para a primeira etapa, gostaria de explicar que:

1 - os dados disponibilizados pelo rBLH estao disponíveis em relatórios .pdf chamados "rBLH em Dados".

2 - Havia um relatório por ano para os anos 2022, 2021 e 2020, além de um relatório que abrangia dados de 2000 a 2019.

3 - Estavam organizados da seguinte forma: uma tabela por Estado com os dados de coleta relativos aos meses do ano a que se referiam, além de 5 tabelas que totalizavam as regiões, e uma tabela com o total do brasil inteiro.

4 - Decidimos primeiro por um corte de 10 anos, tempo que julgamos suficiente para uma análise abrangente dos dados.

5 - Verificamos que faria sentido pegar apenas as páginas dos pdfs que continham apenas os dados relativos aos estados dos anos de 2013 até 2022, portanto 270 páginas, já que para as regiões e para o geral poderíamos realizar cálculos simples, diminuindo a quantidade de dados carregados.

5 - Verificamos que os pdfs tinham as tabelas como imagens, portanto escolhemos utilizar a ferramenta de OCR da AWS Amazon Textract que nos permitiria analizar até 1000 páginas por mês de forma gratuita.

6 - Percebemos que no ano de 2013 o estado Acre não havia registrado coleta no mês de abril, portanto a tabela daquele estado para aquele ano aparecia com uma linha a menos. Dessa forma, decidimos que colocariamos a linha referente ao mês de abril com os valores zerados para que as tabelas ficassem padronizadas com 13 linhas (uma com títulos e depois uma para cada mês do ano).

7 - Percebemos que teríamos dados de 4 tipos: Texto, Data, Números Inteiros e Números Decimais.

8 - Identificamos os dados que possuíamos e as medidas que desejávamos realizar, conforme o modelo que construímos na etapa anterior do projeto.

para a segunda etapa, gostaria de explicar que:

1 - Selecionamos a ferramenta Power BI, utilizando o Power Query para realizar a extração e transformação de dados, visto que o volume de dados não é tão grande, e que, para as próximas etapas, para criação das visualizações, seria uma ferramenta bastante poderosa.

para a terceira etapa gostaria de explicar o que fizemos para extrair, carregar e transformar os dados, num processo de ELT

1 - Para a a extração, realizamos o download dos 4 relatórios: 2022, 2021 e 2020,e o últimos que abrangia dados de 2000 a 2019

2 - Utilizamos o LibreOffice Draw para selecionar as páginas dos pdfs que continham apenas os dados relativos aos estados dos anos de 2013 até 2022, portanto 270 páginas.

3 - Utilizamos a ferramenta de OCR da AWS Amazon Textract para extrair os dados das imagens dos pdfs.

4 - A ferramenta da Amazon se mostrou muito segura e confiável, entregando um .csv para cada página analizada, portanto 270 arquivos nomeados table-x.csv, onde x era o número da página.

5 - Comparamos os valores nos arquivos .csv com os das imagens e não encontramos erros.

6 - Colocamos a linha referente ao mês de abril para o ano de 2013 no estado Acre, que não havia registrado coleta, com os valores zerados para que as tabelas ficassem padronizadas com 13 linhas (uma com títulos e depois uma para cada mês do ano).

7 - Carregamos os dados para o Power BI, utilizando a opção "Obter Dados" a partir da pasta onde estavam localizados os csvs, utilizando a opção “Combinar e Transformar Dados”:

<colocar imagem 1>

A screenshot of a computer

Description automatically generated

8 – Essa opção inicia o Power Query criando consultas, parâmetros e scripts que buscam combinar os dados:

<colocar imagem 2>

A screenshot of a computer

Description automatically generated

9 – Com essa opção selecionada as transformações que realizamos em um arquivo de exemplo são replicadas aos demais arquivos de forma padronizada e automática:

<colocar imagem 3>

A screenshot of a computer

Description automatically generated

10 – Realizamos as seguintes transformações no aquivo de exemplo, de forma a deixá-lo apenas com os dados que extraímos dos pdfs:

**linguagem M:**

let

Fonte = Csv.Document(Parâmetro1,44,"",ExtraValues.Ignore,65001),

#"Outras Colunas Removidas" = Table.SelectColumns(Fonte,{"Column1"}), // Remover colunas sem dados

#"Primeiras Linhas Mantidas" = Table.FirstN(#"Outras Colunas Removidas",13), // Manter apenas as colunas com dados relevantes

#"Dividir Coluna por Delimitador" = Table.SplitColumn(#"Primeiras Linhas Mantidas", "Column1", Splitter.SplitTextByDelimiter(",'", QuoteStyle.Csv), {"Column1.1", "Column1.2", "Column1.3", "Column1.4", "Column1.5", "Column1.6", "Column1.7", "Column1.8", "Column1.9", "Column1.10", "Column1.11", "Column1.12", "Column1.13", "Column1.14"}), // Dividir as colunas a partir do delimitador ,'

#"Valor Substituído" = Table.ReplaceValue(#"Dividir Coluna por Delimitador","'","",Replacer.ReplaceText,{"Column1.1"}), // Retirar os caracteres ' da primeira coluna

#"Valor Substituído1" = Table.ReplaceValue(#"Valor Substituído",",","",Replacer.ReplaceText,{"Column1.14"}), // Retirar os caracteres , da última coluna

#"Cabeçalhos Promovidos" = Table.PromoteHeaders(#"Valor Substituído1", [PromoteAllScalars=true]) // Promover a primeira linha para se tornar cabeçalho

in

#"Cabeçalhos Promovidos"

<colocar imagem 4>

A screenshot of a computer

Description automatically generated

9 - Carregamos uma tabela com as regiões dos estados no excel (tabela Localidade) para permitir a consulta dos dados por região.

10 – Na consulta final, as transformações iniciais são realizadas automaticamente, utilizando as transformações no arquivo de exemplo:

**linguagem M:**

let

Fonte = Folder.Files("C:\PUC - TBD\Eixo 3\rBLH\Produção\2013-2022 csv"),

#"Arquivos Ocultos Filtrados1" = Table.SelectRows(Fonte, each [Attributes]?[Hidden]? <> true),

#"Invocar Função Personalizada1" = Table.AddColumn(#"Arquivos Ocultos Filtrados1", "Transformar Arquivo", each #"Transformar Arquivo"([Content])),

#"Colunas Renomeadas1" = Table.RenameColumns(#"Invocar Função Personalizada1", {"Name", "Nome da Origem"}),

#"Outras Colunas Removidas1" = Table.SelectColumns(#"Colunas Renomeadas1", {"Nome da Origem", "Transformar Arquivo"}),

#"Coluna de Tabela Expandida1" = Table.ExpandTableColumn(#"Outras Colunas Removidas1", "Transformar Arquivo", Table.ColumnNames(#"Transformar Arquivo"(#"Arquivo de Amostra"))),

<colocar imagem 5>

A screenshot of a computer

Description automatically generated

11 – Os nomes dos arquivos correspondem às páginas do PDF, que é o que utilizaremos para identificar os estados da federação a que correspondem. Para renomear os dados criamos uma função “if”, utilizando a ordem dos 27 estados nos pdfs. Para facilitar a digitação, utilizamos o Excel para realizar a tarefa repetitiva:

<colocar imagem 6>

A screenshot of a computer

Description automatically generated

12 – Realizamos as seguintes transformações na consulta final:

**linguagem M:**

#"Colunas Renomeadas" = Table.RenameColumns(#"Coluna de Tabela Expandida1",{{"Nome da Origem", "Estado"}}), // Renomear coluna com o nome dos arquivos para designar nome dos estados

#"Valor Substituído" = Table.ReplaceValue(#"Colunas Renomeadas", each [Estado], // Para cada linha da coluna "Estado"

each //Para cada linha executar a condicional abaixo substituindo o nome do arquivo pelo nome do estado

if [Estado] = "table-1.csv" then

"Acre"

else if [Estado] = "table-2.csv" then

"Amapá"

else if [Estado] = "table-3.csv" then

"Amazonas"

else if [Estado] = "table-4.csv" then

"Pará"

"Bahia"

(…)

else if [Estado] = "table-269.csv" then

"Rio Grande do Sul"

else if [Estado] = "table-270.csv" then

"Santa Catarina"

else

"",

Replacer.ReplaceText,

{"Estado"}

),

#"Tipo Alterado" = Table.TransformColumnTypes(#"Valor Substituído",{{"Atend. em Grupo", Int64.Type}, {"Atend. Individual", Int64.Type}, {"Visita Domiciliar", Int64.Type}, {"Doadoras", Int64.Type}, {"Receptores", Int64.Type}, {"Produção cadastradas no mês", Int64.Type}, {"Acidez Dornic", Int64.Type}, {"Crematócrito", Int64.Type}, {"Exame microbiológico", Int64.Type}, {"Leite Humano Recebido", type number}, {"Leite Humano Transferido", type number}, {"Leite Humano Distibuído", type number}, {"Leite Humano Coletado", type number}, {"Estado", type text}, {"Mês / Ano", type date}}), // Alterar o tipo dos dados para o tipo correto

#"Mês inserido" = Table.AddColumn(#"Tipo Alterado", "Mês", each Text.BetweenDelimiters(Text.From([#"Mês / Ano"], "pt-BR"), "/", "/"), type text), // Extrair o mês da data, criando uma coluna Mês

#"Ano Inserido" = Table.AddColumn(#"Mês inserido", "Ano", each Date.Year([#"Mês / Ano"]), Int64.Type), // Extrair o ano da data, criando uma coluna Ano

#"Colunas Removidas" = Table.RemoveColumns(#"Ano Inserido",{"Mês / Ano"}), // Remover a coluna Mês / Ano

#"Consultas Mescladas" = Table.NestedJoin(#"Colunas Removidas", {"Estado"}, Localidade, {"Estado"}, "Localidade", JoinKind.LeftOuter), // Mesclar a consulta Localidade, que apresenta as regiões de cada estado

#"Localidade Expandido" = Table.ExpandTableColumn(#"Consultas Mescladas", "Localidade", {"Região"}, {"Região"}), // Incluir a coluna Região correspondente a cada Estado

#"Colunas Reordenadas" = Table.ReorderColumns(#"Localidade Expandido",{"Estado", "Região", "Mês", "Ano", "Atend. em Grupo", "Atend. Individual", "Visita Domiciliar", "Doadoras", "Receptores", "Leite Humano Coletado", "Leite Humano Distibuído", "Leite Humano Transferido", "Leite Humano Recebido", "Exame microbiológico", "Crematócrito", "Acidez Dornic", "Produção cadastradas no mês"}) // Reordenar colunas de forma mais lógica

in

#"Colunas Reordenadas"

<colocar imagem 7>

A screenshot of a computer

Description automatically generated

13 – Aplicamos as transformações realizadas, carregando as informações da tabela Dados para utilizarmos no Power BI:

<colocar imagem 8>

A screenshot of a computer

Description automatically generated

14 – Criamos as medidas utilizando a ferramenta “Medida Rápida”:

- Entradas (Leite Humano Coletado + Leite Humano Recebido)

- Saídas (Leite Humano Distribuído + Leite Humano Transferido)

- Produção líquida local (Leite Humano Coletado - Leite Humano Distribuído): não leva em conta as transferências entre bancos de leite.

- Produção líquida total (Entradas - Saídas)

- Média de leite coletado por doadoras (Leite Humano Coletado / Doadoras)

- Média de leite distribuído por receptores (Leite Humano Distribuído / Receptores)

- Média de leite disponível por receptores (Entradas / Receptores)

- Razão entre Doadoras e Receptores (Doadores / Receptores)

<colocar imagem 9>

A screenshot of a computer

Description automatically generated

Dessa forma, os dados ficaram organizados e prontos para gerar visualizações.