ANÁLISE DE DADOS

Pedro Henrique Pedroso da Cruz

MODELAGEM — Boas práticas Modelagem

SQL — Boas práticas SQL para extração de Insights

API — API e Boas Práticas de desenvolvimento em Notebook

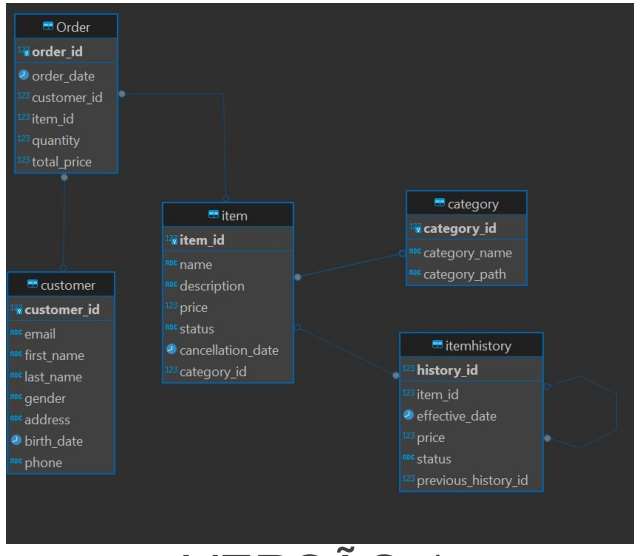
ANÁLISE DE DADOS — Argentina

PROBLEMA

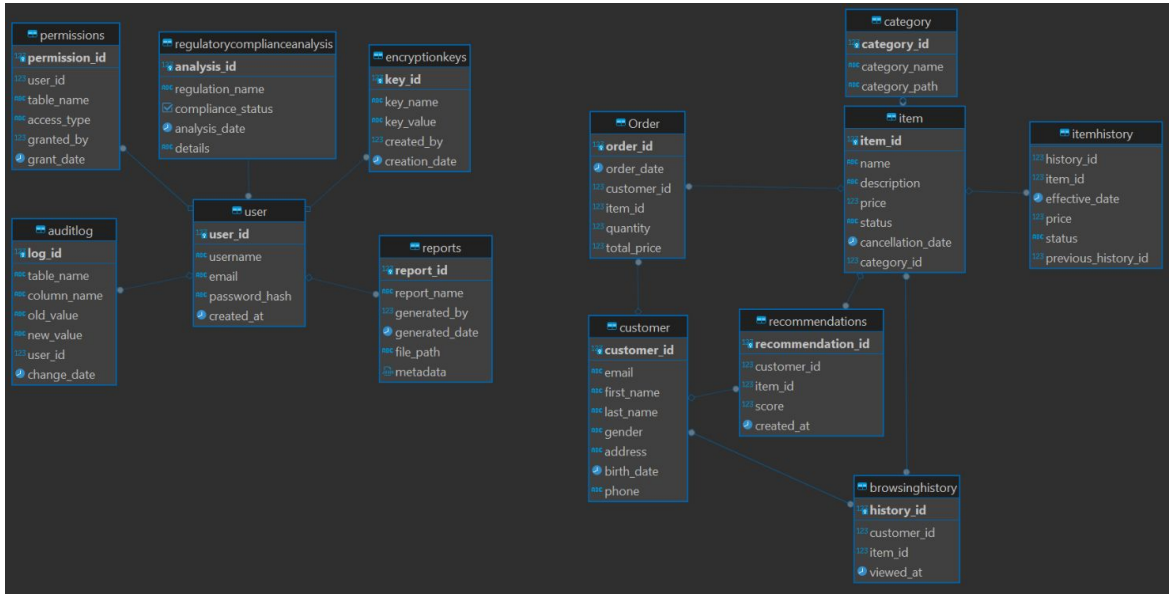
MODELAGEM

TRANSACIONAL

- Normalização dos Dados – 1NF, 2NF, 3NF
- Chaves Pk e FK bem definidas, garantia de unicidade, garantia de integridade referencial.
- Consistência e Integridade dos Dados – Restrições como Unique, Not Null, Check, FK
- Trigger e Stored Procedures para manter regras complexas
- Indices – Criação de índices para WHERE, Join, Order nas colunas mais usadas
- Particionamento de Tabelas
- Documentação dos Campos – Catalogo de Dados, geração automática de documentação
- Governança de Dados – Garantia de acesso aos devidos usuários



VERSÃO 1



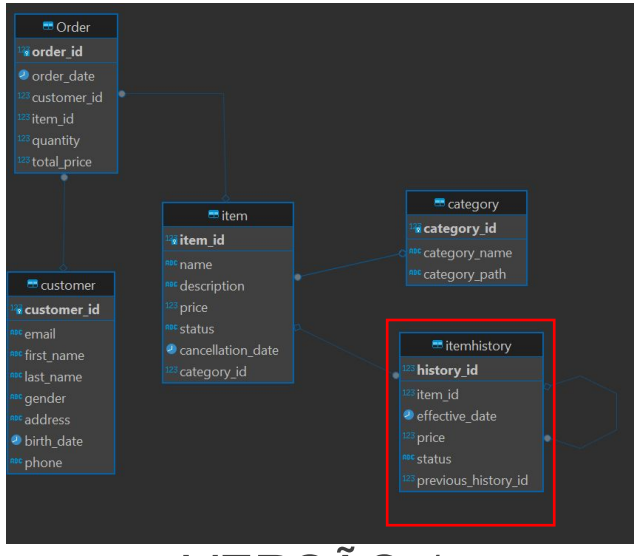
VERSÃO 2

MODELAGEM

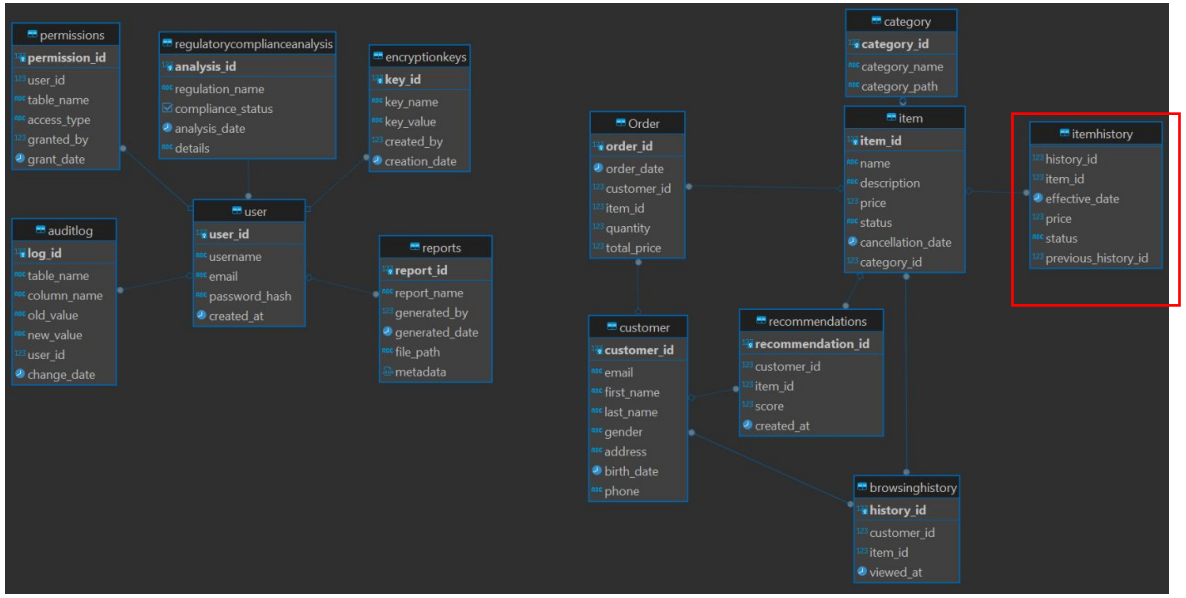
TABELA: ITEMHISTORY

- Itemhistory - É a entidade para armazenar as alterações dos preços dos itens. A solução para resolver o problema de histórico de itens seria a técnica Slowly --Changing Dimensions (SCD) em um data warehouse onde seria possível combinar os conceitos de controle de versão para garantir uma trilha de auditoria transparente.
- A solução permite a visualização de item quando alterado mais de uma vez ao dia, permitindo o tracking de forma transparente e já preparada para outros campos da tabela item.
    - Uma outra solução possível seria criar um TimeStamp para cada alteração e no final do dia executar uma stored procedure (batch e não em tempo real) para recuperar todos os itens alterados. Em ambos os casos devemos analisar a melhor para não impactar o banco de dados transacional durante o período de utilização.

•Estratégia de Hot, Cold Data e Expurgo de Dados



VERSÃO 1



VERSÃO 2

SQL – EXTRAÇÃO DE INSIGHTS

SQL: BOAS PRÁTICAS

- Uso de CTE (Common Table Expression) – Facilita a leitura de sub consultas
- Utilização de alias com “as num_sales”
- Filtragem de Data para dar escalabilidade
- Filtragem Condicional com EXISTS – Usar exists para uma sub consulta de maneira eficiente
- Nomenclatura de fácil entendimento
- Comentário relevante

```sql
-- Liste usuários com aniversário de hoje cujo número de vendas realizadas em janeiro de 2020

WITH January2020Sales AS (
    SELECT customer_id, COUNT(*) AS num_sales
    FROM "Order"
    -- Filtrar aqui o Ano e Mês desejado
    WHERE EXTRACT(year FROM order_date) = 2020
    AND EXTRACT(month FROM order_date) = 1
    GROUP BY customer_id
)
SELECT *
FROM Customer
-- Filtrando aniversario pela data de hoje
WHERE DATE_PART('month', birth_date) = DATE_PART('month', CURRENT_DATE)
AND DATE_PART('day', birth_date) = DATE_PART('day', CURRENT_DATE)
AND EXISTS (
    SELECT 1
    FROM January2020Sales
    WHERE January2020Sales.customer_id = Customer.customer_id
    -- Filtrar aqui o numero de vendas desejado
    AND January2020Sales.num_sales > 1500
);
```

SQL – EXTRAÇÃO DE INSIGHTS

SQL: BOAS PRÁTICAS

- Uso de Funções de Janela (Window Function) – Quando possível usar função de janela em vez de subconsultas.

```
--Para cada mês de 2020, são solicitados os 5 principais usuários que mais venderam (R$) na categoria --Celulares. São obrigatórios o mês e ano da análise

-- View para encontrar os Top 5, adicionando Rank para particionar os dados por cada mês e seu total, retornando assim um rank com os dados por mês
WITH MonthlyTopFiveSellers AS (
    SELECT
        EXTRACT(year FROM o.order_date) AS year,
        EXTRACT(month FROM o.order_date) AS month,
        o.customer_id,
        ROW_NUMBER() OVER (PARTITION BY EXTRACT(year FROM o.order_date), EXTRACT(month FROM o.order_date) ORDER BY SUM(o.total_price) DESC) AS rank,
        c.first_name,
        c.last_name,
        SUM(o.quantity) AS total_quantity,
        SUM(o.total_price) AS total_sales
    FROM
        "Order" o
    JOIN
        Item i ON o.item_id = i.item_id
    JOIN
        Customer c ON o.customer_id = c.customer_id
    JOIN
        Category cat ON i.category_id = cat.category_id
    WHERE
        EXTRACT(year FROM o.order_date) = 2020
        --Adicionar aqui a categoria desejada
        AND cat.category_name = 'Celulares'
    GROUP BY
        EXTRACT(year FROM o.order_date),
        EXTRACT(month FROM o.order_date),
        o.customer_id,
        c.first_name,
```

API E NOTEBOOK

API: DOWNLOAD DE DADOS

- Documentação Clara – Utilizar docstring detalhadas para explicar a rotina
- Separação de Responsabilidade – Divisão entre as rotinas get_item_details e main
- Uso de Context Manager – with open(output_file.... Garante que CSV seja aberto e fechado corretamente.
- Verificação de Resposata da API – Verificar o retorno 200 antes de processar.
- Tratamento de Erro

API: PRÓXIMOS PASSOS

- Parallelismo e Concorrência – Uso de concorrência com threads e async
- Tratamento de Erro mais Robusto
- Gestão de Logs mais robusta
- Lógica de retry
- Limitação de Taxa (Rate Limit)
- Paginação de Resultados
- Parâmetros dos campos buscados e não hardcode.

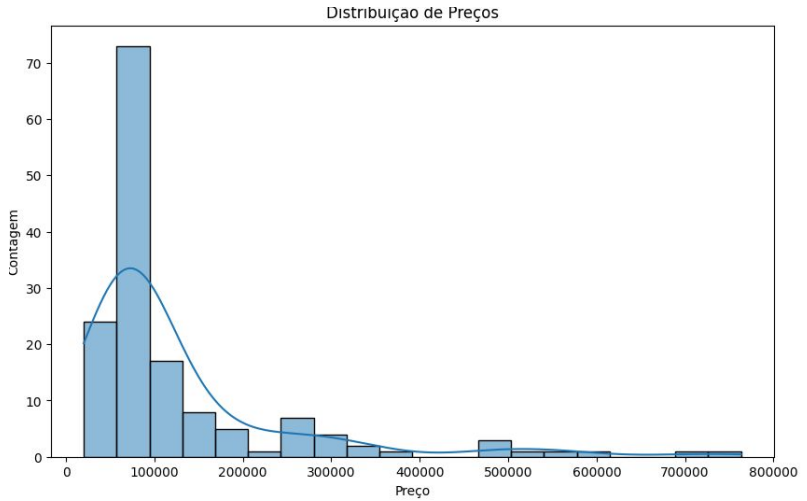
EXPLORAÇÃO DE DADOS

ANÁLISE EXPLORATÓRIA

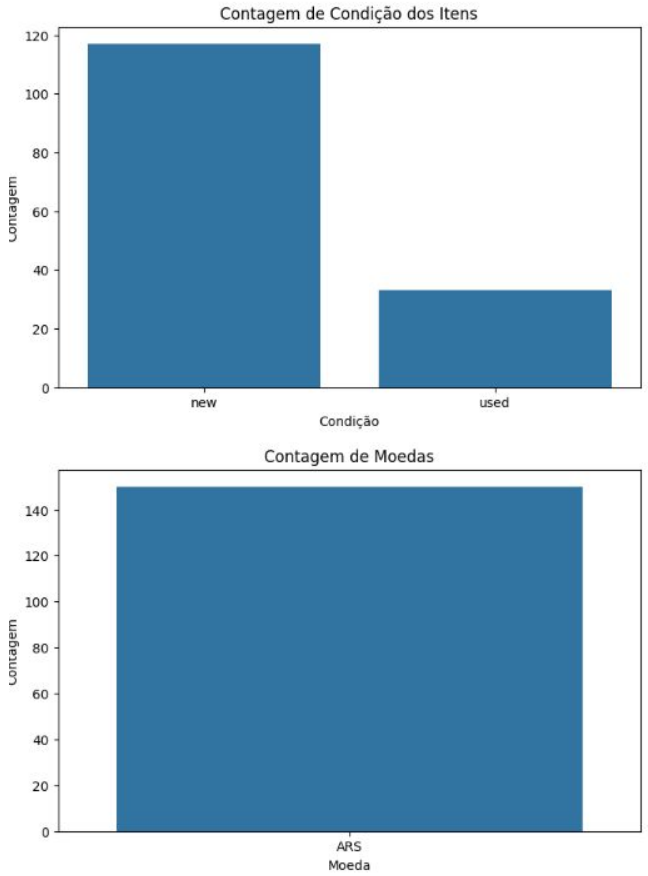
- Carregar e inspecionar os dados
- Limpeza dos Dados
- Análise Descritiva – Estatística Descritiva, Distribuições
- Visualização dos Dados
- Análise de Outliers
- Análise de Correlação
- Segmentação e Agrupamento
- Documentação e Storytelling

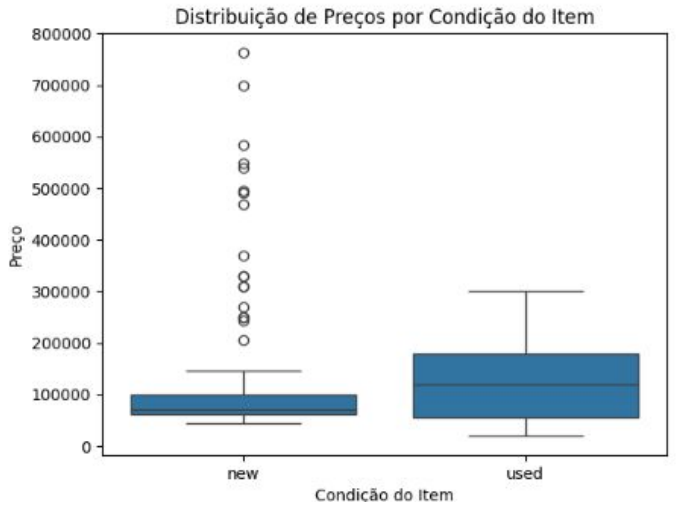
Estatísticas descritivas das variáveis numéricas:

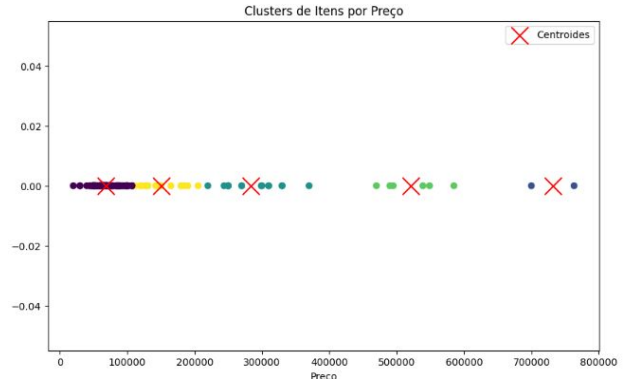
	price	available_quantity
count	150.000000	0.0
mean	127643.046733	NaN
std	129233.476652	NaN
min	20000.000000	NaN
25%	62499.000000	NaN
50%	74500.000000	NaN
75%	127798.550000	NaN
max	763400.000000	NaN

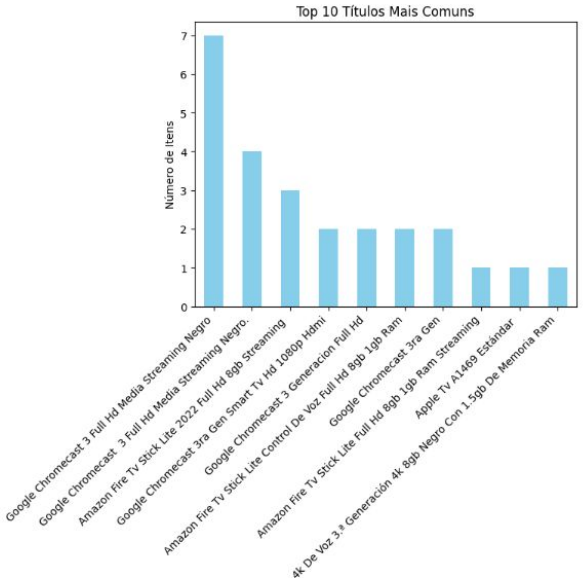


EXPLORAÇÃO DE DADOS









EXPLORAÇÃO DE DADOS

Resultado da Análise

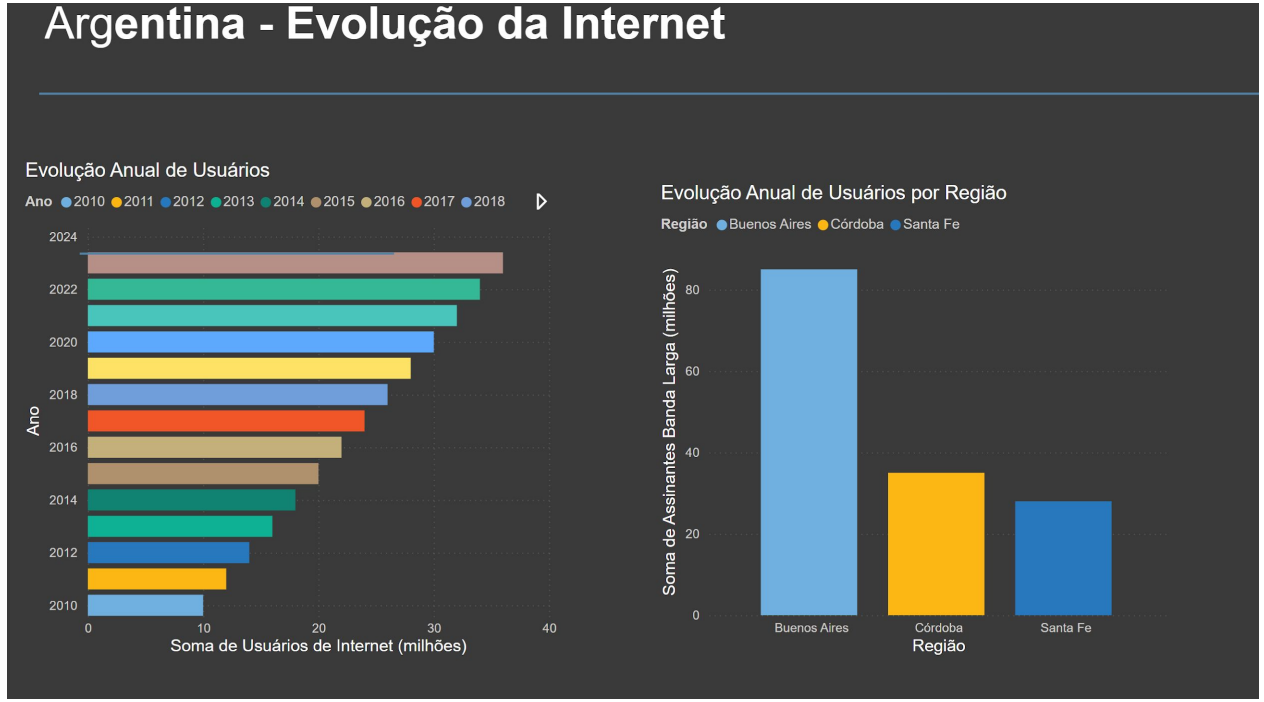
- Analisamos alguns campos apenas que extraimos da API, sendo gerado um schema com as colunas:
  - 0 item_id 150 non-null object 1 title 150 non-null object 2 condition 150 non-null object 3 permalink 150 non-null object 4 price 150 non-null float64 5 currency 150 non-null object 6 available_quantity 0 non-null float64

- Percebemos um problema no campo available_quantity que deixamos proposital para demonstrar que através de análise exploratória encontramos problemas na distribuição dos campos, campos vazios, outliers, entre outros. Uma sequencia de analise foram feitas para encontrar e demonstrar quais seriam os outliers, uma possível anomalia nos dados, ou até mesmo produtos com valores muito acima da média, o que seria uma questão a discutir o que fazer para analise especificas.
- No final analise de insights, como: Top 10 Títulos Mais Comuns Relação entre Preço e Quantidade Disponível Proporção de Moedas Utilizadas

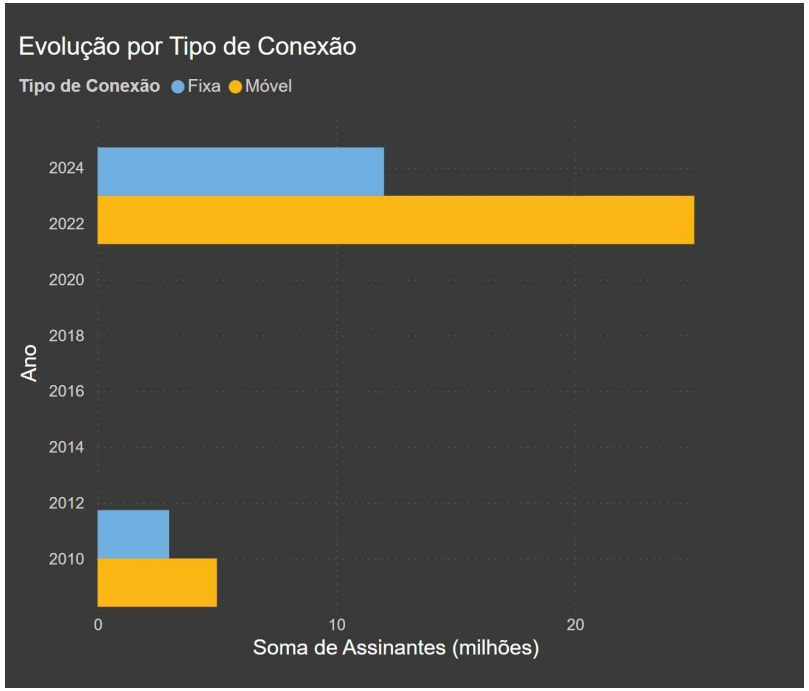
ARGENTINA

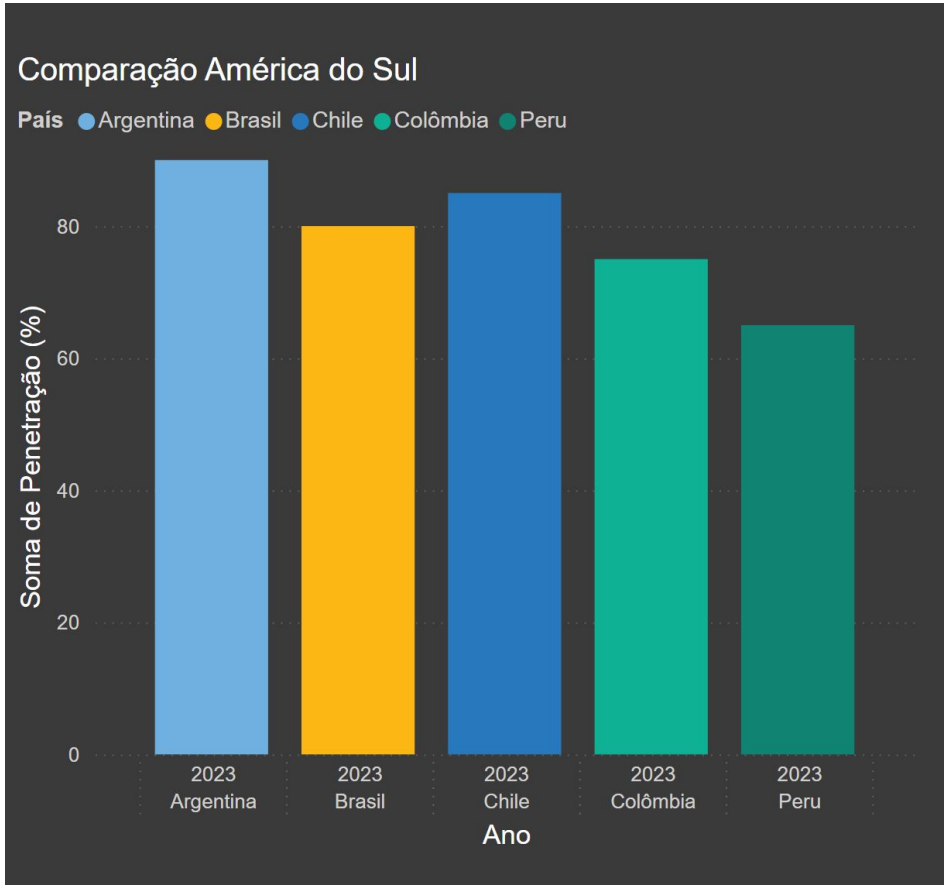
DASHBOARDS

- Objetivo claro e especifico para publico alvo
- Graficos para apresentar as informações na linha do tempo (Graficos de Linhas)
- Gráficos comparativos (Graficos de Barras)
- Possiveis correlações e tendências
- Impacto nas áreas de maior valor
- Gráficos alto nível e baixo nível (Drill down)
- Técnicas do 5 Why´s
- Ações de Predição ou Prescrição

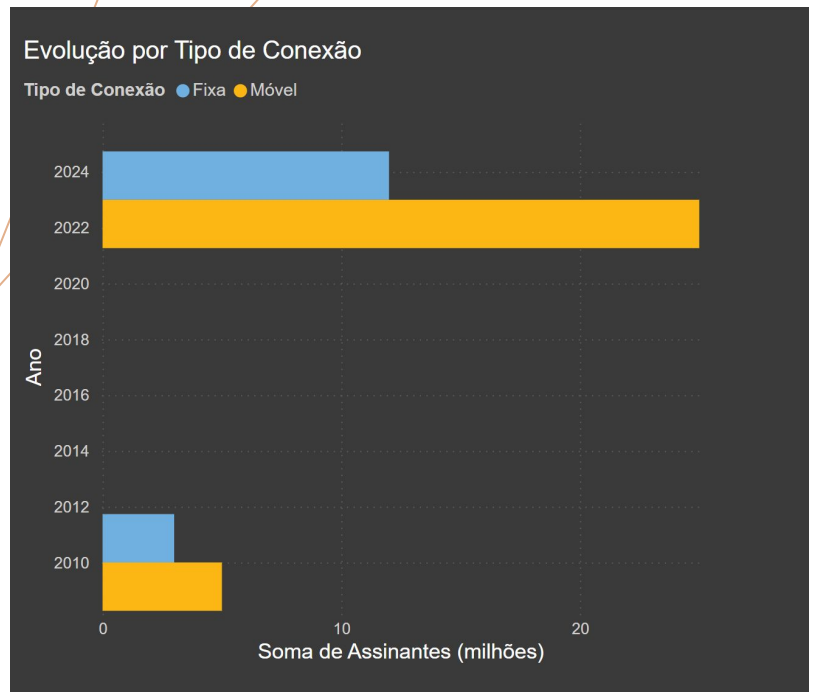


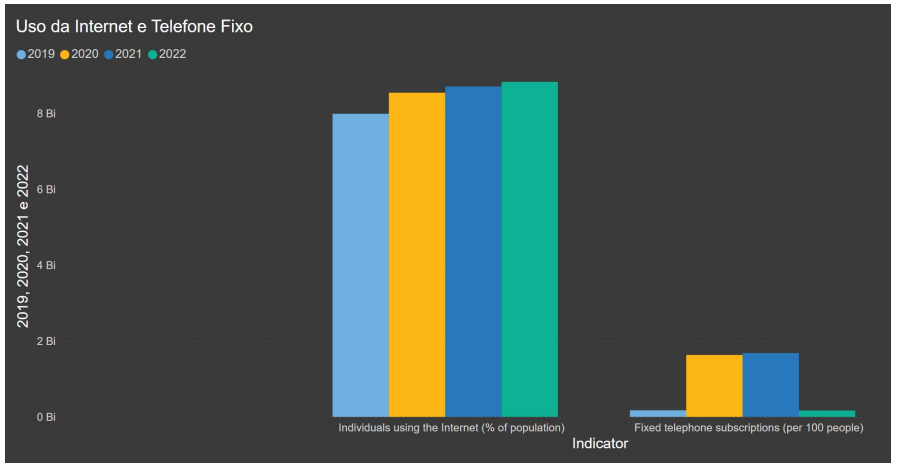
ARGENTINA

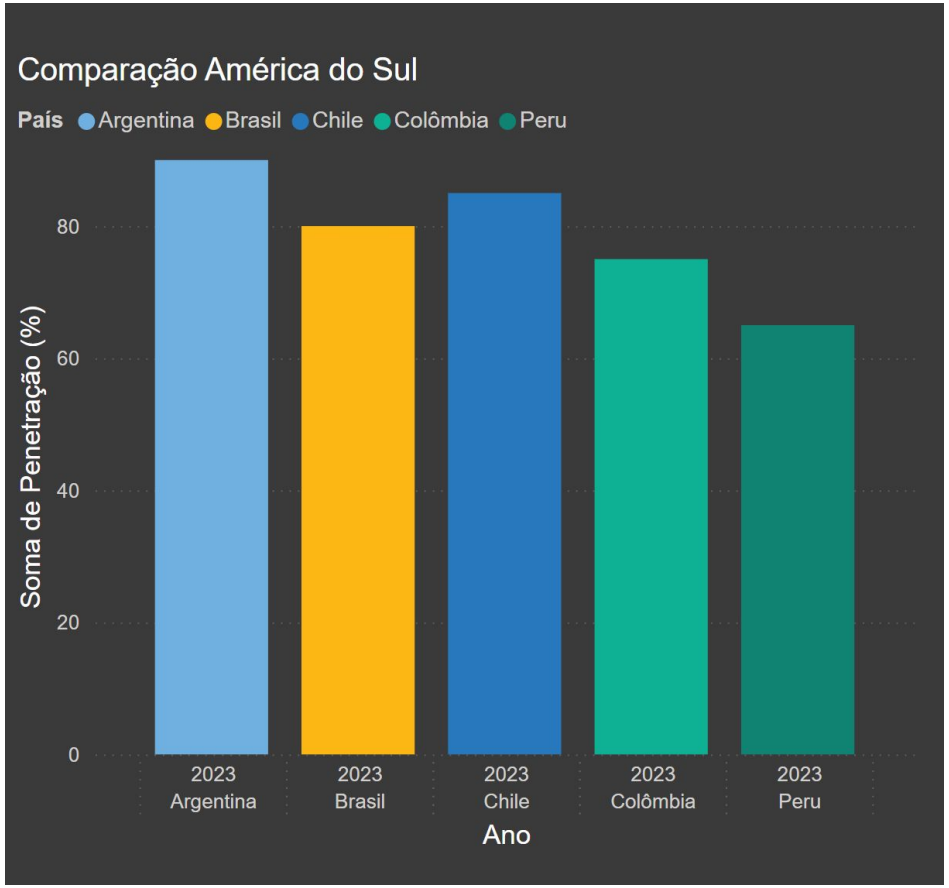

Evolução por Tipo de Conexão
Tipo de Conexão
Fixa
Móvel
2024
2022
2020
2018
2016
2014
2012
2010
Ano
0
10
20
Soma de Assinantes (milhões)


Comparação América do Sul
País
Argentina
Brasil
Chile
Colômbia
Peru
80
60
40
20
0
Soma de Penetração (%)
2023
Argentina
2023
Brasil
2023
Chile
2023
Colômbia
2023
Peru
Ano

ARGENTINA


Evolução por Tipo de Conexão
Tipo de Conexão Fixa Móvel
2024
2022
2020
2018
2016
2014
2012
2010
Ano
0
10
20
Soma de Assinantes (milhões)


Uso da Internet e Telefone Fixo
2019 2020 2021 2022
8 Bi
6 Bi
4 Bi
2 Bi
0 Bi
2019, 2020, 2021 e 2022
Individuals using the Internet (% of population)
Fixed telephone subscriptions (per 100 people)
Indicator


Comparação América do Sul
País Argentina Brasil Chile Colômbia Peru
80
60
40
20
0
Soma de Penetração (%)
2023
Argentina
2023
Brasil
2023
Chile
2023
Colômbia
2023
Peru
Ano

OBRIGADO

Pedro Henrique Pedroso da Cruz