1. Objetivo do projeto

O objetivo deste projeto é analisar o perfil de risco de segurados da categoria tarifária "Automóveis comuns (passeio nacional e importado)" utilizando dados históricos de sinistros. Através da análise de variáveis como tipo de veículo, perfil do cliente e região, busco identificar características de risco que podem impactar a subscrição de seguros. Além disso, o projeto visa otimizar a seleção de riscos e minimizar perdas financeiras, garantindo a confiabilidade das informações através de etapas de limpeza, tratamento e análise exploratória dos dados.

2. A fonte de dados utilizada e suas características

Os dados utilizados são provenientes da base de dados AutoSeg, disponibilizada pela SUSEP (Superintendência de Seguros Privados).

A principal fonte de dados utilizada para a análise foi a tabela **arq_casco_comp**, que contém informações detalhadas sobre sinistros e prêmios de seguros de automóveis.

Feature Engineering

Para a derivação de variáveis de risco, utilizei a tabela **arq_casco_comp** em conjunto com outras tabelas auxiliares que complementam as informações:

1. **auto2_vei**: Contém o código e a descrição de cada modelo de veículo, além do código do grupo a que pertence.
2. **auto2_grupo**: Fornece o código e a descrição dos grupos de modelos.
3. **auto_cat**: Contém o código e a descrição de categorias tarifárias, permitindo a classificação dos veículos.
4. **auto_cau**: Apresenta o código e a descrição das causas de sinistros, ajudando a entender os motivos dos sinistros.
5. **auto_idade**: Contém o código e a descrição de faixas etárias dos veículos, que podem influenciar a probabilidade de sinistros.
6. **auto_reg**: Fornece o código e a descrição das regiões de circulação, crucial para análise territorial dos riscos.

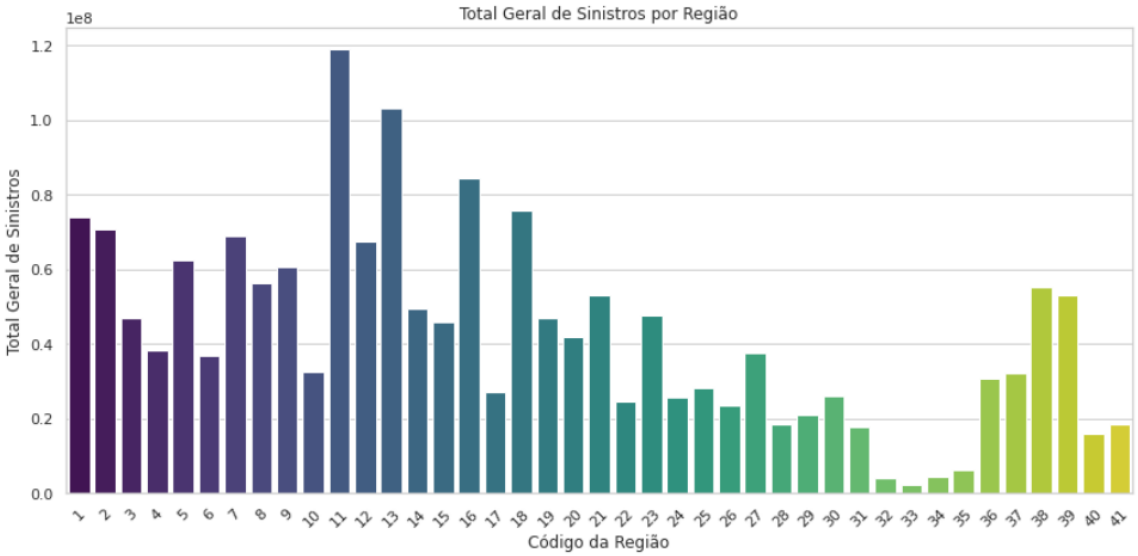
3. Qual é a categoria tarifária que você irá analisar?

"Automóveis comuns (passeio nacional e importado)" - Essa escolha é justificada pelo tamanho da base de dados, que é ampla e permite uma análise mais robusta dos sinistros nessa categoria.

4. Dentro dessa categoria tarifária, quais riscos você não aceitará de acordo com a Política de Subscrição da sua empresa?

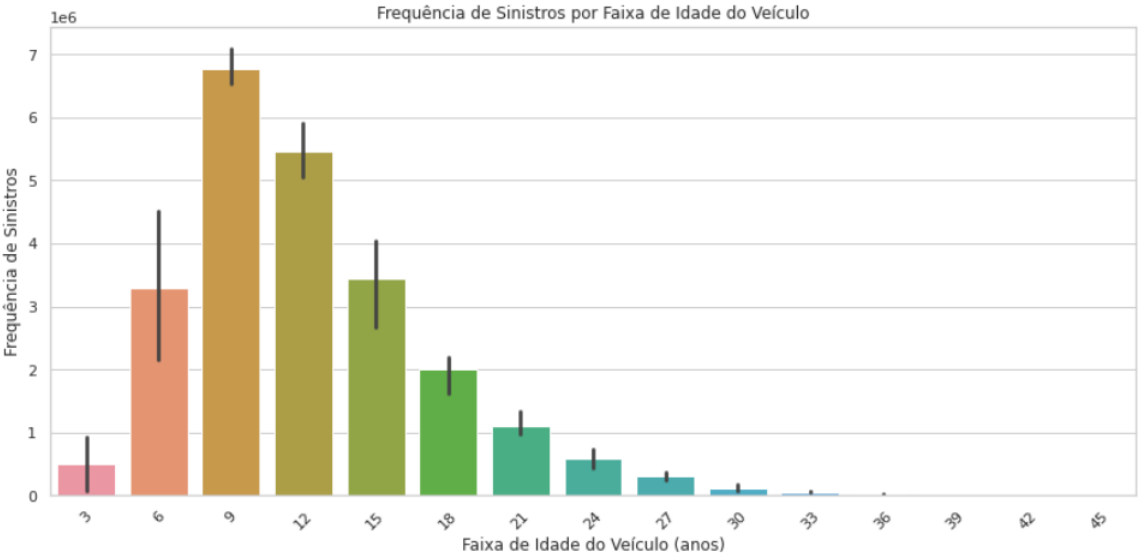
Os riscos que não serão aceitos incluem:

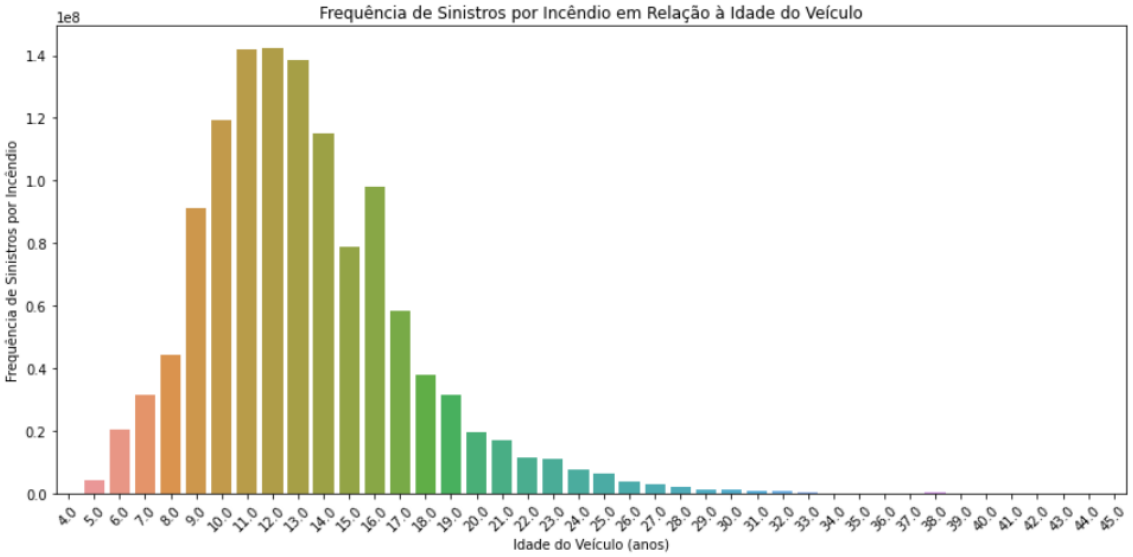
- Veículos registrados nas regiões 11 (Região metropolitana de SP) e 13 (Ribeirão Preto e Campinas), devido ao elevado número de sinistros nessas áreas.

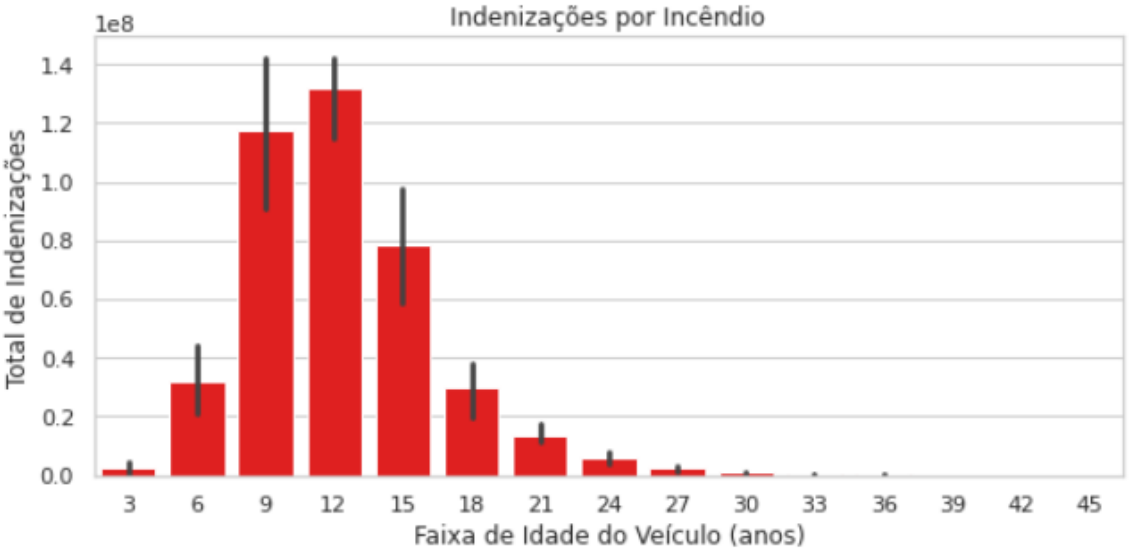


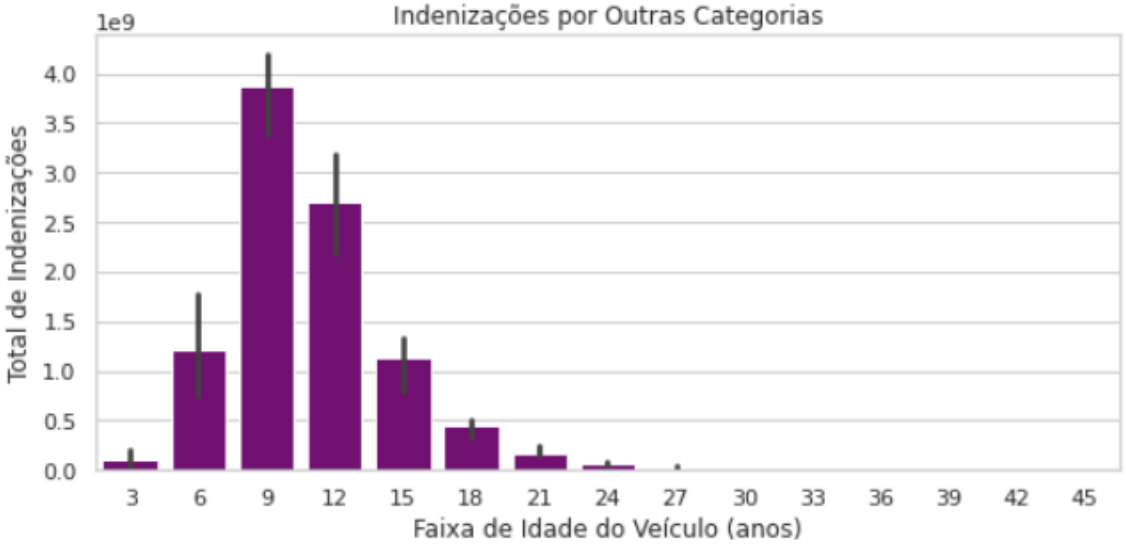
Obs: Os sinistros analisados por categoria demonstraram o mesmo comportamento das regiões 11 e 13 liderando.

- Veículos com mais de 10 anos apresentam uma maior probabilidade de incêndio e falhas mecânicas, e por consequência, aumenta o risco de sinistros.

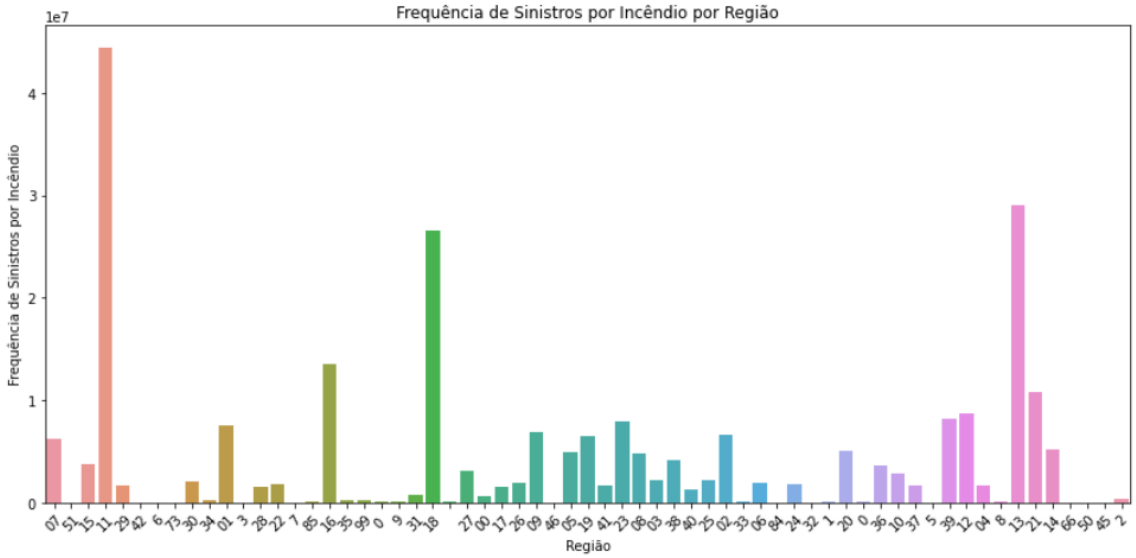



Frequência de Sinistros por Incêndio em Relação à Idade do Veículo
1e8
Frequência de Sinistros por Incêndio
Idade do Veículo (anos)


Indenizações por Incêndio
1e8
Total de Indenizações
Faixa de Idade do Veículo (anos)


Indenizações por Outras Categorias
1e9
Total de Indenizações
Faixa de Idade do Veículo (anos)

Ao analisar a frequência de sinistros por incêndio nas diferentes regiões, o gráfico abaixo enfatiza a necessidade de excluir as regiões 11 e 13. Os gráficos anteriores mostram que veículos mais antigos estão associados a um risco elevado de incêndios e problemas mecânicos, reforçando a decisão de não aceitar esses riscos.



- Não incluir o parâmetro de sexo, uma vez que a categoria tarifária é muito ampla e a segmentação excessiva poderia limitar a base de clientes.

5. Qual é a experiência histórica que você irá utilizar? Por que você escolheu esses períodos?

Utilizei os dados de 2015 a 2019, excluindo 2020 devido à pandemia de COVID-19, que poderia distorcer os números de sinistros devido a lockdowns e outras restrições. Essa escolha permite uma análise mais precisa das tendências históricas sem a interferência de anomalias externas.

6. Como você avaliou a razoabilidade dos dados?

A razoabilidade dos dados foi avaliada por meio de análises de consistência, incluindo a verificação de valores extremos, ausência de dados e comparação de frequências de sinistros em diferentes categorias. Utilizei gráficos de distribuição para visualizar a variabilidade e identificar possíveis outliers. Além disso, analisei a relação entre variáveis para garantir que fossem lógicas.

7. Quais transformações você realizou para melhorar a confiabilidade da informação?

1. Limpeza de Dados: Removi duplicatas e registros incompletos.
2. Tratamento de Valores Ausentes: Imputei valores ausentes com a média ou o valor mais frequente.
3. Normalização de Dados Textuais: Normalizei colunas de texto (minúsculas e remoção de espaços).
4. Tratamento de Outliers: Removi outliers que poderiam distorcer os resultados.
5. Codificação One-Hot: Apliquei codificação one-hot em colunas categóricas.

6. Normalização de Variáveis Numéricas: Utilize o StandardScaler para normalizar variáveis numéricas.

8. Qual é o tamanho da sua amostra?

Período de Análise: 2015 a 2019, excluindo 2020 para evitar distorções da pandemia.

Critérios:

- Tipo de Veículo: Apenas "Automóveis comuns (passeio nacional e importado)".
- Idade do Veículo: Até 10 anos.
- Regiões Excluídas: 11 e 13, devido ao alto número de sinistros.

```
%python
# Contar o número de registros que atendem aos critérios especificados
sample_size_query = """
SELECT COUNT(*) AS Numero_Registros
FROM arq_casco_comp_tratado v
JOIN auto_cat_tratado a ON v.COD_TARIF = a.CODIGO
WHERE a.CODIGO = 1  -- Apenas veículos de passeio
  AND (2024 - v.ANO_MODELO) <= 10  -- Idade do veículo <= 10 anos
  AND v.REGIAO NOT IN (11, 13)  -- Excluir regiões 11 e 13
"""

# Executar a consulta e carregar o resultado em um DataFrame
sample_size_df = spark.sql(sample_size_query).toPandas()

# Mostrar o número de registros
print(f"Número de registros na amostra: {sample_size_df['Numero_Registros'][0]}")
```

```
▸ (3) jobs Spark
Número de registros na amostra: 22078215
```

9. Você acredita que seus dados têm credibilidade, tanto em relação ao tamanho da amostra quanto em relação a sua qualidade?

Sim, os dados têm credibilidade. A amostra é representativa, cobrindo 2015 a 2019, e foi tratada para minimizar inconsistências. Além disso, as fontes, como a SUSEP, são confiáveis, o que reforça a qualidade das informações.

10. Existe alguma limitação dessa informação que você deve documentar e comunicar?

As limitações incluem:

- A exclusão de dados de 2020, que pode resultar em uma visão distorcida das tendências se compararmos com períodos anteriores.
- A falta de uma categoria específica para carros elétricos é uma limitação. No momento, não estou disposto a assumir esse risco, mas sei que os dados são de uma época em que os veículos elétricos e híbridos ainda não eram tão populares. Isso pode impactar nas previsões.

https://github.com/victorvsaraujo/AutoSeg-Risk-Evaluation