RELATÓRIO TÉCNICO

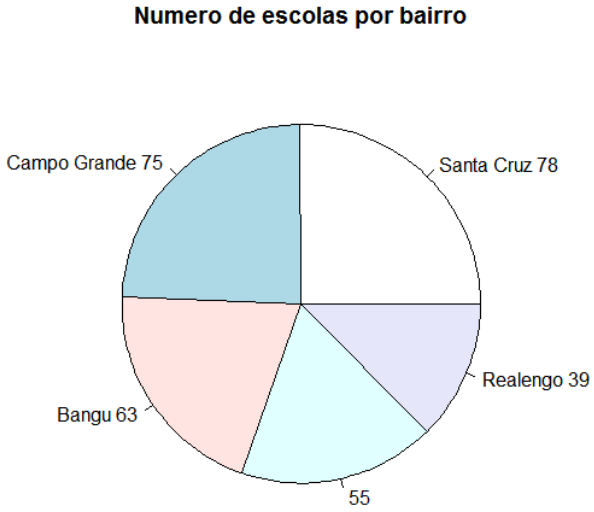
HPCC

- Não foi possível utilizar o HPCC
- Segui rigorosamente o PDF de instalação e configuração, instalei o virtual box, baixei e instalei o Virtual Image File. Porém, ao abrir a máquina virtual, a tela ficou somente na cor preta e absolutamente nada apareceu.
- Após várias horas tentando resolver, optei por não perder mais tempo e iniciei a análise de dados do arquivo escola.xls.

Análise de dados

O arquivo que me foi repassado se refere aos dados de escolas do município do Rio de Janeiro referentes ao ano de 2014. Em tal arquivo, consta 20 variáveis e 1484 observações. Após uma análise geral, comecei a tentar fazer o que me foi solicitado no corpo do e-mail (itens abaixo). Realizei algumas etapas de pré-processamento nos dados afim de torná-los mais aptos para a análise. Optei por utilizar a linguagem R para a análise estatística e geração de gráficos. Embora eu poderia ter utilizado a linguagem Python, escolhi a linguagem R por estar mais recentemente habituado com os comandos e funções em R devido ao curso Formação Cientista de Dados que estou fazendo atualmente.

1. Quais os bairros com o maior número de escolas?
    a. Utilizando o pacote (biblioteca de funções) "sqldf" da linguagem R, foi possível utilizar comandos em SQL para realizar buscas no conjunto de dados. Para isso, converti o conjunto de dados em um tipo de tabela chamado "data.frame" (df).
    b. Segue a query utilizada: query <- sqldf("select Bairro, count(Nome) as qtdEscolas from df group by Bairro order by qtdEscolas DESC"). Essa query retorna os bairros com suas respectivas quantidade de escolas em ordem decrescente, ou seja, o bairro que possui o maior número de escolas aparece primeiro.
    c. Então, plotei 1 gráfico em formato de pizza com os 5 bairros que possuem o maior número de escolas. Segue a função utilizada: pie(query$qtdEscolas[1:5], labels = c('Bairro', 'qtdEscolas'), main = "Numero de escolas por bairro"). Segue abaixo o gráfico plotado. No gráfico, é possível notar que 55 escolas não possuem o bairro cadastrado.



2. Quantas escolas funcionam em tempo integral no bairro Catumbi e que percentagem isso representa?
    a. Utilizando a query abaixo, foi possível obter fazer essa análise e obter esse resultado: query2 <- sqldf("select Bairro, count(Nome) as qtdEscolas_TempoIntegral from df where Bairro = 'Catumbi' and TurnosAtendidos = 'Integral'"). Como o resultado é apenas 1 número, não houve necessidade de plotar um gráfico. A conclusão é que o bairro Catumbi possui 4 escolas e somente 2 delas funcionam em tempo integral. Então, isso representa 50%.
3. Há algum coordenador pedagógico que atua em mais de uma escola?
    a. query3 <- sqldf("select CoordenadorPedagogico, count(Nome) as qtdEscolas from df group by CoordenadorPedagogico order by qtdEscolas DESC"). Essa query faz a busca e retorna o resultado. O resultado mostra que nenhum coordenador pedagógico atua em mais de uma escola, e 537 escolas não possuem o coordenador cadastrado.

Feito isso, o enunciado diz para se sentir livre e realizar outros tipos de análise com o intuito de extrair insights adicionais. Portanto, realizei mais 3 tipos de consultas.

1. Quais escolas oferecem Creche para as crianças?
    a. query4 <- sqldf("select Nome from df where SeriesAtendidas like '%Creche%' order by Nome"). Essa query me retornou o nome das escolas que oferecem Creche como ensino. Verifiquei que 460 escolas do conjunto de dados oferecem Creche.
2. Quais escolas oferecem Pré-Escola?
    a. query5 = sqldf("select Nome from df where SeriesAtendidas like '%Pré-Escola%' order by Nome"). Verifiquei que 714 escolas oferecem Pré-Escola como ensino.
3. Quais escolas oferecem nono ano e possuem o IDEB 1 maior que cinco?
    a. query6 = sqldf("select Nome from df where SeriesAtendidas like '%9º Ano%' and IDEB1 > 5 order by Nome"). Verifiquei que 107 escolas satisfazem essas duas condições.

Segue abaixo a função que plotou o gráfico de barras e o gráfico que mostra a relação dessas três análises, respectivamente:

barras = c(nrow(query4), nrow(query5), nrow(query6))

nomes = c('Creche', 'Pre-Escola', '9 Ano com IDEB > 5')

barplot(barras, legend.text = nomes, col = c('red', 'blue', 'green'), xlab = 'Ensino oferecido', ylab = 'Quantidade de escolas')