



---

**FACULDADE DE TECNOLOGIA DE São José dos Campos**  
**Curso Superior de Tecnologia em Logística**

Pedro H. Hernandes Fonseca  
Matheus Oliveira Alexandre  
Kauê Oliveira Venâncio  
Leonardo Rocha Alves

**RELATÓRIO TÉCNICO REFERENTE A SINISTROS DE TRÂNSITO**

**São José dos campos, SP**

## RESUMO

Este relatório técnico descreve o processo de consolidação e análise de dados de sinistros de trânsito referentes às malhas viárias administradas pela ARTESP (2015-2025), pelo DER (2023-2025) e dados consolidados do RENAEST. O objetivo principal foi criar uma base de dados unificada e higienizada para alimentar um dashboard de Business Intelligence (Power BI), permitindo a identificação de padrões de accidentalidade e o suporte à tomada de decisão. A metodologia envolveu a utilização da linguagem de programação Python e suas bibliotecas para extração, transformação e carregamento (ETL) dos dados brutos, com destaque para a escolha estratégica do RENAEST devido à maturidade de seus dados. A análise fundamenta-se também em estudos do Ipea e literatura acadêmica sobre o impacto dos acidentes na saúde pública e economia. Os resultados obtidos compõem uma ferramenta visual que correlaciona variáveis como localização, tipologia das ocorrências e severidade, visando contribuir para a gestão da segurança viária.

**Palavras-chave:** Sinistros de Trânsito; Python; Power BI; Logística; Segurança Viária.



## ABSTRACT

This technical report describes the consolidation and analysis process of traffic accident data regarding road networks managed by ARTESP (2015-2025), DER (2023-2025), and consolidated data from RENAEST. The main objective was to create a unified and sanitized database to feed a Business Intelligence dashboard (Power BI), enabling the identification of accident patterns and supporting decision-making. The methodology involved using the Python programming language and its libraries for Extract, Transform, and Load (ETL) of raw data, highlighting the strategic choice of RENAEST due to its data maturity. The analysis is also grounded in IPEA studies and academic literature regarding the impact of accidents on public health and the economy. The results comprise a visual tool that correlates variables such as location, occurrence typology, and severity, aiming to contribute to road safety management.

**Keywords:** Traffic Accidents; Python; Power BI; Logistics; Road Safety.

<b>1. INTRODUÇÃO .....</b>	<b>4</b>
1.1 - PROBLEMA E DELIMITAÇÃO DA ÁREA PESQUISADA.....	5
Delimitação da Pesquisa .....	5
1.2. OBJETIVOS .....	6
Objetivo Geral .....	6
Objetivos Específicos .....	6
1.3. Justificativa .....	7
1.4. Metodologia .....	7
<b>2- CARACTERIZAÇÃO DA ORGANIZAÇÃO (FONTES DE DADOS).....</b>	<b>7</b>
2.1. Agência Reguladora de Serviços Públicos Delegados de Transporte do Estado de São Paulo (ARTESP) .....	8
2.2. Departamento de Estradas de Rodagem (DER) .....	8
2.3. Outras Fontes e Contexto Regulatório (IPEA / RENAEST) .....	9
<b>3. FUNDAMENTAÇÃO TEÓRICA .....</b>	<b>9</b>
3.1. Impacto Econômico e Análise Temporal .....	9
3.2. Impacto na Saúde Pública e Classificação de Severidade .....	10
<b>4. PROPOSTA DE SOLUÇÃO E DESENVOLVIMENTO .....</b>	<b>12</b>
4.1. Estratégia de Seleção de Dados: A Escolha do RENAEST .....	12
4.2. Arquitetura da Solução (Pipeline ETL) .....	12
4.3. Detalhamento dos Scripts e Algoritmos .....	13
4.4. Integração com Business Intelligence (Power BI) .....	14



# 1. INTRODUÇÃO

Os **sinistros de trânsito** representam um dos maiores desafios para a saúde pública e para a logística de transportes no Brasil. Além das irreparáveis perdas humanas, tais eventos geram custos econômicos significativos, impactos na infraestrutura viária e sobrecarga no sistema de saúde.

A Organização Mundial da Saúde (OMS) classifica o tema como uma epidemia global. No contexto brasileiro, o Instituto de Pesquisa Econômica Aplicada (**IPEA**) reportou uma média anual superior a **40 mil mortes** na década de 2010 a 2019. Além da mortalidade, a literatura acadêmica aponta para os **impactos crônicos na saúde pública**, incluindo milhões de feridos com sequelas permanentes e as chamadas "sequelas invisíveis" (psicológicas e sociais), resultando em altos custos de reabilitação e sobrecarga do Sistema Único de Saúde (SUS).

A gestão moderna e a formulação de políticas públicas eficazes exigem uma abordagem orientada a dados. Contudo, essa visão sistêmica é dificultada pela **fragmentação das informações**. Os dados de accidentalidade no estado de São Paulo, por exemplo, estão distribuídos em fontes heterogêneas de diferentes órgãos gestores, como a **ARTESP** (rodovias concedidas) e o **DER** (rodovias estaduais).

Dessa forma, o presente relatório técnico descreve a solução desenvolvida para sanar essa dificuldade. O projeto utilizou ferramentas de programação **Python** e o processo de ETL (Extração, Transformação e Carga) para unificar e higienizar *datasets* complexos, criando uma base de dados robusta e consistente para análise estratégica via **Power BI**.

## 1.1 - PROBLEMA E DELIMITAÇÃO DA ÁREA PESQUISADA

O problema central abordado neste relatório é a **dispersão e a heterogeneidade dos dados de sinistros de trânsito** no Estado de São Paulo. As informações encontram-se fragmentadas entre concessionárias (**ARTESP**), gestão pública estadual (**DER**) e bases nacionais, muitas vezes em formatos incompatíveis (CSV, Excel, múltiplas abas) e com padrões de preenchimento divergentes, inviabilizando análises diretas.

Esta dispersão é crítica, pois impede a formação de uma base de conhecimento unificada e histórica que auxilie na prevenção. Tecnicamente, a heterogeneidade se manifesta na **ausência de um dicionário de dados único**, resultando em inconformidades como diferentes nomes para colunas idênticas e a presença de caracteres especiais ou acentos em campos de texto que dificultam a correta agregação e geolocalização (*conforme detectado nos datasets brutos da ARTESP e do DER, exigindo limpeza via Python*).

### Delimitação da Pesquisa

A área de pesquisa e desenvolvimento deste projeto delimita-se pela **análise e unificação dos datasets de acidentes de trânsito** dentro dos seguintes parâmetros:

- **Fontes Primárias:** Dados fornecidos pela ARTESP (referentes ao período de **2015 a 2025**) e pelo DER (referentes ao período de **2023 a 2025**).
- **Foco Metodológico:** O desafio reside em conciliar estes dados históricos e discrepantes, aplicando técnicas de **ETL em Python (Pandas)** para padronização.
- **Resultado Esperado:** Geração de um *dataset* único e limpo para consumo exclusivo em um *dashboard* de *Business Intelligence (Power BI)*.

O escopo limita-se, portanto, à **gestão da informação** (transformação de dados brutos em informação acionável) e não à formulação de políticas públicas, embora o resultado final forneça subsídios para tal.

## 1.2. OBJETIVOS

A finalidade deste projeto é suprir a lacuna informacional e técnica identificada na análise de sinistros de trânsito. O processo visa transformar dados brutos e dispersos em inteligência acionável, alinhada aos princípios de **Logística 4.0** e **Gestão Orientada a Dados**.

### Objetivo Geral

Desenvolver e implementar um **processo automatizado de Extração, Transformação e Carga (ETL)** de dados, utilizando a linguagem **Python**, para unificar bases de sinistros de trânsito heterogêneas. O resultado final é a alimentação de um *dashboard* gerencial de *Business Intelligence (Power BI)* que permita a análise preditiva e a identificação de pontos de intervenção na malha viária.

### Objetivos Específicos

Os objetivos específicos delineiam as etapas críticas do projeto e a aplicação das ferramentas técnicas:

- **Consolidar Bases Históricas:** Efetuar a coleta e o *merge* das bases de dados da **ARTESP (2015-2025)** e do **DER (2023-2025)**, garantindo a compatibilidade de *schema* entre os diferentes períodos e fontes.
- **Implementar Scripts de ETL em Python:** Utilizar bibliotecas como **Pandas** e **Regex** para desenvolver *scripts* customizados, focados na limpeza, padronização e normalização de dados (ex: remoção de acentos, tratamento de *missings*, padronização de datas).
- **Qualificar e Enriquecer Dados:** Integrar, quando aplicável e viável, informações qualificadas de outras fontes relevantes, como o **RENAEST** (Registro Nacional de Acidentes e Estatísticas de Trânsito) ou dados de geolocalização.
- **Entregar Visualização Interativa:** Criar painéis e *dashboards* no **Power BI** com visualizações interativas (gráficos de tendência, mapas de calor, *drill-down*), capacitando os usuários a extrair *insights* rápidos para o suporte à decisão e planejamento de segurança viária.

### *1.3. Justificativa*

A unificação dessas bases é vital para validar hipóteses econômicas e sociais. Segundo o IPEA, existe uma correlação direta entre o desenvolvimento econômico e a taxa de mortalidade no trânsito. Compreender essa dinâmica através de dados higienizados permite aos gestores públicos e privados otimizar recursos de fiscalização e engenharia de tráfego, reduzindo custos logísticos e salvando vidas.

### *1.4. Metodologia*

O projeto utilizou uma abordagem quantitativa e aplicada. A metodologia técnica consistiu no desenvolvimento de scripts na linguagem Python (bibliotecas Pandas, Pathlib, Regex) para a etapa de Engenharia de Dados, seguida pela modelagem dimensional e visualização de dados na ferramenta Microsoft Power BI.

## 2- CARACTERIZAÇÃO DA ORGANIZAÇÃO (FONTES DE DADOS)

Para fins deste relatório técnico, a "organização" estudada compreende o **ecossistema de dados de sinistros de trânsito** geridos pelas principais agências reguladoras e departamentos estaduais no Estado de São Paulo. A caracterização das fontes é crucial para entender a necessidade de aplicação dos *scripts* em Python.

### *2.1. Agência Reguladora de Serviços Públicos Delegados de Transporte do Estado de São Paulo (ARTESP)*

A ARTESP é a agência responsável por fiscalizar e gerir as rodovias concedidas à iniciativa privada.

- **Natureza dos Dados:** Histórico de acidentes nas rodovias pedagiadas sob concessão.
- **Período Abordado: 2015 a 2025.** Esta base é extensa e fundamental para a análise de tendências de longo prazo.
- **Desafio de ETL:** A unificação dos dados da ARTESP reside principalmente na **heterogeneidade temporal**. Ao longo de uma década, o *schema* (estrutura e nomes de colunas) dos arquivos CSV pode ter sofrido alterações, demandando que o *script* (*unificar\_tabelas\_artesp\_2015\_2025.py*) normalize os nomes das colunas e remova inconsistências de grafia de nomes de cidades e tipos de ocorrência.

### *2.2. Departamento de Estradas de Rodagem (DER)*

O DER administra e fiscaliza as rodovias estaduais não concedidas.

- **Natureza dos Dados:** Registros de acidentes nas rodovias geridas diretamente pelo Estado.
- **Período Abordado: 2023 a 2025.** Esta base é mais recente e complementa a visão da ARTESP, cobrindo outras vias.

- **Desafio de ETL:** Diferentemente da ARTESP, os dados do DER frequentemente vêm em formatos menos estruturados (como arquivos **Excel** com **múltiplas abas**), exigindo que o *script* (*unificar\_tabelas\_DER\_2023\_2025.py*) realize uma **leitura robusta** em todas as abas das planilhas para garantir que nenhum registro seja perdido, antes de aplicar a padronização das colunas.

### 2.3. Outras Fontes e Contexto Regulatório (IPEA / RENAEST)

O projeto também se apoia em fontes secundárias para fundamentação teórica e eventual enriquecimento dos dados:

- **IPEA (Instituto de Pesquisa Econômica Aplicada):** Fornece o **contexto macroeconômico e social**, correlacionando o desenvolvimento econômico com a taxa de mortalidade no trânsito, o que é essencial para a discussão dos resultados.
- **RENAEST (Registro Nacional de Acidentes e Estatísticas de Trânsito):** Representa a base nacional. A integração com o RENAEST serviria para **qualificação e validação** dos dados estaduais, um objetivo específico do projeto que busca adicionar uma camada de referência nacional ao *dataset* local.

A principal contribuição do projeto é superar o desafio da **interoperabilidade** entre as bases 2.1 e 2.2, usando Python para transformá-las em um único *dataset* coeso, que possa ser facilmente interpretado e visualizado no Power BI.

### 3. FUNDAMENTAÇÃO TEÓRICA

A base teórica deste trabalho se sustenta em duas vertentes cruciais: a macroeconomia do trânsito e o impacto sanitário dos sinistros, fornecendo o **contexto para a interpretação dos padrões de dados** que o *dashboard* de Power BI irá exibir.

#### 3.1. Impacto Econômico e Análise Temporal

O estudo "**Taxa de mortes no trânsito está associada ao desenvolvimento econômico**" (IPEA, 2025) fornece a base macroeconômica para a análise temporal de accidentalidade.

O IPEA demonstra que a variação nas taxas de sinistros é altamente sensível às flutuações econômicas. A redução de acidentes observada em períodos como o pós-2015 no Brasil, por exemplo, foi em grande parte consequência da **recessão econômica**, que resultou na diminuição da circulação de veículos (principalmente de carga e de longa distância).

#### Implicações para o Projeto:

- **Composição do Tráfego:** Recessões tendem a diminuir o tráfego de veículos pesados (cargas), que estão frequentemente associados a acidentes de maior gravidade. O *dashboard* precisa ser capaz de correlacionar a **composição da frota** envolvida nos sinistros com os ciclos econômicos.
- **Gestão de Política Pública:** A teoria reforça que a queda de acidentes resultante de uma crise econômica é insustentável. O crescimento de renda, sem o devido investimento em infraestrutura e fiscalização, tende a elevar novamente a demanda por transporte e, consequentemente, os sinistros. Isso justifica a necessidade de uma ferramenta de **análise preditiva e monitoramento contínuo** (Power BI), alimentada por dados de longo prazo (ARTESP 2015-2025).

### 3.2. Impacto na Saúde Pública e Classificação de Severidade

Complementarmente, a literatura acadêmica, como a monografia de **Silva (2017)**, reforça que os acidentes de trânsito não devem ser avaliados apenas pelo número de óbitos, mas pelo espectro completo da **morbidade e incapacidade**.

O autor destaca que, para cada fatalidade, há um número substancial de feridos com sequelas físicas, funcionais e as chamadas "sequelas invisíveis" (psicológicas e sociais), que impõem um **ônus financeiro e logístico severo** ao Sistema Único de Saúde (SUS) e à sociedade.

#### Implicações para o Projeto:

- **Padronização da Severidade:** Para que a análise de saúde pública seja útil, é imperativo que os dados brutos da ARTESP e do DER sejam unificados sob uma **classificação rigorosa e padronizada da severidade das vítimas** (Ilesos, Leves, Graves, Fatais). Essa padronização é uma tarefa crítica de **limpeza de dados (ETL)** executada pelos *scripts* em Python.
- **Suporte Logístico:** A análise de dados de severidade no Power BI permite aos gestores públicos identificar "**hotspots**" de **alta gravidade** e planejar a logística de emergência (posicionamento de viaturas, direcionamento de vítimas para hospitais especializados), contribuindo diretamente para a minimização das consequências à saúde.

## 4. PROPOSTA DE SOLUÇÃO E DESENVOLVIMENTO

A solução desenvolvida para este projeto consiste na implementação de um *pipeline* de dados automatizado (ETL - *Extract, Transform, Load*) baseado em *scripting*. Esta abordagem foi escolhida em detrimento do processamento manual em planilhas para garantir três pilares: **reprodutibilidade** (o processo pode ser repetido com novos dados), **auditabilidade** (o código documenta as transformações) e **escalabilidade** (capacidade de processar milhões de registros).

#### *4.1. Estratégia de Seleção de Dados: A Escolha do RENAEST*

Para garantir a robustez das análises apresentadas no *dashboard*, o projeto adotou critérios rigorosos de Qualidade de Dados (*Data Quality*). Embora as bases estaduais (ARTESP e DER) forneçam a granularidade espacial necessária, optou-se estrategicamente pela incorporação e cruzamento com dados do **RENAEST** (**Registro Nacional de Acidentes e Estatísticas de Trânsito**).

A seleção do RENAEST fundamentou-se no nível de maturidade e "polimento" desta base. Diferentemente de coletas manuais ou bases brutas de boletins de ocorrência (B.O.) — que frequentemente apresentam excesso de ruído, campos nulos e inconsistências de tipagem —, os dados do RENAEST passam por uma validação prévia em nível nacional. Essa característica foi decisiva para:

1. **Redução de Ruído:** Minimização do tempo computacional gasto com a limpeza de erros de digitação básicos (ex: nomes de municípios grafados incorretamente).
2. **Confiabilidade Taxonômica:** Garantia de que a classificação dos tipos de acidentes (ex: "colisão frontal" vs "choque") seguisse o padrão normativo da SENATRAN.
3. **Complementaridade:** Servir como *baseline* (linha de base) para validar as informações extraídas via Python das planilhas heterogêneas da ARTESP e do DER.

#### *4.2. Arquitetura da Solução (Pipeline ETL)*

O processo de ETL foi integralmente desenvolvido na linguagem **Python 3.x**, operando em ambiente local. A arquitetura modular permitiu tratar cada fonte de dados com regras de negócio específicas antes da unificação. As principais bibliotecas do ecossistema *Scientific Python* empregadas foram:

- **Pandas (`import pandas as pd`):** O "motor" da solução. Utilizado para a manipulação de *DataFrames*, permitindo operações vetorizadas de alta performance para ler, filtrar e transformar grandes volumes de dados sem a latência de softwares de planilha tradicionais.

- **Pathlib (from pathlib import Path):** Implementada para garantir a navegação agnóstica no sistema de arquivos. O uso da Pathlib facilitou a varredura recursiva de diretórios, permitindo que o script localizasse arquivos de dados independentemente de estarem em subpastas ou da estrutura do Sistema Operacional (Windows/Linux).
- **Unicodedata e Re (Regex):** Essenciais para a higienização de *strings*. Foram criadas funções de normalização (normalize('NFKD', s)) para remover acentos e caracteres especiais, garantindo a integridade referencial necessária para os relacionamentos no Power BI.

#### *4.3. Detalhamento dos Scripts e Algoritmos*

O desenvolvimento foi segmentado em dois scripts principais, adaptados à natureza dos dados de entrada:

**A. Unificação ARTESP (Script: unificar\_tabelas\_artesp\_2015\_2025.py)** O desafio principal desta base foi a *variação temporal do schema* (estrutura das tabelas) ao longo de uma década.

- **Algoritmo:** O script executa uma varredura recursiva (.rglob) para identificar todos os arquivos .csv.
- **Normalização:** Foi implementada uma função de padronização de cabeçalhos que converte nomes de colunas para minúsculas e remove espaços. Isso resolveu conflitos onde, por exemplo, uma coluna chamada "Data Ocorrência" em 2015 mudou para "data\_ocorrecia" em 2020.
- **Tratamento de Tipos:** Conversão forçada de colunas de data para o padrão ISO 8601 (AAAA-MM-DD), eliminando ambiguidades de formato (DD/MM vs MM/DD).

**B. Unificação DER (Script: unificar\_tabelas\_DER\_2023\_2025.py)** A complexidade desta fonte residia no formato de armazenamento: arquivos Excel (.xlsx) contendo múltiplas abas de trabalho.

- **Leitura Robusta:** O script foi programado para iterar sobre todas as abas (sheet\_name=None) de cada arquivo, consolidando-as verticalmente em um único *DataFrame*.
- **Rastreabilidade:** Adicionou-se automaticamente uma coluna arquivo\_origem em cada registro. Isso criou uma trilha de auditoria, permitindo que, ao identificar um erro no Dashboard final, o analista possa rastrear exatamente de qual planilha original aquele dado provém.

#### 4.4. Integração com Business Intelligence (Power BI)

A etapa final do *pipeline* consistiu na exportação dos dados tratados para formatos otimizados (.csv com codificação UTF-8). Esta estratégia de "pré-processamento" em Python gerou benefícios diretos no **Power BI**:

1. **Otimização de Performance:** Ao carregar dados já limpos, eliminou-se a necessidade de etapas complexas e lentas no *Power Query* (ETL nativo do Power BI), reduzindo drasticamente o tempo de atualização do relatório.
2. **Georreferenciamento Preciso:** A limpeza textual realizada pelo Python (remoção de acentos e padronização de nomes como "SAO PAULO" / "SÃO PAULO") garantiu 100% de reconhecimento dos municípios pelos mapas do Power BI, permitindo a criação de *Heatmaps* (mapas de calor) precisos para identificação de zonas críticas.
3. **Modelagem de Dados:** A base unificada permitiu a criação de um modelo de dados simples, facilitando o cálculo de medidas DAX (Ex: Total de Óbitos, Média Móvel de Acidentes) e a implementação de filtros dinâmicos por ano, rodovia e tipo de veículo.

## 5. CONSIDERAÇÕES FINAIS

O presente trabalho demonstrou que a aplicação de técnicas de programação (Python) é fundamental para a logística e gestão de trânsito moderna. A capacidade de unificar fontes de dados heterogêneas (ARTESP, DER e RENAEST) transformou dados brutos em informação estratégica.

Conclui-se que a escolha pelo uso de dados já tratados do RENAEST, combinada com a automação da limpeza dos dados estaduais, resultou em um dashboard de alta confiabilidade. A ferramenta desenvolvida não apenas facilita a visualização dos dados históricos, mas serve como base para previsões futuras e planejamento de políticas de segurança viária, visando a redução dos impactos na saúde pública e economia.

Para trabalhos futuros, sugere-se a implementação de algoritmos de Machine Learning para previsão de severidade de acidentes com base nas variáveis higienizadas neste projeto.

## REFERÊNCIAS

IPEA. Taxa de mortes no trânsito está associada ao desenvolvimento econômico. Portal Ipea, 2025. Disponível em: <https://www.ipea.gov.br>. Acesso em: 23 set. 2025.

SILVA, Wanderley Rodrigues da. Os Acidentes de Trânsito e os Impactos na Saúde Pública. 2017. Monografia (Pós-Graduação Lato Sensu em Segurança Viária Urbana) – Fundação Universidade Federal do Tocantins, Araguaína, 2017.

PANDAS DEVELOPMENT TEAM. pandas: powerful Python data analysis toolkit. Versão 2.0. Disponível em: <https://pandas.pydata.org/>. Acesso em: 2025.

MINISTÉRIO DA INFRAESTRUTURA. RENAEST - Registro Nacional de Acidentes e Estatísticas de Trânsito. Disponível em: <https://www.gov.br/infraestrutura/pt-br>. Acesso em: 2025.

