Engenharia de Dados na Análise de um Dataset de Casos de Câncer no Brasil

Este projeto aplica técnicas de engenharia de dados para analisar e interpretar um vasto conjunto de dados disponível no site kaggle sobre o câncer no Brasil. Com a finalidade de explorar a incidência de casos, relação faixa etária/óbitos por câncer e outros. A análise de dados pode ajudar a responder questões cruciais na área médica, contribuindo para políticas de saúde mais eficazes e melhores práticas de tratamento. Com um enfoque específico na doença do câncer, este projeto tem o potencial de influenciar positivamente a prevenção, diagnóstico e tratamento.

O governo brasileiro, por meio do Instituto Nacional do Câncer (INCA), é responsável pela geração de estimativas relacionadas ao câncer no país. Para isso, o INCA estabelece centros de coleta de dados sistemáticos, conhecidos como Registros de Câncer com Base Populacional (RCBP). Esses registros seguem as leis regionais vigentes e estão disponíveis para solicitação por qualquer indivíduo interessado.

Foi construída uma solução robusta para gerir, armazenar, limpar, modelar e processar dados, com o auxílio de tecnologias da AWS. Essas tecnologias incluíram:

AWS S3: Ferramenta que permitiu armazenar dados brutos e processados com segurança e flexibilidade.

AWS Glue: Utilizamos o Glue para realizar o ETL (Extração, Transformação e Carregamento) dos dados. Extração, processamento, limpeza e modelagem dados com glue job.

AWS Athena: Analise dos dados diretamente do S3. Execução de consultas SQL sem a necessidade de configurar servidores ou clusters.

Para automatizar e gerenciar nossa infraestrutura de maneira eficiente, foi utilizado o Terraform, uma ferramenta de Infraestrutura como Código (IaC). Criação e provisionamento da infraestrutura de forma declarativa, aumentando a produtividade e a eficiência de nas operações.

Na etapa de Análise Exploratória foram descobertos os vários insights importantes abaixo:

Mais mulheres sofrem com câncer no Brasil

A incidência de casos por gênero:

Feminino: 971.471

Masculino: 806.534

Podemos assumir que no Brasil, as mulheres apresentam uma maior incidência de câncer em comparação aos homens.

Relação faixa etária e diagnóstico de câncer.

Faixa Etária	Tipo do Cançer	Casos
0-4	LEUCEMIA LINFOBLASTICA AGUDA	2194
0-4	RIM	849
0-4	RETINA	654
0-4	GLANDULA SUPRA-RENAL, SOE	445
0-4	LEUCEMIA MIELOIDE AGUDA	368

Faixa Etária	Tipo do Cançer	Casos
5-14	LEUCEMIA LINFOBLASTICA AGUDA	2386
5-14	LEUCEMIA MIELOIDE AGUDA	620
5-14	CEREBRO	600
5-14	OSSOS LONGOS DOS M. INFERIORES	585
5-14	CEREBELO	523

Faixa Etária	Tipo do Cançer	Casos
15-24	COLO DO UTERO, SOE	6425
15-24	GLANDULA TIREOIDE	2741
15-24	TESTICULO, SOE	1286
15-24	DOENCA DE HODGKIN, SOE	1172
15-24	LEUCEMIA LINFOBLASTICA AGUDA	1119

Faixa Etária	Tipo do Cançer	Casos
25-34	COLO DO UTERO, SOE	26353
25-34	GLANDULA TIREOIDE	8991
25-34	MAMA , SOE	6080
25-34	PELE DE OUTRAS PARTES DA FACE	2378
25-34	TESTICULO, SOE	2286

Faixa Etária	Tipo do Cançer	Casos
35-44	COLO DO UTERO, SOE	26849
35-44	MAMA , SOE	22947
35-44	GLANDULA TIREOIDE	12449
35-44	PELE DE OUTRAS PARTES DA FACE	7739
35-44	NEOPLASIA MALIGNA DA PELE	4261

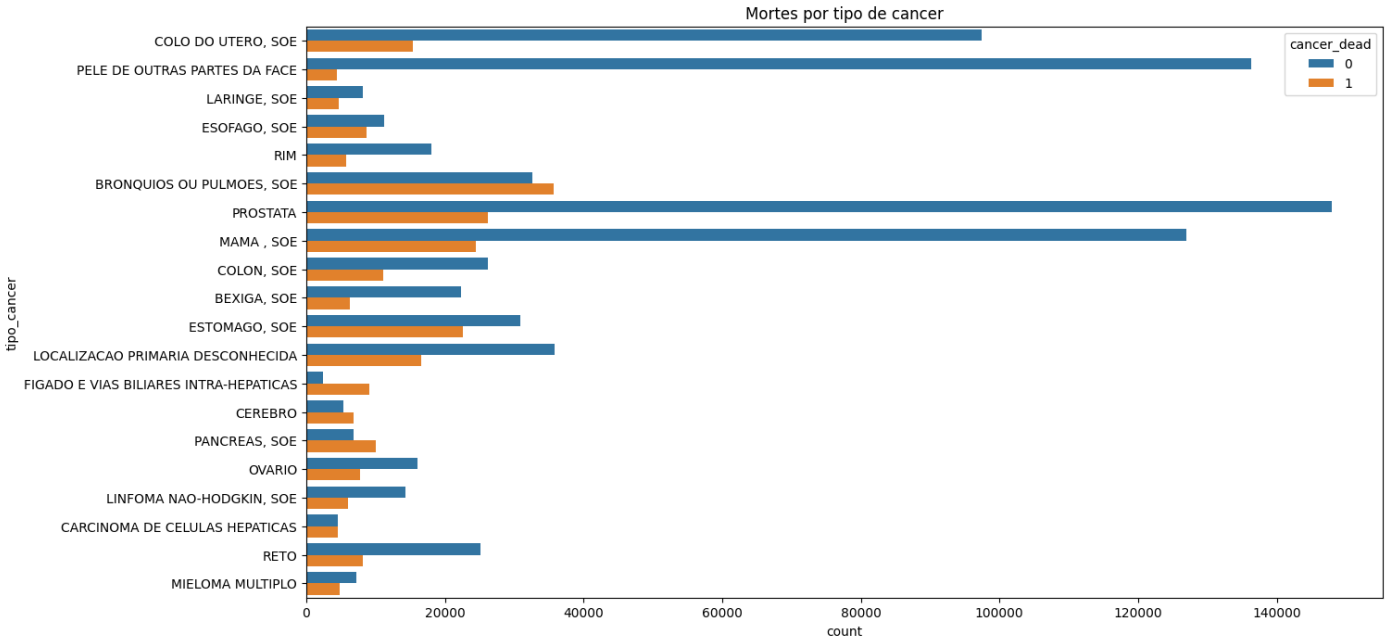
Faixa Etária	Tipo do Cançer	Casos
45-54	MAMA , SOE	38321
45-54	COLO DO UTERO, SOE	20070
45-54	PELE DE OUTRAS PARTES DA FACE	17135
45-54	GLANDULA TIREOIDE	13386
45-54	PROSTATA	12074

Faixa Etária	Tipo do Cançer	Casos
55-64	PROSTATA	43960
55-64	MAMA , SOE	33923
55-64	PELE DE OUTRAS PARTES DA FACE	26243
55-64	BRONQUIOS OU PULMOES, SOE	17533
55-64	COLO DO UTERO, SOE	13928

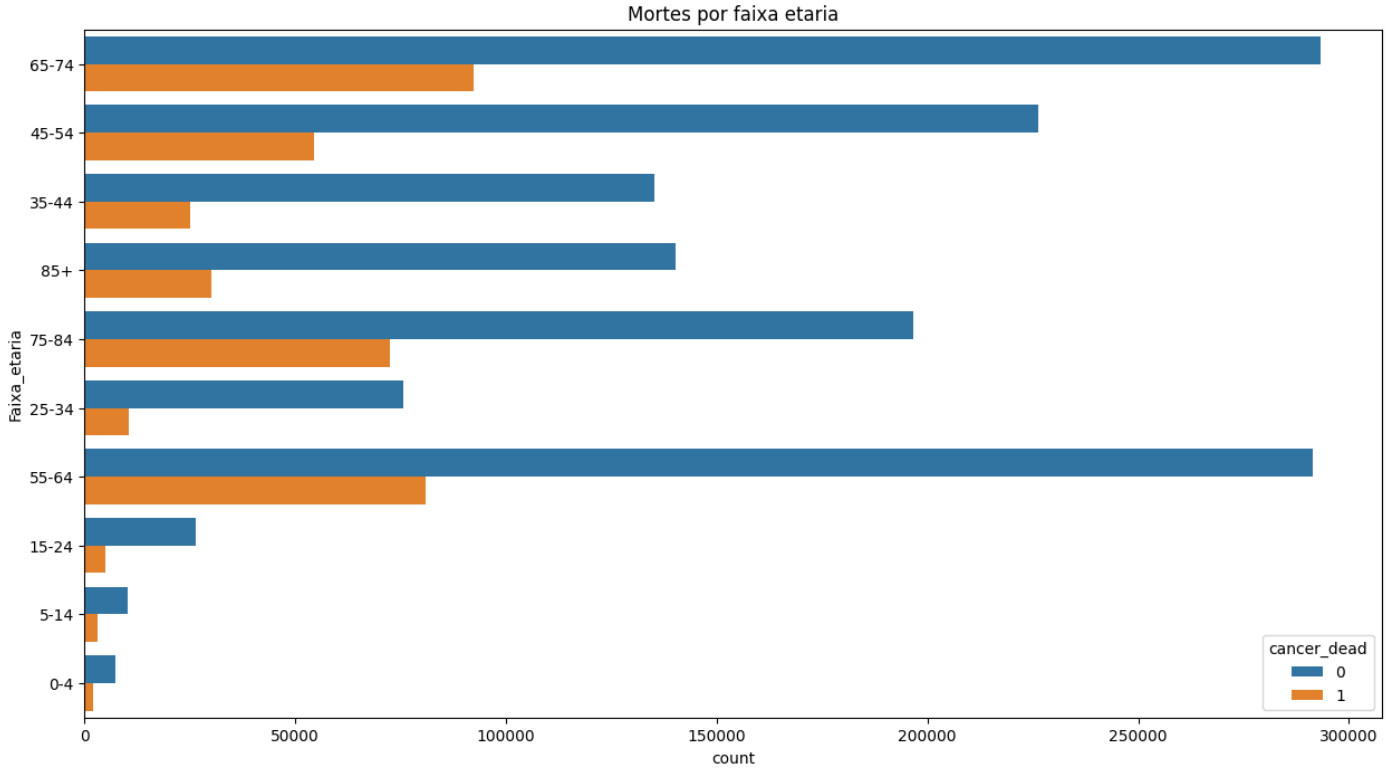
Faixa Etária	Tipo do Cançer	Casos
65-74	PROSTATA	63915
65-74	PELE DE OUTRAS PARTES DA FACE	32730
65-74	MAMA , SOE	24728
65-74	BRONQUIOS OU PULMOES, SOE	20503
65-74	ESTOMAGO, SOE	13567

Faixa Etária	Tipo do Cançer	Casos
75-84	PROSTATA	37224
75-84	PELE DE OUTRAS PARTES DA FACE	30031
75-84	MAMA , SOE	14979
75-84	BRONQUIOS OU PULMOES, SOE	13135
75-84	NEOPLASIA MALIGNA DA PELE	10950

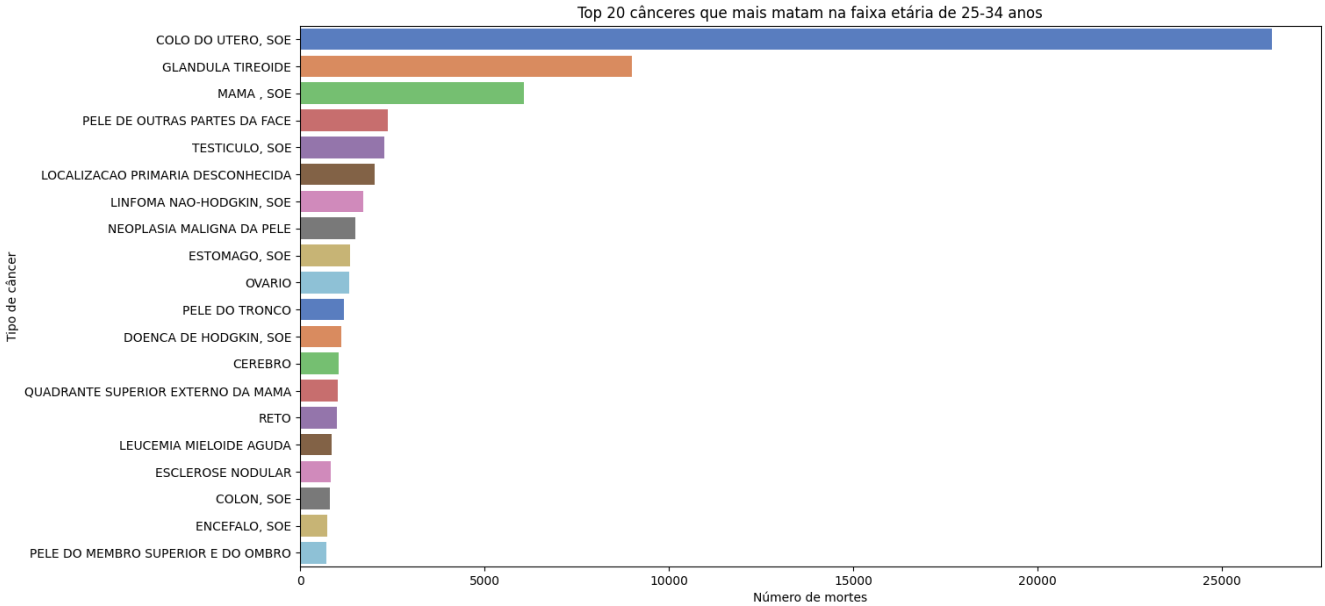
Quando analisamos a relação morte por câncer e tipo do câncer, percebemos altas taxas de mortalidades em tipos de canceres relativamente comuns, como é o caso do câncer nos brônquios ou pulmões, no estomago, no esôfago.



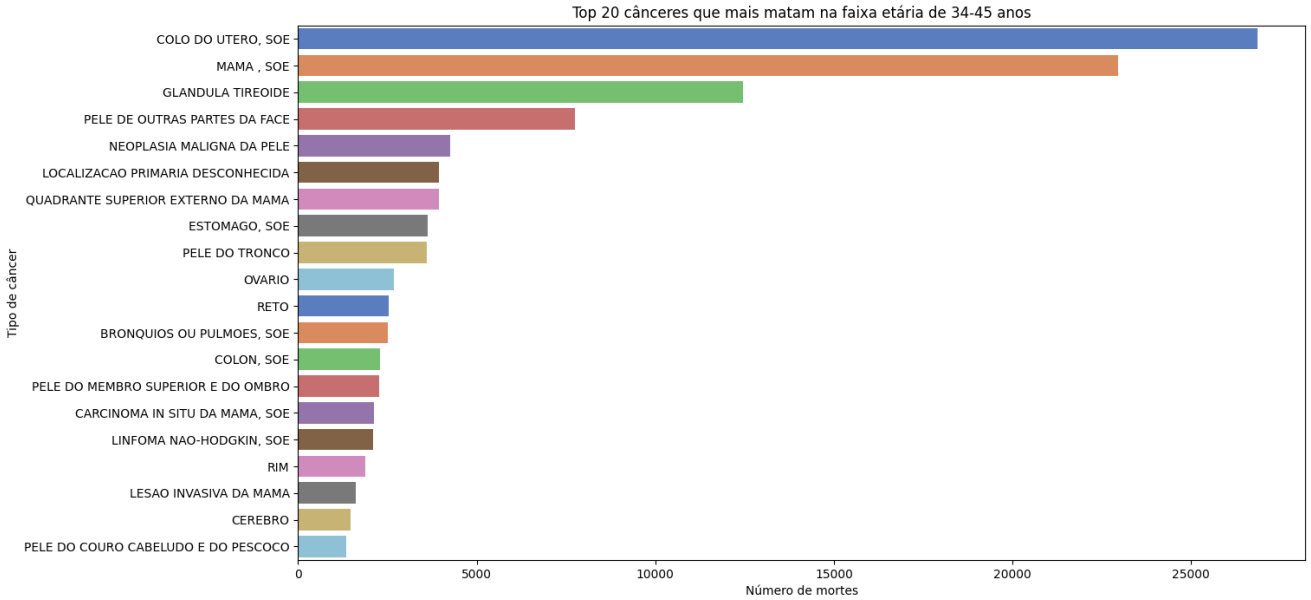
Mortes por faixa etária, altos índices entre 65 a 74 anos, 55 a 64 anos e 75 a 84 anos.



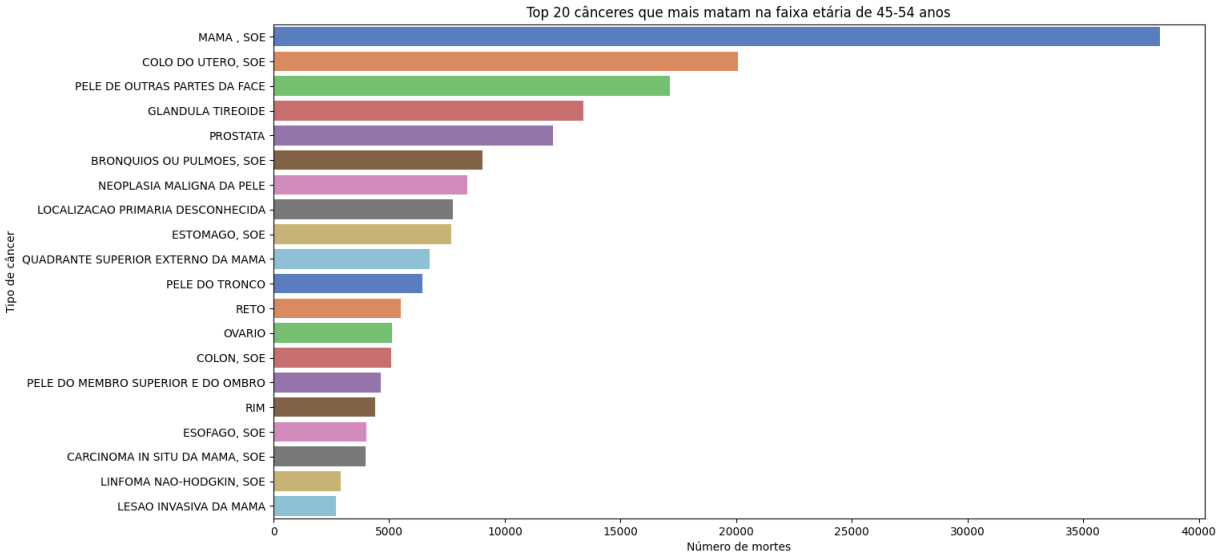
Os 20 principais cânceres que mais levam a óbito na faixa etária de 25-34 anos.



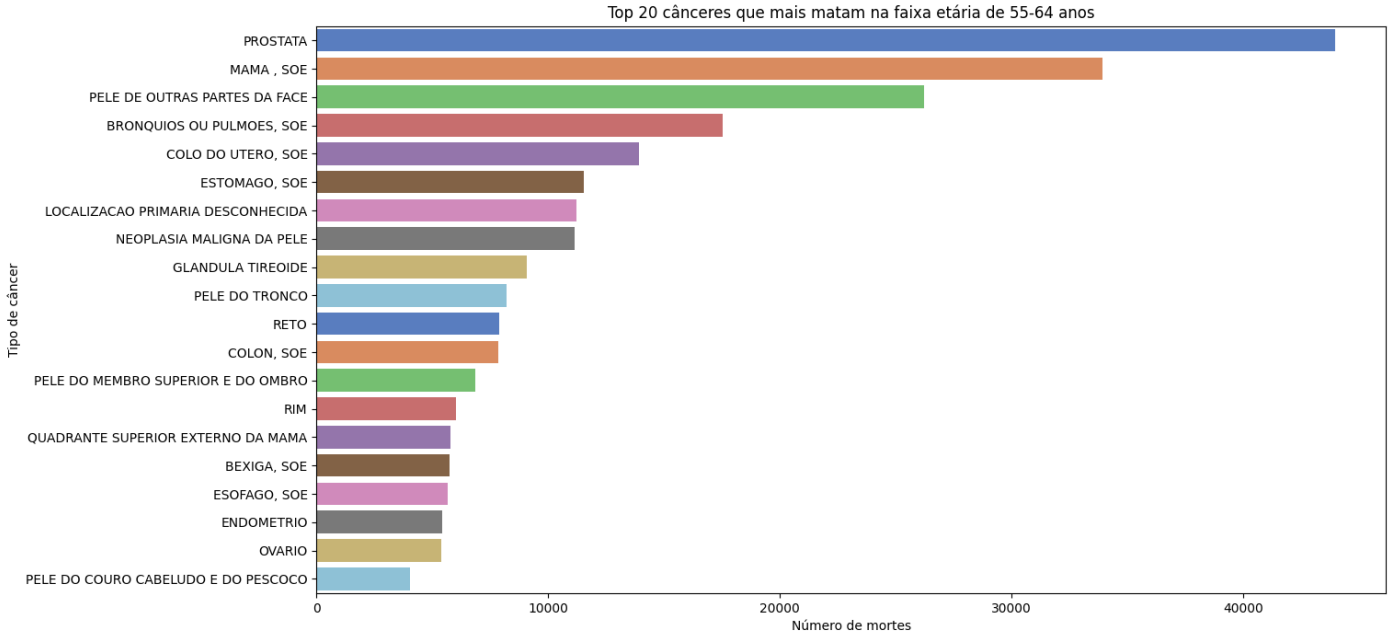
Os 20 principais cânceres que mais levam a óbito na faixa etária de 35-44 anos.



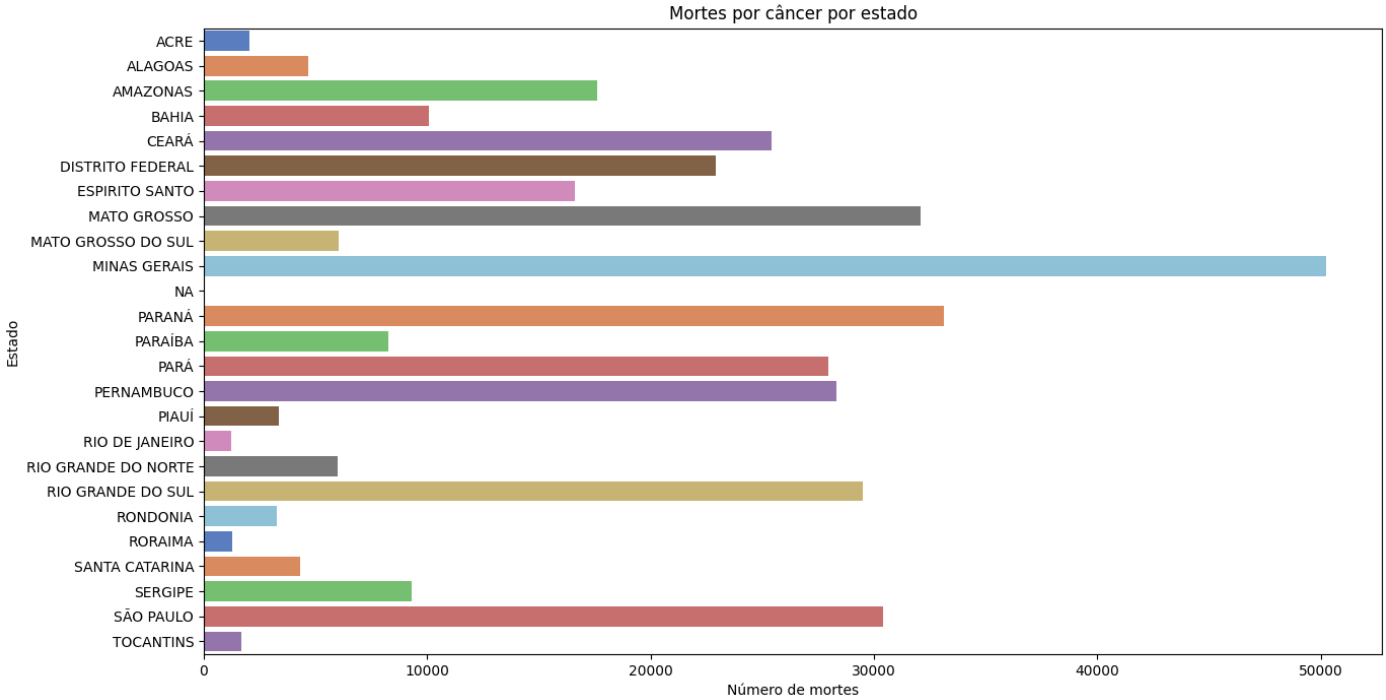
Os 20 principais cânceres que mais levam a óbito na faixa etária de 45-54 anos.



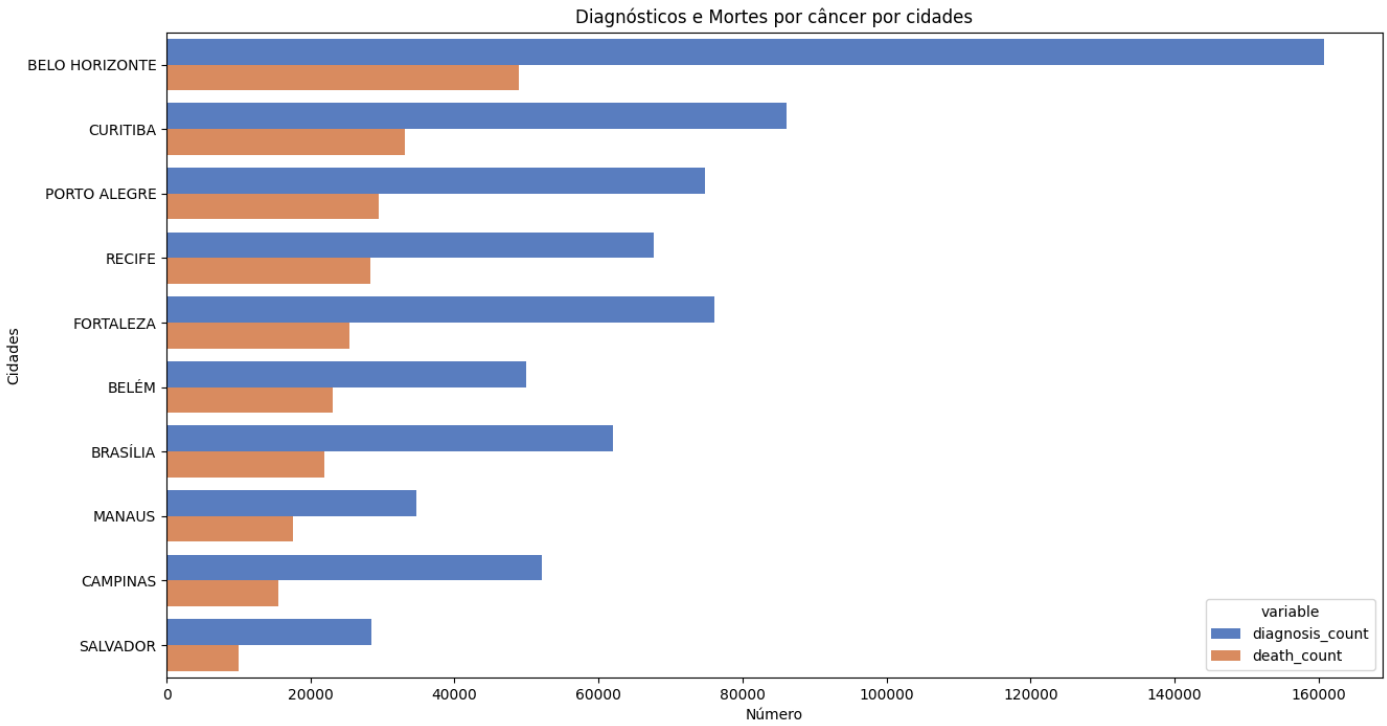
Os 20 principais cânceres que mais levam a óbito na faixa etária de 55-64 anos.



Mortes por câncer por estado, os principais estados para se ter atenção são, Minas Gerais, Paraná, Mato Grosso.



Diagnósticos e óbitos por câncer por cidades, preocupação principalmente para as cidades de Belo Horizonte pelos altos índices de diagnósticos e as cidades de Curitiba, Belém e Manaus por apresentarem uma alta taxa de mortalidade e relação a quantidade de diagnósticos.



Este projeto proporcionou uma oportunidade para aplicar e aprofundar conhecimentos em Ciência e Engenharia de Dados, ao abordar uma questão de importância global: a incidência de câncer. Foi proposto uma solução para um problema persistente e universal, usando recursos para analisar e entender as complexas relações entre os dados relacionados a essa doença.

A melhor compreensão dessas relações tem potencial para contribuir em muitas áreas da sociedade, oferecendo insights sobre o impacto do câncer na vida das pessoas e identificando estratégias eficazes para minimizar a incidência e prevenir as mortes associadas à doença. Os dados, quando corretamente interpretados, podem fornecer as ferramentas necessárias para enfrentar esta questão com uma resposta informada e baseada em evidências.

Em termos de desenvolvimento futuro deste projeto, vemos oportunidades para expandir nossa análise, estabelecendo mais conexões entre os dados disponíveis. Além disso, a automação de processos em cada etapa do projeto permitirá uma eficiência maior e garantirá a relevância contínua de nossas descobertas à medida que novos dados forem disponibilizados.