Definir JSONS

```
profile <- c("~/Documents/eLattes/Fiocruz e Zica/eLattes output/saida/profile.json")
advise  <- c("~/Documents/eLattes/Fiocruz e Zica/eLattes output/saida/advise.json")
publication <- c("~/Documents/eLattes/Fiocruz e Zica/eLattes output/saida/publication.json")
list <- c("~/Documents/eLattes/Fiocruz e Zica/eLattes output/saida/zika.list.json")
graph  <- c("~/Documents/eLattes/Fiocruz e Zica/eLattes output/saida/graph.json")
```

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 pesquisadores científicos que produzem estudos e conhecimentos que possam levar a desenvolvimentos significativos no combate **Zika**.

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 que tratam sobre **Zika** são:

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

Análise estrutural dos dados

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

```
library(jsonlite) #Importado para lidar com arquivos com extensão JSON
library(listviewer) #Importado para lidar com listas
library(ggplot2) #Importado para realizar visualizações
library(tidyr) #Importado par utilizar funções relacionadas a dataframes
library(dplyr)
```

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

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

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

```
library(readxl)
library(stringr)
```

A importação dos dados é mostrada como a seguir, utilizando a função `fromJSON()` do pacote `jsonlite`.

```
zika.profile <- fromJSON(profile)
zika.advise <- fromJSON(advise)
zika.publication <- fromJSON(publication)
zika.list <- fromJSON(list)
zika.graph <- fromJSON(graph)
```

Ao realizar a importação dos arquivos em formato `json`, é possível identificar seu formato, uma lista (ou "large list") de objetos JSON. Esse formato não é o mais simples de ser análisado utilizando as funções e bibliotecas da linguagem R, porém em breve, uma série de passos será aplicada para que sejam simplificadas.

Por meio da biblioteca `dplyr` é possível realizar uma série de verificações relacionadas à disposição dos dados ao serem importados.

Descrição dos dados do Perfil

As seguintes análises são realizadas para os arquivos de perfis.

Por meio do tamanho da lista gerada ao importar os arquivos, é possível obter a quantidade de pesquisadores científicos que produzem estudos e conhecimentos que possam levar a desenvolvimentos significativos no combate **Zika**. *Quantidade de pesquisadores na base*

```
length(zika.profile)
```

```
## [1] 1035
```

Por meio da função `glimpse`, é possível verificar uma apresentação inicial dos dados de perfil pesquisadores. O código a seguir mostra os atributos presentes em um dos pesquisadores.

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

```
## List of 7
##  $ nome                  : chr "Luciana Souza De Arag\u00e3o Fran\u00e7a"
##  $ resumo_cv             : chr "Possui gradua\u00e7\u00e3o em Ci\u00eancias Biol\u00f3gicas Modalidad
##  $ areas_de_atuacao      :'data.frame':  4 obs. of  4 variables:
##   ..$ grande_area  : chr [1:4] "CIENCIAS_BIOLOGICAS" "CIENCIAS_BIOLOGICAS" "CIENCIAS_BIOLOGICAS" "CI
##   ..$ area         : chr [1:4] "Imunologia" "Imunologia" "Parasitologia" "Biotecnologia"
##   ..$ sub_area     : chr [1:4] "Imunologia Aplicada" "Imunologia Celular" "" "Biotecnologia em Sa\u0
##   ..$ especialidade: chr [1:4] "" "" "" ""
##  $ endereco_profissional :List of 8
##   ..$ instituicao: chr "Hospital S\u00e3o Rafael"
##   ..$ orgao      : chr "Centro de Biotecnologia e Terapia Celular"
##   ..$ unidade    : chr ""
##   ..$ DDD        : chr "71"
##   ..$ telefone   : chr "32816970"
##   ..$ bairro     : chr "S\u00e3o Marcos"
##   ..$ cep        : chr "41253900"
##   ..$ cidade     : chr "Salvador"
##  $ producao_bibiografica :List of 1
##   ..$ PERIODICO:'data.frame':    2 obs. of  10 variables:
##   .. ..$ natureza        : chr [1:2] "COMPLETO" "COMPLETO"
##   .. ..$ titulo          : chr [1:2] "Protective effects of mito-TEMPO against doxorubicin cardiotox
```

```
##   .. ..$ periodico       : chr [1:2] "Cancer Chemotherapy and Pharmacology" "AMERICAN JOURNAL OF PATH
##   .. ..$ ano             : chr [1:2] "2015" "2017"
##   .. ..$ volume          : chr [1:2] "77" "187"
##   .. ..$ issn            : chr [1:2] "03445704" "00029440"
##   .. ..$ paginas         : chr [1:2] "659 - " "1134 - 1146"
##   .. ..$ doi             : chr [1:2] "10.1007/s00280-015-2949-7" "10.1016/j.ajpath.2017.01.016"
##   .. ..$ autores         :List of 2
##   .. ..$ autores-endogeno:List of 2
##  $ orientacoes_academicas:List of 1
##   ..$ ORIENTACAO_EM_ANDAMENTO_MESTRADO:'data.frame': 1 obs. of  13 variables:
##   .. ..$ natureza                   : chr "Disserta\u00e7\u00e3o de mestrado"
##   .. ..$ titulo                     : chr "Efeito imunomodulat\u00f3rio do N-acil-hydrazone HAH2 em m
##   .. ..$ ano                        : chr "2015"
##   .. ..$ id_lattes_aluno            : chr ""
##   .. ..$ nome_aluno                 : chr "J\u00e9ssica Vieira Cerqueira"
##   .. ..$ instituicao                : chr "Universidade Federal da Bahia"
##   .. ..$ curso                      : chr "Programa de p\u00f3s gradua\u00e7\u00e3o em Imunologia"
##   .. ..$ codigo_do_curso            : chr "90000048"
##   .. ..$ bolsa                      : chr "SIM"
##   .. ..$ agencia_financiadora       : chr "Coordena\u00e7\u00e3o de Aperfei\u00e7oamento de Pessoal d
##   .. ..$ codigo_agencia_financiadora: chr "045000000000"
##   .. ..$ nome_orientadores          :List of 1
##   .. ..$ id_lattes_orientadores     :List of 1
##  $ senioridade           : chr "4"
```

Uma breve inspeção visual dos atributos anteriormente apresentados permite inferir que o pesquisador **x**, sob análise: 1. A 2. B 3. C 4. D 5. E

Descrição dos dados de orientações

As seguintes descrições possuem, para simplificar, informações relativas às orientações concluídas nos anos 2016 e 2017.

Retorna campos dentro do arquivo ***zika.advise***

```
names(zika.advise)
```

```
## [1] "ORIENTACAO_EM_ANDAMENTO_DE_POS_DOUTORADO"
## [2] "ORIENTACAO_EM_ANDAMENTO_DOUTORADO"
## [3] "ORIENTACAO_EM_ANDAMENTO_MESTRADO"
## [4] "ORIENTACAO_CONCLUIDA_POS_DOUTORADO"
## [5] "ORIENTACAO_CONCLUIDA_DOUTORADO"
## [6] "ORIENTACAO_CONCLUIDA_MESTRADO"
```

Quantidade de orientações de Mestrado concluídas em 2017

```
length(zika.advise$ORIENTACAO_CONCLUIDA_MESTRADO$`2017`$natureza)
```

```
## [1] 371
```

Quantidade de orientações de Doutorado concluídas em 2017

```
length(zika.advise$ORIENTACAO_CONCLUIDA_DOUTORADO$`2017`$natureza)
```

```
## [1] 270
```

Cursos presentes nos Doutorados concluídos em 2017

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

```
##
##              Ci<U+00EA>ncias da Sa<U+00FA>de
##                                           12
##                              Medicina Tropical
##                                           12
##                         Sa<U+00FA>de Coletiva
##                                            7
##                    Sa<U+00FA>de P<U+00FA>blica
##                                            7
##                 Ci<U+00EA>ncias M<U+00E9>dicas
##                                            6
##                               Gen<U+00E9>tica
##                                            6
##    Inova<U+00E7><U+00E3>o Terap<U+00EA>utica
##                                            6
##                  Biologia Celular e Molecular
##                                            5
##              Ci<U+00EA>ncias (Microbiologia)
##                                            5
## Doen<U+00E7>as Infecciosas e Parasit<U+00E1>rias
##                                            5
```

Cursos presentes nos Mestrados concluídos em 2017

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

```
##
##                          Sa<U+00FA>de P<U+00FA>blica
##                                                 18
##                               Sa<U+00FA>de Coletiva
##                                                 14
##                       Ci<U+00EA>ncias da Sa<U+00FA>de
##                                                 11
##                  Dist<U+00FA>rbios do Desenvolvimento
##                                                  9
##                                    Medicina Tropical
##                                                  9
##                          Biologia Parasit<U+00E1>ria
##                                                  7
##                  Ci<U+00EA>ncias Farmac<U+00EA>uticas
##                                                  7
##                                           Virologia
##                                                  7
##                          Biologia Celular e Molecular
##                                                  6
## Biologia de Agentes Infecciosos e Parasit<U+00E1>rios
##                                                  5
```

Descrição dos dados de produção bibliográfica

Quanto aos dados de publicações, as descrições a seguir mostram informações sobre os tipos de produções bibliográficas presentes no arquivo juntantente a títulos de periódicos publicados no ano de 2017.

Retorna campos com tipos de publicação dentro de ***zika.publication***

```
names(zika.publication)
```

```
## [1] "PERIODICO"         "LIVRO"              "CAPITULO_DE_LIVRO"
## [4] "EVENTO"            "ARTIGO_ACEITO"
```

Retorna campos dos periódicos de 2017

```
names(zika.publication$PERIODICO$`2017`)
```

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

Mostra o título dos periódicos publicados em 2017

```
head(sort(table(zika.publication$PERIODICO$`2017`$titulo), decreasing = TRUE), 10)
```

```
##
##                                                 Cardiorespiratory performance of coronary artery diseas
##
##                                                                                    Clinical, laboratory and
##
##                                                                                Early Evidence for Zika Virus
##
##
##
##                         New Virus Genome Sequences of the Guama Serogroup (Genus <i>Orthobunyavirus</i
##
##                                      Bienestar en invertebrados: un par<U+00E1>metro <U+00E9>tico de
##
##
##
##                                  &#916;Np73 overexpression promotes resistance to apoptosis but does
##
## &#945;-Ketoadipic Acid and &#945;-Aminoadipic Acid Cause Disturbance of Glutamatergic Neurotransmissi
##
##              &amp;lt;i&amp;gt;Mycobacterium bovis&amp;lt;/i&amp;gt; BCG as a Delivery System for the
##
```

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.

ZIKA

```
# quantidade de periódicos publicados sobre ZIKA entre 2010 a 2017
infperiodicosZika <- data.frame()
for(i in 1:length(zika.publication$PERIODICO))
infperiodicosZika <- rbind(infperiodicosZika, zika.publication$PERIODICO[[i]])
```

```
# quantidade de livros publicados relacionados a Zika entre 2010 a 2017
inflivrosZika <- data.frame()
for(i in 1:length(zika.publication$LIVRO))
inflivrosZika <- rbind(inflivrosZika, zika.publication$LIVRO[[i]])
```

```
# produção de livros de cada integrante do ZIKA no ano de 2017
autoreslivrosZika <- zika.publication[["LIVRO"]][["2017"]][["autores"]]
tabelaautoresZika <- table(unlist(autoreslivrosZika))
dfautoresZika <- data.frame(tabelaautoresZika)
```

```
# quantidade de periódicos publicados sobre Zika entre 2010 a 2017
infeventosZika <- data.frame()
for(i in 1:length(zika.publication$EVENTO))
infeventosZika <- rbind(infeventosZika, zika.publication$EVENTO[[i]])
```

```
# Monta data frame com orientações completas
orientacaoCompletaZika <- bind_rows(zika.advise$ORIENTACAO_CONCLUIDA_DOUTORADO)
orientacaoCompletaZika <- bind_rows(orientacaoCompletaZika,
zika.advise$ORIENTACAO_CONCLUIDA_POS_DOUTORADO)
orientacaoCompletaZika <- bind_rows(orientacaoCompletaZika,
zika.advise$ORIENTACAO_CONCLUIDA_MESTRADO)
orientacaoCompletaZika <- bind_rows(orientacaoCompletaZika,
zika.advise$OUTRAS_ORIENTACOES_CONCLUIDAS)
```

```
# Monta data frame com orientações incompletas
orientacaoIncompletaZika <-bind_rows(zika.advise$ORIENTACAO_EM_ANDAMENTO_DE_POS_DOUTORADO)
orientacaoIncompletaZika <- bind_rows(orientacaoIncompletaZika,
zika.advise$ORIENTACAO_EM_ANDAMENTO_DOUTORADO)
orientacaoIncompletaZika <- bind_rows(orientacaoIncompletaZika,
```

```
zika.advise$ORIENTACAO_EM_ANDAMENTO_MESTRADO)
orientacaoIncompletaZika <- bind_rows(orientacaoIncompletaZika,
zika.advise$ORIENTACAO_EM_ANDAMENTO_GRADUACAO)
```

```
# data frames com orientações completas e incompletas
orientacaoZika.df <- bind_rows(orientacaoIncompletaZika, orientacaoCompletaZika)
df_orientadores1Zika <- as.data.frame(sort(table(unlist(
orientacaoZika.df$id_lattes_orientadores, recursive = TRUE)),decreasing=TRUE))
```

```
# participação de professores em eventos de cunho internacional
pubZika <- data.frame()
for(profile in zika.profile) {
if(!is.null(profile$`producao_bibiografica`$`EVENTO`)) {
temp <- profile$`producao_bibiografica`$`EVENTO` %>%
filter(classificacao == "INTERNACIONAL")
pubZika <- rbind(pubZika, temp)
}
}
```

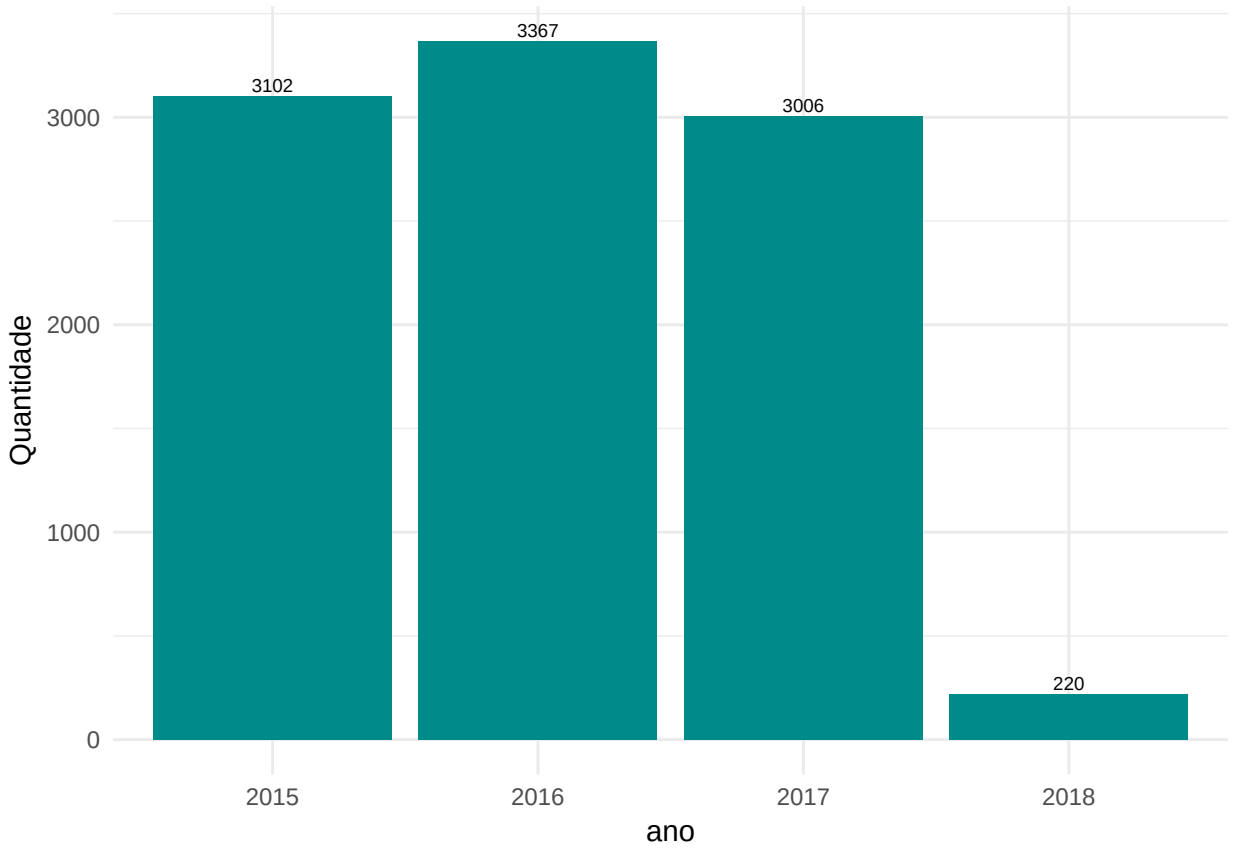
```
# periódicos que possuem mais artigos publicados
perZika <- data.frame()
for (profile in zika.profile) {
if(!is.null(profile$producao_bibiografica$PERIODICO)) {
temp <- profile$`producao_bibiografica`$PERIODICO
perZika <- rbind(perZika, temp)
}
}
top10Zika <- perZika %>%
filter(ano >= 2012) %>%
count(periodico) %>%
top_n(10) %>%
arrange(n, periodico) %>%
mutate(periodico = factor(periodico, levels = unique(periodico)))
```

```
## Selecting by n
```

Resultados

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

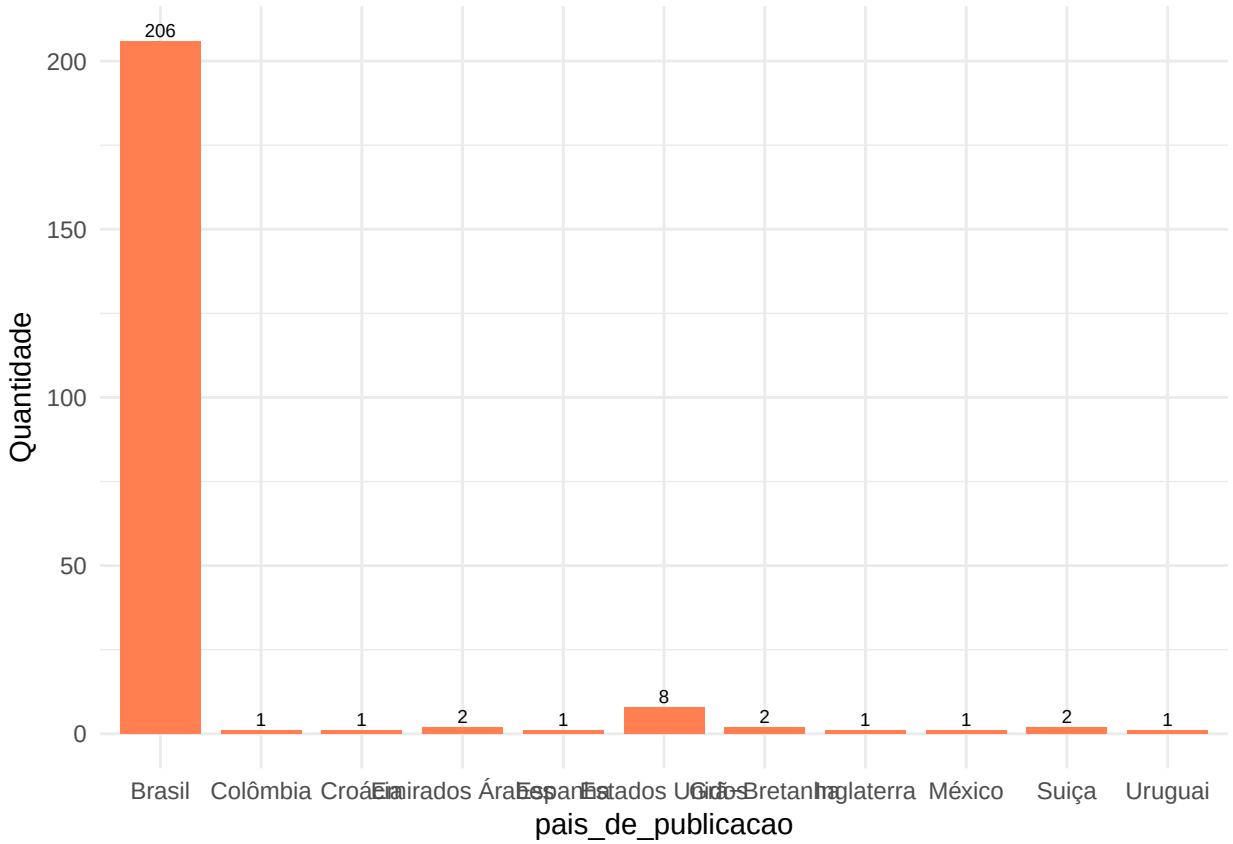
```
infperiodicosZika %>%
group_by(ano) %>%
summarise(Quantidade = n()) %>%
ggplot(aes(x = ano, y = Quantidade)) +
geom_bar(position = "stack",stat = "identity", fill = "darkcyan")+
geom_text(aes(label=Quantidade), vjust=-0.3, size=2.5)+
theme_minimal()
```



É possível notar que o **ZIKA** publicou **x** nos últimos anos. Percebe-se que o máximo de publicações, durante o período estudado, encontra-se no ano de 2016. Isso pode indicar que existem pesquisas com um bom aproveitamento nesse período.

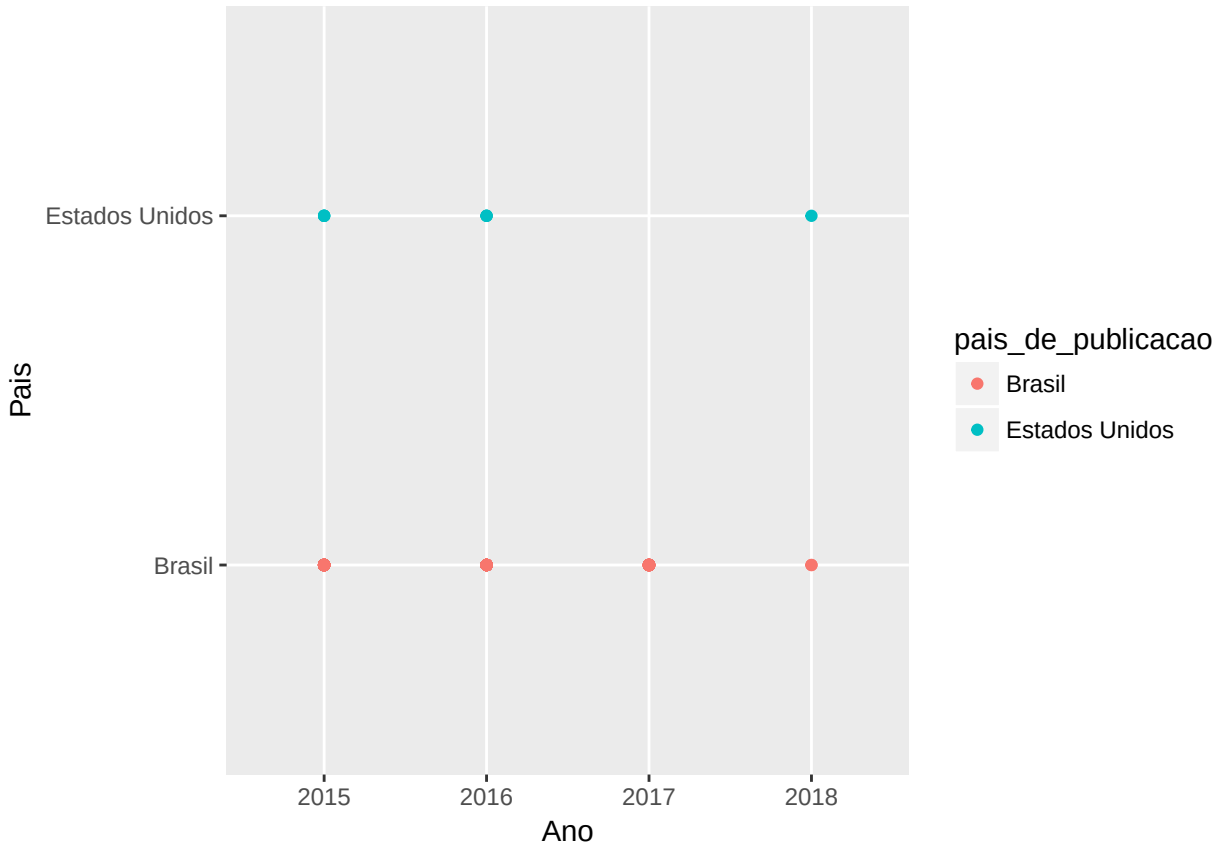
```
inflivrosZika %>%
group_by(pais_de_publicacao) %>%
summarise(Quantidade = n()) %>%
ggplot(aes(x = pais_de_publicacao, y = Quantidade)) +
geom_bar(width=0.8, height = 0.3, position = "stack",stat = "identity", fill = "coral")+
geom_text(aes(label=Quantidade), vjust=-0.3, size=2.5) +
theme_minimal()
```

```
## Warning: Ignoring unknown parameters: height
```



Com relação aos livros publicados no contexto do **ZIKA**, nota-se que existe um número considerável(principalmente se o número de livros publicados em território nacional for levado em conta) de livros publicados no exterior. Com 6 livros publicados nos Estados Unidos nos últimos anos, além de outros países como Alemanha e Holanda, onde há um foco grande na área da computação, é possível perceber que o programa tem caminhado para um ótimo nível de internacionalização, pelo menos no quesito dos livros.

```
inflivrosZika %>%
filter(pais_de_publicacao %in% c("Brasil", "Estados Unidos", "Holanda",
"Grã-Bretanha", "Alemanha", "Suiça")) %>%
group_by(ano,pais_de_publicacao) %>%
ggplot(aes(x=ano,y=pais_de_publicacao, color= pais_de_publicacao)) +
xlab("Ano") + ylab("Pais") + geom_point()
```



Quanto à perspectiva dos livros sendo publicados em diferentes países nos últimos anos, é possível ver que houve um aumento na quantidade de localidades. O ano de 2017 parece ser o auge dessa quantidade, com livros sendo publicados em até 4 países diferentes.

```
ggplot(dfautoresZika, aes(x = Freq, y = Var1, color=Var1)) +
xlab("QUANTIDADE") + ylab("Autor") + geom_point()
```

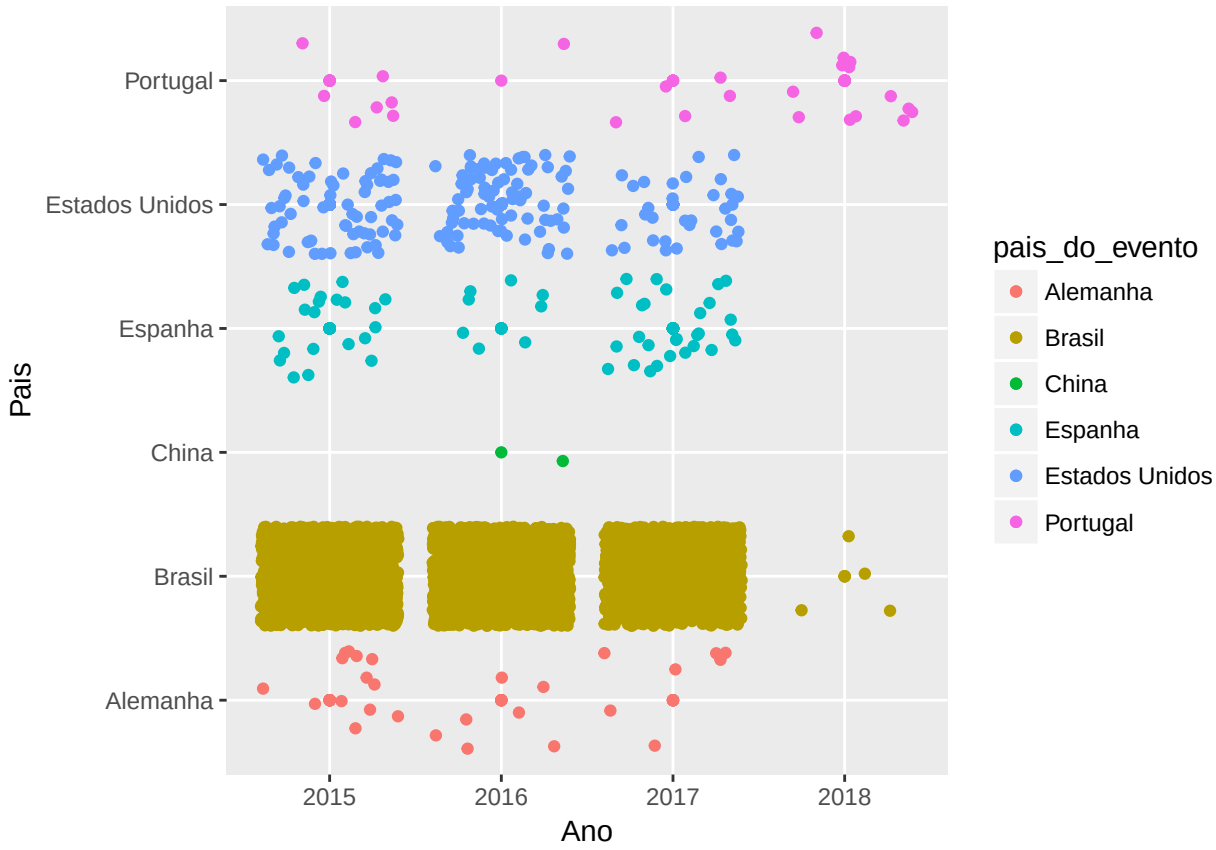
```
## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0x7f

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0x7f
```

FROTA, Mirna Albuquerque.	LACERDA E SILVA, THAIS	MOLEN
GÜNTHER, Wanda Maria Risso	LANA, R. M.	MONTE
GIL, L. H. V.	LARA, A.P.M	MONTE
GOMES−GOUVÊA, MICHELE SOARES	LENT, R.	MORAE
GOMEZ, R S	LUZ, M. T.	MOREIF
GONCALVES, J. C.	Leite, P. M.	MOREIF
GONCALVES, S. J. C.	Lima, E. O.	MOTA, M
GONCALVEZ, Juliana	MACHADO, A. S.	MOTTA,
GRISOTTI, M.	MAIA, T. F.	MUTO, M
GRYSCHEK, RONALDO CESAR BORGES	MALLOY−DINIZ, L. F.	Malheiro
GUERRA, M. J. C.	MARQUES, R.	Marcond
GURGEL, I. G. D.	MARTINS, G. G. Z.	Martins,
Giugliani, E.R.J.	MASSARA, C. L.	McINTY
HADDAD, L. B. P.	MASSARANI, Luisa	Nóbrega
HONORIO, N. A.	MATOS, V. P.	NASCIM
INACIO, Luiz Gustavo	MAYRINK, M. C.	NASCIM
JACOBI, P. R.	MELO, E. S.	NASTRI
Jorge Rezende Filho	MENDES−CORRÊA, MARIA CÁSSIA JACINTHO	NAVARF
KIKUCHI, L.	MENDONCA, M. A.	NEPOTE

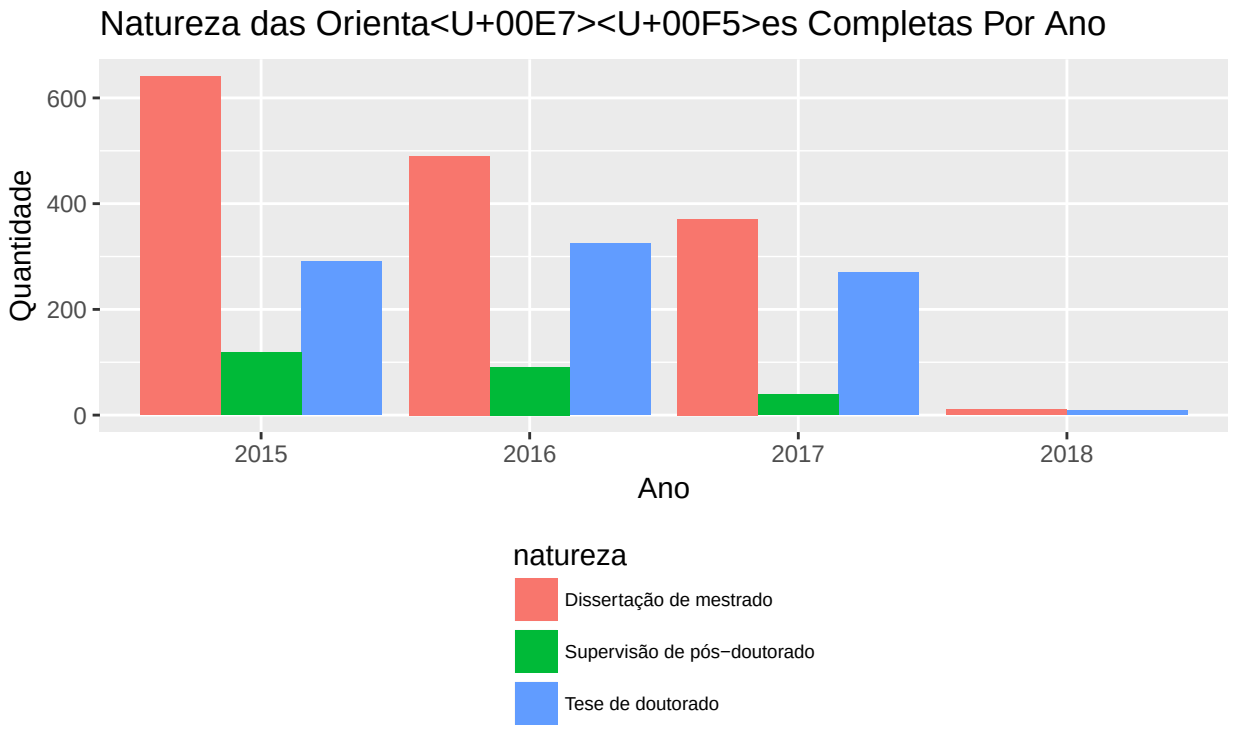
O gráfico acima mostra os autores de livros em função do número de suas publicações.

```
infeventosZika %>%
filter(pais_do_evento %in% c("Brasil", "Estados Unidos", "Japão", "Colômbia",
"Venezuela", "Portugal", "Grã-Bretanha", "França",
"Espanha", "China", "Alemanha")) %>%
group_by(ano_do_trabalho,pais_do_evento) %>%
ggplot(aes(x=ano_do_trabalho,y=pais_do_evento, color= pais_do_evento)) +
xlab("Ano") + ylab("Pais") + geom_point() + geom_jitter()
```



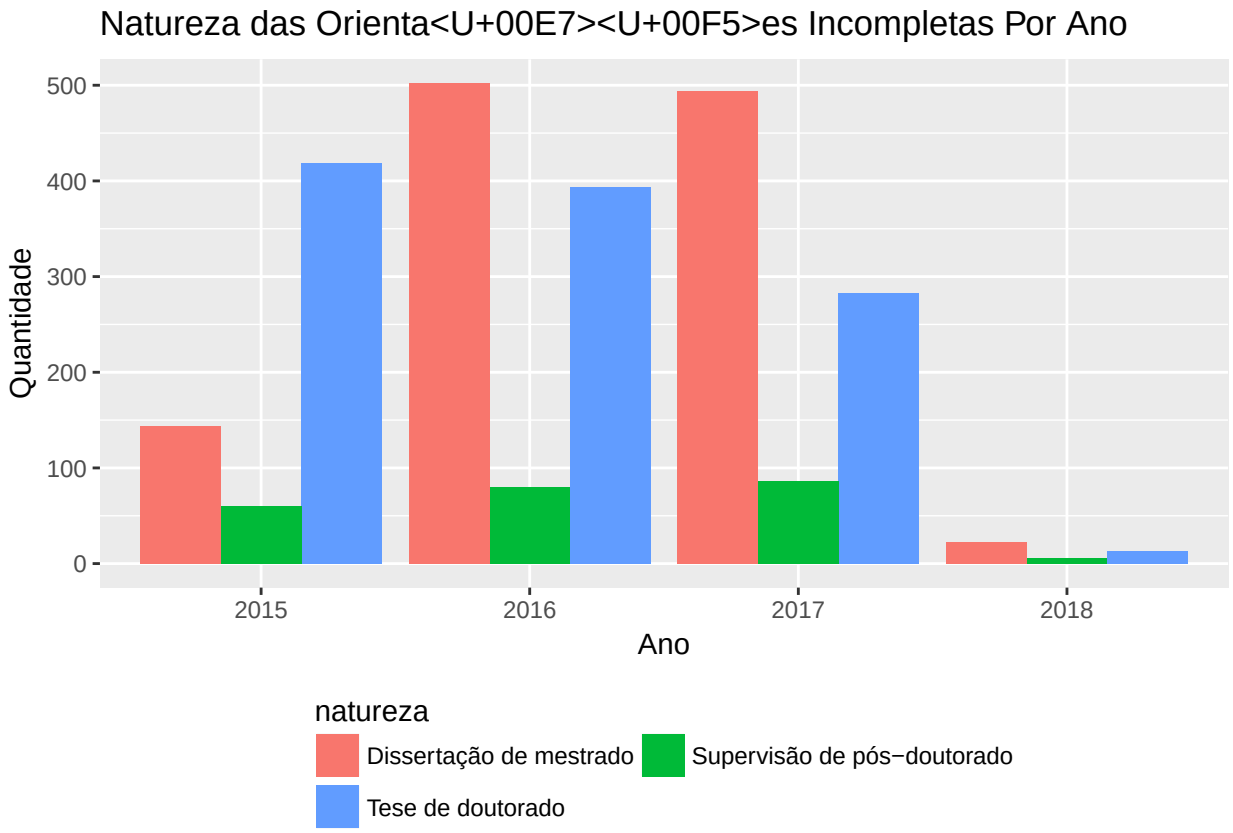
Quanto aos eventos em que houveram participações dos membros do do PPGInf, nota-se uma concentração grande de cunho nacional, com um foco internacional nos Estados Unidos.

```
ggplot(orientacaoCompletaZika,aes(ano,fill=natureza)) +
geom_bar(stat = "count", position="dodge") +
ggtitle("Natureza das Orientações Completas Por Ano") +
theme(legend.position="bottom",legend.text=element_text(size=7)) +
guides(fill=guide_legend(nrow=5, byrow=TRUE, title.position = "top")) +
labs(x="Ano",y="Quantidade")
```



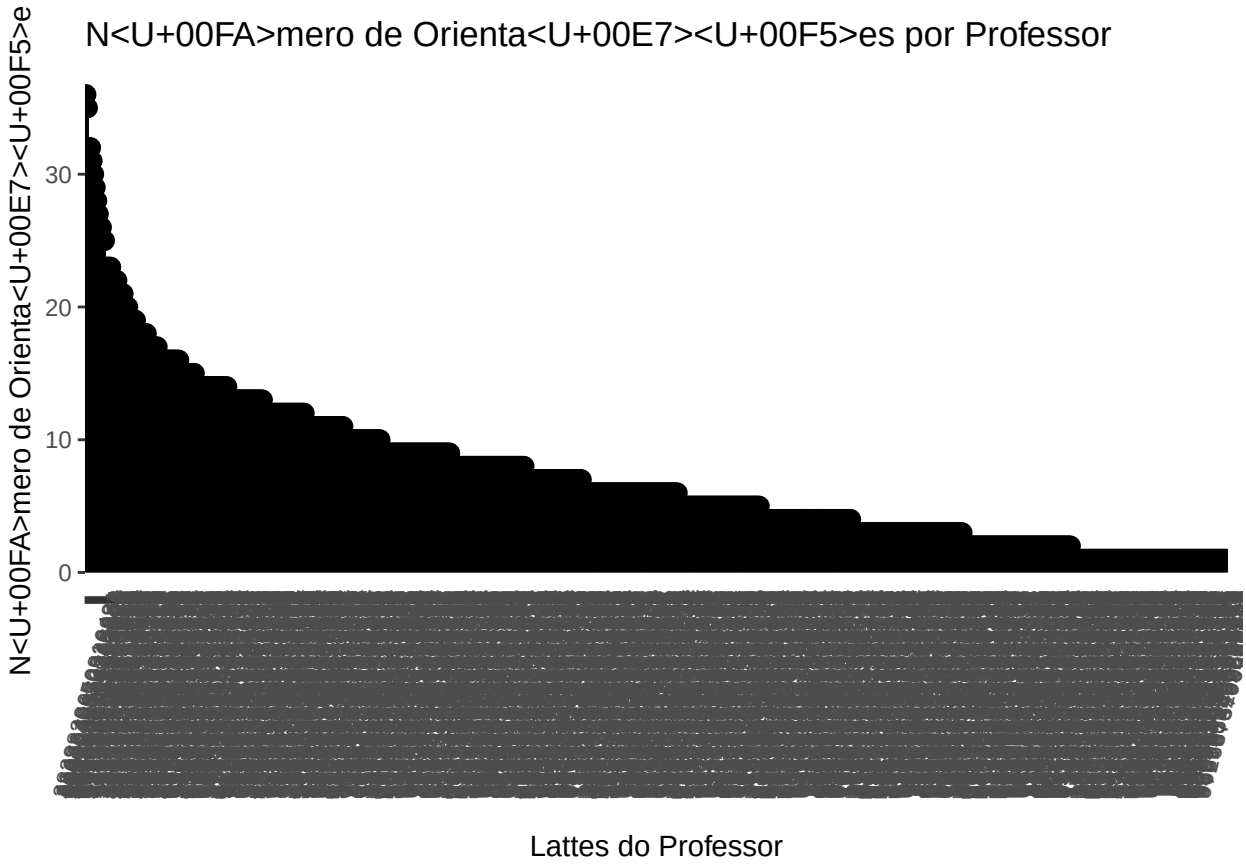
Em relação às orientações completadas durante os últimos anos, nota-se que o foco maior dos professores envolvidos é o trabalho de conclusão de curso. Isso se dá, provavelmente por uma questão numérica. O número de alunos de graduação é naturalmente maior que os de pós graduação. Esses professores que compõem o PPGInf também são, em maioria, professores da graduação. Dado isso, essa disparidade é esperada. Quanto ao número de dissertações de mestrado e doutorado, que é o foco desse programa, nota-se um bom número de orientações dessa natureza, principalmente da primeira mencionada.

```
ggplot(orientacaoIncompletaZika,aes(ano,fill=natureza)) +
geom_bar(stat = "count", position="dodge") +
ggtitle("Natureza das Orientações Incompletas Por Ano") +
theme(legend.position="bottom",legend.text=element_text(size=9)) +
guides(fill=guide_legend(nrow=2, byrow=TRUE, title.position = "top")) +
labs(x="Ano",y="Quantidade")
```



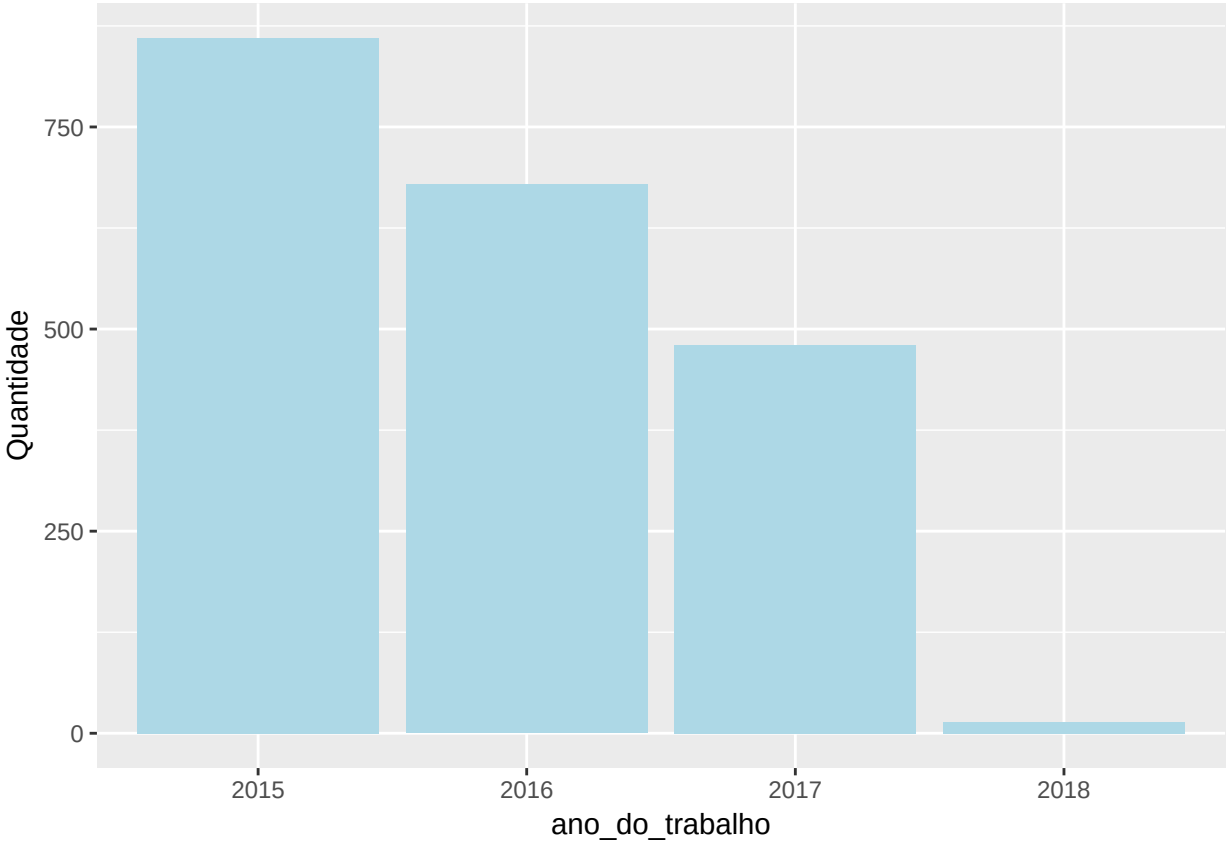
Quanto às orientações incompletas, é possível perceber que o número de orientações de doutorado que aparece nesse gráfico é relativamente maior do que o número que apareceu no gráfico anterior. Um forte indicador do porquê disso acontecer é a maior complexidade de uma pesquisa necessária para a escrita de tese de doutorado.

```
ggplot(df_orientadores1Zika, aes(x=Var1, y=Freq)) +
geom_point(size=3) +
geom_segment(aes(x=Var1,
xend=Var1,
y=0,
yend=Freq)) +
labs(title="Número de Orientações por Professor", x="Lattes do Professor",
y="Número de Orientações") +
theme(axis.text.x = element_text(angle=75, vjust=0.6))
```



O gráfico acima mostra o número de orientações por professor com base em seu identificador de Lattes. Percebe-se que o número de orientações é bastante variado.

```
pubZika %>%
group_by(ano_do_trabalho) %>%
summarise(Quantidade=n()) %>%
ggplot(aes(x=ano_do_trabalho, y=Quantidade)) +
geom_bar(stat="identity", fill="lightblue")
```



O gráfico acima mostra a quantidade geral de publicações do grupo, de todas as naturezas.

Conclusão

Por meio desse trabalho, foi possível ter uma noção mais real de como é trabalhar com a ciência de dados e deixar ainda mais claro a importância de ferramentas de análise. Foi possível perceber como boa parte do tempo empregado é gasto analisando a estrutura dos dados obtidos e bolando estratégias para que se adquira aquilo que realmente importa. Também, há de se ressaltar, como é gratificante atingir o objetivo que se planeja ao iniciar a análise, possibilitando um entendimento maior do que se está estudando. No caso, foi possível diminuir o escopo dos nossos dados, filtrar por informações particulares, plotar gráficos para facilitar a visualização, e, de forma geral, conhecer um volume de dados expressivo com uma pequena análise quantitativa e qualitativa feita em cima dele.