1. Objetivo

Construir um pipeline de dados na nuvem que permita coletar, padronizar, modelar, carregar e analisar contratos públicos federais no período de 2020 até 2024. O propósito é transformar esses dados abertos brutos em um repositório (data lake / data warehouse) analisável para identificar padrões de gastos, detectar anomalias e gerar insights acionáveis sobre contratação governamental.

1.a. Problema

Os dados de compras públicas estão fragmentados em arquivos CSV para cada mês de cada ano, o que dificulta análises consistentes sobre quem contrata, quanto se gasta, tendências temporais e potenciais anomalias (valores concentrados, contratos sem concorrência, fornecedores com volume atípico), etc.

O projeto busca criar uma base única, limpa, confiável e atualizável automaticamente para responder perguntas de auditoria pública e pesquisa.

1.b. Perguntas de negócio

- Gasto total por ano e no período.

- Top 10 fornecedores por valor total e número de contratos no período.

- Distribuição de gastos por órgão, modalidade de compra e grupo de objeto de compra no período.

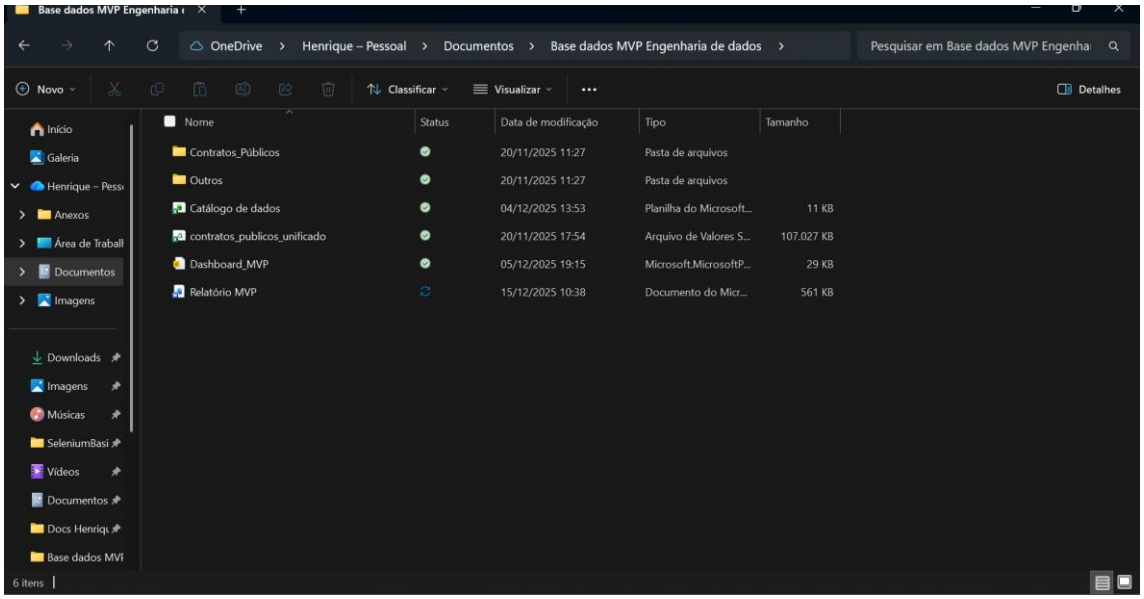
- Evolução ano a ano do valor total contratado por fornecedor (Top 10).

- Sazonalidade (comparação da quantidade de contratos para cada mês no período).

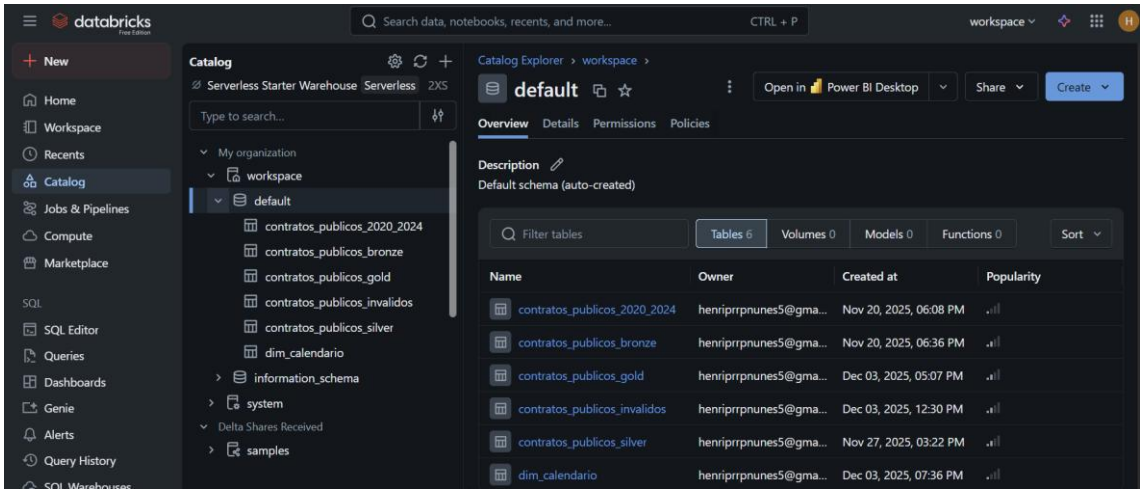
2. Coleta

O conjunto de dados encontra-se no site do Portal da Transparência (https://portaldatransparencia.gov.br/download-de-dados/compras). Diante disso, é preciso gerar arquivos CSV para cada mês de cada ano, gerando 60 arquivos separados (no intervalo dos anos de 2020 a 2024). Desse modo, foi necessário unificar esses documentos e formar uma base de dados singular para que fosse feito o upload dos dados na plataforma do Databricks. Para esse fim, utilizou-se o código abaixo no Jupyter Notebook:

https://github.com/henriqueprrpnunes/MVP_engenharia_de_dados_PUC/blob/main/contratoscsv_unificados.ipynb



Dentro da plataforma Databricks, acessei a seção *Catalog*. Assim, fiz o upload do arquivo *contratos_publicos_unificado.csv* por meio da criação de uma tabela (chamada contratos_publicos_2020_2024) dentro do schema *default* no catálogo *workspace*.



3. Modelagem

O pipeline de dados desenvolvido neste projeto foi estruturado com base no conceito de Arquitetura Medalhão, contemplando as camadas Bronze, Silver e Gold.

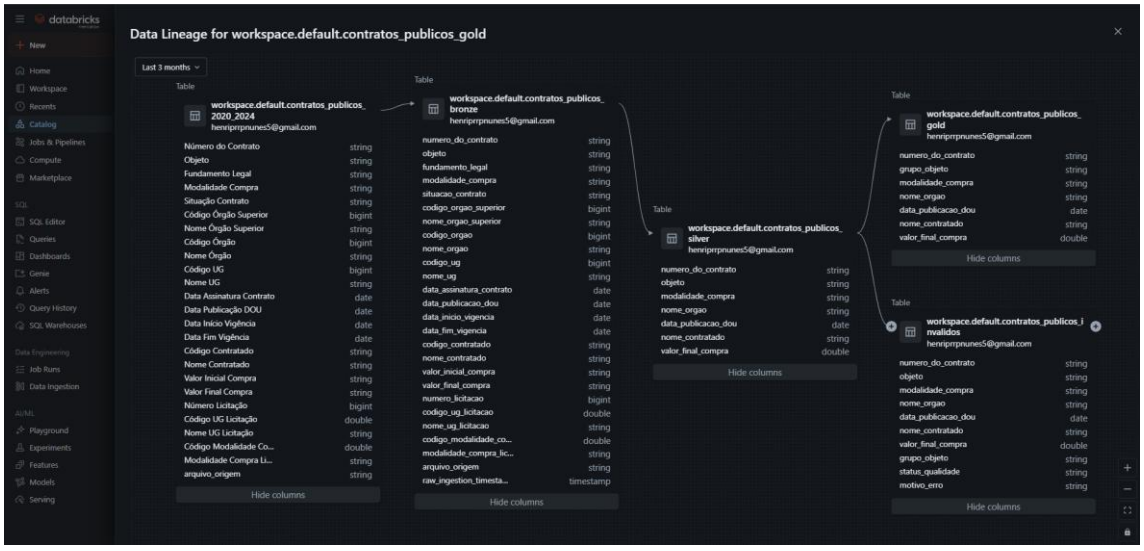
A camada Bronze é composta pelos dados brutos importados diretamente da fonte, a partir do arquivo CSV (*contratos_publicos_unificado.csv*). Nessa etapa, foi realizado apenas um tratamento mínimo, incluindo a normalização automática dos nomes das colunas para garantir consistência, bem como a adição de um timestamp técnico para registrar o momento da ingestão dos dados, resultando na tabela *contratos_publicos_bronze*.

A camada Silver concentra os processos de refinamento dos dados, como a filtragem de colunas relevantes, a padronização de campos textuais e a definição dos tipos de dados adequados. Essas transformações culminaram na materialização da tabela *contratos_publicos_silver*.

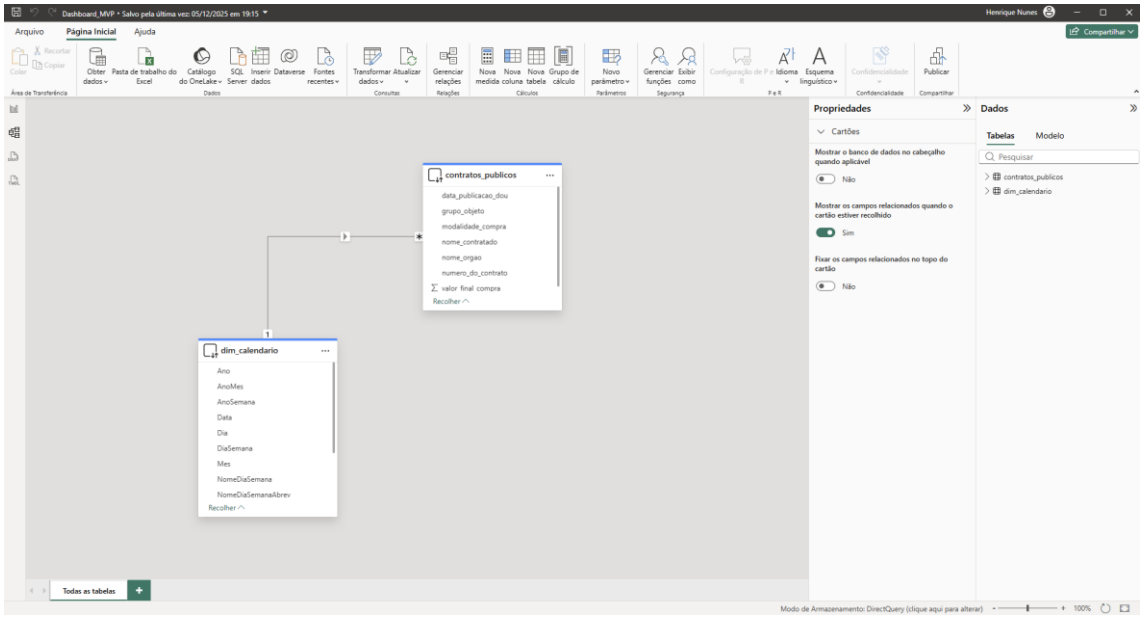
A camada Gold, denominada *contratos_publicos_gold*, representa o conjunto de dados final, preparado para consumo analítico no Power BI. Nessa fase, foram aplicadas regras de qualidade de dados e realizada a inclusão de uma coluna categórica para enriquecimento das análises.

Por fim, foi criada uma tabela dimensão calendário completa, abrangendo todas as datas entre os anos de 2000 e 2050, acompanhadas de seus principais atributos derivados, com o objetivo de suportar análises temporais consistentes no ambiente analítico.

Linhagem de dados:



No ambiente do Power BI, foi implementado um modelo de dados no Esquema Estrela, composto pela tabela *dim_calendario*, que representa a dimensão temporal, e pela tabela *contratos_publicos*, que representa a tabela fato. O relacionamento entre essas tabelas é do tipo um-para-muitos (1:N) e é estabelecido, no nível semântico da ferramenta, por meio da coluna *Data* da dimensão calendário e da coluna *data_publicacao_dou* da tabela fato. Essa modelagem viabiliza análises temporais consistentes dos contratos públicos, estando alinhada às boas práticas de modelagem dimensional para fins analíticos.



Catálogo de dados:

TABELA CONTRATOS_PUBLICOS_GOLD

COLUNA	TIPO	DESCRIÇÃO	DOMÍNIO/VALORES ESPERADOS
numero_do_contrato	string	Número que identifica o contrato no ComprasNet	>=1
grupo_objeto	string	Classificação/Categoria do objeto do contrato.	BENS PATRIMONIAIS; MATERIAIS; OBRAS; SERVIÇOS; OUTROS
modalidade_compra	string	Procedimentos formais que a Administração Pública utiliza para contratar obras, serviços, comprar bens e realizar alienações, os quais definem as regras gerais da competição	Convite; Tomada de Preços; Concorrência; Concorrência Internacional; Pregão; Dispensa de Licitação; Inexigibilidade de Licitação; Concurso; Tomada de Preços por Técnica e Preço; Concorrência por Técnica e Preço; Concorrência Internacional por Técnica e Preço; Pregão - Registro de Preço; Sem Informação
nome_orgao	string	Nome do Órgão	texto livre
data_publicacao_dou	date	Data da publicação do contrato no Diário Oficial da União	01/01/2020 a 31/12/2024
nome_contratado	string	Nome do contratado	texto livre
valor_final_compra	double	Valor final da compra após possíveis reajustes, acréscimos etc	>0

TABELA DIM_CALENDÁRIO

COLUNA	TIPO	DESCRIÇÃO	DOMÍNIO/VALORES ESPERADOS
Data	date	Datas completas	01/01/2000 a 31/12/2050
Ano	int	Ano da data	2000 a 2050
Mes	int	Número do mês da data	1 a 12
Dia	int	Número do dia da data	1 a 31
DiaSemana	int	Número do dia da semana	1 a 7
Trimestre	int	Número do trimestre	1 a 4
SemanaAno	int	Número da semana dentro do ano	1 a 53
Ano Mês	string	Combinação de ano e mês no formato YYYYMM	2000-01 a 2050-12
AnoSemana	string	Combinação de ano e número da semana no formato YYYYSS	2000-01 a 2050-53
NomeMes	string	Nome completo do mês	Janeiro; Fevereiro; Março; Abril; Maio; Junho; Julho; Agosto; Setembro; Outubro; Novembro; Dezembro
NomeMesAbrev	string	Abreviação do nome do mês	Jan; Fev; Mar; Abr; Mai; Jun; Jul; Ago; Set; Out; Nov; Dez
NomeDiaSemana	string	Nome completo do dia da semana	Segunda-feira, Terça-feira, Quarta-feira, Quinta-feira, Sexta-feira, Sábado; Domingo
NomeDiaSemanaAbrev	string	Nome abreviado do dia da semana	Seg; Ter; Qua; Qui; Sex; Sáb; Dom

4. Carga

Essa etapa está explicada no link abaixo:

https://github.com/henriqueprrpnunes/MVP_engenharia_de_dados_PUC/blob/9c56b56c2c6d50f7ea40c0d9e3060409e7848ba9/MVP_contratos_publicos.ipynb

5. Análise

5.a. Qualidade de dados

Para avaliar a qualidade dos dados, foram testados cinco critérios principais:

1) se o número do contrato contém apenas dígitos e não é composto apenas de zeros;
2) se a modalidade de compra pertence a uma lista de valores permitidos (Convite, Tomada de Preços, Concorrência, Concorrência Internacional, Pregão, Dispensa de Licitação, Inexigibilidade de Licitação, Concurso, Tomada de Preços por Técnica e Preço, Concorrência por Técnica e Preço, Concorrência Internacional por Técnica e Preço, Pregão - Registro de Preço, Sem Informação);
3) se a data de publicação está dentro de um intervalo pré-definido (entre 2020 e 2024);
4) se o valor final da compra é maior que zero;
5) se alguma coluna possui campos nulos.

Considerando que todas as perguntas de negócio analíticas são respondidas de forma integrada em um único relatório, optou-se por utilizar exclusivamente registros que atenderam simultaneamente a todos os critérios de qualidade definidos. Essa abordagem garantiu consistência analítica entre os diferentes indicadores apresentados e evitou a introdução de vieses decorrentes de inconsistências estruturais, semânticas ou temporais nos dados.

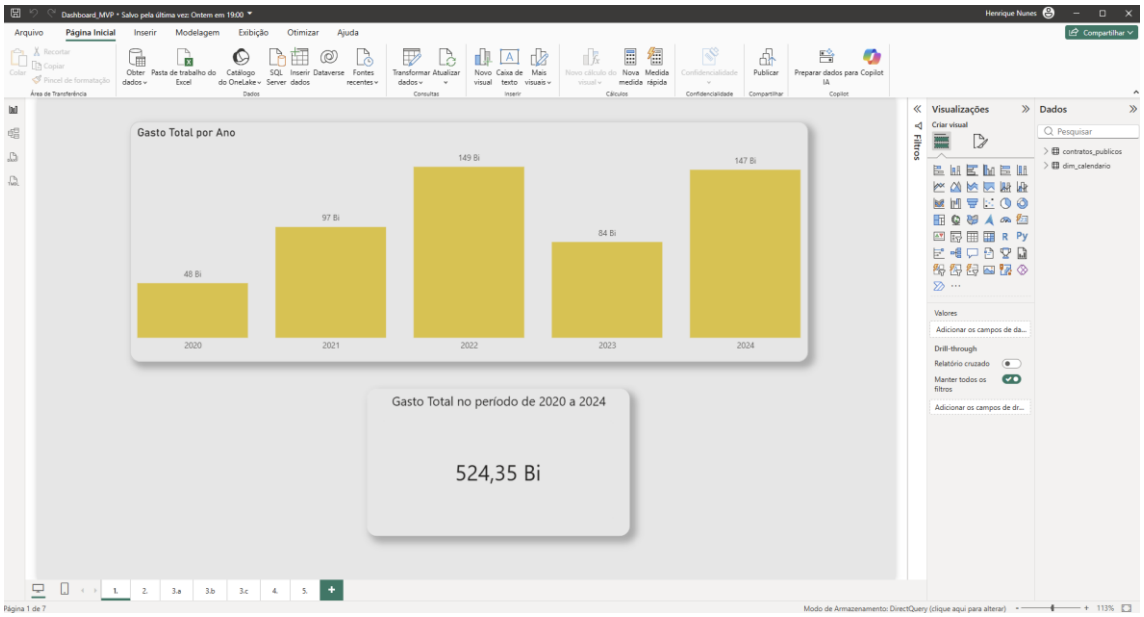
Os registros que não respeitaram a um ou mais critérios foram mantidos separadamente para fins de auditoria e transparência, permitindo a mensuração do impacto da limpeza de dados sobre o conjunto original.

5.b. Solução do problema

O Power BI foi utilizado para responder as perguntas de negócio.

I) Gasto total por ano e no período.

Para obter essas informações, foi selecionado um gráfico de colunas clusterizado em que a coluna *Ano* da tabela *dim_calendario* ficou no eixo X e a coluna *valor_final_compra* (no formato de SOMA) da tabela *contratos_publicos* ficou no eixo Y. Enquanto isso, foi escolhido um cartão com a coluna *valor_final_compra* (no formato de SOMA) para mostrar o total de gastos nos anos de 2020 até 2024.



Os gastos em contratações públicas do Governo Federal entre 2020 e 2024 evidencia variações significativas ao longo do período. Em 2020, o volume de R$ 48 bilhões reflete um patamar reduzido, influenciado pelas incertezas e limitações operacionais decorrentes da pandemia.

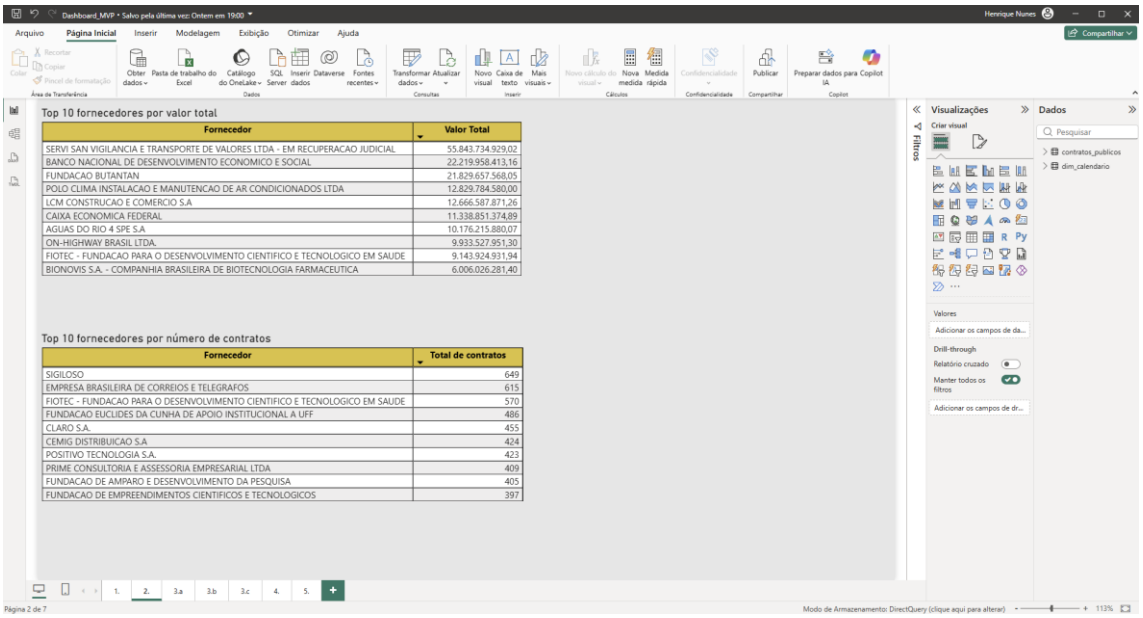
Em 2021, os gastos aumentaram para R$ 97 bilhões, indicando a retomada das contratações e a recomposição de demandas represadas. O ano de 2022 registrou o maior nível do período, com R$ 149 bilhões, sugerindo forte expansão das contratações federais, possivelmente associada à intensificação da execução orçamentária.

Em 2023, observa-se uma retração para R$ 84 bilhões, compatível com um cenário de transição governamental e ajuste das prioridades de gasto. Já em 2024, o retorno ao patamar de R$ 149 bilhões indica a normalização e o fortalecimento das contratações públicas.

De forma geral, os dados revelam que o gasto federal em contratações públicas foi marcado por alta volatilidade, influenciado por fatores conjunturais e institucionais ao longo do período analisado.

II) Top 10 fornecedores por valor total e número de contratos no período.

Para obter essas informações, foi selecionado a tabela. Assim, uma tem as colunas *nome_contratado* e *valor_final_compra* (no formato de SOMA) da tabela *contratos_publicos*. A segunda tabela apresenta *nome_contratado* e *numero_do_contrato* (no formato de CONTAGEM DISTINTA) da tabela *contratos_publicos.* Foi aplicado um filtro de tipo N superior até 10 baseado na soma do *valor_final_compra* e na contagem do *numero_do_contrato* nestes visuais.



Top 10 fornecedores por valor total

Fornecedor	Valor Total
SERVI SAN VIGILANCIA E TRANSPORTE DE VALORES LTDA - EM RECUPERACAO JUDICIAL	55.843.734.929,02
BANCO NACIONAL DE DESENVOLVIMENTO ECONOMICO E SOCIAL	22.219.958.413,16
FUNDACAO BUTANTAN	21.829.657.568,05
POLO CLIMA INSTALACAO E MANUTENCAO DE AR CONDICIONADOS LTDA	12.829.784.580,00
LCM CONSTRUCAO E COMERCIO S.A	12.666.587.871,26
CAIXA ECONOMICA FEDERAL	11.338.851.374,89
AGUAS DO RIO 4 SPE S.A	10.176.215.880,07
ON-HIGHWAY BRASIL LTDA.	9.933.527.951,30
FIOTEC - FUNDACAO PARA O DESENVOLVIMENTO CIENTIFICO E TECNOLOGICO EM SAUDE	9.143.924.931,94
BIONOVIS S.A. - COMPANHIA BRASILEIRA DE BIOTECNOLOGIA FARMACEUTICA	6.006.026.281,40

Top 10 fornecedores por número de contratos

Fornecedor	Total de contratos
SIGILOSO	649
EMPRESA BRASILEIRA DE CORREIOS E TELEGRAFOS	615
FIOTEC - FUNDACAO PARA O DESENVOLVIMENTO CIENTIFICO E TECNOLOGICO EM SAUDE	570
FUNDACAO EUCLIDES DA CUNHA DE APOIO INSTITUCIONAL A UFF	486
CLARO S.A.	455
CEMIG DISTRIBUICAO S.A	424
POSITIVO TECNOLOGIA S.A.	423
PRIME CONSULTORIA E ASSESSORIA EMPRESARIAL LTDA	409
FUNDACAO DE AMPARO E DESENVOLVIMENTO DA PESQUISA	405
FUNDACAO DE EMPREENDIMENTOS CIENTIFICOS E TECNOLOGICOS	397

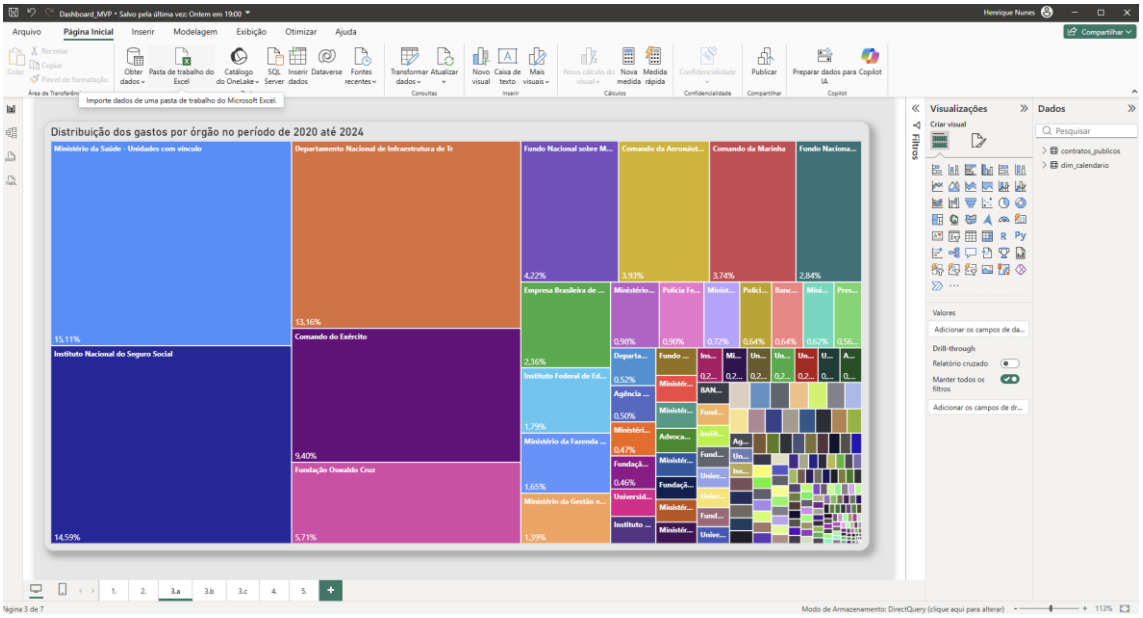
Observa-se uma forte concentração de recursos em poucos fornecedores. Empresas como Serv San Vigilância e Transporte de Valores, BNDES e Fundação Butantan lideram o ranking em valor contratado, com montantes bastante elevados. Esse padrão indica que uma parcela relevante do orçamento é direcionada a contratos de grande porte, geralmente associados a serviços estratégicos, financeiros, científicos, de saúde ou de infraestrutura, que exigem alta capacidade operacional e técnica. A presença de bancos públicos e fundações reforça o papel do Estado como contratante de instituições-chave para políticas públicas estruturantes.

Por outro lado, quando analisado o volume de contratos, o cenário se mostra mais pulverizado. Fornecedores como Correios, Serpro, Fiotec e empresas de tecnologia e serviços aparecem com elevado número de contratos, mas não necessariamente com os maiores valores totais. Isso sugere a existência de contratos recorrentes, de menor valor

individual, voltados à manutenção de serviços continuados, tecnologia da informação, logística e apoio administrativo.

III) Distribuição de gastos por órgão, modalidade de compra e grupo de objeto de compra no período.

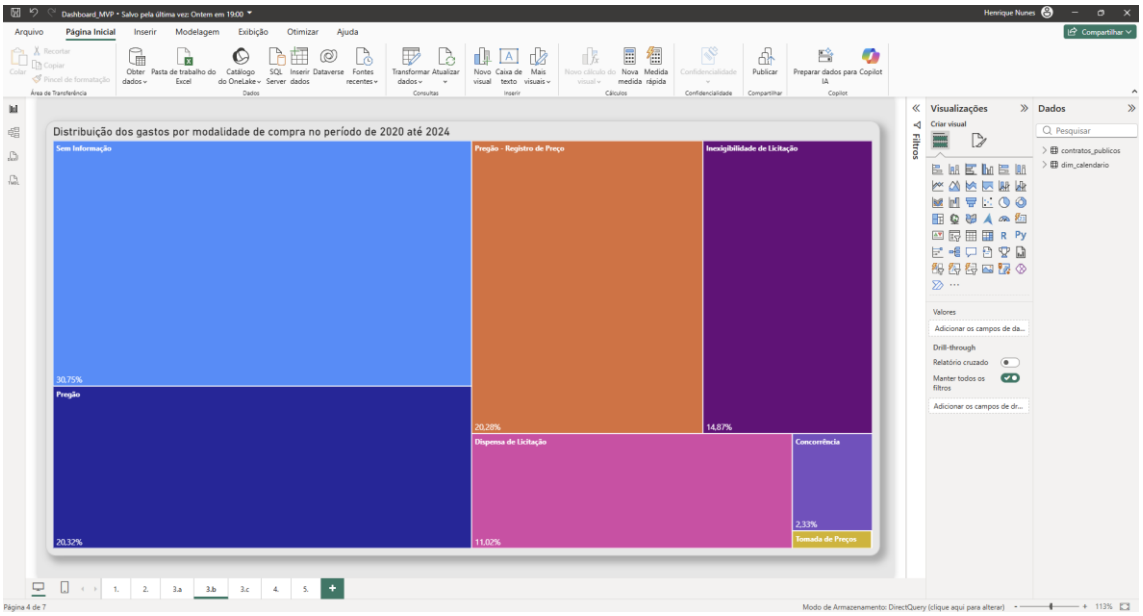
Para obter as informações por órgão, foi selecionado a Treemap, onde o campo de categoria foi representado pela coluna *nome_orgao* da tabela *contratos_publicos* e o campo de valores pela coluna *valor_final_compra* (no formato de SOMA com o valor mostrado como percentual do total geral) da tabela *contratos_publicos*.



Os gastos com contratações públicas do Governo Federal apresentam um padrão claro de concentração institucional. A maior parcela dos recursos é direcionada a poucos órgãos, com destaque para o Ministério da Saúde, o INSS e o DNIT, o que evidencia a priorização de políticas públicas nas áreas de saúde, previdência e infraestrutura.

Na sequência, os comandos das Forças Armadas concentram volumes relevantes de gasto, refletindo a natureza estratégica e contínua das contratações na área de defesa. Paralelamente, instituições como a Fundação Oswaldo Cruz reforçam a importância das contratações voltadas à ciência e à saúde pública.

Para obter as informações por modalidade de compra, foi selecionado a Treemap, onde o campo de categoria foi representado pela coluna *modalidade_compra* da tabela *contratos_publicos* e o campo de valores pela coluna *valor_final_compra* (no formato de SOMA com o valor mostrado como percentual do total geral) da tabela *contratos_publicos*.

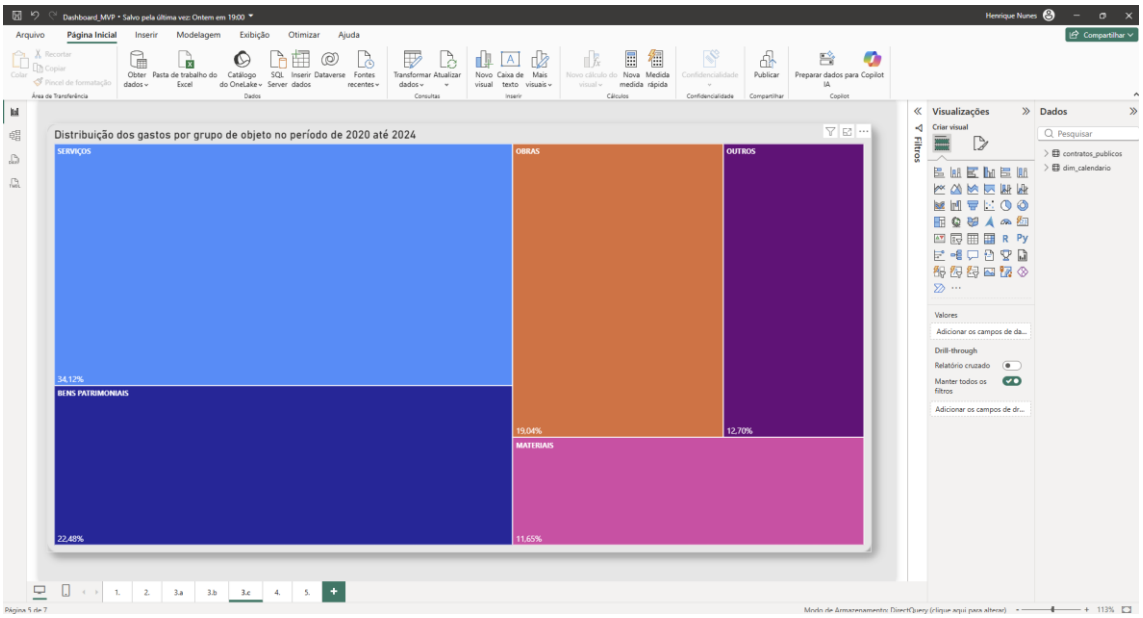


A distribuição dos gastos por modalidade de compra evidencia um padrão relevante nas contratações públicas federais:

- A categoria “Sem Informação” concentra a maior parcela dos gastos (30,75%), o que indica fragilidade na qualidade ou completude dos dados e limita a transparência da análise por modalidade.
- O Pregão e o Pregão – Registro de Preços somam juntos cerca de 40% do total (20,32% e 20,28%, respectivamente), confirmando o pregão como principal instrumento de contratação, alinhado ao seu caráter competitivo e à busca por eficiência.
- A Inexigibilidade de Licitação responde por 14,87%, refletindo contratações em contextos específicos, como fornecedor exclusivo ou serviços técnicos especializados.
- A Dispensa de Licitação representa 11,02%, indicando uso relevante desse mecanismo, geralmente associado a situações emergenciais ou de menor valor.
- Modalidades mais tradicionais, como Concorrência (2,33%) e Tomada de Preços, apresentam participação residual.

Em síntese, os dados mostram predominância de modalidades mais ágeis, especialmente o pregão, mas também revelam um volume expressivo de gastos sem classificação adequada, apontando para oportunidades de melhoria na governança e na qualidade da informação sobre as contratações públicas.

Para obter as informações por grupo de objeto de compra, foi selecionado a Treemap, onde o campo de categoria foi representado pela coluna *grupo_objeto* da tabela *contratos_publicos* e o campo de valores pela coluna *valor_final_compra* (no formato de SOMA com o valor mostrado como percentual do total geral) da tabela *contratos_publicos*.



Os gastos em contratações públicas federais concentram-se majoritariamente em Serviços (34,12%), evidenciando o peso de contratos continuados e de apoio à operação do Estado.

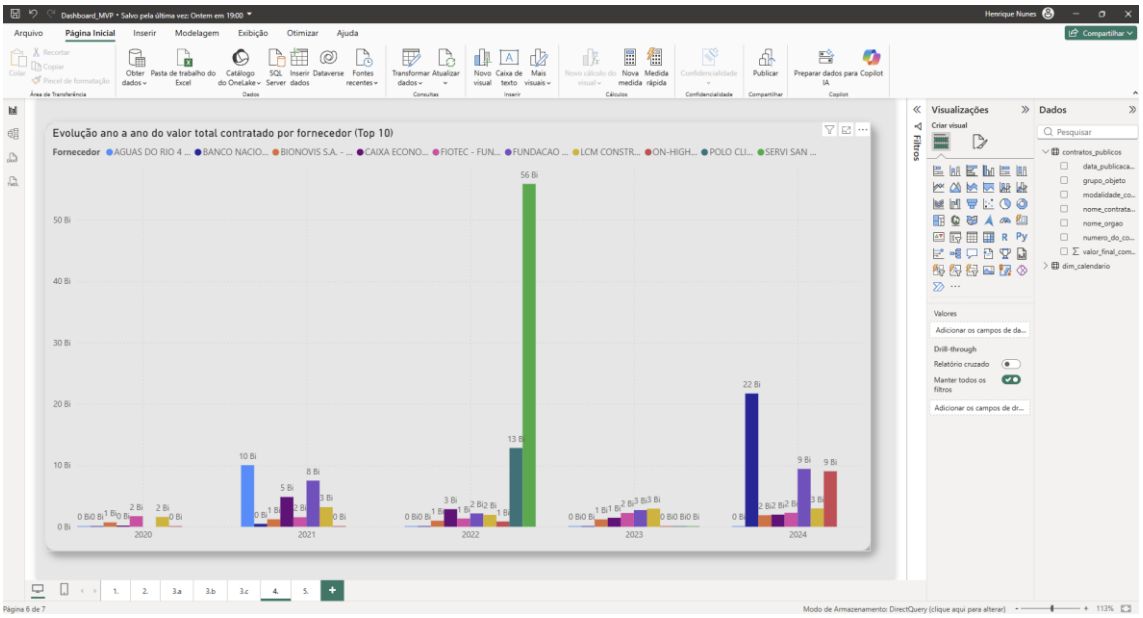
Na sequência, Bens Patrimoniais representam 22,48%, indicando investimentos relevantes em ativos duráveis, enquanto Obras respondem por 19,04%, refletindo a execução de projetos de infraestrutura.

Os Materiais concentram 11,65% dos gastos, e a categoria Outros corresponde a 12,70%, agregando objetos diversos de menor especificidade.

De forma geral, a distribuição aponta para uma priorização de serviços e investimentos estruturais, com menor participação relativa de aquisições de materiais no período analisado.

IV) Evolução ano a ano do valor total contratado por fornecedor (Top 10).

Para obter essas informações, foi selecionado o gráfico de colunas clusterizado com a coluna *Ano* da tabela *dim_calendario* no eixo X, a coluna *valor_final_compra* (no formato de SOMA) da tabela *contratos_publicos* no eixo Y e a coluna *nome_contratado* da tabela *contratos_publicos* no campo de legenda.



A evolução anual do valor contratado dos 10 maiores fornecedores demonstra um padrão de forte concentração pontual e elevada volatilidade.

Em 2022, destaca-se um pico excepcional de contratações, com um fornecedor alcançando cerca de R$ 56 bilhões, valor muito superior aos demais anos e fornecedores, indicando a celebração de contratos atípicos ou de grande porte concentrados em um único exercício. Esse mesmo ano também apresenta outros fornecedores com volumes relevantes, reforçando um cenário de expansão concentrada.

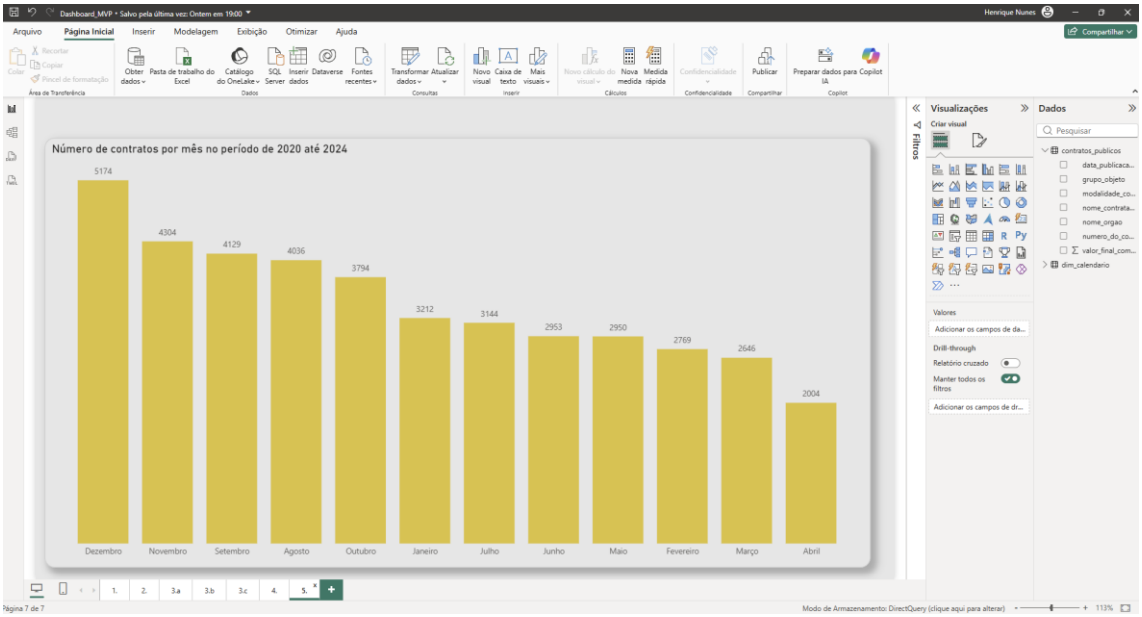
Nos anos de 2020 e 2021, os valores são significativamente menores e mais distribuídos, com contratos que, em geral, não ultrapassam R$ 10 bilhões por fornecedor. Já em 2023, observa-se uma queda generalizada, com volumes reduzidos e ausência de grandes contratos.

Em 2024, ocorre nova retomada parcial, com destaque para um fornecedor que atinge cerca de R$ 22 bilhões, além de outros com valores entre R$ 8 e R$ 9 bilhões, embora sem repetir o nível extremo de 2022.

De forma geral, o gráfico indica que o gasto com grandes fornecedores federais é marcado por eventos pontuais de alto valor, concentrados em poucos fornecedores e anos específicos, mais do que por um crescimento contínuo e uniforme ao longo do tempo.

V) Sazonalidade (comparação da quantidade de contratos para cada mês no período).

Para obter essas informações, foi selecionado o gráfico de colunas clusterizado que a coluna *Ano* da tabela *dim_calendario* ficou no eixo X e a coluna *numero_do_contrato* (no formato de CONTAGEM DISTINTA) da tabela *contratos_publicos* ficou no eixo Y.



O gráfico de número de contratos por mês (2020–2024) revela um padrão claro de sazonalidade nas contratações públicas federais.

Os maiores volumes ocorrem no final do ano, com destaque para dezembro (pico absoluto) e novembro, indicando forte concentração de contratações no encerramento do exercício orçamentário. Meses como setembro, agosto e outubro também apresentam níveis elevados, reforçando a aceleração das contratações no segundo semestre.

Em contraste, o início do ano registra menor atividade. Janeiro, fevereiro e março apresentam volumes reduzidos, sendo abril o mês com menor número de contratos, o que sugere cautela orçamentária e reorganização administrativa no começo do exercício fiscal.

Em síntese, os dados indicam que as contratações federais são fortemente influenciadas pelo ciclo orçamentário, com concentração no fim do ano e retração nos primeiros meses.

6. Autoavaliação

Ao final do desenvolvimento deste trabalho, considero que os objetivos inicialmente delineados foram, em grande medida, atingidos. A proposta de estruturar, tratar e analisar dados de contratações públicas federais foi cumprida, resultando em um modelo de dados funcional e em análises capazes de evidenciar padrões relevantes, como sazonalidade nas contratações e concentração de gastos.

Durante a execução, algumas dificuldades se mostraram centrais. Um dos principais desafios foi a ausência de tabelas de referência oficiais padronizadas para atributos fundamentais, como órgão contratante, modalidade de compra e contratado. Essa limitação impediu a criação de tabelas dimensão baseadas em códigos oficiais, o que levou à decisão de manter esses atributos diretamente na tabela fato, utilizando os nomes textuais disponíveis na base original. Embora essa solução preserve a integridade da informação, ela reduz o nível de normalização do modelo e impõe restrições para análises mais avançadas de governança e integração com outras bases.

Outro desafio relevante foi a classificação dos contratos a partir da descrição do objeto, que se apresenta de forma não estruturada e heterogênea. Para contornar essa limitação, foi adotada uma abordagem baseada em palavras-chave, permitindo categorizar os contratos de forma aproximada. Apesar de funcional, essa estratégia está sujeita a ambiguidades semânticas e possíveis erros de classificação, exigindo cuidado na interpretação dos resultados.

Como trabalhos futuros, este projeto pode ser enriquecido com a incorporação de dicionários oficiais ou bases auxiliares, caso venham a ser disponibilizadas, permitindo a criação de tabelas dimensão mais robustas e padronizadas. Além disso, a aplicação de técnicas mais avançadas de processamento de linguagem natural (NLP) poderia aprimorar significativamente a categorização dos objetos contratuais, aumentando a precisão analítica. Essas evoluções contribuiriam para elevar a maturidade do modelo de dados e ampliar o valor do trabalho no portfólio acadêmico e profissional.