Diretora: Jenifer Milena Pellin da Silva
Assessoras: Alessandra dos Santos Fernandes e Rafaela A. Schneider Hinsching
Assistente de Educação: Marisa R. Reichardt
Orientadora Educacional: Joice Daniela Santana Muraoka
Professor: Wendell Thomas Teske
Estudante: Diogo Zoboli

RELATÓRIO DE GOVERNANÇA DE DADOS – PROJETO PETSHOP

1. INGESTÃO DE DADOS

O desafio proposto consistia em tratar, inicialmente, uma fonte de dados de vendas de um pet shop – resolução de outliers, tipos de dados e missings – e criar um dashboard para ajudar na análise de dados. Primeiramente, a base estava com tipos de dados errados, sem nomes das colunas e muitas linhas nulas ou zeradas. A fonte continha 16 colunas e 250.062 linhas, além de que todos os dados estavam com tipo de texto, mesmo os que aparentavam ser inteiros. O Dicionário de dados da base está localizado abaixo.

Quadro 1 – Dicionário de Dados original: base_vendas-petshop.

base_vendas-petshop

HOT Storage | 14 colunas

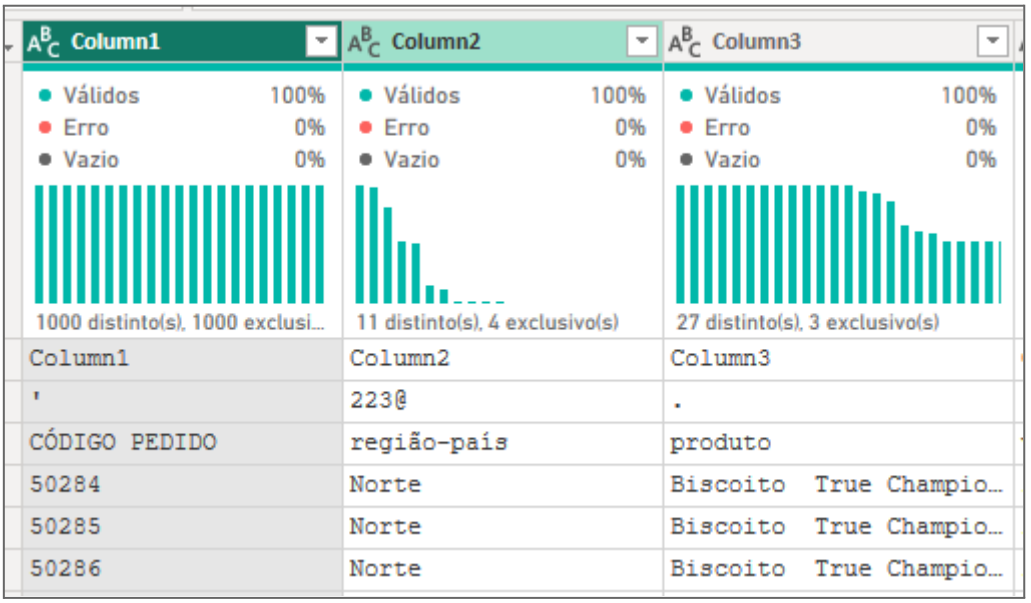
Coluna	Tipo de Dado	Permite Vazios	Categórica	Tipo de Restrição
Column1	VARCHAR	NÃO	NÃO	PK
Column2	VARCHAR	NÃO	SIM	NOT NULL
Column3	VARCHAR	NÃO	NÃO	NOT NULL
Column4	VARCHAR	SIM	NÃO	NULL
Column5	VARCHAR	SIM	NÃO	NULL
Column6	VARCHAR	SIM	NÃO	NULL
Column7	VARCHAR	SIM	NÃO	NULL
Column8	VARCHAR	SIM	NÃO	NULL
Column9	VARCHAR	SIM	NÃO	NULL
Column10	VARCHAR	SIM	SIM	NULL
Column11	VARCHAR	SIM	SIM	NULL
Column12	VARCHAR	SIM	SIM	NULL
Column13	VARCHAR	SIM	NÃO	NULL
Column14	VARCHAR	NÃO	SIM	NULL
Column15	VARCHAR	SIM	NÃO	NULL
Column16	VARCHAR	NÃO	SIM	NOT NULL

Fonte: Elaborado pelos autores.

2. TRATAMENTO DE DADOS

Com os dados inseridos no Power BI, o tratamento de dados inicia-se com a colocação dos nomes dos atributos, usando a função de promover colunas – pois os nomes das colunas estavam juntos aos registros.

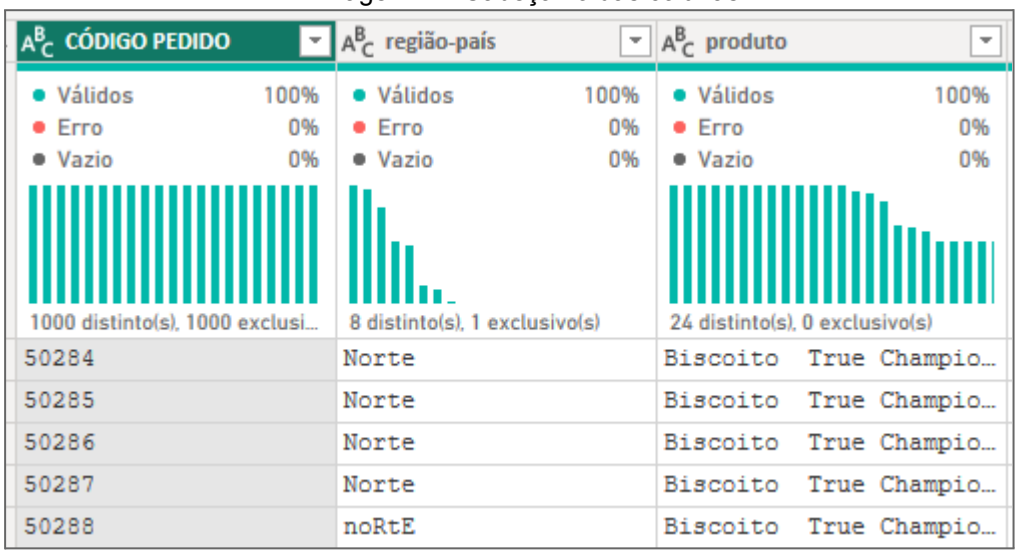
Imagem 1 – Nomes errados das colunas.



Fonte: Elaborado pelos autores.

Com o uso da ferramenta, descobriu-se os nomes de todos os atributos, que, posteriormente, foram padronizados para “ds_nome_coluna” a fim de facilitar o tratamento e a análise de dados.

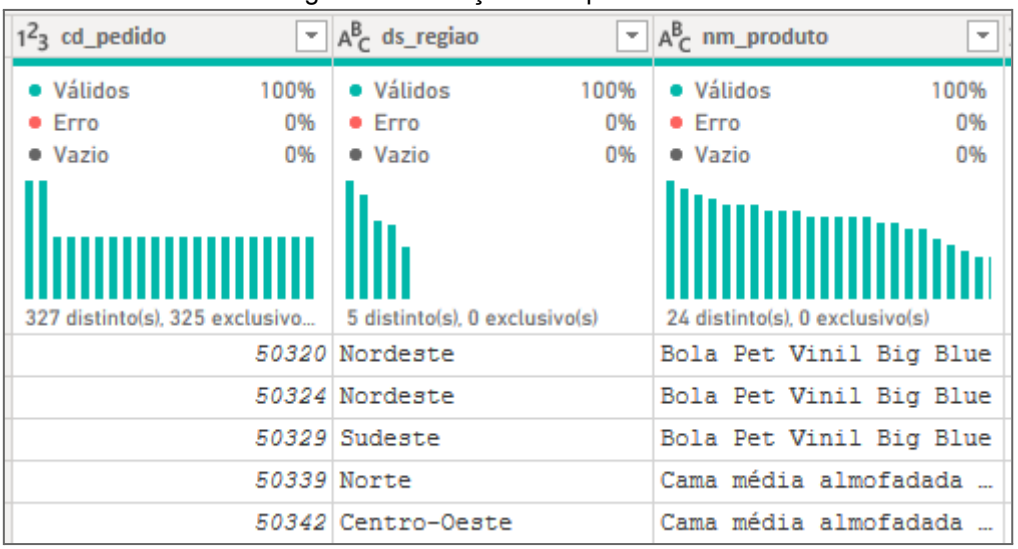
Imagem 2 – Cabeçalho das colunas.



Fonte: Elaborado pelos autores.

Mesmo com os nomes corretos das colunas, os tipos de dados estavam desalinhados à proposta de cada atributo. Logo, foi iniciado o tratamento de cada coluna da tabela, alterando os tipos de dados e revendo os dados para não conter erros.

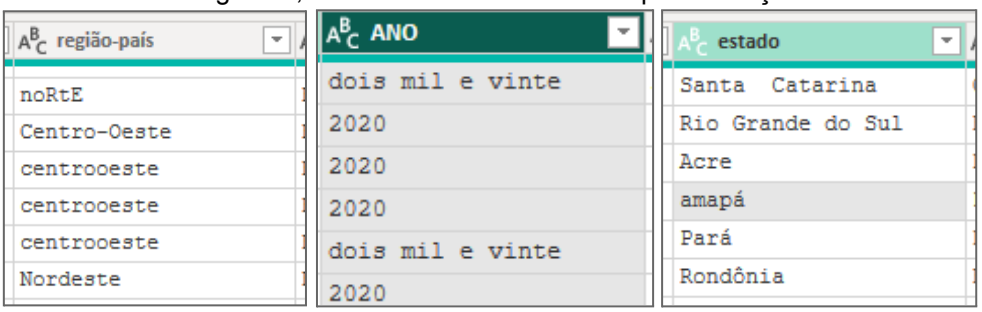
Imagem 3 - Alteração dos tipos de dados.



Fonte: Elaborado pelos autores.

Algumas colunas haviam alguns valores despadronizados (como a região Norte = noRtE ou espaços em branco nos registros), o que, por serem categóricas, não poderia acontecer. Então foi feita a substituição de alguns registros pela sua padronização correta. O mesmo ocorreu na coluna ANO que, apesar de não ser categórica, era um registo de números inteiros e necessitava de substituição.

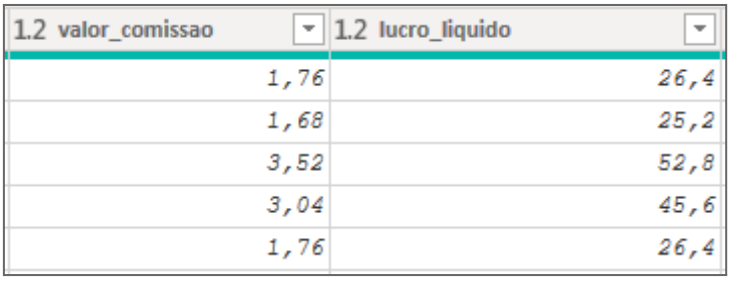
Imagens 4, 5 e 6 - Dados com erros de padronização.



Fonte: Elaborado pelos autores.

As colunas de valores decimais continham dados negativos (coluna de quantidade – inteira – também continha valores negativos, mesmo processo) e majoritariamente tinham dados com pontos finais nos lugares de vírgulas, visto que o Power BI estava em Português e lia pontos como divisão de milhares (Em “1.000,00”, estava escrito “1000.00) – o que impedia-a de ser transformada em tipo decimal, visto que daria erro em todos os valores com ponto final. Para resolver o primeiro erro, foi substituído “-” por “” (nada), e para o segundo erro, foi substituído todos os pontos finais por vírgula e, em seguida, trocado o tipo de dados de texto para números decimais.

Imagem 7 - Dados numéricos padronizados.

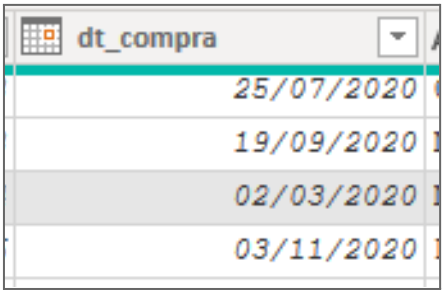


1.2 valor_comissao	1.2 lucro_liquido
1,76	26,4
1,68	25,2
3,52	52,8
3,04	45,6
1,76	26,4

Fonte: Elaborado pelos autores.

As datas de compra estavam todas separadas por colunas (DIA, MES e ANO). Para fazer uma tabela de data de compra, foi necessário transformar as três colunas em números inteiros e mesclá-las, que foi realizado com sucesso. No entanto, havia uma data errada dentro da coluna, dia 31 de fevereiro de 2020. Para resolver esse erro, antes de mudar o tipo de dados do atributo, substituí o dia 31/02/2020 para 02/03/2020 – pequeno cálculo, como fevereiro de 2020 teve apenas 29 dias, adicionei dois dias ao próximo mês – e, assim, mudei o tipo de dados para Date.

Imagem 8 - Atributo de data padronizado.

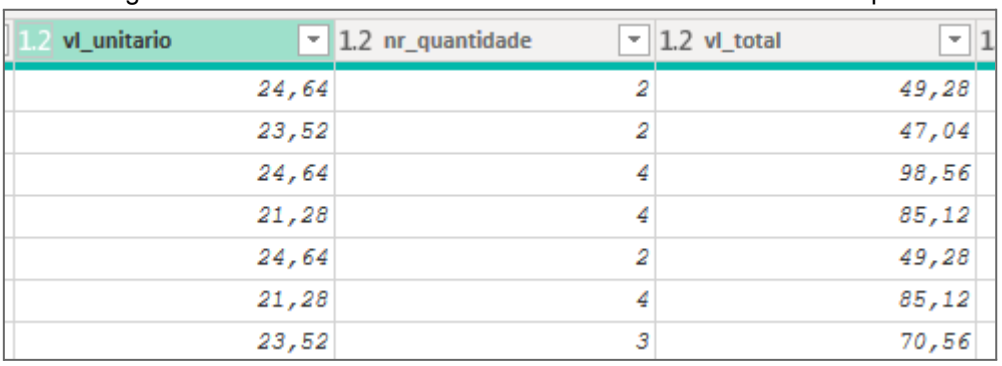


dt_compra
25/07/2020
19/09/2020
02/03/2020
03/11/2020

Fonte: Elaborado pelos autores.

Alguns valores das colunas “nr_quantidade” e “vl_total” estavam faltando. Para resolver, foi criada uma coluna personalizada para calcular o valor total (multiplicando a quantidade pelo valor unitário). Logo, criei outra coluna personalizada que dividia os resultados do valor total pelo atributo do valor unitário para descobrir a quantidade de todos os que faltavam. Todavia, ainda havia erros na coluna de valor total, onde eu multipliquei as novas colunas de quantidade e valor unitário para não haver mais erros. Finalmente, transformei esses dados em números inteiros, resolvendo a falta de dados de alguns atributos.

Imagens 9 e 10 - Gambiarra feita com os dados de valores de compra.



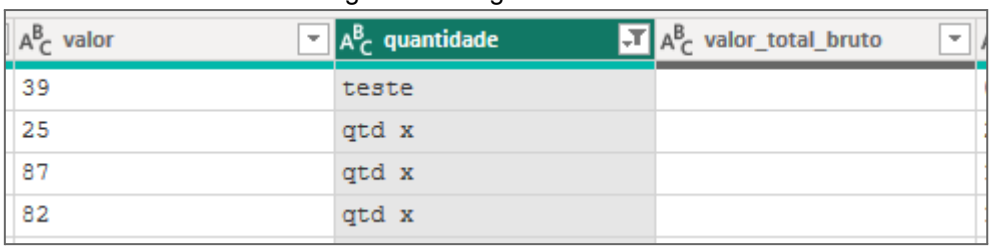
1.2 vl_unitario	1.2 nr_quantidade	1.2 vl_total	1
24,64	2	49,28	
23,52	2	47,04	
24,64	4	98,56	
21,28	4	85,12	
24,64	2	49,28	
21,28	4	85,12	
23,52	3	70,56	

1.2 (não usar) vl_total 2	1.2 (não usar) quantidade	1.2 (não usar) valor_total_bruto
49,28	2	49,28
47,04	2	47,04
98,56	4	98,56
85,12	4	85,12
null	null	49,28
85,12	4	85,12

Fonte: Elaborado pelos autores.

Existiam registros nas colunas de valores que continham palavras como “teste” e “qtd x”, o que indica serem possíveis testes anteriores ao uso da base de dados. Para esses, foi usado a ferramenta de substituir valores, onde substituí todos esses para *null* e logo após apaguei todos os valores nulos e possíveis erros que pudessem ser apresentados na base.

Imagem 11 - Registros de testes.



valor	quantidade	valor_total_bruto
39	teste	
25	qtd x	
87	qtd x	
82	qtd x	

Fonte: Elaborado pelos autores.

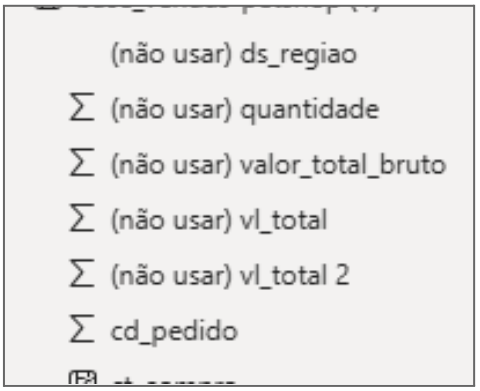
Por fim, renomeei todas as colunas restantes que estavam despadronizadas, criei o Dicionário de dados final com todas as colunas (utilizáveis) para uma possível análise de dados. No dicionário, o cd_pedido será a chave primária da base, visto que é o atributo de identificação de cada registro. Também há algumas colunas categóricas, todas elas seriam usadas com restrições de check, visto que já teriam possíveis valores já escolhidos. Vale ressaltar que além das colunas apresentadas no dicionário, houveram colunas que auxiliaram na análise de dados, tais que não foram apresentadas no dicionário.

Quadro 2 e Imagem 12 - Dicionário de Dados finalizado e imagem com colunas de auxílio.

base_vendas-petshop

HOT Storage | 14 colunas

Coluna	Tipo de Dado	Permite Vazios	Categórica	Tipo de Restrição
cd_pedido	INT	NÃO	NÃO	PK
nm_produto	VARCHAR(250)	NÃO	NÃO	NOT NULL
vl_unitario	DECIMAL(10,2)	NÃO	NÃO	NOT NULL
nr_quantidade	INT	NÃO	NÃO	NOT NULL
vl_total	DECIMAL(10,2)	NÃO	NÃO	NOT NULL
dt_compra	DATE	NÃO	NÃO	NOT NULL
fm_pagamento	VARCHAR(25)	NÃO	SIM	CK
ds_categoria	VARCHAR(25)	NÃO	SIM	CK
ds_regiao	VARCHAR(25)	NÃO	SIM	CK
ds_estado	VARCHAR(25)	NÃO	SIM	CK
ds_centro_distribuicao	VARCHAR(25)	NÃO	SIM	CK
nm_responsavel	VARCHAR(25)	NÃO	SIM	NOT NULL
vl_comissao	DECIMAL(10,2)	NÃO	NÃO	NOT NULL
vl_lucro_liquido	DECIMAL(10,2)	NÃO	NÃO	NOT NULL



Fonte: Elaborado pelos autores.

2.1. ESTRATÉGIA(S) PARA RESOLUÇÃO DE OUTLIERS

Os outliers encontrados foram, majoritariamente, números extremamente grandes, negativos, ou zerados, vindos das tabelas de quantidade de produtos, valores unitários, bruto, lucro líquido e comissão.

Vale lembrar que nem todo outlier é dado automaticamente como um dado errado. Com isso em mente, em todos os casos, seria praticamente impossível conter dados negativos. Para trabalhar nesse quesito, antes de transformar o tipo de dados em numérico, foi usado a ferramenta de substituição para substituir o sinal de negativo (“-”) por, bem, nada (“”).

Nas mesmas colunas anteriormente citadas, também foram substituídos os valores zerados pela média da coluna – pois, não haveria sentido em existir compras de nenhum item – por exemplo: o atributo da quantidade de produtos de uma compra havia alguns registros zerados que, automaticamente, transformava a coluna de valor total bruto zerada, visto que seus dados eram produto da multiplicação da coluna anterior. Sabendo disso, foi criado um cartão de teste no Dashboard que mostrava as médias da coluna e, em seguida, substituído os valores que estivessem zerados.

Por fim, com relação a números extraordinários, não há alguma forma disponibilizada para verificar sua veracidade, então não há o que pudesse ser realmente realizado em prol disso. Ademais, os demais outliers citados acima foram resolvidos e, no quesito de outliers, a base estava com qualidade.

2.2. ESTRATÉGIA(S) PARA RESOLUÇÃO DE MISSINGS

Missings são dados faltantes, tais que, visto que não havia mais formas de descobrir quais seriam os possíveis dados que poderiam ser usados para descobri-los, houve apenas duas opções: apagar os dados ou criar uma média.

Os missings foram encontrados principalmente nas tabelas de valores, mesmas citadas nas estratégias para resolução de outliers. Inicialmente, os outliers apareceram como valores nulos e em branco. Para a resolução desses, foi usado a média da coluna e substituído os valores faltantes pela média que os mesmos teriam. Feito isso, a resolução de missings foi finalizada rapidamente.

3. MANIPULAÇÃO DE DADOS

Após a finalização do tratamento de dados, ainda haviam dados com erros como era o caso da região do país, tal que continha valores de estados fora de sua região, além de ter sido utilizado a linguagem DAX no Power BI para criar colunas condicionais e melhorar a criação de análises de dados.

Para resolver os erros presentes pelas regiões possuírem estados de fora das mesmas, como, por exemplo, a região Sul possuir registros do estado da Bahia, foi necessário usar a função “SWITCH” do DAX, apresentada na imagem abaixo.

Imagem 13 - Comando DAX para arrumar a região.

```
Região Arrumado =
SWITCH(
    TRUE(),
    'base_vendas-petshop (1)'[ds_estado] IN {"Paraná", "Santa Catarina", "Rio Grande do Sul"}, "Sul",
    'base_vendas-petshop (1)'[ds_estado] IN {"São Paulo", "Rio de Janeiro", "Espírito Santo", "Minas Gerais"}, "Sudeste",
    'base_vendas-petshop (1)'[ds_estado] IN {"Bahia", "Pernambuco", "Ceará", "Alagoas", "Sergipe", "Rio Grande do Norte", "Paraíba",
    "Maranhão", "Piauí"}, "Nordeste",
    'base_vendas-petshop (1)'[ds_estado] IN {"Goiás", "Mato Grosso", "Mato Grosso do Sul", "Distrito Federal"}, "Centro-Oeste",
    'base_vendas-petshop (1)'[ds_estado] IN {"Amazonas", "Acre", "Amapá", "Pará", "Rondônia", "Roraima", "Tocantins"}, "Norte",
    "OUTRO"
)
```

Fonte: Elaborado pelos autores.

Além de resolver erros, a linguagem DAX foi de ótima serventia para criar colunas condicionais que serviram de grande ajuda para a posterior análise de dados. Foram utilizadas as funções COUNTROWS (para fazer uma medida de todos os registros da base) e, novamente, a função SWITCH para criar uma coluna condicional que poderá ser utilizada para análises posteriores e resolver – possivelmente – problemas de outliers com números extremos, apesar de que, no resultado final do projeto, não foi utilizada.

Imagem 14 e 15 - Comandos DAX para criar colunas e medidas condicionais.

```
ct_compra =
    SWITCH(
        TRUE(),
        'base_vendas-petshop (1)'[(não usar) vl_total] >= 150000, "Compra Paranormal",
        'base_vendas-petshop (1)'[(não usar) vl_total] >= 100000, "Compra Extrema",
        'base_vendas-petshop (1)'[(não usar) vl_total] >= 50000, "Compra Gigante",
        'base_vendas-petshop (1)'[(não usar) vl_total] >= 20000, "Compra Grande",
        'base_vendas-petshop (1)'[(não usar) vl_total] >= 1000, "Compra Média",
        "Compra Regular"
    )
```

```
Quantidade Vendas = COUNTROWS('base_vendas-petshop (1)')
```

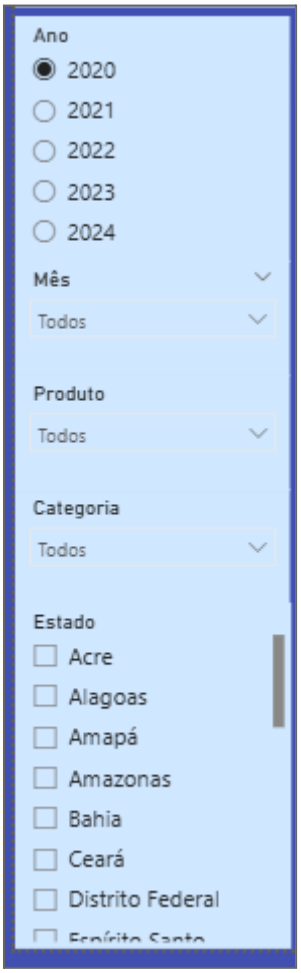
Fonte: Elaborado pelos autores.

4. ANÁLISE DE DADOS & TOMADA DE DECISÃO

Por fim, para criar uma análise de dados, foram usados diversos tipos de gráficos para gerar uma boa tomada de decisão, além de terem sido criados três dashboards com relações de total de vendas melhores funcionários e estados que tiveram mais lucros.

Foram criados filtros no Power BI para permitir a segmentação dos dados. O filtro de **ano** foi configurado com seleção única, evitando a sobreposição de anos em, por exemplo, um gráfico de linhas com todos os meses, fazendo assim um mês ter dados de dois anos. Já os filtros de mês, produto, categoria e estado/região permitem que o destinatário visualize diferentes perspectivas do negócio, tornando o dashboard mais interativo.

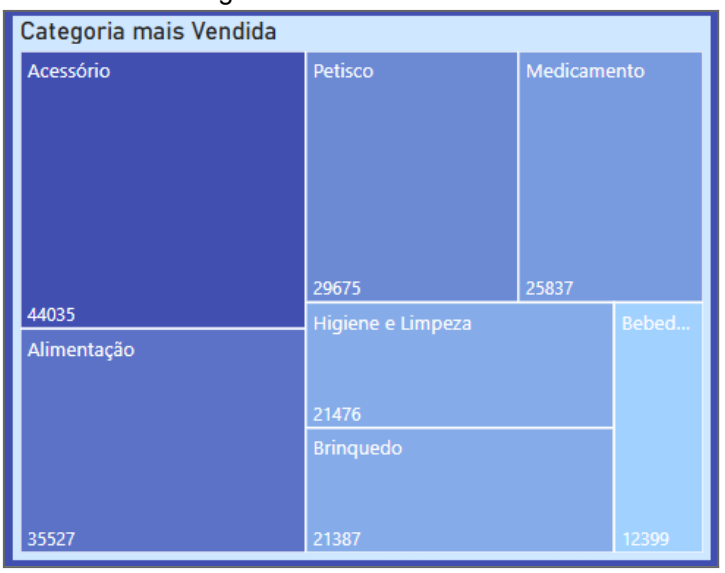
Imagem 16 - Filtros dos dashboards.



Fonte: Elaborado pelos autores.

O gráfico de árvore evidencia as categorias com maior volume de vendas, permitindo identificar rapidamente quais tipos de produtos impulsionam as vendas. Essa visualização hierárquica destaca a quantidade de produtos da categoria que já foram vendidos.

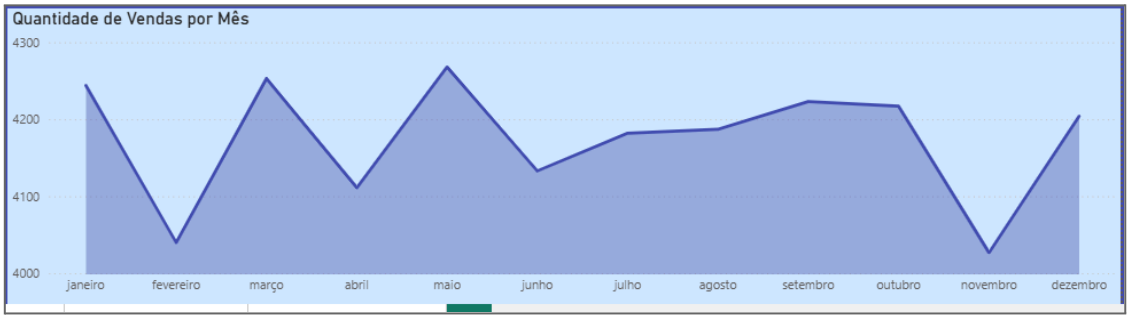
Imagem 17 - Gráfico de Árvore.



Fonte: Elaborado pelos autores.

O gráfico de linhas exibe a evolução temporal das vendas ou do lucro líquido (pois tem duas vertentes desse mesmo gráfico) ao longo do ano. Ele permite identificar tendências e períodos de alta ou baixa performance.

Imagem 18 - Gráfico de Linhas.



Fonte: Elaborado pelos autores.

As tabelas (Papai Jean ama elas) fornecem uma visão detalhada dos principais indicadores por produto, atendente e estado (pois, há três vertentes das tabelas), incluindo quantidade vendida, valor bruto, lucro líquido e comissão. Essa estrutura permite comparações e visão de desempenhos individuais, além de ser útil para a ordenação dos dados.

Imagem 19 - Tabela dos dashboads.

Produto	Total Vendas	Valor Bruto	Lucro Líquido
Nutri Alimentador Inteligente Automático Câmera Google Alexa	2075	R$ 4.981.569,60	R$ 1.335.895,50
Antipulgas e Carrapatos MSD Bravecto para Pet de 2 a 10 Kg	2096	R$ 2.485.735,84	R$ 1.218.202,70
Ração Royal Canin Club Performance para Cães Adultos	2082	R$ 1.953.132,16	R$ 700.294,00
Ração Royal Canin Exigent Gatos Adultos 1,5Kg	2088	R$ 1.009.309,28	R$ 448.170,40
Cama Coração Coroa Pet Nest Almofada Lavável	2075	R$ 778.692,32	R$ 417.894,00
Vitamina E Granulado BigForce	2100	R$ 697.930,24	R$ 338.774,15
Kit Banho e Tosa com Escova	2083	R$ 451.772,16	R$ 319.410,40
Total	**50089**	**R$ 16.340.401,28**	**R$ 6.613.876,49**

Fonte: Elaborado pelos autores.

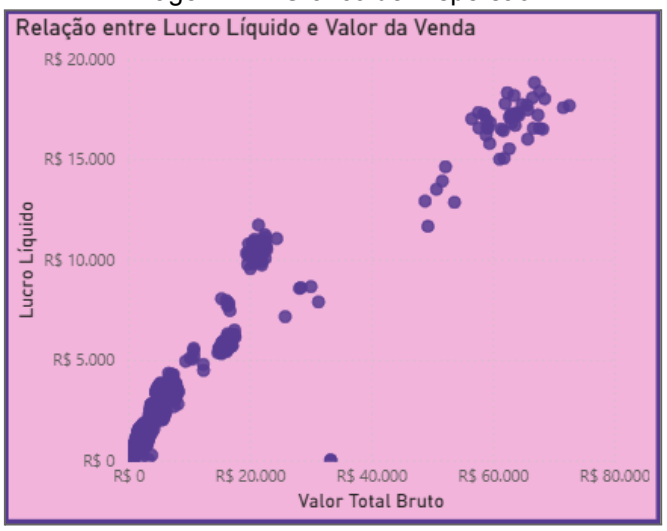
Os cartões destacam dados solitários: funcionário do mês, lucro líquido do mês/ano e estado com maior desempenho em vendas ou lucro (várias vertentes para esse gráfico). Esses indicadores fazem as informações mais relevantes terem uma leitura rápida e objetiva para a tomada de decisão.

Imagem 20 - Cartões dos dashboards.

Fonte: Elaborado pelos autores.

O gráfico de dispersão demonstra a relação entre o valor bruto das vendas e o lucro líquido, evidenciando que vendas de maior valor tendem a gerar lucros mais elevados. Essa análise permite identificar possíveis outliers, antes mesmo de criar a análise de dados (foi muito útil para encontrar registros fora da curva, e, mensagem pessoal do autor, gostei bastante que deu certo esse gráfico, o Jean gosta do de dispersão).

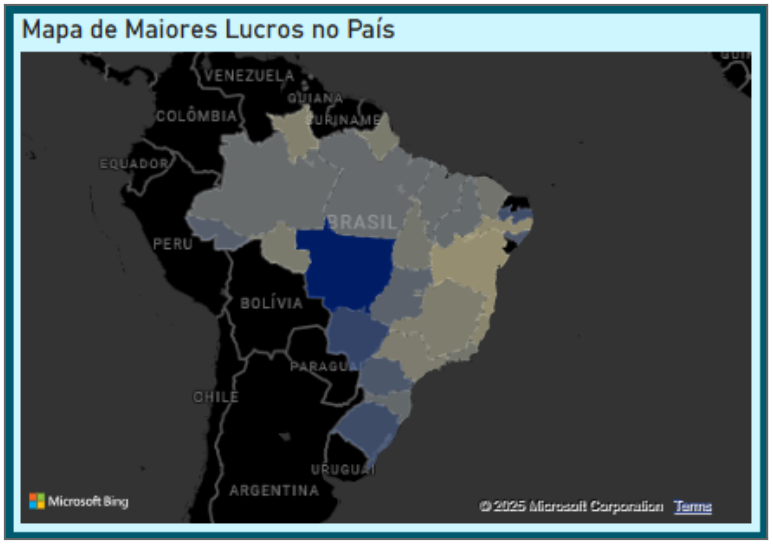
Imagem 21 - Gráfico de Dispersão.



Fonte: Elaborado pelos autores.

O mapa coroplético traz uma análise geográfica das vendas e dos lucros (pois há duas vertentes dele), destacando visualmente os estados com melhor e pior desempenho e facilitando a identificação de oportunidades regionais.

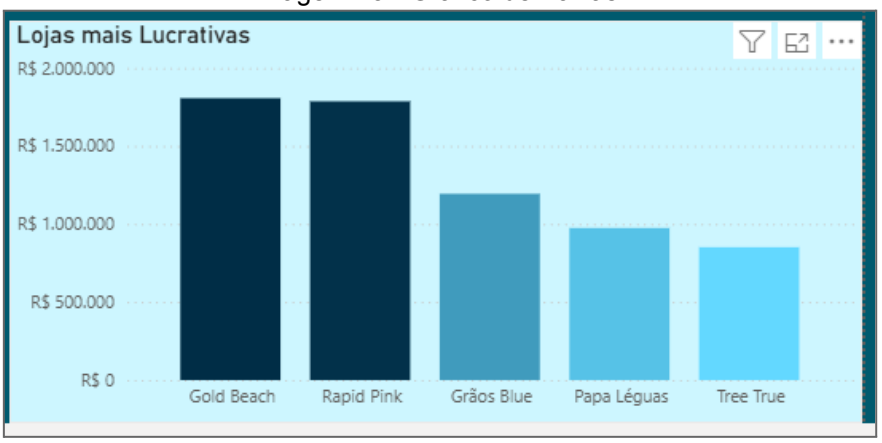
Imagem 22 - Mapa Coroplético.



Fonte: Elaborado pelos autores.

O gráfico de barras mostra as lojas que geram maior lucro ao pet shop. A visualização facilita decisões sobre os locais de melhores oportunidades de investimento para aumentar a lucratividade.

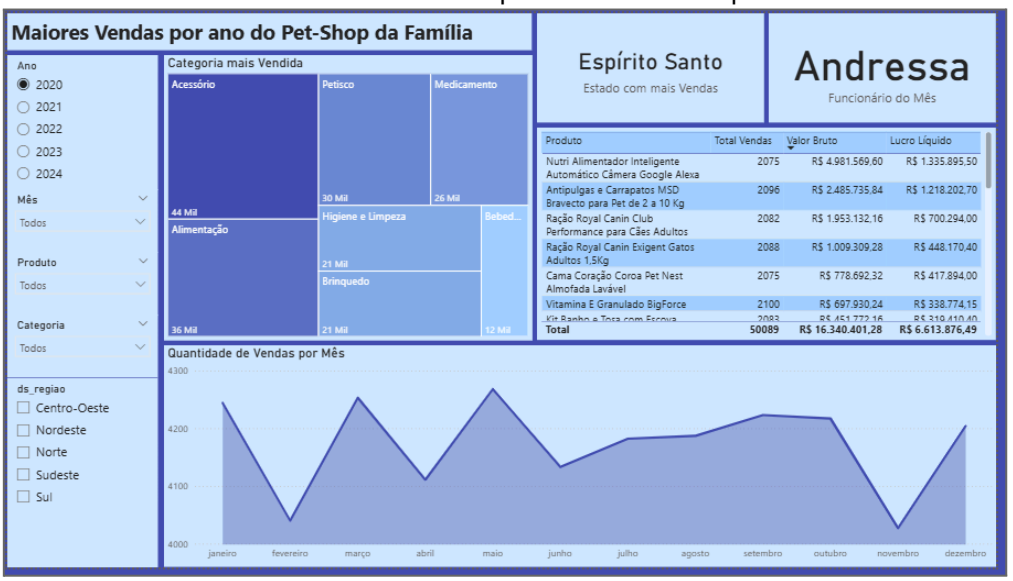
Imagem 23 - Gráfico de Barras.



Fonte: Elaborado pelos autores.

O projeto finalizou com três dashboards — Vendas do Pet Shop, Melhores Funcionários e Desempenho por Região/Estado — oferecendo uma visão ampla do negócio. Combinando gráficos, tabelas e mapas, é possível entender a performance da causa e tomar decisões assertivas. Abaixo, há uma representação não interativa dos três dashboards realizados.

Dashboard 1 - Maiores Vendas por ano do Pet-Shop da Família.

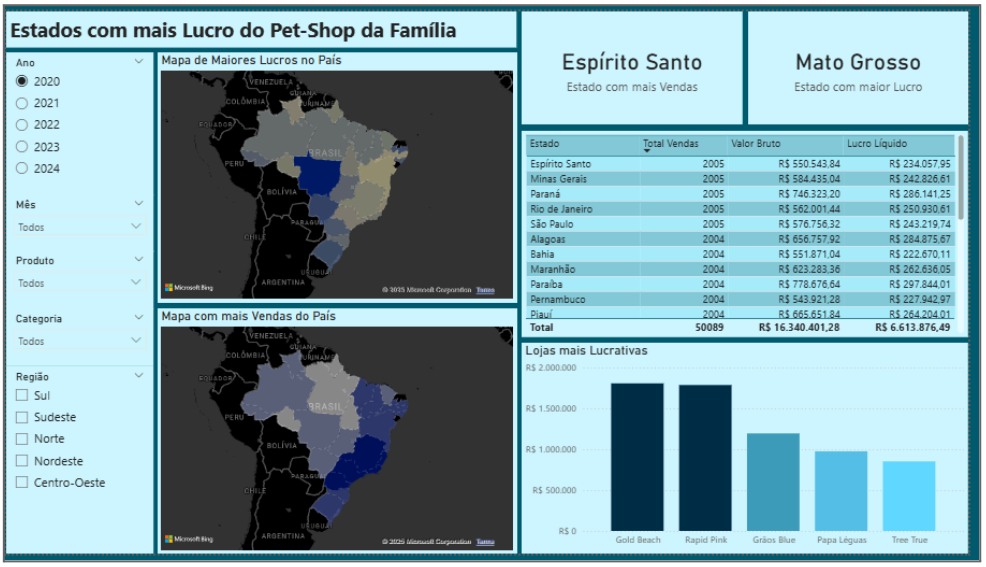


Fonte: Elaborado pelos autores.

Dashboard 2 - Vendedores e lucros por ano do Pet-Shop da Família.

Fonte: Elaborado pelos autores.

Dashboard 3 - Estados com mais Lucro do Pet-Shop da Família.



Fonte: Elaborado pelos autores.

Finalmente, os dashboards possibilitam identificar regiões com baixo desempenho e produtos mais lucrativos, permitindo ações como reforçar campanhas nessas áreas, ajustar a demanda de produtos específicos e valorizar vendedores de melhor resultado, podendo usá-los como inspiração para outros e promovê-los. Assim, o projeto se torna uma ferramenta para orientar decisões que aumentem as vendas e promovam o crescimento do negócio.