Curadoria dos Dados - Data Dream

A seguinte curação tem como objetivo selecionar e organizar os principais dados que constituirão na Sprint 4 o Data Lake e Data Warehouse da solução. A curadoria será feita levando em consideração os três critérios destacados pelo cliente: categoria, canal e região. O tamanho máximo esperado do conjunto de dados será de 15 gb.

Códigos Postais (zipcode-datadream) (< 1MB):

Arquivos:

BR.csv BR.txt

| | BR | 69945-000 | Column 1 | Acrelândia | Acre | Column 2 | 01 | 1200013 - | |
|----|---------|--------------|----------|------------|-----------------|----------|---------|-----------|---------|
| 1 | 9.9805 | -66.8439 | 2 | | | | | | |
| 2 | BR | 69935-000 | | | Assis Brasil | Acre | | 01 | 1200054 |
| 3 | | -10.8833 | | -70.0131 | 2 | | | | |
| 4 | BR | 69932-000 Br | | | 01 | | 1200104 | - | |
| 5 | 10.7677 | -69.0114 | 2 | | | | | | |
| 6 | BR | 69923-000 | | Bujari | Acre | | 01 | 1200138 - | |
| 7 | 9.5786 | -68.172 | 2 | | | | | | |
| 8 | BR | 69922-000 | | Capixaba | Acre | | 01 | 1200179 - | |
| 9 | 10.4878 | -67.8483 | 2 | | | | | | |
| 10 | BR | 69980-000 | | | Cruzeiro do Sul | Acre | | 01 | 1200203 |
| 11 | | -8.0159 | | -72.9298 | 2 | | | | |
| 12 | BR | 69934-000 | | | Epitaciolândia | Acre | | 01 | 1200252 |
| 13 | | -10.9354 | | -68.4441 | 2 | | | | |
| 14 | BR | 69960-000 | | Feijó Acre | | 01 | | 1200302 | -8.905 |
| 15 | | -70.9149 | 2 | | | | | | |
| 16 | BR | 69975-000 | | Jordão | Acre | | 01 | 1200328 - | |
| 17 | 9.0917 | -71.8407 | 2 | | | | | | |

Nenhum dos arquivos acima será útil. Se tratando apenas de um conjunto de endereços postais brasileiro que detalham critérios muito específicos de geolocalização, não se referindo a estabelecimentos de venda de alimentos. Além da inutilização dos critérios oferecidos, a tabela possui

uma organização atípica, sendo inviável para a condição atual do projeto realizar um script com tamanha complexidade.

CNPJs (cnpj-datadream) (5,4 GB):

Arquivos:

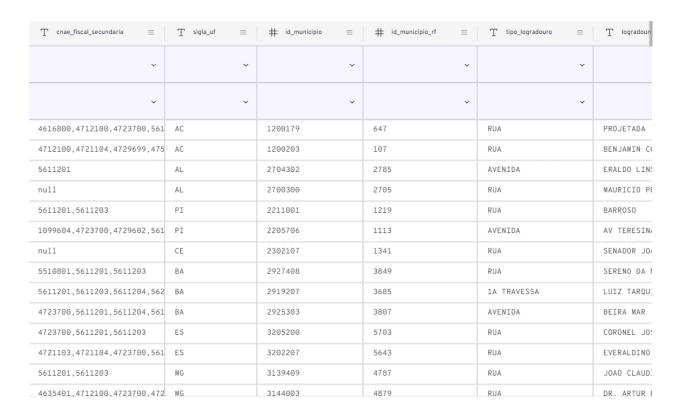
cnpjs_1.csv

cnpjs_2.csv

cnpjs_3.csv

cnpjs_4.csv

cnpjs_5.csv



Os arquivos referentes ao CPNJs são extremamente úteis e cruciais para a solução. As cinco tabelas se referem aos estabelecimentos alimentícios distribuídos por todo o Brasil, caracterizando o tipo do estabelecimento

(mercado, café, etc) e sua exata localização (estado, cidade, bairro, número). A organização da base está boa, com fácil manipulação. O tamanho do arquivo é grande, mas justifica-se pelo grande detalhamento necessário para esse aspecto da solução.

Os arquivos de CNPJ contemplam e satisfazem a identificação dos canais e regiões, sendo suficientes para uma análise sólida.

Receita Federal (receita-datadream) (18,3 MB)

Arquivos:

distribuicao-renda-socios-exclusiva.csv distribuicao-renda-socios.csv distribuicao-renda.csv

| $=$ T Ente Federativo \equiv | # Centil ≡ | $\#$ Quantidade de Contribuintes \equiv | T Rendimentos Tributaveis - Limite Superior da RB2 do Centil [RS milhões] |
|--------------------------------|------------|-------------------------------------------|---------------------------------------------------------------------------|
| ТО | 15 | 1.118 | 4.470,00 |
| ТО | 16 | 1.119 | 4.470,00 |
| ТО | 17 | 1.119 | 4.602,00 |
| ТО | 18 | 1.118 | 5.281,86 |
| ТО | 19 | 1.119 | 6.000,00 |
| ТО | 20 | 1.119 | 6.680,00 |
| ТО | 21 | 1.119 | 7.439,00 |
| ТО | 22 | 1.118 | 8.271,28 |
| ТО | 23 | 1.119 | 8.940,00 |
| ТО | 24 | 1.119 | 9.600,00 |
| ТО | 25 | 1.118 | 10.140,00 |
| ТО | 26 | 1.119 | 10.778,00 |
| ТО | 27 | 1.119 | 11.295,80 |
| ТО | 28 | 1.119 | 11.883,00 |
| ТО | 29 | 1.118 | 12.000,00 |
| ТО | 30 | 1.119 | 12.390,00 |
| ТО | 31 | 1.119 | 12.886,35 |

Os seguintes arquivos se referem à distribuição de impostos conforme renda e bens. Os arquivos possuem três problemas principais que fazem com que eles não sejam ideais para futuras análises. O primeiro problema é o fato deles limitarem a coleta de dados a nível estadual, não permitindo a visão minuciosa de cidades ou bairros. O segundo problema é a alta complexidade do entendimento dos dados, muitos colunas se referem a impostos extremamente específicos, sendo necessário fazer um cálculo entre mais de 20 colunas diferentes para reunir um dado utilizável (possível renda local), o que leva ao terceiro problema: alto risco de dados imprecisos ou com erros, prejudicando a qualidade dos infográficos.

Pesquisa de Orçamento Familiar (pofmain-datadream) (900 MB)

Arquivos:

aluguel estimado.csv caderneta coletiva.csv caracteristicas dieta.csv condicoes vida.csv consumo alimentar.csv despesa coletiva.csv despesa_individual.csv domicilio.csv inventario.csv morador.csv outros.csv quali vida.csv rendimento.csv restricao.csv servico pof2.csv servico pof4.csv

Por ser uma das bases principais, disponibilizada pelo parceiro bem no início do projeto, será evitado tirar qualquer tabela pertencente a esse conjunto, ocupando, de qualquer forma, um pequeno tamanho na memória.

IBGE (ibge-datadream) (0,99 MB)

Arquivos:

gini_geral_s3.csv gini_industria_s3.csv pib_s3.csv

| File Edit Inse | ert Format Dat | a Help | | gini_geral_s | 3.csv ② | Q+ ! |
|-------------------|----------------|------------------------|------------------|--------------------|----------|--------|
| ☐ Reset Sheet | ₩ Hide Columns | ∀ Filter ⊞ Group | ↓↑ Sort ⊗ Enrich | ments 📶 Chart Data | | |
| T A = | Т в ≡ | T c ≡ | T D = | T ε = | T F ≡ | T G |
| Unidade da Federa | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
| Rondônia | 0;663476 | 0;642692 | 0;655898 | 0;661969 | 0;664114 | 0;6680 |
| Acre | 0;733505 | 0;720977 | 0;735187 | 0;733279 | 0;732895 | 0;7260 |
| Amazonas | 0;891846 | 0;898752 | 0;907582 | 0;906511 | 0;904798 | 0;9054 |
| Roraima | 0;754187 | 0;751256 | 0;745649 | 0;746504 | 0;757404 | 0;7490 |
| Pará | 0;728359 | 0;725016 | 0;735970 | 0;731493 | 0;737730 | 0;7350 |
| Amapá | 0;802527 | 0;787800 | 0;795061 | 0;782708 | 0;791405 | 0;7778 |
| Tocantins | 0;709031 | 0;689553 | 0;692404 | 0;694704 | 0;704422 | 0;7021 |
| Maranhão | 0;726878 | 0;726222 | 0;733106 | 0;747359 | 0;754244 | 0;7406 |
| Piauí | 0;769564 | 0;757789 | 0;758241 | 0;767973 | 0;784127 | 0;7612 |
| Ceará | 0;786772 | 0;782014 | 0;792612 | 0;792617 | 0;786051 | 0;7942 |
| Rio Grande do Nor | 0;798872 | 0;795312 | 0;799420 | 0;803622 | 0;804342 | 0;8041 |
| Paraíba | 0;777902 | 0;774334 | 0;770663 | 0;768940 | 0;774127 | 0;7783 |
| Pernambuco | 0;795667 | 0;791996 | 0;797682 | 0;793659 | 0;786755 | 0;7923 |

| File Edit Ins | ert Format Dat | a Help | | gini_industria | _s3.csv ② | Q+ SI |
|-------------------|----------------|-----------------------|------------------|--------------------|-----------|---------|
| C⁴ Reset Sheet | ₩ Hide Columns | ∀ Filter ⊞ Group | ↓↑ Sort ⊗ Enrich | ments 📶 Chart Data | | |
| T A = | Т в ≡ | T c ≡ | T D = | T ε ≡ | T F ≡ | T G |
| Unidade da Federa | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
| Rondônia | 0;790535 | 0;768464 | 0;774926 | 0;776917 | 0;788858 | 0;79028 |
| Acre | 0;839545 | 0;837409 | 0;846886 | 0;868354 | 0;865321 | 0;86820 |
| Amazonas | 0;970171 | 0;972946 | 0;973139 | 0;975159 | 0;975062 | 0;97306 |
| Roraima | 0;859175 | 0;881024 | 0;871617 | 0;860453 | 0;853717 | 0;87285 |
| Pará | 0;872753 | 0;873611 | 0;867314 | 0;870365 | 0;865593 | 0;85799 |
| Amapá | 0;827082 | 0;789789 | 0;807837 | 0;802792 | 0;816648 | 0;78933 |
| Tocantins | 0;868443 | 0;854813 | 0;856579 | 0;844340 | 0;847100 | 0;85992 |
| Maranhão | 0;898132 | 0;902194 | 0;912669 | 0;919352 | 0;905834 | 0;91526 |
| Piauí | 0;935639 | 0;955225 | 0;935620 | 0;951367 | 0;937481 | 0;92962 |
| Ceará | 0;894928 | 0;902645 | 0;904652 | 0;903154 | 0;898827 | 0;90026 |
| Rio Grande do Nor | 0;894844 | 0;898712 | 0;902302 | 0;904250 | 0;903543 | 0;90317 |
| Paraíba | 0;928720 | 0;927039 | 0;914130 | 0;926048 | 0;920529 | 0;92178 |
| Pernambuco | 0;912906 | 0;914245 | 0;912312 | 0;915161 | 0;901490 | 0;90625 |

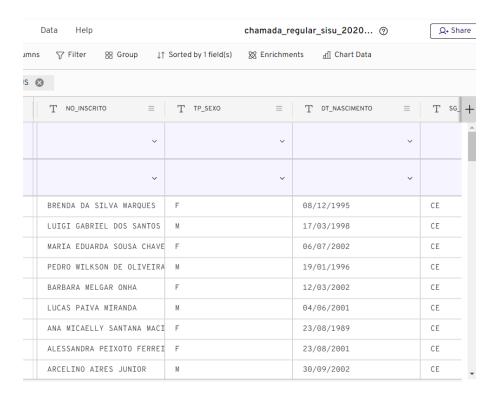
| T A = | # в ≡ | # c ≡ | # D = | # E ≡ | # F ≡ | # G |
|-------------------|--------|--------|--------|---------|---------|---------|
| Município | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
| Alta Floresta D'C | 111291 | 143222 | 173991 | 167127 | 168805 | 191364 |
| Ariquemes (RO) | 449593 | 539636 | 657193 | 749021 | 790697 | 905203 |
| Cabixi (RO) | 31768 | 40985 | 43392 | 49130 | 46884 | 49166 |
| Cacoal (RO) | 474443 | 622437 | 622415 | 758960 | 743194 | 814890 |
| Cerejeiras (RO) | 79174 | 99983 | 121366 | 129107 | 124415 | 143270 |
| Colorado do Oeste | 87254 | 103363 | 114815 | 126534 | 126670 | 137899 |
| Corumbiara (RO) | 45165 | 57284 | 67309 | 70936 | 68935 | 72291 |
| Costa Marques (RC | 37308 | 50996 | 53905 | 62986 | 60812 | 74215 |
| Espigão D'Oeste (| 119312 | 153425 | 180676 | 203257 | 195271 | 220452 |
| Guajará-Mirim (RC | 174680 | 247248 | 302287 | 327643 | 325665 | 374116 |
| Jaru (RO) | 306158 | 385507 | 420126 | 569682 | 614008 | 693762 |
| Ji-Paraná (RO) | 657400 | 774035 | 933474 | 1066022 | 1110409 | 1257350 |
| Machadinho D'Oest | 83828 | 129476 | 156743 | 186678 | 167421 | 183666 |

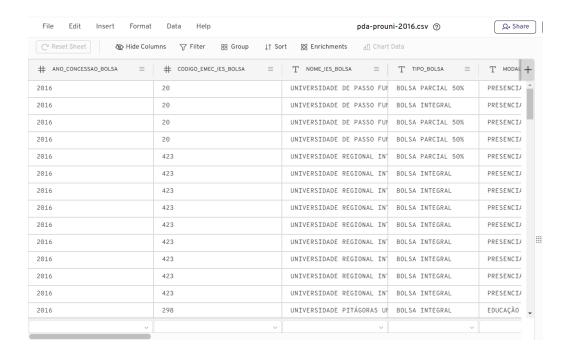
As tabelas gini_geral e gini_industria são generalistas a nível de região, detalhando apenas a nível estadual. Essas informações não servem diretamente para a atual solução que deseja identificar especificamente cidades e bairros com potencial mercado. Contudo, a tabela gini_geral ainda pode ser útil para análises e visões mais macros. Já a tabela pib_s3.csv, ela mostra o PIB por cidade, possibilitando uma análise geral um pouco menos macro.

MEC (dadosmec-datadream) (7,29 GB)

Arquivos:

chamada_regular_sisu_2020_2.csv chamada_regular_sisu_2021_2.csv chamada_regular_sisu_2022_2.csv pda-prouni-2016.csv pda-prouni-2017.csv pda-prouni-2018.csv pda-prouni-2019.csv ProuniRelatorioDadosAbertos2020.csv





O conjunto de dados referente ao MEC é completamente inviável para o contexto do projeto. A intenção era inferir a capacidade de consumo a partir do acesso à cursos superiores, contudo, o tema da educação é extremamente complexo. Essa métrica é afetada pelo sistema de cotas que em certa medida equaliza esse acesso educacional, normalizando a distribuição. Por esse e vários outros motivos relacionados à complexidade do tema educação, não será usado nenhum dado referente ao MEC na atual solução.

INEP (dadosinep-datadream) (1,04 gb)

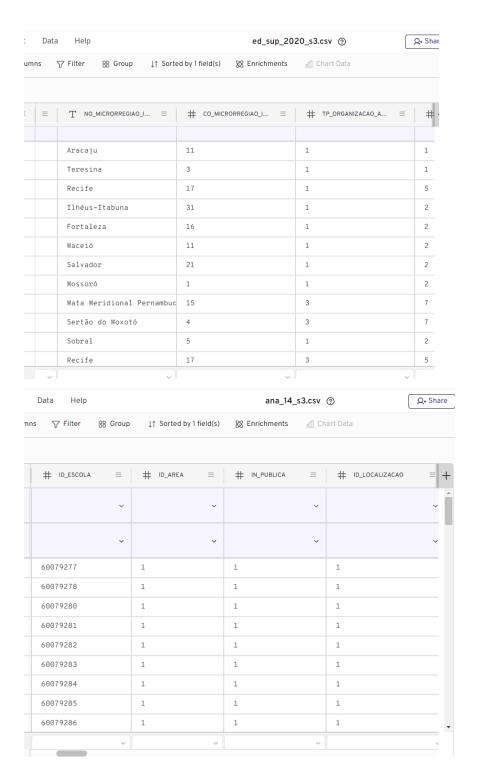
Arquivos:

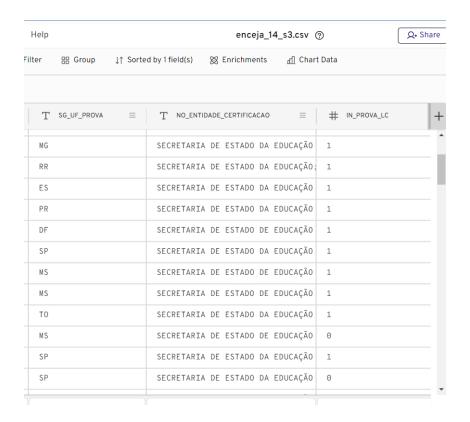
```
ana_14_s3.csv
ana_16_s3.csv
enceja_14_s3.csv
enceja_20_s3.csv
enceja_22_s3.csv
```

ed_sup_2009_s3.csv

. . .

ed_sup_2022_s3.csv





O conjunto de dados referente ao INEP não é aplicável pela complexidade do tema educacional, fugindo do escopo do projeto. Além disso, os dados educacionais indicam padrões de apenas uma parcela da população, a mais jovem. Assim, os dados do INEP não serão utilizados na solução final.

SUS (datasus-datadream) (14,5 gb)

Arquivos:

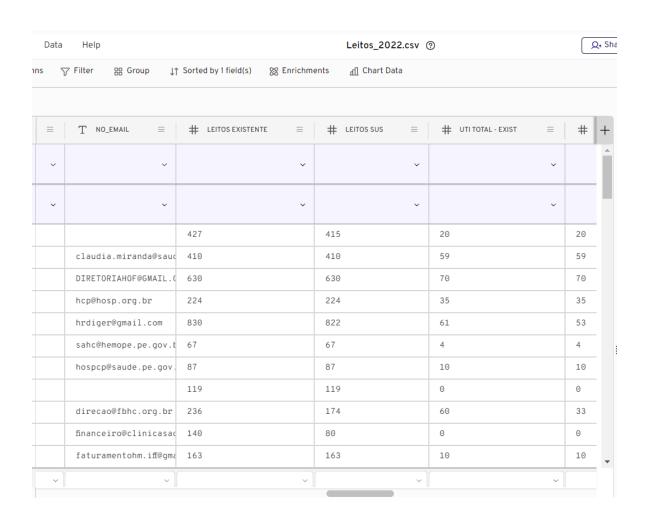
INFLUD21-01-05-2023.csv INFLUD22-03-04-2023.csv INFLUD23-16-10-2023.csv

esus-vepi.LeitoOcupacao_2020.csv esus-vepi.LeitoOcupacao_2021.csv esus-vepi.LeitoOcupacao_2022.csv Leitos_2007.csv

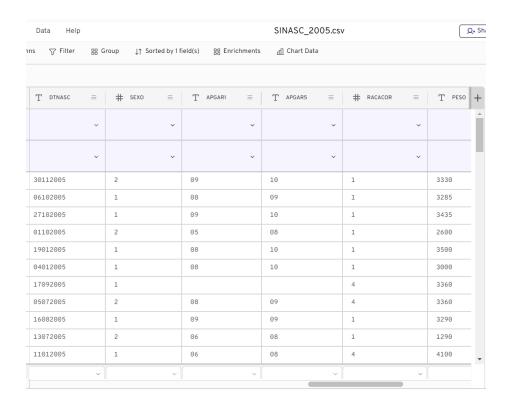
. . .

Leitos_2023.csv

DNOPEN22.csv DNOPEN23.csv SINASC_2005.csv SINASC_2021.csv



| | ↓↑ | Sorted by 1 field(s) 💢 Enrich | hments | 而 Chart Data | | | |
|-------------------------|------------|-------------------------------|--------|-----------------------|-------|--------------------|---|
| | | | | | | | |
| # ocupacaoHospitalarUti | = | # ocupacaoHospitalarCli | ≡ ‡ | ‡ saidaSuspeitaObitos | ≡ # | saidaSuspeitaAltas | = |
| | ~ | | ~ | | ~ | | ~ |
| | ~ | | ~ | | ~ | | ~ |
| 0 | | 0 | 0 | | 0 | | (|
| Θ | | 0 | 0 | | 0 | | (|
| Θ | | 1 | 1 | | 0 | | (|
| 0 | | 0 | 0 | | 0 | | 6 |
| 0 | | 0 | 0 | | 0 | | 6 |
| 0 | | 0 | 0 | | 0 | | 6 |
| 0 | | 0 | 0 | | 0 | | 6 |
| 0 | | 0 | Θ | | 0 | | 6 |
| Θ | | 0 | 1 | | 0 | | (|
| 0 | | 0 | 0 | | 0 | | 6 |
| 0 | | 2 | 0 | | 0 | | 6 |



A prerrogativa é avaliar o poder socioeconômico de uma cidade a partir do número de hospitais, número de leitos e proporção de leitos ocupados. Os outros dados se referem a fichas muito específicas de pacientes em particular ou outros dados micro dos hospitais, em suma, eles não se referem diretamente a poder econômico, os que os tornam desconexos ou muito complexos para o projeto.

Revisão dos Dados Selecionados

CNPJs (cnpj-datadream) (5,4 gb):

Arquivos:

cnpjs_1.csv (1.3 GB) cnpjs_2.csv (1.5 GB) cnpjs_3.csv (483.2 MB) cnpjs_4.csv (225.8 MB cnpjs_5.csv (1.9 GB)

total: 5,4 GB

Pesquisa de Orçamento Familiar (pofmain-datadream)

Arquivos:

aluguel_estimado.csv (4.9 MB) caderneta_coletiva.csv (97.3 MB) caracteristicas_dieta.csv (6.0 MB) condicoes_vida.csv (10.6 MB) consumo_alimentar.csv (394.3 MB) despesa_coletiva.csv (97.9 MB) despesa_individual.csv (311.2 MB)

domicilio.csv (7.8 MB)

inventario.csv (70.2 MB)

morador.csv (51.1 MB)

outros.csv (29.9 MB)

quali_vida.csv (38.3 MB)

rendimento.csv (22.5 MB)

restricao.csv (3.0 MB)

servico_pof2.csv (2.5 MB)

servico_pof4.csv (19.5 MB)

total: 1.167 MB

IBGE (ibge-datadream)

Arquivos:

gini_geral_s3.csv pib_s3.csv

total: 870,6 KB (<1MB)

SUS (datasus-datadream) (14,5 gb)

Arquivos:

Da pasta leitos:

Leitos 2015.csv (23.7 MB)

Leitos 2016.csv (23.5 MB)

Leitos_2017.csv (23.6 MB)

Leitos 2018.csv (23.5 MB)

Leitos_2019.csv (23.1 MB)

```
Leitos_2020.csv (23.7 MB)
Leitos_2021.csv (24.4 MB)
Leitos_2022.csv (24.3 MB)
```

Leitos_2023.csv (17.8 MB)

total: 207,6 MB

Da pasta LeitoOcupacao:

```
esus-vepi.LeitoOcupacao_2020.csv (119 MB)
esus-vepi.LeitoOcupacao_2021.csv (159,2 MB)
esus-vepi.LeitoOcupacao_2022.csv (62,7 MB)
```

total: 340,9 MB

total SUS: 548,5 MB

Tamanho Final dos Conjuntos de Dados Selecionados (em MB)

5400 - CPNJ

1167 - POF

1 - IBGE

548,5 - SUS

Total dos Conjuntos: 7.116,5 MB ou 7,12 GB