Universidade Federal do Rio de Janeiro
Instituto de Computação
Professor: Geraldo Xexéo

Relatório - Jônatas Luís Ramos Simões
GIT - https://github.com/simoesjonatas/dw_individual

Análise de dados sobre o Enade

Questão 1

Para iniciar o trabalho, foi decidido usar o Knime para fazer o Download dos dados através da função "*Unzip files (legacy)*" para baixar o arquivo e descompactar automaticamente salvando em uma pasta local do Knime (Enad). No trabalho é usado essa função três vezes, uma para cada ano de dados do Enade (2017, 2018, 2019). Sendo que no ano de 2019, os dados excepcionalmente estão no github, porque ele não vem no padrão dos outros anos. Os outros anos estão com o link direto do site de download dos dados.

Lembrando que para usar o Unzip Files, no Input file colocamos o link da onde queremos baixar e no Output Diretory colocamos o caminho onde vamos salvar, nesse caso coloquei o output no local que eu criei dentro do meu workflow do Knime.

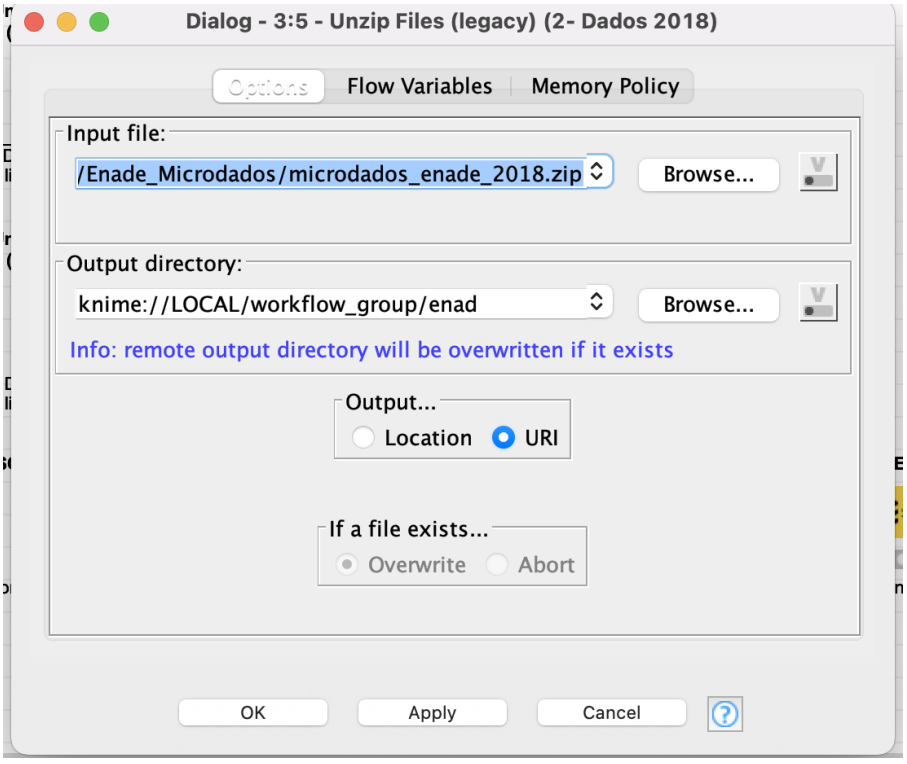


Figura 1 - Configuração do "*Unzip files (legacy)*"

Os link usados em cada Unzip Files foi :

- 2017 -- dados 2017
- 2018 -- dados 2018
- 2019 -- dados 2019

Questão 2

Utilizei a ferramenta *Db Diagram* (https://dbdiagram.io/d), para a construção das tabelas do modelo estrela, na página de anexos em Script do db diagram possui o link direto para o diagrama do modelo relacional. O modelo relacional é constituído de 10 dimensões e uma tabela fato e uma tabela auxiliar para facilitar a inclusão de dados.

Questão 3

Foi criado um servidor MYSQL externo no google cloud, *mysql 5.7*, fui obrigado a usar a versão 5.7, porque um nó usado no *Knime* para carregar os dados não suportava o *MySQL* com versão maior, devido o nó está *Deprecated* (DBLoader do knime). O script da estrutura do banco relacional, que se encontra no anexo ,foi gerado pela ferramenta dbdiagram e depois o script foi executado na ferramenta DataGrip. Lembrando que no Datagrip primeiro foi feita a conexão do banco e depois foi executado o script.

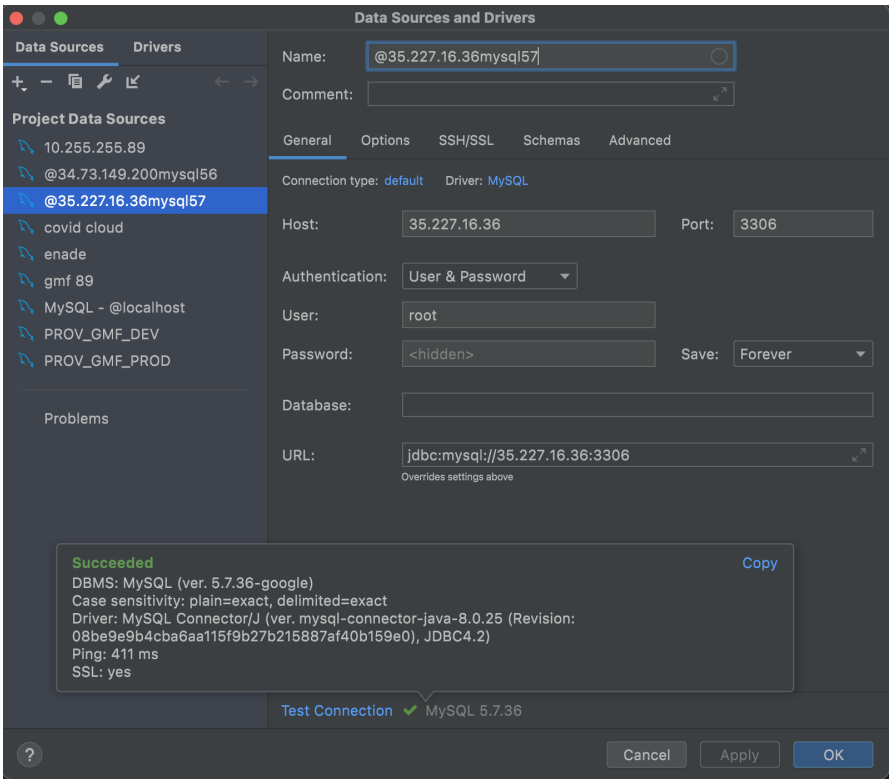


Figura 2 - Configuração do connect do DataGrip

Questão 4

Depois de usar o Unzip Files, foi usado no Knime o CSV Reader para pegar os dados baixados, nesse caso foram utilizados três csv reader referente a cada ano.

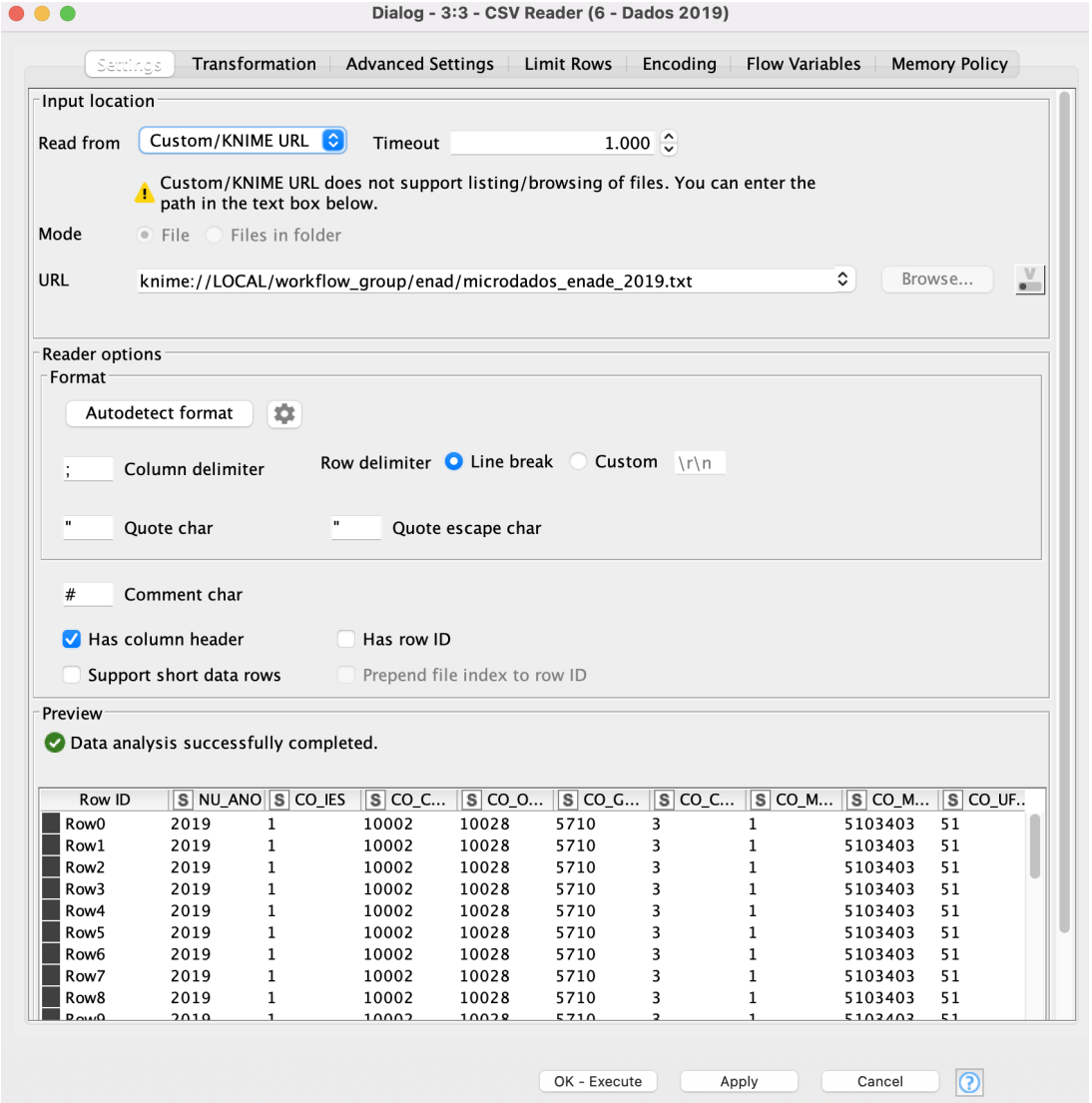


Figura 3 - Configuração do CSV Reader

Através de um fluxo no *KNIME* é possível que a carga de dados fosse feita de uma vez só, desde o *download* dos arquivos até o tratamento dos dados (figura 4). Nesse fluxo faço o download dos arquivos, depois que o download estiver concluído, faço a leitura desses dados no nó do CSV Reader e depois uso um no do Knime para fazer a junção dos três CSVs baixados (Concatenate). Depois disso os dados estão prontos para subir para o banco.

Na parte do banco de dados, através do Knime faço um conector com o banco de dados e depois eu executo os scripts de montagem do modelo relacional (que se encontra no anexo 'Script usado no Knime'), adiciono também uma tabela auxiliar, funções e triggers para me auxiliar na manipulação dos dados para se encaixar na minha base relacional criada.

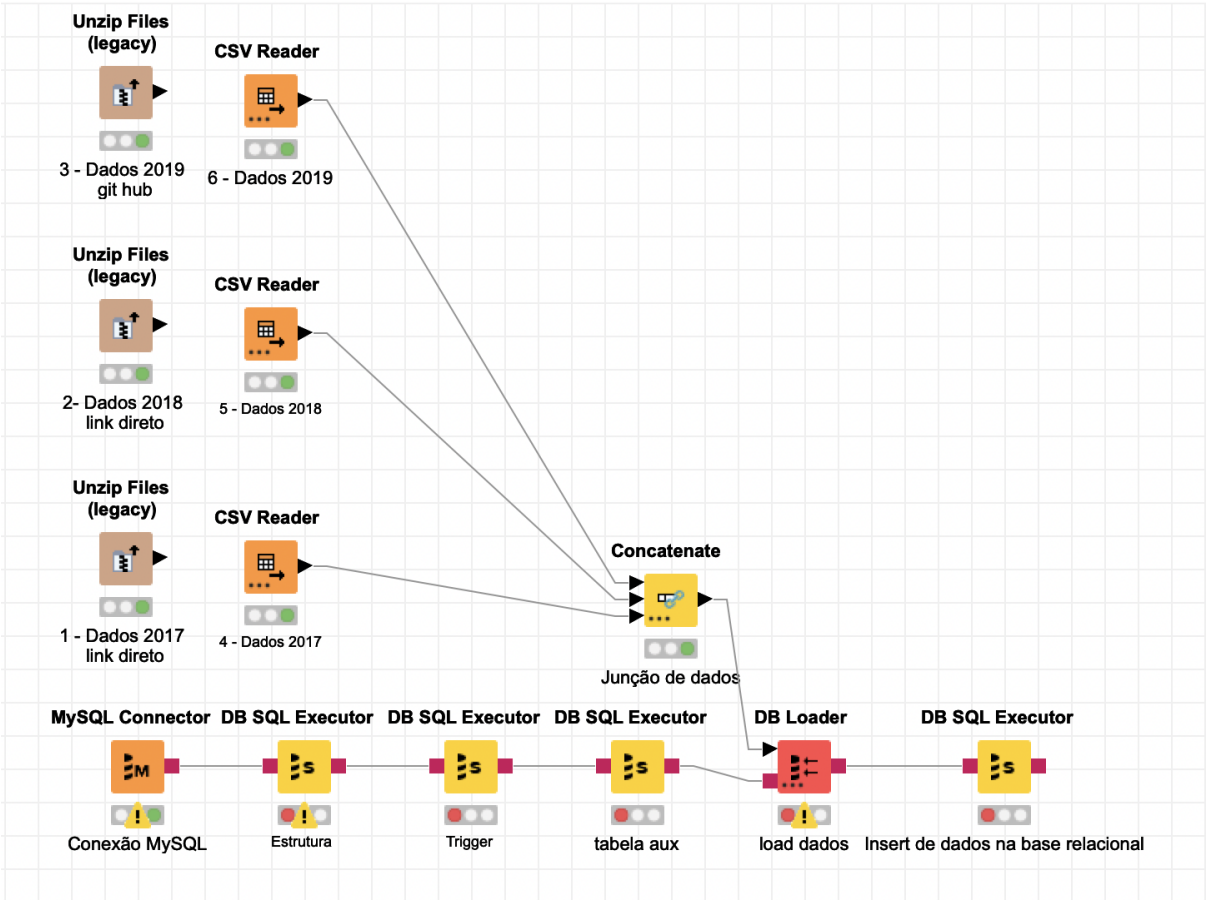


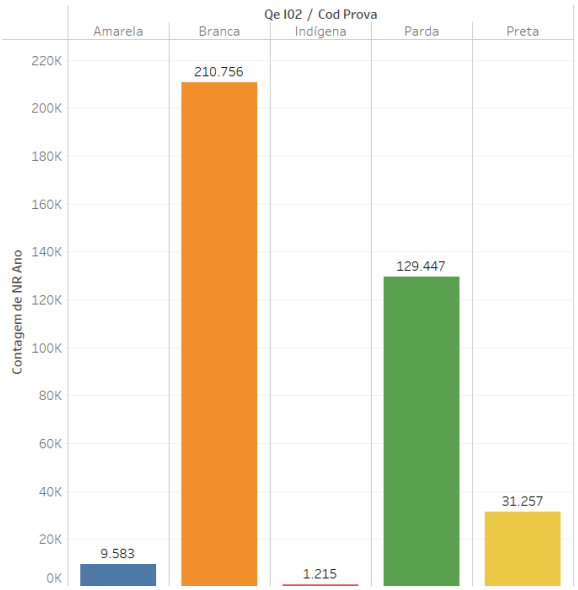
Figura 4 - Fluxo do Knime

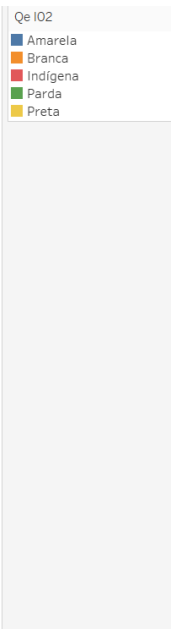
Lembrando que foi criada a tabela auxiliar da mesma forma que os dados vieram para facilitar/agilizar a inclusão de dados do arquivo csv para o banco de dados. E só depois que os dados do arquivo csv estiverem dentro do banco de dados, usamos então recursos do banco de dados para inclusão dos dados no nosso modelo relacional. Com auxílio de uma função e uma trigger para incluir todos os dados nas dimensões corretas.

Questão 5

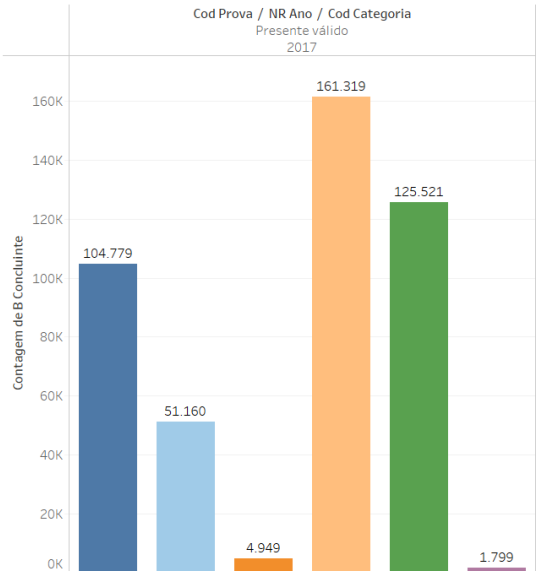
Foi utilizado a ferramenta *Tableau* para a realização dos gráficos e tabelas. O banco de dados criado nas questões anteriores foi conectado ao Tableau para que fosse possível utilizar estes dados e assim efetuar as análises.

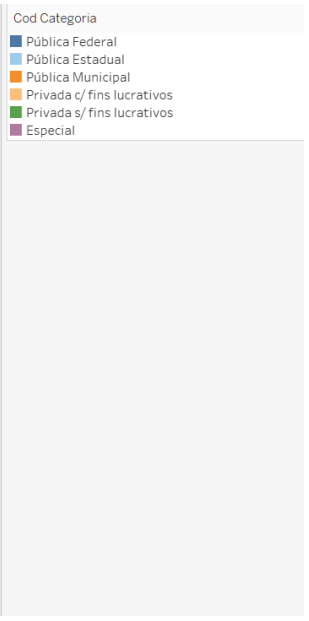
Cor ou raça com mais participação em 2019



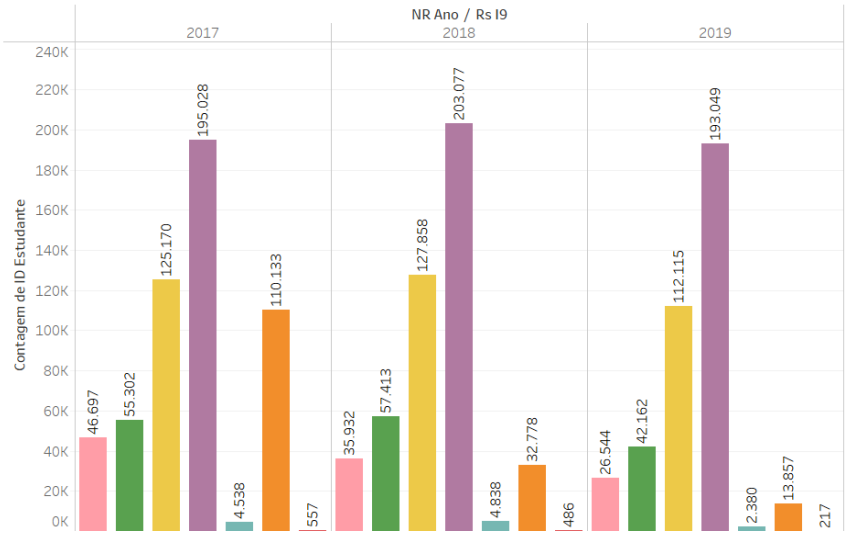


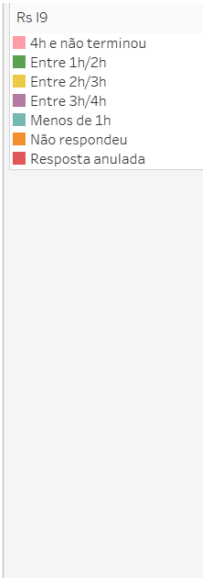
Instituição com mais concluintes



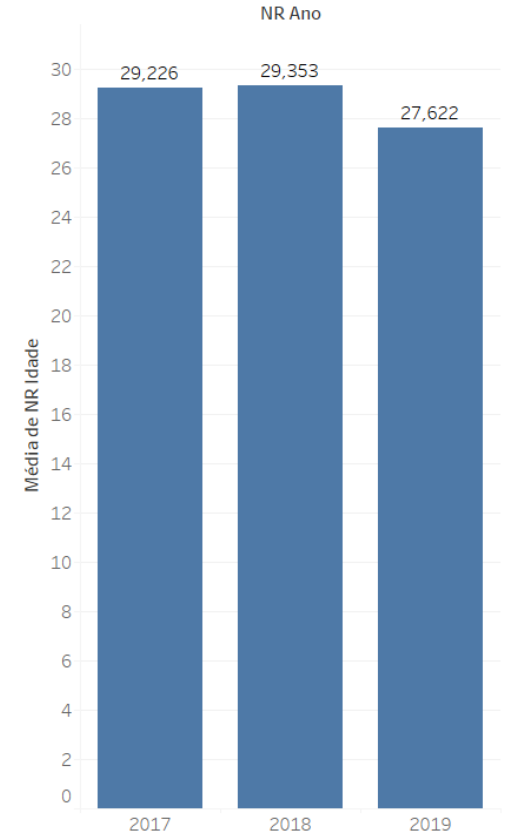


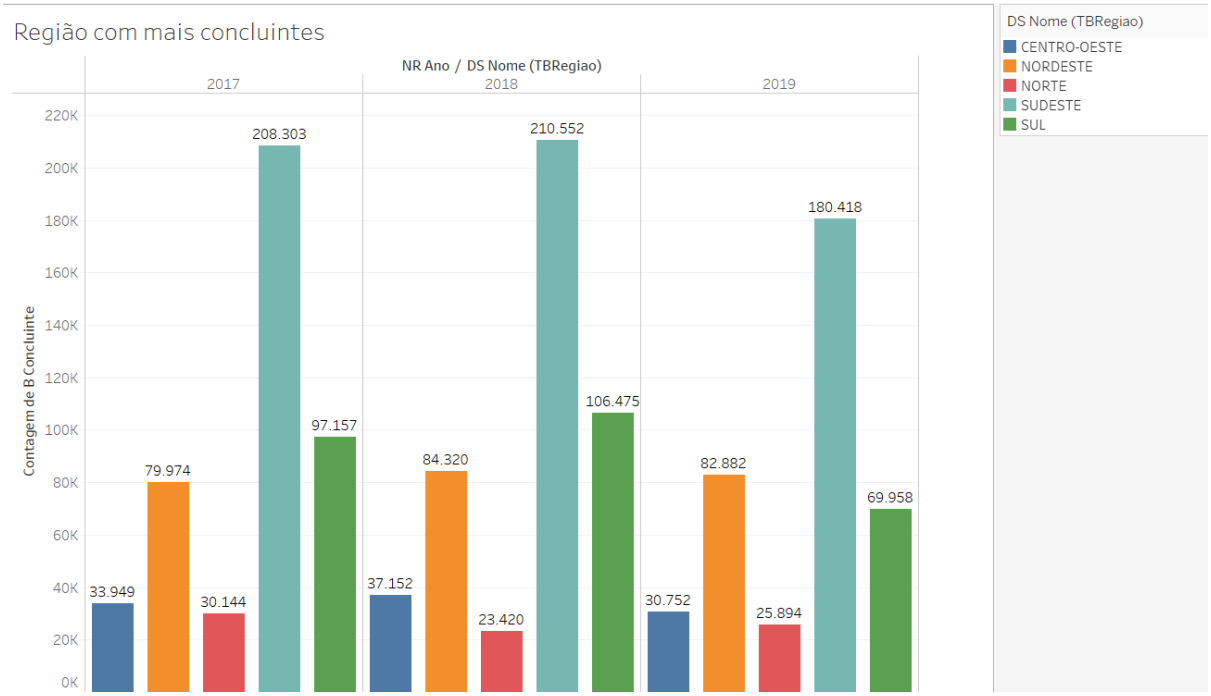
Média de tempo gasto para concluir a prova





Média das idades por ano





Questão 6

Questão 7

- Db Diagram: https://dbdiagram.io/d - Essa ferramenta foi escolhida por conta da sua facilidade em montar o modelo dimensional e exportar o script do modelo;
- DataGrip: https://www.jetbrains.com/datagrip/ - Essa ferramenta foi escolhida por ser mais intuitiva que o MySQL workbench;
- *Knime*: https://www.knime.com/ - Essa ferramenta foi utilizada para fazer o download automático dos dados e para o aprendizado de máquina;
- Tableau: https://www.tableau.com/pt-br - Essa ferramenta é muito utilizada para a geração de gráficos e tabelas;
- GitHub: https://github.com/simoesjonatas/dw_individual - O GitHub foi utilizado para colocar o repositório dos dados analisados;
- Google Cloud: https://cloud.google.com - Foi utilizado para alocar o banco de dados MySQL 5.7;
- Word - Utilizado para fazer o relatório.

Anexo

script do *db diagram*

Link para acessar na ferramenta

Script da base Relacional

Link direto para o arquivo no GitHub

Script usados no knime

Link direto para a pasta no GitHub