98H07-04 Visualização de Dados

Professora Isabel Harb Manssour

Especificação do Projeto Final da Disciplina

Descrição Geral do Projeto

Inicialmente, a turma deverá se organizar em 3 grupos, cada um responsável por desenvolver a solução de um problema distinto. Para isso, foi disponibilizado no Moodle um recurso de "escolha de grupo". **Cada aluno deve obrigatoriamente selecionar o tema com o qual irá trabalhar**.

O objetivo principal deste projeto é proporcionar aos alunos a oportunidade de desenvolver soluções de visualização de dados interativas e analíticas que contemplem o mapeamento dinâmico de demanda e de mercado na pesquisa clínica, atendendo às necessidades estratégicas da startup Let Me Trial.

Os objetivos específicos são:

- Integrar dados públicos (CNES, ANS, DATASUS) para mapeamento georreferenciado da demanda de especialidades e dos centros de pesquisa no Brasil;
- Construir *dashboards* que facilitem o monitoramento de mercado, incluindo operadoras de saúde, beneficiários, indústria biofarmacêutica e potenciais parceiros estratégicos;
- Estimular a aprendizagem ativa e multidisciplinar dos alunos por meio da resolução de problemas reais apresentados pela startup;
- Promover o desenvolvimento de soluções com potencial impacto social, econômico e de saúde para os usuários finais da Let Me Trial.

Cada grupo deverá abordar um dos seguintes temas:

1. **Mapeamento Dinâmico da Demanda de estudos clínicos hematologia e oncologia**
   Descrição/Objetivo: Buscar por estudos clínicos de hematologia e oncologia que estão sendo realizados no Brasil e identificar Centros de Assistência de Alta Complexidade em Oncologia (CACONs), Unidades de Assistência de Alta Complexidade em Oncologia (UNACONs), hospitais, centros de pesquisa no Brasil que possam atender às demandas destes estudos.
   Dados: Serão fornecidos arquivos contendo dados sobre os estudos clínicos com recrutamento no Brasil, registrados na plataforma https://clinicaltrials.gov/ e relacionados a 'cancer' e 'oncology''. Estes arquivos terão dados como número de protocolo, instituição facilitadora, status da pesquisa, cidade, estado, região, zip, país, contatos e geolocalização. Além disso, é preciso "enriquecer" estes dados com informações dos estabelecimentos de saúde e dos centros de pesquisas clínicas ativos no Brasil (verificar como obter estes dados no final deste documento no itens "II) Dados dos estabelecimentos de saúde do Brasil" e "III) Dados dos centros de pesquisa clínica ativos no Brasil").
   Questões iniciais: Quais estados possuem mais estudos clínicos de hematologia e oncologia? Onde há uma maior carência de centros de assistência e de pesquisa que possam dar suporte a estes estudos clínicos?

2. **Mapeamento Dinâmico da Demanda de estudos clínicos outras especialidades reumatologia, cardiologia, dermatologia, neurologia.**
   Descrição/Objetivo: Buscar por estudos clínicos de reumatologia, cardiologia, dermatologia e neurologia que estão sendo realizados no Brasil e identificar hospitais e centros de pesquisa no Brasil que possam atender às demandas destes estudos.

Dados: Serão fornecidos arquivos contendo dados sobre os estudos clínicos com recrutamento no Brasil, registrados na plataforma https://clinicaltrials.gov/ e relacionados a reumatologia, cardiologia, dermatologia e neurologia. Estes arquivos terão dados como número de protocolo, instituição facilitadora, status da pesquisa, cidade, estado, região, zip, país, contatos e geolocalização. Além disso, é preciso "enriquecer" estes dados com informações dos estabelecimentos de saúde e dos centros de pesquisas clínicas ativos no Brasil (verificar como obter estes dados no final deste documento no itens "II) Dados dos estabelecimentos de saúde do Brasil" e "III) Dados dos centros de pesquisa clínica ativos no Brasil").
Questões iniciais: Como está a distribuição dos estudos clínicos de cada especialidade no Brasil? Onde há uma maior carência de hospitais e centros de pesquisa que possam dar suporte a cada um destes estudos clínicos?

3. **Mapeamento Dinâmico de Mercado: operadoras de plano de saúde, indústria biofarmacêutica, Academic Research Organization (ex.: HCOR), Clinical Research Organization e centros de pesquisa**
Descrição/Objetivo: Buscar operadoras de saúde, indústrias biofarmacêuticas e centros de pesquisa no Brasil, independentemente da especialidade, para analisar regiões com mais infraestrutura para realização dos estudos clínicos.
Dados: Buscar por dados: (1) das operadoras de plano de saúde através da ANS TabNet; (2) dos estabelecimentos de saúde do Brasil através do CNES; (3) dos centros de pesquisa no Brasil; e (4) das indústrias biofarmacêuticas. Informações detalhadas para a coleta destes dados estão no final deste documento.
Questões iniciais: Onde se concentram as operadoras e qual o percentual de mercado por município e estado de pessoas com plano de saúde, indústrias e centros de pesquisa? Quais possuem mais hospitais e centros de atendimento de alta complexidade próximos?

O detalhamento de cada tema, bem como o processo de coleta de dados, será apresentado pelos proponentes na aula do dia 04/09, mas os entregáveis de cada parte do projeto são apresentados a seguir.

Parte I

Descrição

A primeira parte do projeto, simplificadamente, consiste na coleta, análise e preparação dos dados; na definição das perguntas de negócio; na especificação do ambiente de desenvolvimento; e no projeto da solução.

Considerando o processo CRISP-DM, o *pipeline* de visualização de dados e o processo de *visual analytics*, nesta primeira parte do desenvolvimento do projeto devem ser seguidas as seguintes etapas:

1. Entendimento do negócio e do tema proposto;
2. Coleta dos dados;
3. Entendimento dos dados através de uma análise exploratória visual;
4. Preparação e pré-processamento de dados, que pode também incluir o enriquecimento dos dados se necessário;
5. Estudo do ambiente de desenvolvimento e projeto das representações visuais que serão implementadas.

A coleta de dados deverá ser feita através de um programa na linguagem de programação Python. Depois, os dados coletados deverão ser analisados, pré-processados e enriquecidos. Espera-se que sejam utilizadas as ferramentas apresentadas nas aulas (por exemplo, ydata-profiling) para auxiliar a analisar as

características e a qualidade dos dados, e a necessidade de fazer algum pré-processamento ou enriquecimento dos dados.

A partir do tema proposto e da coleta e análise dos dados, devem ser definidas as perguntas de negócio, ou seja, o que se quer descobrir a partir das visualizações propostas. Algumas perguntas já foram apresentadas, mas cada grupo deve propor ao menos duas questões adicionais. Não há um número máximo de perguntas, cada grupo deve elaborar a quantidade de perguntas que achar mais adequada.

Também deve ser escolhido e estudado o ambiente de desenvolvimento do *dashboard* que será criado na parte II do projeto. É obrigatório o uso de um ambiente gratuito, como, por exemplo, o Streamlit (https://streamlit.io/) ou o Dash (https://dash.plotly.com/).

A última etapa da parte I do projeto consiste na escolha, justificativa e descrição das visualizações interativas que serão implementadas na parte II do projeto. Cada grupo deve escolher diferentes formas de visualização (*heatmaps*, gráficos de linha, *small multiples*, etc.), **justificar** as escolhas e descrever como os dados serão utilizados/mapeados, identificando qual pergunta de negócio responde e porque a técnica de visualização escolhida é adequada para analisar os dados.

Os entregáveis da parte I do projeto são:

- **Relatório com as seguintes informações**: (1) Nomes dos componentes do grupo; (2) descrição do conjunto de dados, dos *insights* obtidos com a análise exploratória visual (incluindo gráficos) e do enriquecimento dos dados (se realizado); (3) apresentação das perguntas de negócio; (4) ambiente de desenvolvimento escolhido; (5) descrição das visualizações e interações que serão implementadas com uma **justificativa** para escolha de cada uma e os dados que serão utilizados. Não há um formato pré-definido para o relatório, mas ele deve ser um documento escrito (não uma apresentação de PowerPoint) que deve conter todas as informações especificadas apresentadas de uma forma objetiva. Cada item descrito acima deve ser colocado como um título. Recomenda-se o uso de folha tamanho A4 com orientação retrato, uma coluna, fonte tamanho 12 e espaçamento simples de parágrafo. Títulos devem estar com uma fonte maior e em negrito.
- ***Pitch***: o grupo deve preparar e apresentar um *pitch* de 12 a 15 minutos. O objetivo é mostrar a análise dos dados, apresentar as perguntas de negócio e as vantagens da solução proposta para o problema apresentado, justificando as escolhas feitas.

Além disso, na aula de *checkpoint* é obrigatório que cada grupo apresente os dados já coletados para a professora e discuta a solução que será proposta. O grupo que não estiver presente ou não estiver com os dados coletados e prontos para a análise terá um desconto de 1 ponto na nota do trabalho.

Datas

Checkpoint para apresentação dos dados e discussão da solução: 25/09/2025 (no horário de aula)

Apresentação do *Pitch* e entrega do relatório no Moodle: 07/10/2025 (durante o horário de aula).

Critérios de avaliação

Serão considerados para avaliação da parte I os seguintes critérios:

- *Checkpoint*: dados já coletados para discussão da solução.
- Relatório: inclusão de todo conteúdo solicitado; relevância e profundidade da análise dos dados; relevância, originalidade, eficácia e **justificativa** das visualizações interativas propostas; capacidade de síntese e escrita.
- *Pitch*: qualidade da apresentação (clara e fluente); respeito ao tempo estabelecido; capacidade de síntese para apresentação da solução proposta e justificativa das escolhas.

Parte II

Descrição

A segunda parte do projeto consiste na implementação, teste e validação da solução proposta na parte I. Considerando o processo CRISP-DM, o *pipeline* de visualização de dados e o processo de *visual analytics*, inicialmente deve ser feito o mapeamento dos dados para as representações visuais especificadas na parte I. A implementação das visualizações e interações projetadas deve ser feita usando o ambiente de desenvolvimento escolhido. Estas visualizações interativas devem ser organizadas, obrigatoriamente, em um ou mais *dashboards* e permitir a exploração dos dados de maneira efetiva e intuitiva, possibilitando responder as perguntas de negócio especificadas. É importante explorar diferentes formas de visualização, como, por exemplo, *heatmaps*, mapa de símbolos, e *small multiples*, além de outros gráficos que possam apresentar informações complementares, tais como informações estatísticas. Também devem ser consideradas as boas práticas que serão vistas em aula para o desenvolvimento das visualizações e do(s) *dashboard*(*s*), incluindo os *design patterns* para *dashboards* (https://dashboarddesignpatterns.github.io/).

Depois, a solução implementada deve ser testada e avaliada para verificar se responde a todas as perguntas de negócio definidas na parte I do projeto. Nesta etapa pode ser necessário fazer alguns ajustes nas visualizações interativas propostas, pois é importante que elas possibilitem extrair as informações necessárias, permitindo solucionar os problemas propostos.

Os entregáveis da parte II do projeto são:

- ***Link* e código da solução proposta**: a solução deve estar disponível *online*, o código deve estar no Github, e os *links* devem ser incluídos no relatório.
- **Relatório com as seguintes informações**: (1) Nomes dos componentes do grupo; (2) descrição das visualizações e interações implementadas, demonstrando a facilidade de uso das mesmas e os padrões utilizados para o desenvolvimento do(s) *dashboard(s)*; (3) análise dos dados utilizando as visualizações implementadas, incluindo as respostas para as perguntas de negócio e os *insights* obtidos com o uso da solução implementada; (4) discussão dos benefícios e limitações da solução proposta e apresentação de possíveis melhorias como trabalho futuro; (5) *links* para a solução implementada e para o código disponível no Github. Não há um formato pré-definido para o relatório, mas ele deve ser um documento escrito (não uma apresentação de PowerPoint) e todas as informações especificadas devem ser apresentadas de uma forma objetiva. Recomenda-se o uso de folha tamanho A4 com orientação retrato, uma coluna, fonte tamanho 12 e espaçamento simples de parágrafo. Títulos devem estar com uma fonte maior e em negrito.
- ***Pitch***: o grupo deve preparar e apresentar um *pitch* de 12 a 15 minutos. O objetivo é mostrar o funcionamento, as vantagens e a inovação da solução implementada, bem como os resultados obtidos (respostas das perguntas de negócio e *insights* identificados). As limitações da solução também devem ser abordadas.

Além disso, na aula de *checkpoint* é obrigatório que cada grupo apresente as visualizações que estão sendo implementadas para a professora e discuta como será a elaboração do *dashboard* final. O grupo que não estiver presente ou não estiver com algumas visualizações já implementadas terá um desconto de 1 ponto na nota do trabalho.

Datas

Checkpoint para apresentação das visualizações e discussão do *dashboard*: 06/11/2025 (no horário de aula)

Apresentação do *Pitch* e entrega do relatório no Moodle: 27/11/2025 (durante o horário de aula).

Critérios de avaliação

Serão considerados para avaliação da parte II os seguintes critérios:

- *Checkpoint*: visualizações já implementadas para discussão sobre a elaboração do *dashboard* final.
- Solução implementada: qualidade visual seguindo as boas práticas estudadas, variedade de representações visuais interativas, funcionalidades desenvolvidas e *insights* obtidos.
- Relatório: inclusão de todo conteúdo solicitado; relevância, originalidade e eficácia das visualizações interativas propostas; relevância e profundidade da análise dos resultados obtidos; criatividade e qualidade da solução proposta; capacidade de síntese e escrita.
- *Pitch*: qualidade da apresentação (clara e fluente); respeito ao tempo estabelecido; capacidade de síntese para demonstração da solução proposta e dos *insights* obtidos.

Informações sobre as coletas de dados

I) Dados das operadoras de plano de saúde

Nome do site: ANS TabNet — Informações em Saúde Suplementar (Agência Nacional de Saúde Suplementar).
URL inicial: https://www.ans.gov.br/anstabnet/index.htm

Passos

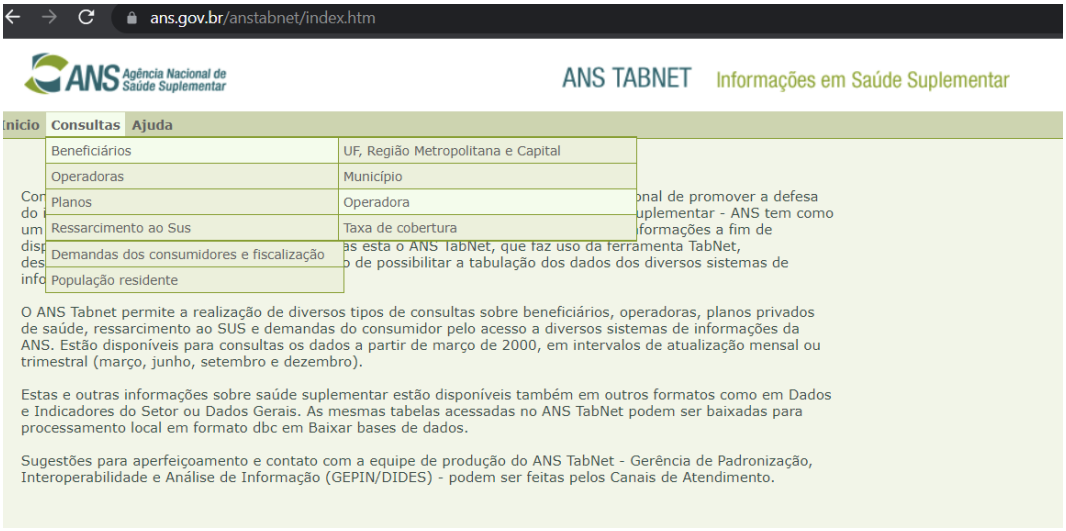
1. **Entrar no ANS TabNet:** Abra a página acima. No menu superior, passe o mouse em **Consultas** e clique em **Operadoras** (como na imagem).
2. **Ir para a tela de tabulação:** Você será levado à interface TabNet das operadoras: https://www.ans.gov.br/anstabnet/cgi-bin/dh?dados/tabnet_cc.def
   (É a página de seleção de parâmetros para gerar tabelas dinâmicas.)
3. **Configurar a tabulação**
   - **Linha:** selecione **Operadora**.
   - **Coluna:** selecione **UF**.
   - **Conteúdo:** selecione **Benef. assist. médica** (beneficiários de planos médico-hospitalares).
   - **Períodos disponíveis:** marque **Jun/2025** (ou o período mais atual)
4. **Formato da saída**
   - Em **Formato**, escolha "**Tabela com bordas"**.
5. **Gerar a tabela**
   - Clique em **Mostrar** (ou **Processar**) para exibir os resultados.
6. **Exportar para CSV**
   - Abaixo/ao lado da tabela, clique em **Copiar CSV** (no site pode aparecer como **"cópia em .CSV"**).
   - Salve o arquivo com extensão **.csv**.

Checagem final

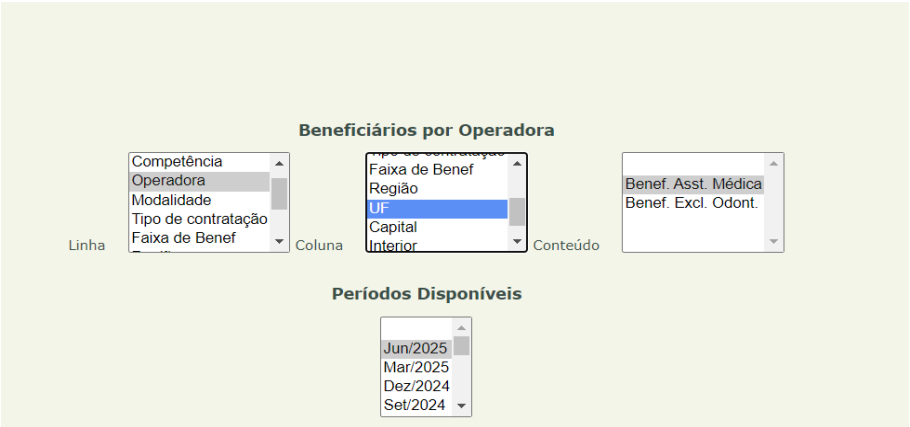
- Linha = Operadora
- Coluna = UF
- Conteúdo = Benef. assist. médica
- Período = Jun/2025 (ou o período mais atual)

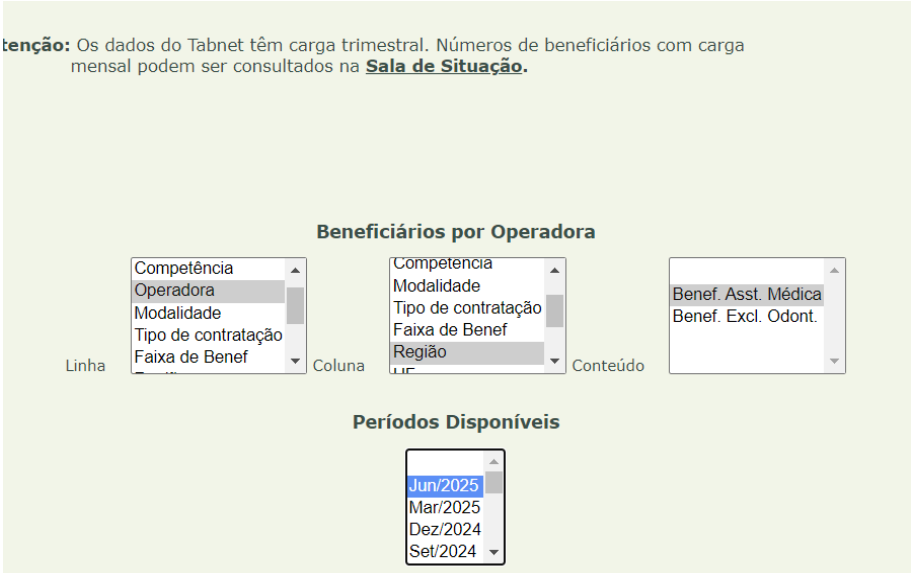
- Formato = Tabela com bordas
- Exportado via **Copiar CSV**

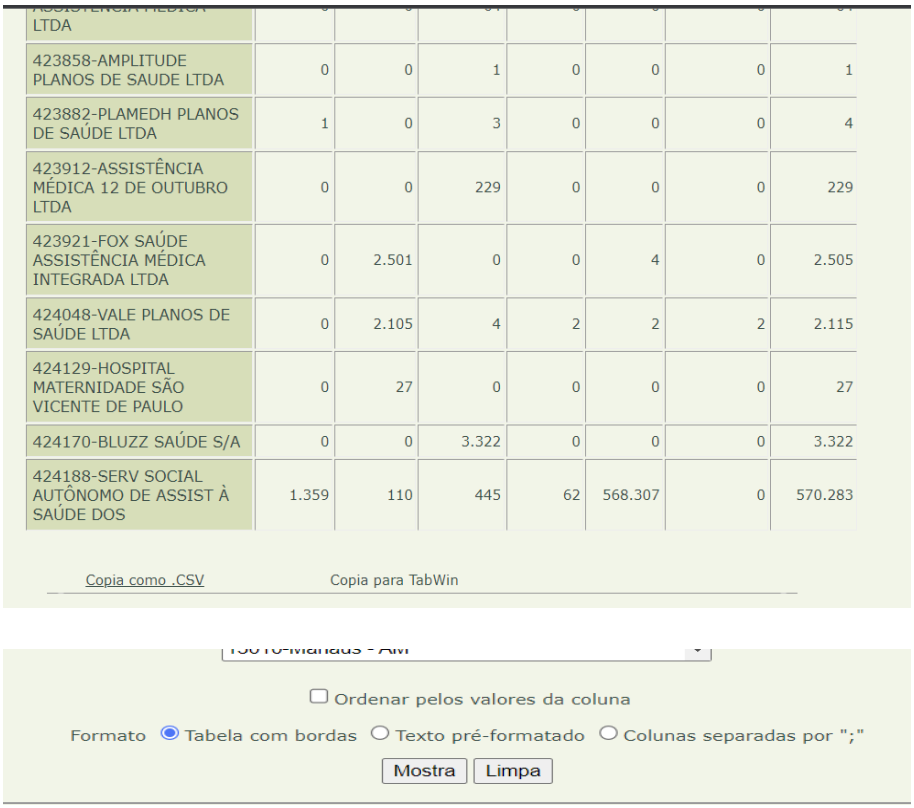
https://www.ans.gov.br/anstabnet/index.htm



https://www.ans.gov.br/anstabnet/cgi-bin/dh?dados/tabnet_cc.def







LTDA							
423858-AMPLITUDE PLANOS DE SAUDE LTDA	0	0	1	0	0	0	1
423882-PLAMEDH PLANOS DE SAÚDE LTDA	1	0	3	0	0	0	4
423912-ASSISTÊNCIA MÉDICA 12 DE OUTUBRO LTDA	0	0	229	0	0	0	229
423921-FOX SAÚDE ASSISTÊNCIA MÉDICA INTEGRADA LTDA	0	2.501	0	0	4	0	2.505
424048-VALE PLANOS DE SAÚDE LTDA	0	2.105	4	2	2	2	2.115
424129-HOSPITAL MATERNIDADE SÃO VICENTE DE PAULO	0	27	0	0	0	0	27
424170-BLUZZ SAÚDE S/A	0	0	3.322	0	0	0	3.322
424188-SERV SOCIAL AUTÔNOMO DE ASSIST À SAÚDE DOS	1.359	110	445	62	568.307	0	570.283

Copia como .CSV Copia para TabWin

II) Dados dos estabelecimentos de saúde do Brasil

O arquivo “arquivosCNES.zip” disponível no Moodle possui dados sobre os estabelecimentos de saúde do Brasil e informações de geolocalização para fazer o mapeamento destes estabelecimentos. Siga as instruções abaixo para preparar os dados para utilização no trabalho.

1. Unificar os arquivos DBF que contêm o código CNES e o nome do hospital, consolidando em um único dataset.
2. Unificar os Parquets da pasta nome_cnes, mantendo apenas as colunas de código do município e CNES, removendo duplicatas.
3. Cruzar os dois resultados (do passo 1 e 2) para obter um dataset com CNES + Nome do Hospital + Código do Município.
4. Unificar com o arquivo municipios.csv, que traz o código do município e suas coordenadas (latitude e longitude).

III) Dados dos centros de pesquisa clínica ativos no Brasil

Estes dados podem ser obtidos através do site https://clinicaltrials.gov/. Um código para obter um arquivo com estas informações está disponível em: https://colab.research.google.com/drive/13Wk4BPmJ9V_diDz1SpuBk1CFV2AYio3D?usp=sharing

IV) Dados das indústrias biofarmacêuticas

Estes dados serão fornecidos pela startup Let Me Trial.