes_aluno : chr [1:57] "" "" "" "" ...
## .. ..$ nome_aluno : chr [1:57] "Pedro Henrique Lopes Batista" "Cristiano Nascimento" ...
## .. ..$ instituicao : chr [1:57] "Universidade Católica de Brasília" "Universidade Católica de Brasília" ...
## .. ..$ curso : chr [1:57] "Engenharia Civil" "Engenharia Civil" "Engenharia Civil" ...
## .. ..$ codigo_do_curso : chr [1:57] "90000020" "90000020" "90000020" "90000022" ...
## .. ..$ bolsa : chr [1:57] "SIM" "SIM" "NAO" "NAO" ...
## .. ..$ agencia_financiadora : chr [1:57] "Universidade Católica de Brasília" "Universidade Católica de Brasília" ...
## .. ..$ codigo_agencia_financiadora: chr [1:57] "001000000998" "001000000998" "" "" ...
## .. ..$ nome_orientadores :List of 57
## .. ..$ id_lattes_orientadores :List of 57
## $ senioridade : chr "8"
```

Uma breve inspeção visual dos atributos anteriormente apresentados permite inferir que o pesquisador sob análise pode ser inferir que o professor da área da engenharia por formação, tendo geotécnica e fenômenos de transporte como sua subárea, especialidades em Geotecnia Ambiental e Mineração e Remediação de Áreas Contaminadas e atualmente trabalha no campus Darcy Ribeiro da UnB com uma senioridade de 8.

Descrição dos dados de orientações

Carregando os dados do arquivo de orientação de geotecnia, podemos as seguintes descrições:

```
advise <- fromJSON("Geotecnia/advise.json")
```

Os tipos de orientação que estarão sob análise:

```
names(advise)
```

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

O período no qual será feita a análise:

```
names(advise$ORIENTACAO_CONCLUIDA_DOUTORADO)
```

```
## [1] "2010" "2011" "2012" "2013" "2014" "2015" "2016" "2017" "2018" "2019"
```

A quantidade de orientações de doutorado concluídas em um ano específico (ex. 2016)

```
length(advise$ORIENTACAO_CONCLUIDA_DOUTORADO$`2016`$natureza)
```

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

Quais os cursos que mais contribuíram com teses concluídas em doutorado em um ano específico (ex. 2017):

```
head(sort(
  table(advise$ORIENTACAO_CONCLUIDA_DOUTORADO$`2017`$curso),
  decreasing = TRUE
), 10)
```

```
## Geotecnia
##          11
```

Quais os cursos que mais contribuíram com teses concluídas em mestrado em um ano específico (ex. 2017):

```
head(sort(
  table(advise$ORIENTACAO_CONCLUIDA_MESTRADO$`2017`$curso),
  decreasing = TRUE
), 10)
```

```
##
##                               Geotecnia
##                               13
##      Programa de Pós-Graduação em Geotecnia
##                               2
##                               Civil Engineering
##                               1
##                               Mestrado em Geotecnia
##                               1
## Mestrado Profissional em Engenharia Ambiental
##                               1
```

Descrição dos dados de produção bibliográfica

Carregando os dados do arquivo de produção bibliográfica de Geotecnia, podemos as seguintes descrições:

```
publication <- fromJSON("Geotecnia/publication.json")
```

Os tipos de produção bibliográfica a serem analisados nessa pesquisa:

```
names(publication)
```

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

Os tipos de informação que cada periódico contém em um ano específico (ex. 2012)

```
names(publication$PERIODICO$`2012`)
```

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

Os periódicos com maior número de publicações para cada programa em um determinado ano (ex. 2017):

```
head(sort(table(publication$PERIODICO$`2017`$periodico), decreasing = TRUE), 10)
```

```
##
##           Computational Particle Mechanics
##                               3
##                DYNA (MEDELLÍN)
##                               2
##           ENVIRONMENTAL GEOTECHNICS
##                               2
## International Journal of Geomechanics
##                               2
##                SOILS & ROCKS
##                               2
##           Transportes (Rio de Janeiro)
##                               2
##           Boletim Goiano de Geografia
##                               1
##           CANADIAN GEOTECHNICAL JOURNAL
##                               1
##           Construção Magazine
##                               1
## Electronic Journal of Geotechnical Engineering
##                               1
```

As editoras com maior número de publicações para cada programa em um determinado ano (ex. 2015):

```
head(sort(
  table(publication$LIVRO$`2015`$nome_da_editora),
  decreasing = TRUE
), 10)
```

```
## ABMS
##      1
```

4.3 Análise exploratória dos dados

A seguir será mostrado a análise exploratória dos dados nos datasets um entendimento de qualidade mais profundo da relação estatística existente para os objetivos do projeto.

4.3.1 Arquivo Perfil

Número de áreas de atuação cumulativo:

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

```
## [1] 63
```

Número de pessoas por áreas de atuação:

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

```
##
## 3 4 5 6
## 1 2 2 7
```

Número de pessoas por grande área:

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

```
##
## CIENCIAS_EXATAS_E_DA_TERRA CIENCIAS_SOCIAIS_APLICADAS
##                               5                               2
##                               ENGENHARIAS
##                               56
```

Número de pessoas que produziram os específicos tipos de produção:

```
table(unlist(sapply(profile, function(x)
  names(x$producao_bibliografica))))
```

```
##
##                ARTIGO_ACEITO
##                    3
##            CAPITULO_DE_LIVRO
##                    9
##  DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA
##                    4
##                    EVENTO
##                   12
##                    LIVRO
##                    6
##                PERIODICO
##                   12
##            TEXTO_EM_JORNAIS
##                    5
```

Número de publicações por artigo aceito:


```
sum(sapply(profile, function(x)
  length(x$producao_bibliografica$ARTIGO_ACEITO$ano)))
```

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

Número de publicações por capítulo de livro:

```
sum(sapply(profile, function(x)
  length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano)))
```

```
## [1] 68
```

Número de publicações por livro:

```
sum(sapply(profile, function(x)
  length(x$producao_bibliografica$LIVRO$ano)))
```

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

Número de publicações por periódico:

```
sum(sapply(profile, function(x)
  length(x$producao_bibliografica$PERIODICO$ano)))
```

```
## [1] 201
```

Número de publicações por texto em jornais:

```
sum(sapply(profile, function(x)
  length(x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano)))
```

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

Número de pessoas por quantitativo de produções por artigo aceito:

```
table(unlist(sapply(profile, function(x)
  length(x$producao_bibliografica$ARTIGO_ACEITO$ano))))
```

```
##
## 0 1 2
## 9 2 1
```

Número de pessoas por quantitativo de produções por capítulo de livro:

```
table(unlist(sapply(profile, function(x)
  length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano))))
```

```
##
## 0 1 2 5 6 8 38
## 3 3 1 1 2 1 1
```

Número de pessoas por quantitativo de produções por livro:

```
table(unlist(sapply(profile, function(x)
  length(x$producao_bibliografica$LIVRO$ano))))
```

```
##
## 0 1 3 7
## 6 4 1 1
```

Número de pessoas por quantitativo de produções por periódico:

```
table(unlist(sapply(profile, function(x)
  length(x$producao_bibliografica$PERIODICO$ano))))
```

```
##
## 3 5 14 15 26 27 31 32
## 2 2 2 1 2 1 1 1
```

Número de pessoas por quantitativo de produções por texto em jornais:

```
table(unlist(sapply(profile, function(x)
  length(x$producao_bibliografica$TEXTO_EM_JORNAIS$ano))))
```

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

Número de produções de artigo aceito por ano:

```
table(unlist(sapply(profile, function(x)
  (x$producao_bibliografica$ARTIGO_ACEITO$ano))))
```

```
##
## 2014 2018
## 1 3
```

Número de produções de capítulo de livro por ano:

```
table(unlist(sapply(profile, function(x)
  (x$producao_bibliografica$CAPITULO_DE_LIVRO$ano))))
```

```
##
## 2012 2013 2014 2015 2016 2017
## 33 1 3 11 2 18
```

Número de produções de livro por ano:

```
table(unlist(sapply(profile, function(x)
  (x$producao_bibliografica$LIVRO$ano))))
```

```
##
## 2010 2011 2012 2014 2015 2016 2017
##    2    1    3    1    1    2    4
```

Número de produções de periódico por ano:

```
table(unlist(sapply(profile, function(x)
  (x$producao_bibliografica$PERIODICO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017 2018
##    20   11   21   27   27   32   30   26    7
```

Número de produções de texto em jornais por ano:

```
table(unlist(sapply(profile, function(x)
  (x$producao_bibliografica$TEXTOS_EM_JORNAIS$ano))))
```

```
##
## 2011 2012 2013 2014 2015 2016
##    2    2    3    1    1    5
```

Número de pessoas que realizaram diferentes tipos de orientações:

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

```
## [1] 67
```

Número de pessoas por tipo de orientação:

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

```
##
##          ORIENTACAO_CONCLUIDA_DOUTORADO
##                                11
##          ORIENTACAO_CONCLUIDA_MESTRADO
##                                12
##          ORIENTACAO_CONCLUIDA_POS_DOUTORADO
##                                4
##          ORIENTACAO_EM_ANDAMENTO_DOUTORADO
##                                11
##          ORIENTACAO_EM_ANDAMENTO_GRADUACAO
##                                1
## ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA
##                                6
##          ORIENTACAO_EM_ANDAMENTO_MESTRADO
##                                10
##          OUTRAS_ORIENTACOES_CONCLUIDAS
##                                12
```

Número de orientações concluídas em mestrado:

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

```
## [1] 153
```

Número de orientações concluídas em doutorado:

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

```
## [1] 89
```

Número de orientações concluídas em pós-doutorado:

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

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

Número de pessoas por quantitativo de orientações concluídas em mestrado:

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

```
##
##  6  7  8 10 11 12 13 15 19 25
##  1  1  1  1  1  2  1  2  1  1
```

Número de pessoas por quantitativo de orientações concluídas em doutorado:

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

```
##
##  0  1  5  7  8  9 10 11 12 13
##  1  1  2  1  2  1  1  1  1  1
```

Número de pessoas por quantitativo de orientações concluídas em pós-doutorado:

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

```
##
## 0 1 2
## 8 2 2
```

Número de orientações em mestrado por ano:

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

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017 2018
##    8   19   20   21   16   20   25   19    5
```

Número de orientações em doutorado por ano:

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

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##    5    8   11   12   13   14   14   12
```

Número de orientações em pós-doutorado por ano:

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

```
##
## 2010 2015 2016 2017
##    3    1    1    1
```

4.3.2 Arquivo Publicação

Criando um data-frame com todos os anos:

```
publication.df <- data.frame()
for (i in 1:length(publication[[1]]))
  publication.df <-
    rbind(publication.df, publication$PERIODICO[[i]])
glimpse(publication.df)
```

```
## Observations: 171
## Variables: 10
## $ natureza      <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo        <chr> "Metodologia para o controle de qualidade d...
## $ periodico     <chr> "Revista Fundações e Obras Geotécnicas", "R...
```

```
## $ ano          <chr> "2010", "2010", "2010", "2010", "2010", "20...
## $ volume       <chr> "1", "01", "15", "34", "4", "17", "28", "33...
## $ issn         <chr> "21780668", "21780668", "0120548X", "037039...
## $ paginas      <chr> "50 - 57", "50 - 57", "16880 - ", "209 - 22...
## $ doi          <chr> "", "", "", "", "", "10.1680/gein.2010.17.1...
## $ autores      <list> [<"SILVA, Carlos Medeiros", "CAMAPUM DE CA...
## $ `autores-endogeno` <list> [<"2245433059787601", "2245433059787601">,...
```

Limpando o data-frame de listas:

```
publication.df$autores <-
  gsub("\\", "\\|\\", "\\\"", "; ", publication.df$autores)
publication.df$autores <-
  gsub("\\|c\\(|\\)", "", publication.df$autores)
publication.df$`autores-endogeno` <-
  gsub(",", ";", publication.df$`autores-endogeno`)
publication.df$`autores-endogeno` <-
  gsub("\\|c\\(|\\)", "", publication.df$`autores-endogeno`)
glimpse(publication.df)
```

```
## Observations: 171
## Variables: 10
## $ natureza     <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo       <chr> "Metodologia para o controle de qualidade d...
## $ periodico    <chr> "Revista Fundações e Obras Geotécnicas", "R...
## $ ano          <chr> "2010", "2010", "2010", "2010", "2010", "20...
## $ volume       <chr> "1", "01", "15", "34", "4", "17", "28", "33...
## $ issn         <chr> "21780668", "21780668", "0120548X", "037039...
## $ paginas      <chr> "50 - 57", "50 - 57", "16880 - ", "209 - 22...
## $ doi          <chr> "", "", "", "", "", "10.1680/gein.2010.17.1...
## $ autores      <chr> "SILVA, Carlos Medeiros; CAMAPUM DE CARVALH...
## $ `autores-endogeno` <chr> "2245433059787601; 2245433059787601", "2245...
```

4.3.3 Arquivo Orientação

Reunir todos os anos e orientações concluídas em um mesmo data-frame:

```
advise.tipo.df <- data.frame()
advise.df <- data.frame()
for (i in 1:length(advise[[1]]))
  advise.tipo.df <-
    rbind(advise.tipo.df, advise$ORIENTACAO_CONCLUIDA_POS_DOUTORADO[[i]])
advise.df <-
  rbind(advise.df, advise.tipo.df)
advise.tipo.df <- data.frame()
for (i in 1:length(advise[[1]]))
  advise.tipo.df <-
    rbind(advise.tipo.df, advise$ORIENTACAO_CONCLUIDA_DOUTORADO[[i]])
advise.df <-
  rbind(advise.df, advise.tipo.df)
advise.tipo.df <- data.frame()
for (i in 1:length(advise[[1]]))
```

```

advise.tipo.df <-
  rbind(advise.tipo.df, advise$ORIENTACAO_CONCLUIDA_MESTRADO[[i]])
advise.df <- rbind(advise.df, advise.tipo.df)
glimpse(advise.df)

## Observations: 226
## Variables: 13
## $ natureza          <chr> "Supervisão de pós-doutorado", "Su...
## $ titulo            <chr> "ESTUDO DA MELHORIA DE UM SOLO TRO...
## $ ano               <chr> "2010", "2010", "2010", "2015", "2...
## $ id_lattes_aluno   <chr> "", "", "", "", "", "", "963332712...
## $ nome_aluno        <chr> "Yamile Valencia González", "Raul ...
## $ instituicao        <chr> "Universidade de Brasília", "Unive...
## $ curso             <chr> "", "", "", "", "", "", "Geotecnia...
## $ codigo_do_curso   <chr> "", "", "", "", "", "", "51500329"...
## $ bolsa             <chr> "SIM", "SIM", "SIM", "SIM", "SIM",...
## $ agencia_financiadora <chr> "Conselho Nacional de Desenvolvime...
## $ codigo_agencia_financiadora <chr> "002200000000", "002200000000", "0...
## $ nome_orientadores <list> ["Jose Camapum de Carvalho", "Jos...
## $ id_lattes_orientadores <list> ["2245433059787601", "22454330597...

```

Numero de orientações por ano:

```
table(advise.df$ano)
```

```

##
## 2010 2011 2012 2013 2014 2015 2016 2017 2018
##   13   25   26   32   26   29   40   30   5

```

Tabela com nome de professor e numero de orientações:

```

head(sort(table(rbind(
  advise.df$ori1, advise.df$ori2
)), decreasing = TRUE), 20)

```

```
## integer(0)
```

4.4 Verificação da qualidade dos dados

Com os dados coletados na seção 2.3 tem-se que os dados para serem submetidos a uma análise apresentam uma qualidade satisfatória para resolver os problemas propostos.

5 Fase 3 - Preparação dos Dados

Nessa fase são realizadas 5 atividades genéricas para de preparação dos dados

5.1 Seleção dos dados.

Para a principal utilização dos dados foram definidos os dataframes advise, que contém informações de qual a natureza da pesquisa produzida, qual aluno produziu, seus respectivos orientadores, o ano e etc. A seguir foi definido os dados importantes do arquivo profile.json, gerando o mais dataframes, chamados de profile, que já está limpo, com as principais colunas definidas.

5.2 Limpeza dos dados

Nessa etapa se anexa com a próxima etapa a construção dos dados. Durante a construção, foi realizada uma limpeza dos dados.

5.3 Construção dos dados

Na etapa construção dos dados é realizada criação de novas variáveis a partir de outras presentes nos datasets para análise.

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

```
## Observations: 12
## Variables: 15
## $ idLattes      <chr> "1515779118499986", "224543..."
## $ nome          <chr> "André Luís Brasil Cavalcant..."
```



```
## $ resumo_cv <chr> "André L. Brasil Cavalcante..."
## $ senioridade <chr> "8", "10", "10", "10", "8",...
## $ ORIENTACAO_CONCLUIDA_DOUTORADO <int> 5, 12, 5, 10, 7, 8, 13, 1, ...
## $ ORIENTACAO_CONCLUIDA_MESTRADO <int> 15, 10, 15, 19, 7, 13, 12, ...
## $ ORIENTACAO_CONCLUIDA_POS_DOUTORADO <int> 1, 2, NA, 2, 1, NA, NA, NA,...
## $ OUTRAS_ORIENTACOES_CONCLUIDAS <int> 57, 9, 3, 13, 24, 17, 12, 2...
## $ CAPITULO_DE_LIVRO <int> 5, 38, 6, 6, 1, 1, 2, 1, NA...
## $ EVENTO <int> 68, 91, 44, 55, 49, 34, 37,...
## $ LIVRO <int> 1, 7, 1, 3, NA, NA, NA, NA,...
## $ OUTRAS_PRODUCOES <int> 1, 1, 1, NA, NA, NA, NA, 5,...
## $ PERIODICO <int> 31, 26, 15, 27, 3, 14, 32, ...
## $ TEXTO_EM_JORNAIS <int> NA, 2, NA, 3, NA, NA, NA, 3...
## $ num_areas <int> 2, 2, 2, 1, 1, 2, 1, 1, 1, ...
```

5.4 Integração dos dados

Não havia necessidade de realizar integração entre os dados, somente realizada a construção dos dataframes como descrito na etapa 3.3.

5.5 Formatação dos dados

A formatação dos dados já foi realizada nas etapas anteriores. Para esses casos mais simples, a formatação será realizada conforme a necessidade.

6 Fase 4 - Modelagem

Com os dados preparados, na fase de modelagem várias técnicas de modelagem são selecionadas e aplicadas e seus parâmetros são calibrados. Usualmente, mais de uma técnica pode ser aplicado ao conjunto de dados disponível, ou uma técnica requer um formato específico dos dados, necessitando nova preparação dos dados. Uma estratégia comum é dividir o conjunto de dados, utilizando uma porção dos dados para o desenvolvimento do modelo e outra para o teste do modelo obtido. Em alguns casos, utiliza-se uma terceira porção dos dados para validação.

6.1 Análise de Redes

A Análise de Redes é a área de tecnologia da informação e das ciências sociais que trata do processo de analisar qualquer tipo de rede por meio da teoria das redes. As redes podem ser social, de transporte ou tecnológicas, como a internet.

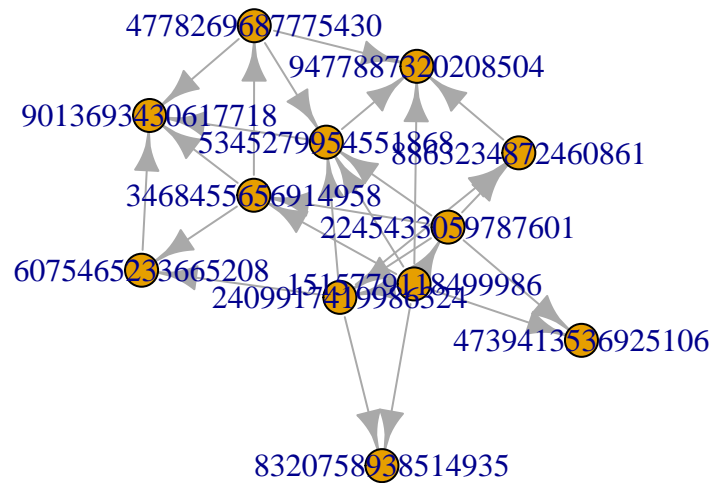
```
colnames(geotecnia_nodes) <- c("IdLattes",  
                                "Index",  
                                "Docente")  
kable(geotecnia_nodes[, c(2, 1, 3)], caption = "Tabela de referência da análise de redes")
```

Table 1: Tabela de referência da análise de redes

Index	IdLattes	Docente
1	1515779118499986	André Luís Brasil Cavalcante
2	2245433059787601	Jose Camapum de Carvalho
3	2409917419986324	Hernán Eduardo Martínez Carvajal
4	3468455656914958	Ennio Marques Palmeira
5	4739413536925106	Luis Fernando Martins Ribeiro
6	4778269687775430	Manoel Porfirio Cordão Neto
7	5345279954551868	Marcio Muniz de Farias
8	6075465233665208	Gregorio Luis Silva Araujo
9	8320758938514935	Juan Felix Rodriguez Rebolledo
10	8863234872460861	Newton Moreira de Souza
11	9013693430617718	Renato Pinto da Cunha
12	9477887320208504	André Pacheco de Assis

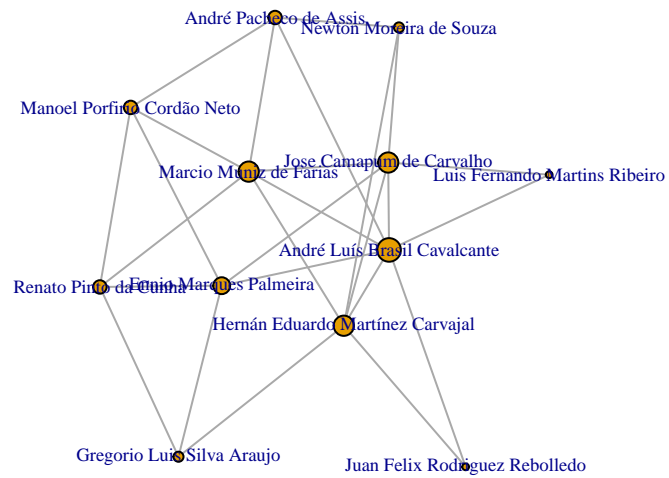
O grafo ilustra que todos os 12 pesquisadores trabalham em conjunto, nenhum trabalha sozinho

```
igraph <-  
  graph_from_data_frame(geotecnia_relations,  
                        directed = TRUE,  
                        vertices = geotecnia_nodes)  
plot(igraph)
```



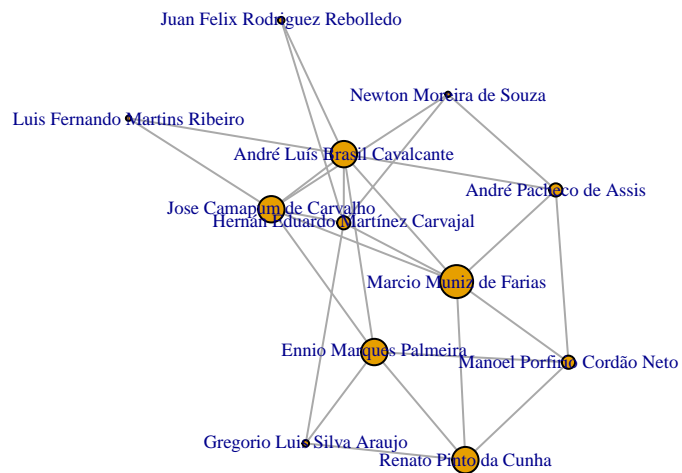
O grafo ilustra o número de colaborações de cada pesquisador pelo tamanho do nó relacionado a seu docente da tabela de autores:

```
plot(
  g,
  vertex.size = V(g)$degree * 1.5,
  vertex.label = geotecnia_nodes$label,
  layout = layout_nicely(g),
  vertex.label.cex = 0.6
)
```



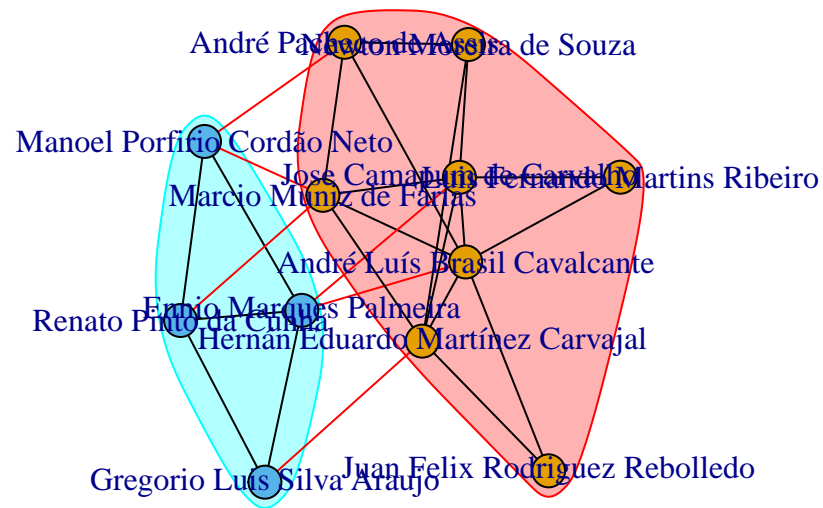
O grafo ilustra o número de publicações de cada autor pelo tamanho do nó relacionado a seu docente da tabela de autores:

```
plot(
  g,
  vertex.size = na.omit(round(V(g)$publicacao / 7) * 3),
  vertex.label = geotecnia_nodes$label,
  layout = layout_nicely(g),
  vertex.label.cex = 0.6
)
```



O grafo ilustra as comunidades entre os autores com agrupamentos no grafo por cores:

```
kc = fastgreedy.community(g)
plot(kc,
     g,
     vertex.label = geotecnia_nodes$label,
     layout = layout_nicely(g))
```



Podemos concluir com a análise das redes do programa de pós-graduação em Geotecnia é interessante notar facilmente há uma colaboração entre todos pesquisadores.

7 Fase 5 - Avaliação

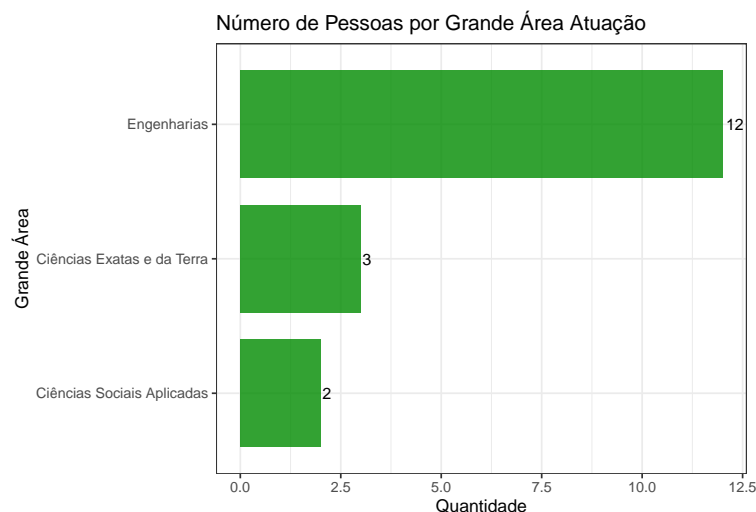
A fase de avaliação consiste na verificação de questões em relação aos datasets que ainda não foram abordadas o suficiente. Dessa forma, consiste em uma análise mais “profunda”, por assim dizer, dos dados em questão. A avaliação foi realizada de forma gráfica, sendo apresentados a seguir: ## Avaliação dos resultados

7.0.1 Arquivo Perfil

O gráfico ilustra a quantidade de professores por grande área de atuação (em forma de colunas):

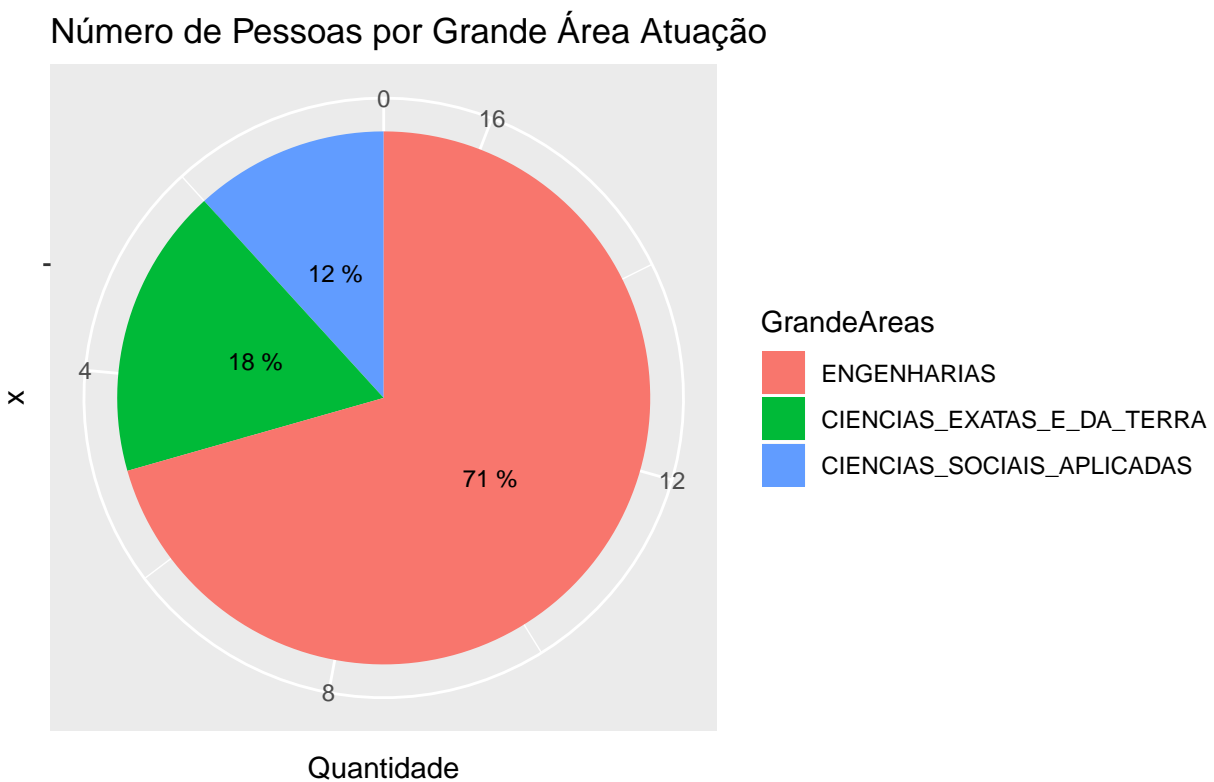
```
profile %>%
  sapply(function(x)
    unique(x$areas_de_atuacao$grande_area)) %>%
  unlist() %>% table() %>% sort() %>% as.data.frame() %>% filter(!. == "") %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "green4",
                                         alpha = 0.8,
                                         width = 0.8) + coord_flip() + geom_text(aes(label = Freq),
                                                                                       hjust = -0.2,
                                                                                       vjust = 0.5,
                                                                                       size = 3.5) +

  labs(title = "Número de Pessoas por Grande Área Atuação", y = "Quantidade", x =
        "Grande Área") + theme_bw() + scale_y_continuous() +
  scale_x_discrete(
    labels = c(
      'CIENCIAS_DA_SAUDE' = 'Ciências da Saúde',
      'CIENCIAS_BIOLOGICAS' = 'Ciências Biológicas',
      'CIENCIAS_HUMANAS' = 'Ciências Humanas',
      "CIENCIAS_EXATAS_E_DA_TERRA" = "Ciências Exatas e da Terra",
      "CIENCIAS_SOCIAIS_APLICADAS" = "Ciências Sociais Aplicadas",
      "CIENCIAS_AGRARIAS" = "Ciências Agrárias",
      "OUTROS" = "Outros",
      "ENGENHARIAS" = "Engenharias",
      "LINGUISTICA_LETRAS_E_ARTES" = "Linguística, Letras e Artes"
    )
  )
)
```



O gráfico ilustra a quantidade de professores por grande área de atuação (em forma de pizza):

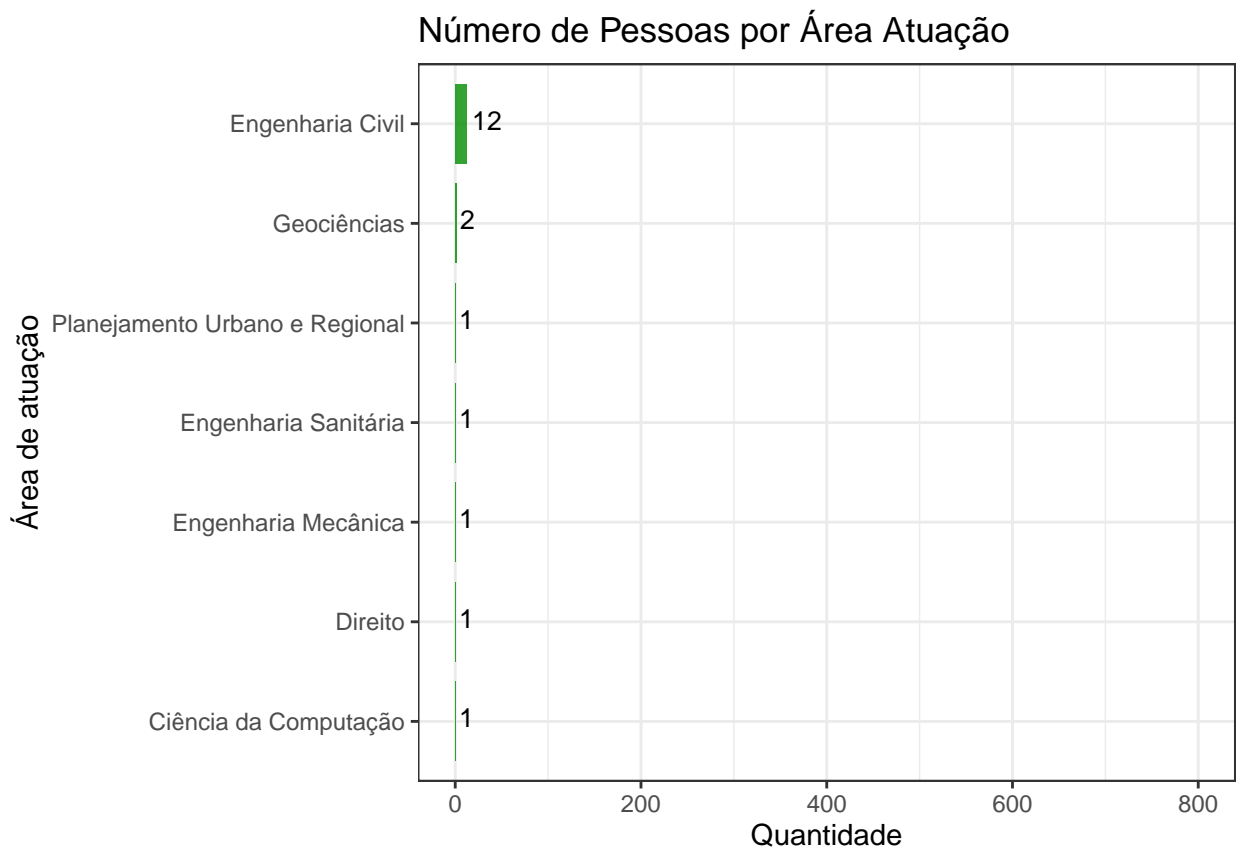
```
ggplot(grande_areas_atuacao, aes(x = "", y = Quantidade, fill = GrandeAreas)) +  
  geom_bar(width = 1, stat = "identity") +  
  coord_polar("y", start = 0, direction = -1) +  
  labs(title = "Número de Pessoas por Grande Área Atuação") +  
  geom_text(  
    data = grande_areas_atuacao,  
    aes(  
      x = "",  
      y = Quantidade,  
      label = paste(Porcentagem, "%")  
    ),  
    position = position_stack(vjust = 0.5),  
    size = 3  
  )  
)
```



Com os graficos podemos observar que todos os 12 professores são da grande área de engenharias e apenas 29% dos professores são de outras grandes áreas.

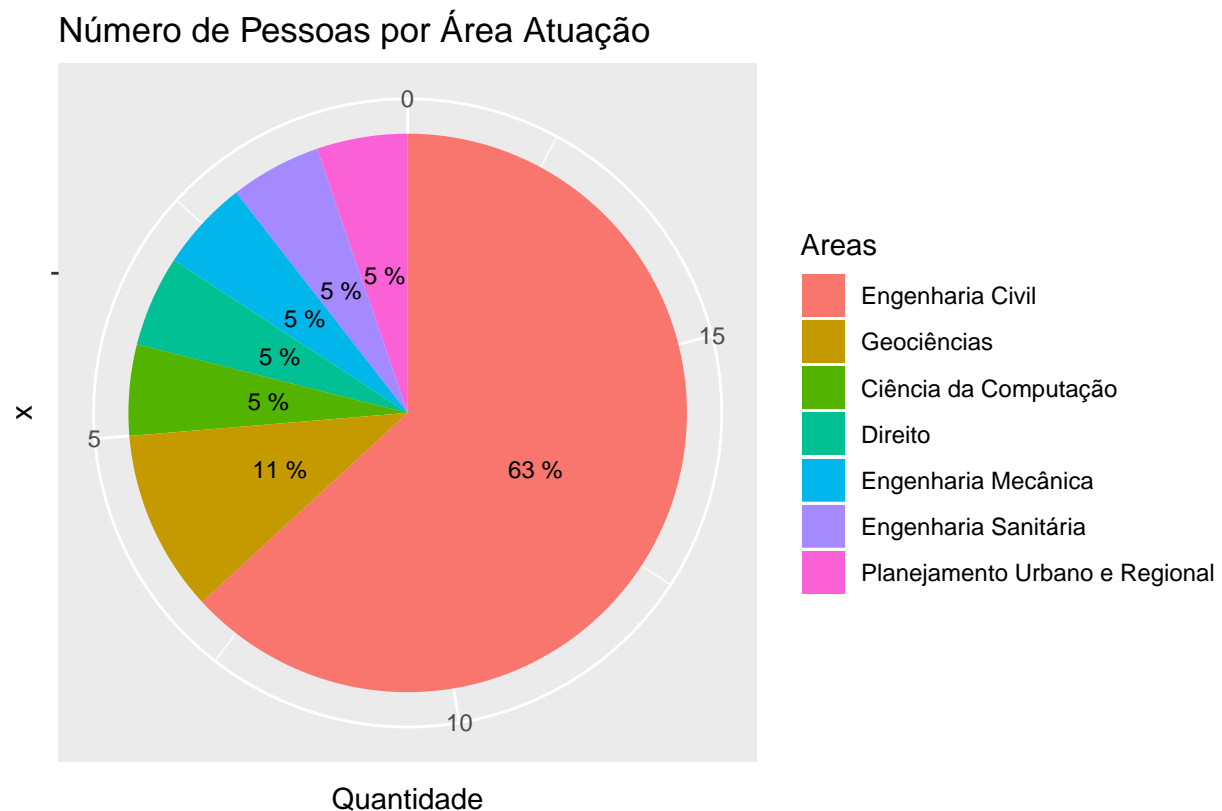
O gráfico ilustra a quantidade de professores por área de atuação (em forma de colunas):

```
profile %>%
  sapply(function(x)
    unique(x$areas_de_atuacao$area)) %>%
  unlist() %>% table() %>% sort() %>% as.data.frame() %>% filter(!. == "") %>% tail(20) %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "green4",
                                         alpha = 0.8,
                                         width = 0.8) + coord_flip() +
  labs(title = "Número de Pessoas por Área Atuação", x = "Área de atuação", y =
        "Quantidade") +
  geom_text(aes(label = Freq),
            hjust = -0.2,
            vjust = 0.3,
            size = 3.5) +
  scale_y_continuous(limits = c(0, 800)) + theme_bw()
```



O gráfico ilustra a quantidade de professores por área de atuação (em forma de pizza):

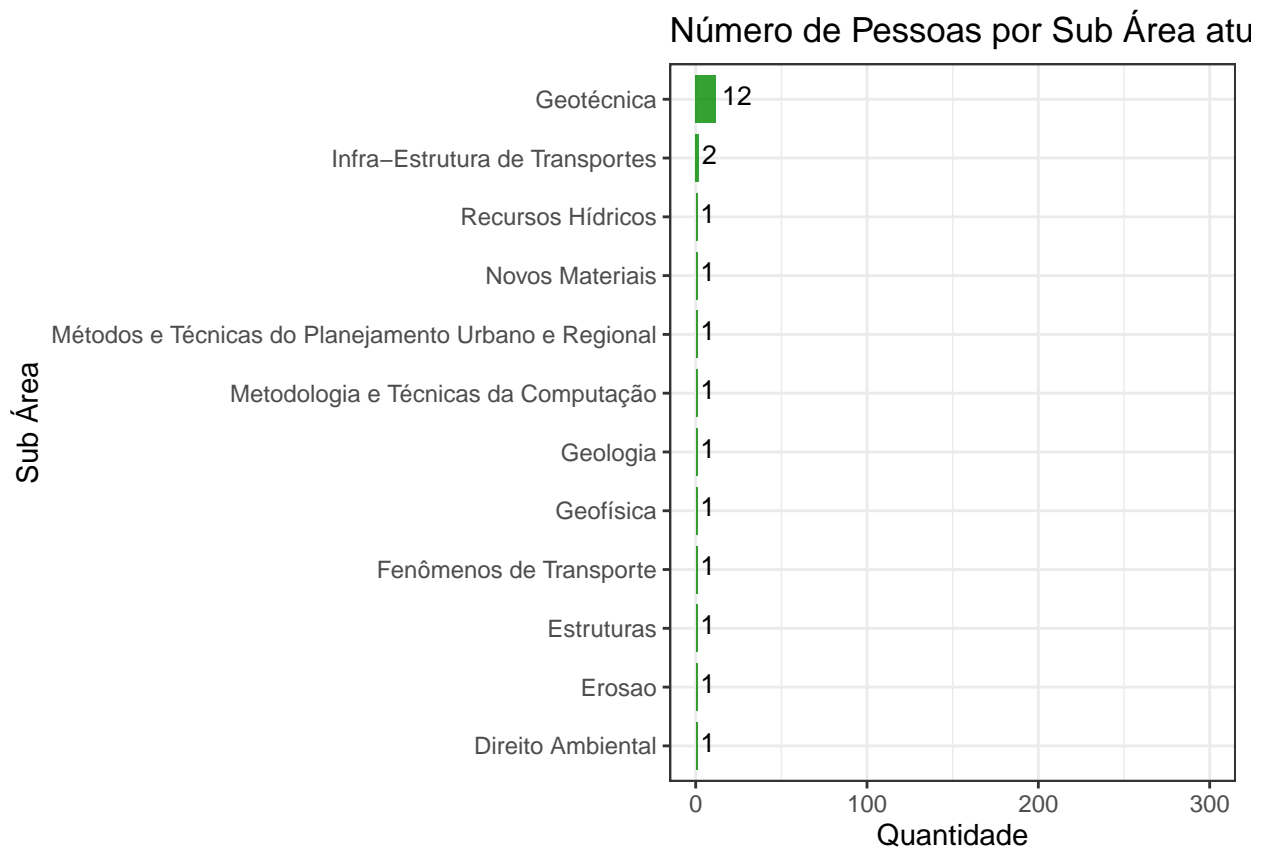
```
ggplot(areas_atuacao, aes(x = "", y = Quantidade, fill = Areas)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0, direction = -1) +
  labs(title = "Número de Pessoas por Área Atuação") +
  geom_text(
    data = areas_atuacao,
    aes(
      x = "",
      y = Quantidade,
      label = paste(Porcentagem, "%")
    ),
    position = position_stack(vjust = 0.5),
    size = 3
  )
```



Com os graficos podemos observar que todos os 12 professores são da área de engenharia civil e apenas 37% dos professores são de outras áreas.

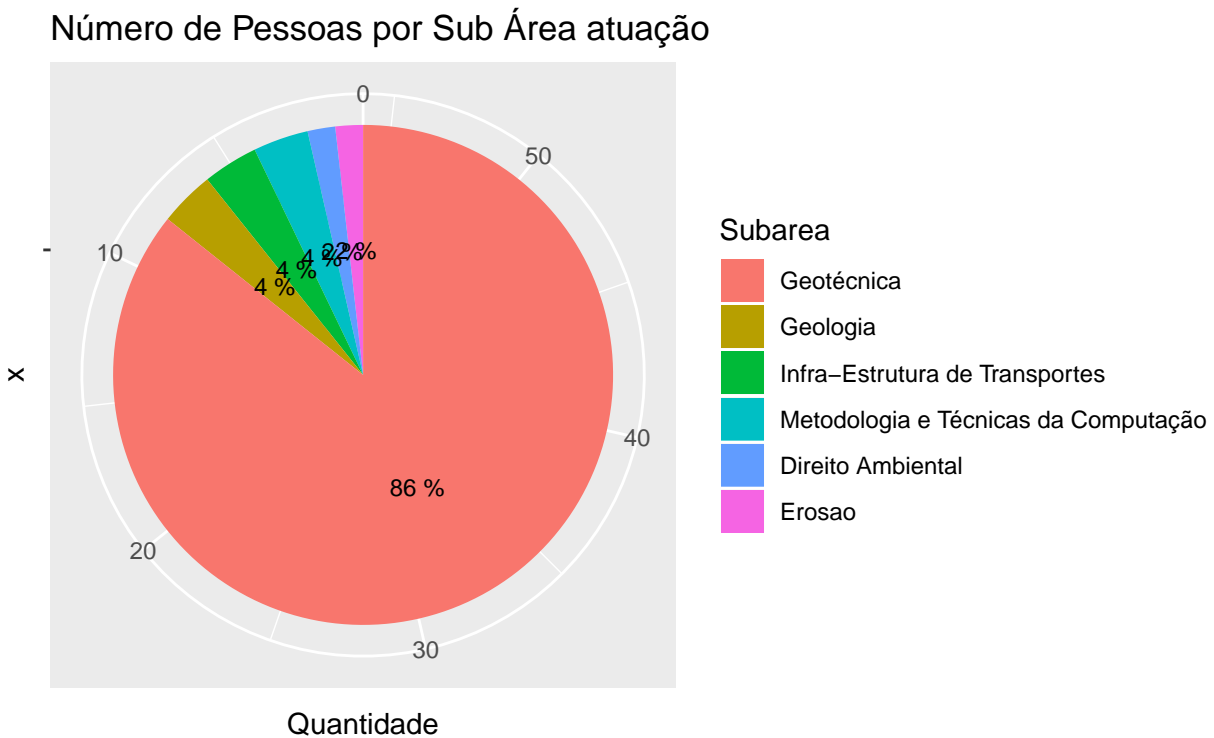
O gráfico ilustra a quantidade de professores por sub área de atuação (em forma de colunas):

```
profile %>%
  sapply(function(x)
    unique(x$areas_de_atuacao$sub_area)) %>%
  unlist() %>% table() %>% sort() %>% as.data.frame() %>% filter(!. == "") %>% tail(30) %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "green4",
    alpha = 0.8,
    width = 0.8) + coord_flip() +
  geom_text(aes(label = Freq),
    hjust = -0.2,
    vjust = 0.3,
    size = 3.5) +
  labs(title = "Número de Pessoas por Sub Área atuação", x = "Sub Área", y =
    "Quantidade") +
  scale_y_continuous(limits = c(0, 300)) + theme_bw()
```



O gráfico ilustra a quantidade de professores por grande área de atuação (em forma de pizza):

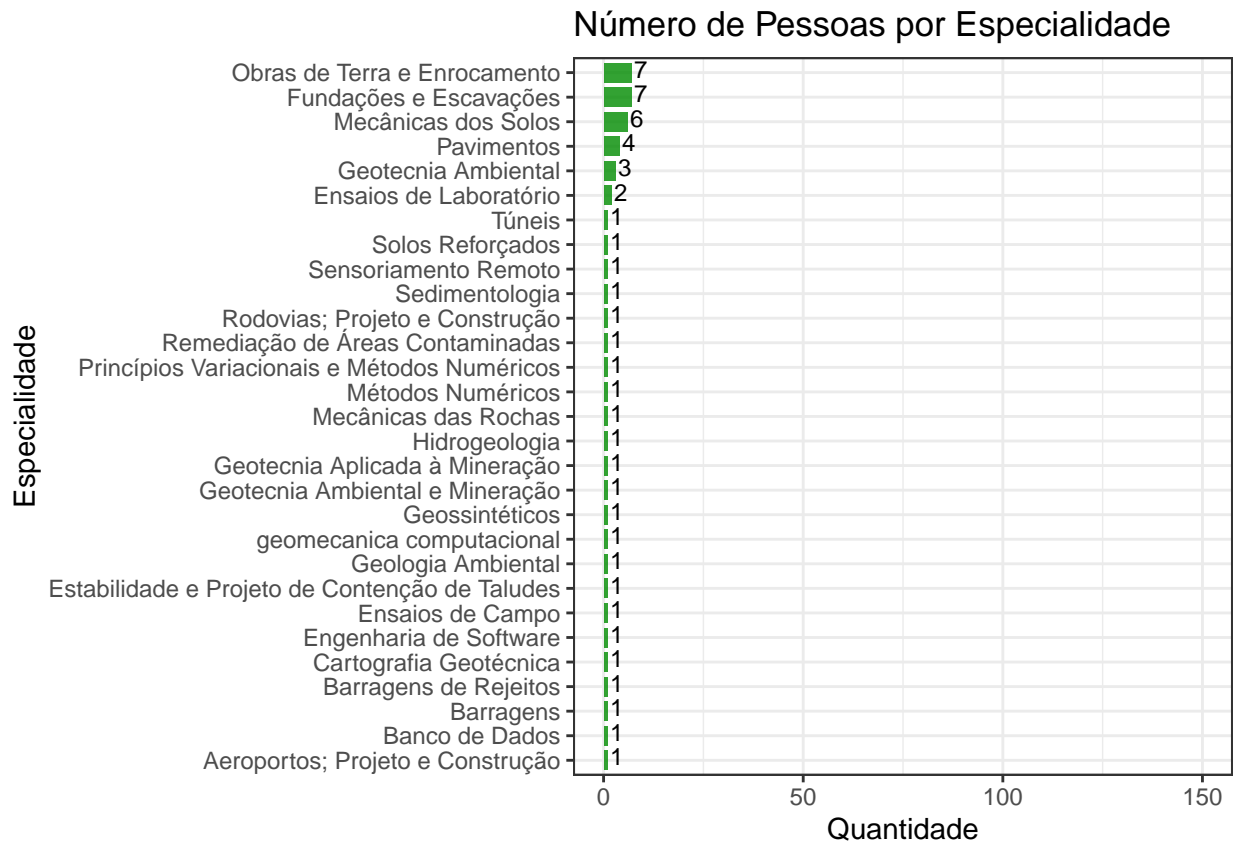
```
ggplot(subarea, aes(x = "", y = Quantidade, fill = Subarea)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0, direction = -1) +
  labs(title = "Número de Pessoas por Sub Área atuação") +
  geom_text(
    data = subarea,
    aes(
      x = "",
      y = Quantidade,
      label = paste(Porcentagem, "%")
    ),
    position = position_stack(vjust = 0.5),
    size = 3
  )
```



Com os graficos podemos observar que todos os 12 professores são da grande área de engenharias e apenas 14% dos professores são de outras sub áreas.

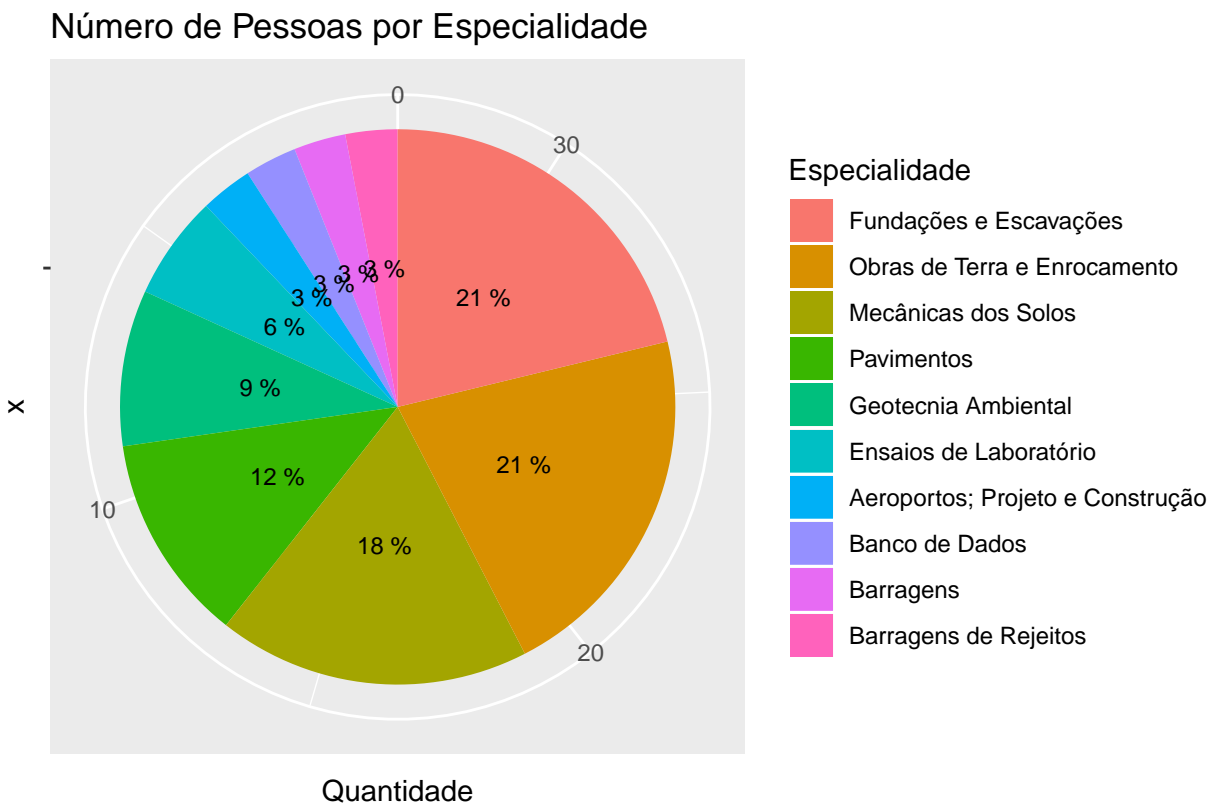
O gráfico ilustra a quantidade de professores por especialidade (em forma de colunas):

```
profile %>%
  sapply(function(x)
    unique(x$areas_de_atuacao$especialidade)) %>%
  unlist() %>% table() %>% sort() %>% as.data.frame() %>% filter(!. == "") %>% tail(29) %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "green4",
    alpha = 0.8,
    width = 0.8) + coord_flip() +
  labs(title = "Número de Pessoas por Especialidade", x = "Especialidade", y =
    "Quantidade") +
  geom_text(aes(label = Freq),
    hjust = -0.2,
    vjust = 0.3,
    size = 3.0) +
  theme_bw() +
  scale_y_continuous(limits = c(0, 150))
```



O gráfico ilustra a quantidade de professores por especialidade (em forma de pizza):

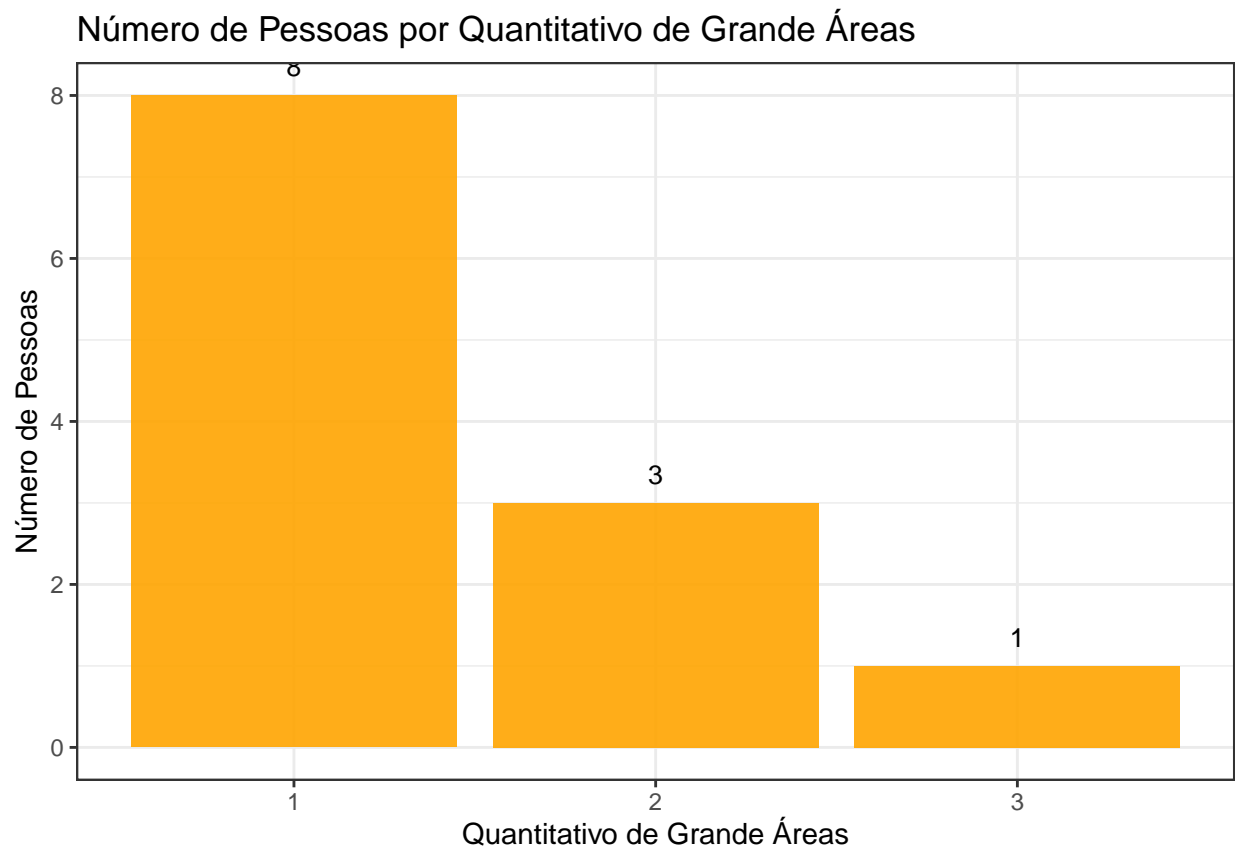
```
ggplot(especialidades_frequentes, aes(x = "", y = Quantidade, fill = Especialidade)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0, direction = -1) +
  labs(title = "Número de Pessoas por Especialidade") +
  geom_text(
    data = especialidades_frequentes,
    aes(
      x = "",
      y = Quantidade,
      label = paste(Porcentagem, "%")
    ),
    position = position_stack(vjust = 0.5),
    size = 3
  )
```



Com os graficos podemos observar que 21% professores tem especialidade em Obras de Terra ou Enrocamento e Fundações e Escavações e apenas 58% dos professores tem outras especialidade.

O gráfico ilustra a quantidade de professores por quantitativo de grande áreas (em forma de barras):

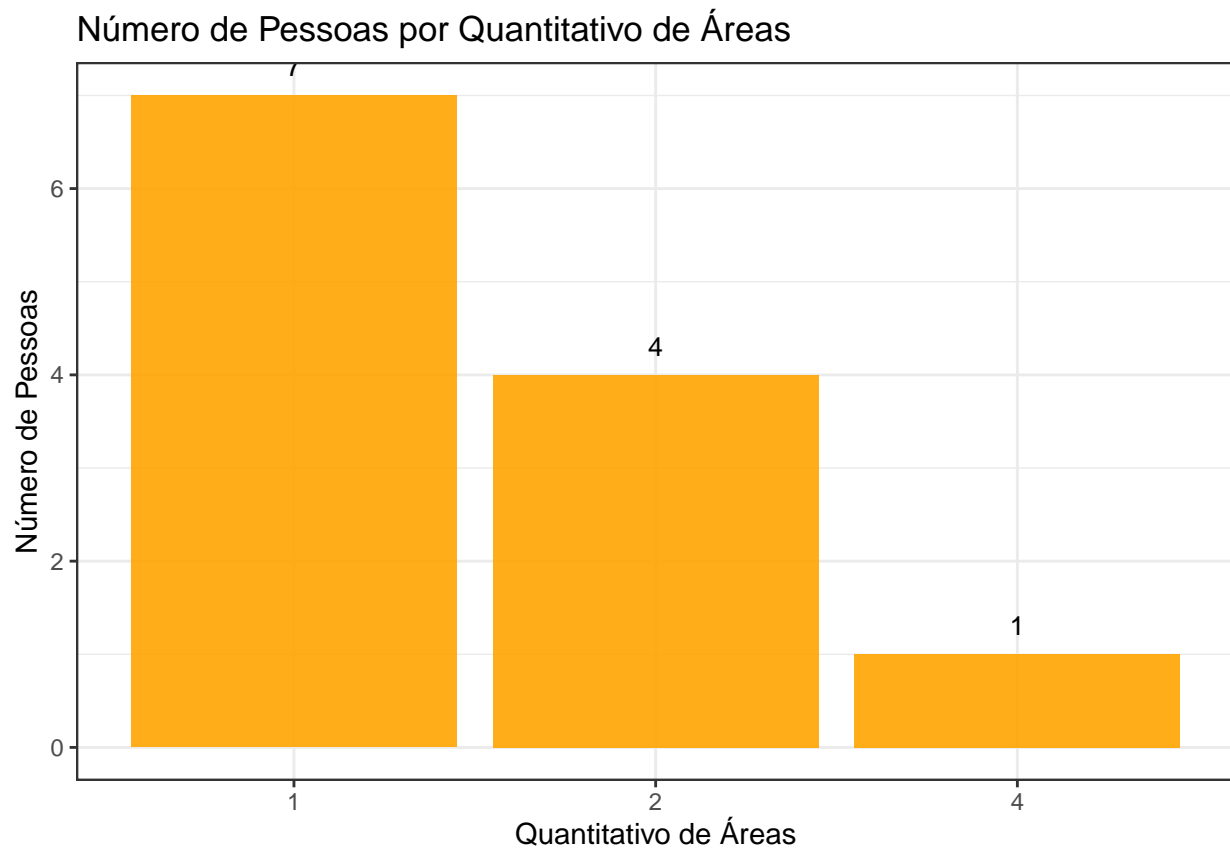
```
profile %>%
  sapply(function(x)
    length(unique(x$areas_de_atuacao$grande_area))) %>%
  unlist() %>% table() %>% sort() %>% rev() %>% as.data.frame() %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "orange", alpha = 0.9) +
  geom_text(aes(label = Freq), size = 3.5, vjust = -1) +
  labs(title = "Número de Pessoas por Quantitativo de Grande Áreas",
    y = "Número de Pessoas", x = "Quantitativo de Grande Áreas") + theme_bw()
```



Podemos observar que tem uma contagem de 8 perfis nas grandes áreas.

O gráfico ilustra a quantidade de professores por quantitativo de áreas (em forma de barras):

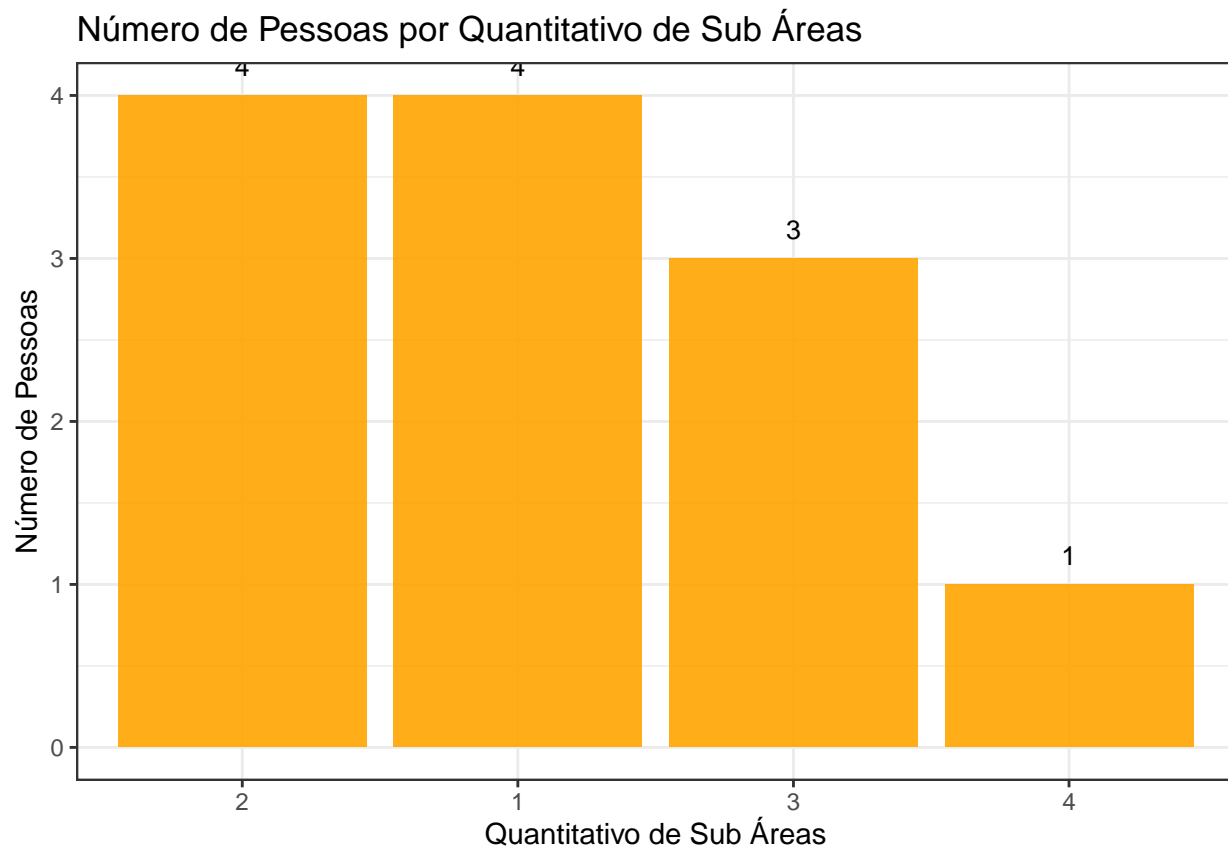
```
profile %>%
  sapply(function(x)
    length(unique(x$areas_de_atuacao$area))) %>%
  unlist() %>% table() %>% sort() %>% rev() %>% as.data.frame() %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "orange", alpha = 0.9) +
  geom_text(aes(label = Freq), size = 3.5, vjust = -1) +
  labs(title = "Número de Pessoas por Quantitativo de Áreas",
    y = "Número de Pessoas", x = "Quantitativo de Áreas") + theme_bw()
```



Podemos observar que tem uma contagem de 7 perfis nas áreas.

O gráfico ilustra a quantidade de professores por quantitativo de sub-áreas (em forma de barras):

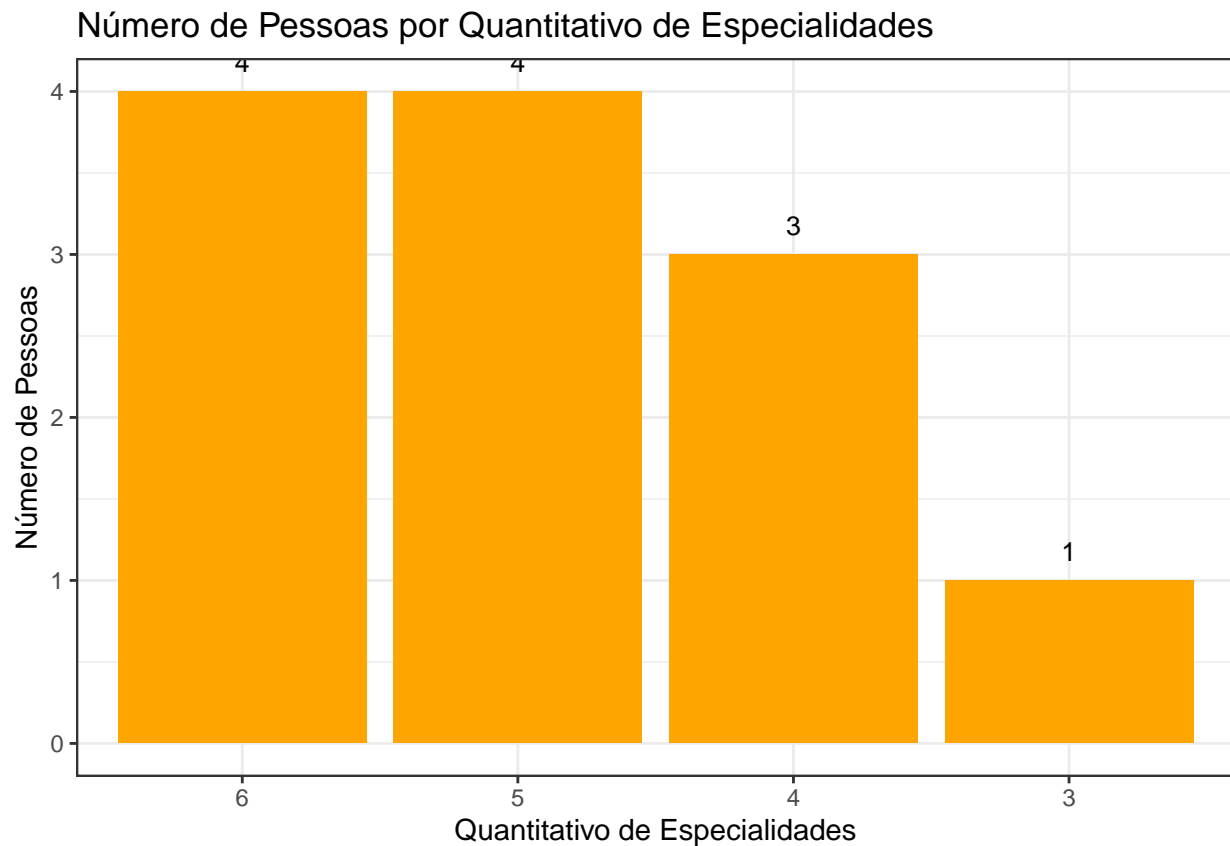
```
profile %>%
  sapply(function(x)
    length(unique(x$areas_de_atuacao$sub_area))) %>%
  unlist() %>% table() %>% sort() %>% rev() %>% as.data.frame() %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "orange", alpha = 0.9) +
  geom_text(aes(label = Freq), size = 3.5, vjust = -1) +
  labs(title = "Número de Pessoas por Quantitativo de Sub Áreas",
    y = "Número de Pessoas", x = "Quantitativo de Sub Áreas") + theme_bw()
```



Podemos observar que tem uma contagem de igual de 4 perfis nas sub áreas.

O gráfico ilustra a quantidade de professores por quantitativo de especialidades (em forma de barras):

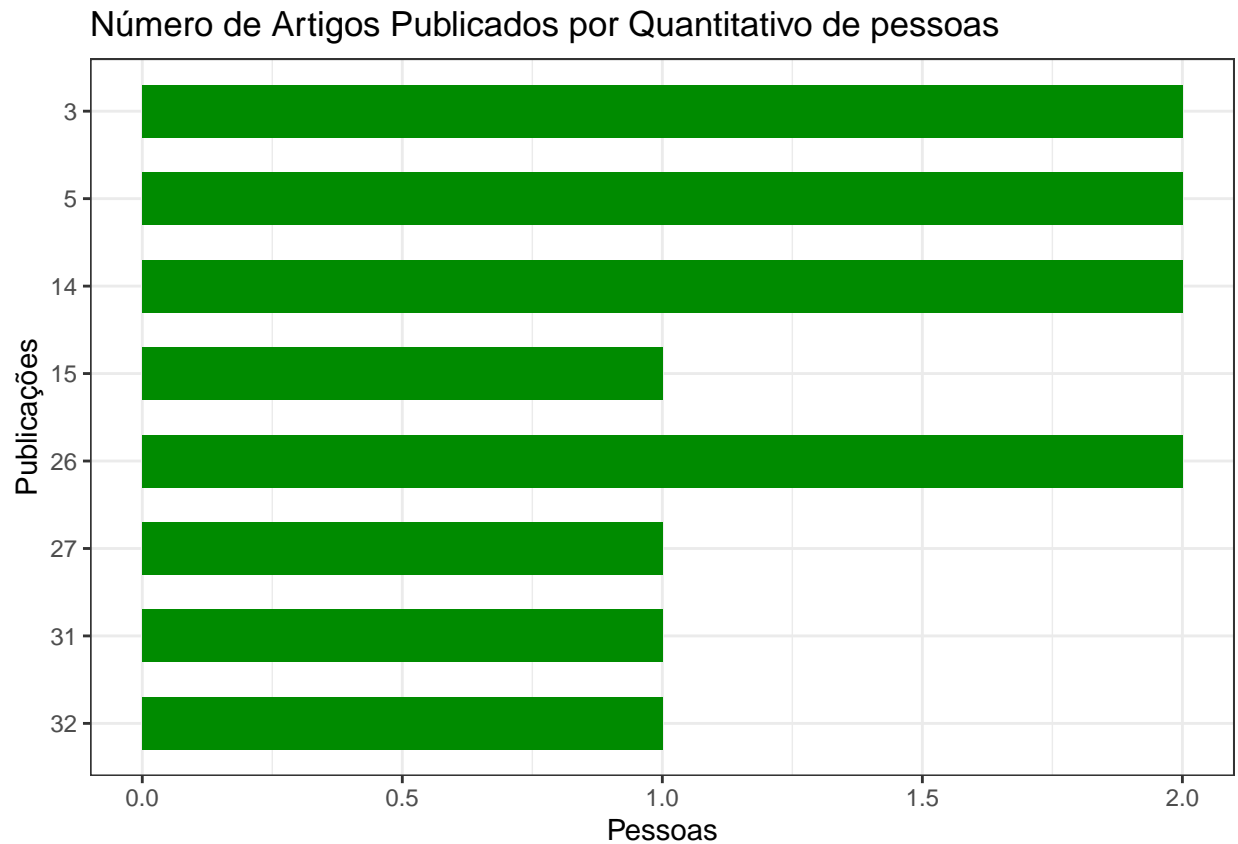
```
profile %>%
  sapply(function(x)
    length(unique(x$areas_de_atuacao$especialidade))) %>%
  unlist() %>% table() %>% sort() %>% rev() %>% as.data.frame() %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "orange") +
  geom_text(aes(label = Freq), size = 3.5, vjust = -1) +
  labs(title = "Número de Pessoas por Quantitativo de Especialidades",
    y = "Número de Pessoas", x = "Quantitativo de Especialidades") + theme_bw()
```



Podemos observar que tem uma contagem de igual de 4 perfis por especialidades

O gráfico ilustra a quantidade de artigos publicados por quantitativo de professores (em forma de colunas):

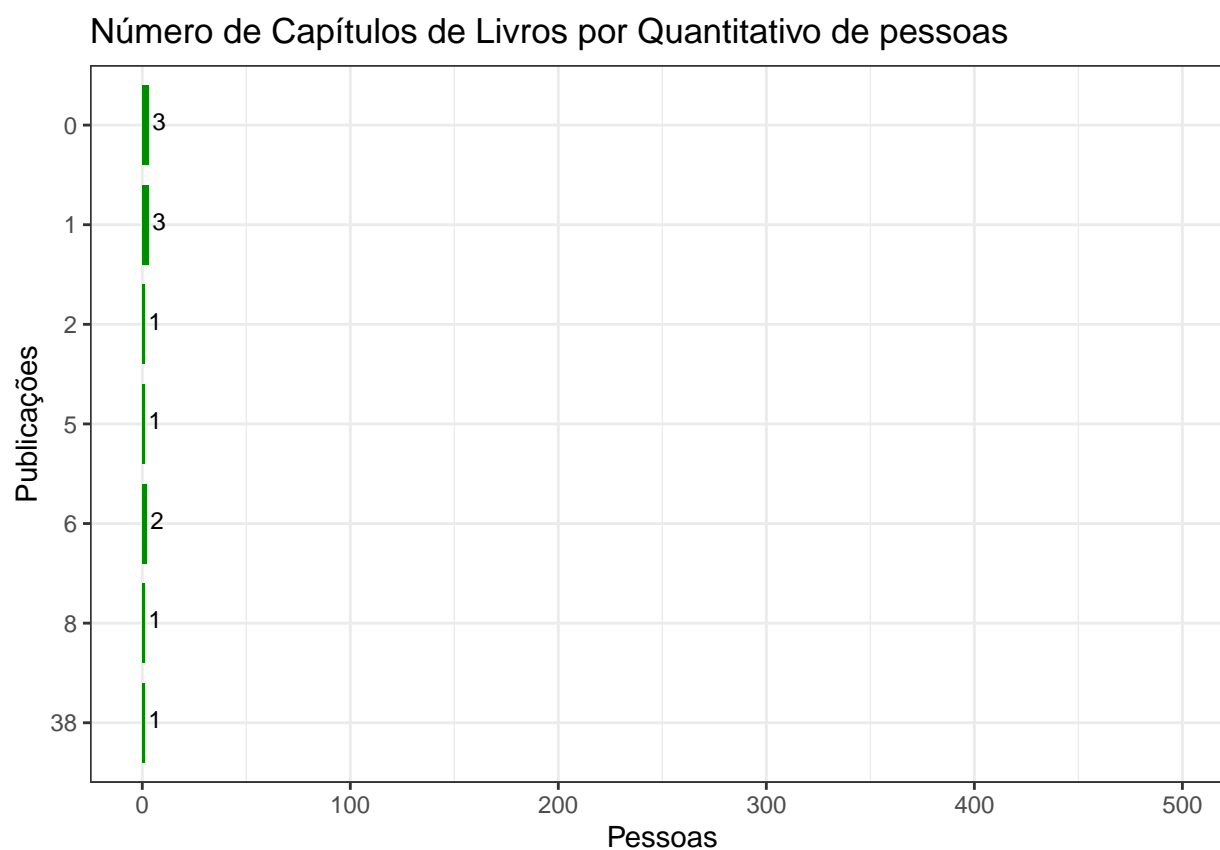
```
profile %>%
  sapply(function(x)
    length(x$producao_bibliografica$PERIODICO$ano)) %>%
  unlist() %>% table() %>% rev() %>% as.data.frame() %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "green4", width = 0.6) +
  labs(title = "Número de Artigos Publicados por Quantitativo de pessoas", y = "Pessoas", x = "Publicações") +
  theme_bw() + coord_flip()
```



Podemos observar que tem uma contagem de 2 perfis já publicaram 26 artigos e só um conseguiu publicar 32 artigos.

O gráfico ilustra a quantidade de capítulo publicados por quantitativo de professores (em forma de colunas):

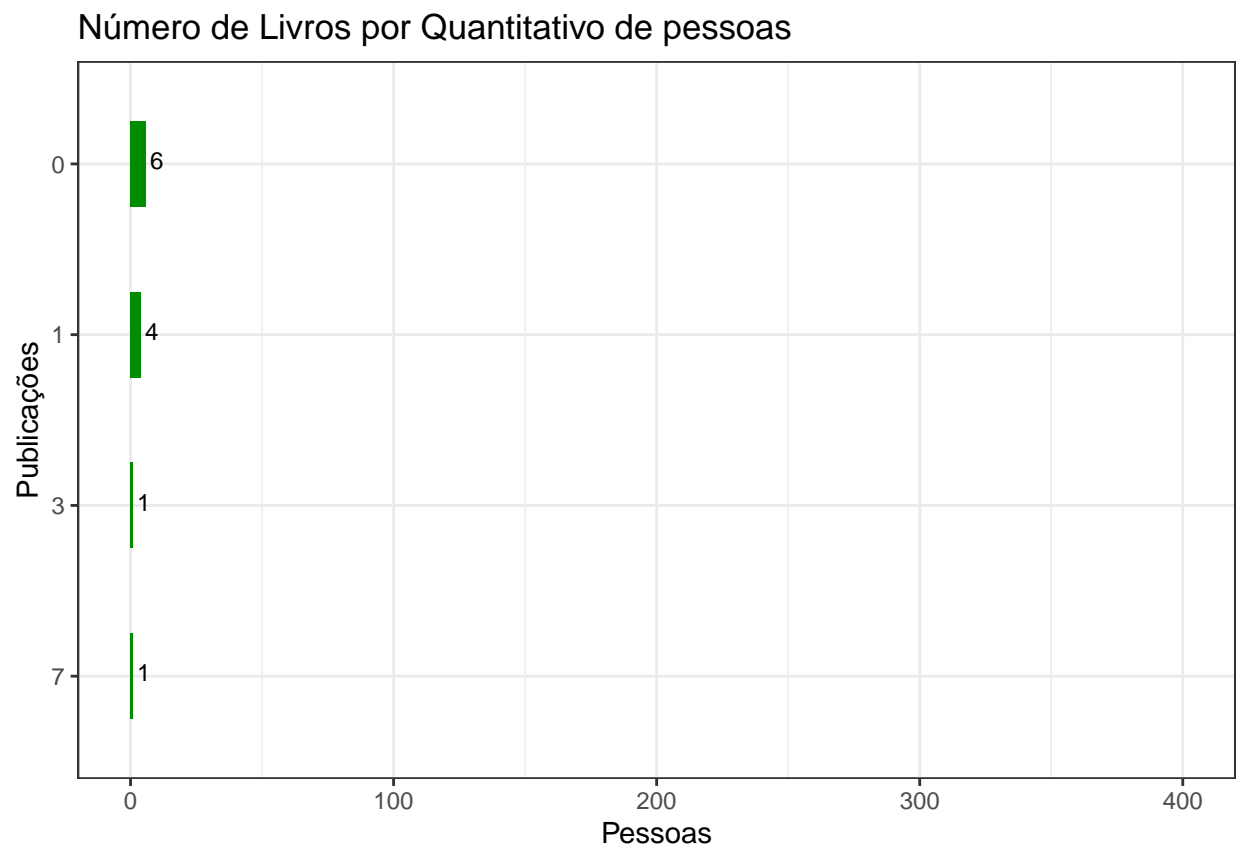
```
profile %>%
  sapply(function(x)
    length(x$producao_bibliografica$CAPITULO_DE_LIVRO$ano)) %>%
  unlist() %>% table() %>% rev() %>% as.data.frame() %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "green4", width = 0.8) + coord_flip() +
  labs(title = "Número de Capítulos de Livros por Quantitativo de pessoas", y = "Pessoas", x = "Publicações") +
  scale_x_discrete() + theme_bw() + geom_text(aes(label = Freq),
    hjust = -0.3,
    vjust = 0.3,
    size = 3.1) + scale_y_continuous(limits = c(0, 500))
```



Podemos observar que tem uma contagem de 2 perfis já publicaram capitulos de livros e 3 não conseguiram publica capitulo de livros.

O gráfico ilustra a quantidade de livros publicados por quantitativo de professores (em forma de colunas):

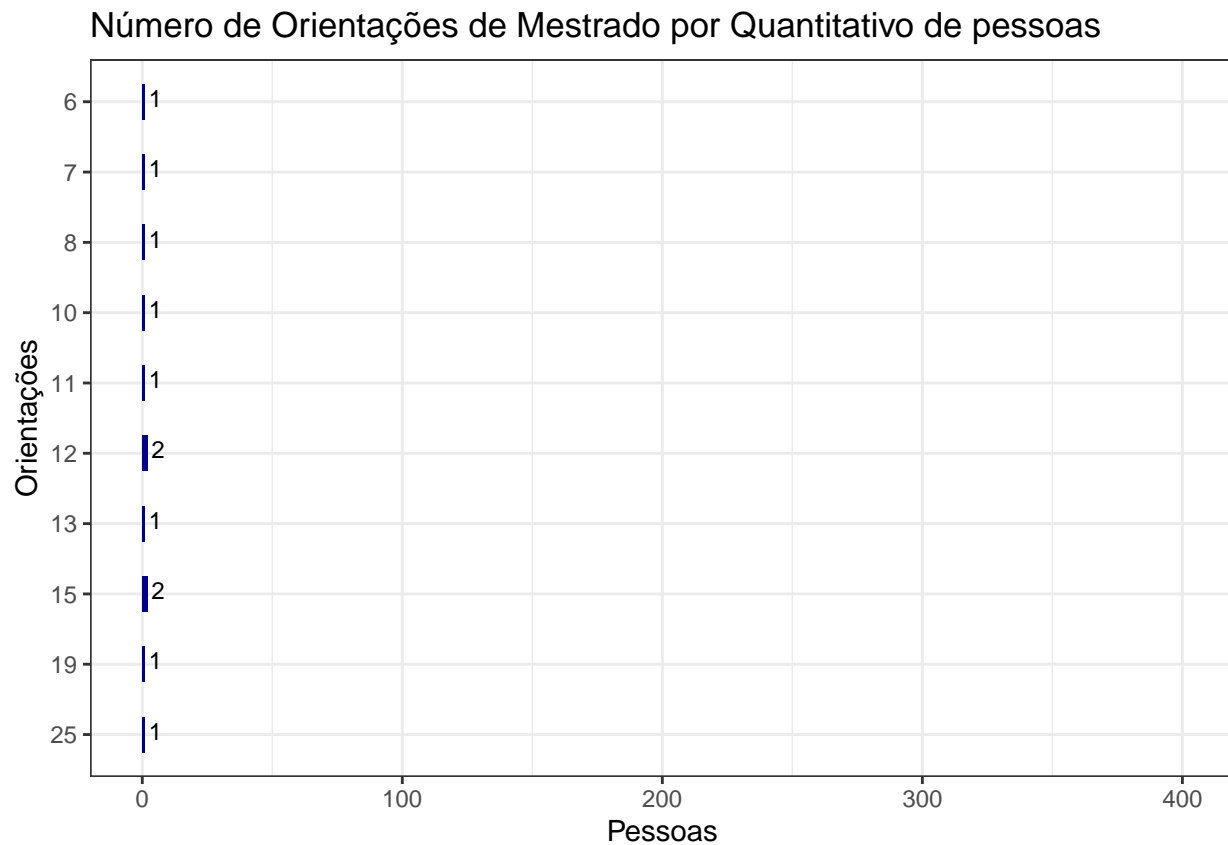
```
profile %>%
  sapply(function(x)
    length(x$producao_bibliografica$LIVRO$ano)) %>%
  unlist() %>% table() %>% rev() %>% as.data.frame() %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "green4", width = 0.5) + coord_flip() +
  labs(title = "Número de Livros por Quantitativo de pessoas", y = "Pessoas", x = "Publicações") +
  theme_bw() + geom_text(aes(label = Freq),
    hjust = -0.3,
    vjust = 0.3,
    size = 3.1) + scale_y_continuous(limits = c(0, 400))
```



Podemos observar que tem uma contagem de 4 perfis já publicaram 1 livro e 3 não conseguiram publica livro.

O gráfico ilustra quantidade de orientações de mestrado por quantitativo de professores (em forma de colunas):

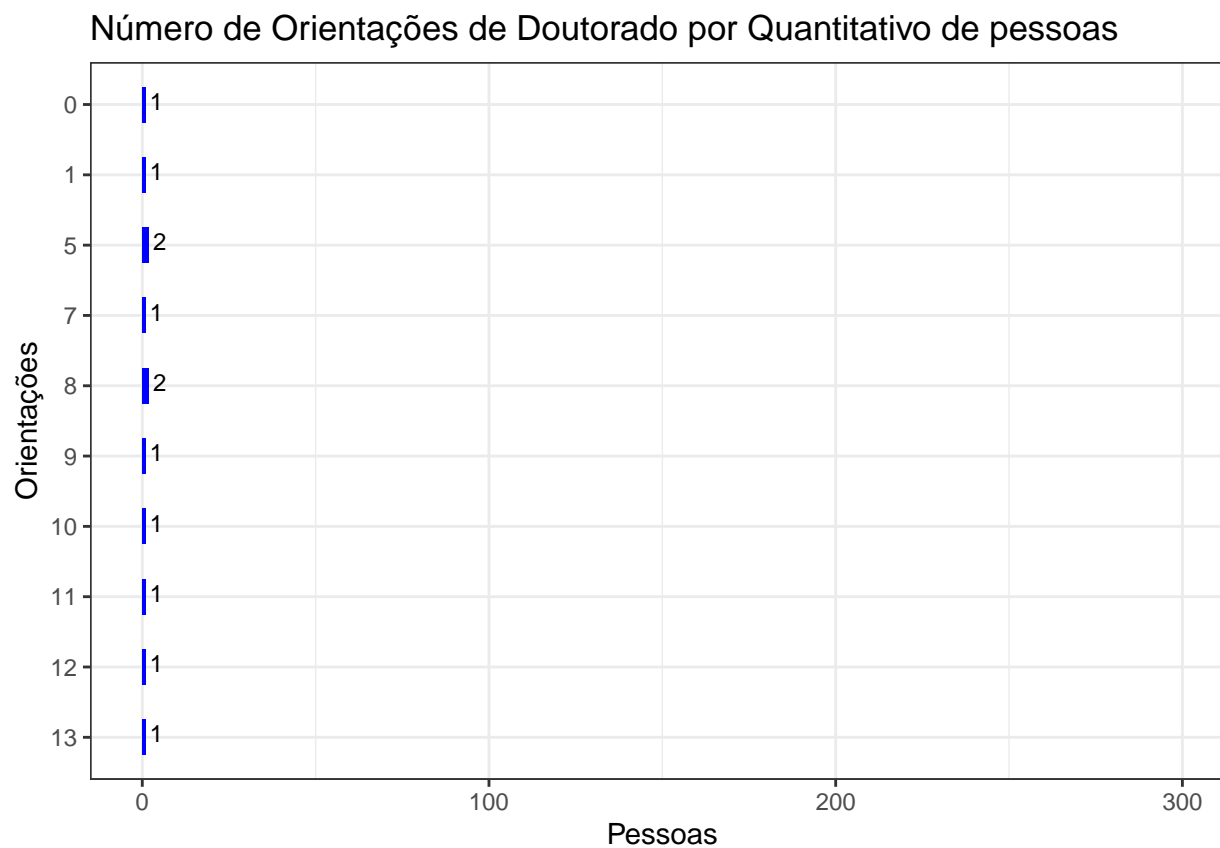
```
profile %>%
  sapply(function(x)
    length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano)) %>%
  unlist() %>% table() %>% rev() %>% as.data.frame() %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "blue4", width = 0.5) + coord_flip() +
  labs(title = "Número de Orientações de Mestrado por Quantitativo de pessoas",
    y = "Pessoas", x = "Orientações") + scale_y_continuous(limits = c(0, 400)) +
  theme_bw() +
  geom_text(aes(label = Freq),
    hjust = -0.3,
    vjust = 0.3,
    size = 3.1)
```



Podemos observar que tem uma contagem de 2 perfis já tiveram 15 orientações de mestrado.

O gráfico ilustra a quantidade de orientações de doutorado por quantitativo de professores (em forma de colunas):

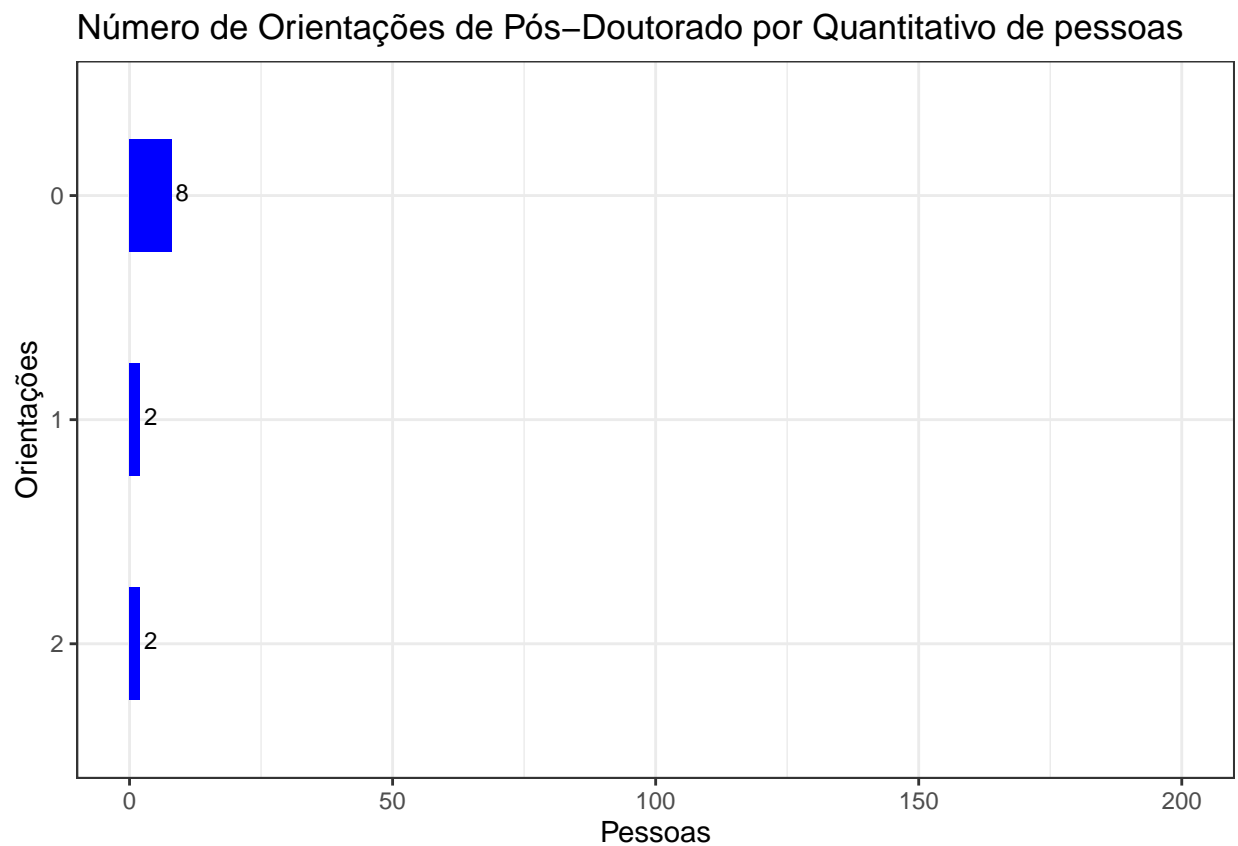
```
profile %>%
  sapply(function(x)
    length(
      x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano
    )) %>%
  unlist() %>% table() %>% rev() %>% as.data.frame() %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "blue", width = 0.5) + coord_flip() +
  labs(title = "Número de Orientações de Doutorado por Quantitativo de pessoas",
       y = "Pessoas", x = "Orientações") + scale_y_continuous(limits = c(0, 300)) +
  theme_bw() +
  geom_text(aes(label = Freq),
            hjust = -0.3,
            vjust = 0.3,
            size = 3.1)
```



Podemos observar que tem uma contagem de 2 perfis já tiveram 8 orientações de doutorado

O gráfico ilustra a quantidade de orientações de pós-doutorado por quantitativo de professores(em forma de colunas):

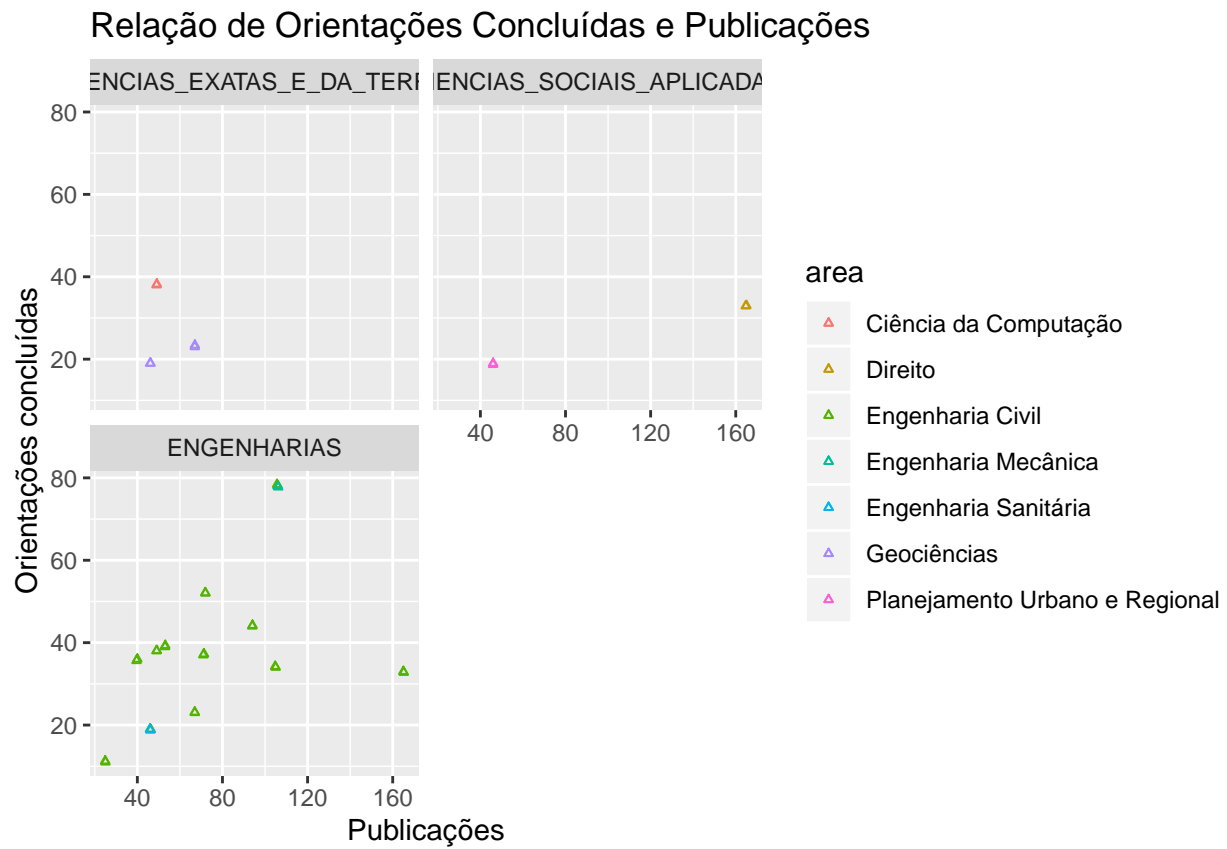
```
profile %>%
  supply(
    function(x)
      length(
        x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano
      )
  ) %>%
  unlist() %>% table() %>% rev() %>% as.data.frame() %>%
  ggplot(aes(x = ., y = Freq)) + geom_col(fill = "blue", width = 0.5) + coord_flip() +
  labs(title = "Número de Orientações de Pós-Doutorado por Quantitativo de pessoas",
       y = "Pessoas", x = "Orientações") + scale_y_continuous(limits = c(0, 200)) +
  theme_bw() +
  geom_text(aes(label = Freq),
            hjust = -0.3,
            vjust = 0.3,
            size = 3.1)
```



Podemos observar que tem uma contagem de 2 perfis já tiveram 2 orientações de pós-doutorado e 8 perfis não tiveram.

O gráfico ilustra as as relações de Orientações Concluídas e Publicações:

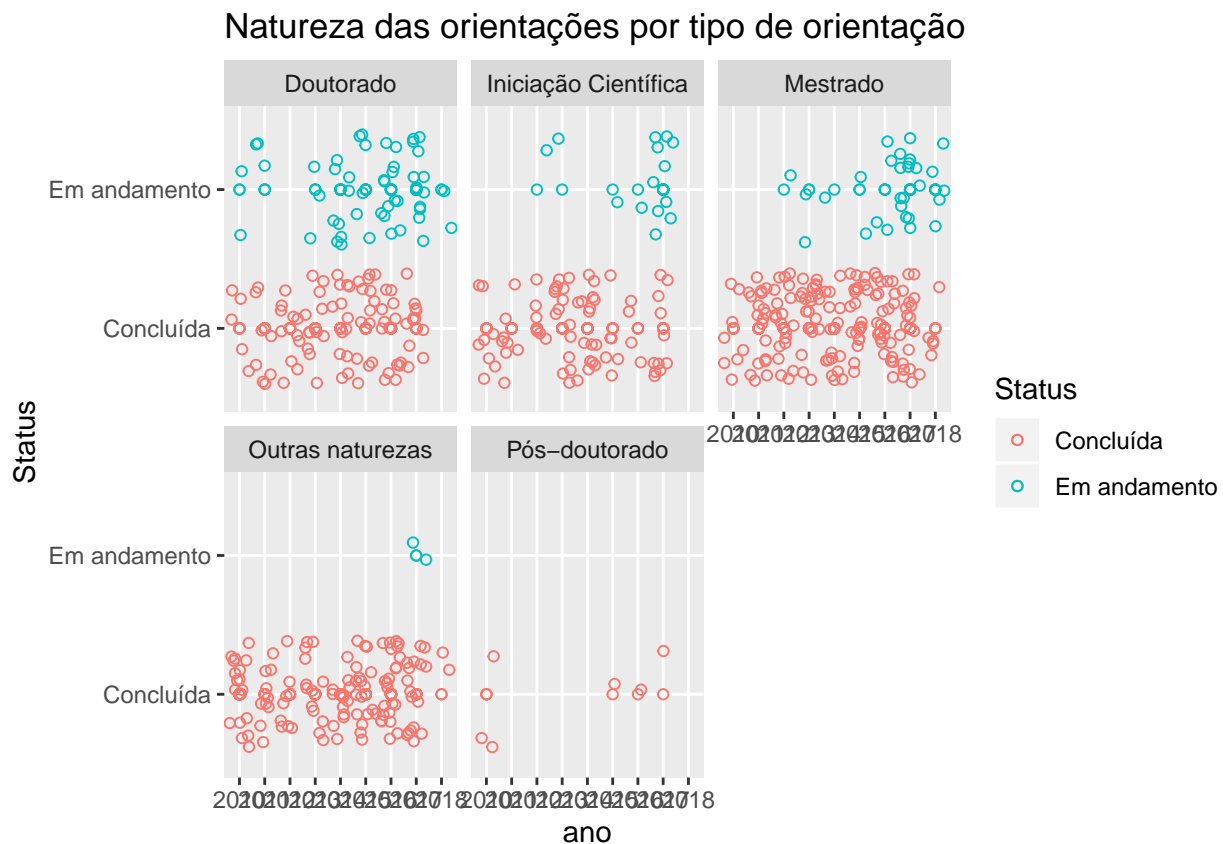
```
profile.areas %>%
  select(-sub_area, -especialidade) %>%
  distinct() %>%
  group_by(publicacoes) %>%
  ggplot(aes(publicacoes, orientacoes_concluidas, color = area)) +
  geom_point(shape = 2, size = .8) + geom_jitter(shape = 2, size = .8) +
  ggtitle('Relação de Orientações Concluídas e Publicações')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



O gráfico ilustra a natureza das orientações por tipo de orientação (em forma de pontos):

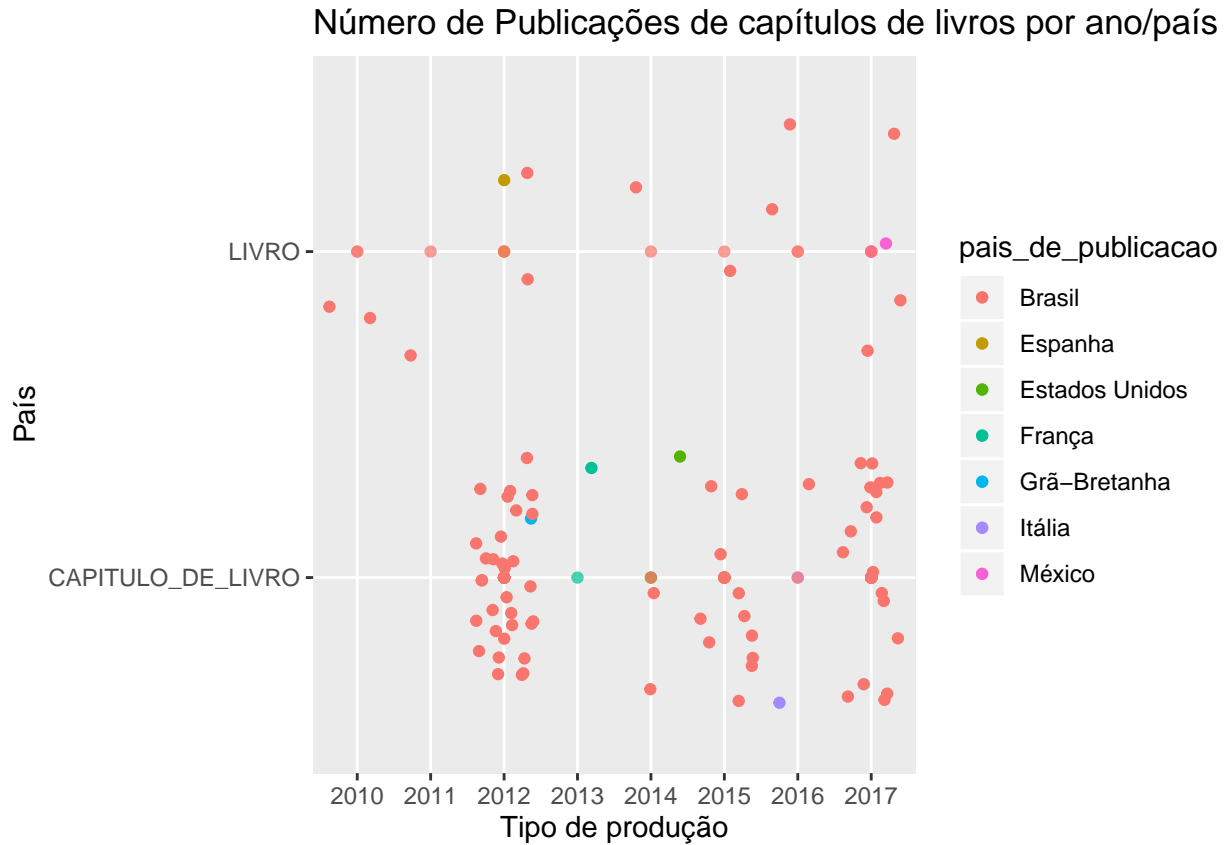
```
profile.df.orientacoes <- extrai.orientacoes(profile) %>%
  select(id_lattes_orientadores, natureza, ano, orientacao, everything()) %>%
  mutate(Status = ifelse(grepl("CONCLUIDA", orientacao), "Concluída", "Em andamento")) %>%
  mutate(Natureza = case_when(grepl("MESTRADO", str_to_upper(natureza))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

profile.df.orientacoes %>% group_by(ano, Status) %>%
  ggplot(aes(x = ano, y = Status, color = Status)) +
  labs(title = "Natureza das orientações por tipo de orientação") +
  geom_point(shape = 1) + geom_jitter(shape = 1) + facet_wrap(. ~ Natureza)
```



O gráfico ilustra a quantidade publicações de capítulos de livros por ano/país (em forma de pontos):

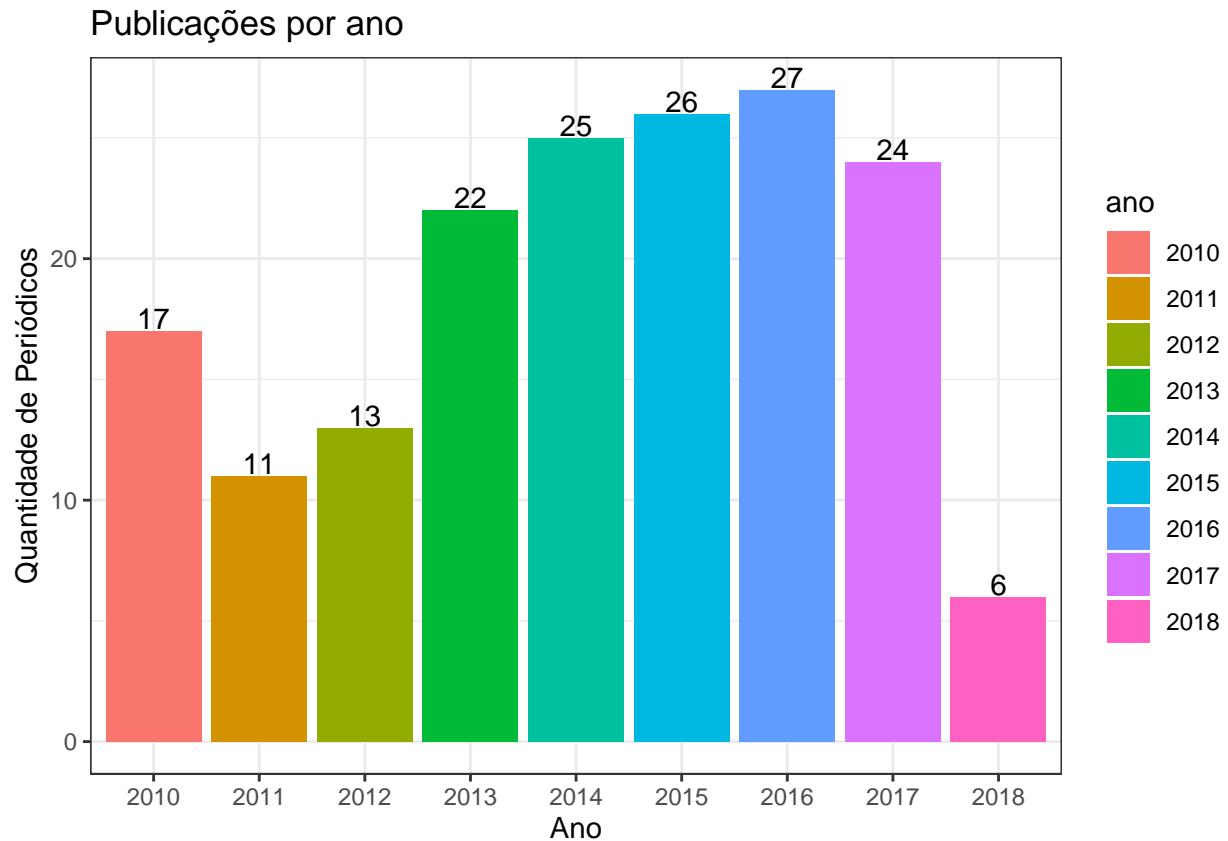
```
profile.df.publicacoes %>%
  filter((tipo_producao %in% c('LIVRO', 'CAPITULO_DE_LIVRO'))) %>%
  group_by(tipo_producao, pais_de_publicacao) %>%
  ggplot(aes(ano, tipo_producao, col = pais_de_publicacao)) +
  labs(title = "Número de Publicações de capítulos de livros por ano/país") +
  geom_point(alpha = 0.7) + geom_jitter() +
  labs(x = 'Tipo de produção', y = 'País')
```



7.0.2 Arquivo Publicação

O gráfico ilustra a quantidade de publicações por ano (em forma de barras):

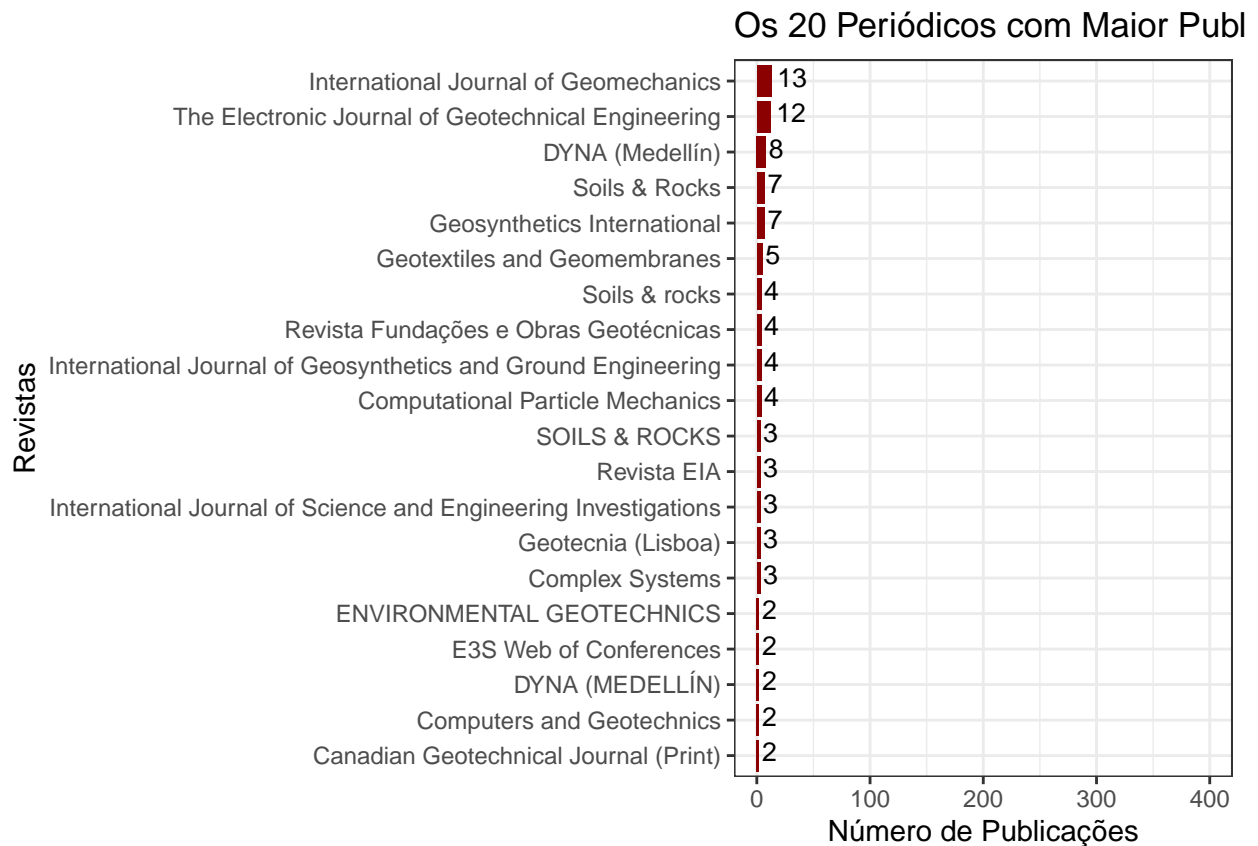
```
publication.periodico.df %>%  
  ggplot(aes(x = ano)) + geom_bar(aes(fill = ano)) +  
  geom_text(stat = "count", aes(label = formatC(..count.., big.mark = ",")), vjust =  
    -0.1) +  
  theme_bw() + labs(title = "Publicações por ano", x = "Ano", y = "Quantidade de Periódicos") +  
  scale_y_continuous(labels = comma)
```



Podemos observar teve um crescimento alto de periódicos em 2016, mas teve muito decréscimo na quantidade em 2018.

O gráfico ilustra as 20 revistas publicadas mais publicadas:

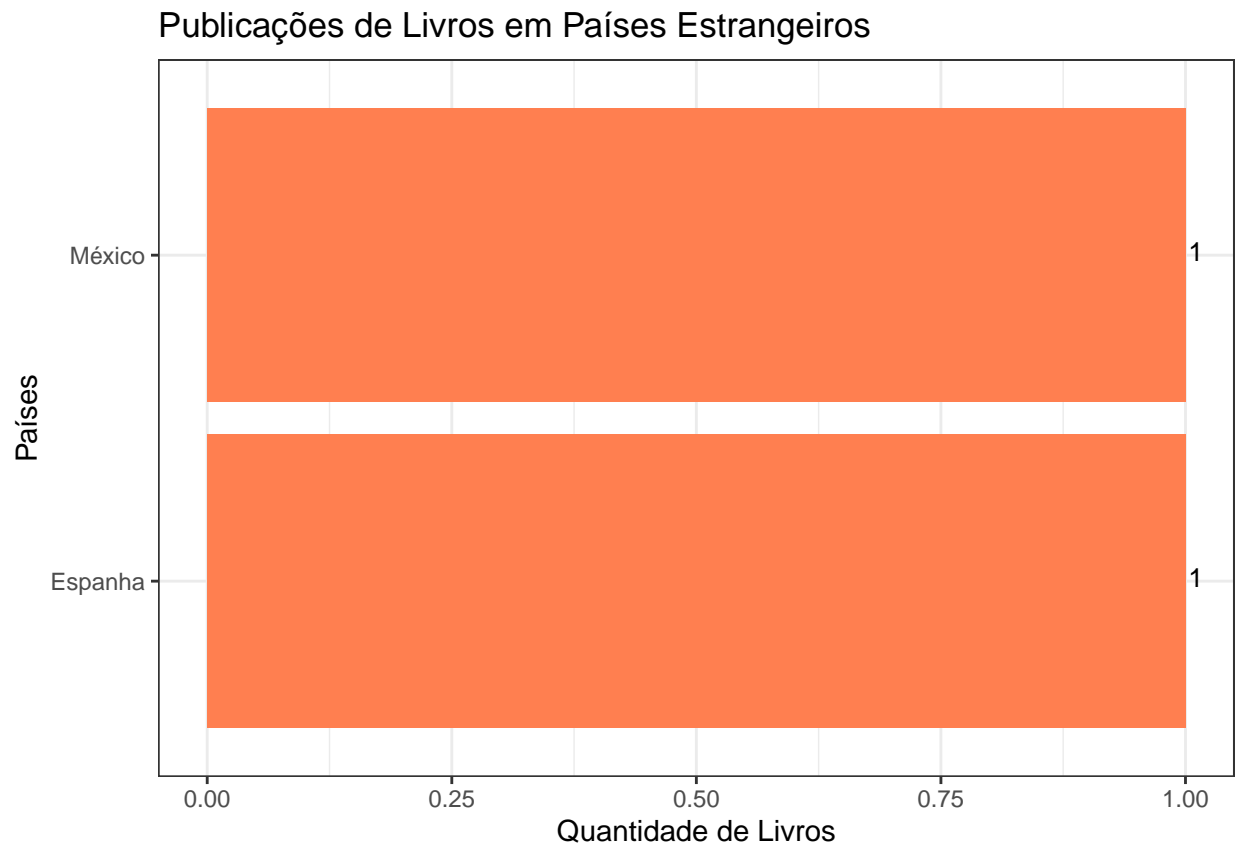
```
publication.periodico.df %>% select(periodico) %>% table() %>% as.data.frame() %>% arrange(desc(Freq)) %>%
head(20) %>% ggplot(aes(x = reorder(., (Freq)), y = Freq)) + geom_col(fill = "red4") + coord_flip() +
labs(title = "Os 20 Periódicos com Maior Publicações",
y = "Número de Publicações", x = "Revistas") + geom_text(
aes(label = comma(Freq)),
hjust = -0.2,
vjust = 0.3,
size = 3.5
) + theme_bw() +
scale_y_continuous(limits = c(0, 400))
```



Podemos observar que revistas International Journal of Geomechanics e The Electronic Journal of Geotechnical Engineering são as mais publicas.

O gráfico ilustra a quantidade de publicações de livros em países estrangeiros (em forma de colunas):

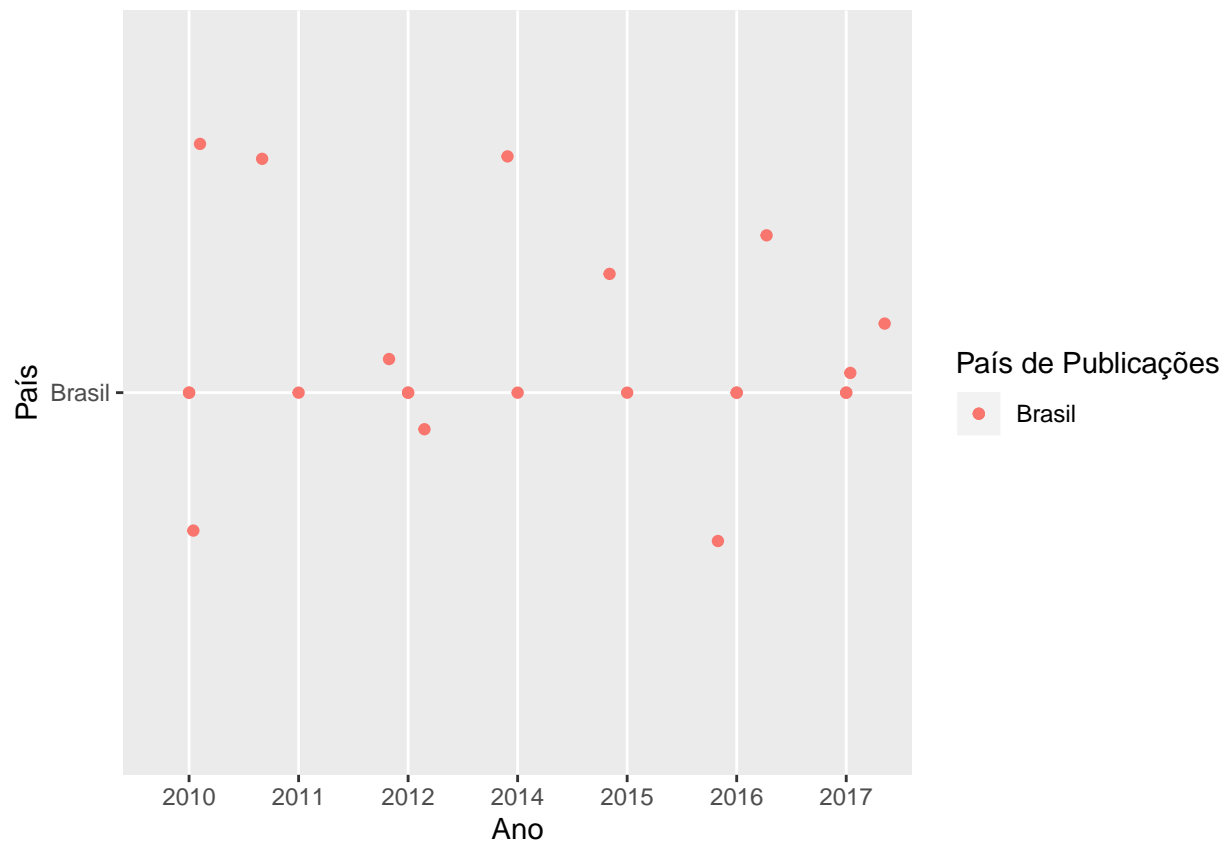
```
publication.livros.df %>%  
  group_by(pais_de_publicacao) %>%  
  summarise(Quantidade = n()) %>%  
  filter(pais_de_publicacao != "Brasil") %>%  
  ggplot(aes(x = reorder(pais_de_publicacao, (Quantidade)), y = Quantidade)) +  
  geom_col(fill = "coral") + geom_text(  
    aes(label = comma(Quantidade)),  
    hjust = -0.2,  
    vjust = 0.3,  
    size = 3.5  
  ) + coord_flip() +  
  labs(title = "Publicações de Livros em Países Estrangeiros", x = "Países", y = "Quantidade de Livros")  
  theme_bw()
```



Podemos observar só tiveram publicações de livros no México e espanha.

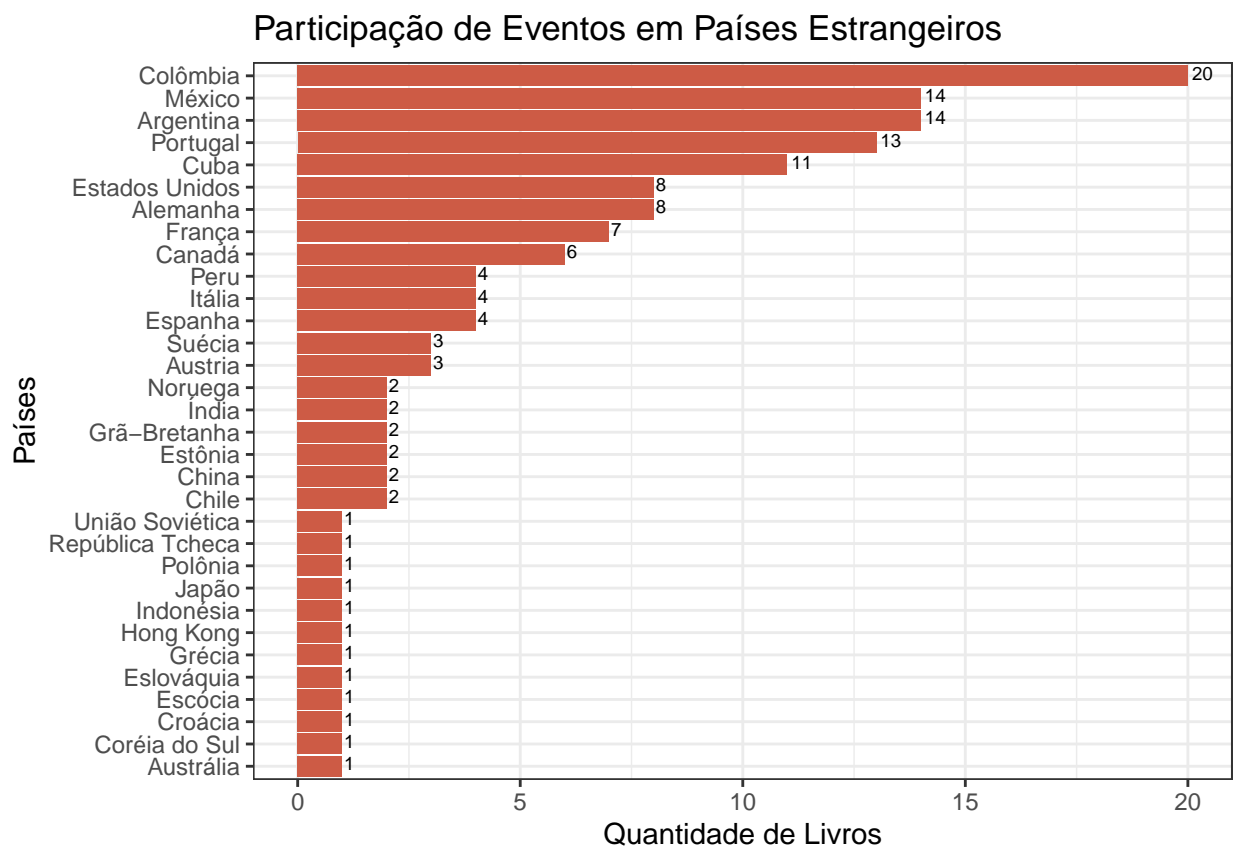
O gráfico ilustra a quantidade de publicações de livros no Brasil (em forma de pontos):

```
publication.livros.df %>%
  filter(
    pais_de_publicacao %in% c(
      "Brasil",
      "Estados Unidos",
      "Holanda",
      "Grã-Bretanha",
      "Alemanha",
      "Suíça"
    )
  ) %>%
  group_by(ano, pais_de_publicacao) %>%
  ggplot(aes(x = ano, y = pais_de_publicacao, color = pais_de_publicacao)) +
  xlab("Ano") + ylab("País") + geom_point() + geom_jitter() +
  labs(color = "País de Publicações")
```



O gráfico ilustra a quantidade de participação de eventos de livros em países estrangeiros (em forma de colunas):

```
publication.eventos.df %>%
  group_by(pais_do_evento) %>%
  summarise(Quantidade = n()) %>%
  filter(pais_do_evento != "Brasil") %>%
  ggplot(aes(x = reorder(pais_do_evento, (Quantidade)), y = Quantidade)) +
  geom_col(fill = "coral3") + geom_text(
    aes(label = comma(Quantidade)),
    hjust = -0.2,
    vjust = 0.3,
    size = 2.5
  ) + coord_flip() +
  labs(title = "Participação de Eventos em Países Estrangeiros", x = "Países", y = "Quantidade de Livros") +
  theme_bw()
```

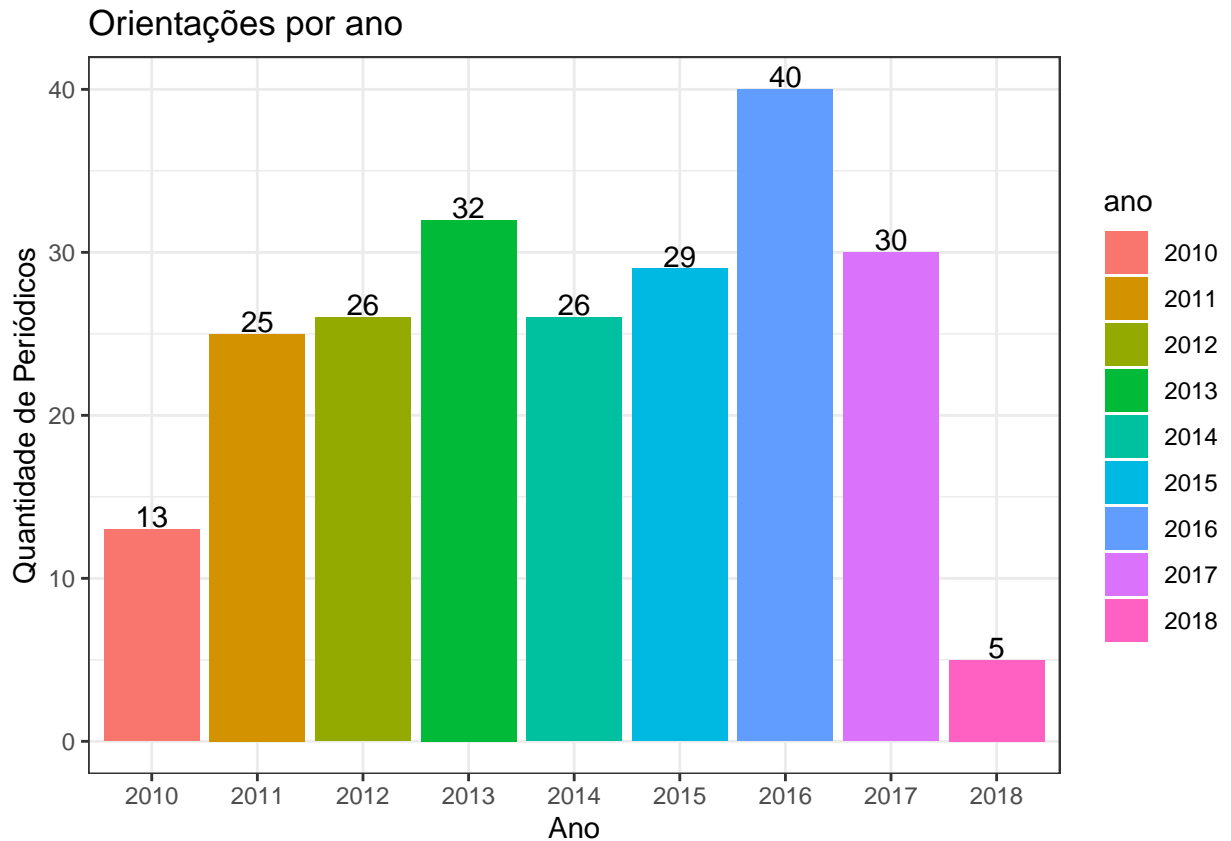


Podemos observar as participações de eventos de livros ocorreu na Colômbia.

7.0.3 Arquivo Orientação

O gráfico ilustra a quantidade de orientações por ano (em forma de barras):

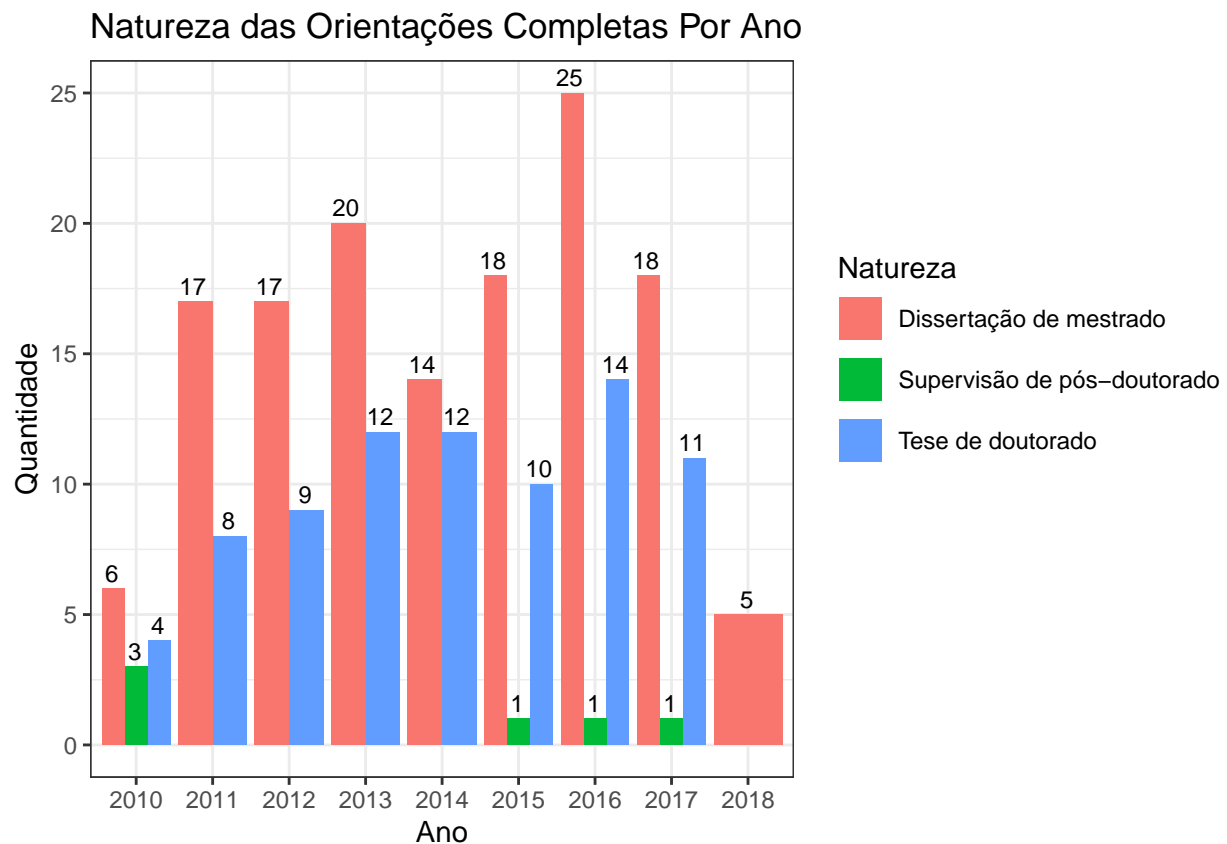
```
orient.df %>%  
  ggplot(aes(x = ano)) + geom_bar(aes(fill = ano)) +  
  geom_text(stat = "count", aes(label = formatC(..count.., big.mark = ",")), vjust =  
    -0.1) +  
  theme_bw() + labs(title = "Orientações por ano", x = "Ano", y = "Quantidade de Periódicos") +  
  scale_y_continuous(labels = comma)
```



Podemos observar teve um crescimento alto de orientações em 2016, mas teve decrescimento muito baixo em 2018.

O gráfico ilustra a quantidade de orientações completas por ano (em forma de barras):

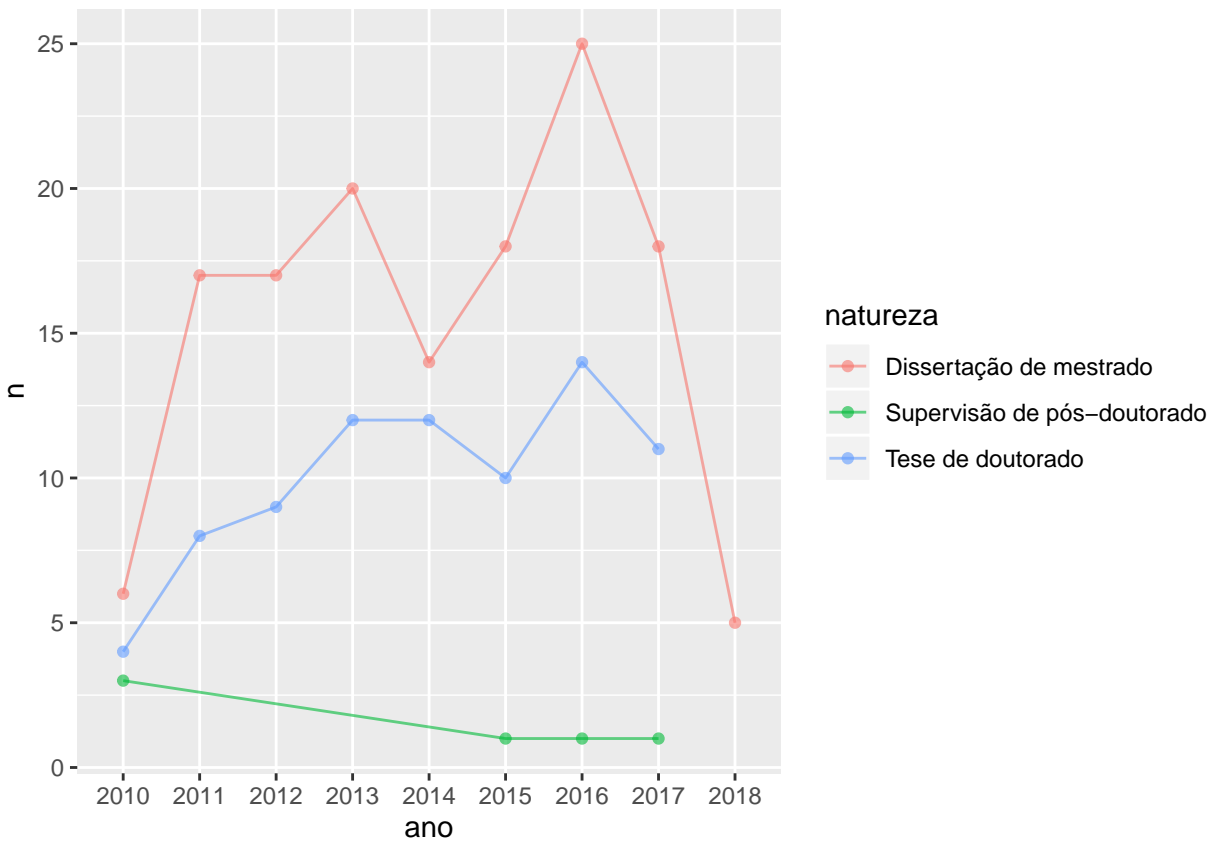
```
ggplot(orient.df, aes(ano, fill = factor(natureza))) +  
  geom_bar(stat = "count", position = "dodge") +  
  ggtitle("Natureza das Orientações Completas Por Ano") +  
  theme(legend.position = "right", legend.text = element_text(size = 7)) +  
  guides(fill = guide_legend(  
    nrow = 5,  
    byrow = TRUE,  
    title.position = "top"  
  )) +  
  labs(x = "Ano", y = "Quantidade") + labs(fill = "Natureza") + theme_bw() +  
  geom_text(  
    hjust = 0.6,  
    vjust = -0.4,  
    size = 3,  
    color = 'black',  
    position = position_dodge(width = 0.9),  
    stat = "count",  
    aes(  
      group = factor(natureza),  
      label = formatC(..count.., big.mark = ",")  
    ), check_overlap = TRUE)  
)
```



Podemos observar teve um crescimento alto de dissertação de mestrado e tese de doutorado em 2016 e pouca supervisão de pós-doutorado.

O gráfico ilustra a quantidade de orientações completas por ano (em forma de linhas e pontos):

```
ggplot(orient_sum, aes(  
  x = ano,  
  y = n,  
  group = natureza,  
  color = natureza  
)) +  
  geom_line(alpha = 0.6) +  
  geom_point(alpha = 0.6)
```



7.1 Revisão do processo

Por meio das análises realizadas nas seções anteriores, tem-se que a modelagem de análise com os modelos estatísticos foram adequados para esse projeto, uma vez que permitiu a obtenção de resultados não triviais em relação aos dois programas de pós-graduação analisados.

Tem-se ainda que os modelos e os datasets em que se aplicaram os modelos nesse relatório são facilmente verificáveis e replicáveis por meio da leitura desse documento.

8 Fase 6 - Implantação (deployment)

Na fase de implantação, realiza-se o planejamento da implantação dos scripts desenvolvidos para o ambiente operacional. Os scripts desenvolvidos nesse trabalho permitem uma análise de dados do programa de pós-graduação sob outros pontos de vista, que podem trazer estatísticas e relações não triviais.

9 Conclusão

Esse trabalho mostrou uma forma de analisar dados de arquivos .json utilizando as técnicas da CRISP-DM. Dessa forma foi possível aprender sobre a ciência de dados e as possíveis análises que ela pode proporcionar; como os dados devem ser preparados; como deve-se buscar informações úteis e analisa-las; como gerar gráficos visualmente mais adequados, usa técnicas grafos para análise de redes e a dificuldade que a falta de padronização pode causar, principalmente quando existem diversos arquivos para analisar de diferentes fontes.

Tendo como base modelo do relatório disponibilizados no aprender UnB da disciplina e implementando os scripts em R com técnicas de análise de redes com igraph e manipulação de dados em R, análise crítica das melhores formas de gerar os gráficos de acordo com os tipos de dados, foi possível fazer uma análise de dados dos datasets de Geotecnia seguindo as 6 fases do CRISP-DM.

Por fim, os resultados obtidos podem ser considerados relevantes e bem-sucedidos para o conjunto de dados, e os resultados são coerentes ao comparar as pontuações de avaliação quadrienal, que colocam o programa Geotecnia como um programa nota 6 que possui um número alto de publicações acadêmicas e orientações acadêmicas.

Referências

- Bharat, Mishra. 2019. “**Understanding Crisp-Dm Using Video Game Sales Data.**” <https://medium.com/@imBharatMishra/understanding-crisp-dm-using-video-game-sales-data-a2d55c7a2593>.
- Capes. 2006. “**Sucupira.**” <https://sucupira.capes.gov.br/sucupira/>.
- CRISP-DM. 2006. “**CRISP-Dm.**” https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm.
- Fernandes, Jorge H C, and Ricardo Barros Sampaio. 2011. “**Unb.elattes.**” <http://unb.elattes.com.br/>.
- Geotecnia, Sucupira. 2013. “**Sucupira Geotecnia.**” https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/programa/viewPrograma.jsf?popup=true&cd_programa=53001010032P2.
- Sucupira. 2006. “**Sucupira.**” <https://www.capes.gov.br/acessoainformacao/91-conteudo-estatico/avaliacao-capes/6871-caracterizacao-do-sistema-de-avaliacao-da-pos-graduacao>.
- Todamateria. 2006. “**O Que é Ciência?**” <https://www.todamateria.com.br/o-que-e-ciencia/>.
- UnB. 2006. “**UnB.**” <https://www.unb.br/pesquisa>.