MVP3 – Análise Dados Seguros

ANÁLISE DE DADOS DE UMA SEGURADORA

1 - INTRODUÇÃO

A análise de dados desempenha um papel fundamental no setor de seguros, permitindo uma melhor compreensão dos riscos, aprimoramento da precificação de apólices e otimização dos processos operacionais. Para este estudo, utilizaremos as informações de:

- Prêmios Emitidos
- Sinistros Pagos

Com essas informações, queremos identificar a "saúde financeira" da seguradora. Para uma melhor análise, seriam necessários outros dados (por exemplo, comissão paga, despesas administrativas etc.), mas infelizmente esses dados não estavam disponíveis.

Sendo assim, o objetivo desta análise consiste em identificar:

- *Os prêmios emitidos por competência em 2024*
  - Identificar se a seguradora está conseguindo distribuir seus produtos no mercado
- *Os Sinistros pagos por competência em 2024*
  - Identificar se os produtos possuem um risco alto de sinistros
- *Quais os produtos que mais vendem (3 mais vendidos)*
  - Identificar os produtos com mais participação no mercado, para possíveis ajustes dos mesmos e aumento de participação (futuro)
- *Comparação entre Prêmios Emitidos X Sinistros Pagos dos 3 produtos mais vendidos em 2024*
  - Identificar se os produtos mais comercializados da empresa possuem altos riscos

2 - ARQUIVOS

As informações a serem observadas estão nos arquivos:

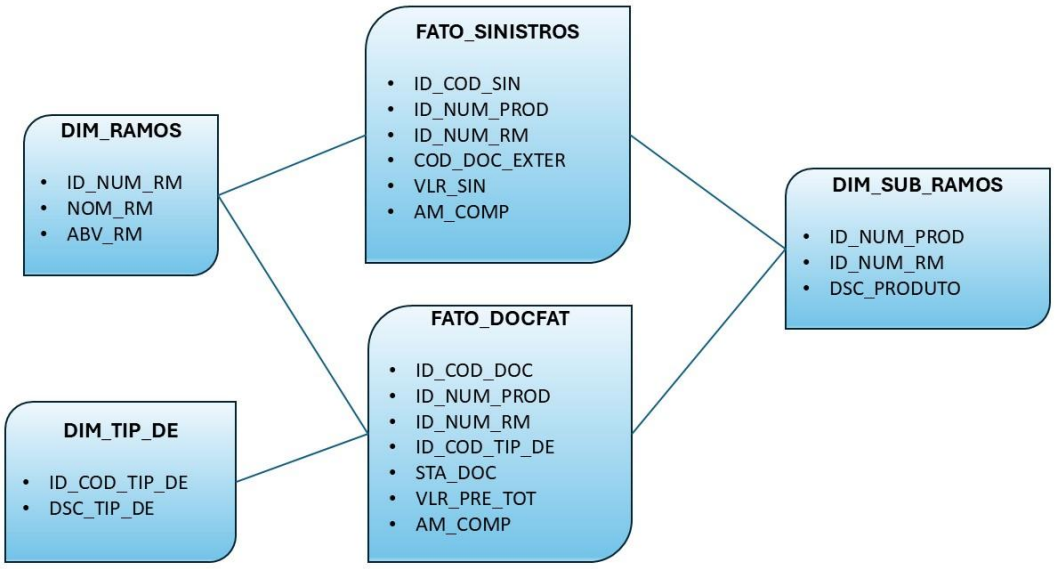
- DOCFAT.zip
  - Refere-se a todos os documentos e faturas emitidas em 2024
- RAMOS.zip
  - Informações referentes aos ramos (Dado SUSEP)
- SUB_RAMOS.zip
  - Esta informação, em conjunto com a RAMOS identifica o produto cadastrado na seguradora
- SINISTROS.zip
  - Refere-se aos sinistros avisados em 2024
- SINISTROS_PGTO.zip
  - Informações sobre os pagamentos realizados

3 – MODELO DE DADOS

Com o crescimento exponencial dos dados, as empresas buscam soluções eficientes para armazenar, organizar e analisar grandes volumes de informações. Nesse contexto, os Data Warehouses desempenham um papel fundamental, permitindo a integração de dados de diferentes fontes para apoiar a tomada de decisões estratégicas.

Para o estudo em questão, utilizaremos um Data Warehouse modelo Estrela. Esse modelo é caracterizado por uma tabela fato, que armazena métricas numéricas e transacionais, e tabelas dimensão, que fornecem informações descritivas para análise, permitindo a construção de visões detalhadas e agregadas dos dados.

Abaixo, segue o modelo do projeto



Descrição das tabelas:

Tabela	Descrição	Campos	Tipo	Descrição
DIM_RAMOS	Contêm os códigos e os nomes dos ramos de seguro	ID_NUM_RM	Numérico	Código do Ramo
		NOM_RM	String	Nome do ramo
		ABV_RM	String	Abreviatura do nome do Ramo
DIM_SUB_RAMOS	Contêm os códigos e os nomes dos produtos vendidos na seguradora	ID_NUM_PROD	Numérico	Código do Produto
		ID_NUM_RM	Numérico	Código do Ramo
		DSC_PRODUTO	String	Descrição do Produto - Nome
FATO_DOCFAT	Contêm todas as apólices, endossos e faturas emitidas no período de JAN/2024 até DEZ/2024	ID_COD_DOC	String	Código da Apólice
		ID_NUM_PROD	Numérico	Código do Produto
		ID_NUM_RM	Numérico	Número do Ramo
		ID_COD_TIP_DE	Numérico	Tipo de Documento (Proposta, Endosso, Fatura, etc)
		STA_DOC	String	Statusn do Documento
		VLR_PRE_TOT	Numérico	Prêmio toral emitido
		AM_COMP	Numérico	Competência do documento
FATO_SINISTROS		ID_COD_SIN	String	Código do Sinistro

	Contêm todas os sinisttros ocorridos e pagos no período de JAN/2024 até DEZ/2024	ID_NUM_PROD	Númérico	Código do Produto
		ID_NUM_RM	Numérico	Número do Ramo
		COD_DOC_EXTER	String	Código da Apólice
		VLR_SIN	Numérico	Valor do Sinistro
		AM_COMP	Numérico	Competência do Sinistro

4 – PLATAFORMA

A plataforma utilizada foi a Databricks Community Edition

5 – DETALHAMENTO

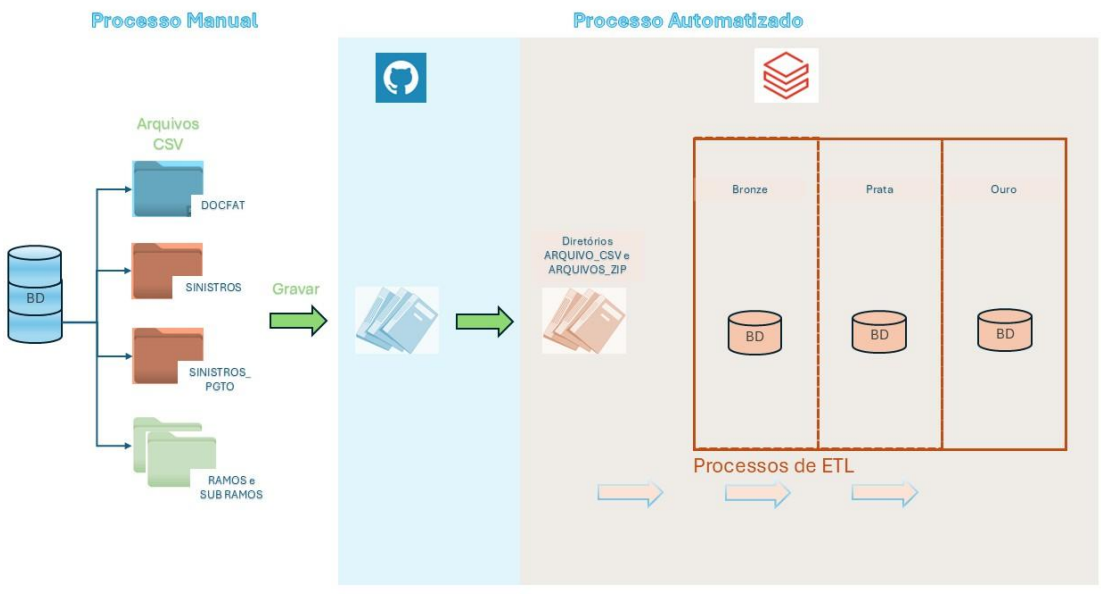
5.1 – Busca pelos Dados

Foram usados dados fictícios (base de desenvolvimento) de uma seguradora. As informações foram descaracterizadas para confidencialidade.

5.2 – Coleta

As informações foram retiradas de suas fontes originais e gravados em arquivos CSVs.. Os arquivos foram dispostos zipados no GITHUB. O processo inicia buscando os arquivos neste diretório do GITHUB e gravando dentro do Databricks no diretório ***dbfs:/FileStore/tables/arquivos_zip.*** Depois os arquivos ZIP foram abertos e jogados no diretório ***dbfs:/FileStore/tables/arquivos_csv***

Abaixo, o desenho do pipeline da operação



5.3 – Modelagem

O modelo construído foi um Datawarehouse em Esquema Snowflake conforme mostrado no item 3 – Modelo de Dados

5.4 – Carga

Abaixo, segue a descrição do processo ETL utilizado

- *Leitura dos arquivos*
  - Leitura dos arquivos brutos CSV

- *Transformação/Inclusão de campos*
  - Ajustes realizados

TABELA	SCHEMA	CAMPO	O QUE FOI REALIZADO
DOCFAT	bronze para prata	AM_COMP	• Inclusão da informação AM_COMP para os casos em que ela estava como 0. O cálculo foi baseado na DT_EMISS do mesmo arquivo
DOCFAT	bronze para prata	VLR_PRE_TOT	• Verificar se a informação é numérica, caso contrário ajustar a informação como 0 (zero). Isso foi necessário para ajustar os casos em que a informação estava como NULL
DOCFAT	prata para ouro	Campos Necessários	• Selecioneis apenas as informações ("COD_SIN", "COD_DOC_EXTER", "VLR_SIN", "AM_COMP","NUM_RM", "NUM_SUB_RM") para serem armazenadas no schema
DOC_SINISTROS	bronze para prata	AM_COMP	• Inclusão do campo AM_COMP a partir da informação DT_AVISO. Esta ação foi necessária para ajudar no gráfico para "quebrar" por competência (AM_COMP)
DOC_SINISTROS	bronze para prata	VLR_SIN	• Inclusão da informação VLR_SIN na tabela DOC_SINISTROS. A informação VLR_SIN estava na tabela DOC_SINISTROS_PGTO • Verificar se a informação VLR_SIN é NULL. Se SIM, ajustar como zero (0)
DOC_SINISTROS	bronze para prata	NUM_SUB_RM	• Inclusão da informação NUM_SUB_RM na tabela DOC_SINISTROS. Esta

			informação estava na tabela DOCFAT
DOC_SINISTROS	Prata para ouro	Campos Necessários	• Selecionei apenas as informações ("COD_SIN","NUM_RM", "COD_DOC_EXTER","DT_AVISO","AM_COMP","VLR_SIN", "NUM_SUB_RM") para serem armazenadas no schema

5.4 – Análise

A- Qualidade dos Dados

Analisando o arquivo, foram identificados alguns ajustes que precisaram ser realizados

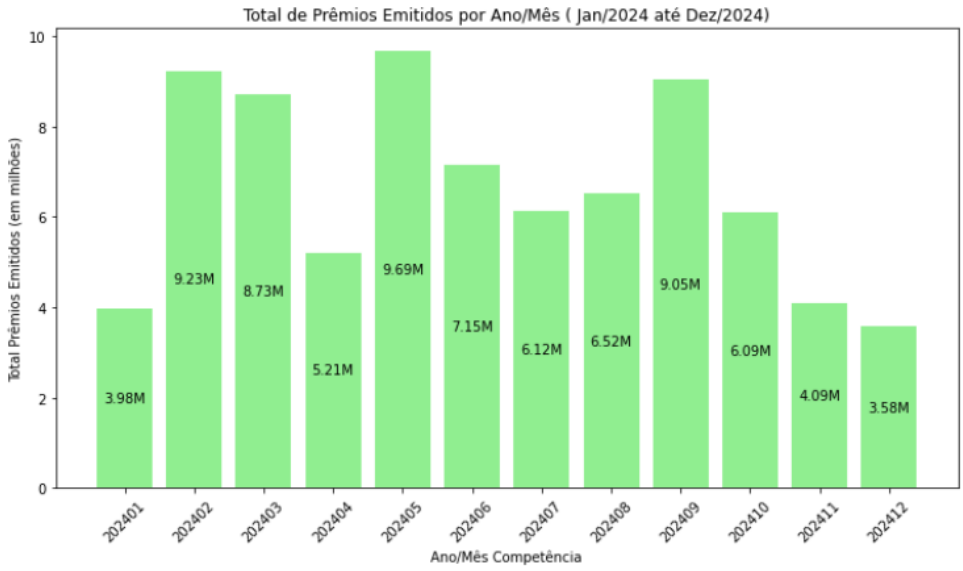
- Existia uma linha no meio do arquivo, com informações truncadas e a mesma estava influenciando nos cálculos do programa (parava com erro). A solução foi retirar esta linha já que foi identificado que se tratava de "lixo"
- Ajustes mencionados no item 5.4, de forma a facilitar a análise
- Retirada de colunas com as informações NULL que não seriam necessários para o estudo

Fora estas questões, os dados apresentavam-se com ótima qualidade

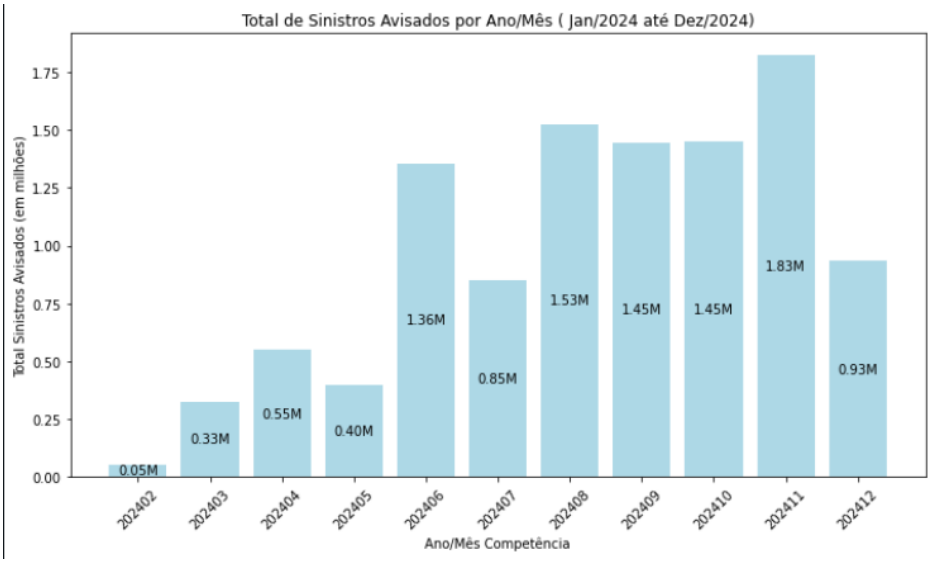
B – Solução do Problema

Abaixo, seguem as respostas para os problemas mencionados

- Os prêmios emitidos por competência em 2024



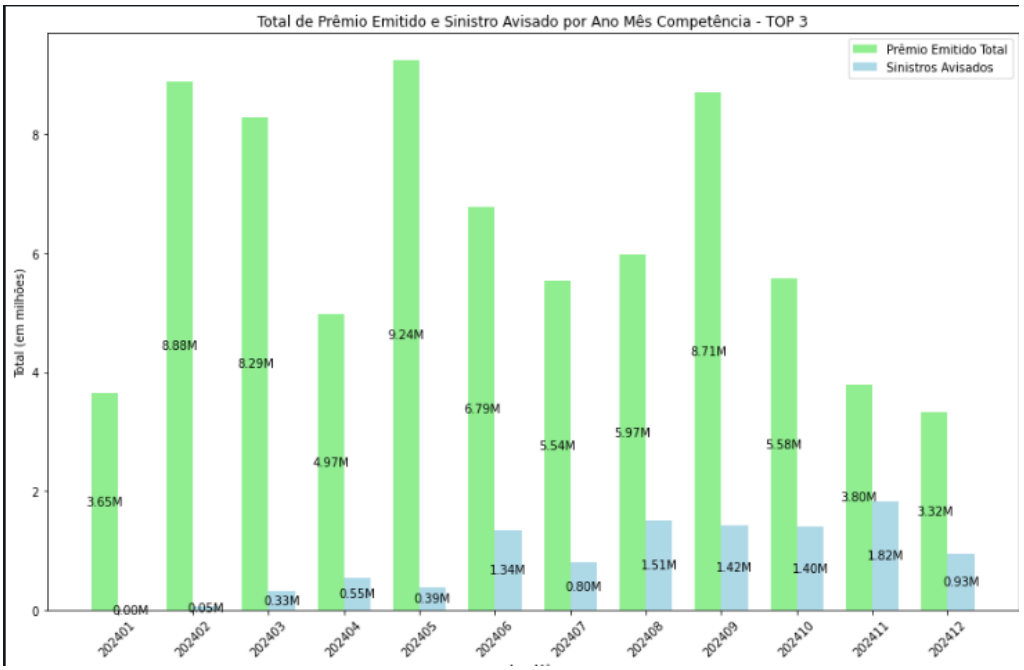
- Os Sinistros Pagos por competência em 2024



- Quais os produtos que mais vendem (3 mais vendidos)

NUM_RM	NUM_SUB_RM	TOTAL_VENDAS	DSC_SUB_RM
53	0	4.5744e+07	RESPONSABILIDADE CIVIL
53	1	1.59127e+07	RCF-V - GARANTIA ÚNICA (DM/DC)
93	1	1.32632e+07	VG

- Comparação entre Prêmios Emitidos X Sinistros Pagos dos 3 produtos mais vendidos em 2024



5 – AUTOAVALIAÇÃO

De acordo com os objetivos solicitados, acredito que tenha conseguido atingir os objetivos delineados.

A minha maior dificuldade foi a utilização do Databricks, já que se tratava de uma plataforma que eu nunca tinha trabalhado. Tive muitas dificuldades iniciais (queria acessar o Databricks usando o Google Colab, mas não consegui). Sendo assim, usei o notebook do próprio Databricks para a realização dos trabalhos. Foi esse trabalho que me fez definir sobre os próximo passos futuros: Investir em Engenharia de Dados

6 – PROGRAMA

O programa encontra-se no GitHub:
https://github.com/phfgomes1969/PUC-RJ-MVP3/blob/main/MVP3%20-%20Seguros.pdf