Equipe 6 - NutriSquad

Desafio escolhido: "**Qual o impacto do Suplemento Nutricional na prevenção de quedas na população idosa?" (Desafio 2)**

No contexto do desafio selecionada pela equipe, o primeiro passo para respondermos à questão é detectar e isolar as variáveis e desfechos de interesse. Portanto, trabalhamos nas bases de dados disponibilizadas em busca de ferramentas que nos auxiliasse na resposta ao desafio proposto.

O desfecho "quedas" foi selecionado dentro da base de dados da CIHA (Comunicação de Informação Hospitalar e Ambulatorial), parte do DATASUS, com a filtragem de CID (Classificação internacional de doenças) considerando desfecho quedas (W10). Outros desfechos secundários podem ser filtrados e inseridos durante a análise de dados, que estejam diretamente associados com quedas.

Para avaliar a qualidade da alimentação do ponto de vista nutricional, a base de dados da Pesquisa de Orçamentos Familiares (POF 2017/2018) foi utilizada. O tratamento inicial de ambas as bases de dados para cruzamento futuro (análises seguintes) se deu conforme segue abaixo:

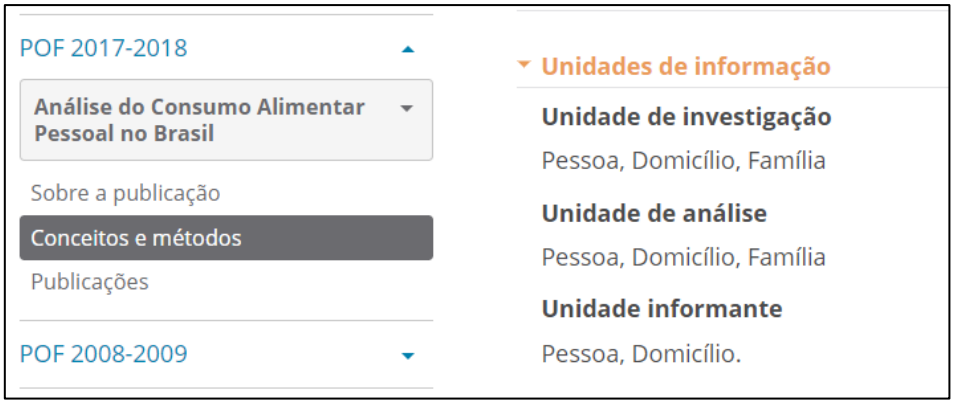
TRATAMENTO DATASET – POF 2017/2018

Bases selecionadas com variáveis de interesse:

- **CONSUMO ALIMENTAR**

Base contendo 1.175.390 observações, referentes a listagem de alimentos consumidos pelo entrevistado. Dentre as variáveis de interesse, todas as informações referentes ao consumo de macronutrientes e micronutrientes, como proteínas, energia, vitaminas e minerais foram mantidos.

Para a obtenção da informação de consumo a nível individual (total por indivíduo e não por alimento) foi realizada a soma dos alimentos consumidos por cada indivíduo. A unidade de informação para a amostra de consumo alimentar na POF considera até o nível indivíduo (ou unidade informante – pessoa), conforme exposto na metodologia de amostragem:



Fonte: https://www.ibge.gov.br/estatisticas/sociais/saude/24786-pesquisa-de-orcamentos-familiares-2.html?edicao=28523&t=conceitos-e-metodos

Portanto, foi gerado um **número identificador** (id) de cada indivíduo, considerando até a menor unidade de informação da amostra (pessoa, dentro da unidade informante ou código do informante): "COD_UPA + NUM_DOM + NUM_UC + COD_INFORMANTE"

Após a geração do ID, os nutrientes foram agregados (função 'collapse') por tal ID, gerando, portanto, a informação nutricional de consumo total do indivíduo, considerando a soma de todos os alimentos reportados. Após a agregação, o novo 'Banco_Consumo', a nível individual contém 46.164 observações ou indivíduos entrevistados:

```
. sum energia_kcal, d

                    (rawsum) energia_kcal
-------------------------------------------------------------
      Percentiles      Smallest
 1%        705.08         254.83
 5%        1182.8         292.45
10%       1511.48         301.72       Obs                46,164
25%       2187.85         310.35       Sum of Wgt.        46,164

50%        3036.5                      Mean               3228.229
                          Largest       Std. Dev.          1505.222
75%       4026.855        17163.95
90%       5126.355         17274.6     Variance           2265693
95%        5905.06        17626.86     Skewness           1.216972
99%        7924.17        17877.43     Kurtosis           6.809019
```

Considerando a publicação oficial do IBGE com os resultados gerais da avaliação do consumo, temos a informação do total de indivíduos entrevistados com dados de consumo alimentar, validando, portanto, a conversão inicial e limpeza do *dataset* de origem para a unidade de interesse.

Tabela 1 - Unidades primárias de amostragem selecionadas, domicílios entrevistados na amostra e com consumo alimentar pessoal e pessoas na subamostra, segundo as Unidades da Federação - período 2017-2018

Unidades da Federação	Unidades primárias de amostragem selecionadas	Domicílios entrevistados na amostra		Pessoas na subamostra
		Total	Consumo alimentar pessoal (subamostra)	
Brasil	**5 504**	**57 920**	**20 112**	**46 164**
Rondônia	93	951	332	772
Acre	94	950	344	770
Amazonas	181	1 833	615	1 614
Roraima	78	765	240	623

Fonte: IBGE, 2020

Outras bases de dados contêm variáveis de interesse para comparação com o desfecho, sendo também filtradas para a construção do *dataset* de trabalho. As demais bases selecionadas e, consequentemente, as variáveis de interesse filtradas nessas bases compreendem:

- **CARACTERÍSTICAS DA DIETA**

  - Peso referido
  - Altura referida

Com essas informações podemos trabalhar para obter o índice de massa corporal (IMC) do indivíduo, como um indicador do estado nutricional do mesmo.

Coleta obtida juntamente com o módulo de consumo alimentar, portanto, a base contém automaticamente, os dados dos 46.164 indivíduos com dados de consumo.

- **MORADOR**

Na base de dados de morador, podemos filtrar e selecionar algumas variáveis socioeconômicas de interesse, como:

  - UF do domicílio
  - Estrato do domicílio (zona rural, urbana, capital ou região metropolitana)
  - Situação do domicílio (rural ou urbano)

- Idade
- Sexo
- Raça
- Anos de estudo

- **DOMICÍLIO**

Dentro da base de dados do domicílio, foram selecionadas algumas variáveis que são apontadas na literatura como associadas com o risco de quedas em idosos, como o tipo de habitação, por exemplo, apartamentos com escadas ao invés de casa. Outra variável para teste foi o tipo de piso da habitação, que pode aumentar ou não o risco de queda em idosos com fragilidade.

- Tipo do domicílio – casa, apartamento, outras habitações
- Material de revestimento do piso (zona rural, urbana, capital ou região metropolitana)

As bases iniciais de morador e domicílio possuem dados de 178.431 entrevistados e 57.920 domicílios, respectivamente. Como primeiro filtro, as informações de interesse são aquelas apenas dos indivíduos que possuem dados de consumo alimentar. Portanto, o primeiro passo dentro dessas duas bases (morador e domicílio) é a seleção de IDs referentes aos indivíduos da amostra da base de consumo.

Base mestre: Consumo alimentar

1º passo – Juntar as duas variáveis

```
merge two datasets by observations
```

2º passo – Excluir as observações que não estão presentes nas duas bases

```
drop if merge =!= 3
```

No banco ‘mestre’ temos a junção das variáveis selecionadas em cada uma das bases acima, filtradas apenas para os indivíduos com dados de consumo alimentar.

Considerando que o interesse do desafio é, exclusivamente, na pessoa idosa, o próximo filtro aplicado já na base mestre é o filtro de faixa etária, para considerar apenas indivíduos a partir dos 60 anos de idade.

```
drop if v0403 < 60
```
(Exclusão de observações quando a variável idade, em anos, é menor que 60)

```
. sum v0403 , d

                            V0403
-------------------------------------------------------------
      Percentiles      Smallest
 1%           0              0
 5%           3              0
10%           6              0       Obs                46,164
25%          14              0       Sum of Wgt.        46,164

50%          29                      Mean              31.45113
                        Largest       Std. Dev.         20.82674
75%          46            102
90%          61            103       Variance           433.753
95%          70            105       Skewness          .5242526
99%          83            108       Kurtosis          2.502667

. drop if v0403 < 60
(40,865 observations deleted)
```

Um total de 40.865 observações excluídas, dado que foram indivíduos com coleta de dados de consumo alimentar, mas fora da faixa etária de idosos.

AMOSTRA FINAL = 5.299 INDIVÍDUOS IDOSOS, DE AMBOS OS SEXOS, COM DADOS DE CONSUMO ALIMENTAR

- NOVO DICIONÁRIO DE VARIÁVEIS ALTERADAS NO BANCO FINAL

v0403 – idade

v0404 – sexo

v0405 – raca

v72c01 – peso em kg

v72C02 – altura em cm

v0201 – tipo_domicilio

v0204 – tipo_piso

Comunicação de Informação Hospitalar e Ambulatorial (CIHA)

A CIHA surgiu da necessidade de incluir, no Sistema CIH, a possibilidade de registro dos atendimentos ambulatoriais, não informados no Sistema de Informações Ambulatoriais do SUS (SIA/SUS).

O banco está no formato .db, tem cerca de 2,3 GB compactado e 20 GB expandido.

Origem dos dados:

http://repositorio-datathon.distrito.me/datathon/dados_exportados/datasus_ciha/bd_convertido_ciha_20201201_CO.db.zip

O repositório original é composto por 26 bases referentes aos estados brasileiros (O estado de Roraima não está presenta na base original, por isso temos apenas 26). A primeira etapa do tratamento foi converter o formato em que os arquivos se encontravam (.db) para um que fosse de mais fácil manipulação (.csv).

Além da conversão, também foi feito um filtro na coluna do ano de competência, selecionando as informações referentes ao anos de 2017 e 2018, para ser compatível com as demais bases a serem utilizadas.

Pelo fato do tema do desafio abordar o público idoso, também se fez necessário mais um filtro, onde apenas os indivíduos com 60 anos ou mais seriam levados em consideração.

O script abaixo, feito em Python, é um exemplo do tratamento feito na tabela do estado de São Paulo para analisar apenas pessoas com idade a partir de 60 anos. Este procedimento foi repetido para os demais estados.

df_sp = df_sp.loc[df_sp['IDADE'] > 59]

Após isso, as bases geradas foram concatenadas através do seguinte comando em Python:

df_brasil = pd.concat([df_ac, df_al, df_am, … , df_sp, df_to])

Com a nova base criada e tratada, um novo arquivo .csv foi gerado, por meio do código feito em Python:

df_brasil.to_csv('BRASIL.csv')

O arquivo gerado, com os dados tratados, tem 1,5 GB expandido.

- ## Próximos passos:

Com as bases limpas e selecionadas pelos filtros iniciais, o próximo passo é definir a granularidade das análises (nível de análise). Uma vez que a base original da POF a granularidade de localização geográfica não especifica municípios, mas sim estratos, uma possibilidade é a utilização desses estratos como ponto de corte na base da CIHA. Outra possibilidade a ser testada é a seleção de apenas uma região ou estado. Após a junção das bases, essas decisões serão avaliadas do ponto de vista de distribuição dos dados, presença dos eventos de interesse (não escolher um nível com um "n" muito pequeno), dentre outros.