Relatório Técnico de Acompanhamento: Streamline ETL - Instituto CONDES - Grupo 06

2024-12-10

Contents

Introdução

Resumo Executivo

Este relatório documenta o progresso do projeto integrador Streamline ETL, desenvolvido pelo grupo 06. O objetivo principal é aplicar os conhecimentos adquiridos em sala de aula em um ambiente prático, proporcionando aprendizado em trabalho em equipe, metodologias ágeis e interação com stakeholders.

Por conta de complicações na etapa de comunicação com os representantes do Instituto CONDES, os trabalhos feitos nesta fase do projeto integrador foram utilizados dados da base de oncologia, disponibilizados através do portal DATASUS.

Principais destaques: - Adaptação de metodologias ágeis para melhor gestão de projetos. - Colaboração eficaz entre os membros dos grupos. - Desafios enfrentados e próximos passos a serem adotados.

Objetivos do Projeto

O Projeto Integrador tem como objetivo proporcionar aos alunos uma experiência prática e abrangente no campo da ciência de dados. Através da aplicação de conceitos teóricos em projetos reais, os estudantes têm a oportunidade de consolidar conhecimentos adquiridos durante o curso, enfrentando desafios que simulam situações do mercado. Isso inclui a manipulação de grandes volumes de dados, a utilização de ferramentas analíticas e a construção de modelos preditivos.

Além disso, o projeto busca promover o desenvolvimento de habilidades interpessoais como o trabalho em equipe e a comunicação eficaz. Em ambientes corporativos, a capacidade de colaborar com outros profissionais e comunicar insights de forma clara é tão importante quanto o domínio da técnica. Por isso, o Projeto Integrador incentiva a interação entre os participantes, estimulando a troca de ideias, a resolução conjunta de problemas e a divisão de responsabilidades.

Outro pilar fundamental do projeto é a adoção de metodologias ágeis. Essas metodologias permitem uma abordagem iterativa e incremental, garantindo a entrega contínua de valor e facilitando ajustes ao longo do caminho. A utilização dessas práticas assegura que o grupo esteja sempre alinhado com os objetivos do projeto, promovendo transparência e eficiência. Por fim, o projeto incentiva a interação direta com stakeholders. Isso inclui desde o levantamento de requisitos até a apresentação de soluções desenvolvidas. O objetivo é garantir que as entregas finais atendam às necessidades reais dos usuários ou clientes, criando impacto e agregando valor ao contexto em que o projeto está inserido.

O Projeto Streamline ETL tem como principal objetivo auxiliar as instituições de saúde a melhorarem tanto o funcionamento administrativo quanto o prático nas áreas relacionadas à oncologia. A ideia central é fornecer subsídios para que decisões mais acertadas possam ser tomadas, contribuindo diretamente para a eficiência dessas instituições nos setores que têm contato com câncer.

Para alcançar esse objetivo, é fundamental trabalhar os dados em conformidade com os interesses dos gestores, obedecendo a Lei Geral de Proteção de Dados (LGPD) e outros requisitos legais. Nesse contexto, a próxima fase do projeto, focada na criação de visualizações, será importante para aumentar o alcance desses dados, de forma que mais pessoas conseguirão compreender sobre os processos dos setores de oncologia. Nesse sentido, executivos e diretores poderão tomar mais decisões baseadas em dados.

Um dos focos é otimizar a alocação de recursos, identificando regiões ou populações com maior incidência de câncer para direcionar melhor as equipes médicas e equipamentos. Outro objetivo é reduzir custos de tratamento, utilizando dados históricos para prever despesas associadas a diferentes tipos de tratamento, permitindo uma gestão mais eficiente dos recursos financeiros.

Adicionalmente, o projeto buscará aumentar a taxa de sobrevivência, desenvolvendo insights que apoiem a criação de protocolos clínicos mais eficazes e intervenções mais rápidas. A análise de fatores de risco também será uma prioridade, explorando dados para identificar características específicas, como idade ou histórico familiar, que possam influenciar o desenvolvimento do câncer.

Por fim, o projeto visa aprimorar políticas de saúde pública e facilitar estudos epidemiológicos, oferecendo informações valiosas para campanhas de conscientização, triagens preventivas e subsídios para pesquisa. A automação da geração de relatórios também será uma ferramenta importante para garantir conformidade regulatória, agilizando auditorias e assegurando a transparência das instituições de saúde.

Grupo Alvo do projeto

O projeto é direcionado a vários perfis de usuários, cada um com necessidades específicas e objetivos diferentes no uso dos dados. Entre esses, destacam-se os profissionais de saúde, como médicos, enfermeiros e especialistas em oncologia, que precisam de dados bem organizados para tomar decisões clínicas mais informadas. Esses dados são essenciais para planejar tratamentos personalizados e monitorar a evolução dos pacientes, garantindo maior precisão e eficácia nas intervenções.

Outro público-alvo importante são os pesquisadores e cientistas de dados na área da saúde, que utilizam essas informações para realizar estudos epidemiológicos e identificar padrões de incidência de câncer. Com os dados adequados, esses profissionais podem desenvolver algoritmos preditivos que auxiliem na detecção precoce da doença, bem como melhorar os tratamentos disponíveis, promovendo avanços na área médica.

Além disso, os gestores e administradores de instituições de saúde dependem desses insights para otimizar a alocação de recursos, garantindo que as equipes e equipamentos sejam direcionados às áreas mais necessitadas. Eles também utilizam as análises para aprimorar a qualidade dos serviços prestados e controlar os custos associados ao tratamento do câncer, promovendo uma gestão mais eficiente e sustentável.

As secretarias e o Ministério da Saúde também são usuários críticos desses dados, pois precisam deles para formular políticas de saúde pública efetivas. Essas instituições utilizam os dados para alocar orçamentos de forma estratégica, implementar programas de prevenção e monitorar a eficácia das políticas de saúde em vigor, assegurando melhores resultados populacionais.

Por fim, os planos de Saúde e seguradoras podem explorar essas informações para avaliar riscos e custos associados ao tratamento oncológico. Esses dados permitem o desenvolvimento de planos de cobertura mais acessíveis e com foco preventivo, beneficiando tanto os pacientes quanto as empresas que prestam serviços de saúde. Dessa forma, os dados tratados pelo projeto atendem a uma ampla gama de necessidades, impactando positivamente diversas áreas do setor de saúde.

Estrutura do Grupo

O grupo que desenvolveu o projeto é composto por cinco integrantes, cada um trazendo experiências e habilidades complementares que contribuíram para o sucesso do trabalho. À frente do time estava

Pedro Gabriel, atuando como líder. Pedro trabalha no FNDE como assistente administrativo, cargo no qual desenvolveu habilidades organizacionais e de comunicação que foram fundamentais para o gerenciamento do projeto. Ele foi o principal ponto de contato com os stakeholders, coletando feedbacks e garantindo que as necessidades e expectativas fossem claramente compreendidas e repassadas para os demais integrantes. Sua liderança foi essencial para manter o grupo alinhado e focado nos objetivos.

Lucas Araujo, integrante do grupo, atualmente é estagiário de BI na BBTS, onde é responsável por desenvolver painéis interativos e informativos. Sua expertise em visualização de dados foi valiosa para criar dashboards e relatórios no projeto, ajudando a traduzir informações complexas em insights acessíveis e visuais para os usuários.

Outro integrante, **Felipe Martins**, atua como estagiário de Gestão de Dados no CGEE, onde é responsável pelo desenvolvimento de processos de ETL. Sua experiência prática com fluxos de dados foi um diferencial para o projeto, principalmente nas etapas de transformação e integração de dados. Felipe também trouxe conhecimento técnico adquirido em sua passagem pela América Tecnologia, onde trabalhou como estagiário em operações de suporte técnico, enriquecendo o grupo com uma visão ampla sobre infraestrutura e suporte.

Gabriel Barreto trouxe ao grupo sua vivência prática como estagiário em banco de dados nos Ministérios da Economia e no Ministério da Gestão e da Inovação em Serviços Públicos. Com essa bagagem, ele contribuiu significativamente para a modelagem e organização das bases de dados utilizadas no projeto, assegurando que as estruturas fossem otimizadas para análises eficientes.

Por fim, a equipe contou com **Milena Soares**, apesar de estar em início de carreira, Milena demonstrou grande potencial ao aplicar os conhecimentos adquiridos em sala de aula. Ela foi responsável por colaborar nas etapas de análise de dados e desenvolvimento de algoritmos, demonstrando um excelente domínio técnico e uma abordagem inovadora na solução de problemas.

Com uma equipe diversificada e altamente qualificada, o projeto beneficiou-se da soma das experiências profissionais e acadêmicas dos integrantes, permitindo o desenvolvimento de uma solução robusta e alinhada às demandas do setor.

Metodologia

Metodologias Utilizadas

Para desenvolver o projeto o grupo adotou a **metodologia ágil**, que foi amplamente utilizada para gerenciar tarefas e etapas dos processos. Com a aplicação de práticas como **Scrum** e **Kanban**, o time pôde dividir o trabalho em ciclos curtos e entregáveis, promovendo maior flexibilidade e capacidade de adaptação às mudanças.

O **Kanban** foi usado de forma específica para o gerenciamento visual das tarefas, com o auxílio da plataforma **Figma**. Essa abordagem permitiu que os grupos tivessem uma visão clara do progresso das atividades, identificando gargalos e garantindo a execução fluida das etapas planejadas. Outra metodologia relevante foi o **benchmarking**, que envolveu uma análise comparativa com ferramentas e soluções disponíveis no mercado. Esse processo foi essencial para identificar boas práticas e adaptar funcionalidades e estratégias de sucesso aos projetos, enriquecendo as soluções desenvolvidas.

Além disso, o grupo recebeu feedback contínuo, com sessões semanais de orientação e revisão do andamento do trabalho. Essas sessões permitiram ajustes constantes no escopo e nas estratégias, garantindo que o projeto persistisse alinhado aos objetivos iniciais e às expectativas dos stakeholders. Essas práticas não só asseguraram o progresso consistente do projeto, mas também fomentaram o aprendizado contínuo e a colaboração entre os participantes.

Resumo das Ferramentas Utilizadas

O desenvolvimento do projeto contou com o uso de diversas ferramentas, cuidadosamente escolhidas para atender às necessidades de comunicação, organização, colaboração, e execução técnica. Para facilitar a comunicação interna entre os membros da equipe, foi utilizado o **WhatsApp**, que garantiu a troca ágil de informações.

A organização das tarefas e a colaboração visual ficaram a cargo do **Figma**, que possibilitou o planejamento estruturado e uma visão compartilhada do progresso das atividades. O **GitHub** desempenhou um papel crucial no controle de versão e na documentação do projeto, assegurando que todas as alterações fossem registradas e que os arquivos estivessem acessíveis a todos os integrantes.

Para a parte técnica, o **Python** foi a principal linguagem utilizada para manipulação e transformação dos dados, proporcionando flexibilidade e eficiência nas etapas analíticas do projeto.

A elaboração do relatório final foi feita no **Google Docs** e também no **RMarkdown**. Foram construídos dois relatórios diferentes usando essas ferramentas por questões de flexibilidade e maior alcance dos documentos. O processo de ETL foi orquestrado com o **Airflow**, uma ferramenta essencial para organizar e automatizar as etapas do fluxo de dados. Para conseguir instalar e usar o **Airflow**, foi necessário subir um contêiner no **Docker**.

Por fim, o armazenamento seguro e escalável dos dados foi realizado no **Amazon S3**, que serviu como uma solução robusta para manter os dados na nuvem, facilitando o acesso e o gerenciamento de grandes volumes de informações. Essas ferramentas, integradas de maneira estratégica, foram fundamentais para o sucesso do projeto. Abaixo segue uma síntese de relação das ferramentas com o projeto.

Table 1: Tabela de Ferramentas Utilizadas

Nome_da_Ferramenta	Utilidade	Versão_da_Ferramenta
Whatsapp	Comunicação Interna	24.20.71
Figma	Organização de tarefas e colaboração visual	124.4.7
Github	Controle de versão e documentação	2.47.1
Python	Manipulação/Transformação dos dados	3.11.9
R	Elaboração do EDA	4.2.2
Google Docs	Elaboração do relatório	Atual
RMarkDown	Elaboração do relatório	2.29
Airflow	Orquestração do processo de ETL	2.9.0
Docker	Orquestração de contêiner do AIRFLOW	4.32
S3	Armazenamento (na nuvem) dos dados	Atual

Dicionário da Base de Dados

A base de dados escolhida é o Painel de Oncologia, este painel fornece informações detalhadas sobre diagnósticos, tratamentos e características demográficas dos pacientes atendidos no sistema de saúde pública do Brasil. A escolha dessa base de dados visa oferecer uma visão abrangente e confiável para análises sobre o atendimento oncológico no país. Abaixo segue o dicionário da base de dados.

Table 2: Tabela de Colunas do Dataset

Nome_da_Coluna	Descrição	Tipo
ano_diagn	Ano de diagnóstico no formato AAAA	Numérico
anomes_dia	Ano e mês do diagnóstico no formato AAAAMM	Numérico
ano_tratam	Ano do primeiro tratamento registrado no formato AAAA	Numérico
anomes_tra	Ano e mês do primeiro tratamento registrado no formato AAAAMM	Numérico
uf_resid	Cód. da Unidade de Federação de residência, conforme código IBGE	Numérico
mun_resid	Cód. da Unidade de Federação de residência + Cód. do Município de residência, conforme código IBGE	Numérico
uf_tratam	Cód. da Unidade de Federação do estabelecimento onde foi registrado o primeiro tratamento, conforme código IBGE	Numérico
mun_tratam	Cód. da Unidade de Federação de residência + Cód. do Município de residência do estabelecimento onde foi registrado o primeiro tratamento, conforme código IBGE	Numérico
uf_diagn	Cód. da Unidade de Federação do estabelecimento onde foi registrado o diagnóstico, conforme código IBGE	Numérico
mun_diag	Cód. da Unidade de Federação de residência + Cód. do Município de residência do estabelecimento onde foi registrado o diagnóstico, conforme código IBGE	Numérico
tratamento	Primeiro tratamento registrado: 1 - Cirurgia, 2 - Quimioterapia, 3 - Radioterapia, 4 - Quimioterapia + Radioterapia, 5 - Sem informação de tratamento	Numérico
diagnostic	Categorias de diagnóstico: 1 - Neoplasias Malignas, 2 - Neoplasias in situ, 3 - Neoplasias de comportamento incerto ou desconhecido, 4 - C44 e C73	Numérico
idade	Idade do(a) paciente no momento do diagnóstico, de 000 a 999, sendo 999 = idade ignorada	Numérico
sexo	Sexo do paciente: F - Feminino, M - Masculino	Textual
estadiam	Estadiamento registrado da doença: 0 - 0, 1 - I, 2 - II, 3 - III, 4 - IV, 5 - Não se aplica, 9 - Ignorado	Numérico
cnes_diag	Código do Estabelecimento no CNES onde foi registrado o diagnóstico	Numérico
cnes_trat	Código do Estabelecimento no CNES onde foi registrado o primeiro tratamento	Numérico
tempo_trat	Intervalo de tempo entre o diagnóstico e o primeiro tratamento registrados. Inclui indicador se o tratamento ocorreu antes ou depois do laudo de diagnóstico	Numérico
cns_pac	Cartão Nacional de Saúde	Numérico
diag_deth	Diagnóstico detalhado registrado, lista de CID-10 (C00 ao C97, D00 ao D09, D37 ao D48)	Textual
dt_diag	Data detalhada do diagnóstico no formato DD/MM/AAAA	Numérico
dt_trat	Data detalhada do primeiro tratamento no formato DD/MM/AAAA	Numérico
dt_nasc	Data de nascimento do paciente no formato DD/MM/AAAA	Numérico

Processo de ETL

O processo de ETL foi cuidadosamente projetado para garantir a qualidade, integridade e acessibilidade dos dados, permitindo a realização de análises consistentes e a geração de insights relevantes para o projeto.

Extração Os dados utilizados neste projeto foram extraídos diretamente do portal **DATASUS**, que disponibiliza informações relevantes sobre o sistema de saúde pública no Brasil. O arquivo correspondente ao Painel de Oncologia foi fornecido no formato .dbc (DataBase Container), um formato binário específico utilizado pelo DATASUS para compactação e distribuição de dados.

Para garantir a acessibilidade dos dados, foi necessário realizar a conversão do formato .dbc para o formato .csv, utilizando a linguagem Python e a biblioteca **dbc-reader (versão 1.0.5)**, especialmente projetada para leitura e manipulação desse tipo de arquivo. Após a conversão, os dados foram salvos em arquivos .csv e mesclados em um único arquivo, facilitando as etapas subsequentes de transformação e análise.

Transformação A etapa de transformação teve como objetivo preparar os dados extraídos para análises e visualizações. A base de dados bruta já apresenta um excelente nível de organização e tratamento, com estruturação adequada e valores padronizados. No entanto, para atender aos objetivos específicos do projeto, foram realizadas transformações adicionais que envolveram a adaptação e a criação de novas colunas, bem como ajustes em valores existentes.

Essas transformações foram implementadas utilizando a biblioteca pandas (versão 2.2.3) da linguagem **Python**, uma ferramenta robusta para manipulação e análise de dados. A flexibilidade do pandas permitiu realizar operações complexas de forma eficiente, mantendo a integridade e a consistência dos dados durante todo o processo.

O código usado foi colocado no apêndice. Nesse sentido segue descrição das etapas de transformação da base:

1. Transformamos a coluna **ANO_MESDIA**, que vinha com os valores representando o ano e o mes do diagnóstico (201809, 201811, ...), agora passou a representar somente o mes do diagnóstico e foi nomeada para **MES_DIAGN**.

2. Como nem todos fazem ou registram o tratamento do câncer, as colunas com informações relacionadas à tratamento podem conter valores nulos. Como é o caso da coluna **ANO_TRATAM**, nela optamos por preencher os valores nulos com **'DESCONHECIDO'**.

3. Na coluna **ANOMES_TRA**, aplicamos a mesma lógica da coluna **MES_DIAGN**, em que recortamos os valores e pegamos só o mês do início do tratamento. Para os valores nulos preenchemos também com **'DESCONHECIDO'**.

4. Em seguida, ignoramos as colunas que não possuem valores nulos, pois a base veio tratada de forma que certos valores inseridos em algumas colunas indicam valores faltantes (por exemplo, na coluna 'tratamento' o valor 5 indica 'sem informação de tratamento') nesse sentido, focamos nos valores nulos.

5. Nas colunas **UF_TRATAM** e **MUN_TRATAM** existem valores nulos, visto que nem todos iniciam o tratamento. Nesse sentido preenchemos eles com **'DESCONHECIDO'**.

6. Renomeamos as colunas referentes à diagnóstico e tratamento para um padrão. todas elas vão terminar com **DIAGN (para diagnóstico)** e **TRATAM (para tratamento)**, visto que havia uma mesclagem dessas nomenclaturas, o que poderia causar confusão no momento de manipulação dos dados.

7. Na coluna **IDADE**, quando o valor for 999 é porque a idade do paciente foi ignorada no momento do registro. Como foram poucos os registros com o valor 999 (certa de 10). Eles não teriam muito impacto. Alteramos esses valores para **DESCONHECIDO** para facilitar a manipulação de dados nessa coluna. Se um analista fosse calcular a média, esses valores 999 atrapalhariam.

8. Verificamos que a coluna **ESTADIAM** (estadiamento registrado da doença) tem valores nulos, fizemos o tratamento preenchendo os valores nulos por 9, que no dicionário de dados da base indica que é o valor correspondente ao ignorado. Além disso transformamos a coluna para inteiro, afim de dar uniformidade aos dados.

9. A coluna **CNES_TRATAM** (Código do estabelecimento do tratamento no cadastro nacional) tem valores nulos (já que nem todos fazem o tratamento da doença), fizemos o tratamento da base preenchendo os valores nulos com 0, depois passamos a coluna para inteiro (uniformidade) e depois substituimos onde havia 0 por **'DESCONHECIDO'**.

10. A coluna **TEMPO_TRATAM** tem valores nulos (já que nem todos fazem o tratamento) preenchemos esses valores com 99999, que no dicionário de dados é indicado ser o número para quando não há informação de tratamento.

11. Deletamos a coluna **CNS_PAC**, que é o identificador da **Carteira Nacional de Saúde**. Por questão de dados sensíveis, a coluna foi removida.

12. A coluna **DT_TRATAM** (data do início do tratamento) tem valores nulos, preenchemos eles como DESCONHECIDO.

Carga Na etapa de carga, os dados transformados foram armazenados em um bucket na **Amazon S3**, garantindo que estivessem disponíveis para consultas, análises e integrações futuras. Este armazenamento em nuvem foi escolhido por sua escalabilidade, confiabilidade e fácil integração com outras ferramentas de análise e visualização de dados.

A operação de carga foi realizada utilizando o **Apache Airflow**, uma ferramenta robusta de orquestração de workflows, que permitiu automatizar e monitorar o processo de ETL. O Airflow foi configurado para transferir os dados do ambiente local para o bucket S3 de forma eficiente, garantindo que todo o pipeline fosse executado de maneira consistente e sem interrupções.

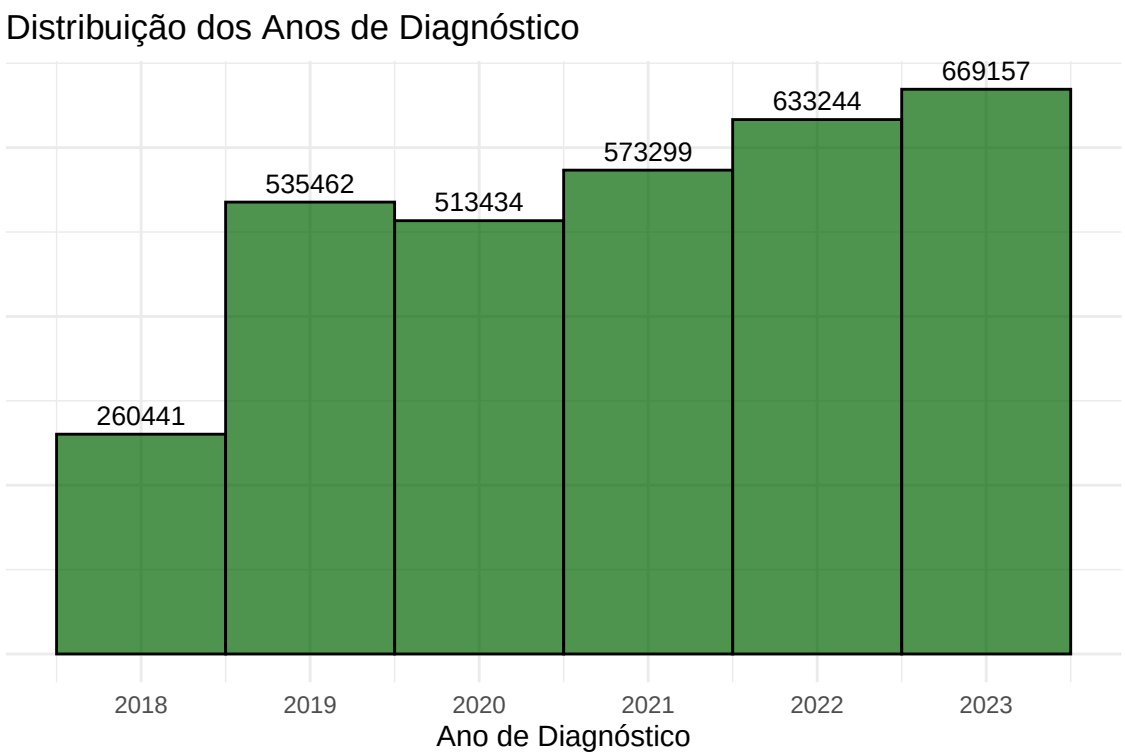
EDA

Realizamos uma análise exploratória de dados (EDA) para compreender melhor a estrutura e as características da base. Essa etapa foi essencial para identificar padrões, outliers e avaliar a distribuição das variáveis. A EDA nos permitiu validar a qualidade dos dados transformados e adquirir insights preliminares que orientaram as próximas etapas do projeto.

Para visualizar os dados e explorar suas distribuições, utilizamos a linguagem **R** e o pacote **ggplot2** (versão 3.5.1), que oferece ferramentas poderosas para a criação de gráficos e visualizações de alta

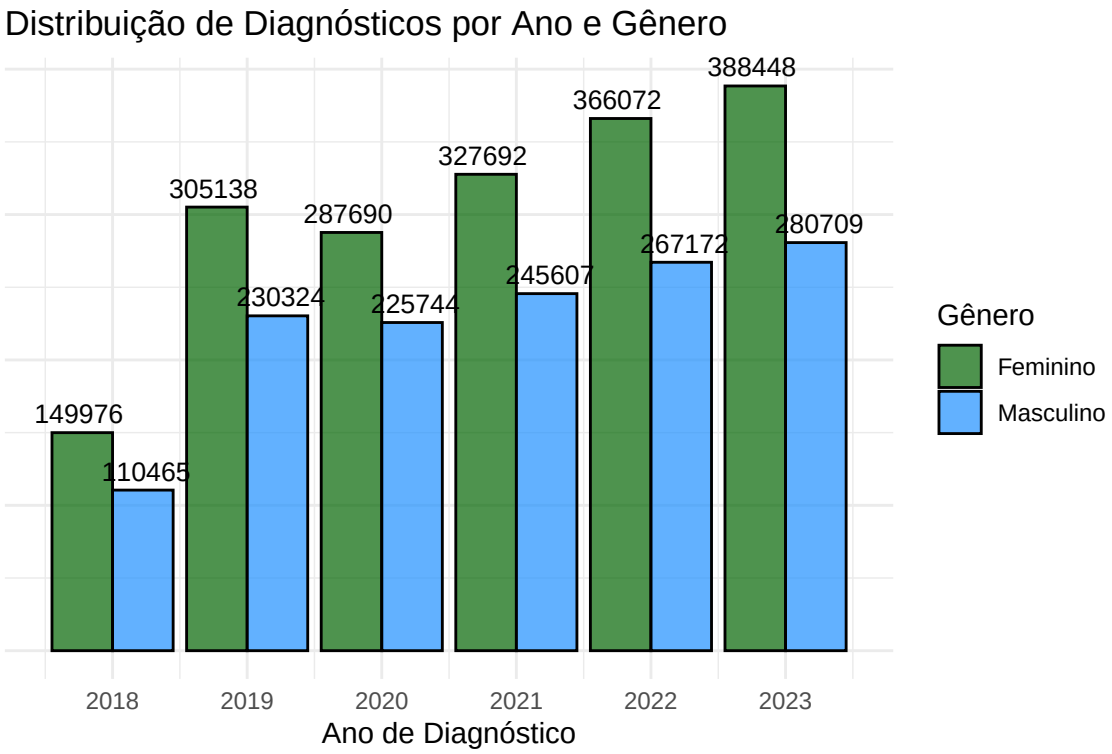
qualidade. Essas visualizações foram fundamentais para destacar tendências e peculiaridades na base de dados, permitindo uma interpretação mais intuitiva e informada.

1º Gráfico - Relação de anos com casos de câncer



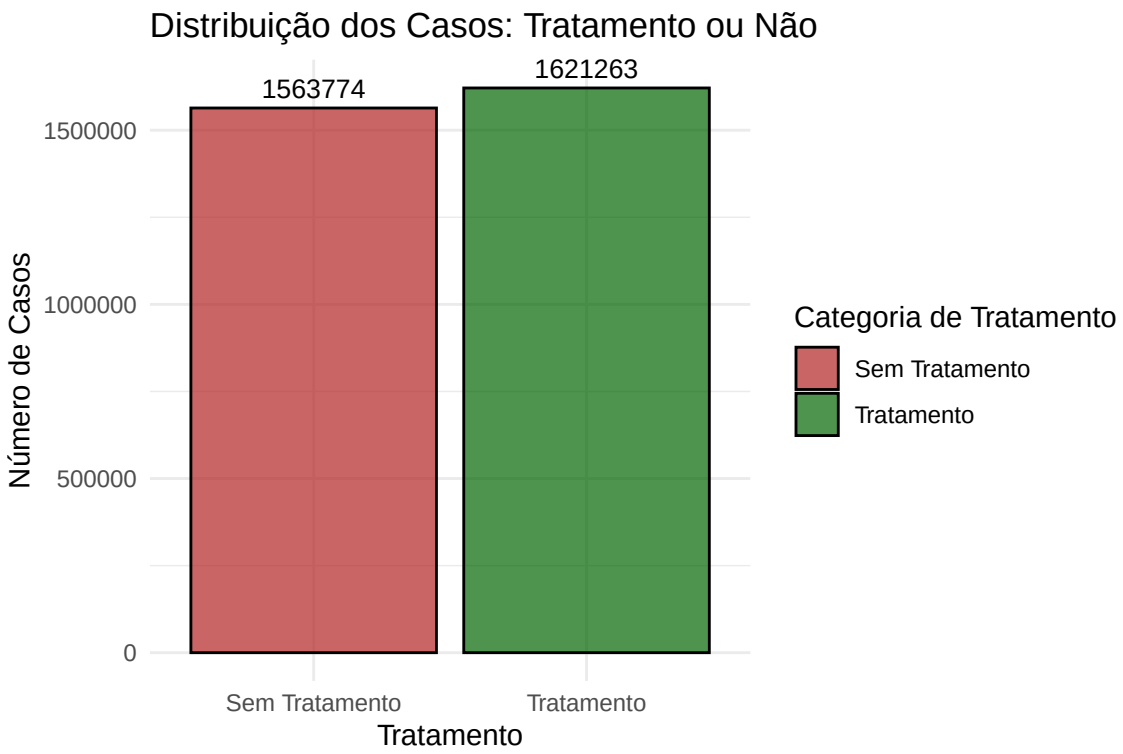
No gráfico acima vemos que há uma tendência de crescimento de casos de diagnóstico de câncer conforme os anos passam. Podemos ver um salto gritante do ano de 2018 para 2019, o que pode ser atribuido a falhas técnicas, pois não faz sentido esses casos terem dobrado de um ano para o outro. Além disso, percebemos uma queda no ano de 2019 a 2020 por conta da pandemia do COVID-19. Faz sentido dizer que como o vírus sobrecarregou o sistema de sáude e também forçou as pessoas a ficarem em casa, menos diagnósticos foram feitos.

2º Gráfico - Casos de câncer por ano e gênero



Analisando o gráfico acima constata-se que as mulheres têm mais câncer do que os homens, o que não condiz com a realidade. Segundo pesquisas, os homens têm maior chance de ter essa doença do que as mulheres. No gráfico o sexo feminino está em evidência e não podemos tirar conclusões disso.

3º Gráfico - Casos de câncer que foi registrado tratamento



Com esse gráfico vemos que muitos dos casos diagnosticados ficam sem tratamento. Isso se dá por conta de vários motivos, podemos associar o medo de não chegar à cura, os custos do tratamento, a dor que os exames e tratamentos pode causar, etc...

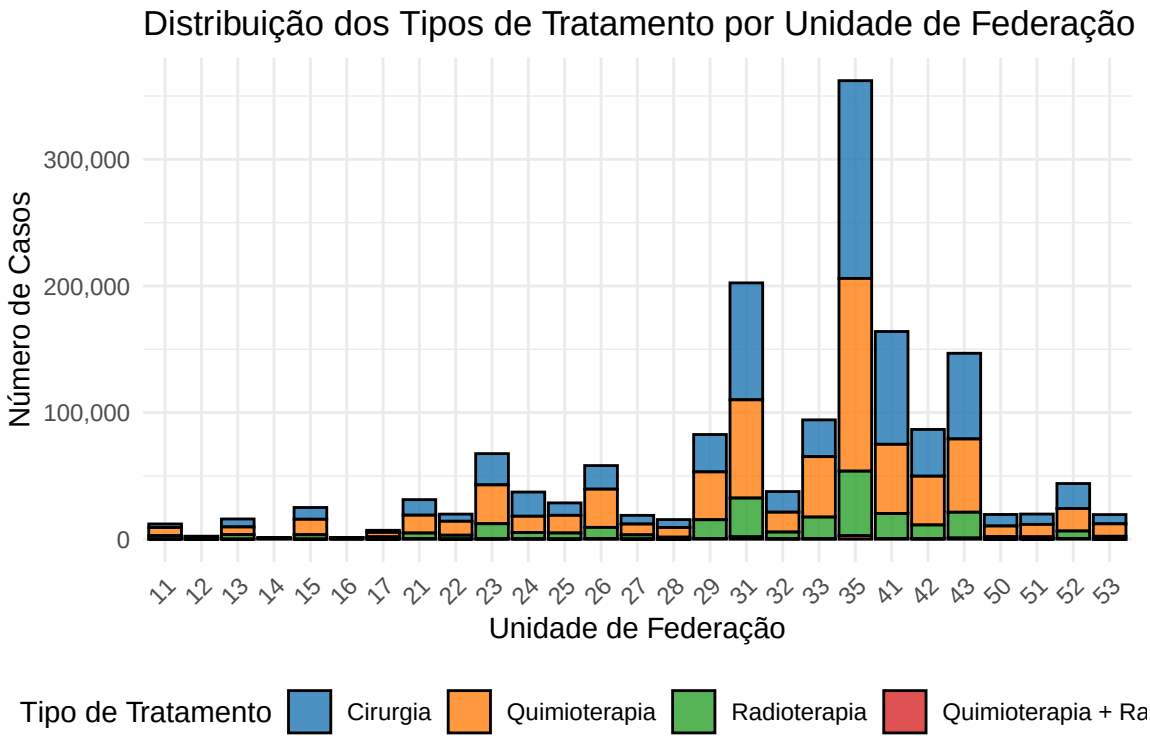
Resultados

Na seção de Resultados, apresentaremos as visualizações gráficas e interpretações das análises realizadas ao longo deste projeto. Os gráficos, gerados a partir dos dados tratados, permitem observar padrões e tendências importantes sobre as informações de diagnóstico e tratamento na base de oncologia. Além disso, serão discutidas as descobertas mais relevantes, destacando aspectos significativos observados nos dados, bem como as limitações do estudo, que podem incluir aspectos relacionados à qualidade dos dados, possíveis vieses nas informações ou restrições metodológicas durante o processo de análise.

Visualizações Gráficas

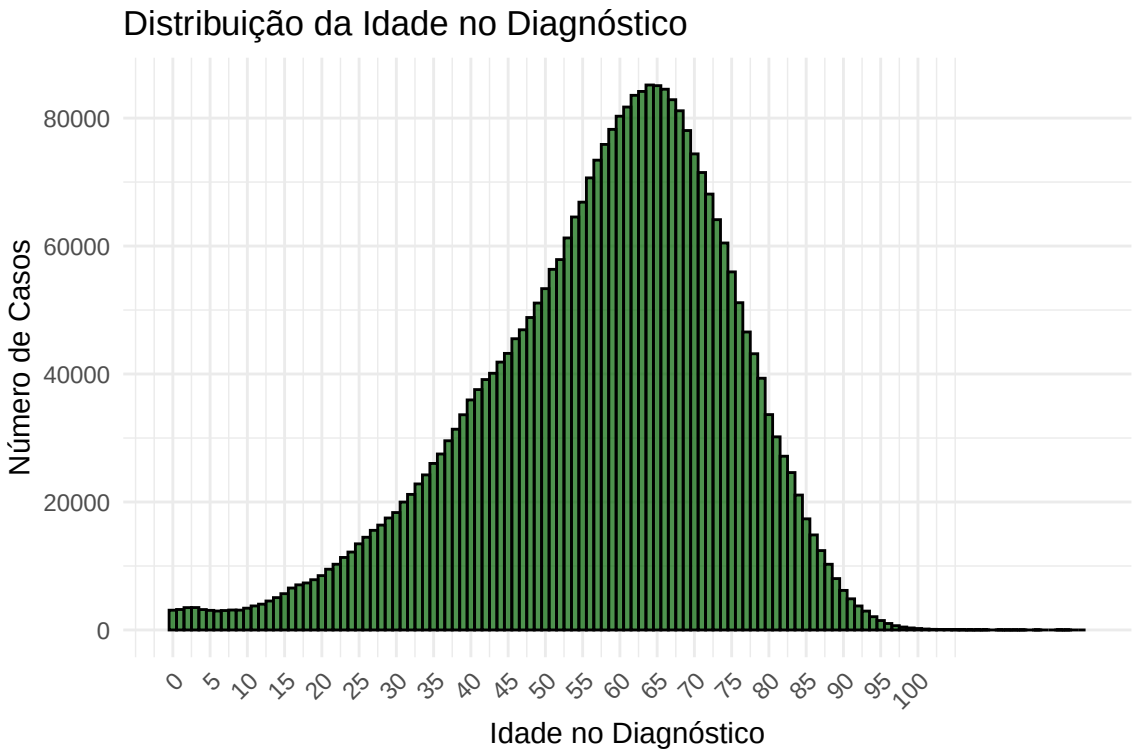
Desenvolvemos cinco gráficos mais elaboradoss em cima dos dados. A fim de tirarmos insights e descobrirmos padrões que o dataset contenha.

1º - Distribuição dos tipos de tratamento por unidade da federação



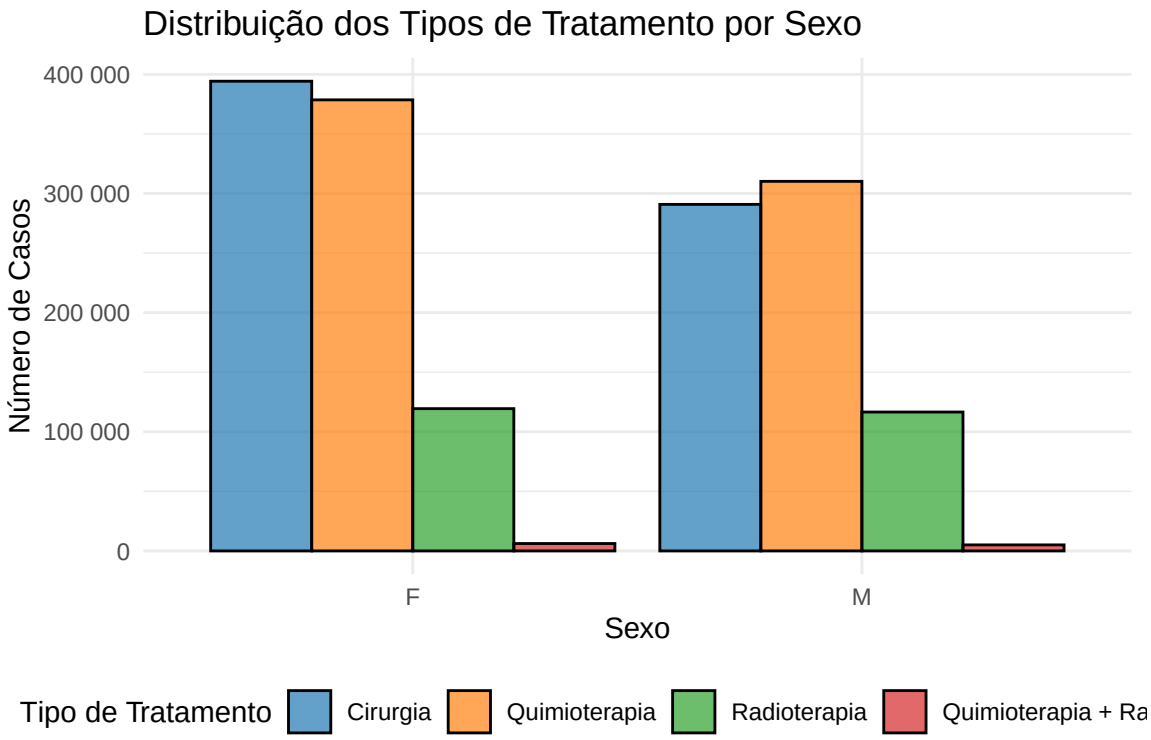
Neste primeiro gráfico vemos que a unidade federativa com maior número de casos registrados durante os anos de 2018 a 2023 foi a de SP (código 35), seguida de Minas Gerais (código 31) e depois o Paraná (código 41). Então os estados onde ocorre maior incidência da doença são da região sudeste e da sul. Pesquisas dizem que a maior causa do câncer está relacionada aos ambientes, com o ambiente em geral (água, terra e ar), ambiente ocupacional (indústrias químicas e afins) e o ambiente de consumo (alimentos, medicamentos). Cada um com os seus fatores de risco diferentes, nesse sentido, podemos fazer um paralelo com a qualidade de vida nos estados da região Sul e Sudeste, que por serem mais desenvolvidos e portanto, industrializados, acabam afetando mais os seus habitantes.

2º - Distribuição dos casos por idade



O gráfico de distribuição da idade no diagnóstico revela que a maior parte dos pacientes diagnosticados com câncer se concentra em faixas etárias mais avançadas, com uma distribuição próxima a uma curva normal. Observa-se um pico nas idades em torno de 65 a 70 anos, sugerindo que esta é uma faixa etária mais prevalente entre os casos diagnosticados. A distribuição apresenta uma leve assimetria à esquerda, indicando que há menos pacientes diagnosticados em idades mais jovens, com uma queda gradual nas extremidades. Esses padrões podem refletir fatores como o aumento do risco de câncer com a idade, bem como o tipo de câncer mais comum nas diferentes faixas etárias.

3º - Tratamentos por sexo



O gráfico revela diferenças notáveis na distribuição dos tipos de tratamento entre os sexos. Para o sexo feminino, observa-se que a cirurgia é o tratamento mais prevalente, superando a quimioterapia, enquanto a radioterapia mantém uma distribuição constante. Por outro lado, entre os pacientes do sexo masculino, a quimioterapia se destaca como o tratamento mais comum, com um número superior ao de cirurgias, e novamente, a radioterapia se apresenta de forma equilibrada em ambas as categorias de sexo. Essa análise sugere que as preferências ou necessidades de tratamento podem variar entre os sexos, com as mulheres tendo uma maior tendência a receber cirurgias e os homens mais frequentemente optando por tratamentos quimioterápicos.

Conclusão

Este projeto proporcionou uma análise profunda sobre o tratamento de pacientes diagnosticados com câncer, permitindo identificar padrões importantes, como a prevalência de certos tratamentos entre diferentes faixas etárias e sexos, além de destacar a relação entre o tipo de tratamento e os dados demográficos dos pacientes. A análise visual e estatística revelou insights valiosos, como a predominância de tratamentos como cirurgia em mulheres e quimioterapia em homens, além da distribuição das idades no momento do diagnóstico, que mostra uma maior concentração de casos em faixas etárias mais avançadas. Esses resultados não apenas ampliam o entendimento sobre as escolhas e resultados de tratamento, mas também oferecem uma base para aprimorar políticas de saúde e personalizar o tratamento para grupos específicos.

Como aprendizado, ficou claro que a análise de dados de saúde pode revelar tendências e padrões úteis, mas também apresenta desafios, como o tratamento de dados ausentes e a necessidade de uma manipulação cuidadosa para garantir a precisão das conclusões. A utilização de visualizações

gráficas foi fundamental para facilitar a compreensão de grandes volumes de dados e destacar tendências que não seriam facilmente observáveis apenas com estatísticas descritivas.

Para trabalhos futuros, seria interessante explorar mais a fundo a relação entre variáveis como a faixa de renda e os fatores que provocam câncer, buscando identificar outras influências no acesso e escolha de tratamentos. Além disso, a análise de dados de resultados de tratamento, como a taxa de sucesso e sobrevida, poderia fornecer informações ainda mais valiosas para a melhoria do tratamento oncológico. A combinação de dados clínicos com fatores psicossociais tam bém poderia oferecer uma visão mais holística do impacto das terapias, contribuindo para uma abordagem de tratamento mais personalizada.

Referências

- Portal do DATASUS. Disponível em https://datasus.saude.gov.br/transferencia-de-arquivos/#.
- Documentação do Amazon S3. Disponível em https://docs.aws.amazon.com/s3/.
- Documentação do Apache AIRFLOW. Disponível em: https://airflow.apache.org/docs/.
- DOCKER. Docker Documentation. Disponível em: https://docs.docker.com/.

Apêndice

Para conseguir reproduzir o código que trata a base, é preciso que você tenha o Docker instalado na sua máquina.

1. Com o Docker instalado, clone o repositório do projeto: https://github.com/GabrielTelles4K/StreamLitETL-ProjetoIntegrador.git
2. Certifique-se de que as pastas estão na mesma organização do repositório. Os arquivos .dbc precisam estar dentro da pasta etl_process_airflow/data/dados_brutos para que o processo funcione.
3. Abra um terminal e navegue até a pasta 'etl_process_airflow'.
4. Suba o contêiner usando:

```
docker compose up -d
```

5. Após o contêiner terminar de subir, navegue até a URL e inicie a DAG 'etl_streamline' clicando no botão com símbolo de 'play'.
6. Isso será o suficiente para fazer a transformação dos arquivos brutos e gerar a base final (que estará na pasta dados_transformados_finais, dentro de data). Para inserir os dados no S3, é necessário configurar uma conexão e fazer outras configurações, como criar uma conta para acesso ao bucket. Se o usuário preferir essa alternativa, favor informar no email: felipe_mmmachado@hotmail.com
7. Essa etapa de transformação foi resumida no notebook ETL.py, para ficar mais fácil a visualização do processo.
8. Os gráficos foram feitos no R, diretamente no arquivo do relatório RMarkdown. Segue código para replicá-los:

```
# Carregar o ggplot2
library(ggplot2)
# Criar o arquivo .rds com a base tratada
base_tratada_final <- readRDS("base_tratada_final.rds")

# Criar o gráfico de histograma para a coluna ANO_DIAGN
grafico_ano_diagn <- ggplot(base_tratada_final, aes(x = ANO_DIAGN)) +
  geom_histogram(binwidth = 1, fill = "#006400", color = "black", alpha = 0.7) +
  scale_x_continuous(breaks = seq(min(base_tratada_final$ANO_DIAGN, na.rm = TRUE),
                                  max(base_tratada_final$ANO_DIAGN, na.rm = TRUE), by = 1)) +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5, color = "black", size = 3.5)
  labs(
    title = "Distribuição dos Anos de Diagnóstico",
    x = "Ano de Diagnóstico",
    y = NULL # Remove o rótulo do eixo Y
  ) +
  theme_minimal() +
  theme(
    axis.text.y = element_blank(),  # Remove os valores do eixo Y
    axis.ticks.y = element_blank()  # Remove as marcas do eixo Y
  )

# Exibir o gráfico
print(grafico_ano_diagn)
```

```
# Criar o gráfico de barras agrupadas para a coluna SEXO por ano
grafico_sexo_ano <- ggplot(base_tratada_final, aes(x = ANO_DIAGN, fill = SEXO)) +
  geom_bar(position = "dodge", color = "black", alpha = 0.7) +
  scale_fill_manual(values = c("F" = "#006400", "M" = "#1E90FF"),
                    labels = c("Feminino", "Masculino")) + # Cores distintas para gêneros
  scale_x_continuous(breaks = seq(min(base_tratada_final$ANO_DIAGN, na.rm = TRUE),
                                  max(base_tratada_final$ANO_DIAGN, na.rm = TRUE), by = 1)) +
  geom_text(stat = "count", aes(label = ..count..),
            position = position_dodge(width = 1),
            vjust = -0.5, color = "black", size = 3.5) +
  labs(
    title = "Distribuição de Diagnósticos por Ano e Gênero",
    x = "Ano de Diagnóstico",
    y = NULL, # Remove o rótulo do eixo Y
    fill = "Gênero" # Legenda da cor
  ) +
  theme_minimal() +
  theme(
    axis.text.y = element_blank(),  # Remove os valores do eixo Y
    axis.ticks.y = element_blank()  # Remove as marcas do eixo Y
```

```
  )

# Exibir o gráfico
print(grafico_sexo_ano)
```

```
# Criar uma nova variável para representar a categoria de tratamento
base_tratada_final$tratamento_categoria <- ifelse(base_tratada_final$TRATAMENTO == 5, "Sem Trat

# Criar o gráfico de barras para a coluna TRATAMENTO
grafico_tratamento <- ggplot(base_tratada_final, aes(x = tratamento_categoria, fill = tratament
  geom_bar(color = "black", alpha = 0.7) +
  scale_fill_manual(values = c("Tratamento" = "#006400", "Sem Tratamento" = "#B22222")) +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5, color = "black", size = 3.5)
  labs(
    title = "Distribuição dos Casos: Tratamento ou Não",
    x = "Tratamento",
    y = "Número de Casos",
    fill = "Categoria de Tratamento"
  ) +
  theme_minimal()

# Exibir o gráfico
print(grafico_tratamento)
```

```
library(dplyr)
library(scales) # Para formatação dos números

# Filtrar apenas os registros com tratamento
base_tratada_final_tratamento <- base_tratada_final %>%
  filter(TRATAMENTO != 5) # Excluir registros sem tratamento

# Criar o gráfico de barras empilhadas por Unidade de Federação (UF_TRATAM)
grafico_uf_tratamento <- ggplot(base_tratada_final_tratamento, aes(x = UF_TRATAM, fill = as.fa
  geom_bar(position = "stack", color = "black", alpha = 0.8) +
  scale_fill_manual(
    values = c(
      "1" = "#1f77b4", # Cirurgia
      "2" = "#ff7f0e", # Quimioterapia
      "3" = "#2ca02c", # Radioterapia
      "4" = "#d62728"  # Quimio + Radio
    ),
    labels = c(
      "1" = "Cirurgia",
```

```
      "2" = "Quimioterapia",
      "3" = "Radioterapia",
      "4" = "Quimioterapia + Radioterapia"
    )
  ) +
  scale_y_continuous(labels = label_comma()) + # Formatar os números no eixo Y
  labs(
    title = "Distribuição dos Tipos de Tratamento por Unidade de Federação",
    x = "Unidade de Federação",
    y = "Número de Casos",
    fill = "Tipo de Tratamento"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1), # Girar os rótulos do eixo X
    legend.position = "bottom"
  )

# Exibir o gráfico
print(grafico_uf_tratamento)
```

```
# Substituir valores "DESCONHECIDO" por NA
base_tratada_final$IDADE[base_tratada_final$IDADE == "DESCONHECIDO"] <- NA

# Converter a coluna IDADE para numérica
base_tratada_final$IDADE <- as.numeric(base_tratada_final$IDADE)

# Criar o gráfico de histograma para a distribuição de idades no diagnóstico
grafico_idade_diagn <- ggplot(base_tratada_final, aes(x = IDADE)) +
  geom_histogram(binwidth = 1, fill = "#006400", color = "black", alpha = 0.7) +
  labs(
    title = "Distribuição da Idade no Diagnóstico",
    x = "Idade no Diagnóstico",
    y = "Número de Casos"
  ) +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 100, by = 5)) +  # Ajuste os intervalos no eixo X
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotacionar os números do eixo X

# Exibir o gráfico
print(grafico_idade_diagn)
```

```
library(dplyr)
```

```
library(scales)

# Filtrando os dados, mantendo os tratamentos válidos
base_tratada_final_tratamento <- base_tratada_final %>%
  filter(TRATAMENTO != 5)  # Exclui registros sem tratamento

# Gráfico de barras agrupadas para distribuição de tratamento por sexo
grafico_sexo_tratamento <- ggplot(base_tratada_final_tratamento, aes(x = SEXO, fill = as.factor
  geom_bar(position = "dodge", color = "black", alpha = 0.7) +
  scale_fill_manual(
    values = c(
      "1" = "#1f77b4", # Cirurgia
      "2" = "#ff7f0e", # Quimioterapia
      "3" = "#2ca02c", # Radioterapia
      "4" = "#d62728"  # Quimioterapia + Radioterapia
    ),
    labels = c(
      "1" = "Cirurgia",
      "2" = "Quimioterapia",
      "3" = "Radioterapia",
      "4" = "Quimioterapia + Radioterapia"
    )
  ) +
  labs(
    title = "Distribuição dos Tipos de Tratamento por Sexo",
    x = "Sexo",
    y = "Número de Casos",
    fill = "Tipo de Tratamento"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 0, hjust = 0.5), # Ajustar rótulos do eixo X
    legend.position = "bottom"
  ) +
  scale_y_continuous(labels = label_number(scale = 1))  # Remove a notação científica

# Exibir o gráfico
print(grafico_sexo_tratamento)
```