

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS**NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

Pós-graduação *Lato Sensu* em Analytics e Business Intelligence

**RELATÓRIO TÉCNICO**

ORGANIZAÇÃO DOS DADOS DA PREVIDÊNCIA SOCIAL

Bruna Fernandes de Medeiros

Belo Horizonte

2021

**SUMÁRIO**

[1. Introdução 4](#_Toc87290320)

[1.1. Contexto 4](#_Toc87290321)

[1.2. Objetivos 4](#_Toc87290322)

[1.3. Público alvo 4](#_Toc87290323)

[2. Modelo de Dados 5](#_Toc87290324)

[2.1. Modelo Dimensional 5](#_Toc87290325)

[2.2. Fatos e Dimensões 5](#_Toc87290326)

[3. Integração, Tratamento e Carga de Dados 6](#_Toc87290327)

[4. Camada de Apresentação 17](#_Toc87290328)

[4.1 Dashboard 17](#_Toc87290329)

[4.2 Análises avançadas 17](#_Toc87290330)

[5. Registros de Homologação 17](#_Toc87290331)

[5. Conclusões 17](#_Toc87290332)

[6. Links 18](#_Toc87290333)

[REFERÊNCIAS 19](#_Toc87290334)

# 1. Introdução

## 1.1. Contexto

O Brasil é uma país que distribui muitos tipos de benefícios da previdência, são elas: por aposentadoria acidentária, por idade, invalidez, entre outras. Porém não é fornecida ferramenta que facilite o acesso a essas informações por categorias. E com isso não é possível diferenciar os valores destinados a cada tipo.

Entretanto as informações estão disponibilizadas em formato CSV de forma completa no site do governo, onde consta ano, sexo, quantidade de benefícios concedidos.

Sendo assim possível que as empresas de forma geral possam acessar essas informações mais detalhadas com gráficos e painéis para trabalha-las com o objetivo de obter insights de forma rápida.

## 1.2. Objetivos

Tornar as informações sobre os benefícios concedidos pela previdência social mais fácil de serem analisadas e visualizadas fornecendo um dashboard interativo.

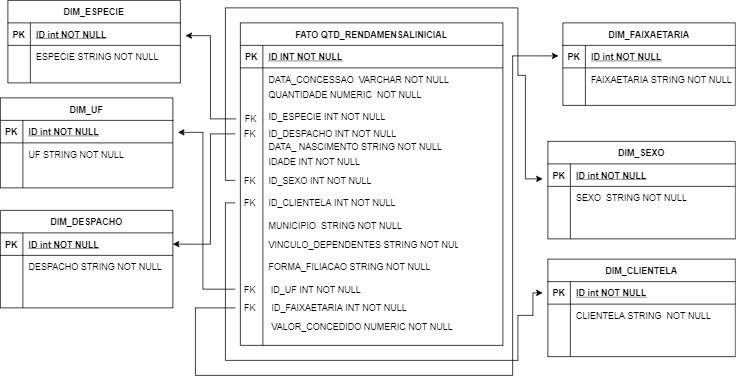
## 1.3. Público alvo

Pessoas interessadas em assuntos relacionados a previdência social, que tenham acesso à internet, redes sociais, e que estejam por dentro das notícias e se sintam dispostas a análises mais profundas do que o governo lhes oferece.

# 2. Modelo de Dados

## 2.1. Modelo Dimensional

O diagrama do modelo dimensional oferece uma visão sobre como os dados serão organizados para atender às necessidades identificadas por meio dos objetivos do projeto.



## 2.2. Fatos e Dimensões

Tabela Fato “QTD\_RENDAMENSALINICIAL”: contém os registros da quantidade renda recebida de acordo com as colunas: ID; DATA\_CONCESSAO; QUANTIDADE; ID\_ESPECIE; ID\_FAIXAETARIA; ID\_DESPACHO; DATA\_NASCIMENTO; ID\_SEXO; ID\_CLIENTELA; MUNICIPIO; VINCULO\_DEPENDENTES; FORMA\_FILIACAO, VALOR\_CONCEDIDO; IDADE E ID\_UF.

As dimensões são 5, divididas em:

“DIM\_ESPECIE”, onde contém os registros do tipo do benefício. Exemplo: “Pensão por morte previdenciária”, “Auxilio salário maternidade”, “Auxilio doença previdenciário”, entre outros. Representada pelas colunas ID e ESPECIE.

“DIM\_UF”, nessa tabela consta os estados brasileiros. Exemplo: Bahia, São Paulo, etc. Representada pelas colunas ID e UF.

“DIM\_DESPACHO”, nessa tabela consta a forma de concessão. Exemplo: Concessão normal, decorrente de ação judicial, entre outras. Representada pelas colunas ID e DESPACHO.

“DIM\_SEXO”, representa o sexo do beneficiário. Exemplo: feminino ou masculino. Representada pelas colunas ID e SEXO.

“DIM\_CLIENTELA”, consta a área geográfica em que o beneficiário reside. Exemplo: urbano ou rural. Representada pelas colunas ID e CLIENTELA.

“DIM\_FAIXAETARIA”, onde consta a divisão das idades. Exemplo: 15-25, 26-39.

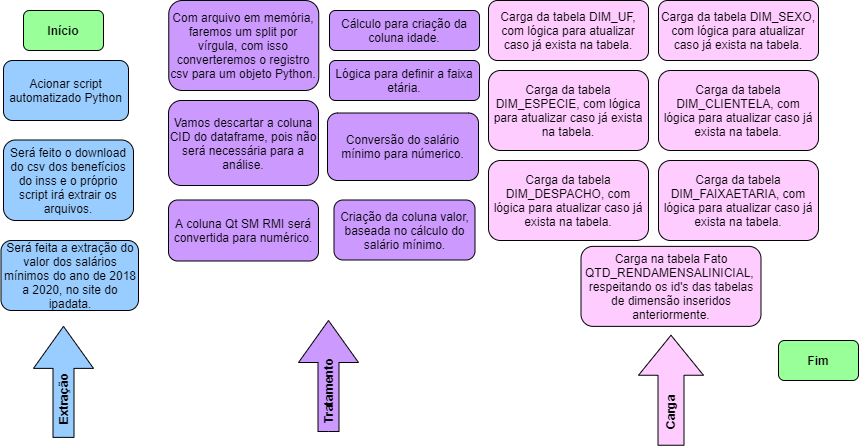
# 3. Integração, Tratamento e Carga de Dados

O processo de extração, tratamento e carga de dados será feito, nos arquivos do ano de 2018, 2019 e 2020. Com script Python, utilizando biblioteca Pandas, o arquivo será acessado no site: <https://dados.gov.br/dataset/inss-beneficios-concedidos> onde fará o download dos arquivos, também será feito download da base do salário mínimo no site: http://www.ipeadata.gov.br/exibeserie.aspx?stub=1&serid1739471028=1739471028 finalizando a etapa de extração.

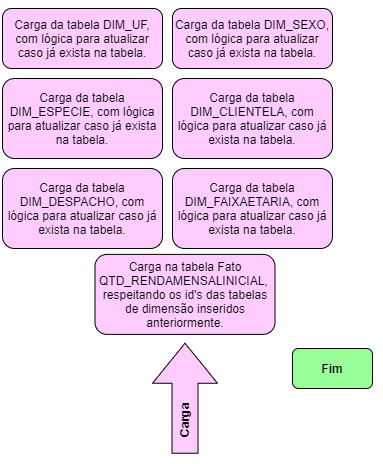
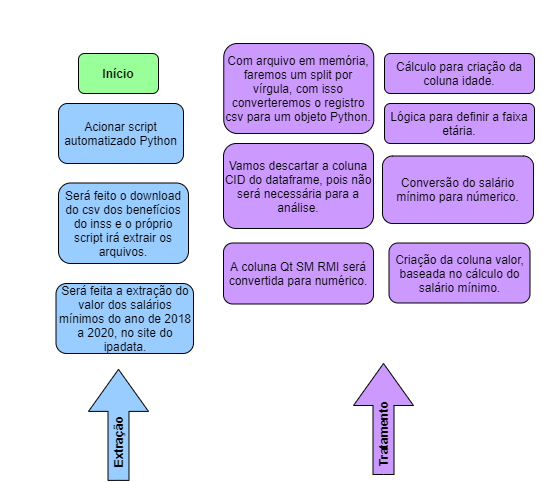
Para iniciar a etapa de tratamento com os arquivos em memória será feito um split por vírgula, convertendo o registro CSV para um objeto Python, descartando a coluna CID, pois não será interessante para a nossa análise. A coluna QT SM RMI deve ser convertida para numérico, pois usaremos a mesma para calcular o valor da concessão com base no salário mínimo criando a coluna Valor. Criação das colunas ID\_FAIXAETARIA e IDADE com base na coluna DT NASCIMENTO do CSV, para melhor análise e enriquecimento dos dados no processo de ETL que será orquestrado pelo Python3.

Já no processo de carga de dados, iremos inserir nas tabelas de dimensão, com a lógica para atualizar caso já exista na tabela e ao final do processo iremos popular a tabela de fato respeitando os ids das tabelas de dimensão inseridos anteriormente. Todo o processo de carga será feito no banco de dados relacional MySql.

Abaixo fluxograma das etapas:



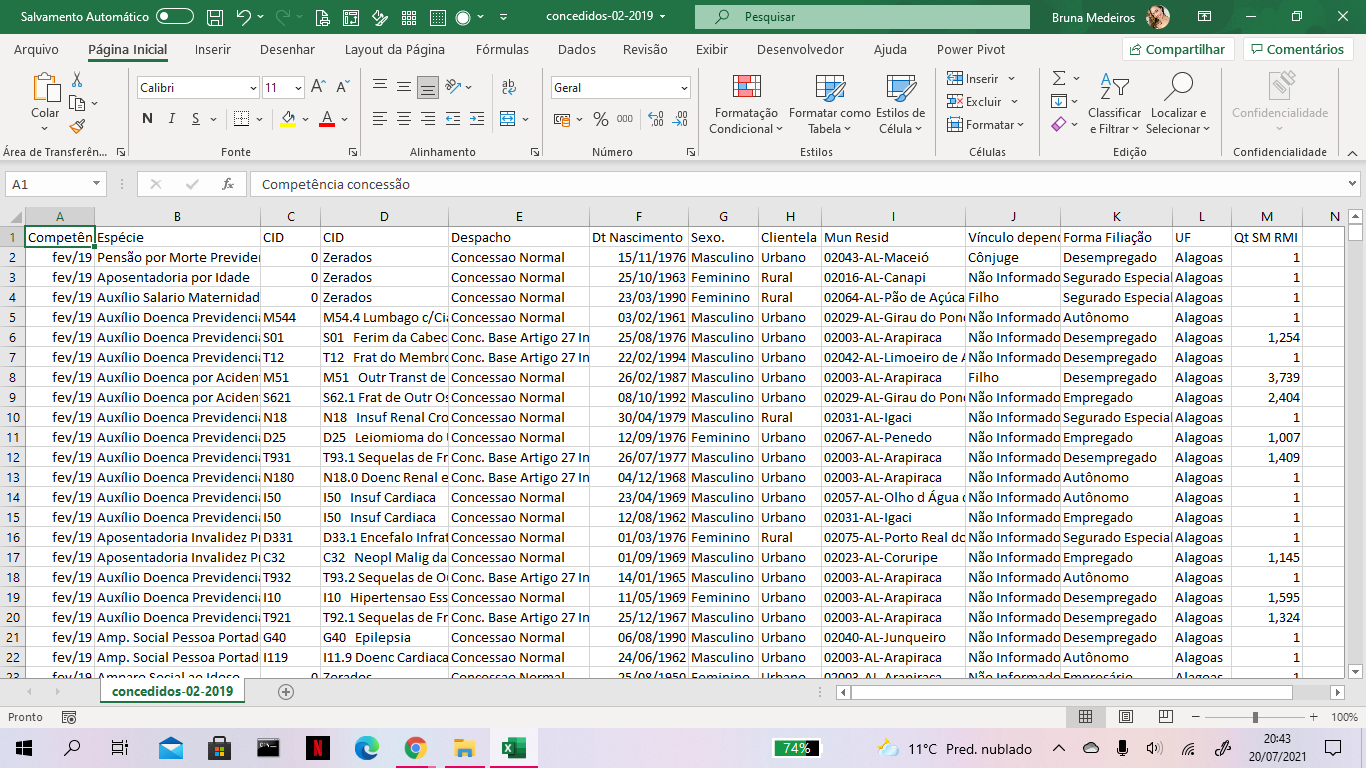
Fluxograma com zoom das etapas:



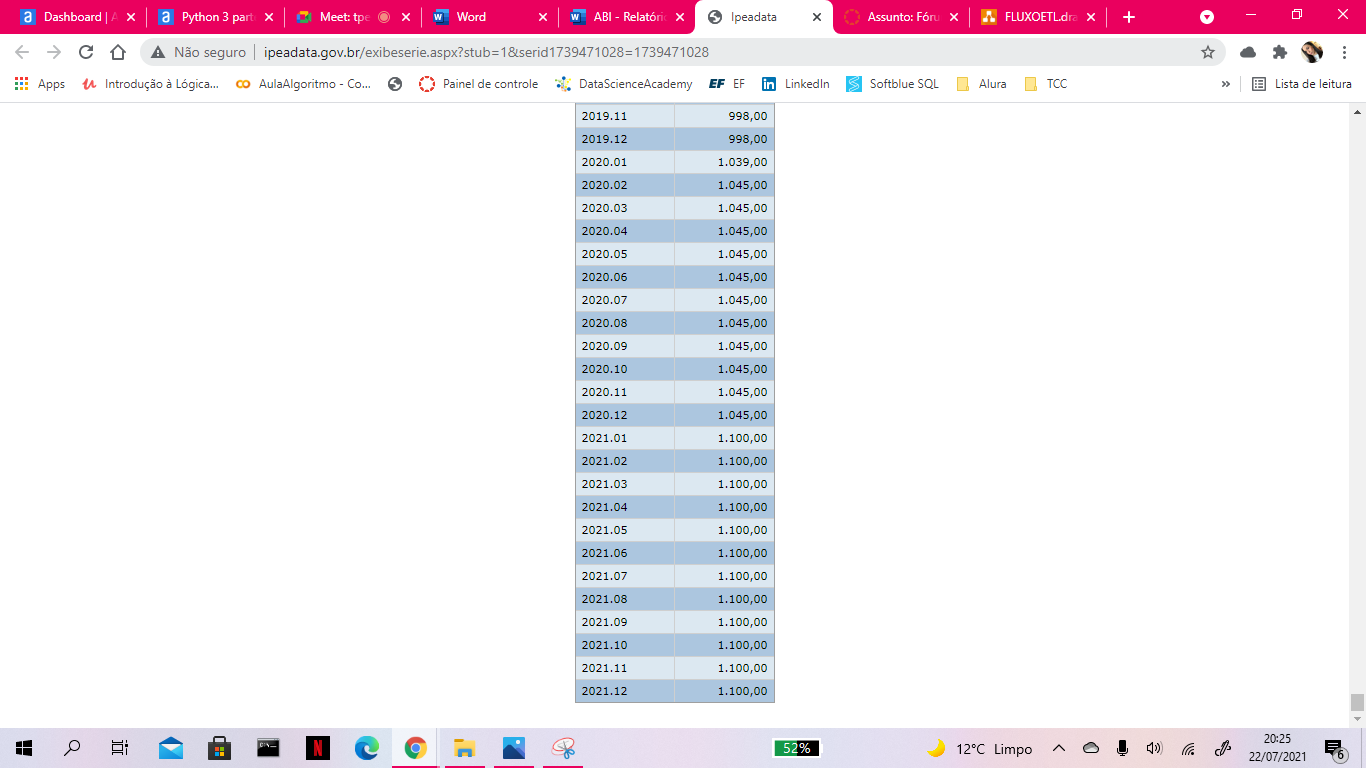
Abaixo print do site onde os arquivos foram extraídos:



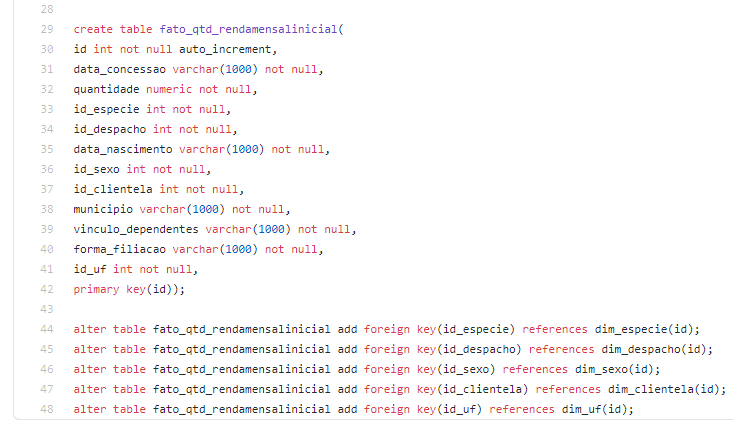
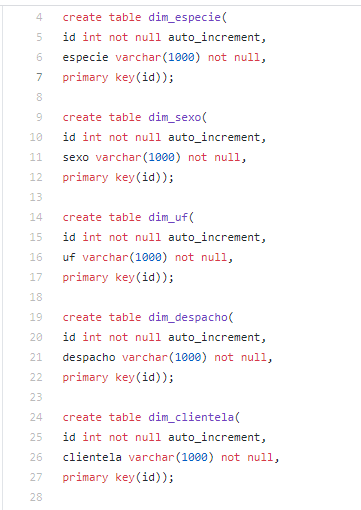
Exemplo do csv :



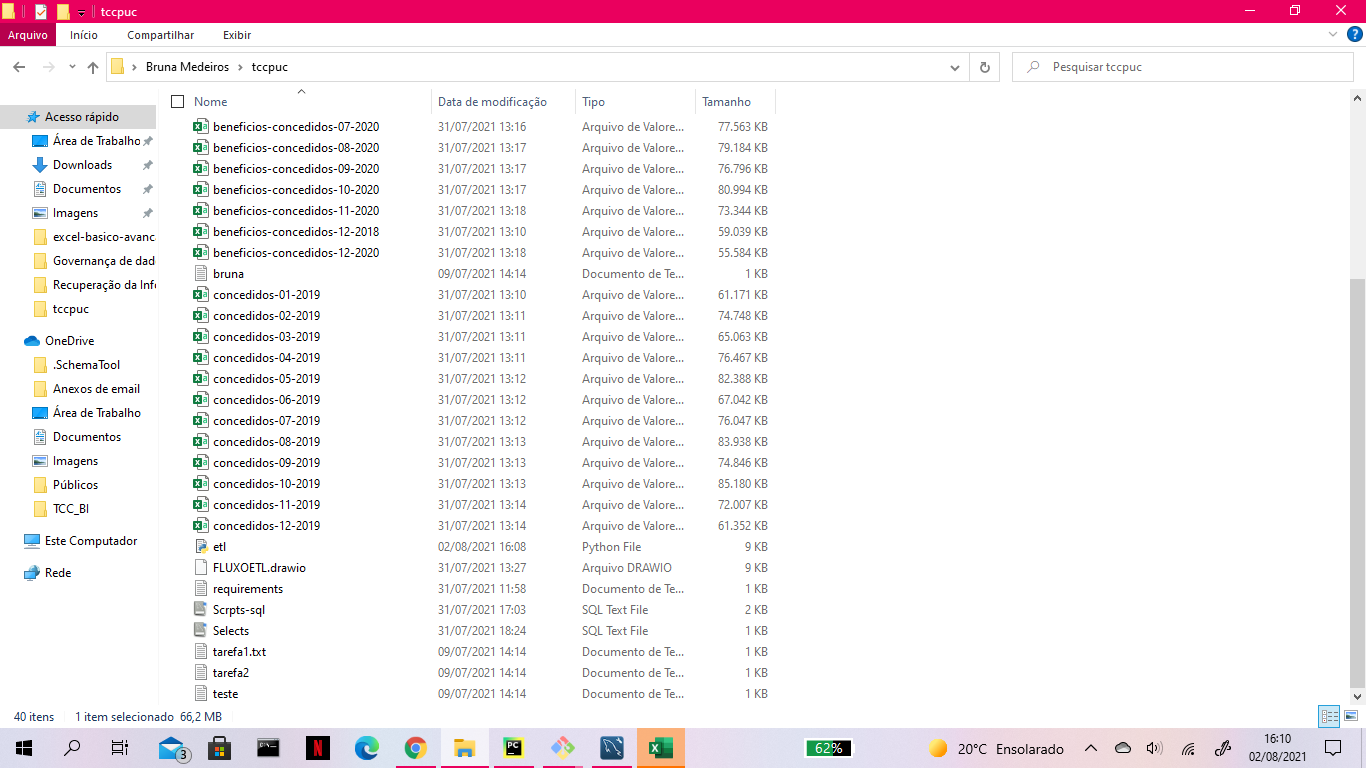
Print do site onde foi extraída a base do salário mínimo:



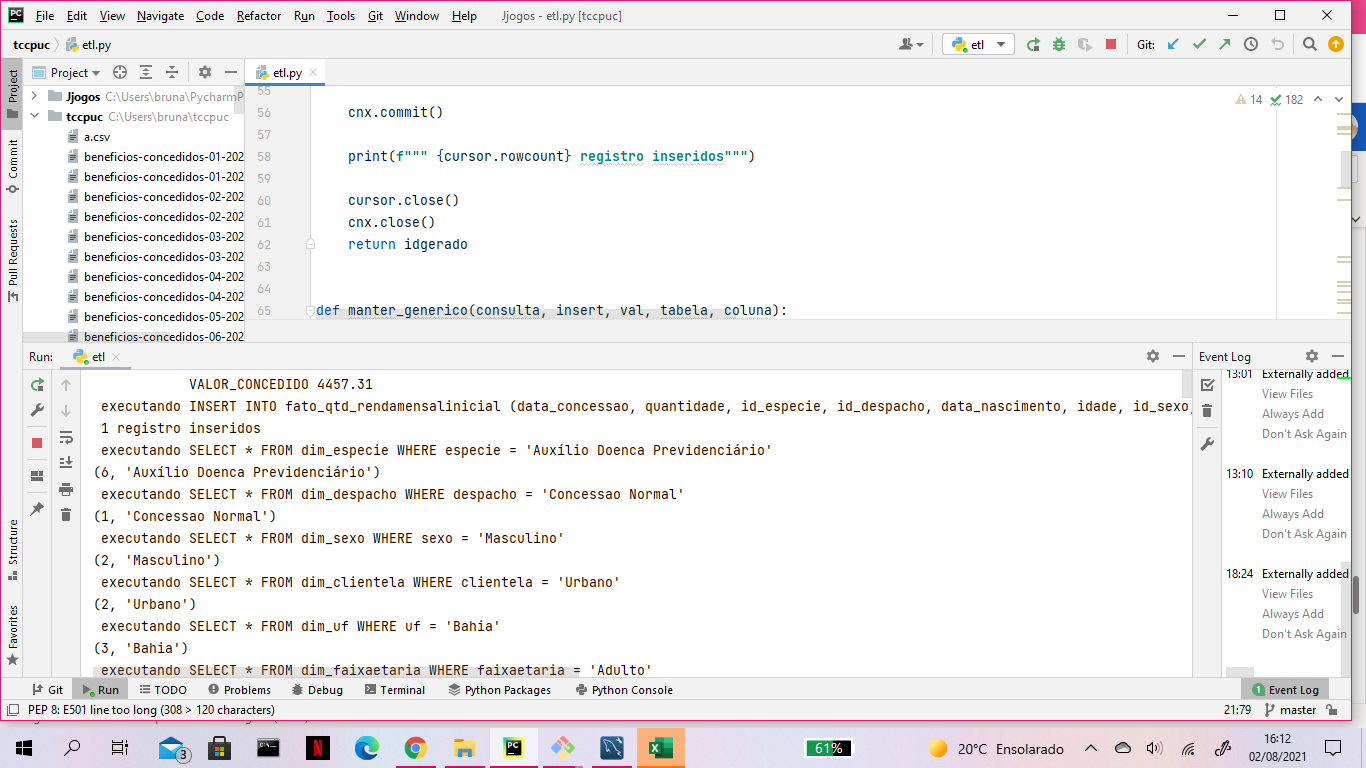
Scripts da criação das tabelas no MySql que faz parte da carga de dados:



Diretório do script etl.py após o download automatizado dos CSVs diretamente do site do governo:

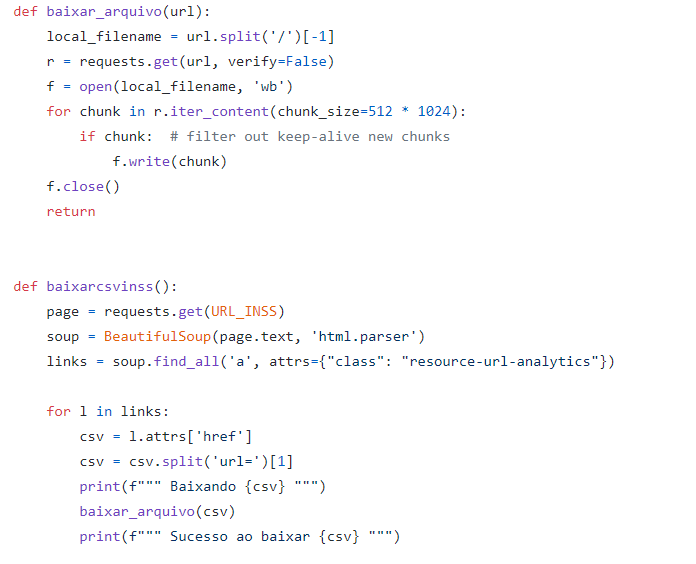


Logs do script de etl.py em execução:

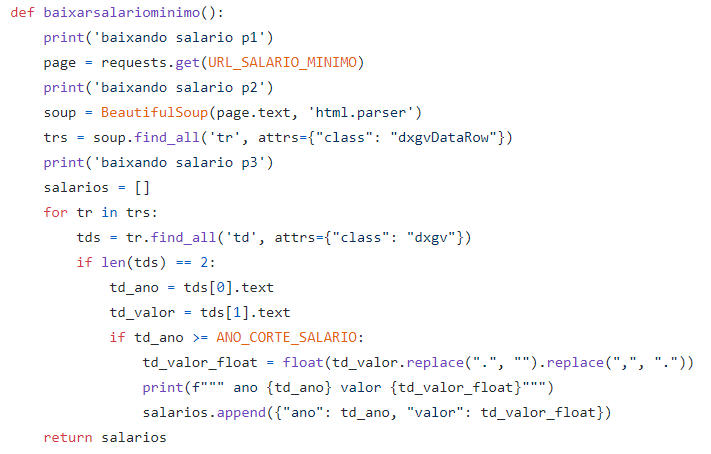


Link do github contendo os códigos fonte <https://github.com/brunamedeiiros/tccpuc> .

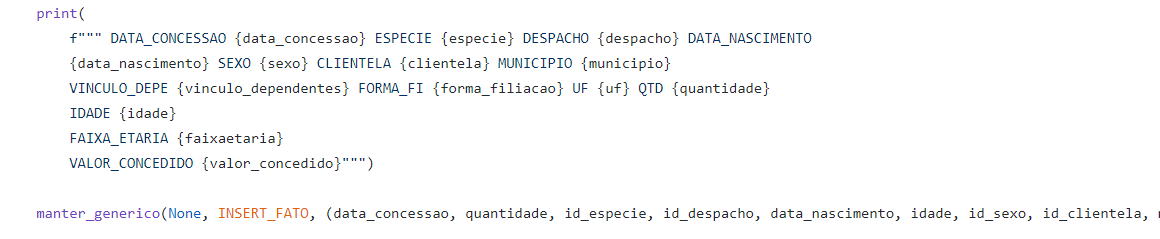
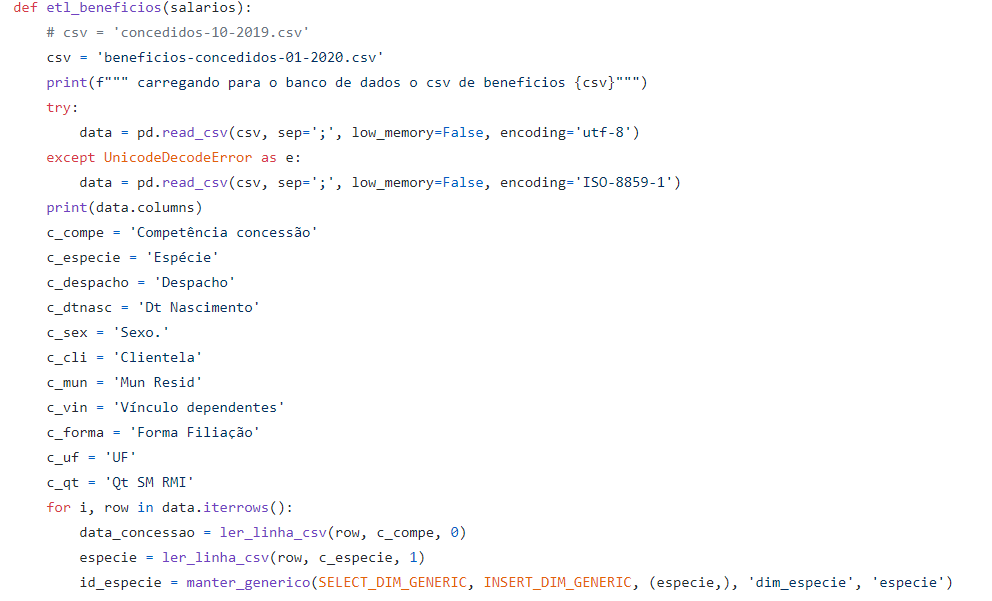
Trecho do código criado para baixar os arquivos do site do INSS para a máquina local. Primeiro é realizado um crawl e com o HTML em memória utilizando a funcionalidade de ler CSS do BeatifulSoup, os links de download são extraídos facilmente e assim os CSVs são baixados pelo método baixar\_arquivo.



Trecho do código que acessa o site do governo do salário mínimo e realiza a raspagem do HTML, e converte essa informação para uma estrutura amigável em Python utilizando a biblioteca BeautifulSoap.



Parte do código contendo o coração da ETL, o método etl\_beneficios recebe a lista de salários e o nome do CSV com isso utiliza a biblioteca Pandas para auxiliar no parse do mesmo. Para cada linha do arquivo carrega as tabelas no banco de dimensões e fato fato\_qtd\_rendamensalinicial no banco de dados MySQL.



Tabelas após a execução do script etl.py :

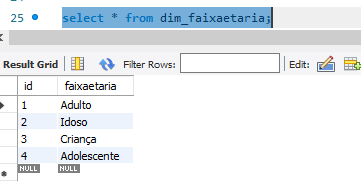
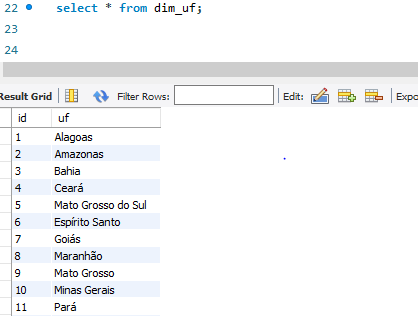
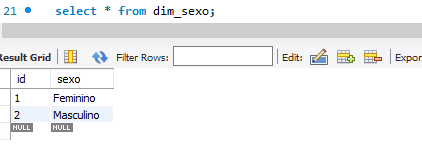
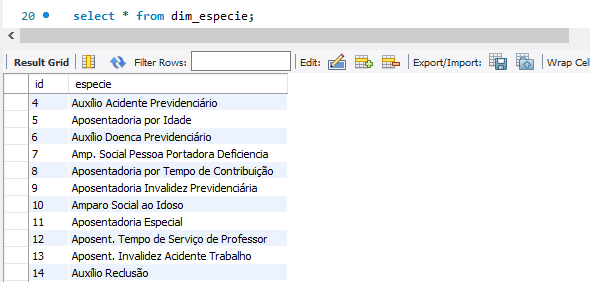
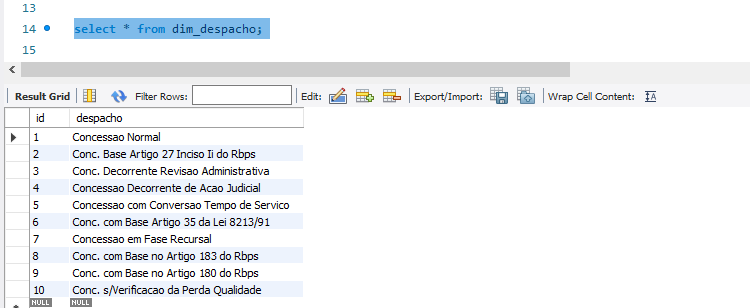
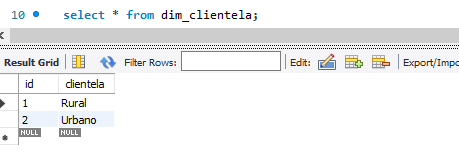
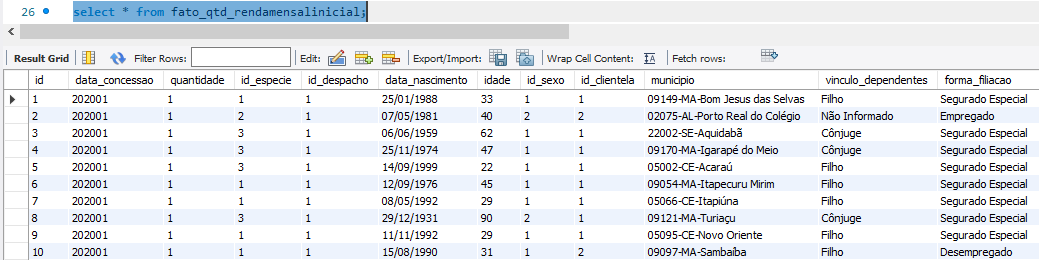


Tabela fato\_qtd\_rendamensalinicial :



# 4. Camada de Apresentação

## 4.1 Dashboard

O Dashboard para análise dos dados coletados foi criado de forma que possibilitasse uma noção de total gasto, por região (clientela), sexo, faixa etária e UF.

As visualizações foram empregadas da melhor forma para a apresentação, usamos cartão para que o valor total ficasse em maior destaque, foi usada a rosca pois são comparações entre dois valores, já o gráfico de barras empilhadas foi usado para o total por faixa etária

Descreva os painéis de controle (dashboard) criados detalhando as visões estratégica, tática e operacional e como são interligadas. Descreva os indicadores e suas fórmulas de cálculo, as métricas, as dimensões de análise, as visualizações empregadas (mapas, gráficos, gauges, entre outros), bem como os filtros aplicáveis à solução, com base nas dimensões utilizadas.

## 4.2 Análises avançadas

Padrões e/ou Tendências: análises de dados criadas a partir da aplicação de técnicas de aprendizado de máquina. As análises podem ser independentes ou incorporadas no dashboard criado anteriormente.

# 5. Registros de Homologação

Testes da solução desenvolvida mostrando que o dado apresentado no dashboard é o mesmo dos sistemas fonte. Estes testes são apresentados, normalmente, por meio de consultas SQL feitas contra as fontes de dados e pelo confrontamento dos resultados com o que é exibido nas visualizações de dados apresentadas nos dashboards

# 5. Conclusões

Este item deve apresentar os seguintes itens: (1) análise crítica apresentando os achados mais relevantes nos dados, feitos a partir do uso do dashboard e das experiências adquiridas no processo de desenvolvimento; (2) proposta de intervenção, possíveis ações a serem tomadas por gestores do contexto analisado no intuito de melhorar o desempenho da organização; e, por fim, (3) as lições aprendidas no processo que possam ajudar nos projetos futuros.

Aponte, ainda, as limitações do trabalho e possíveis pontos de extensão para que outros possam utilizar como ponto de partida em novos projetos.

# 6. Links

Aqui devem ser disponibilizar os links para os artefatos criados (processos de carga, códigos-fonte, painéis, entre outros) e para o vídeo com a apresentação de 5 minutos e para o repositório contendo os códigos fontes ou os artefatos construídos no projeto.

# REFERÊNCIAS

http://www.ipeadata.gov.br/exibeserie.aspx?stub=1&serid1739471028=1739471028

<https://dados.gov.br/dataset/inss-beneficios-concedidos> 22.07.21 20:27

Como um projeto aplicado, as referências, levantadas no processo de revisão bibliográfica, são opcionais e estimulamos que o responsável pelo projeto apresente aquilo que for relevante para complementar o entendimento do trabalho. Dessa forma, relacione-as de acordo com o modelo a seguir.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.