

ME115 - Trabalho Final

159874 - Maria Helena G. B. Reis 170409 - Elizabeth Borgognoni Souto

15 de junho de 2018

Introdução

O Censo Escolar é um levantamento de dados estatístico-educacionais de âmbito nacional realizado todos os anos e coordenado pelo Inep. Ele é feito com a colaboração das secretarias estaduais e municipais de Educação e com a participação de todas as escolas públicas e privadas do país.

Trata-se do principal instrumento de coleta de informações da educação básica, que abrange as suas diferentes etapas e modalidades: ensino regular (educação Infantil e ensinos fundamental e médio), educação especial e educação de jovens e adultos (EJA). O Censo Escolar coleta dados sobre estabelecimentos, matrículas, funções docentes, movimento e rendimento escolar.

Essas informações são utilizadas para traçar um panorama nacional da educação básica e servem de referência para a formulação de políticas públicas e execução de programas na área da educação, incluindo os de transferência de recursos públicos como merenda e transporte escolar, distribuição de livros e uniformes, implantação de bibliotecas, instalação de energia elétrica, Dinheiro Direto na Escola e Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (Fundeb).

Além disso, os resultados obtidos no Censo Escolar sobre o rendimento (aprovação e reprovação) e movimento (abandono) escolar dos alunos do ensino Fundamental e Médio, juntamente com outras avaliações do Inep (Saeb e Prova Brasil), são utilizados para o cálculo do Índice de Desenvolvimento da Educação Básica (IDEB), indicador que serve de referência para as metas do Plano de Desenvolvimento da Educação (PDE), do Ministério da Educação.

Iremos trabalhar com as informações das escolas do país, que estão no banco de dados ‘Microdados Censo Escolar 2017’ disponibilizado por Marcelo Pacote na plataforma Kaggle, disponível através do link: <https://www.kaggle.com/marcelopacote/microdados-censo-escolar-2017/version/1>.

A tabela ESCOLAS.CSV contém informações sobre as escolas cadastradas no censo 2017 no Brasil: ESCOLAS.CSV

Proposta de análise: Como está o funcionamento das escolas no país em diferentes dependências administrativas?

Iremos dividir nossa análise em 4 etapas, dadas da seguinte forma:

1. Relação do número de salas de aula existentes e utilizadas nas escolas por dependência administrativa
2. Relação do numero de funcionários das escolas por dependência administrativa
3. Recursos relacionados à infraestrutura
4. Relação do destino do lixo das escolas (reciclagem) por localização (rural/urbana)
5. Relação do abastecimento de água por localização (rural/urbana)

Leitura dos dados

```
## [1] "Classe: tbl      Variáveis: 166      Observações: 282362"
```

Manipulação e limpeza dos dados

Analisamos as variáveis da tabela escola e escolhemos estas variáveis para trabalharmos, 3 variáveis numéricas e o restante variáveis categóricas, esta escolha foi feita devido ao banco de dados *ESCOLAS.CSV* ter muitas variáveis categóricas e poucas variáveis numéricas para análise.

```
## Observations: 282,362
## Variables: 23
## $ TP_DEPENDENCIA <int> 3, 2, 3, 3, 3, 2, 2, 3, 3, 3, 3, 3, ...
## $ TP_LOCALIZACAO <int> 1, 1, 2, 1, 2, 1, 1, 2, 1, 1, 1, 2, ...
## $ NU_SALAS_EXISTENTES <int> 10, 26, 1, 9, 4, 13, 14, 3, 2, 8, 4...
## $ NU_SALAS_UTILIZADAS <int> 10, 26, 1, 9, 4, 12, 12, 3, 5, 7, 4...
## $ NU_FUNCIONARIOS <int> 62, 81, 2, 16, 5, 88, 40, 11, 25, 4...
## $ NU_COMPUTADOR <int> 4, 39, 0, 2, 0, 40, 52, 0, 1, 5, 1, ...
## $ IN_SALA_DIRETORIA <int> 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, ...
## $ IN_SALA_PROFESSOR <int> 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, ...
## $ IN_BIBLIOTECA_SALA_LEITURA <int> 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, ...
## $ IN_AREA_VERDE <int> 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, ...
## $ IN_LABORATORIO_INFORMATICA <int> 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, ...
## $ IN_LABORATORIO_CIENCIAS <int> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...
## $ IN_AGUA_REDE_PUBLICA <int> 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, ...
## $ IN_AGUA_INEXISTENTE <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_LIXO_COLETA_PERIODICA <int> 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, ...
## $ IN_LIXO_QUEIMA <int> 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, ...
## $ IN_AGUA_POCO_ARTESIANO <int> 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_AGUA_CACIMBA <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, ...
## $ IN_AGUA_FONTE_RIO <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_LIXO_ENTERRA <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_LIXO_OUTROS <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_LIXO_JOGA_OUTRA_AREA <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_LIXO_RECICLA <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...
```

Manipulamos os dados que estavam desorganizados com o pacote *dplyr* para melhor compreensão.

```
## Observations: 282,362
## Variables: 23
## $ TP_DEPENDENCIA <fct> Municipal, Estadual, Municipal, Mun...
## $ TP_LOCALIZACAO <fct> Urbana, Urbana, Rural, Urbana, Rura...
## $ NU_SALAS_EXISTENTES <int> 10, 26, 1, 9, 4, 13, 14, 3, 2, 8, 4...
## $ NU_SALAS_UTILIZADAS <int> 10, 26, 1, 9, 4, 12, 12, 3, 5, 7, 4...
## $ NU_FUNCIONARIOS <int> 62, 81, 2, 16, 5, 88, 40, 11, 25, 4...
## $ NU_COMPUTADOR <int> 4, 39, 0, 2, 0, 40, 52, 0, 1, 5, 1, ...
## $ IN_SALA_DIRETORIA <fct> 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, ...
## $ IN_SALA_PROFESSOR <fct> 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, ...
## $ IN_BIBLIOTECA_SALA_LEITURA <fct> 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, ...
## $ IN_AREA_VERDE <fct> 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, ...
## $ IN_LABORATORIO_INFORMATICA <fct> 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, ...
## $ IN_LABORATORIO_CIENCIAS <fct> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...
## $ IN_AGUA_REDE_PUBLICA <fct> 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, ...
## $ IN_AGUA_INEXISTENTE <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_LIXO_COLETA_PERIODICA <fct> 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, ...
## $ IN_LIXO_QUEIMA <fct> 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, ...
## $ IN_AGUA_POCO_ARTESIANO <fct> 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_AGUA_CACIMBA <fct> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, ...
## $ IN_AGUA_FONTE_RIO <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
```

```

## $ IN_LIXO_ENTERRA           <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_LIXO_OUTROS            <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_LIXO_JOGA_OUTRA_AREA  <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ IN_LIXO_RECICLA           <fct> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...

```

Analises estatísticas dos dados numéricos

```

##   TP_DEPENDENCIA  TP_LOCALIZACAO  NU_SALAS_EXISTENTES
## Estadual : 37668  Rural :123439    Min.   : 1.00
## Federal  : 766   Urbana:158923   1st Qu.: 4.00
## Municipal:180248                         Median : 7.00
## Privada  : 63680                         Mean   : 8.26
##                                         3rd Qu.: 11.00
##                                         Max.   :2504.00
##                                         NA's   :96906
##   NU_SALAS_UTILIZADAS  NU_FUNCIONARIOS  NU_COMPUTADOR  IN_SALA_DIRETORIA
##   Min.   : 1.00       Min.   : 1.00       Min.   : 0.00   0   : 57416
##   1st Qu.: 3.00       1st Qu.: 10.00      1st Qu.: 1.00   1   :128509
##   Median : 6.00       Median : 23.00      Median : 3.00   NA's: 96437
##   Mean   : 7.77       Mean   : 31.35      Mean   : 10.73
##   3rd Qu.: 10.00      3rd Qu.: 42.00      3rd Qu.: 15.00
##   Max.   :2504.00      Max.   :3636.00      Max.   :3188.00
##   NA's   :96437        NA's   :96437        NA's   :96437
##   IN_SALA_PROFESSOR  IN_BIBLIOTECA_SALA_LEITURA  IN_AREA_VERDE
##   0   : 77736          0   :91684          0   :132495
##   1   :108189          1   :94241          1   : 53430
##   NA's: 96437          NA's:96437          NA's: 96437
##
##   IN_LABORATORIO_INFORMATICA  IN_LABORATORIO_CIENCIAS  IN_AGUA_REDE_PUBLICA
##   0   :111832             0   :166214             0   : 49663
##   1   : 74093              1   : 19711              1   :136262
##   NA's: 96437              NA's: 96437              NA's: 96437
##
##   IN_AGUA_INEXISTENTE  IN_LIXO_COLETA_PERIODICA  IN_LIXO_QUEIMA
##   0   :180078              0   : 41838              0   :146150
##   1   : 5847                1   :144087              1   : 39775
##   NA's: 96437              NA's: 96437              NA's: 96437
##
##   IN_AGUA_POCO_ARTESIANO  IN_AGUA_CACIMBA  IN_AGUA_FONTE_RIO  IN_LIXO_ENTERRA
##   0   :158860              0   :168220              0   :177319              0   :182324
##   1   : 27065               1   : 17705               1   :  8606               1   :  3601
##   NA's: 96437              NA's: 96437              NA's: 96437              NA's: 96437
##
```

```

## 
## 
## 
##   IN_LIXO_OUTROS IN_LIXO_JOGA_OUTRA_AREA IN_LIXO_RECICLA
##   0     :183176      0     :179947          0     :173444
##   1     : 2749      1     : 5978          1     : 12481
##   NA's: 96437     NA's: 96437        NA's: 96437
## 
## 
## 
## 
```

Observando as estatísticas sumárias das variáveis escolhidas, vemos a grande quantidade de valores NA que a tabela tem, nos mostrando que a coleta dos dados está com vários valores faltantes que podem interferir nos cálculos. Com a média das variáveis NU_SALAS_EXISTENTES e NU_SALAS_UTILIZADAS percebemos que as escolas, em média, possuem mais salas de aula do que utilizam. Analisando as variáveis IN_LABORATORIO_INFORMATICA e IN_LABORATORIO_CIENCIAS percebemos que existem poucos laboratórios de informática e de ciências nas escolas brasileiras, cerca de 40% e 11% (respectivamente) de escolas possuem esse tipo de infraestrutura. E com as variáveis IN_LIXO_RECICLA que as escolas brasileiras praticam bem pouco a reciclagem com o lixo de suas escolas, apenas 7%.

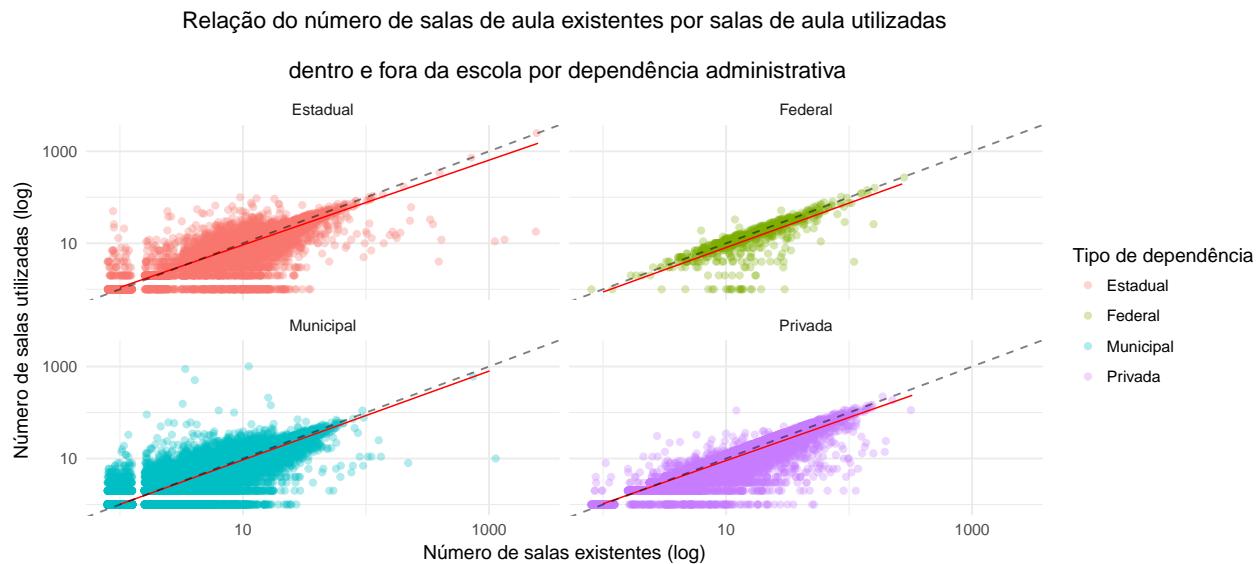
```

## # A tibble: 4 x 5
##   TP_DEPENDENCIA media_numero_de_s~ media_numero_de_sa~ media_numero_de_f~
##   <fct>           <dbl>            <dbl>            <dbl>
## 1 Estadual         11.4             10.6            50.4
## 2 Federal          21.5             18.6            127.
## 3 Municipal        6.22             6.03            25.7
## 4 Privada          11.4             10.3            31.2
## # ... with 1 more variable: media_numero_de_computadores <dbl>
```

As escolas federais possuem as maiores médias, nas escolas municipais a situação se inverte e nas escolas privadas e estaduais a média de salas utilizadas e existentes são as mais próximas. Além disso, apenas nas escolas federais a média da quantidade de computadores é maior que a média de funcionários.

Gráficos

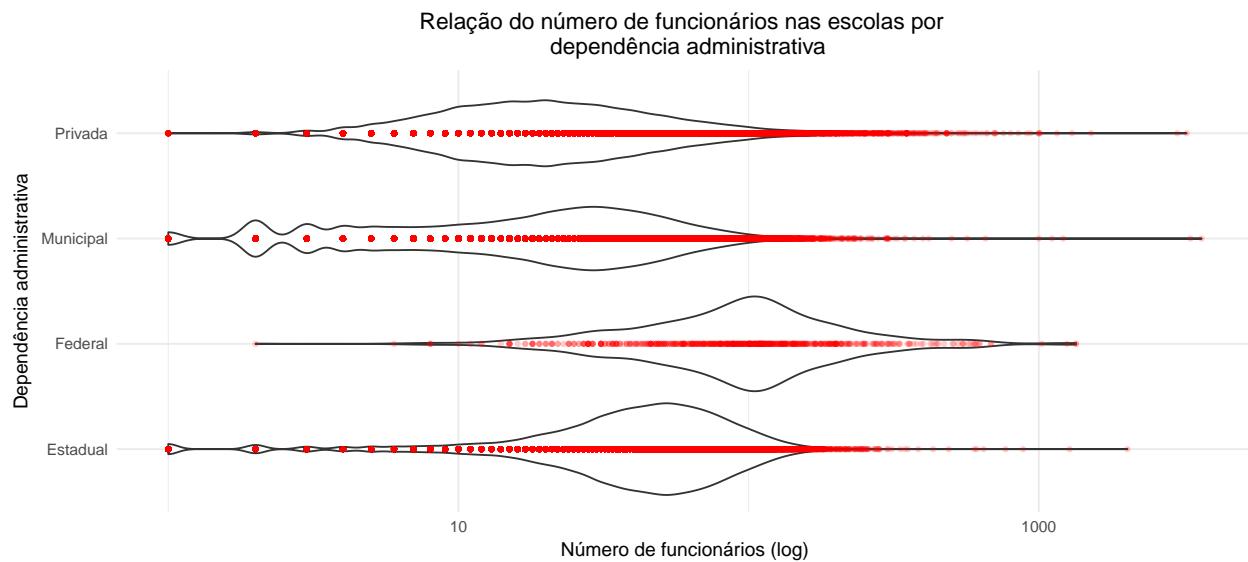
Relação do número de salas de aula existentes e utilizadas nas escolas por dependência administrativa



Aqui plotamos um gráfico log x log da relação do número de salas de aula existentes por salas de aula utilizadas dentro e fora da escola por dependência administrativa. Em vermelho, temos o ajuste linear dos pontos, e a linha preta tracejada representa a reta $y = x$, que implica em *salas existentes* = *salas utilizadas*. Um ponto estar abaixo da reta $y = x$ significa que a escola possui mais salas do que utiliza, e estar acima da reta implica que a escola necessita de mais salas de aula do que as que há nas dependências do colégio.

Podemos percebermos que, enquanto as escolas municipais possuem o ajuste linear mais próximo da reta $y = x$, são as escolas federais que têm pontos de maior acurácia com relação à reta $y = x$. Além disso, é possível afirmar que nas escolas federais o uso das salas de aula é mais organizado, sendo a maioria das salas existentes utilizadas dentro das próprias escolas. Nas escolas estaduais e municipais esse cenário já é o contrário, possuindo grande parcela das salas utilizadas fora da escola. Nas escolas privadas esse cenário já é um pouco melhor, mas possui um número grande de salas inutilizadas.

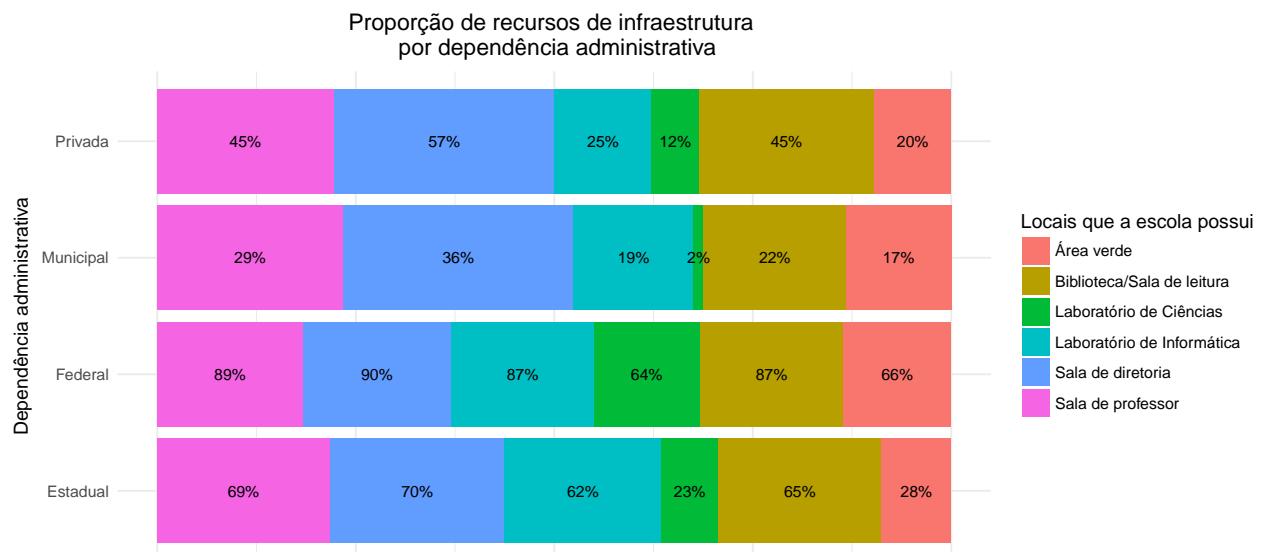
Relação do número de funcionários nas escolas por dependência administrativa



Neste gráfico analisamos a relação do número (em log) de funcionários nas escolas por dependência administrativa. Os pontos possuem opacidade, então quanto mais forte é a cor, mais pontos estão localizados nessa área. Optamos por acrescentar o gráfico de violino, que é um gráfico de densidade espelhada exibido da mesma forma que um boxplot.

A maior concentração de funcionários são nas escolas federais e estaduais. Também é perceptível que há muitos outliers no gráfico e todas as densidades possuem *tails* muito longos.

Recursos relacionados à infraestrutura

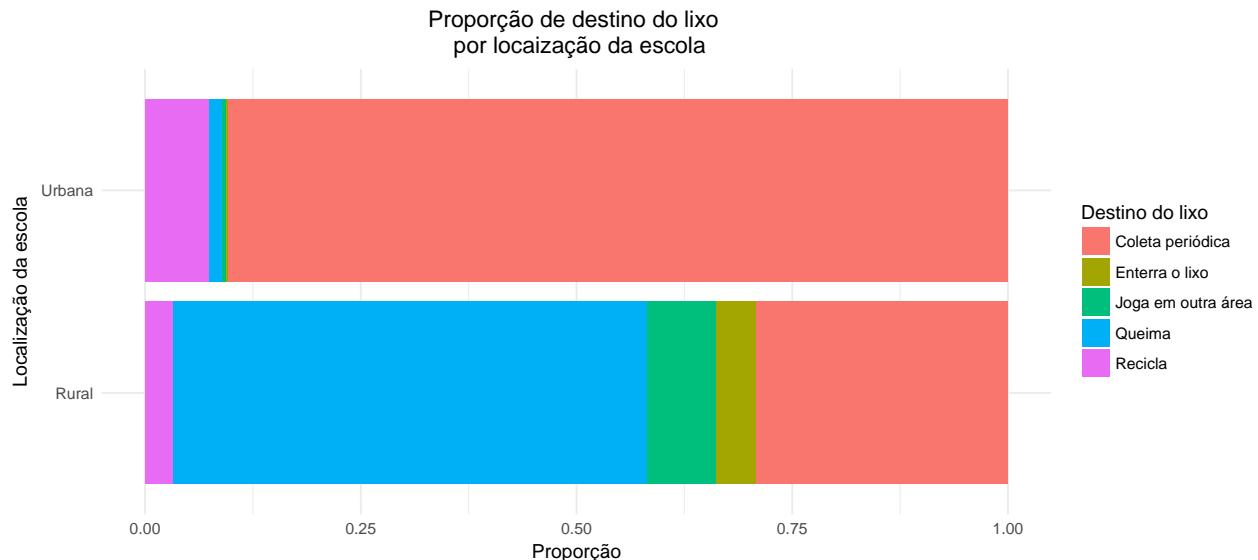


Neste gráfico é necessário estudar cada dependência administrativa separadamente. As porcentagens representam a quantidade de escolas que possuem aquele tipo de infraestrutura.

Analizando o gráfico, percebemos que as escolas dão preferência à construção de salas para a diretoria, uma

vez que é a infraestrutura cuja maioria das escolas possuem, em todos os tipos de dependência. Logo em seguida temos as salas de professores e bibliotecas ou salas de leitura. Vemos também que as escolas federais são as que mais possuem os diferentes tipos de locais, e as municipais as que menos possuem. Os investimentos em área verde e, principalmente, laboratório de Ciências são escassos em todas as dependências, menos a federal. Vale ressaltar que as escolas estaduais e federais apresentam um número alto de laboratório de informática, o que mostra o investimento na tecnologia.

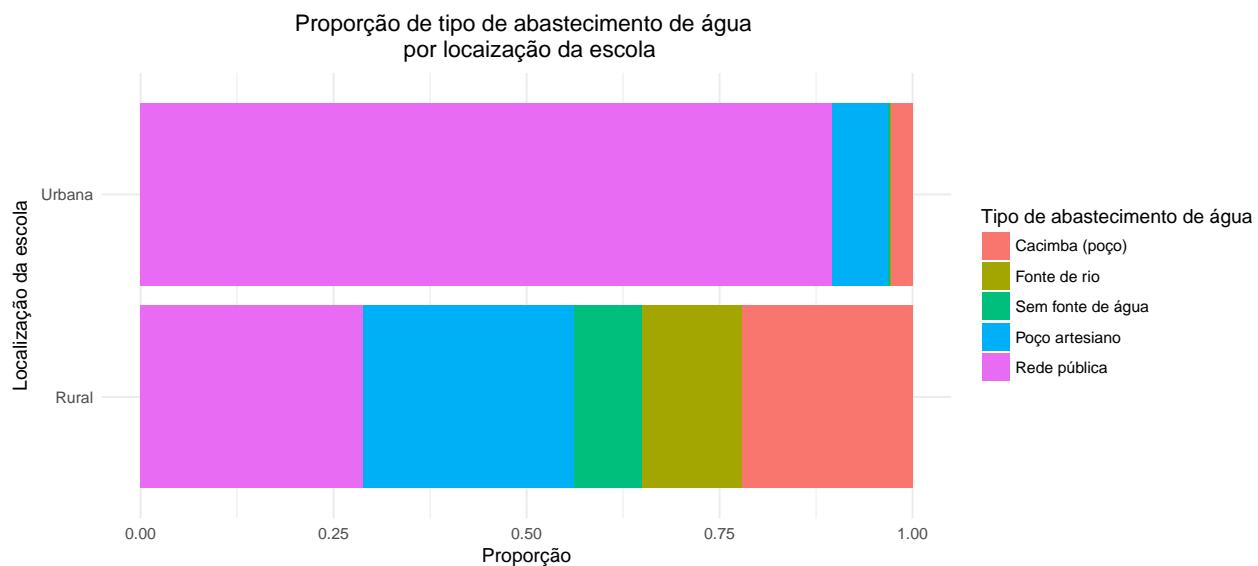
Relação do destino do lixo das escolas (reciclagem) por localização (rural/urbana)



Neste gráfico é necessário estudar as proporções de cada localização separadamente, uma vez que elas não estão na mesma escala. As proporções representam a quantidade de escolas que possuem aquele tipo destino de lixo.

Nas escolas urbanas, a coleta periódica é a escolha predominante, enquanto que a queima, enterrar ou jogar em outra área são as opções com as menores taxas, quase não significativas. Nas escolas rurais, vemos que o principal destino do lixo é a queima, o principal problema oriundo desse processo é a poluição do ar e, portanto, degradação do meio ambiente. A menor taxa é da reciclagem, método que também possui baixo índice nas escolas urbanas. Além disso, as escolas rurais encaminham seu lixo para enterramento ou outras áreas numa parcela pequena, mas ainda significativa.

Relação com o abastecimento de água por localização (rural/urbana)



Neste gráfico é necessário estudar as proporções de cada localização separadamente, uma vez que elas não estão na mesma escala. As proporções representam a quantidade de escolas que possuem aquele tipo de abastecimento de água.

Em ambas as localizações a fonte principal de fornecimento de água é a rede pública, seguida dos poços artesianos. Um fator muito relevante é que um número grande de escolas rurais (cerca de 5%) não possuem nenhum tipo de abastecimento de água, o que sugere uma situação extremamente precária.

Bibliografia

Tabela auxiliar do censo escolar 2017 que utilizamos para ver o significado das variáveis: <https://drive.google.com/file/d/1I0Z8V3lmwycRAejmHQtnDETts4qb-G4s/view?usp=sharing>

Considerações finais

Pelos fatores analisados, organização e infraestrutura, o funcionamento das escolas federais é melhor do que as outras dependências administrativas, e as escolas municipais têm as piores condições. Percebemos que existem poucos laboratórios de informática e de ciências nas escolas brasileiras, cerca de 40% e 11% (respectivamente) de escolas possuem esse tipo de infraestrutura. Nas escolas privadas e estaduais a média de salas utilizadas e existentes são as mais próximas. Além disso, vimos que um número pequeno, mas significativo, de escolas rurais não possuem abastecimento de água, além de dependerem de um tipo de disposição de lixo que mais agride o meio ambiente. E, por fim, as escolas brasileiras praticam pouco a reciclagem com o lixo de suas escolas, apenas 7%.