

BRUNO DE SOUZA MAIA DE CARVALHO

EDUARDO DA SILVA COUTINHO

RICARDO DE SOUZA DOS REIS SILVA

**ANÁLISE SOCIOECONÔMICA DOS CANDIDATOS DO ENEM ATRAVÉS DE MICRODADOS DISPONIBILIZADOS PELO PORTAL INEP**

RIO DE JANEIRO

2020



BRUNO DE SOUZA MAIA DE CARVALHO

EDUARDO DA SILVA COUTINHO

RICARDO DE SOUZA DOS REIS SILVA

**ANÁLISE SOCIOECONÔMICA DOS CANDIDATOS DO ENEM ATRAVÉS DE MICRODADOS DISPONIBILIZADOS PELO PORTAL INEP**

Trabalho de conclusão de curso apresentado à Faculdade Senac como requisito para a obtenção de aprovação no curso de Pós-graduação lato sensu em

Big Data.

Docente orientador: Clayton Escouper das Chagas

RIO DE JANEIRO

2020

FOLHA QUE A BIBLIOTECA VAI FAZER

Ficha catalográfica. A ficha catalográfica é o elemento que contém as informações bibliográficas necessárias para a identificação e localização de uma obra no acervo de uma biblioteca. Além disso, também é obrigatória e recomendada pela ABNT nos trabalhos acadêmicos convencionais. Deve seguir as normas do Código de Catalogação Anglo Americano (AACR2). A elaboração da ficha catalográfica será realizada pela Bibliotecária e, solicitada à biblioteca da Faculdade de Tecnologia Senac Rio, pelo e-mail: **marcele.lima@rj.senac.br**. A biblioteca disponibiliza o serviço desde que solicitado com antecedência de dois dias.  Os dados registrados na ficha catalográfica fornecerão uma visão sumária da temática do trabalho acadêmico e de seus aspectos de ordem física. A ficha catalográfica será incluída no verso da folha de rosto do trabalho.



**Faculdade Senac**

**Curso:** Pós-Graduação lato sensu em Big Data

**Ano:** 2020

**Nome do aluno:**

Bruno de Souza Maia de Carvalho

Eduardo da Silva Coutinho

Ricardo de Souza dos Reis Silva

**Título:** Análise Socioeconômica dos Candidatos do Enem Através de Microdados Disponibilizados Pelo Portal INEP

**Nome do docente orientador:** Clayton Escouper das Chagas

**Conceito:** \_\_\_\_\_\_\_\_

Recomendações: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Rio de Janeiro, \_\_\_\_\_\_\_ de \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ de 2020.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Orientador

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Co-orientador

**RESUMO**

Nos dias de hoje é possível notar o aumento do número de pessoas que buscam aprender mais sobre finanças, pessoais e investimentos. Em dias que a informação está ao alcance da maioria, não é difícil se deparar com anúncios de corretoras de investimentos, guias para gerir finanças pessoais, livros de “como investir na bolsa” ou, até mesmo, acerca da compra de imóveis. Com isso, mais pessoas vem se dando conta de que é importante cuidar da vida financeira, e de que é possível até se tornar um futuro investidor.

Tudo isso leva uma pessoa a pensar que talvez esteja gastando seu dinheiro da forma errada e que deveria ter mais controle sobre seus gastos. Muitas vezes uma pessoa adquire despesas das quais nem lembra o dia, quem dirá o motivo. Daí surge a necessidade do controle sobre as finanças pessoais, e até mesmo de alguma ferramenta que auxilie esse gerenciamento.

Neste contexto, apresentamos a solução de uma ferramenta de prático acesso e fácil usabilidade para apoiar a necessidade de quem deseja cuidar do próprio dinheiro e investimento. Nela o usuário conseguirá gerir suas receitas, gastos, investimentos e consultar gráficos de consolidação e  as principais fontes de dados e notícias para suportá-lo nesta empreitada.

Para ajudar na criação deste produto, utilizamos a metodologia Design Thinking, que nos ajudou a entender as necessidades deste público que está iniciando a administração das finanças pessoais e para sua criação utilizamos das mais atuais práticas de gerenciamento de projetos de software ágil e suas ferramentas, como o Scrum e o Kanban Board para assegurar o desenvolvimento de um backlog bem elaborado, bem como da metodologia Design Thinking para entender como essa nova classe de investidores necessita desta demanda.

Palavras-chave: Usabilidade. Necessidades do usuário. Design Thinking. Agile. Scrum. Kanban Board. Framework. Finanças pessoais. Investimento. Aplicativo.

**ABSTRACT**

Nowadays it is possible to notice the increase in the number of people looking to learn more about personal finance and investments. On days when information is available to the majority, it is not difficult to come across advertisements from investment brokers, guides for managing personal finances, books on "how to invest in the stock market" or even about buying real estate. With this, more people are realizing that it is important to take care of the financial life, and that it is even possible to become a future investor.

All of this leads a person to think that they may be spending their money the wrong way, and that they should have more control over their spending. Often a person acquires expenses that he does not even remember the day, who will say the reason. Hence, there is a need for control over personal finances, and even for some tool that helps this management.

In this context, we present the solution of a tool of practical access and easy usability to support the need of those who want to take care of their own money and investment. In it, the user will be able to manage his revenues, expenses, investments and consult consolidation charts and the main sources of data and news to support him in this endeavor.

To help in the creation of this product, we used the Design Thinking methodology, which helped us to understand the needs of this public that is starting the administration of personal finances and for its creation we use the most current practices of agile software project management and its tools, such as Scrum and Kanban Board to ensure the development of a well-designed backlog, as well as the Design Thinking methodology to understand how this new class of investors needs this demand.

Keywords: Usability. User needs. Design Thinking. Agile. Scrum. Kanban Board. Framework. Personal finances. Investment. App.

**LISTA DE FIGURAS**

|  |  |
| --- | --- |
| 1 - Arquitetura de dados …………………………………………………… 15  2 - Processo do KDD ………………………………………………………. 16  3 - Aplicação de bibliotecas - Linguagem Python ………………………. 22  4 - Criação dos diretórios - Salvar datasets ……………………………... 22  5 - Criação da variável para recebimento dos dados da ‘’url’’................ 23  6 - Download dos arquivos para importação dos microdados ………... 23  7 - Apresentação do tamanho dos arquivos importados do INEP ….… 24  8 - Instanciando o serviço do Spark ……………………………………… 24  9 - Importação dos microdados para o Data Frame …………………… 24  10 - Quantidade de registros por data frame……………………………. 25  11 - Impressão de duas linhas com referência ao dataframe 2015…... 26  12 - Impressão da estrutura do dataframe ao ano de 2015 ……………26  13 - Apresentação da descrição do dataframe para o ano de 2015…. 26  14 - Visualização das colunas para realização da análise exploratória. 26  15 - Criação dos data frames ano de 2015 a 2019…………………....... 27  16 - Quantidade de inscritos por idade ………………............................  17 - Quantidade de inscritos por ano…………………………….............  18 - Quantidade de candidatos por ano e sexo…………………...........  19 - Quantidade de inscritos por raça…………………………………....  20 - Quantidade de inscritos por raça e ano………………………….....  21 - Quantidade de alunos baseados pelo status de ensino................ 22 - Quantidade de alunos por tipo de escola......................................  23 - Quantidade de candidatos com deficiência visual...................................  24 - Quantidade de candidatos com deficiência física.....................................  25 - Quantidade de gestantes de 2015 a 2019...............................................  26 - Quantidade de candidatos que necessitam de intervenção em Libras....  27 - Condição de ensino ao pai do candidato.................................................  28 - Condição de ensino a mãe do candidato................................................. 29 - Quantidade de celulares por residência..................................................  30 - Quantidade de inscrições por residência com computador....................  31 - Quantidade de inscrições por residência com internet...........................  32 - Quantidade de aprovados no exame por ano.........................................  33 - Quantidade de candidatos por renda mensal família por ano................ |  |

**SUMÁRIO**

**1 - INTRODUÇÃO………………………………………………………………….…. 9**

  1.1 - Perguntas e respostas sobre a proposta……….…...........…….…... 10

**2 - LINK DO PROJETO…………………………………………….......…......…....11**

**3 - MODELO DE NEGÓCIO…………………………………...............…………...12**

3.1 - Estudo do Dataset …………………....………........................……...12

               3.1.1 - Primeira etapa da análise ……….............................................13

               3.1.2 - Segunda etapa da análise …………………....………...……...13

               3.1.3 - Terceira etapa da análise ………………….…….........……….14

**4 - ARQUITETURA DE DADOS……………………………….............………….15**

**5 - MINERAÇÃO DE DADOS …………………………………...……………...….16**

**6 - PRÉ-PROCESSAMENTO………………………………....……...…….……...17**

6.1 - Apache Spark……………………………………....…......…………….18

          6.2 - Jupyter Notebook…………………………….……………..................18

          6.3 - Banco de Dados MYSQL……………………….…….……....…...….19

          6.4 - Visual Studio Code………………………………………….………....19

          6.5 - Visualização de Dados Com PowerBI…………………….….……...19

**7 - DATA CLEANING……………………………………………..........……...…...20**

**8 - ANÁLISE EXPLORATÓRIA………………………...………………......……...21**

8.1 - Desenvolvimento da aplicação dos códigos……………...…....…...21

          8.2 - Regressão Linear............................................................................30

**9 - CONFIGURAÇÃO DO AMBIENTE....………………...………...….........…...31**

9.1 - Softwares……………………………………………………………….

9.1.1 Instalação do Python……………………………………….

9.1.2 Configurar Banco de Dados……………………………….

9.1.3 Instalar e configurar IDE……………………….......……….

9.1.4 Carregar o projeto na IDE…………………………………….

**9.2 - Infraestrutura do ambiente……………………………....………...**

9.2.1 - Instalação Python…………………………………...……….

9.2.2 - Instalação Java - JDK…………………………...………….

9.2.3 - Instalação Apache Spark………………………...………...

9.2.4 - Instalação Jupyter Notebook………………………….

9.2.5 - Configuração de variáveis de ambiente - Windows...

**10 - ANÁLISE E DISCUSSÃO DE RESULTADOS……..................................**

**11 - CONCLUSÃO………………………………………...........................……...**

**12 - REFERÊNCIAS……………………………………….…..................……….**

1. **INTRODUÇÃO**

A extração dos dados educacionais tem se tornado um campo em amplo desenvolvimento para estudos e obtenção de padrões pedagógicos em diversas instituições públicas e particulares. Através dessa análise é possível um reconhecimento no processo de aprendizado em bases da educação de nível básico à nível médio.

O Exame Nacional do Ensino Médio (Enem), foi criada em 1998. Ela é utilizada para avaliar a qualidade do ensino médio no país. Seu resultado serve para acesso ao ensino superior em universidades públicas brasileiras, através do Sistema de Seleção Unificada (SiSU), assim como em algumas universidades no exterior. O Enem é o maior exame vestibular do Brasil e o segundo maior do mundo, atrás somente do Gāo Kǎo, o exame de admissão do ensino superior da República Popular da China.

É uma prova realizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), sendo uma autarquia federal vinculada ao Ministério da Educação (MEC). Seu objetivo é promover estudos, pesquisas e avaliações periódicas sobre o sistema educacional brasileiro, com o objetivo de subsidiar a formulação e implementação de políticas públicas para a área educacional.

O INEP realiza levantamentos estatísticos e avaliações em todos os níveis e modalidades de ensino. Suas atividades mais conhecidas são: a realização do Exame Nacional do Ensino Médio (Enem); a realização do Exame Nacional de Desempenho de Estudantes (Enade); a organização das avaliações periódicas dos ensinos básico e superior brasileiro.

Diversas universidades do Brasil, utilizam-se como critério de admissão a nota de qualificação obtida na prova e posteriormente é possível realizar o download dos microdados disponibilizado referente a cada ano de prova feita pelos alunos.

Neste estudo realizaremos uma verificação sobre a possibilidade de previsão das notas dos alunos inscritos, aplicando somente a categoria

socioeconômica atribuída ao questionário no momento da inscrição para realização da prova.

Abordaremos diversos temas relacionados ao assunto de big data, dentre eles : coleta de dados, através dos microdados disponibilizados pelo INEP, limpeza dos dados, onde será aplicada uma correlação entre tabelas e dicionários com informações referentes as bases de dados disponíveis, pré-processamento, onde faremos a eliminação de anomalias que de alguma forma possa comprometer por total o resultado e qualidade de todo processo, mineração dos dados, utilizando Apache Spark (ferramenta Big Data que tem o objetivo de processar grandes conjuntos de dados de forma paralela e distribuída) e Jupyter Notebook para a verificação dos dados já totalmente validados e sólidos, onde será feita a conexão à um banco de dados MySQL.

**1.1 Perguntas e respostas sobre a proposta:**

**Pergunta:** O que será feito? (**what**)

**Resposta:** Uma análise exploratória em datasets do ENEM disponibilizados pelo portal INEP.

**Pergunta:** Por que será feito? (**why**)

**Resposta:** Será feito com o objetivo de diagnosticar problemas socioeconômicos aos candidatos que irão prestar o exame nacional do ensino médio, de acordo com métricas relacionadas ao que o mesmo informa na ficha de inscrição.

**Pergunta:** Onde será feito? (**where**)

**Resposta:** O ambiente de desenvolvimento será feito em ferramenta Open Source - Visual Studio Code, utilizando de forma integrada com banco de dados relacional - Mysql.

**Pergunta:** Quando será feito? (**when**)

**Resposta:** O objetivo é entregar em sprints e concluir na data final do curso.

**Pergunta:** Por quem será feito? (**who**)

**Resposta:** Bruno Carvalho, Eduardo Coutinho, Ricardo Reis.

**Pergunta:** Como será feito? (**how**)

**Resposta:** Será feito em um processo de análise exploratória com base em datasets disponibilizados pelo INEP, a uma amostragem dos anos de 2015 a 2019.

**Pergunta:** Quanto custa fazer? (**how much**)

**Resposta**: N/A. Levará em torno de 16 semanas, totalizando 4 meses para a data final da apresentação dos sprints do projeto.

**2 LINK DO PROJETO**

* Repositório do código fonte   
  Sistema: GitHub

Url: <https://github.com/projetobigdatasenac2020>

**3 MODELO DE NEGÓCIO**

O presente estudo aplica-se a verificação e validação de *datasets* provenientes do Exame Nacional do Ensino Médio – ENEM, através do qual será feita uma análise exploratória dos dados, possibilitando a identificação de precariedade de estudo em determinadas regiões do Brasil e relacioná-las.

O objetivo inicial é aferir se:

a) Os candidatos estão em condições de desenvolver todos os seus conhecimentos científicos e tecnológicos para os padrões exigidos da prova;

b) Identificar a geolocalização de instituições de ensino médio e correlacionar com as notas providas da base de dados do INEP;

c) Apresentar a ordenação de resultados por áreas de aplicação, dentre elas: Matemática, Ciências Humanas, Ciências da Natureza, Linguagens e Códigos e Redação;

d) Apresentar correspondência com a infraestrutura familiar na interferência direta e indireta do resultado dos alunos ao acesso às Universidades.

**3.1 ESTUDO DO DATASET**

O exame passou por alterações estruturais entre os exercícios de 2014 e 2015, algumas dessas refletindo na análise dos microdados.

O projeto baseia-se em uma base de dados, que contém em torno 20 gigabytes, utilizando microdados entre os anos de 2015 a 2019, provida pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), onde a estrutura dos dados são geradas através de arquivos com extensão (.csv). Um arquivo CSV (Valores Separados por Vírgula) é um tipo especial de arquivo que você pode criar ou editar no Excel. Em vez de armazenar informações em colunas, os arquivos CSV armazenam informações separadas por vírgulas.

Em vez de armazenar informações em colunas, os arquivos CSV armazenam informações separadas por vírgulas.  sendo o processo de estudo dividido em etapas de atuação como descrito abaixo:

3.1.1 A primeira etapa da análise do *dataset*, consiste na verificação dos dados que o candidato deve preencher no formulário de inscrição para a prova, com isso, obtêm-se diversos campos de informação, como: município de residência, cidade, nacionalidade, município de nascimento, idade, cor/raça, sexo, estado civil, grau de escolaridade do pai e da mãe, tipo de escola oriunda (instituição pública ou privada) e se o estudante apresenta algum tipo de deficiência.

3.1.2 Segunda etapa da análise do *dataset*, apresentam-se dados referentes à questões socioeconômicas do candidato, realizadas no ato da inscrição do exame. Abaixo apresentamos campos que são preenchidos:

   i. Até que série seu pai, ou o homem responsável por você, estudou?

    ii. Até que série sua mãe, ou a mulher responsável por você, estudou?

   iii. Incluindo você, quantas pessoas moram atualmente em sua residência?

   iv. Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)

   v. Na sua residência tem telefone celular?

  vi. Na sua residência tem computador?

 vii. Na sua residência tem acesso à Internet?

viii. Você já concluiu ou está concluindo o Ensino Médio?

      ix. Em que tipo de escola você frequentou o Ensino Médio?

3.1.3 Terceira etapa da análise do *dataset*, consiste na exposição ao rendimento dos alunos nas provas objetivas e redação. Com isso, é realizado um estudo nos seguintes campos abaixo:

            i.        Presença na prova objetiva de Ciências da Natureza

                     ii.        Presença na prova objetiva de Ciências Humanas

 iii.        Presença na prova objetiva de Linguagens e Códigos

   iv.        Presença na prova objetiva de Matemática

  v.        Nota da prova de Ciências da Natureza

vi.        Nota da prova de Ciências Humanas

                 vii.        Nota da prova de Linguagens e Códigos

               viii.         Nota da prova de Matemática

                  ix.         Língua Estrangeira

x.         Situação da redação do participante

            xi.         Nota da prova de redação

**4 ARQUITETURA DE DADOS**

Com isso, para uma melhor análise e visualização na aplicabilidade aos termos e técnicas de todo processo do nosso estudo, foi desenvolvido uma arquitetura simplificada, vide (Fig.1) objetivando todas as ferramentas e procedimentos que serão aplicados e abordados em todo projeto.

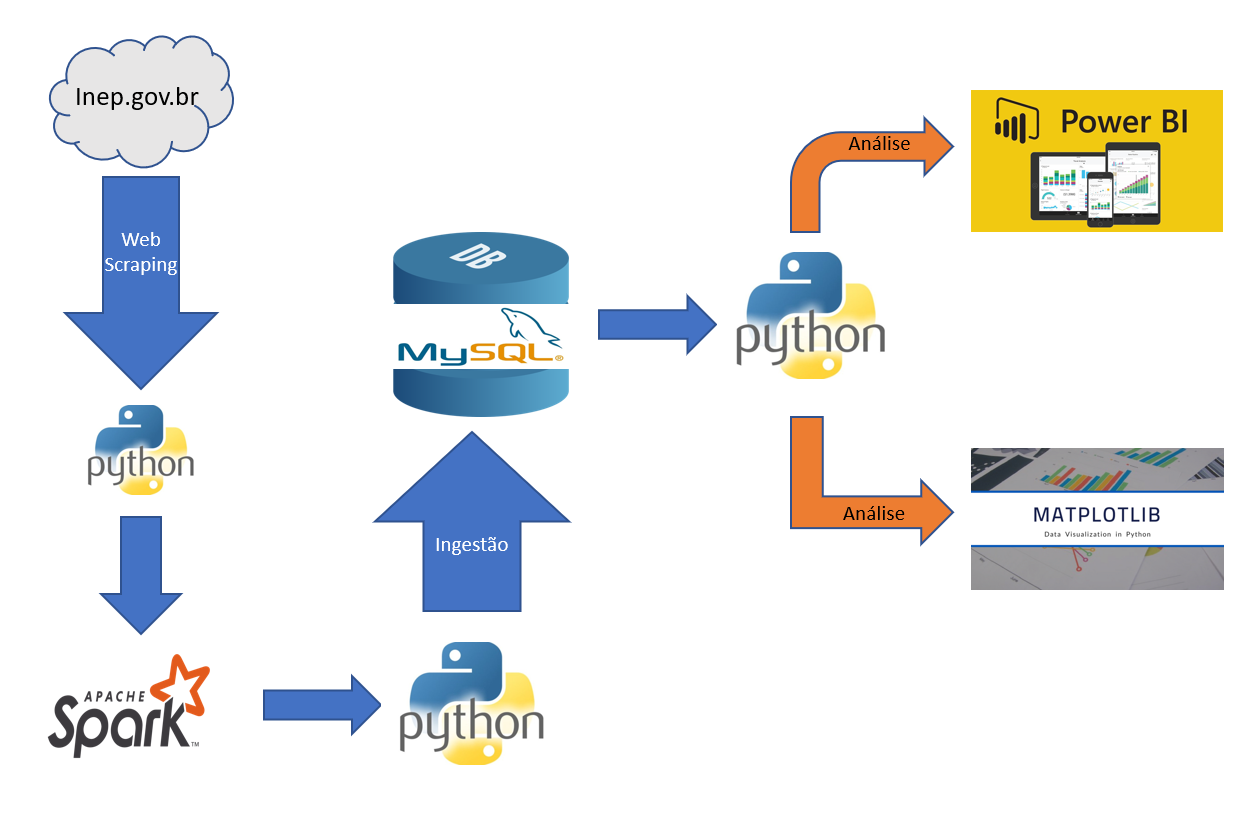
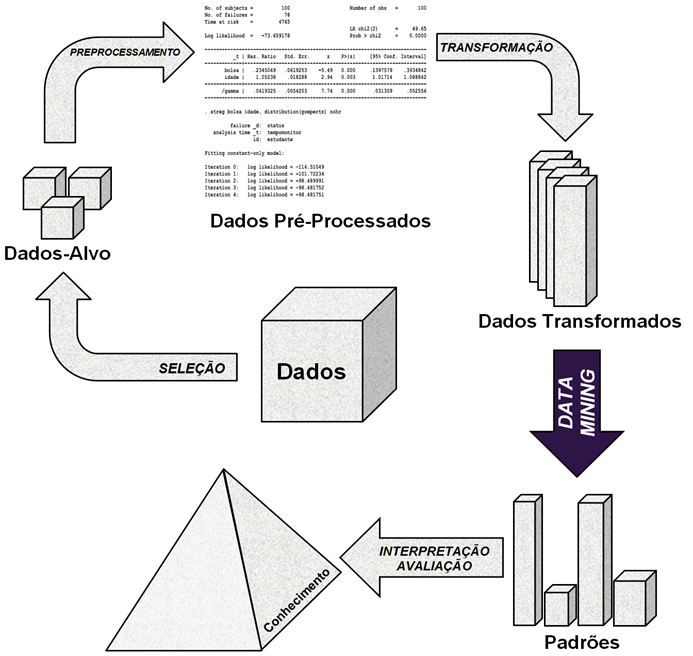
****

Fig.1 (Fonte: Elaboração dos Autores)

**5 MINERAÇÃO DE DADOS**

A mineração de dados é dita como um processo utilizado para extração de dados de um conjunto maior de dados brutos e se tem a aplicação em diversos campos como: ciência e pesquisa. Envolve coleta, armazenamento de dados eficazes, pré-processamento, transformação dos dados (limpeza), busca de padrões e visualização final dos dados, gerando o conhecimento.

Abaixo temos (Fig.2), que apresenta todo o procedimento da fase de início ao fim do método aplicado.

****

*Fig.2*

*Estágios do Processo KDD e Data Mining.*

*FONTE: Fayyad, Piatetsky-Shapiro e Smyth. From data mining to knowledge discovery in databases.*

**6 PRÉ-PROCESSAMENTO**

A fase de pré-processamento talvez seja a mais importante a fim de se obter um bom resultado, o pré-processamento nada mais é do que o processo de preparação, organização e estruturação dos nossos dados, além de ser o momento ideal para escolhermos quais dados fazem sentido estarem sendo analisados em nosso *dataset*.

É importante realçar a importância desta etapa, dito que, a qualidade de dados pode influenciar diretamente no resultado na modelagem, muitas das vezes pensamos que o problema da nossa solução é o algoritmo usado para gerar o modelo, porém o grande vilão é o seu próprio conjunto de dados que podem possuir muitos atributos com valores faltantes, outliers e escalas de valores contradizentes e por fim nenhum modelo será capaz de trabalhar com esses dados e gerar resultados de qualidade.

Para a análise de estudo aos microdados disponibilizados pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) em relação ao ENEM, teve-se a aplicação das seguintes ferramentas:

* Apache Spark
* Jupyter Notebook
* Banco de Dados – MySQL
* Visual Studio Code
* Microsoft Power BI

Para analisar e validar as bases de dados foram aplicadas de forma local (máquina física), onde tivemos alguns problemas de processamento de arquivos do *dataset*, por possuírem uma quantidade excessiva de informações. Para isso, feita uma análise e retiradas colunas que não seriam de suma importância para a abrangência do estudo, com isso restando 49 colunas onde faremos uma abordagem socioeconômica de acordo com itens do item 6.1. Anteriormente havíamos pensado em estar exportando para uma *engine* de banco de dados em Cloud, porém ficaria totalmente inapropriado em ter um custo muito elevado para aplicação desses dados dessa forma. Com isso decidimos aplicar em desktop robusto e com configurações que iriam suportar todo processamento dos microdados.

**6.1 – Apache Spark**

O Apache Spark é uma ferramenta Big Data que tem o objetivo de processar grandes conjuntos de dados de forma paralela e distribuída. Visto que a massa de dados utilizada é de grande quantidade vide item 6.1, com isso a arquitetura utilizada pelo Spark em nosso projeto, conta com verificação do conjunto de informações e dados disponibilizados e em paralelo sendo feito o desenvolvimento da aplicação com o Banco de Dados Mysql e linguagem de programação Python.

**6.2 – Jupyter Notebook**

O Jupyter Notebook é uma ferramenta de computação alfabetizada, extremamente eficiente, pois com ela é possível unir códigos e textos. Assim é possível explicar cada funcionalidade de forma detalhada. É possível gerar gráficos em tempo real dentro da ferramenta. Em nosso estudo foi desenvolvido para desenvolvimento e documentação dos códigos de forma inicial, gerando um grande fluxo e rapidez na visualização do resultado da análise do *dataset*.

**6.3 – Banco de Dados Mysql**

O MySQL é um sistema de gerenciamento de banco de dados (SGBD), que utiliza a linguagem SQL (Linguagem de Consulta Estruturada, do inglês *Structured* Query *Language*) como interface. É atualmente um dos sistemas de gerenciamento de bancos de dados mais populares da Oracle Corporation, com mais de 10 milhões de instalações pelo mundo.

Após análise e limpeza dos dados, foram exportadas as colunas para uma base de dados (Mysql), onde será feita a conexão com uma ferramenta de visualização para exposição dos dados gerados por *querys* no banco de dados.

**6.4 – Visual Studio Code**

O Visual Studio Code é um editor de código-fonte desenvolvido pela Microsoft para Windows, Linux e macOS. Ele é um software livre e de código aberto, apesar do download oficial estar sob uma licença proprietária.

Após utilização inicial do Jupyter Notebook, resolvemos fazer a migração para o Visual Studio *Code* e com isso tivemos uma performance no processamento. Ferramenta onde desenvolvemos o código-fonte do projeto e toda documentação linha a linha.

**6.5 – Visualização de Dados com PowerBI**

O PowerBI é um serviço de Business Analytics baseado em nuvem que oferece um ponto de visão única dos dados comerciais mais críticos. É possível também realizar um monitoramento da saúde do negócio da empresa usando Dashboards interativos e em tempo real, criar relatórios e muito mais.

Com isso no nosso projeto, a parte visualização dos dados finais foi abordada com o software Microsoft PowerBI, prestando uma automação e sincronização com a base de dados Mysql, retornando com uma performance praticamente instantânea e mantendo a integridade dos dados após realização de procedimentos internos. Foi possível a geração de relatórios quantitativos e qualitativos através da ferramenta e apresentação de Dashboards com aplicação em Cloud.

**7 DATA CLEANING**

Em virtude da análise feita no *dataset*, foram encontrados campos de diversos estudantes que “faltaram à prova”, campos não preenchidos de forma correta, “idade” que não se tinha justificativa (Ex: um ano de vida) onde se tornou necessária a remoção dessas informações e as que não se fazem correlacionais ao tema do nosso projeto.

Para tal procedimento de limpeza dos *datasets*, foi feita a seleção das colunas através do dicionário disponibilizado em conjunto com a base de dados, onde a documentação estrutural está contida e com isso retiramos as que seriam úteis o e que se fazem correlacionais com a questão socioeconômicas do candidato.

Toda essa transformação e alteração dos dados, foram feitas através de *querys* na aplicação do banco de dados estruturado Mysql e com isso, retornando informações concretas após longo tempo de processamento na base de dados.

No início do estudo foi verificado que o dataset do ano de 2015 apresenta em sua base bruta no total de 166 colunas e após análise total, verificamos que para abordagem a nossa correlação socioeconômica iríamos utilizar ao total de 96 colunas, as quais, resultam em dados mais claros e objetivos na informação final e visualização dos gráficos. Essa validação, foi replicada aos demais anos para uma verificação final.

**8 ANÁLISE EXPLORATÓRIA**

Em suma a análise exploratória de dados é considerada um dos pontos mais cruciais para o processo de análise em bases de dados. Quer dizer que é um processo que adota: organização, resumo, aplicação de alguns ou vários cálculos e visualização dos dados presentes na análise. Esse processo foi desenvolvido pelo grande estatístico, *John W. Tukey* e com o livro *Exploratory Data* *Analysis (EDA)* lançado em 1977, fez uma contribuição memorável para o mundo científico e social, unindo pensamento estatístico aos processos de transformação e exploração de dados.

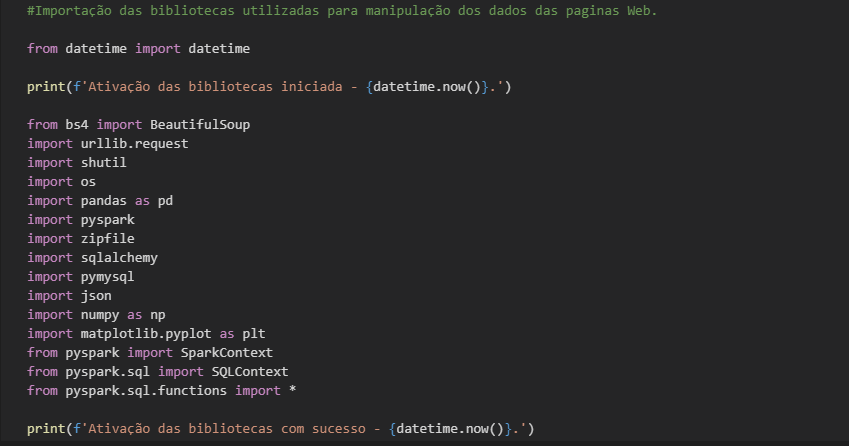
Com isso em nossa análise exploratória está sendo feita a aplicação em *datasets* providos do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP e realizando a verificação para a possibilidade de diminuirmos a motivo da falta de acesso ao ensino superior derivado de questões socioeconômicas.

Para isso adotamos em nosso projeto, aplicações em linguagem Python, desenvolvendo os códigos com integração a *engine* de banco de dados relacional e com isso possamos compreender, gerar tabelas de dados, visualizações concretas, exposição de fatos numéricos com base em fórmulas e modelos estatísticos, para enfim podermos apresentar a visualização dos dados.

Em torno do processo abordaremos a identificação dos padrões de cada item, saber quais seus pontos fortes e fracos, diminuir a desigualdade e com isso buscando uma padronização final na análise dos *datasets* colhidos através da base de dados.

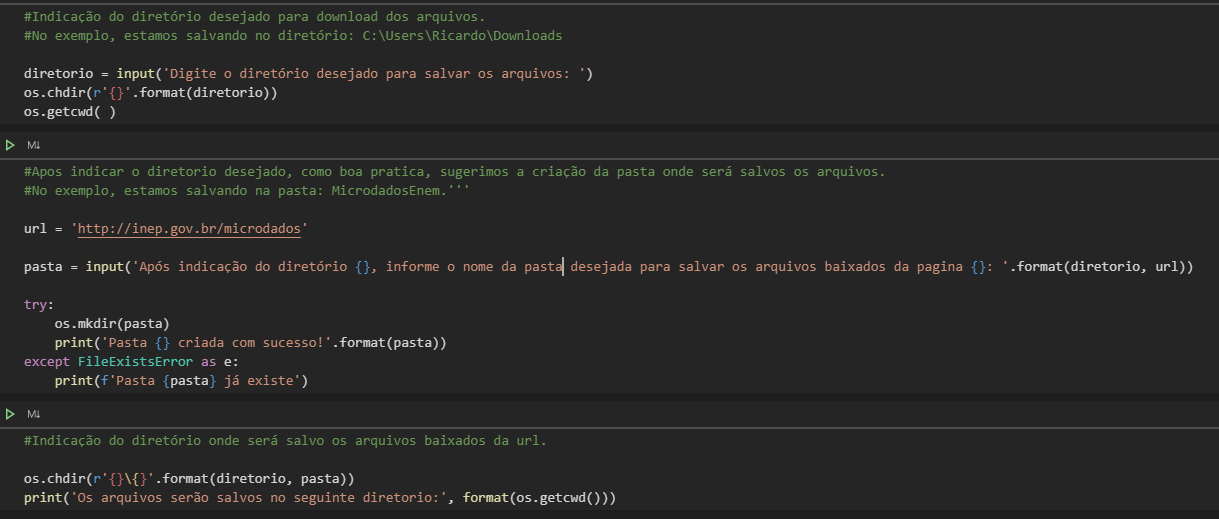
**8.1 Desenvolvimento da aplicação dos códigos:**

**8.1.1** -> Criação das bibliotecas para início da execução do código em linguagem Python.



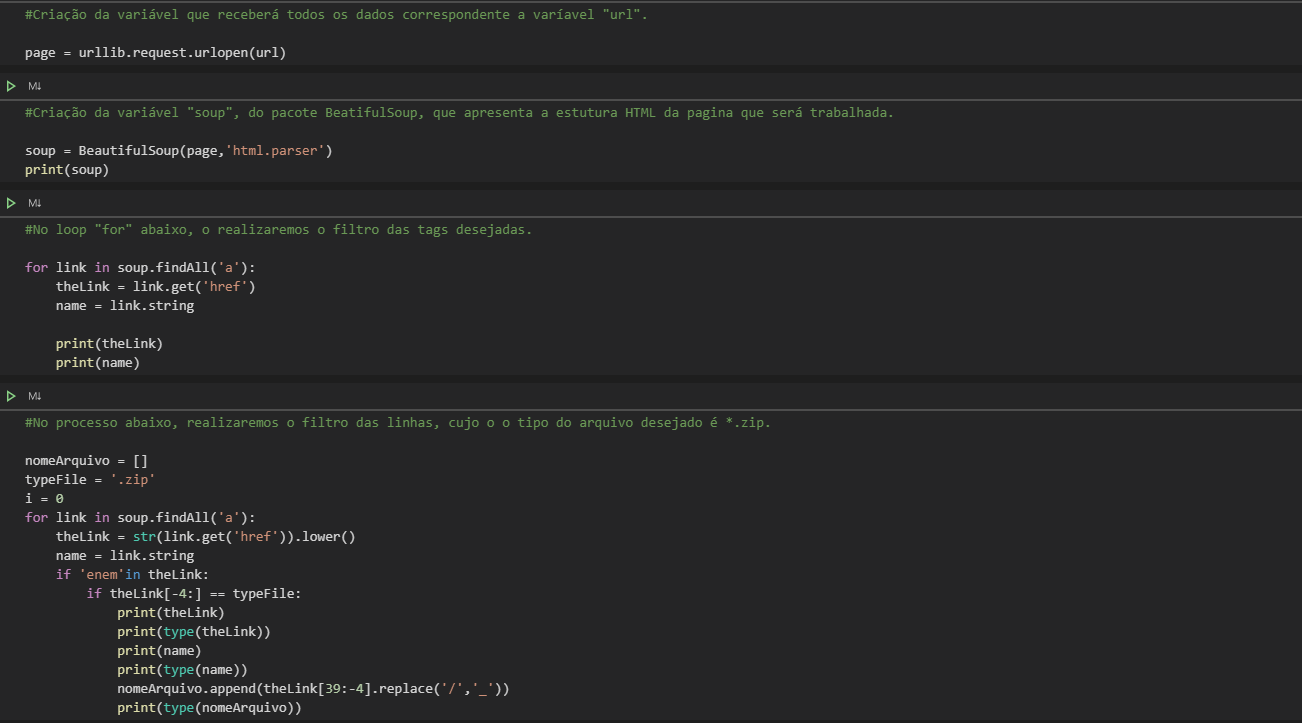
**Fig. 3 - Aplicação de bibliotecas - Linguagem Python**

**8.1.2** -> Criação e configuração dos diretórios locais para salvar os datasets que serão importados da base de dados do INEP.

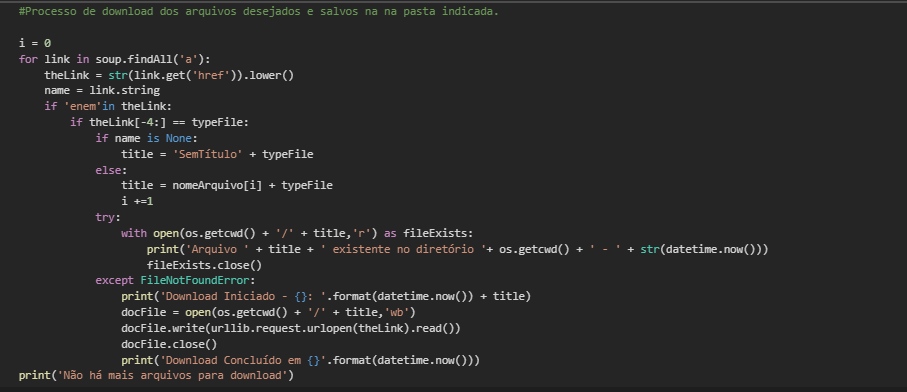


**Fig. 4 - Criação dos diretórios para salvar datasets**

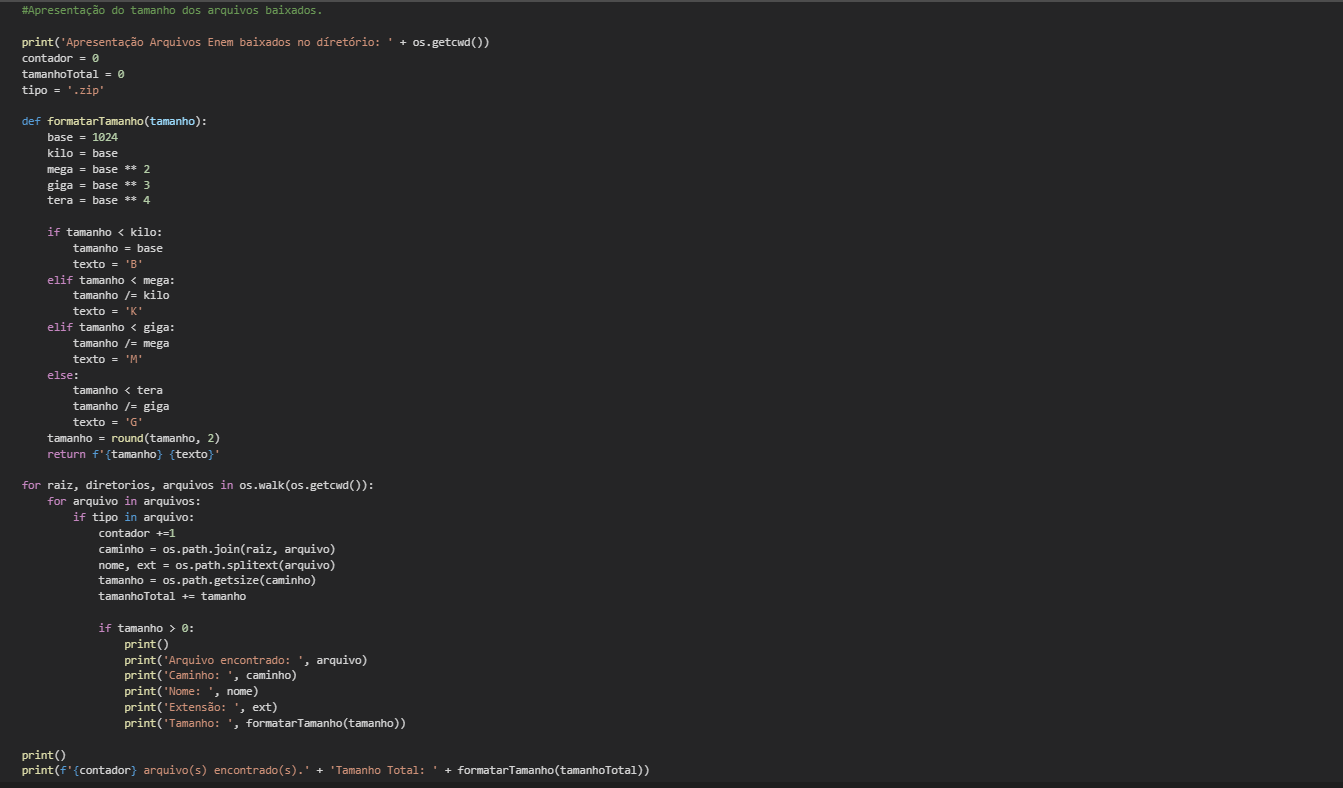
**8.1.3** -> Processo de Web Scraping no site do INEP, buscando importar microdados referente ao ENEM



**Fig. 5 - Criação da variável para recebimento dos dados da “url”**

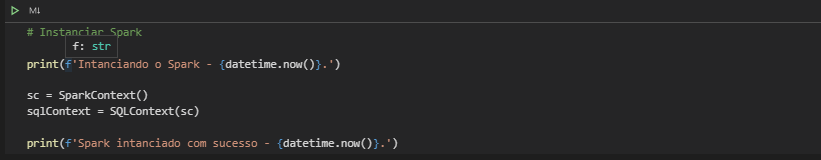


**Fig. 6 - Download dos arquivos para importação dos microdados**



**Fig. 7 - Apresentação do tamanho dos arquivos importados do INEP**

**8.1.4** -> Processo de inicialização do serviço Spark



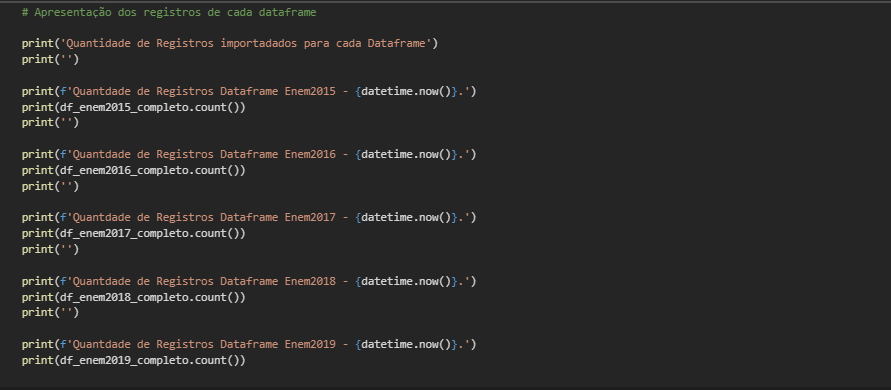
**Fig. 8 - Instanciando o serviço Spark**

**8.1.5** -> Processo de importação para o dataframe criado, para verificação do separados dos microdados, dos anos que serão feitos o estudo de 2015 à 2019



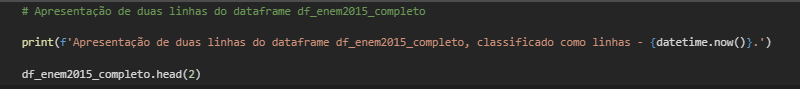
**Fig. 9 - Importação dos microdados para o Dataframe**

**8.1.6** -> Feito processo de verificação de registros para cada data frame que será analisado. De 2015 a 2019.



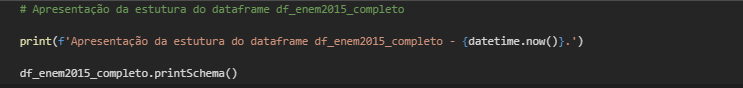
**Fig. 10 - Quantidade de registros por cada dataframe**

**8.1.7** -> Apresentação de duas linhas do dataframe de referência ao ano de 2015.



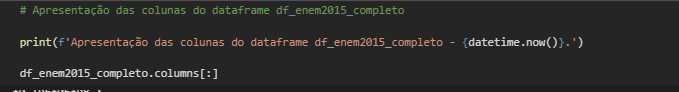
**Fig. 11 - Impressão de duas linhas referentes ao dataframe 2015**

**8.1.8** -> Apresentação da estrutura para o dataframe de referência ao ano de 2015.



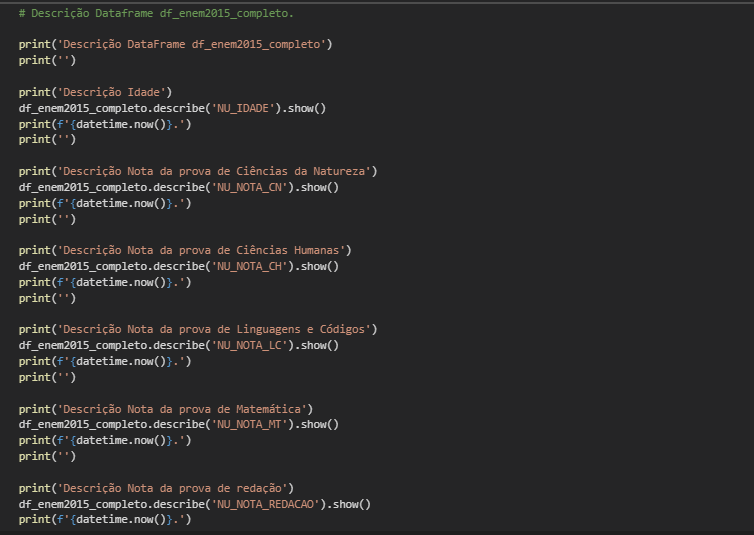
**Fig. 12 - Impressão da estrutura do dataframe ao ano de 2015**

**8.1.9** -> Apresentação das colunas para o dataframe de referência ao ano de 2015.



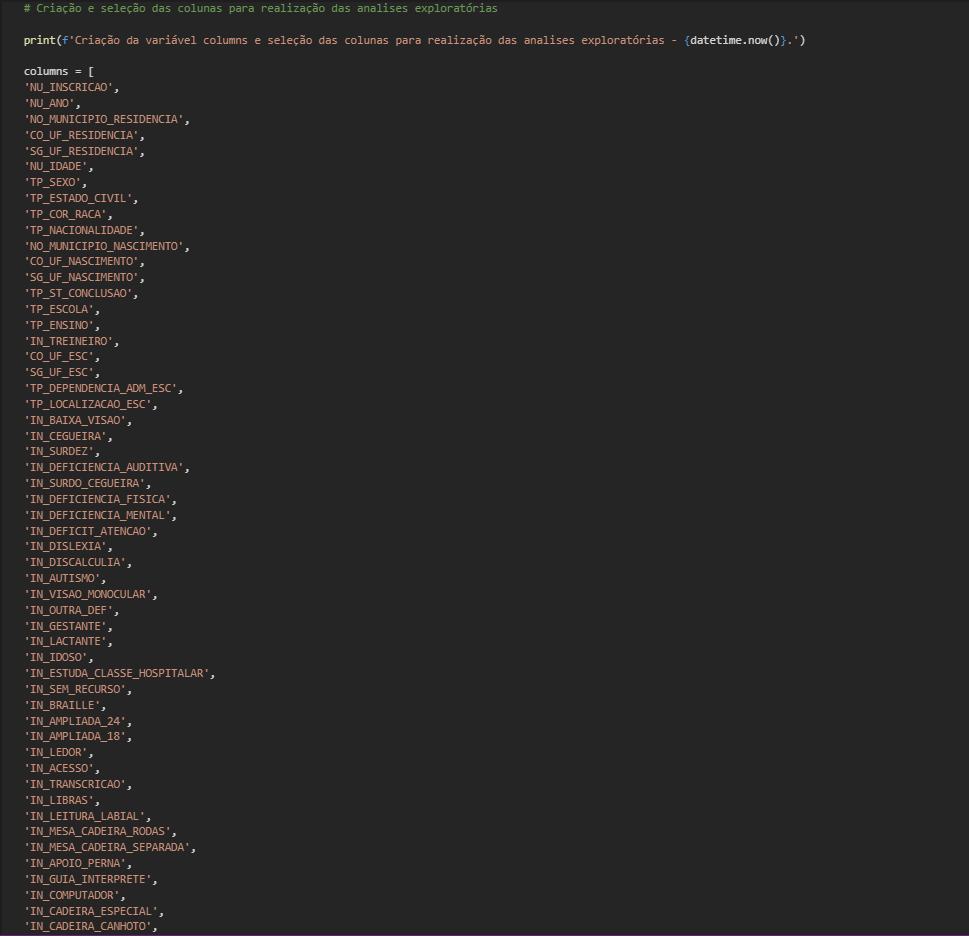
**Fig. 13 - Impressão das colunas referentes ao dataframe de 2015**

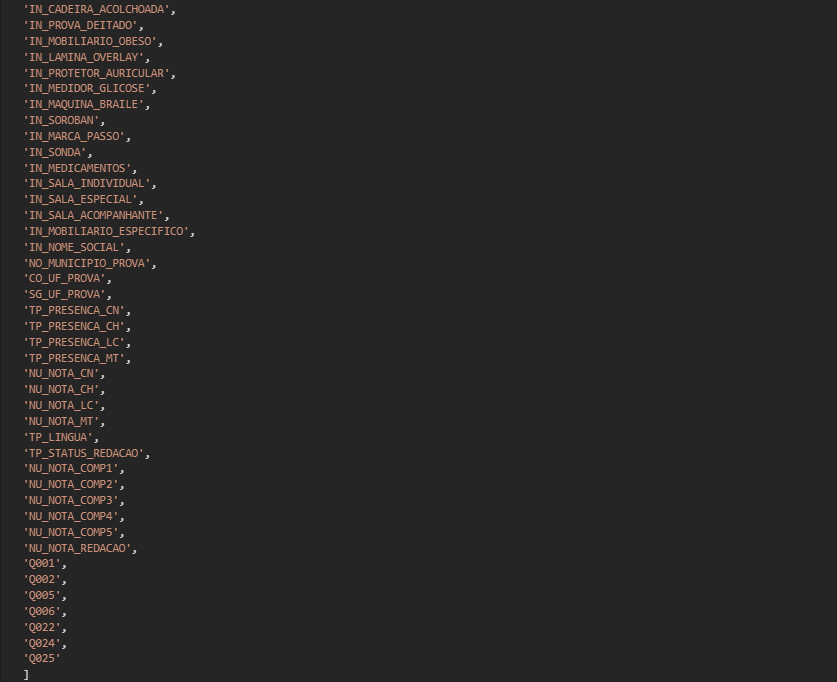
**8.1.10** -> Descrição do dataframe ao ano de 2015, realizando referência as disciplinas exigidas no exame nacional do ensino médio. Em relação a quantidade de alunos presentes área da prova, nota (realizando mínimo e máximo), idade (realizando mínimo e máximo).



**Fig. 11 -> Apresentação da descrição do dataframe para referência do ano de 2015**

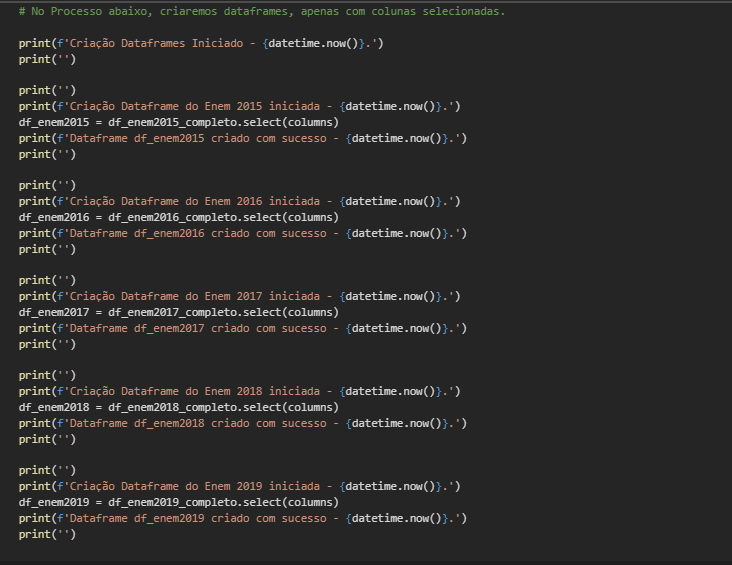
**8.1.11** -> Criação e seleção das colunas para realização das análises exploratórias em cada data frame, do ano de 2015 a 2019.





**Fig. 12->Apresentação das colunas para realização da análise exploratória**

**8.1.12** -> Criação dos data frames, com as colunas que serão utilizadas no processo de análise exploratória.



**Fig. 13 -> Criação dos data frames, do ano de 2015 a 2019**

**8.2 Regressão Linear**

Sendo uma grande aliada para o acompanhamento de possíveis cenários de um negócio e suporte para uma tomada de decisão. Uma aplicação da regressão linear se deve em entender um determinado conjunto de variáveis independentes como por exemplo: sexo, raça, nacionalidade, estado civil, idade, oriundos do nosso projeto e está estatisticamente relacionado com o exame e com isso gerando uma amostra de resultado. A nossa abordagem realiza a análise de regressão para ver o quão relacionada é cada variável e isso geralmente requer uma interação para encontrar uma possível combinação certa de variáveis para aplicação a um melhor modelo.

Com a técnica de regressão linear realizaremos um apontamento de quais regiões do país apresentam maior defasagem socioeconômica em relação a questões familiares, de falta de acesso à informação ou uma base social para uma qualidade de ensino. Com isso aplicamos códigos de integração com os *datasets* apontando colunas que fazem referência direta a níveis de estudo dos familiares de cada candidato.

**9 CONFIGURAÇÃO DO AMBIENTE**

**9.1 - Softwares**

9.1.1 Instalação do Python:

1. Acesse o site:

[**https://www.python.org/downloads/**](https://www.python.org/downloads/)

2. Realize o download e instale a versão *Python 3.7.6 64-bit*

9.1.2 Configurar Banco de Dados:

1. Acesse o site: **<https://dev.mysql.com/downloads/workbench/>**
2. Realize o download e instale da versão 8.0.20 do *MySQL Workbench*;
3. Crie o usuário da aplicação;
4. *CREATE USER 'admin'@'%' IDENTIFIED BY ‘root;*
5. *GRANT ALL PRIVILEGES ON \*.\* TO 'admin'@'%';*

9.1.3 Instalar e configurar IDE

1. Acesse o site: <https://code.visualstudio.com/>
2. Realize o download e instale a versão 1.48.0

9.1.4 Carregar o projeto na IDE

1. Baixe o projeto:

<https://github.com/projetobigdatasenac2020>

**9.2 - Infraestrutura do ambiente**

**9.2.1 - Instalação Python**

1 - Acesse o site:

<https://www.python.org/ftp/python/3.8.5/python-3.8.5-amd64.exe>

2 - Realizar o download da versão:

Versão 3.8.5

**9.2.2 - Instalação Java - JDK**

1 - Acesse o site:

<https://www.oracle.com/br/java/technologies/javase/javase-jdk8-downloads.html#license-lightbox>

2 - Realizar o download da versão:

    Versão: JDK - 1.8.0

**9.2.3 - Instalação Apache Spark**

1 - Acesse o site:

<https://archive.apache.org/dist/spark/spark-3.0.0/spark-3.0.0-bin-hadoop2.7.tgz>

2 - Realizar o download da versão:

     Versão: Apache Spark 3.0.0

3 - Procedimentos:

3.1 - Descompactar o arquivo para a pasta de destino

    3.2 - Salvar o arquivo winutils.exe na pasta bin

3.3 - Salvar o arquivo mysql-connector-java-8.0.20 na pasta jars, para conexão ao banco de dados Mysql

4 - Instalação biblioteca python

* Instalação matplotlib - via cmd
* Abrir console cmd => Digitar - pip install matplotlib
* Para teste da instalação da biblioteca no python, se faz necessário a verificação abaixo:
* Abrir console cmd e conectar ao python e aplicar import a biblioteca que deseja.

Segue abaixo:

* Abrir console cmd => Digitar - python
* Após digitar => import matplotlib.pyplot as plt
* Caso não ocorra nenhum erro, pois, a instalação se concluiu com sucesso.

**9.2.4 - Instalação Jupyter Notebook**

Executar console cmd = > digitar - pip install pupyter

**9.2.5 - Configuração de variáveis de ambiente - Windows**

1 - Acesso

* Acesso ao computador local:
  + Abrir meu computador
  + Propriedades
  + Configurações avançadas do sistema
  + Variáveis de ambiente
  + Variáveis de usuário
  + Opção path

**10 ANÁLISE DESCRITIVA**

Para esta etapa,utilizaremos bibliotecas da linguagem de programação Python, conhecida como Matplotlib e Seaborn. Sendo o matplotlib uma biblioteca para gerar gráficos 2D de arrays em Python. Embora Matplotlib seja escrito principalmente em Python puro, ele faz uso intenso de NumPy e de outros códigos de extensão para fornecer um bom desempenho, mesmo para matrizes grandes. Já o seaborn é uma biblioteca de visualização de dados Python baseada em matplotlib. Ele fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos.

Temos uma melhor diagramação e apresentação do estudo, para que os insights gerados possam promover cada vez mais a tomada de decisão em algum órgão público ou privado, sendo ele de amplitude estadual ou federal, dentro do escopo de nossa atuação do projeto. Após a verificação feita na etapa de análise exploratória dos dados é possível apresentar padrões ou conexões com maior rapidez e eficiência.

Com isso, teve-se a possibilidade de realizar a correlação entre informações contidas nos datasets dos anos de 2015 a 2019 e gerar uma análise descritiva de forma totalmente aprofundada fazendo-se referência ao tema de nosso projeto que se deriva as questões socioeconômicas de cada candidato, que está à prestar o exame do ENEM. Diversas questões foram abordadas, onde tivemos algumas dificuldades para enquadramento das possíveis colunas disponibilizadas e com bastante cautela, realizamos o apontamento para as que mais se referenciam ao estudo, objetivando clareza na informação que for repassada e concretização dos dados contidos em cada gráfico.

Na sequência, será apresentada informações geradas através do estudo e repassada de uma forma intuitiva e gráfica para uma melhor visão e entendimento na amplitude do tema proposto ao projeto.

**10.1** – Análise realizada para quantidade de inscrições dos candidatos por ano, de 2015 a 2019.

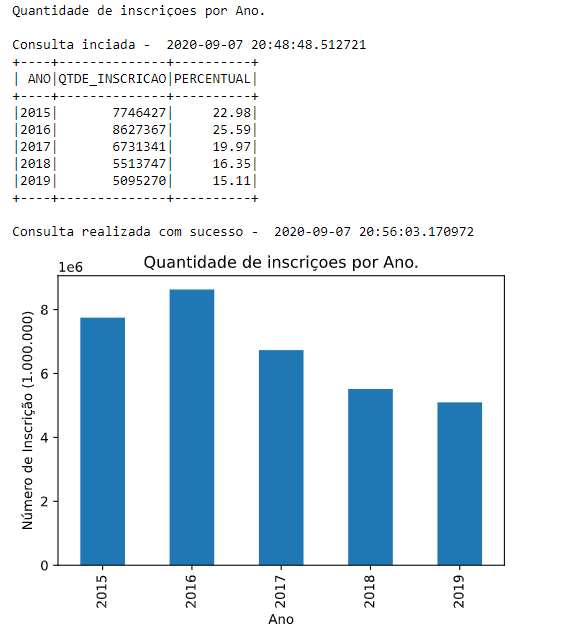


Fig. 14 -> Quantidade de alunos por ano

**10.2** – Análise realizada para a quantidade de inscrições realizadas por estado em todo território nacional

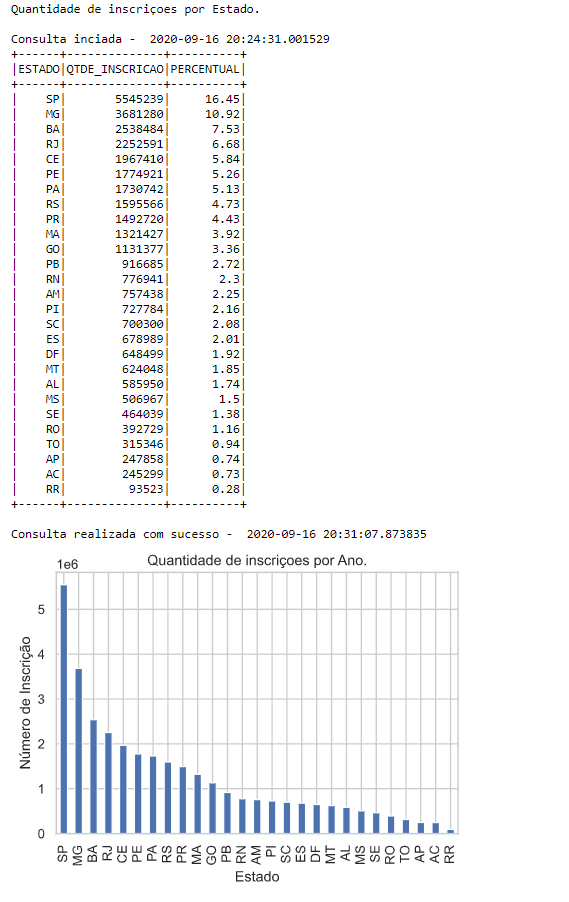


Fig. 15 -> Quantidade de inscrições por Estado

**10.3** – Análise realizada para a verificação da idade de cada candidato ao exame prestado.

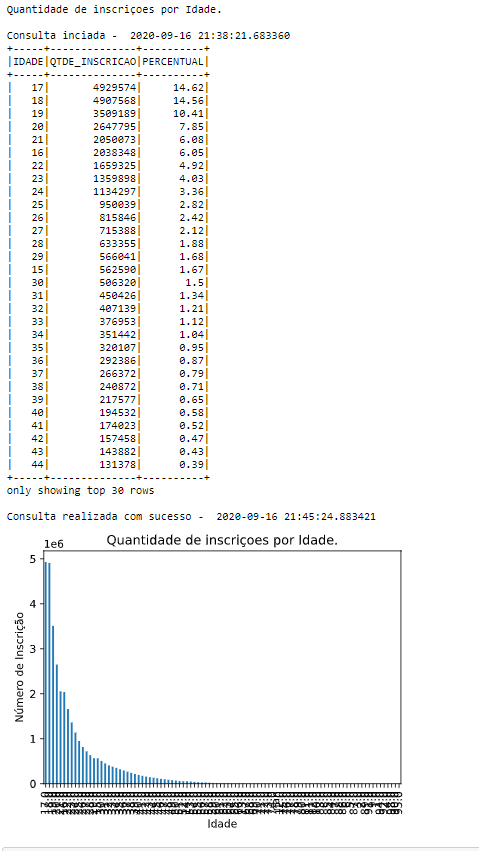


Fig. 16 -> Quantidade de candidatos por idade

**10.4** – Análise realizada par verificação por sexo de cada candidato ao exame.

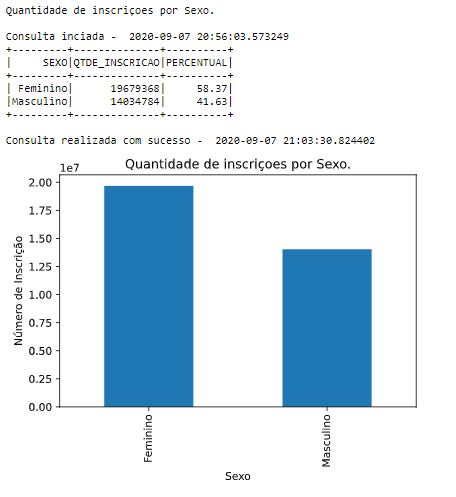


Fig. 17 -> Quantidade de candidatos por sexo

**10.5** – Análise realizada para verificação por sexo a cada ano, compreendendo o período de 2015 a 2019.

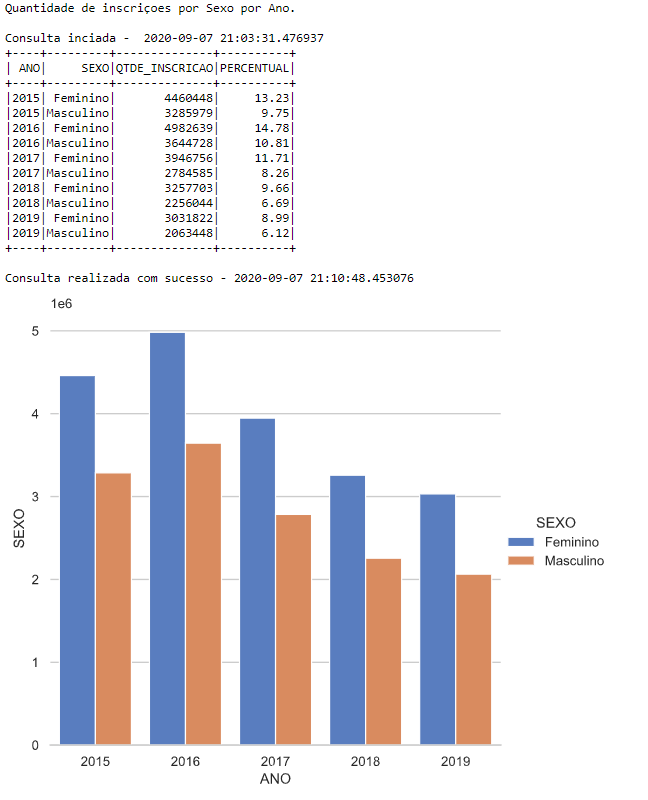


Fig. 18 -> Quantidade de candidatos por ano e sexo

**10.6** -> Análise realizada para referência a raça de cada candidato

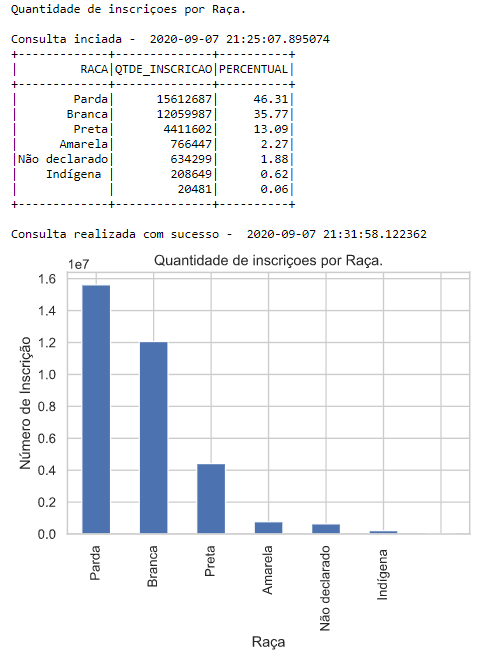


Fig. 19 -> Quantidade de inscritos por raça

**10.7** -> Análise realizada para averiguação em relação a raça de cada candidato por ano prestando, compreendendo o período de 2015 a 2019.

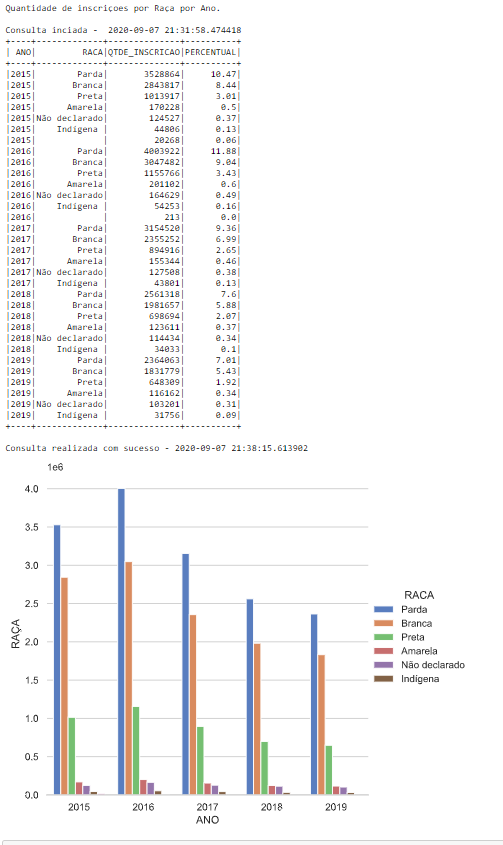


Fig. 20 -> Quantidade de inscritos por raça de 2015 a 2019

**10.8** -> Análise realizada para levantamento de informações abrangendo a quantidade de alunos, a qual, verifica o status de ensino para cada candidato ao exame, compreendendo os anos de 2015 a 2019.

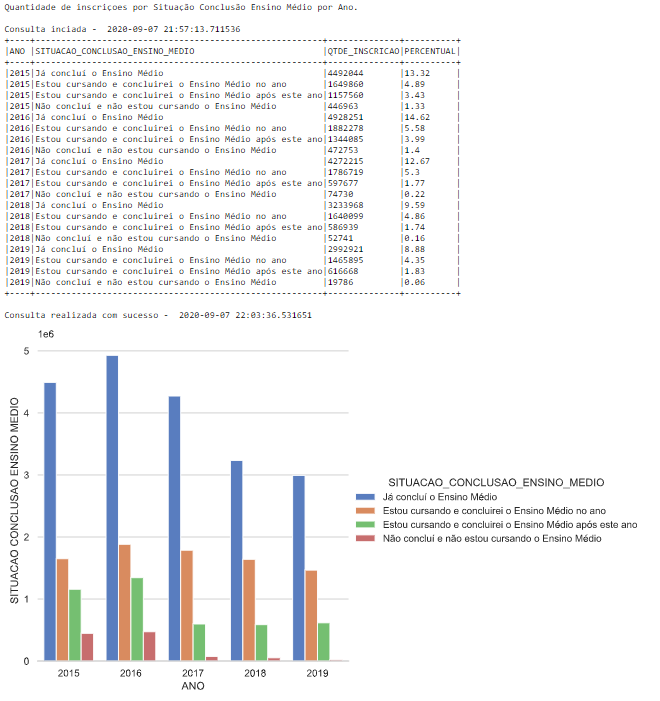


Fig. 21 -> Verificação da quantidade de alunos baseados pelo status de ensino

**10.9** -> Análise realizada para levantamento de qual tipo de instituição o candidato desenvolveu durante sua trajetória acadêmica, compreendendo os anos de 2015 a 2019.

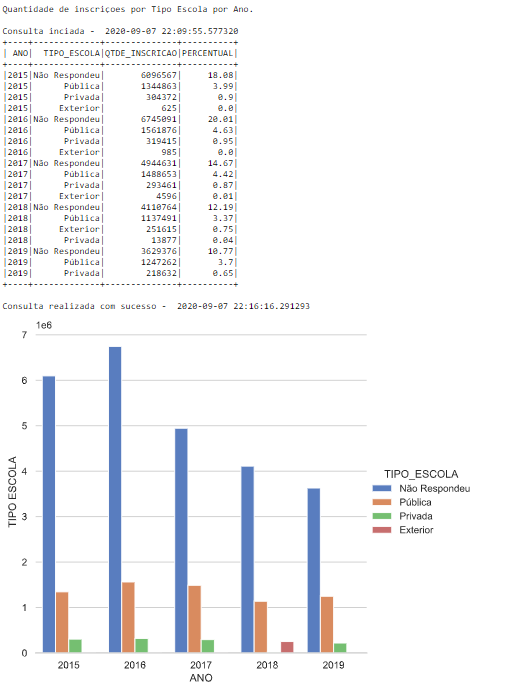


Fig. 22 -> Quantidade de alunos por tipo de escola

**10.10** -> **ACESSIBILIDADE AO EXAME**

Realizamos a análise de diversas tabelas contidas nos datasets dos anos de 2015 a 2019 e foi verificado itens interessantes para serem repassados e abordados neste capítulo. Com isso, disponibilizaremos gráficos informando questões de acessibilidade ao exame do ENEM, para todos os candidatos que necessitam de algum tipo de cuidado e atenção durante a prova.

**10.10.1**-> Análise realizada para verificação da quantidade de candidatos com cegueira por ano, compreendendo de 2015 a 2019

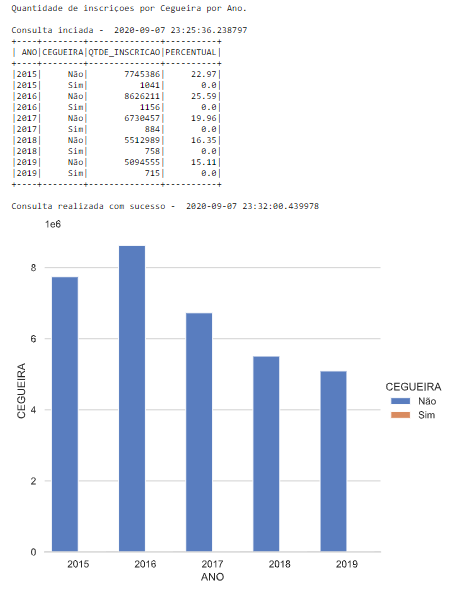


Fig. 23 -> Quantidade de candidatos com deficiência visual

**10.10.2** -> Análise realizada para verificação da quantidade de candidatos com deficiência física, compreendendo os anos de 2015 a 2019.

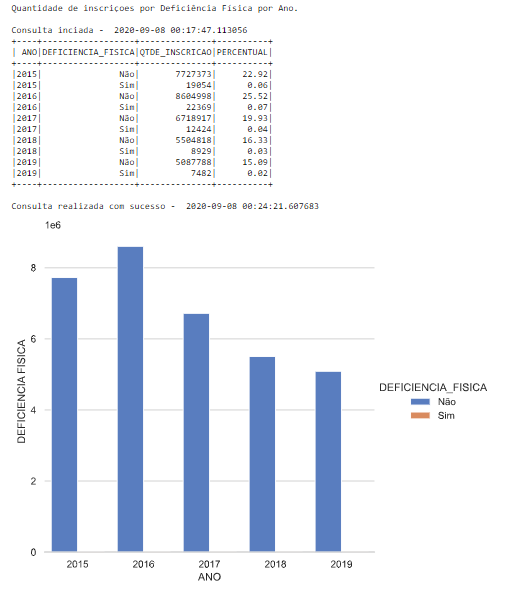


Fig. 24 -> Quantidade de candidatos com deficiência física

**10.10.3** -> Análise realizada para verificação da quantidade de candidatos gestantes, compreendendo os anos de 2015 a 2019.

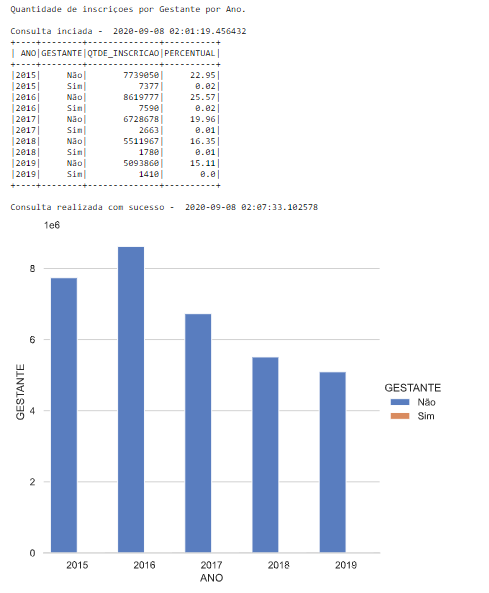


Fig. 25 -> Quantidade de gestantes de 2015 a 2019

**10.10.4** -> Análise realizada para verificação da quantidade de candidatos que necessitam de apoio com Libras, compreendendo os anos de 2015 a 2019.

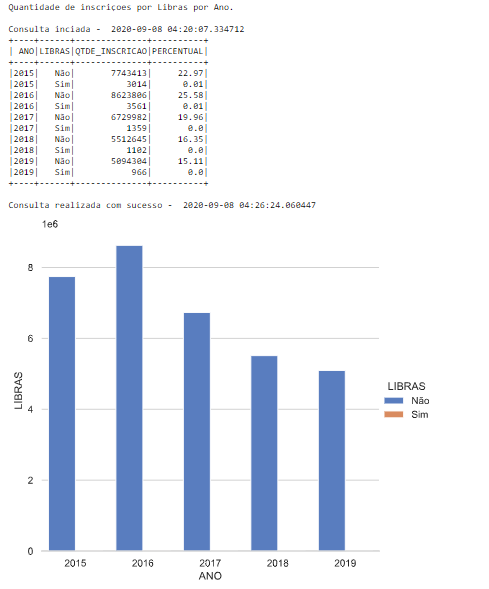


Fig. 26 -> Quantidade de candidatos que necessitam de intervenção em Libras

**10.11** -> CORRELAÇÃO SÓCIO ECONÔMICA AO CANDIDATO

Realizado o procedimento de correlação das informações entre tabelas contidas nos datasets compreendendo os anos de 2015 a 2019, onde com isso podemos apurar questões socioeconômicas de cada candidato, envolvendo seu âmbito familiar, acesso à tecnologia e até onde irá atingir no resultado do exame nacional do ensino médio. Com isso, abaixo, segue todo estudo de forma gráfica e descrita com legendas para que seja objetiva e clara a informação dos dados repassados desse projeto.

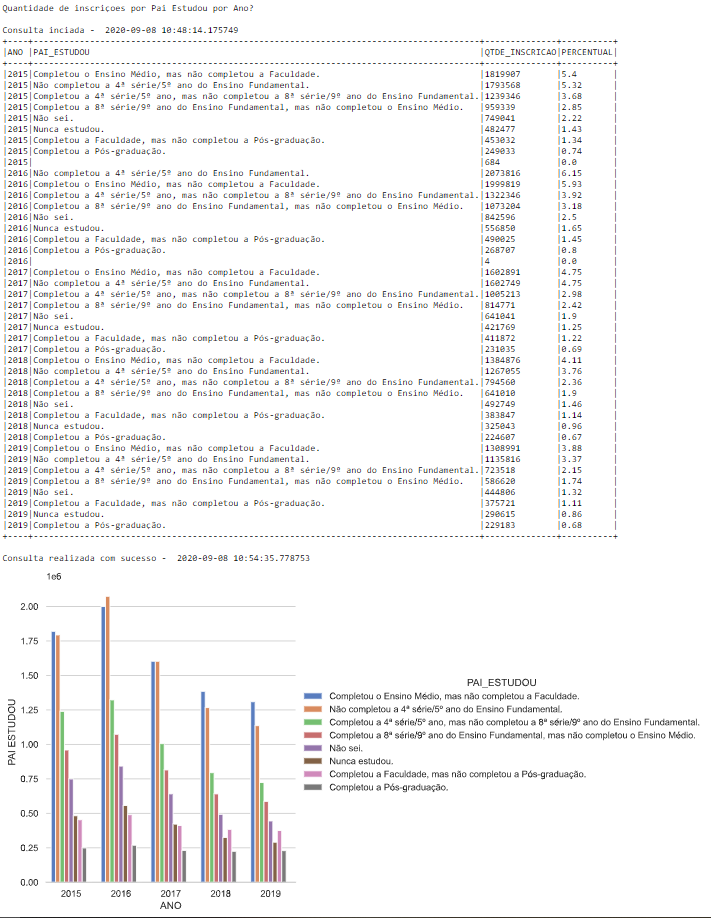
**10.11.1** -> Análise realizada para verificação do status de ensino do pai do candidato, podendo influenciar no resultado, compreendendo os anos de 2015 a 2019.

Fig. 27 -> Condição de ensino ao pai do candidato

**10.11.2** -> Análise realizada para verificação do status de ensino da mãe do candidato, podendo influenciar no resultado, compreendendo os anos de 2015 a 2019



Fig. 28 -> Condição de ensino a mãe do candidato

10.11.3 -> Análise realizada para verificação da quantidade de celular por residência, compreendendo o período de 2015 a 2019.

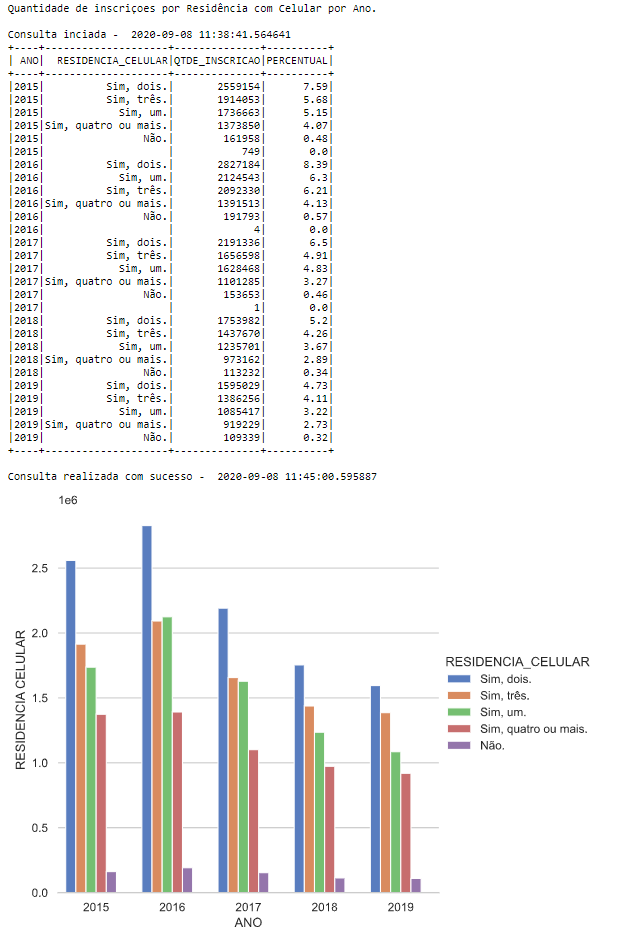


Fig. 29 -> Quantidade de celulares por residência

10.11.4 - Análise realizada para verificação da quantidade de inscrições por residência com computador, compreendendo os anos de 2015 a 2019.

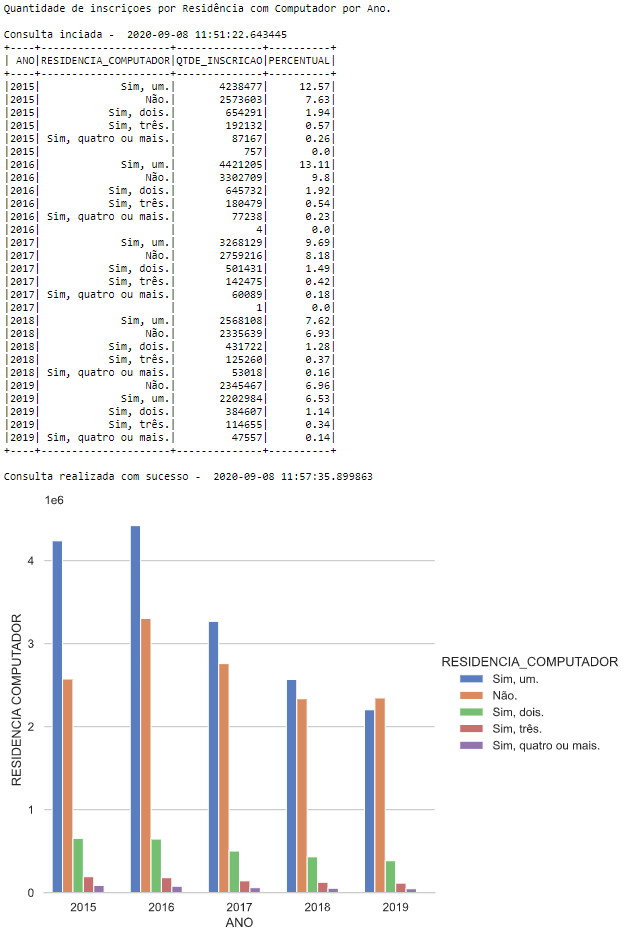


Fig. 30 -> Quantidade de inscrições por residência com computador

10.11.5 – Análise realizada para levantamento da quantidade de inscritos que possuem internet em sua residência, possibilitando maior acesso à informação, compreendendo os anos de 2015 a 2019.

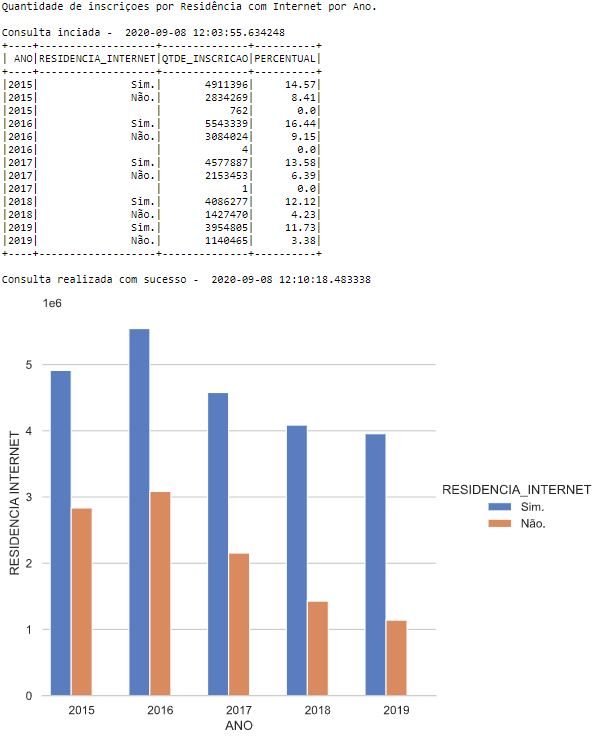


Fig. 31 -> Quantidade de inscrições por residência com internet

10.11.6 -> Análise realizada para verificação da quantidade de aprovados no exame, a regra atribuída foi que as notas em todas as áreas, somando-se seja maior que quatrocentos e cinquenta pontos, caso o aluno não obtenha essa pontuação mínima, o mesmo encontra-se reprovado e não podendo se candidatar a uma vaga. Com isso, o candidato possui seu status como aprovado e podendo ao mínimo se candidatar a uma vaga. Os anos compreendidos para estudo, foram de 2015 a 2019.

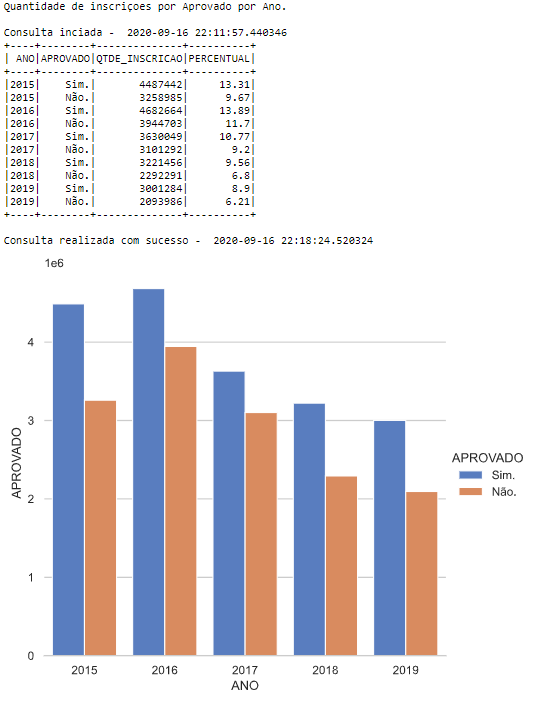
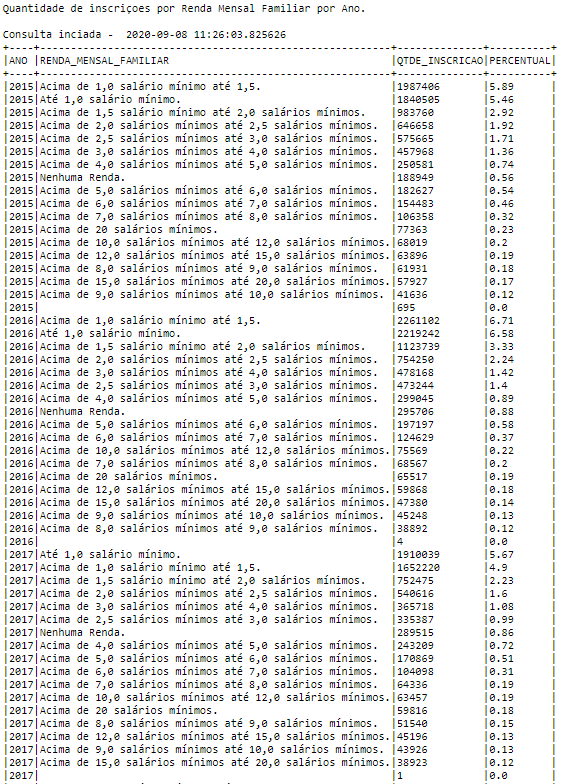


Fig. 32 -> Quantidade de aprovados no exame por ano

10.11.7 -> **Análise realizada para verificação da renda familiar de cada candidato para o exame nacional do ensino médio.**

Com essa verificação é possível fazer uma correlação mais aprofundada dos principais problemas na área da educação em nosso país. Em virtude de baixas condições financeiras, que milhares de candidatos possuem, é possível se ter uma breve ideia do baixo acesso à informação, de uma baixa qualidade de ensino em nível fundamental e médio, de uma falta de interesse das instituições de ensino em promover os devidos programas de intensificação ao exame do ENEM no último ano em que o candidato esteja cursando o ensino médio, dentre vários outros fatores que acarretam o desinteresse do aluno em prestar a prova, por fatores sociais provocados em virtude de questões financeiras familiares.



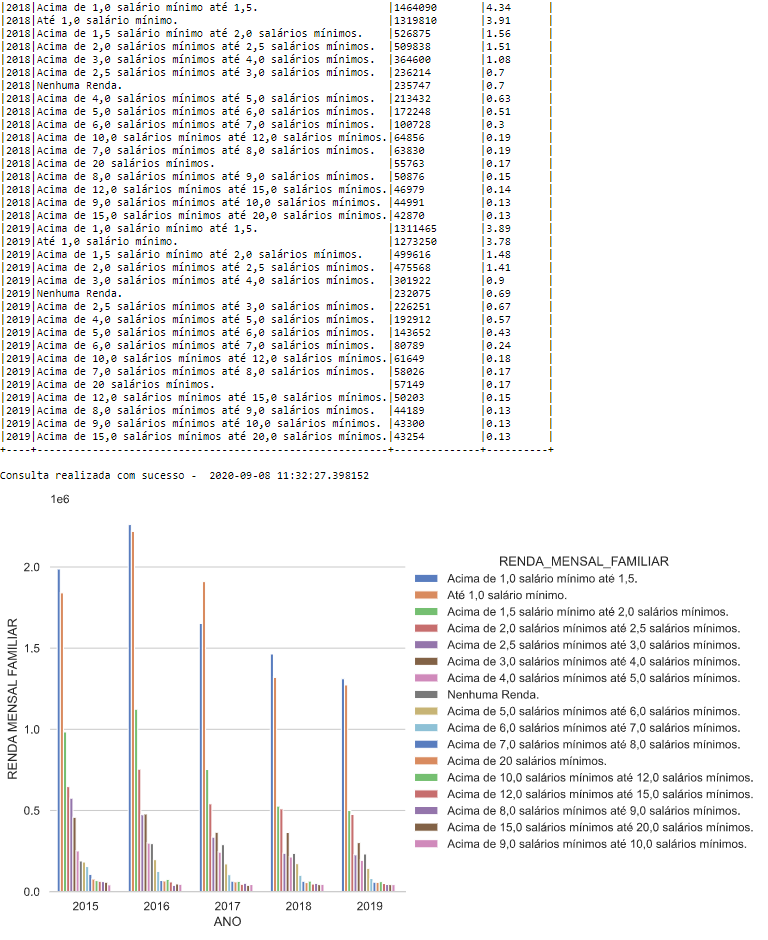


Fig. 33 -> Quantidade de candidatos por renda mensal família por ano

**11 CONCLUSÃO**

O presente estudo mostrou que realmente há uma degradação no acesso ao ensino superior em referência às escolas públicas e particulares no país, decorrente de critérios socioeconômicos e oriunda do nível de instrução educacional familiar e de critérios como acesso à tecnologia, dificultando um maior envolvimento com a disseminação de conhecimento e informações.

O projeto apresenta também uma aplicação de limpeza e tratamento dos dados utilizados inicialmente. Posteriormente, com as informações geradas através de ferramentas de Big Data e as correlacionando entre *datasets* disponibilizados pelo INEP, chegamos a resultados estatísticos de alta relevância no índice da educação brasileira e de total prospecção e base para análises futuras de outros exames do Exame Nacional do Ensino Médio (ENEM).

Deve-se, também, ao fato de que a intensificação do repasse de ensino aos familiares de cada estudante que futuramente vá se candidatar ao exame é de extrema importância, pois, o apoio e apontamento do jovem em relação a este âmbito é totalmente espelhado aos de seu âmbito social e familiar.

A solução principal seria um investimento em políticas educacionais de acesso e sociais para que se possa delinear estratégias de aprimoramento no ensino básico de acesso e intensificar a disseminação pelo repasse de informações em paralelo com a tecnologia.

Necessita-se de um aprimoramento de trabalho entre o governo federal e o estadual de cada região do pais, para que as instituições de ensino público e particular de todo território brasileiro tenham uma melhoria expressiva na qualidade de repasse em seus ensinos e seus parques tecnológicos.

**12 REFERÊNCIAS**

**INEP - Microdados**

* http://inep.gov.br/microdados

**Scielo - desempenho de estudantes no Enem 2010**

* <https://www.scielo.br/scielo.php?pid=S2176-66812013000200005&script=sci_arttext>

**Indicadores de Níveis Socio Econômicos dos inscritos do ENEM - INEP**

* <http://td.inep.gov.br/ojs3/index.php/td/issue/view/390/31>

**Congresso da Sociedade Brasileira de Computação - Previsão desempenho dos candidatos do ENEM**

* https://sol.sbc.org.br/index.php/ctic/article/view/3244/3206

**A reforma do ensino médio e a implantação do Enem no Brasil**

**Autores: Maria Helena Guimarães de Castro e Sergio Tiezzi**

* http://www.schwartzman.org.br/simon/desafios/4ensinomedio.pdf

**Repositório UNB: Limites e possibilidades do Enem no processo de democratização do acesso a educação superior brasileira**

* https://repositorio.unb.br/handle/10482/20129

**Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - História**

* https://pt.wikipedia.org/wiki/Instituto\_Nacional\_de\_Estudos\_e\_Pesquisas\_Educacionais\_An%C3%ADsio\_Teixeira

**Etapas do processo de Big Data**

* https://blog.leucotron.com.br/etapas-de-big-data/

**Matlotlib**

* https://pt.wikipedia.org/wiki/Matplotlib

**Introdução ao Apache Spark**

* https://www.devmedia.com.br/introducao-ao-apache-spark/34178

**Revista Brasileira de Ensino de Física**

* https://www.scielo.br/pdf/rbef/2015nahead/1806-1117-rbef-S1806-11173710001.pdf

**Arquivos Analíticos de Políticas Educativas - AAPE**

* https://www.redalyc.org/pdf/2750/275043450074.pdf

**KDD - Data Mining**

* https://itforum365.com.br/colunas/kdd-e-data-mining-mais-do-que-apenas-conceitos/

**Etapas Pré-Processamento de dados**

* https://www.datageeks.com.br/pre-processamento-de-dados/

**Pré-Processamento com Machine Learning**

* https://medium.com/@caiquecoelho/um-guia-completo-para-o-pr%C3%A9-processamento-de-dados-em-machine-learning-f860fbadabe1

**MYSQL - História / Características**

* https://pt.wikipedia.org/wiki/MySQL

**Visual Studio Code - História / Características**

* https://pt.wikipedia.org/wiki/Visual\_Studio\_Code

**Utilização da Ferramenta Jupyter Notebook**

* https://medium.com/@suzana.svm/por-que-usar-jupyter-notebook-77d5a59b42a1

**Influência de características socioeconômicas no desempenho dos candidatos - Enem 2012**

* http://www.abrapecnet.org.br/enpec/x-enpec/anais2015/resumos/R1960-1.PDF

**Revista Espacios - Saint-Jean-de-Védas, Occitania - França**

* https://www.revistaespacios.com/a17v38n05/a17v38n05p12.pdf

**Data Science Academy - Microsoft PowerBI Para Data Science**

* https://www.datascienceacademy.com.br/course?courseid=microsoft-power-bi-para-data-science

**Compreendendo dados com Análise Exploratória**

* https://medium.com/@gabriel.stankevix/analise-explorat%C3%B3ria-de-dados-732007ddbfaf

**Aplicabilidade da Regressão Linear**

* https://medium.com/@ligiagalvao/regress%C3%A3o-linear-um-dos-tipos-mais-importantes-de-an%C3%A1lise-de-dados-827cfdbd2229

**Utilização de biblioteca MatPlotLib**

* https://matplotlib.org/users/history.html

**Aplicação de biblioteca Seaborn - Visualização de Gráficos**

* https://seaborn.pydata.org/