BReATH

PRESENTATION

Brazilian
Research of
Atmosphere
Towards
Health

INTEGRANTES

233840

Elton Cardoso do Nascimento

Implementação e integração do Banco de dados

234720

Gabriel Costa Kinder

Limpeza, tratamento e inserção do dataSUS

218733

João Pedro de Moraes Bonucci

Limpeza, tratamento e inserção dos dados climáticos

240106

Lucas Otávio Nascimento de Araújo

Ferramentas de análise sob o banco de dados

01 PROBLEMA

Relacionando o clima com doenças respiratórias

Bruno Rocha / FotoAerena / Estadão Conteúdo

TEMA

O data set consiste em um banco de dados relacional que agrega

- Dados climaticos
- Dados de qualidade do ar
- Dados de doenças respiratórias

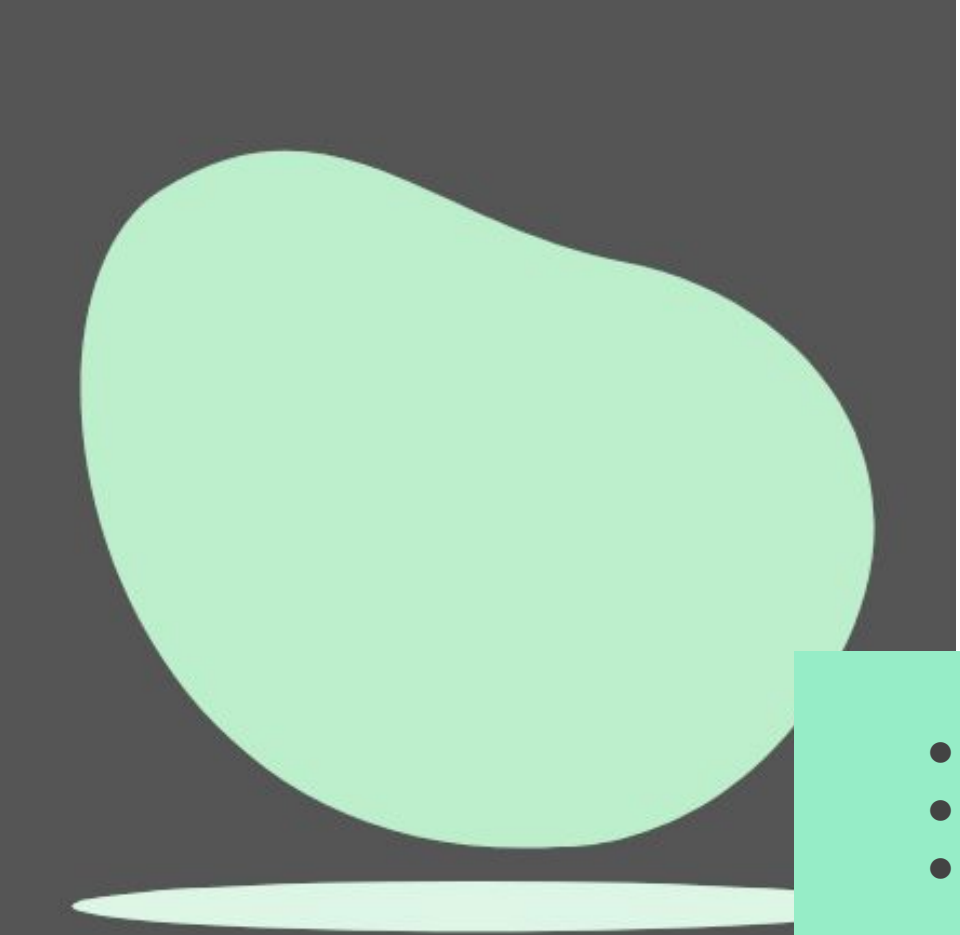
Do território Brasileiro entre os anos de 2000 e 2020 (nem todos os datasets contém todo o periodo). Com isso, ambicionamos encontrar ou fortalecer relações entre condições ambientais e doenças respiratórias no nosso país fornecendo dados específicos por cidade e mês ao longo de anos.

MOTIVAÇÃO E CONTEXTO GERADOR

Nossa motivação para o problema nasceu devido a duradoura estiagem que estamos passando somado ao contexto pandêmico onde doenças respiratórias são um tema de foco. Por isso queríamos trazer algo relacionado a saúde e ao tema debatido, mas contribuindo com bancos ainda não tão explorados.

Em 2017, dois problemas respiratórios estavam entre as dez maiores causas de morte do país. Considerando isso, escolhemos como objetivo predizer a incidência de doenças respiratórias por meio de dados ambientais (clima, poluição).

Nosso objetivo se alinha com os Objetivos de Desenvolvimento Sustentável da ONU, mais especificamente com o objetivo 3, “Garantir o acesso à saúde de qualidade e promover o bem-estar para todos, em todas as idades”.



02

DATASET

- **Fontes**
- **Modelo conceitual**
- **Modelo lógico**

Business illustrations by Storyset

FONTES DE DADOS UTILIZADAS

DADOS CLIMÁTICOS

DADOS DE SAÚDE

01

Climate Weather Surface of Brazil - Hourly

02

Banco de Coordenadas Geográficas das Cidades Brasileiras

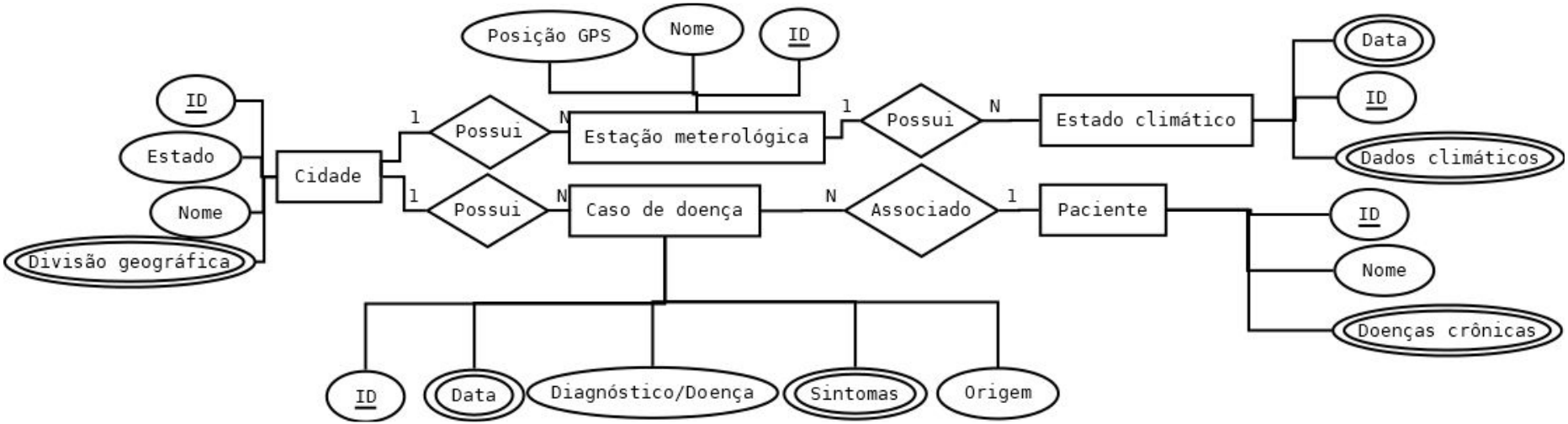
03

Plataforma da Qualidade do Ar

04

SRGA Banco de dados síndrome respiratória aguda grave 2013 - 2018

MODELO CONCEITUAL

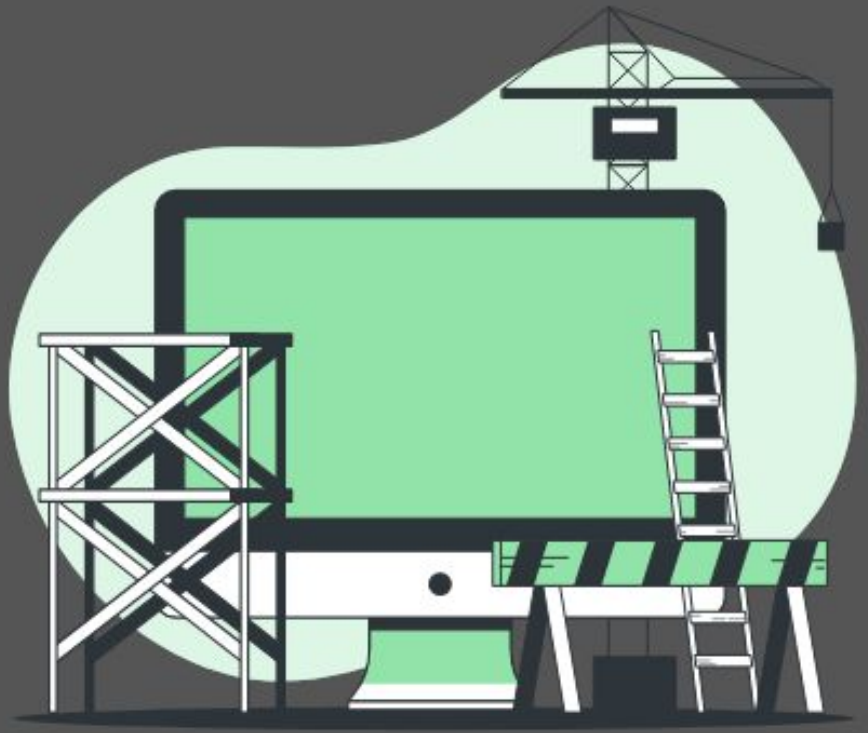


MODELO LÓGICO

03

OPERACÕES

- Operações de preparo
- Implementação física
- Integralização



Work Illustrations by Storyset

DATASUS

- Extração
  - Dados de diagnósticos referentes aos anos de 2013-2018 obtidos do dataSUS
- Agregação
  - Unificar as tabelas de diferentes anos
  - Juntar informações da tabela do IBGE (nome das Cidades)
  - Juntas informações de coordenadas geográficas (latitude e longitude)
- Tratamento
  - Remover cidades inválidas
  - Arrumar índices
- Integração
  - Integrar os dados do dataSUS com os outros dados de nosso BD

DADOS CLIMÁTICOS

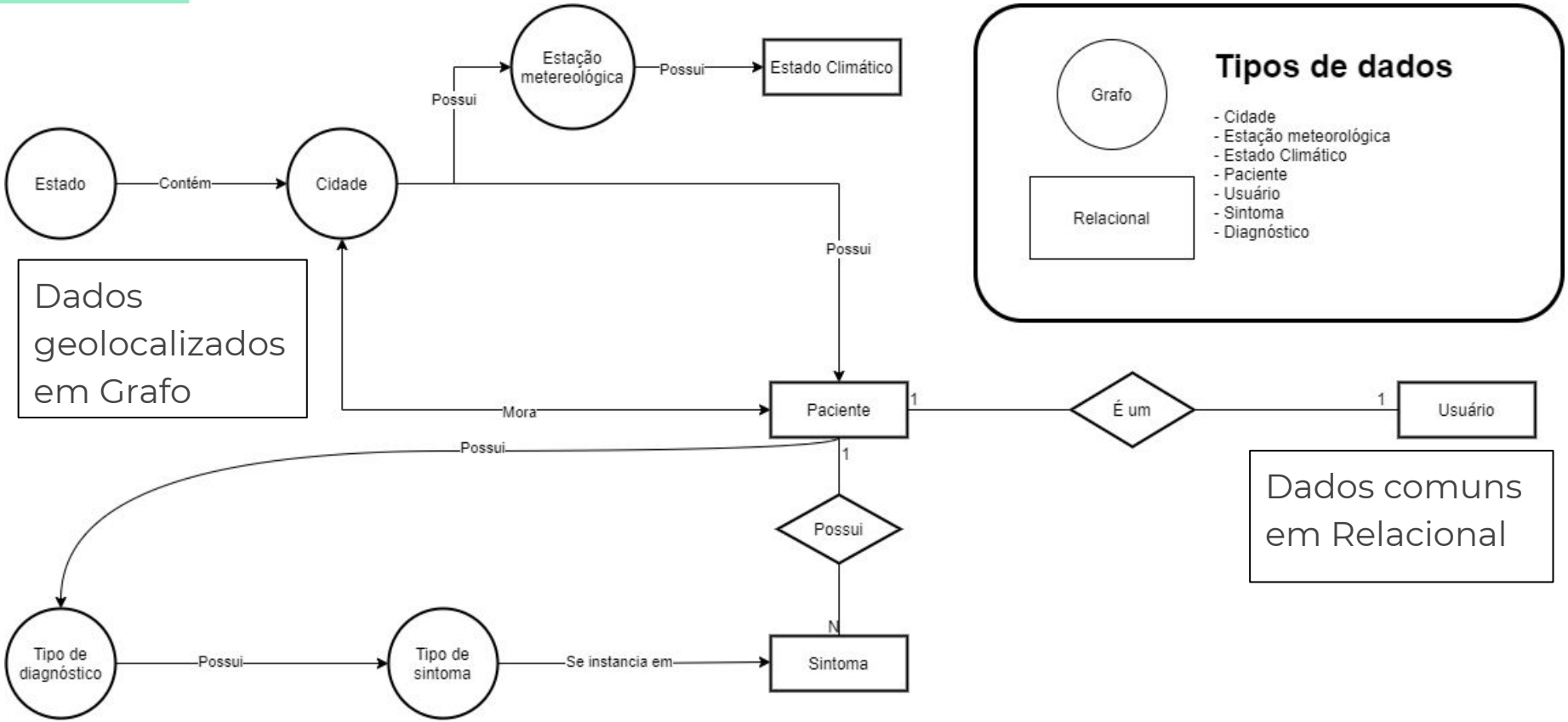
- Extração
  - Extrair dados provenientes de estações climáticas
- Transformação
  - Padronização das colunas entre as tabelas de diferentes regiões do país
- Agregação
  - Junção dos dados climáticos de diferentes estações de coletas com suas respectivas estações e localidades
- Tratamento
  - Remover dados incompletos
  - Normalizar dados
  - Correção de dados com erros de digitação
  - Renomeação e exclusão de colunas do banco de dados
- Integração
  - Integração com os dados de saúde do banco SRAG

IMPLEMENTAÇÃO FÍSICA

Relacional

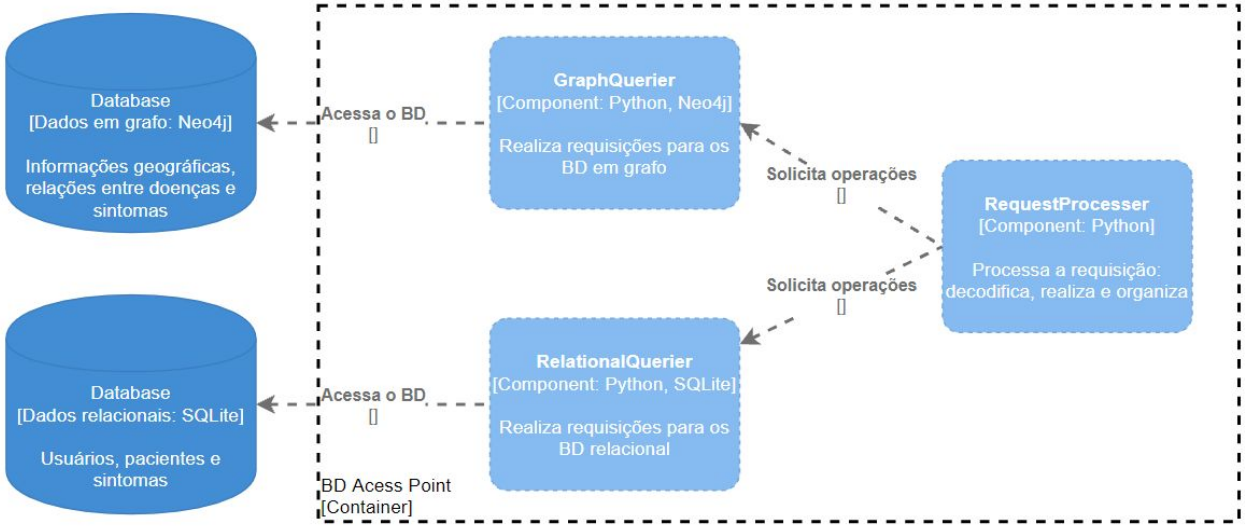
Grafo

DIVISÃO DOS DADOS



INTEGRALIZAÇÃO

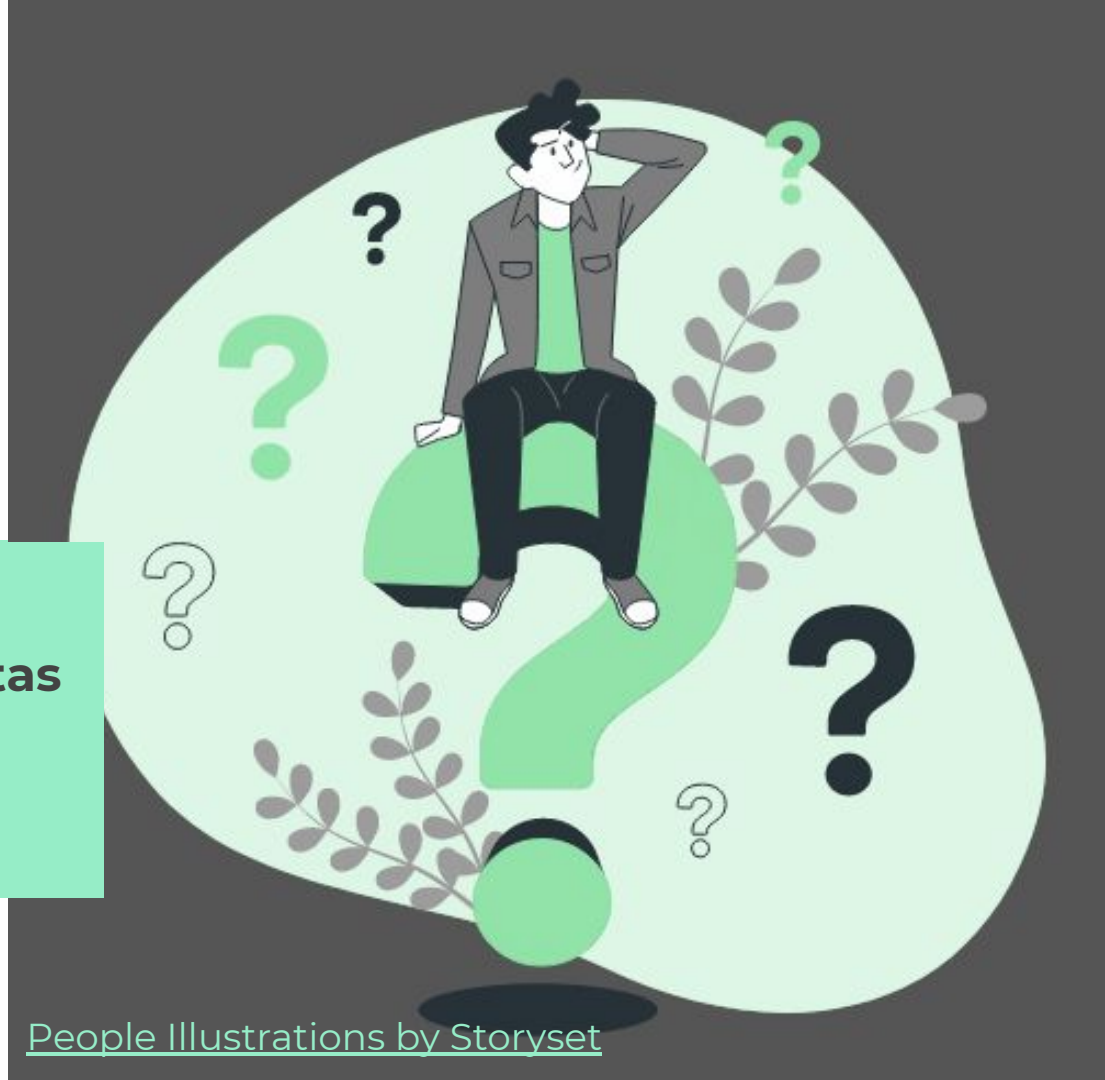
- Acesso como serviço para outras aplicações



04

ANÁLISE

- Conjunto de perguntas
- Consultas iniciais
- Machine Learning



People Illustrations by Storyset

PERGUNTAS

- Quais os sintomas mais comuns?
- Como a sazonalidade impacta a incidência de casos ou sintomas?
- Como o região impacta a ocorrência de casos?
- Quais fatores do tempo atmosférico afetam a incidência de casos? Como?
- É possível prever o aumento de casos de doenças respiratórias?

MACHINE LEARNING

Duas operações principais

- Clusterizar sintomas em um período de tempo, e depois comparar com o tempo
  - K-means
  - DBSCAN
  - Hierarchical Clustering

- Predizer a quantidade de casos, segundo o clima
  - Regressão linear
  - RNN

Todos os casos, clusterizados com K-means

k = 5

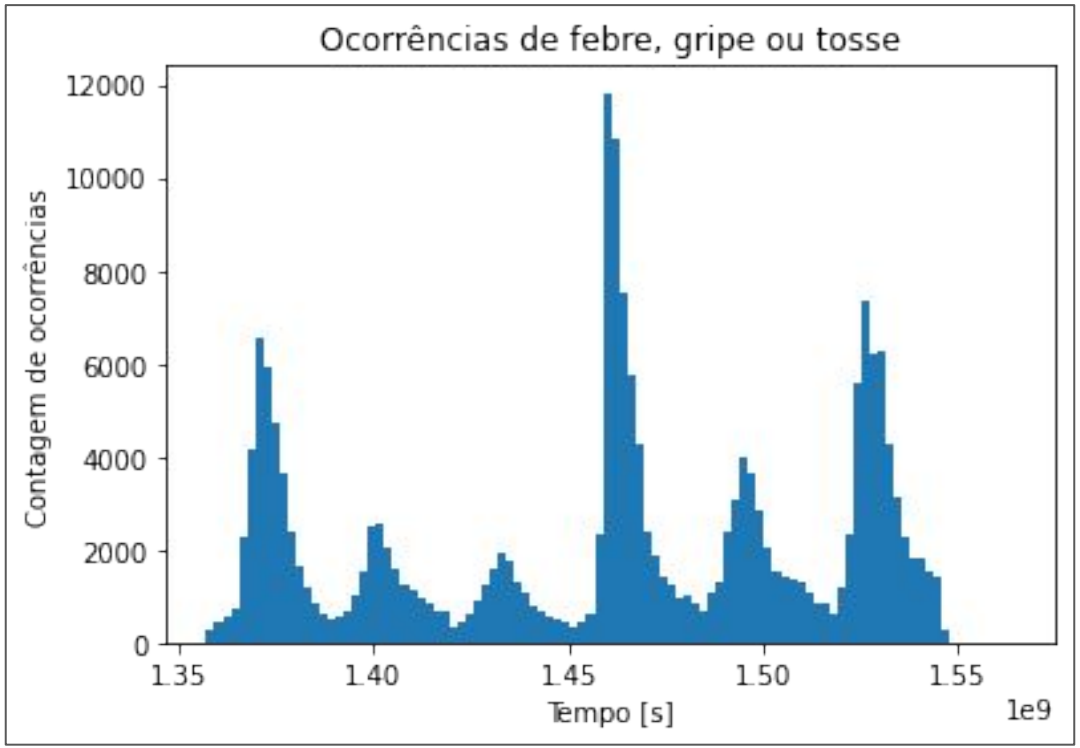
k = 27

Teste inicial

10% dos casos, selecionados aleatoriamente, regionalizados pelo Hierarchical Clustering

Teste inicial

O gráfico abaixo representa o total de casos por intervalo de tempo com síntomas específicos entre 2013 até 2018.



Teste inicial

05 PERGUNTAS

- **Map/Paralelização**
  - **Pandas vs SQL**
- **Uso de memória x eficiência**
  - **Pandas vs SQL**