Uma Análise da Correlação Entre Qualidade do Ar e Câncer

Alunos: Daniel Lopes
Gabriel Mattar
Vitor Lemos

Introdução

- Existem estudos que apontam correlação entre a concentração de material particulado presente no ar e doenças graves
- As maiores fontes antropogênicas de particulados são a queima de combustíveis fósseis em motores de combustão interna de veículos, termoelétricas e indústrias e as poeiras de construção e de áreas onde a vegetação natural foi removida
- As pequenas partículas e gotículas presentes no material particulado, principalmente no PM2.5, são responsáveis por uma série de problemas de saúde

Objetivos

- Através das visualizações realizadas será possível evidenciar se uma alta taxa de concentração de materiais particulados correlacionam com um alto índice de câncer na população
- Além disso será possível analisar se um tipo específico de material particulado está associado a determinados tipos de câncer

Caracterização - Data Mart do Câncer

- O Data Mart disponibilizado pelo NPCR (National Program Of Cancer Registries) contempla as estatísticas de câncer nos Estados Unidos (1999 - 2014)
- A fonte principal dos dados de incidência de câncer advém de registros médicos, os dados de câncer são registrados por um funcionário do hospital e então são enviados ao registro estadual
- A fonte dos dados de mortalidade são baseadas em estatísticas de certidões de óbito dos 50 estados
- 10 tabelas - Foi escolhida a tabela que agrupa os dados por estado (961776 registros)

Data Mart do Câncer

- Esse Data Mart contém 14 colunas com informações estatísticas sobre câncer nos EUA. As colunas utilizadas no trabalho foram:
  - Localidade de ocorrência
  - População
  - Tipo de câncer
  - Número de ocorrências
  - Tipo de ocorrência
  - Raça
  - Sexo
  - Ano

Qualidade dos dados - Data Mart do Câncer

- Não existe informação sobre o que interpretar no caso de instâncias preenchidas com conteúdo '~', '.', '-'
- Para alguns estados não possui informações de incidência e mortalidade, apresentando somente um dos tipos de evento

Caracterização - Data Mart Qualidade do Ar

- Possui todas as medições contidas no programa realizado pela EPA (United States Environmental Protection Agency)
- Possui 2.4 bilhões de valores de medição que datam desde 1957
- Dados possuem consistência nacional a partir do ano de 1980
- 32 tabelas - Foi escolhida a tabela com dados anuais (2038710 registros)

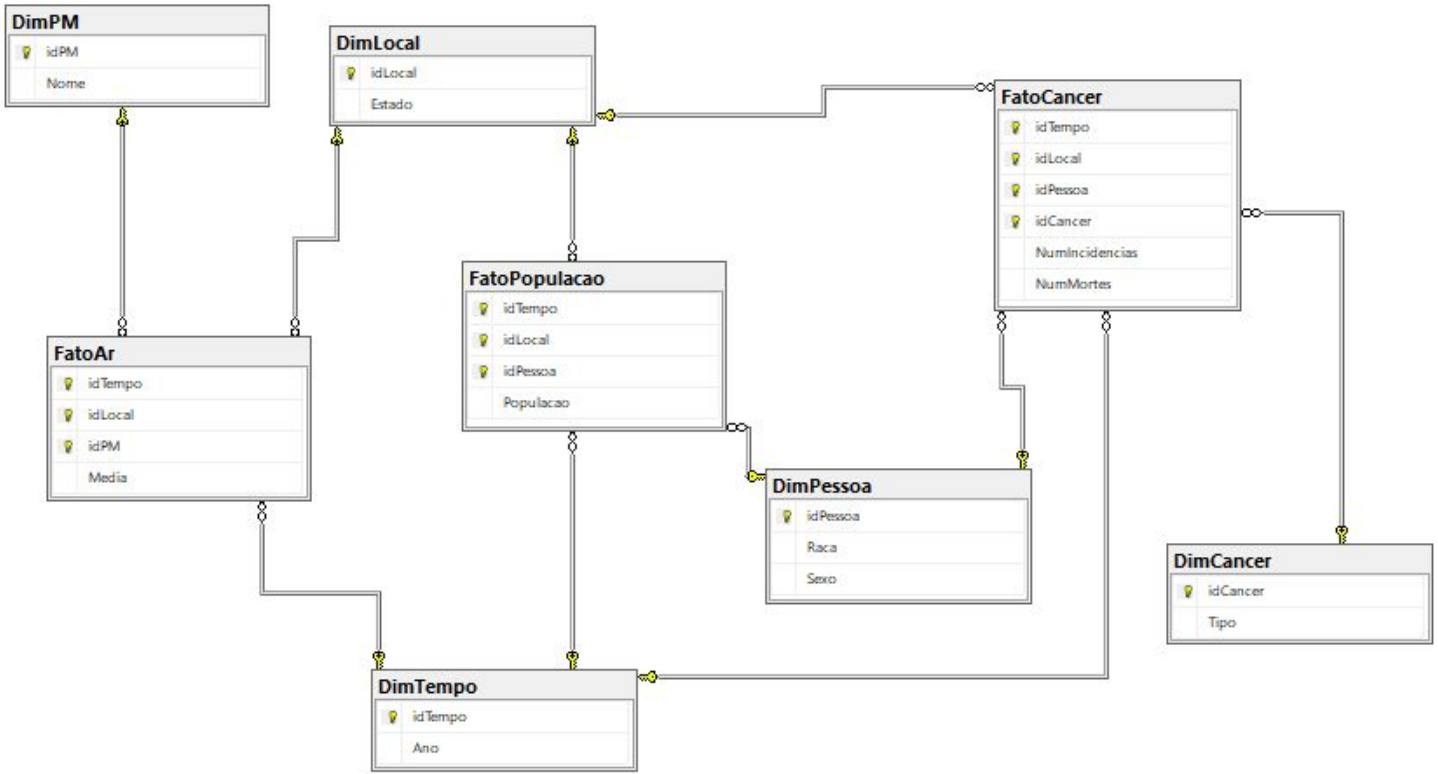
Data Mart Qualidade do Ar

- Esse Data Mart possui 55 colunas com diversas informações sobre a coleta e parâmetros de qualidade do ar. As colunas utilizadas no trabalho foram:
  - Parâmetro (poluente)
  - Quantidade presente no ar
  - Localidade da coleta do dado
  - Ano da medição
  - Unidade de medida

Qualidade dos dados - Data Mart Qualidade do Ar

- Falta de padronização nos nomes dos parâmetros
  - Monóxido de Carbono - CO
- Utilização de várias unidades de medida para uma mesma métrica
  - Microgramas por metro cúbico - Nanogramas por metro cúbico
- Parâmetros com diferentes frequências de medições
  - Medidos a cada hora - oito em oito horas - diariamente

Modelo Estrela



Processo de ETL

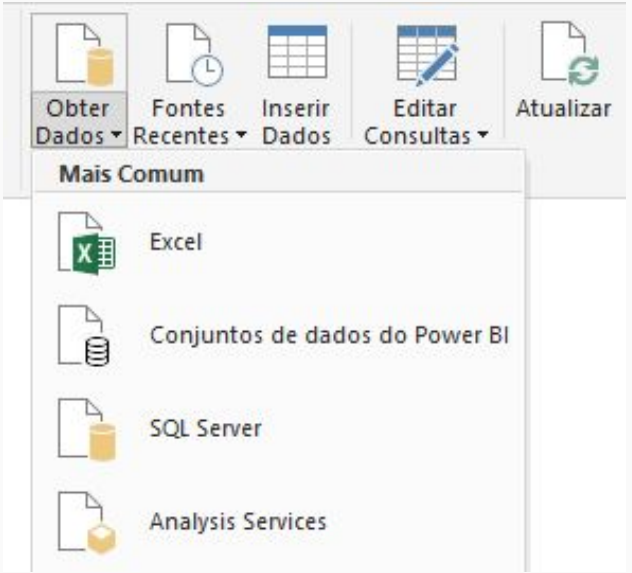
- O processo de ETL foi feito via SQL através do *SQL Server Management Studio*
- Foram geradas duas tabelas iniciais populadas com os dados dos dois Data Marts (fonte de arquivos simples - CSV e TXT)
    - Em seguida foram escritos scripts para tratar esses dados e popular as tabelas descritas no modelo estrela proposto

Exemplo de consulta do processo de ETL

```
INSERT INTO FatoCancer(idTempo, idLocal, idPessoa, idCancer, NumIncidencias, NumMortes)
SELECT T.idTempo, L.idLocal, P.idPessoa, C.idCancer, Aux.NumIncidencias, Aux.NumMortes
FROM (
    SELECT ISNULL(I.Ano, M.Ano), ISNULL(I.Estado, M.Estado), ISNULL(I.Raca, M.Raca), ISNULL(I.Sexo, M.Sexo),
        ISNULL(I.Tipo, M.Tipo), CASE WHEN I.NumIncidencias IN ('~', '.', '-') THEN NULL ELSE I.NumIncidencias END,
        CASE WHEN M.NumMortes IN ('~', '.', '-') THEN NULL ELSE M.NumMortes END
    FROM (
        SELECT TRIM(YEAR), TRIM(AREA), TRIM(RACE), TRIM(SEX), REPLACE(REPLACE(SITE, '<i>', ''), '</i>', ''), COUNT
        FROM dbo.BYAREA
        WHERE LEN(TRIM(YEAR)) = 4 AND TRIM(RACE) <> 'All Races' AND TRIM(SEX) <> 'Male and Female'
            AND SITE <> 'All Cancer Sites Combined' AND TRIM(EVENT_TYPE) = 'Incidence'
    ) AS I(Ano, Estado, Raca, Sexo, Tipo, NumIncidencias)
    FULL OUTER JOIN (
        SELECT TRIM(YEAR), TRIM(AREA), TRIM(RACE), TRIM(SEX), REPLACE(REPLACE(SITE, '<i>', ''), '</i>', ''), COUNT
        FROM dbo.BYAREA
        WHERE LEN(TRIM(YEAR)) = 4 AND TRIM(RACE) <> 'All Races' AND TRIM(SEX) <> 'Male and Female'
            AND SITE <> 'All Cancer Sites Combined' AND TRIM(EVENT_TYPE) = 'Mortality'
    ) AS M(Ano, Estado, Raca, Sexo, Tipo, NumMortes)
        ON I.Ano = M.Ano AND I.Estado = M.Estado AND I.Raca = M.Raca AND I.Sexo = M.Sexo AND I.Tipo = M.Tipo
) AS Aux(Ano, Estado, Raca, Sexo, Tipo, NumIncidencias, NumMortes)
JOIN DimTempo T ON T.Ano = Aux.Ano
JOIN DimLocal L ON L.Estado = Aux.Estado
JOIN DimPessoa P ON P.Raca = Aux.Raca AND P.Sexo = Aux.Sexo
JOIN DimCancer C ON C.Tipo = Aux.Tipo;
GO
```

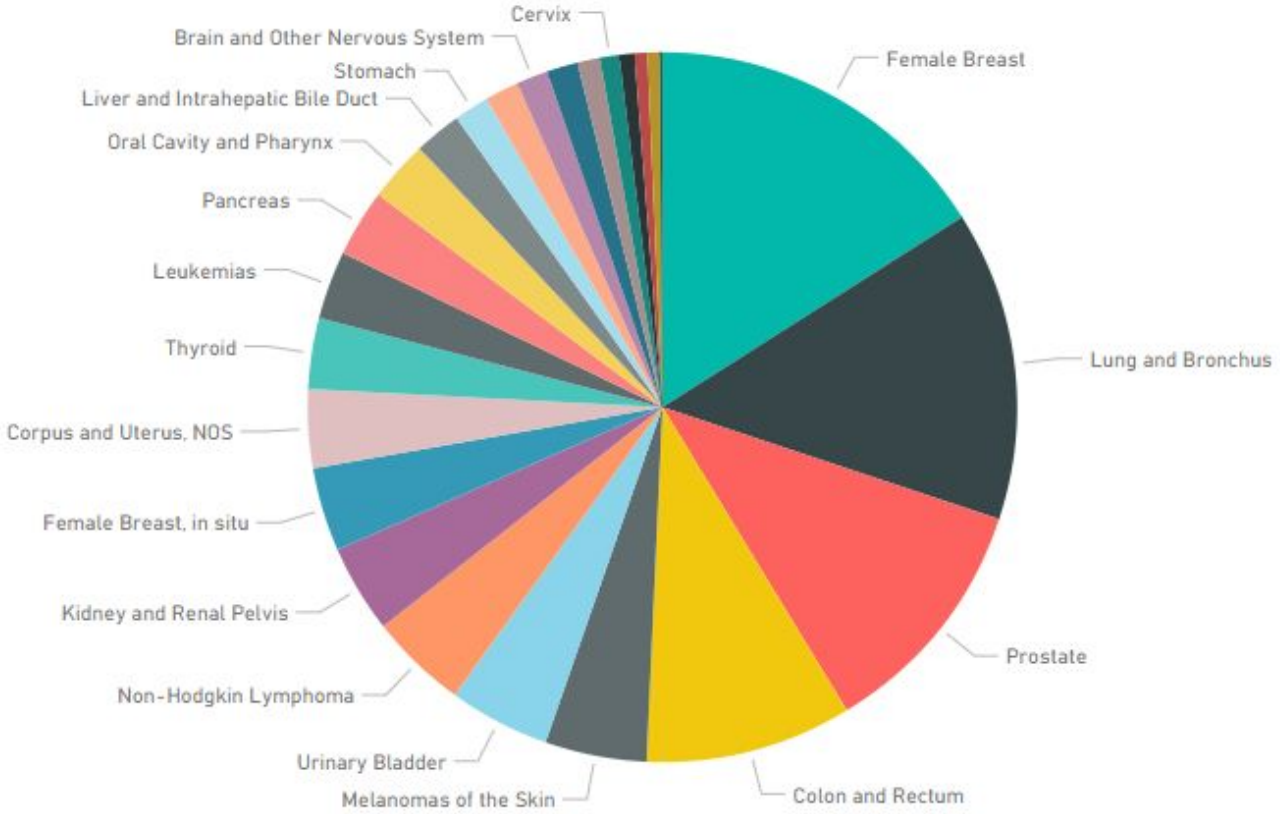
Importação dos dados

- As tabelas geradas pelo processo de ETL foram importadas diretamente através do software *PowerBI* (possui integração com o *SQL Server*).

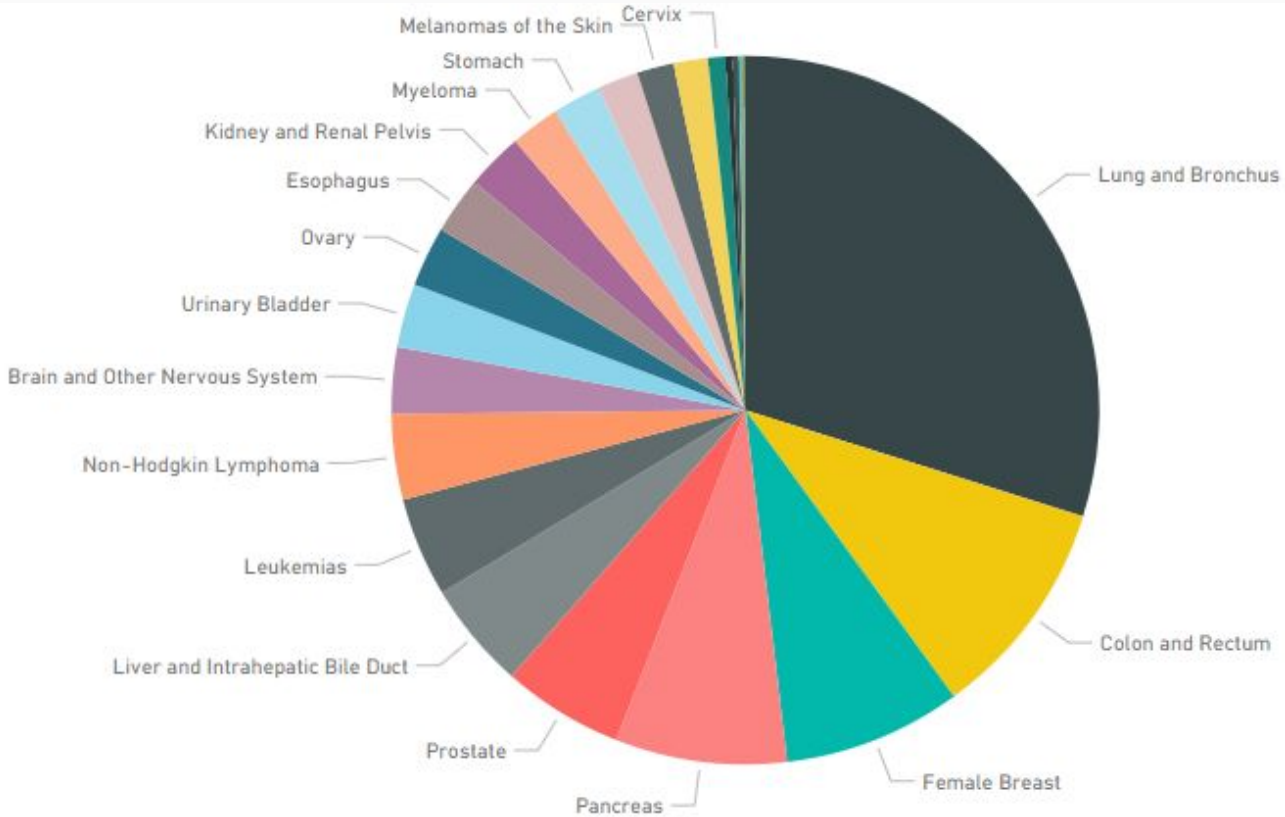


Visualizações dos Dados

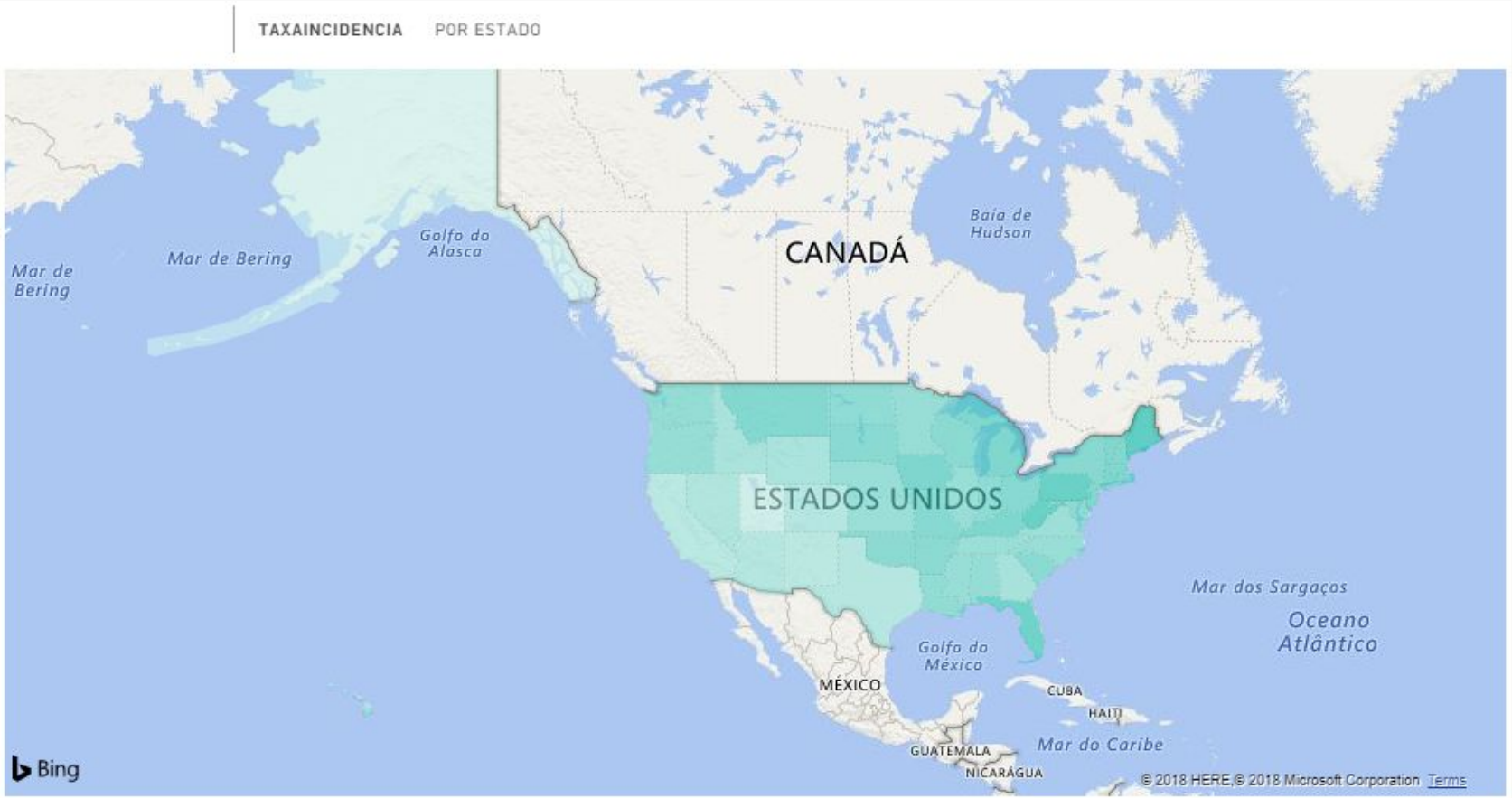
Número de Incidências por Tipo de Câncer



Número de Mortes por Tipo de Câncer



Taxa de Incidência de Câncer em 2003

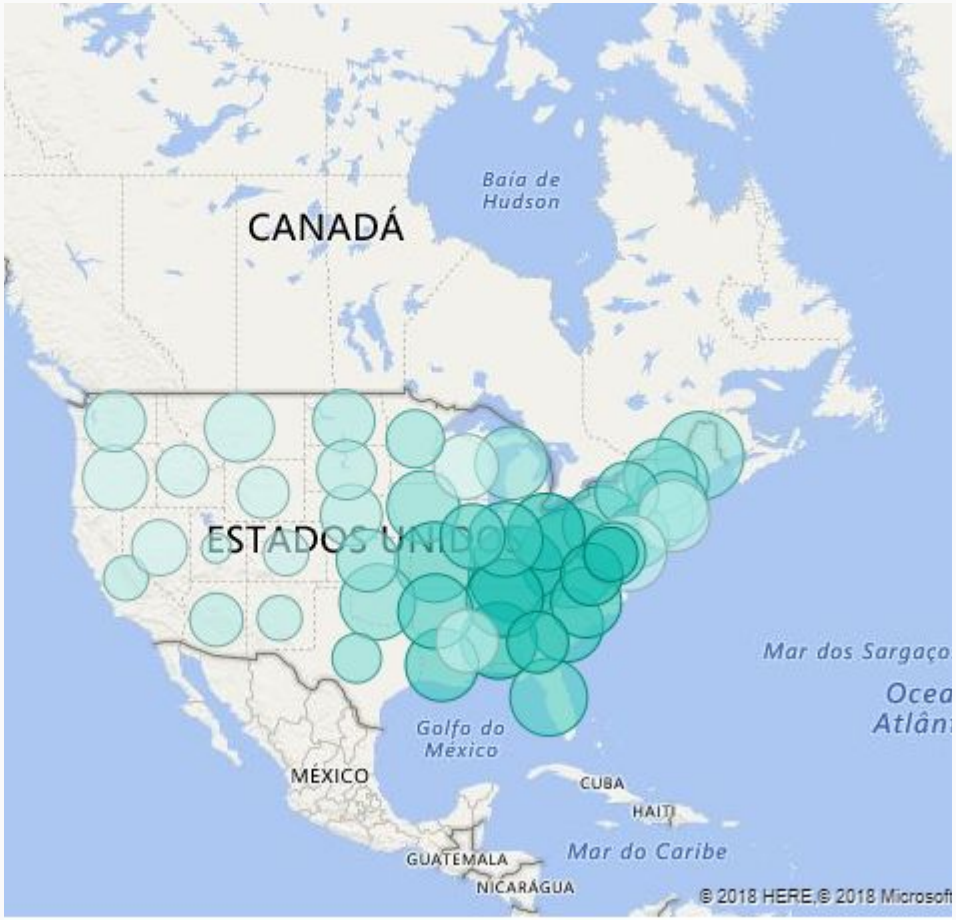


Taxa de Incidência de Câncer em 2014

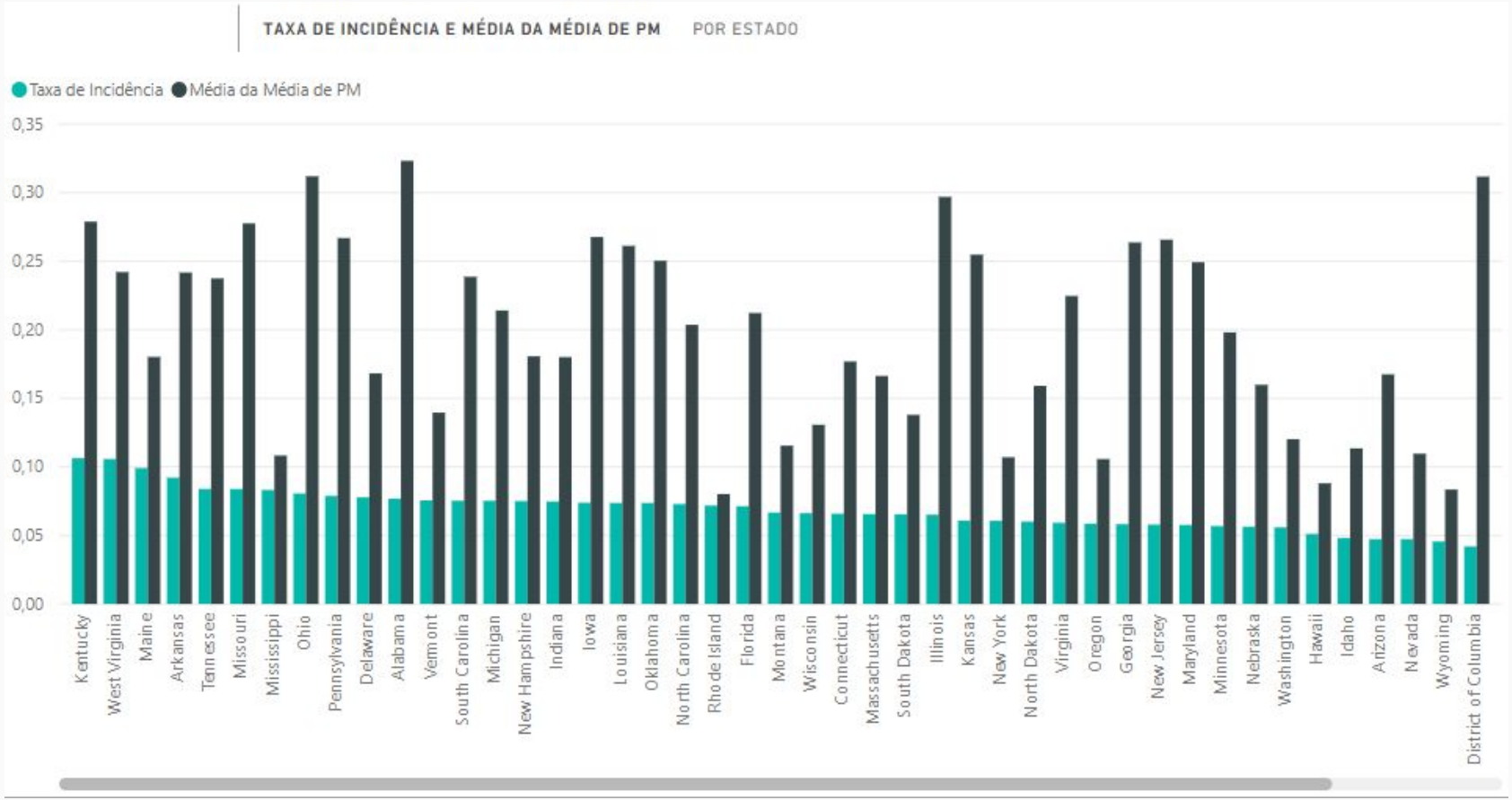
Legenda

Tamanho dos círculos:
Taxa de Incidência de Câncer de Pulmão

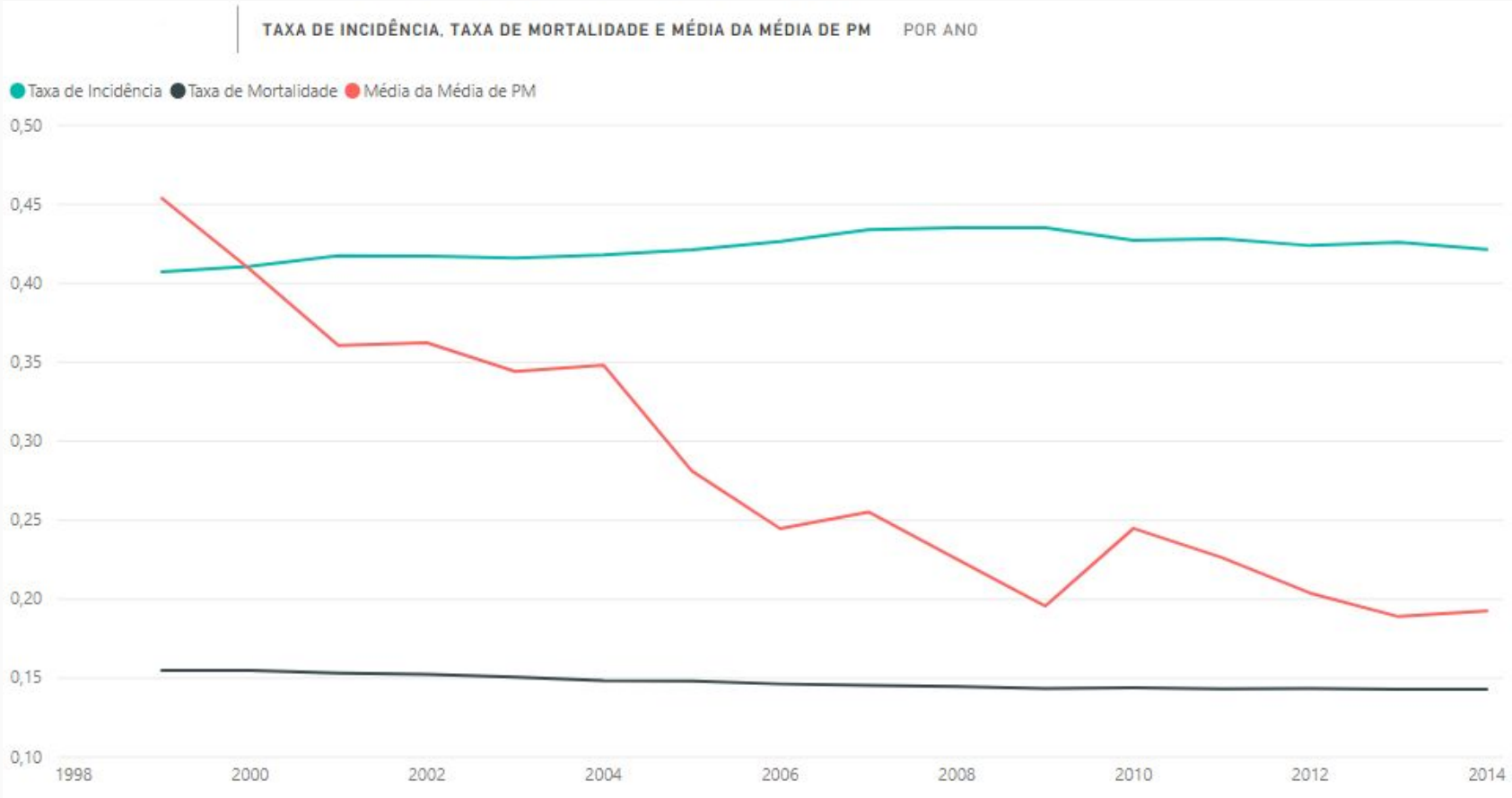
Saturação dos círculos:
Média de poluentes



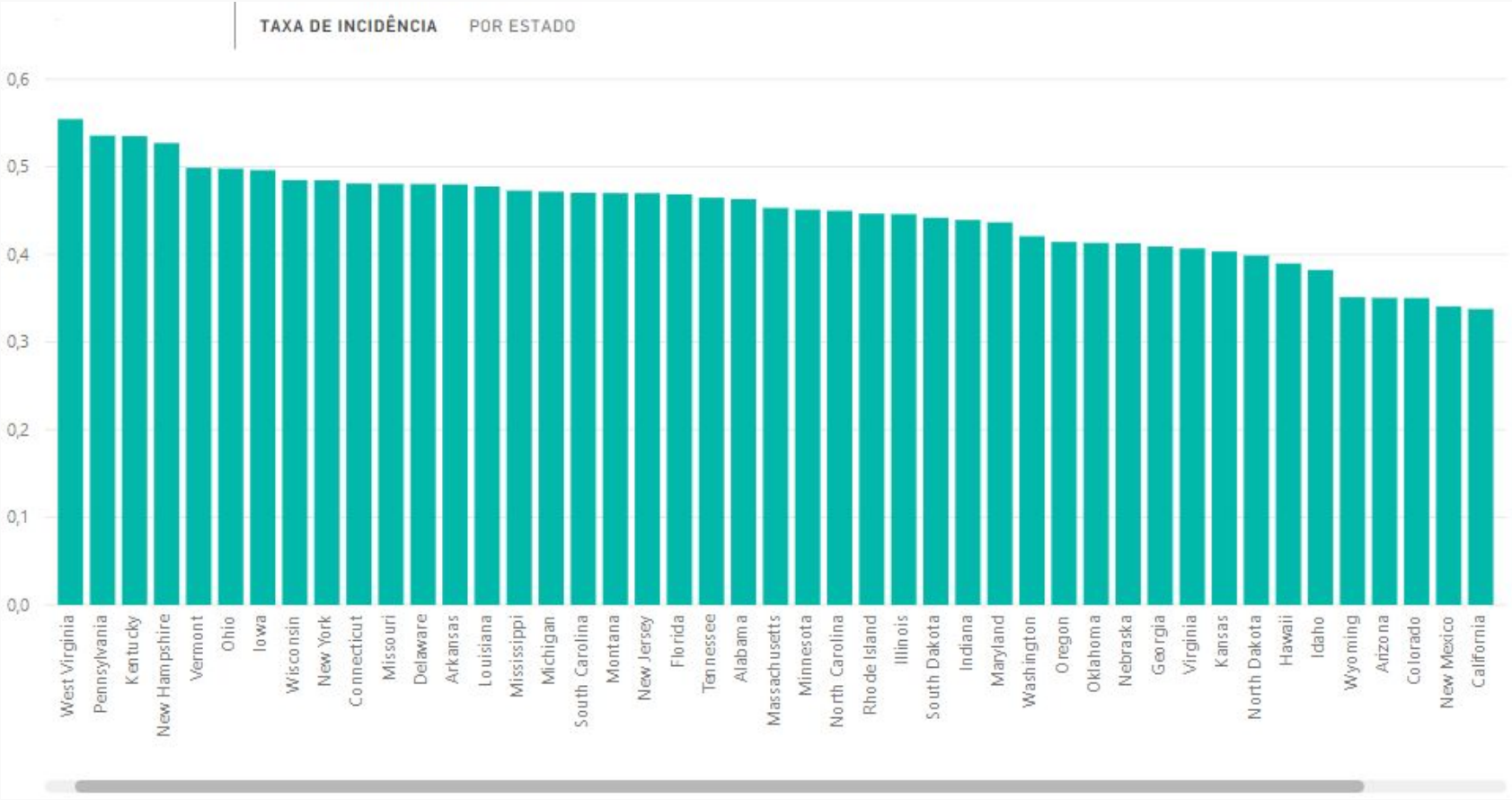
Taxa de Incidência e Média de Poluentes por Estado



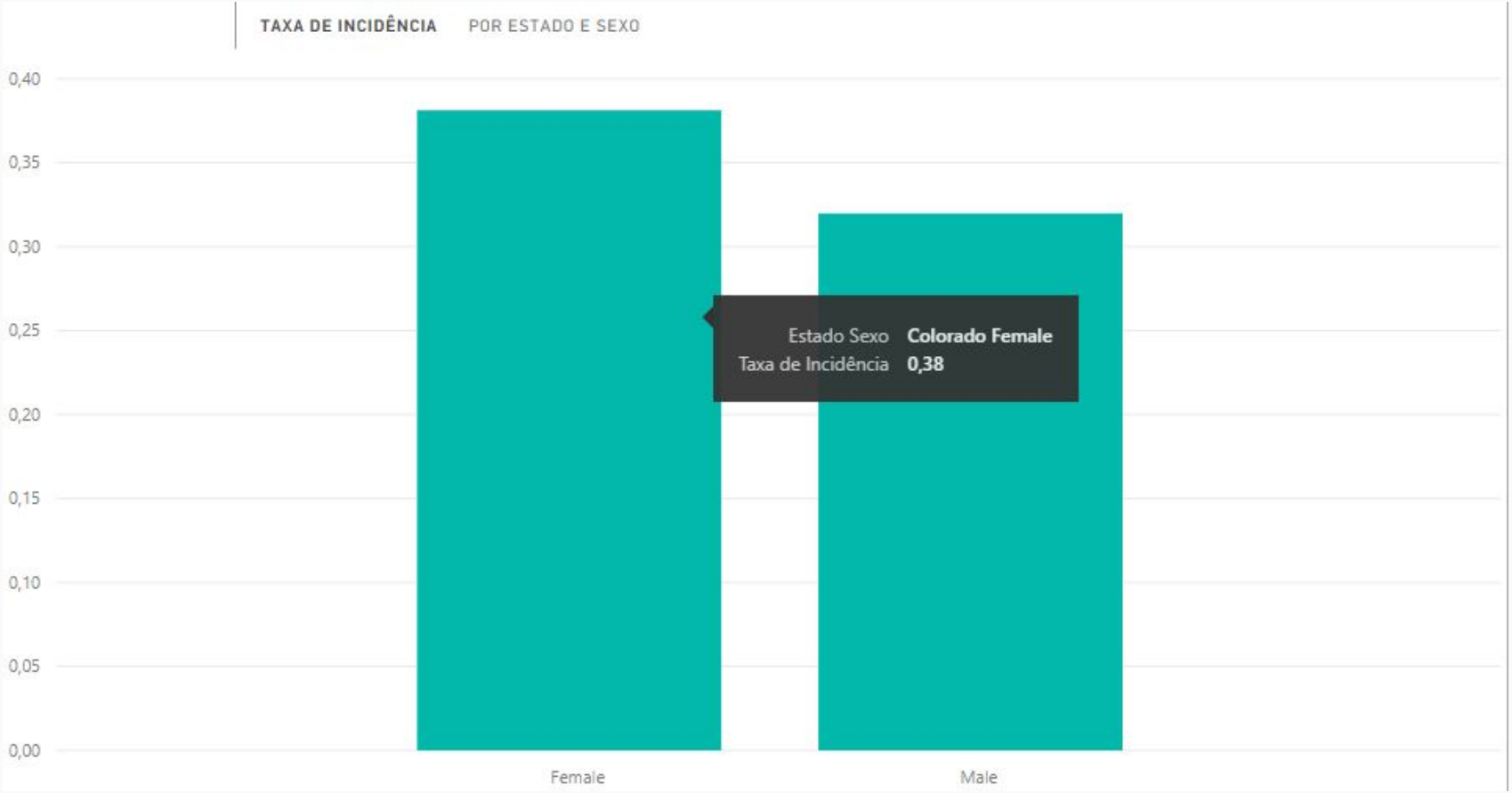
Tentativa Frustrada de Estabelecer Correlação Entre PM e Taxa de Câncer



Taxa de Incidência de Câncer Por Estado

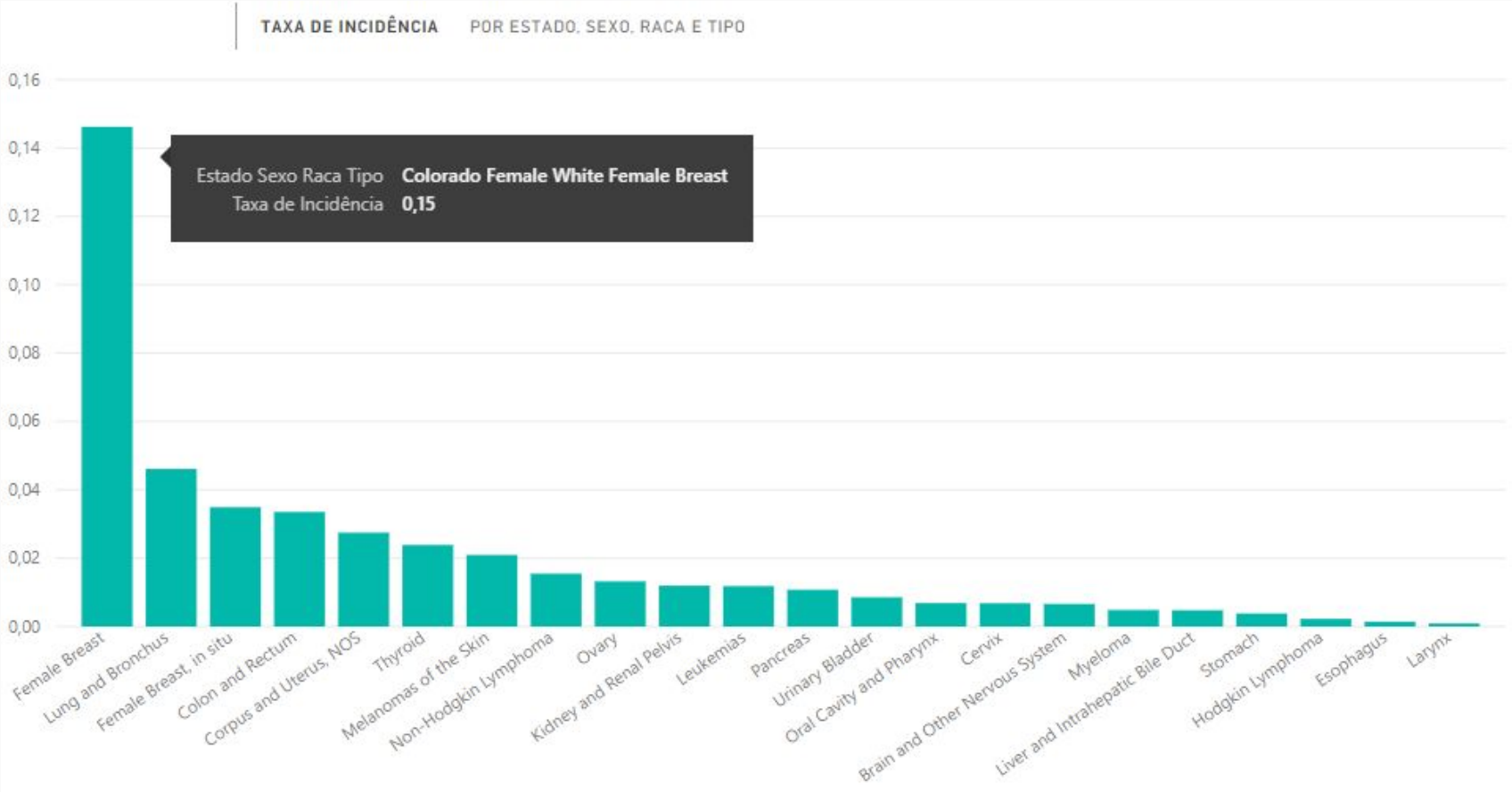


Taxa de Incidência por Estado e Sexo (Exemplo de *Drill Down*)



Taxa de Incidência por Estado, Sexo e Raça (*Drill Down*)

Taxa de Incidência por Estado, Sexo, Raça e Tipo de Câncer (*Drill Down*)



Conclusão

- Foi possível explorar diversos aspectos fornecidos pelos dois *DataSets*
  - No *DataSet* de câncer foi possível observar correlações entre raça e determinado tipo de câncer, tipos de câncer que afetam mais homens do que mulheres
  - No *DataSet* de ar foi possível observar as variações de poluentes ao longo dos anos e como a concentração desses poluentes estão distribuídas nos estados do EUA
- Não foi possível observar correlação significativa entre a concentração de poluentes e um aumento na quantidade de câncer na população

Referências

https://www.ecodebate.com.br/2017/11/01/poluicao-do-ar-esta-associada-mortalidade-por-canceres-nao-pulmonares/

https://www.kaggle.com/epa/epa-historical-air-quality/data

https://www.cdc.gov/cancer/npcr/uscs/download_data.htm