l

**TRƯỜNG CAO ĐẲNG FPT POLYTECHNIC**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**DỰ ÁN TỐT NGHIỆP**

**CHUYÊN NGÀNH: XỬ LÝ DỮ LIỆU**

**ĐỀ TÀI: PHÂN TÍCH DỰ ĐOÁN TÌNH HÌNH TAI NẠN GIAO THÔNG**

**GV HƯỚNG DẪN :** VĂN CÔNG KHANH

**LỚP :** PRO2291.01

**NHÓM :** PRO 2291

**THÀNH VIÊN NHÓM :**

1. Nguyễn Phạm Đăng Khoa – PS39283 (Nhóm trưởng)
2. Lê Hoàng Bảo – PS38302
3. Phạm Minh Hải – PS33057
4. Nguyễn Phúc Hậu – PS42890
5. Nguyễn Đăng Khoa – PS39225
6. Cao Đặng Nhật Anh – PS40602

TP.HCM 12 – 2025

MỤC LỤC

[Chương 1: Giới thiệu dự án 4](#_Toc216086368)

[1.1 Giới thiệu 4](#_Toc216086369)

[1.1.1 Hiện trạng 4](#_Toc216086370)

[1.1.2 Thông tin bộ dữ liệu: 4](#_Toc216086371)

[1.2 Mục tiêu và phạm vi (Kế hoạch ban đầu) 5](#_Toc216086372)

[1.2.1 Mục tiêu tổng thể 5](#_Toc216086373)

[1.2.2 Mục tiêu cụ thể 6](#_Toc216086374)

[1.2.3 Phạm vi dự án: 7](#_Toc216086375)

[1.2.4 Đối tượng phục vụ (Stakeholders) 8](#_Toc216086376)

[1.2.5 Tiêu chí thành công 8](#_Toc216086377)

[1.2.6 Lý do chưa hoàn thành Giai đoạn 2 8](#_Toc216086378)

[1.3 Sản phẩm mục tiêu 9](#_Toc216086379)

[1.4 Cấu trúc báo cáo 10](#_Toc216086380)

[1.5 Lập kế hoạch dự án 11](#_Toc216086381)

[Chương 2: Phân tích yêu cầu 13](#_Toc216086382)

[2.1 Phân tích yêu cầu 13](#_Toc216086383)

[2.2 Câu chuyện dữ liệu 15](#_Toc216086384)

[2.2.1 Đặt vấn đề 15](#_Toc216086385)

[2.3 Kiến trúc hệ thống 16](#_Toc216086386)

[2.3.1 Giải thích 16](#_Toc216086387)

[2.4 Giải thích về bộ dữ liệu 18](#_Toc216086388)

[2.4.1 Các khái niệm 18](#_Toc216086389)

[2.4.2 Các trường dữ liệu 18](#_Toc216086390)

[Chương 3: Làm sạch và chuyển đổi dữ liệu 22](#_Toc216086391)

[3.1 Cài đặt cấu hình (Module: config.py) 22](#_Toc216086392)

[3.2 Phân tích dữ liệu (Module: eda.py) 24](#_Toc216086393)

[3.3 Làm sạch dữ liệu & xây dựng đặc trưng (Module: cleaner.py) 29](#_Toc216086394)

[3.4 Xây dựng Mô hình Dữ liệu (Star Schema) (Module: Splitter.py) 40](#_Toc216086395)

[3.5 Tạo bảng tổng hợp (Module: Aggregate.py) 47](#_Toc216086396)

[Chương 4: Xây dựng và phát triển sản phẩm 54](#_Toc216086397)

[4.1 Aggregate Tables Thiết kế 54](#_Toc216086398)

[4.1.1 Nguyên tắc thiết kế Aggregates 54](#_Toc216086399)

[4.1.2 Agg\_federal 54](#_Toc216086400)

[4.1.3 Agg\_state\_anomaly 54](#_Toc216086401)

[4.1.4 Agg\_infrastructure 55](#_Toc216086402)

[4.1.5 Agg\_weather\_deep\_dive 55](#_Toc216086403)

[4.2 Dashboard Thiết kế 56](#_Toc216086404)

[4.2.1 Dashboard 1: National Overview 56](#_Toc216086405)

[4.2.2 Dashboard 2: State Anomaly Detection 58](#_Toc216086406)

[4.2.3 Dashboard 3: Infrastructure Statistics 59](#_Toc216086407)

[4.2.4 Dashboard 4: Weather Deep Dive 61](#_Toc216086408)

[4.2.5 Thiết kế Principles 62](#_Toc216086409)

[Chương 5: Kết quả và đánh giá 62](#_Toc216086410)

[5.1. Kết quả xử lý dữ liệu 62](#_Toc216086411)

[5.1.1. Data Pipeline Output 62](#_Toc216086412)

[5.1.2. Star Schema Output 63](#_Toc216086413)

[5.1.3. Aggregate Output 63](#_Toc216086414)

[5.2 So sánh Kế hoạch ban đầu vs Thực tế 63](#_Toc216086415)

[5.3. Đánh giá hiệu quả 65](#_Toc216086416)

[5.3.1. So sánh trước/sau xử lý 65](#_Toc216086417)

[5.3.2. So với mục tiêu đề ra 65](#_Toc216086418)

[5.3.3. Performance 66](#_Toc216086419)

[5.4. Recommendations 66](#_Toc216086420)

[CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 67](#_Toc216086421)

[6.1. Tổng kết nội dung thực hiện 67](#_Toc216086422)

[6.2. Đánh giá tổng quan 67](#_Toc216086423)

[6.3. Hướng phát triển 68](#_Toc216086424)

[6.3.1. Ngắn hạn (1-3 tháng) 68](#_Toc216086425)

[6.3.2. Trung hạn (3-6 tháng) 69](#_Toc216086426)

[6.5. Lời kết 70](#_Toc216086427)

[Tài liệu tham khảo 71](#_Toc216086428)

[Phụ lục 71](#_Toc216086429)

# Chương 1: Giới thiệu dự án

## Giới thiệu

### Hiện trạng

* Tai nạn giao thông là một trong những nguyên nhân hàng đầu gây tử vong tại Hoa Kỳ với hàng loạt ca tử vong mỗi năm. Việc hiểu rõ các yếu tố ảnh hưởng đến tai nạn giao thông như thời gian, địa điểm, thời tiết và cơ sở hạ tầng là then chốt để các cơ quan quản lý giao thông đưa ra các biện pháp can thiệp hiệu quả.
* Hiện nay, dữ liệu về tai nạn giao thông được thu thập từ nhiều nguồn khác nhau nhưng chưa được tổng hợp và phân tích một cách có hệ thống. Các báo cáo thường mang tính mô tả, thiếu đi khả năng đào sâu vào các mẫu hình (patterns) ẩn sau dữ liệu. Điều này dẫn đến việc các quyết định chính sách thiếu cơ sở dữ liệu vững chắc.
* Nhu cầu cấp thiết hiện tại là xây dựng một hệ thống phân tích dữ liệu toàn diện, cho phép:
  + Xác định các điểm nóng (thành phố) về tai nạn (hotspots) theo địa lý
  + Phân tích xu hướng theo thời gian
  + Đánh giá tác động của thời tiết và cơ sở hạ tầng
  + Cung cấp thông tin để phân bổ nguồn lực và hoạch định chính sách

### Thông tin bộ dữ liệu:

* **Nguồn dữ liệu:** US Accidents Dataset từ Kaggle
* **Tác giả**: Moosavi, Sobhan et al. (2019)
* **Đường dẫn**: <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>

A qr code on a white background

AI-generated content may be incorrect.

#### Đặc điểm tổng quan

|  |  |
| --- | --- |
| **Tiêu chí** | **Giá trị** |
| Quy mô | 7.7 triệu bản ghi tai nạn giao thông |
| Phạm vi thời gian | Từ tháng 2/2016 đến tháng 3/2023 |
| Phạm vi địa lý | 49 bang tại Hoa Kỳ |
| Nguồn thu thập | API giao thông thời gian thực, camera giao thông, báo cáo từ sở giao thông |

#### Cấu trúc dữ liệu

* **47 trường dữ liệu** bao gồm:
* Thông tin định danh: ID, Source
* Thông tin thời gian: Start\_Time, End\_Time
* Thông tin vị trí: Tọa độ (Lat/Lng), State, City, County, Street, Zipcode
* Thông tin mức độ: Severity (thang đo 1-4)
* Thông tin thời tiết: Temperature, Visibility, Precipitation, Weather\_Condition
* Thông tin cơ sở hạ tầng: Traffic\_Signal, Junction, Crossing, Railway, Station, Stop, Amenity

## Mục tiêu và phạm vi (Kế hoạch ban đầu)

### Mục tiêu tổng thể

Dự án "Phân tích và Dự đoán Tình hình Tai nạn Giao thông Hoa Kỳ" được thực hiện nhằm giải quyết các vấn đề thực tiễn sau:

**Vấn đề 1: Thiếu cái nhìn tổng quan về tình hình tai nạn**

Các cơ quan quản lý giao thông liên bang cần hiểu rõ xu hướng tai nạn theo thời gian để đánh giá hiệu quả chính sách và phân bổ ngân sách. Tuy nhiên, dữ liệu hiện tại nằm rải rác, chưa được tổng hợp thành các chỉ số dễ theo dõi như tổng số tai nạn theo năm, tỷ lệ thay đổi YoY, hay tỷ lệ tai nạn nghiêm trọng.

**Vấn đề 2: Khó xác định các bang/khu vực có mức độ nghiêm trọng bất thường**

Chính quyền các bang cần biết liệu địa phương mình có đang gặp vấn đề về an toàn giao thông so với mặt bằng chung quốc gia hay không. Việc so sánh đơn thuần số lượng tai nạn không phản ánh đúng mức độ nghiêm trọng, cần có các chỉ số chuẩn hóa như z-score để phát hiện anomaly.

**Vấn đề 3: Chưa hiểu rõ mối quan hệ giữa cơ sở hạ tầng và tai nạn**

Lực lượng cảnh sát và cơ quan quản lý giao thông cần xác định các điểm nóng về hạ tầng (ngã tư, đèn tín hiệu, đường ray) để triển khai tuần tra và cải thiện. Dữ liệu hiện tại chưa được phân tích để đưa ra risk score cho từng loại hạ tầng.

**Vấn đề 4: Thiếu phân tích tác động của thời tiết**

Các công ty bảo hiểm và người tham gia giao thông cần hiểu rõ thời tiết ảnh hưởng như thế nào đến mức độ nghiêm trọng của tai nạn. Thông tin này giúp định giá bảo hiểm chính xác hơn và cảnh báo người dân trong điều kiện thời tiết xấu.

**Vấn đề 5: Chưa có khả năng dự đoán (Kế hoạch ban đầu)**

Ngoài phân tích quá khứ, việc dự đoán mức độ nghiêm trọng của tai nạn dựa trên các yếu tố đầu vào (thời tiết, địa điểm, thời gian) sẽ giúp các cơ quan chức năng chủ động phòng ngừa thay vì chỉ phản ứng sau khi tai nạn xảy ra.

### Mục tiêu cụ thể

Để giải quyết các vấn đề trên, dự án đặt ra hai nhóm mục tiêu:

**Nhóm mục tiêu 1 - Phân tích dữ liệu (Analytics):**

|  |  |  |
| --- | --- | --- |
| **Vấn đề cần giải quyết** | **Mục tiêu** | **Giải pháp** |
| Thiếu cái nhìn tổng quan | Cung cấp dashboard National Overview | Aggregate theo Year, tính YoY change |
| Khó xác định anomaly | Phát hiện bang có severity bất thường | Tính z-score so với national average |
| Chưa hiểu rõ tác động hạ tầng | Đánh giá risk theo loại hạ tầng | Tính Infrastructure Risk Score |
| Thiếu phân tích thời tiết | Phân tích impact của từng loại thời tiết | So sánh với baseline (Clear weather) |

**Nhóm mục tiêu 2 - Dự đoán (Predictive Modeling) - Kế hoạch ban đầu:**

|  |  |  |
| --- | --- | --- |
| **Vấn đề cần giải quyết** | **Mục tiêu** | **Giải pháp** |
| Chưa dự đoán được severity | Classification model | Random Forest/XGBoost với features: weather, time, location, infrastructure |
| Chưa dự báo được xu hướng | Time series forecasting | Prophet/ARIMA dự báo số tai nạn theo tháng |

### Phạm vi dự án:

**Phạm vi dữ liệu:**

|  |  |
| --- | --- |
| **Tiêu chí** | **Phạm vi** |
| Thời gian phân tích | 2019-2022 (4 năm có đầy đủ dữ liệu) |
| Số lượng bản ghi | ~ 4.8 triệu bản ghi sau khi lọc |
| Phạm vi địa lý | 49 bang tại Hoa Kỳ |

**Phạm vi kỹ thuật theo kế hoạch ban đầu:**

|  |  |  |
| --- | --- | --- |
| **Giai đoạn** | **Nội dung** | **Trạng thái** |
| Giai đoạn 1: Data Engineering & Analytics | ETL pipeline, Star Schema, Aggregate tables, Dashboard specs | Hoàn thành |
| Giai đoạn 2: Predictive Modeling | Classification model, Time series forecasting | Chưa thực hiện |

**Giai đoạn 2 - Predictive Modeling (Kế hoạch ban đầu - chưa thực hiện):**

* Classification Model: Dự đoán Severity (1-4) dựa trên các features về thời tiết, địa điểm, thời gian, cơ sở hạ tầng
* Time Series Model: Dự báo số lượng tai nạn theo tháng/quý
* Feature Engineering cho ML: Tạo các features phái sinh phù hợp cho mô hình
* Model Evaluation: Đánh giá và so sánh hiệu năng các mô hình

### Đối tượng phục vụ (Stakeholders)

Dự án hướng đến phục vụ 4 nhóm đối tượng chính:

|  |  |  |
| --- | --- | --- |
| **Stakeholder** | **Nhu cầu thông tin** | **Dashboard/Output tương ứng** |
| Chính phủ Liên bang | Xu hướng tổng quan, so sánh YoY, phân bổ ngân sách | National Overview |
| Chính quyền Bang/Địa phương | Xác định anomaly, so sánh với mức trung bình quốc gia | State Anomaly Detection |
| Lực lượng Cảnh sát | Điểm nóng hạ tầng, triển khai tuần tra | Infrastructure Statistics |
| Người trực tiếp tham gia giao thông | Tác động thời tiết, định giá rủi ro | Weather Deep Dive |

### Tiêu chí thành công

**Giai đoạn 1 (Đã hoàn thành):**

|  |  |
| --- | --- |
| **Tiêu chí** | **Mục tiêu** |
| Data Quality | <5% missing values sau xử lý |
| Pipeline Performance | Xử lý 4.8M records trong <15 phút |
| Schema Integrity | 100% FK relationships hợp lệ |
| Phạm vi phân tích cho các Stakeholder | 4 dashboards cho 4 nhóm đối tượng |

**Giai đoạn 2 (Chưa thực hiện):**

|  |  |
| --- | --- |
| **Tiêu chí** | **Mục tiêu dự kiến** |
| Classification Accuracy | ≥70% |
| Forecasting MAPE | ≤15% |

### Lý do chưa hoàn thành Giai đoạn 2

Giai đoạn Predictive Modeling chưa được thực hiện do:

1. **Hạn chế về thời gian:** Giai đoạn 1 đã chiếm phần lớn thời lượng để đảm bảo chất lượng pipeline và data model.
2. **Ưu tiên về foundation:** Nhóm quyết định tập trung hoàn thiện nền tảng dữ liệu vững chắc trước khi phát triển ML models.
3. **Độ phức tạp kỹ thuật:** Việc xây dựng ML models yêu cầu kiến thức chuyên sâu mà nhóm cần thêm thời gian để nghiên cứu.

Giai đoạn 2 được đưa vào **Hướng phát triển** tại Chương 6.

## Sản phẩm mục tiêu

1. **Data Pipeline (Python Scripts)**

|  |  |  |
| --- | --- | --- |
| **Module** | **Chức năng** | **Output** |
| config.py | Cấu hình tham số, đường dẫn | Configuration settings |
| loader.py | Load và filter dữ liệu thô | Filtered DataFrame |
| eda.py | Tổng quan khám phá và phân tích toàn bộ dữ liệu | Phần 3.2 Phân tích dữ liệu đã được làm rõ |
| cleaner.py | Làm sạch và feature engineering | accidents\_cleaned.csv |
| splitter.py | Tạo Star Schema | 3 dims + 1 fact table |
| aggregate.py | Tạo bảng tổng hợp | 4 aggregate tables |

1. **Star Schema (Data Model)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Type** | **Records** | **Primary Key** |
| dim\_time | Dimension | ~1,461 | Date |
| dim\_location | Dimension | ~1.2M | Location\_id |
| dim\_weather | Dimension | ~120 | Weather\_id |
| accident\_detail | Fact | ~4.8M | ID |

1. **Aggregate Tables**

|  |  |  |  |
| --- | --- | --- | --- |
| **Aggregate** | **Stakeholder** | **Level** | **Key Metrics** |
| agg\_federal | Chính phủ liên bang | Year | YoY change, total accidents |
| agg\_state\_anomaly | Chính phủ tại 1 bang cụ thể | State & Year | Anomaly detection (phát hiện điểm bất thường) |
| agg\_infrastructure | Cơ quan an ninh và Cơ sở hạ tầng (CISA) | State & Urban/Rural | Infrastructure Analysis |
| agg\_weather\_deep | Người tham gia giao thông | Weather & State | Weather Impact |

## Cấu trúc báo cáo

Báo cáo được tổ chức thành 6 chương với nội dung như sau:

**Chương 1: Giới thiệu đề tài**

* Trình bày lý do chọn đề tài, mục tiêu và phạm vi dự án
* Mô tả sản phẩm mục tiêu và cấu trúc báo cáo

**Chương 2: Phân tích yêu cầu**

* Phân tích yêu cầu từ các nhóm stakeholder
* Xây dựng câu chuyện dữ liệu (Data Story)
* Thiết kế kiến trúc hệ thống
* Mô tả chi tiết bộ dữ liệu

**Chương 3: Làm sạch và chuyển đổi dữ liệu**

* Module cấu hình (config.py)
* Phân tích dữ liệu khám phá (eda.py)
* Làm sạch và xây dựng đặc trưng (cleaner.py)
* Xây dựng mô hình dữ liệu Star Schema (splitter.py)
* Tạo bảng tổng hợp (aggregate.py)

**Chương 4: Xây dựng và phát triển sản phẩm**

* Thiết kế Aggregate Tables
* Thiết kế Dashboard chi tiết

**Chương 5: Kết quả và đánh giá**

* Kết quả xử lý dữ liệu
* So sánh kế hoạch ban đầu vs thực tế
* Đánh giá hiệu quả
* Recommendations

**Chương 6: Kết luận và hướng phát triển**

* Tổng kết nội dung thực hiện
* Đánh giá tổng quan
* Hướng phát triển (bao gồm ML roadmap)
* Lời kết

**Phụ lục**

* Cấu trúc thư mục dự án
* Data Dictionary
* Tableau Calculated Fields
* Link sản phẩm

## Lập kế hoạch dự án

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TT** | **Hạng mục** | **Bắt đầu** | **Kết thúc** | **Kết quả** |
| **1** | **Giới thiệu dự án** | **15/09/2024** | **21/09/2024** | **Hoàn thành** |
| 1.1 | Giới thiệu và khảo sát dataset | 15/09/2024 | 17/09/2024 | Xác định scope: 2019-2022, 3M records |
| 1.2 | Phân tích yêu cầu dự án | 18/09/2024 | 19/09/2024 | Xác định 10 output files, star schema |
| 1.3 | Lập kế hoạch và phân công | 20/09/2024 | 21/09/2024 | Timeline 10 tuần, phân chia 6 thành viên |
| **2** | **Phân tích yêu cầu** | **22/09/2024** | **05/10/2024** | **Hoàn thành** |
| 2.1 | Phân tích yêu cầu Stakeholders | 22/09/2024 | 25/09/2024 | Xác định 4 stakeholder groups |
| 2.2 | Xây dựng câu chuyện dữ liệu | 26/09/2024 | 29/09/2024 | 5 analytical narratives |
| 2.3 | Thiết kế kiến trúc hệ thống | 30/09/2024 | 02/10/2024 | 5-layer pipeline architecture |
| 2.4 | Phân tích bộ dữ liệu | 03/10/2024 | 05/10/2024 | Data dictionary 46 fields |
| **3** | **Làm sạch và chuyển đổi dữ liệu** | **06/10/2024** | **26/10/2024** | **Hoàn thành** |
| 3.1 | Chuẩn bị dữ liệu | 06/10/2024 | 09/10/2024 | Setup project structure, load raw data |
| 3.2 | Exploratory Data Analysis (EDA) | 10/10/2024 | 13/10/2024 | Phát hiện 12% missing, outliers |
| 3.3 | Làm sạch dữ liệu | 14/10/2024 | 19/10/2024 | Handle nulls, cap outliers, remove invalid |
| 3.4 | Chuyển đổi và tạo features | 20/10/2024 | 26/10/2024 | 14 features mới (time, weather, infra) |
| **4** | **Xử lý dữ liệu** | **27/10/2024** | **16/11/2024** | **Hoàn thành** |
| 4.1 | Chuẩn hóa dữ liệu | 27/10/2024 | 31/10/2024 | Hoàn thành |
| 4.2 | Mô hình hóa dữ liệu (Star Schema) | 01/11/2024 | 08/11/2024 | 3 dims + 1 fact tables |
| 4.3 | Validation và testing | 09/11/2024 | 12/11/2024 | Hoàn thành |
| 4.4 | Tạo bảng aggregate | 13/11/2024 | 16/11/2024 | 4 aggregate tables cho Tableau |
| **5** | **Trực quan hóa dữ liệu** | **17/11/2024** | **30/11/2024** | **Hoàn thành** |
| 5.1 | Thiết kế dashboard mockups | 20/11/2024 | 23/11/2024 | 3-4 dashboard layouts |
| 5.2 | Xây dựng visuals trong Tableau | 24/11/2024 | 28/11/2024 | Column - bar charts, pie – donut, maps, … |
| 5.3 | Testing và optimization | 29/11/2024 | 30/11/2024 | Performance tuning |
| **6** | **Xây dựng báo cáo** | **15/09/2024** | **20/12/2024** | **Hoàn thành** |
| 6.1 | Thiết kế dashboard tổng hợp | 01/12/2024 | 06/12/2024 | 3-4 interactive dashboards |
| 6.2 | Tạo detailed reports | 07/12/2024 | 12/12/2024 | State, city |
| 6.3 | Viết documentation | 13/12/2024 | 17/12/2024 | Technical + user documentation |
| 6.4 | Review và finalization | 18/12/2024 | 20/12/2024 | Final deliverables readyz |
| 7 | **Hoàn thiện và bàn giao** | **21/12/2024** | **27/12/2024** | **Hoàn thành** |
| 7.1 | Tổng hợp kết quả | 21/12/2024 | 23/12/2024 | Hoàn thành |
| 7.2 | Chuẩn bị presentation | 24/12/2024 | 25/12/2024 | Demo với thầy |
| 7.3 | Bàn giao và đóng dự án | 27/12/2024 | 27/12/2024 | Nộp document |

# Chương 2: Phân tích yêu cầu

## Phân tích yêu cầu

1. **Dữ liệu:**
   1. **Đặc điểm dữ liệu hiện có:**

* **Quy mô:** 7.7 triệu records, ~3.5GB (raw CSV)
* **Cấu trúc:** 46 columns, bao gồm thông tin thời gian, địa lý, thời tiết, cơ sở hạ tầng
* **Định dạng:** CSV file, dữ liệu semi-structured
* **Phạm vi thời gian**: 2016-2023
  1. **Yêu cầu về dữ liệu:**

1. **Lọc dữ liệu theo thời gian:** Chỉ lấy 2019-2022
2. **Làm sạch dữ liệu:** Xử lý missing, outliers, invalid values
3. **Chuẩn hóa**: Đưa về star schema để query nhanh
4. **Pre-aggregate:** Tạo bảng tổng hợp sẵn cho dashboard
5. **Quản lý và lưu trữ:**
   1. **Yêu cầu lưu trữ:**

* Dữ liệu raw không được thay đổi (immutable)
* Dữ liệu processed phải có version control
* Cấu trúc thư mục rõ ràng: raw → cleaned → dimensions/fact → aggregates
  1. **Yêu cầu quản lý**:
* Code phải modular, có thể chạy lại với dữ liệu mới
* Documentation đầy đủ cho mỗi bước xử lý
* Git version control cho code và documents
* Backup strategy cho dữ liệu quan trọng

1. **Công nghệ:**
   1. **Yêu cầu công nghệ:**
2. **Xử lý dữ liệu:** Cần tool xử lý 3M+ records hiệu quả
3. **Visualization:** Dashboard interactive, real-time filtering
4. **Performance:** Query response time < 3 giây
5. **Accessibility:** Không yêu cầu database server phức tạp
6. **Cost:** Ưu tiên open - source hoặc low-cost solutions
7. **Quyết định công nghệ:**

* Dữ liệu - Chọn: **Python (pandas)**
* **Lý do chọn:**
  + **Xử lý hiệu quả:** pandas với 3M records trong memory
  + **Ecosystem phong phú:** Thư viện xử lý data sẵn có
  + **Miễn phí:** Open - source, không có licensing cost
  + **Cộng đồng lớn:** Dễ tìm giải pháp khi gặp vấn đề
* **Đáp ứng yêu cầu:**
  + **Yêu cầu 1 (Lọc 2019-2022)**
  + **Yêu cầu 2 (Làm sạch)**
  + **Yêu cầu 3 (Feature engineering)**
  + **Yêu cầu 4 (Star schema):** merge(), groupby(), splitter
  + **Yêu cầu 5 (Aggregates):** groupby().agg() với custom functions

1. **Quản lý và lưu trữ - Chọn: File-based (CSV) + Git**

* **Lý do chọn:**
  + **Version control:** Git quản lý và kiểm soát trong code và documentation
  + **No infrastructure:** Không cần setup database server
  + **Portability:** Dễ share, backup
* **Cấu trúc lưu trữ**:

A screenshot of a computer program

AI-generated content may be incorrect.

1. **Công nghệ visualization - Chọn: Tableau Desktop**

* **Lý do chọn:**
  + **Interactive dashboards:** Drag-and-drop
  + **Dễ tiếp cận:** Non-technical users có thể sử dụng
* **Đáp ứng yêu cầu:**
  + **Yêu cầu interactive:** Filters, drill-down, tooltips native support
  + **Yêu cầu performance:** Connect trực tiếp CSV, in-memory caching
  + **Yêu cầu <3s query**: Pre-aggregates + star schema optimize query
  + **Yêu cầu accessibility:** Desktop version, không cần server

## Câu chuyện dữ liệu

### Đặt vấn đề

**Mô tả thực trạng:**

* Hiện nay, tai nạn giao thông tại Hoa Kỳ gây ra hang loạt ca tử vong và 4.4 triệu ca bị thương mỗi năm, gây ra thiệt hại kinh tế. Các cơ quan quản lý giao thông bang và liên bang đang đối mặt với thách thức:

**Thiếu thông tin chi tiết**: Báo cáo hiện tại chỉ tổng hợp ở mức quốc gia/bang, không đủ chi tiết để can thiệp cụ thể

* + **Không thấy được patterns**: Xu hướng theo thời gian (giờ, ngày, mùa) chưa được phân tích có hệ thống
  + **Chưa định lượng tác động**: Vai trò của thời tiết và cơ sở hạ tầng chưa được thống kê định lượng rõ ràng
  + **Phản ứng chậm**: Không có dashboard real-time để quan sát và phản ứng nhanh với hotspots
* Với bộ dữ liệu 7.7 triệu records từ nhiều nguồn, chúng ta có cơ hội phân tích sâu và cung cấp insight cho hành động cụ thể.

## Kiến trúc hệ thống

### Giải thích

* **Tại sao thiết kế 6 layers?**

**Layer 1 - Data Source**:

* Single source of truth: Chỉ có 1 raw file, không merge từ nhiều nguồn
* Immutable: Raw data không bao giờ được modify

**Layer 2 - Ingestion**:

* Early filtering: Giảm 7.7M đến 3M ngay từ đầu (tiết kiệm memory)
* Dtype optimization: Downcast int64 đến int8, float64 đến float32 (giảm 50% memory)

**Layer 3 - Exploration**:

* Understand before transform: Không clean mù quáng, phải hiểu distribution trước
* Document quality issues: EDA report làm evidence cho cleaning decisions

**Layer 4 - Transformation**:

* Separate concerns: Cleaning khác Modeling, tách thành 2 modules
* Feature engineering: Tạo derived columns để enrich analysis

**Layer 5 - Modeling**:

* Star schema for queries: Optimize read performance cho Tableau
* Denormalization trade-off: Chấp nhận redundancy để giảm JOINs

**Layer 6 - Aggregation**:

* Pre-compute patterns: Tính trước metrics phổ biến
* Dashboard performance: Aggregate tables, response time dưới 1s
* **Thiết kế Decisions**

**Decision 1**: Tại sao dùng CSV thay vì Database?

Reasoning:

* Simplicity: Không cần setup PostgreSQL hoặc MySQL
* Portability: Dễ share, backup
* Tableau compatibility: Native CSV connector

**Decision 2:** Tại sao cần cả star schema VÀ aggregates?

Reasoning:

* Star schema: Flexible, cho ad-hoc queries (analysts explore)
* Aggregates: Fast, cho dashboards (executives monitor)
* Use case:
  + Dashboard connects to aggregates (real-time KPIs)
  + Detailed reports connect to star schema (drill-down)

## Giải thích về bộ dữ liệu

### Các khái niệm

|  |  |  |
| --- | --- | --- |
| **Tên khái niệm** | **Định nghĩa** | **Chi tiết** |
| Severity (Mức độ nghiêm trọng) | Thang đo 1-4 về tác động của tai nạn đến traffic flow (Lưu ý: Không phải thương vong người) | Severity 1: Minor (<15p, không chặn làn); Severity 2: Moderate (15-60p, chặn 1 phần); Severity 3: Serious (1-4h, chặn nhiều làn); Severity 4: Severe (>4h, đóng đường) |
| Duration (Thời lượng) | Khoảng thời gian từ khi tai nạn được báo cáo đến khi traffic trở về bình thường | Tương quan dương với Severity; Dữ liệu > 24h (1440 phút) thường được coi là lỗi dữ liệu (data error) |
| Infrastructure Features (Đặc điểm hạ tầng) | Tập hợp 7 biến Boolean (True/False) chỉ sự hiện diện của hạ tầng giao thông gần vị trí tai nạn | 7 Flags: Traffic\_Signal, Junction (Giao lộ), Crossing (Vạch kẻ), Railway, Station (Trạm xe), Stop (Biển báo), Amenity (Tiện ích) |
| Weather Condition (Điều kiện thời tiết) | Biến phân loại mô tả trạng thái thời tiết tại thời điểm tai nạn (gốc: 150 giá trị) | Nhóm chính: Clear (60%), Rain (25%), Snow (8%), Fog (5%), Other (2%) |

### Các trường dữ liệu

**Bảng Data Dictionary đầy đủ**

**IDENTIFIERS (2 fields)**

|  |  |  |
| --- | --- | --- |
| Tên trường | Kiểu dữ liệu | Mô tả |
| ID | string | Định danh duy nhất cho vụ tai nạn. |
| Source | string | Nguồn API cung cấp dữ liệu. |

**TIME (2 fields)**

|  |  |  |
| --- | --- | --- |
| Tên trường | Kiểu dữ liệu | Mô tả |
| ID | string | Định danh duy nhất cho vụ tai nạn. |
| Source | string | Nguồn API cung cấp dữ liệu. |
| Start\_Time | datetime | Thời điểm bắt đầu xảy ra tai nạn. |
| End\_Time | datetime | Thời điểm giao thông được giải phóng hoàn toàn. |

**LOCATION (9 fields)**

|  |  |  |
| --- | --- | --- |
| Tên trường | Kiểu dữ liệu | Mô tả |
| ID | string | Định danh duy nhất cho vụ tai nạn. |
| Source | string | Nguồn API cung cấp dữ liệu. |
| Start\_Time | datetime | Thời điểm bắt đầu xảy ra tai nạn. |
| End\_Time | datetime | Thời điểm giao thông được giải phóng hoàn toàn. |
| Start\_Lat | float | Tọa độ vĩ độ nơi xảy ra tai nạn. |
| Start\_Lng | float | Tọa độ kinh độ nơi xảy ra tai nạn. |
| End\_Lat | float | Tọa độ vĩ độ cuối cùng của đoạn đường bị ảnh hưởng (nếu có). |
| End\_Lng | float | Tọa độ kinh độ cuối cùng của đoạn đường bị ảnh hưởng (nếu có). |
| Street | string | Tên đường xảy ra tai nạn. |
| City | string | Thành phố xảy ra tai nạn. |
| County | string | Quận/Hạt xảy ra tai nạn. |
| State | string | Tên viết tắt của tiểu bang (ví dụ: OH). |
| Zipcode | string | Mã bưu chính của khu vực. |

**SEVERITY (1 field)**

|  |  |  |
| --- | --- | --- |
| Tên trường | Kiểu dữ liệu | Mô tả |
| Severity | int | Thang đo mức độ nghiêm trọng (1-4) về tác động của tai nạn đối với luồng giao thông. |

**WEATHER (8 fields)**

|  |  |  |
| --- | --- | --- |
| Tên trường | Kiểu dữ liệu | Mô tả |
| Temperature(F) | float | Nhiệt độ môi trường (đơn vị Fahrenheit). |
| Humidity(%) | float | Độ ẩm tương đối (đơn vị phần trăm). |
| Pressure(in) | float | Áp suất không khí (đơn vị inch thủy ngân). |
| Visibility(mi) | float | Khoảng cách nhìn thấy (đơn vị dặm). |
| Wind\_Speed(mph) | float | Tốc độ gió (đơn vị dặm/giờ). |
| Precipitation(in) | float | Lượng mưa (đơn vị inch). |
| Weather\_Condition | string | Mô tả chi tiết về điều kiện thời tiết (ví dụ: Light Rain, Clear). |

**INFRASTRUCTURE (7 fields)**

|  |  |  |
| --- | --- | --- |
| Tên trường | Kiểu dữ liệu | Mô tả |
| Amenity | bool | Có gần tiện ích công cộng (bệnh viện, trường học,...) hay không. |
| Crossing | bool | Có gần vạch qua đường cho người đi bộ hay không. |
| Junction | bool | Có nằm tại giao lộ hoặc ngã tư hay không. |
| Railway | bool | Có gần đường ray xe lửa hay không. |
| Station | bool | Có gần trạm xe bus hoặc ga tàu/tàu điện ngầm hay không. |
| Stop | bool | Có biển báo dừng (Stop Sign) hay không. |
| Traffic\_Signal | bool | Có tín hiệu đèn giao thông hay không. |

**DERIVED FIELDS (Created in cleaning)**

**Time Features (9 fields)**

|  |  |  |
| --- | --- | --- |
| Tên trường | Kiểu dữ liệu | Mục đích |
| Year | int | Phân tích xu hướng theo năm (Trend analysis). |
| Month | int | Phân tích tính thời vụ (Seasonality). |
| Day | int | Phân tích chi tiết theo ngày (Day-level analysis). |
| Date | date | Ngày xảy ra của tai nạn |

**Infrastructure Features (1 field)**

|  |  |  |
| --- | --- | --- |
| Tên trường | Kiểu dữ liệu | Mục đích |
| Infra\_Score | int | Điểm số đánh giá độ phức tạp của cơ sở hạ tầng tại vị trí tai nạn (tổng của các cờ Amenity, Crossing, Junction, Railway, Station, Stop, Traffic\_Signal). |

**Location Identifier (1 field)**

|  |  |  |
| --- | --- | --- |
| Tên trường | Kiểu dữ liệu | Mục đích |
| Location\_ID | string | Khóa ngoại (Foreign Key) được tạo bằng cách ghép các trường State, City, và Street. |

* **Relationships giữa các trường**

Temporal relationships:

* Start\_Time → Derive → Year, Month, Day
* End\_Time - Start\_Time → Duration\_min

Spatial relationships:

* (Street, City) → Unique → Location\_ID

Infrastructure relationships:

* All infrastructure booleans → Aggregate → Infra\_Score

# Chương 3: Làm sạch và chuyển đổi dữ liệu

## 3.1 Cài đặt cấu hình (Module: config.py)

Mục tiêu và vai trò

* **Trung tâm cấu hình:** Tập trung định nghĩa đường dẫn dữ liệu, khoảng thời gian phân tích, ngưỡng phát hiện ngoại lệ và tham số tổng hợp để mọi module khác tham chiếu đồng nhất.
* **Tính linh hoạt:** Cho phép thay đổi cấu hình mà không chỉnh sửa logic xử lý trong cleaner/eda.
* **An toàn đường dẫn:** Sử dụng pathlib để thao tác path một cách nhất quán, tránh lỗi phân tách trên các hệ điều hành.

Nội dung cấu hình chính

* **Đường dẫn dữ liệu:**

**A computer code with text

AI-generated content may be incorrect.**

* **Ý nghĩa:** Chuẩn hóa vị trí file thô và các thư mục kết quả (cleaned/dim/fact/aggregates) phục vụ phân tách dữ liệu theo kiến trúc dữ liệu chuẩn (dim/fact).
* **Tham số phân tích thời gian:**

A number of numbers and symbols

AI-generated content may be incorrect.

* **Ý nghĩa**: Hạn chế phạm vi thời gian được phân tích và báo cáo, đảm bảo tính tập trung và hiệu năng.
* **Ngưỡng ngoại lệ (domain knowledge):**

**A white background with green text

AI-generated content may be incorrect.**

* **Ý nghĩa:** Capping/corralling các biến số để loại bỏ các giá trị không thực tế (ví dụ: nhiệt độ vượt phạm vi hợp lý, thời lượng tai nạn > 24h)
* **Tham số tổng hợp:**

**A math equations on a white background

AI-generated content may be incorrect.**

* **Ý nghĩa:** Chuẩn bị cho các phân tích tập trung (top-K, quy tắc 80/20) ở các bước tổng hợp/BI.

## 3.2 Phân tích dữ liệu (Module: eda.py)

Mục tiêu chuyên sâu

* **Xác định cấu trúc và chất lượng dữ liệu:** Kiểu dữ liệu thời gian, phạm vi biến số, độ đầy đủ.
* **Nhận diện mẫu hình sơ bộ:** Theo thời gian, địa lý, mức độ nghiêm trọng để định hướng bước làm sạch và tạo đặc trưng tiếp theo.
* **Phát hiện rủi ro:** Dữ liệu thiếu, ID trùng, giá trị phạm vi bất thường (min/max).

Giải thích chi tiết theo từng phương thức

**Phân tích thời gian: analyze\_temporal**

* **Tác vụ:** Trích xuất Hour, Month, DayOfWeek từ Start\_Time; đếm tai nạn theo Year; nhận diện top 5 giờ cao điểm.
* **Điểm cần chú ý:** Cần bảo đảm Start\_Time đã ở dạng datetime
* **Trích đoạn code:**

**A screen shot of a computer code

AI-generated content may be incorrect.**

**Giá trị đem lại:** Xác định nhịp thời gian (giờ cao điểm, mùa/tháng ảnh hưởng), phục vụ phân bổ nguồn lực hoặc làm feature cho mô hình.

**Kết quả**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Phân tích địa lý: analyze\_geographic**

* **Tác vụ:** Thống kê top 10 bang; nhóm theo cặp (State, City) để tìm 10 địa điểm có nhiều tai nạn nhất.
* **Trích đoạn code:**

**A computer code with text

AI-generated content may be incorrect.Giá trị đem lại:** Hiểu tập trung không gian, phát hiện điểm nóng theo bang/thành phố để ưu tiên phân tích.

**Kết quả**

A screenshot of a computer

AI-generated content may be incorrect.A table of numbers and letters

AI-generated content may be incorrect.

**Phân tích mức độ nghiêm trọng: analyze\_severity**

* **Tác vụ:** Gom nhóm theo Severity, tính số lượng (count) và thời lượng trung bình Duration\_min.
* **Trích đoạn code:**

**A computer code with red and black text

AI-generated content may be incorrect.Giá trị đem lại:** Liên hệ giữa severity và thời lượng sự kiện, dùng để đánh giá tác động và làm feature.

**Kết quả**

**A close-up of a number

AI-generated content may be incorrect.**

**Kiểm tra chất lượng dữ liệu: check\_data\_quality**

* **Tác vụ:** Liệt kê các cột có missing; đếm ID trùng; báo cáo min/max cho 3 biến số chính.
* **Trích đoạn code:**

**A screen shot of a computer code

AI-generated content may be incorrect.**

**Giá trị đem lại:** Bằng chứng tình trạng dữ liệu, định hướng chính xác cho bước cleaning.

**Kết quả**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Quy trình tổng hợp: run\_all**

* **Tác vụ:** Chuẩn hóa datetime cho Start\_Time, End\_Time; tạo Duration\_min nếu thiếu; gọi tất cả các phân tích.
* **Trích đoạn code:A screen shot of a computer code

  AI-generated content may be incorrect.Lưu ý:** errors='coerce' biến giá trị không hợp lệ thành NaT, giúp tránh crash ở bước tính thời lượng.

## 3.3 Làm sạch dữ liệu & xây dựng đặc trưng (Module: cleaner.py)

* **Chuẩn hóa dữ liệu:** Lấp đầy giá trị thiếu, điều chỉnh ngoại lệ theo domain thresholds.
* **Tăng thông tin:** Tạo features theo thời gian, thời tiết, hạ tầng, vị trí giúp mô hình/BI hiểu ngữ cảnh tai nạn tốt hơn.
* **Bảo toàn pipeline:** Lưu kết quả, đảm bảo có thể tái sử dụng và kiểm chứng.

Giải thích chi tiết theo phương thức

**Xử lý missing values: handle\_missing**

* **Giải pháp:**
* **Numeric:** giá trị median cho Temperature(F), Visibility(mi), Precipitation(in) để ổn định trước outliers.
* **Boolean:** điền False cho các cột hạ tầng (ví dụ Traffic\_Signal) nhằm giả định “không có” khi thiếu.
* **Categorical thời tiết:** điền 'Clear' như trạng thái mặc định an toàn.
* **Trích đoạn code:**

**A screenshot of a computer code

AI-generated content may be incorrect.Lưu ý:** Median giảm ảnh hưởng của phân phối lệch còn với bool\_cols ‘False‘ sẽ là hợp lý nhất với các giá trụ null

**Kết quả**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Xử lý ngoại lệ: handle\_outliers**

**1. Temperature (Nhiệt độ)**

**Ngưỡng hợp lý:**

* Tối thiểu: TEMP\_MIN = -20°F
* Tối đa: TEMP\_MAX = 120°F

**Xử lý:**

* Các giá trị nhỏ hơn -20°F được ép về -20.
* Các giá trị lớn hơn 120°F được ép về 120.

**• Trích đoạn code:**

**A computer code with many colored text

AI-generated content may be incorrect.**

**Kết quả :**

**A black text on a white background

AI-generated content may be incorrect.**

**2. Visibility (Tầm nhìn)**

**Ngưỡng hợp lý:**

* Tối đa: VISIBILITY\_MAX = 10 miles

**Xử lý:**

* Các giá trị lớn hơn 10 được ép về 10.

**Code minh họa:**

**A close-up of a computer code

AI-generated content may be incorrect.Kết quả:**

****

**3. Duration (Thời lượng tai nạn)**

**Ngưỡng hợp lý:**

* Tối đa: DURATION\_MAX\_HOURS = 24 giờ → 1,440 phút
* Tối thiểu: 0 phút

**Xử lý:**

* Các giá trị lớn hơn 1,440 phút được ép về 1,440.
* Các giá trị âm được ép về 0.

**• Trích đoạn code:**

**A computer code with text

AI-generated content may be incorrect.**

**4. Remove invalid severity (Loại bỏ mức độ nghiêm trọng không hợp lệ)**

**Giá trị hợp lệ:**

* Chỉ chấp nhận Severity ∈ {1, 2, 3, 4}

**Xử lý:**

* Loại bỏ các bản ghi có giá trị Severity nằm ngoài tập hợp hợp lệ.

**• Trích đoạn code:**

**A computer code with text

AI-generated content may be incorrect.**

**Tạo đặc trưng thời gian: create\_time\_features**

**• Giải pháp:**

Hàm này tạo ra các đặc trưng thời gian quan trọng từ cột Start\_Time của dữ liệu nhằm hỗ trợ mô hình học máy hiểu rõ hơn bối cảnh thời gian của tai nạn. Cụ thể:

* Tách các thành phần thời gian từ Start\_Time:
  + - Year – năm xảy ra tai nạn
    - Month – tháng
    - Day – ngày
    - Hour – giờ
    - DayOfWeek – thứ (0 = Monday, 6 = Sunday)
    - DayName – tên thứ dạng chữ ("Monday", "Tuesday", ...)
* Tạo đặc trưng Duration\_min:  
   Nếu cột Duration\_min chưa có, hàm sẽ tự tính thời lượng tai nạn bằng:
* End\_Time - Start\_Time (quy đổi sang phút)

Đây là một đặc trưng quan trọng mô tả mức độ nghiêm trọng của tai nạn.

**• Trích đoạn code:**

**A screen shot of a computer code

AI-generated content may be incorrect.**

**• Lưu ý:**

* Việc tách nhỏ các thành phần thời gian giúp mô hình dễ học được chu kỳ theo giờ, theo ngày, theo mùa, hoặc ngày trong tuần.
* Duration\_min chỉ được tạo khi chưa tồn tại ➝ tránh overwrite dữ liệu gốc.
* Cần đảm bảo Start\_Time và End\_Time đã được chuyển về dạng datetime trước khi gọi hàm.

**Tạo khoảng thời gian: get\_time\_period**

**• Giải pháp:**

Hàm này tiếp tục bổ sung các đặc trưng thời gian quan trọng giúp mô hình hiểu rõ bối cảnh giao thông theo khung giờ và cuối tuần:

Phân nhóm khung giờ (Time\_Period):  
Dựa trên giá trị của Hour, tạo ra 4 nhóm:

* **Morning\_Rush (6h–9h):** giờ cao điểm buổi sáng
* **Day (9h–17h):** giờ làm việc ban ngày
* **Evening\_Rush (17h–20h):** cao điểm buổi chiều
* **Night (20h–6h):** ban đêm, ít giao thông hơn
* Xác định cuối tuần (Is\_Weekend):
* Thứ Bảy (5) và Chủ nhật (6) được đánh dấu là cuối tuần
* Chuyển về dạng nhị phân (0/1) hỗ trợ mô hình học máy

**Trích đoạn code:**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Kết quả**

**A black text on a white background

AI-generated content may be incorrect.**

**Tạo đặc trưng thời tiết: create\_weather\_features**

* **Giải pháp:**
  + Tạo cờ (flag) cho mưa (Is\_Rain), tuyết/ice (Is\_Snow), sương mù/mist (Is\_Fog) dựa trên từ khóa trong Weather\_Condition.
  + Đánh dấu Low\_Visibility khi Visibility(mi) < 1.
* **Trích đoạn code:**

**A screenshot of a computer code

AI-generated content may be incorrect.Lưu ý:** na=False tránh lỗi khi giá trị NaN, và .astype(int) chuẩn hóa về 0/1 cho mô hình.

**Kết quả**

**A black text on a white background

AI-generated content may be incorrect.**

**Tạo đặc trưng hạ tầng: create\_infrastructure\_features**

* **Giải pháp:** Cộng dồn các cờ hạ tầng để thành Infra\_Score phản ánh độ “phức tạp” hạ tầng nơi xảy ra tai nạn.
* **Trích đoạn code:**

A computer code with red and blue text

AI-generated content may be incorrect.

**Giá trị đem lại:** Một thước đo đơn giản về môi trường hạ tầng, hữu ích cho phân tích tương quan với severity/frequency.

**Kết quả**

**A black text on a white background

AI-generated content may be incorrect.**

**Tạo định danh vị trí: create\_location\_id**

* **Giải pháp:** Kết hợp State, City, Street thành chuỗi Location\_ID, đảm bảo không bị vỡ bởi giá trị thiếu.
* **Trích đoạn code:**

**A computer code with text

AI-generated content may be incorrect.Giá trị đem lại:** Khóa vị trí dạng text phục vụ phân nhóm/aggregate hoặc join với bảng dim sau này.

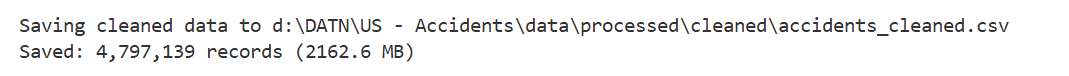
**Lưu dữ liệu sạch: save\_cleaned\_data**

* **Giải pháp:** Đảm bảo thư mục tồn tại, xuất CSV, báo cáo số lượng bản ghi và dung lượng file.
* **Trích đoạn code:**

**A computer code with text

AI-generated content may be incorrect.Giá trị đem lại:** Tạo artifact ổn định cho các bước downstream (EDA chi tiết, modeling, BI).

**Kết quả**



**Pipeline tổng thể: run\_all**

* **Tác vụ:** Ghi log, chạy theo trình tự hợp lý, báo cáo trước/sau xử lý.
* **Trích đoạn code:**

A computer code with text

AI-generated content may be incorrect.

**Lưu ý:** Trình tự này đảm bảo:

* Điền missing trước khi capping outliers (tránh capping trên NaN).
* Tạo features sau khi dữ liệu số đã ổn (giảm rủi ro feature lệch).
* Lưu cuối cùng để kết quả nhất quán.

**Kết quả**

**A number with black text

AI-generated content may be incorrect.**

**Kết quả tổng hợp**

Kết quả làm sạch dữ liệu

**Missing values:**

* + Các biến số (Temperature(F), Visibility(mi), Precipitation(in)) đã được điền bằng giá trị median.
  + Các biến boolean (Amenity, Crossing, Junction, Railway, Station, Stop, Traffic\_Signal) đã được điền bằng False.
  + Biến phân loại Weather\_Condition đã được điền bằng "Clear".  
    → Kết quả: Không còn giá trị thiếu trong các cột quan trọng.

**Outliers:**

* Nhiệt độ được giới hạn trong khoảng -20 đến 120 °F.
* Tầm nhìn được giới hạn tối đa 10 miles.
* Thời lượng tai nạn (Duration\_min) được giới hạn tối đa 24 giờ và tối thiểu 0 phút.
* Các bản ghi có Severity ngoài tập {1, 2, 3, 4} đã bị loại bỏ.  
  → Kết quả: Dữ liệu số nằm trong phạm vi hợp lý, loại bỏ các giá trị cực trị không thực tế.

Kết quả tạo đặc trưng

**Thời gian:**

* + Tạo thêm các cột: Year, Month, Day, Hour, DayOfWeek, DayName.
  + Phân loại khung thời gian thành: Morning\_Rush, Day, Evening\_Rush, Night.
  + Thêm biến Is\_Weekend để đánh dấu cuối tuần.  
    → Kết quả: Bộ dữ liệu có thêm 9 đặc trưng thời gian, hỗ trợ phân tích theo chu kỳ ngày/tuần.

**Thời tiết:**

* + Tạo các cột nhị phân: Is\_Rain, Is\_Snow, Is\_Fog.
  + Thêm biến Low\_Visibility khi tầm nhìn < 1 mile.  
    → Kết quả: Bộ dữ liệu có thêm 4 đặc trưng thời tiết, phản ánh điều kiện môi trường.

**Hạ tầng:**

* + Tính toán Infra\_Score bằng tổng số yếu tố hạ tầng (đèn giao thông, trạm, đường ray, …).  
    → Kết quả: Có thêm 1 đặc trưng tổng hợp phản ánh mức độ phức tạp hạ tầng.

**Vị trí:**

* Tạo Location\_ID bằng cách kết hợp State, City, Street.  
  → Kết quả: Có thêm 1 định danh vị trí duy nhất cho mỗi bản ghi.

Kết quả lưu trữ

* Dữ liệu đã được lưu thành file accidents\_cleaned.csv trong thư mục processed/cleaned.
* Báo cáo số lượng bản ghi trước và sau khi xử lý, cùng dung lượng file xuất ra.

→ Kết quả: Bộ dữ liệu sạch, có đầy đủ đặc trưng mới, sẵn sàng cho bước phân tích và mô hình hóa.

Kết quả phân tích khám phá (EDA)

* **Thời gian:** Xác định được phân phối tai nạn theo năm, giờ cao điểm, ngày trong tuần.
* **Địa lý:** Liệt kê top bang và thành phố có số vụ tai nạn cao nhất.
* **Severity:** Thống kê số lượng và thời lượng trung bình theo mức độ nghiêm trọng.
* **Chất lượng dữ liệu:** Báo cáo tình trạng missing values, duplicate IDs, phạm vi giá trị số.

## 3.4 Xây dựng Mô hình Dữ liệu (Star Schema) (Module: Splitter.py)

**1. Giải pháp tổng quan**

Module DataSplitter thực hiện chuyển đổi dữ liệu tai nạn giao thông đã làm sạch thành **mô hình sao (Star Schema)**, phục vụ phân tích  
 Lược đồ hình sao gồm:

**Dimension tables:**

* **dim\_time** Chứa thông tin ngày: Date (PK), Day, Month, Quarter, Year.
* **dim\_location** Chứa thông tin địa điểm: Location\_id (PK), Street, City, County, State, Timezone
  + kèm các đặc trưng dạng boolean đã được tổng hợp theo từng vị trí (Amenity, Crossing, Junction, Stop, Traffic\_Signal).

**dim\_weather:**

* Chứa thông tin điều kiện thời tiết:
  + weather\_id (PK)
  + Weather\_Condition.

**Fact table – accident\_detail**

* Chứa các sự kiện (tai nạn):
  + ID (PK)
  + Location\_id (FK)
  + weather\_id (FK)
  + full\_date (FK)
  + Start\_Time
  + End\_Time
  + Severity
  + Duration\_min
  + Description.

Module cũng bao gồm bước kiểm tra toàn vẹn tham chiếu nhằm bảo đảm **liên kết** giữa fact và các dimension.

**2. Tạo bảng thời gian – dim\_time**

**Mục tiêu:**

Tách thông tin thời gian từ Start\_Time và chuẩn hóa về dạng bảng thời gian với **Date là khóa chính (PK)**.

**Xử lý chính:**

* Chuyển Start\_Time sang datetime.
* Tách các trường Day, Month, Quarter, Year.
* Lấy giá trị duy nhất → tạo dimension.
* Xuất thành file dim\_time.csv.

**Trích đoạn code:**

A screen shot of a computer program

AI-generated content may be incorrect.

**3. Tạo bảng địa điểm – dim\_location**

**Mục tiêu:**

Chuẩn hóa địa điểm tai nạn theo từng Location\_id, đồng thời tổng hợp số lần xuất hiện các đặc trưng cơ sở hạ tầng.

**Xử lý chính:**

* Tạo khóa chính: Location\_id = Street + "\_" + City.
* Gom nhóm theo vị trí, đếm số True của: Amenity, Crossing, Junction, Stop, Traffic\_Signal.
* Xuất file dim\_location.csv.

**Trích đoạn code:**

A computer code with many colorful text

AI-generated content may be incorrect.

**4. Tạo bảng thời tiết – dim\_weather**

**Mục tiêu:**

Chuẩn hóa danh sách điều kiện thời tiết thành bảng dimension nhỏ gọn.

**Xử lý chính:**

* Lấy các Weather\_Condition duy nhất.
* Tạo khóa chính dạng: W1, W2, W3, …
* Xuất dim\_weather.csv.

**Trích đoạn code:**

A computer code with text

AI-generated content may be incorrect.

**5. Tạo bảng fact – accident\_detail**

**Mục tiêu:**

Lưu thông tin chi tiết từng sự kiện tai nạn, gồm timestamp, độ nghiêm trọng, thời lượng, và liên kết với các bảng dimension.

**Xử lý chính:**

* Gán weather\_id bằng merge với dim\_weather.
* Tạo lại Location\_id nếu chưa tồn tại.
* Tạo full\_date = Start\_Time.date (FK đến dim\_time).
* Tính Duration\_min theo phút.
* Chỉ chọn các cột quan trọng để tạo fact table.
* Xuất file accident\_detail.csv.

**Trích đoạn code:**

A screen shot of a computer code

AI-generated content may be incorrect.

**6. Kiểm tra toàn vẹn tham chiếu**

**Mục tiêu:**

Xác minh rằng tất cả khóa ngoại trong Fact Table đều khớp với Dimension Tables.

* So sánh fact.full\_date với dim\_time.Date
* Kiểm tra Location\_id trong dim\_location
* Kiểm tra weather\_id trong dim\_weather

**Trích đoạn code:**

A screen shot of a computer code

AI-generated content may be incorrect.

**Kết quả**

**A close up of text

AI-generated content may be incorrect.**

**Kết luận**

Quá trình xây dựng Star Schema với các bảng **dim\_time**, **dim\_location**, **dim\_weather** và bảng sự kiện **accident\_detail** đã chuẩn hóa dữ liệu tai nạn giao thông theo mô hình dữ liệu kho (Data Warehouse). Việc tách dữ liệu thành các bảng dimension và fact giúp:

* Giảm trùng lặp dữ liệu và tăng hiệu quả lưu trữ.
* Hỗ trợ phân tích đa chiều theo thời gian, địa điểm và điều kiện thời tiết.
* Đảm bảo toàn vẹn dữ liệu nhờ kiểm tra quan hệ khóa ngoại.

Mô hình dữ liệu hình sao được xây dựng đầy đủ, sạch sẽ và sẵn sàng cho các bước phân tích tiếp theo như trực quan hóa, thống kê chuyên sâu.

**Kết quả**

**A number on a white background

AI-generated content may be incorrect.**

## 3.5 Tạo bảng tổng hợp (Module: Aggregate.py)

**Mục đích**

Aggregation Module được xây dựng nhằm tạo ra các bảng tổng hợp phục vụ trực tiếp cho việc phân tích và xây dựng dashboard. Module giúp giảm kích thước dữ liệu, tăng tốc độ truy vấn và tối ưu cho các công cụ BI như Tableau. Các phép tính quan trọng (YoY, z-score, risk score, time patterns) được chuẩn hoá và thực hiện trước để đảm bảo tính nhất quán. Đồng thời, module tạo các bảng riêng cho từng nhóm người dùng như chính phủ, cảnh sát và bảo hiểm. Kết quả cuối cùng là bộ dữ liệu tổng hợp, sạch và sẵn sàng sử dụng cho trực quan hoá và ra quyết định.

**1. Tổng hợp cấp liên bang – agg\_federal**

**Mục tiêu:** Tạo bảng thống kê tổng quan toàn quốc theo từng năm, phục vụ phân tích xu hướng tai nạn ở cấp Liên Bang (Federal Level).

**Xử lý chính:**

* Gom nhóm dữ liệu theo năm (Year).
* Tính tổng số tai nạn, độ nghiêm trọng trung bình và thời lượng trung bình.
* Chuẩn hóa dữ liệu thành bảng phẳng để sử dụng trực tiếp cho dashboard.
* Xuất file: agg\_federal.csv.

**Trích đoạn code:**

**A computer code with text

AI-generated content may be incorrect.**

**2. Phát hiện bất thường theo bang – agg\_state\_anomaly**

**Mục tiêu:** Xác định các bang có số lượng tai nạn tăng hoặc giảm bất thường so với trung bình nhiều năm trước.

**Xử lý chính:**

* Gom nhóm dữ liệu theo State và Year.
* Tính trung bình lịch sử của mỗi bang.
* So sánh năm hiện tại với trung bình quá khứ để phát hiện điểm bất thường.
* Gắn cờ bất thường (anomaly\_flag).
* Xuất file: agg\_state\_anomaly.csv.

**Trích đoạn code:**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**3. Tổng hợp rủi ro cơ sở hạ tầng – agg\_infrastructure**

**Mục tiêu:**Đánh giá mức độ rủi ro dựa trên các yếu tố hạ tầng như Crossing, Junction, Traffic Signal,… để hỗ trợ phân tích an toàn giao thông.

**Xử lý chính:**

* Gom nhóm theo Location\_id.
* Đếm số lần xuất hiện của các biến boolean hạ tầng.
* Tính điểm rủi ro bằng tổng số điều kiện hạ tầng liên quan.
* Xuất file: agg\_infrastructure.csv.

**Trích đoạn code:**

**A computer screen shot of text

AI-generated content may be incorrect.**

**4. Phân tích ảnh hưởng thời tiết – agg\_weather\_deep**

**Mục tiêu:**  
Đánh giá xem các điều kiện thời tiết nào có tác động lớn nhất tới mức độ nghiêm trọng và tần suất tai nạn.

**Xử lý chính:**

* Gom nhóm theo Weather\_Condition.
* Tính tổng số tai nạn, độ nghiêm trọng trung bình và thời lượng trung bình.
* Sắp xếp theo mức độ ảnh hưởng.
* Xuất file: agg\_weather\_deep.csv.

**Trích đoạn code:**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**5. Pareto 80/20 theo thành phố – agg\_city\_pareto**

**Mục tiêu:**  
Xác định 20% thành phố đóng góp tới 80% tổng số tai nạn để ưu tiên phân bổ nguồn lực.

**Xử lý chính:**

* Gom nhóm theo City để đếm số tai nạn.
* Sắp xếp giảm dần theo số lượng tai nạn.
* Tính tỷ lệ tích lũy (cumulative percentage).
* Xác định nhóm thành phố thuộc vùng Pareto.
* Xuất agg\_city\_pareto.csv.

**Trích đoạn code:**

A computer code with text

AI-generated content may be incorrect.

**6. Phân tích mẫu thời gian – agg\_time\_pattern (Supplementary)**

**Mục tiêu:**  
Xác định các khung giờ và ngày trong tuần có tần suất tai nạn cao, hỗ trợ bố trí lực lượng tuần tra và tối ưu hoá nguồn lực an toàn giao thông.

**Xử lý chính:**

* Gom nhóm dữ liệu theo:  
  **Hour + DayOfWeek + DayName**.
* Tính:
  + tổng số tai nạn (total\_accidents)
  + mức độ nghiêm trọng trung bình
  + thời lượng xử lý trung bình
* Gắn các nhãn phân loại thời gian:
  + Giờ cao điểm (rush hour) 7–9 AM, 4–6 PM
  + Cuối tuần (weekend)
  + Các khoảng thời gian: Morning, Lunch, Afternoon, Evening Rush, Night
* Tính hotspot\_score (thang 0–10) dựa trên mức độ tập trung tai nạn.
* Xuất file: **agg\_time\_pattern.csv**.

**Trích đoạn code:**

A white background with red text

AI-generated content may be incorrect.

**7. Sinh bảng tóm tắt – generate\_summary**

**Mục tiêu:**  
Tạo bảng tổng hợp cấu trúc của tất cả các file tổng hợp (Aggregates) để phục vụ kiểm tra nhanh kích thước, số bản ghi, số cột — đặc biệt hữu ích cho team Dashboard/BI.

**Xử lý chính:**

* Đọc từng file aggregate đã xuất.
* Lưu metadata:
  + số bản ghi (Records)
  + số cột (Columns)
  + dung lượng file (KB)
  + dashboard liên quan
  + stakeholder (nhóm người dùng cuối)
* Xuất file: **\_aggregate\_summary.csv**.

**Trích đoạn code:**

A screen shot of a computer program

AI-generated content may be incorrect.

**Kết quả**

**A screenshot of a computer

AI-generated content may be incorrect.**

# Chương 4: Xây dựng và phát triển sản phẩm

* Các bước xây dựng sản phẩm
* Mô hình dự đoán / phân tích / báo cáo
* Triển khai dashboard, công cụ, giao diện *(nếu có)*

## Aggregate Tables Thiết kế

### Nguyên tắc thiết kế Aggregates

1. **Pre-calculate metrics**: Tính sẵn các metrics phức tạp
2. **Match dashboard grain**: Mỗi aggregate tương ứng 1 dashboard
3. **Include derived fields**: YoY change, percentages, categories

### Agg\_federal

**Stakeholder:** Federal Government (NHTSA, DOT)

**Grain:** Year

|  |  |  |
| --- | --- | --- |
| **Metric** | **Công thức** | **Mô tả** |
| total\_accidents | COUNT(\*) | Tổng số tai nạn |
| total\_high\_severity | COUNT(\*) WHERE Severity >= 3 | Số tai nạn nghiêm trọng |
| avg\_severity | AVG(Severity) | Severity trung bình |
| avg\_duration\_min | AVG(Duration\_min) | Duration trung bình |
| yoy\_accidents\_change | (Current - Prev) / Prev \* 100 | % thay đổi YoY |
| high\_severity\_pct | high\_severity / total \* 100 | % tai nạn nghiêm trọng |
| cumulative\_accidents | RUNNING\_SUM(total) | Tích lũy |

### Agg\_state\_anomaly

Stakeholder: State/Local Government

Grain: State + Year

|  |  |  |
| --- | --- | --- |
| **Metric** | **Công thức** | **Mô tả** |
| is\_anomaly | If [state\_avg] > national\_avg | Flag bất thường |
| severity\_zscore | (state\_avg - national\_avg) / national\_std | Z-score |
| anomaly\_category | Based on Z-score | Critical/High/Elevated/Normal |
| pct\_of\_national | state\_total / national\_total \* 100 | % so với quốc gia |
| yoy\_change | (Current - Prev) / Prev \* 100 | % thay đổi YoY |

**Anomaly Category Rules:**

Z-score > 2.0 → Critical

Z-score > 1.0 → High

Z-score > 0.0 → Elevated

Z-score ≤ 0.0 → Normal

### Agg\_infrastructure

**Stakeholder:** Cơ quan an ninh và cơ sở hạ tầng

**Grain:** State + Urban/Rural

|  |  |  |
| --- | --- | --- |
| **Metric** | **Công thức** | **Mô tả** |
| infra\_risk\_score | Weighted sum / total\_accidents | Điểm rủi ro |
| is\_high\_risk | 1 if score > median | Flag rủi ro cao |
| risk\_category | Based on score + severity | Critical/High/Medium/Low |
| high\_severity\_rate | high\_sev / total \* 100 | Tỷ lệ nghiêm trọng |

**Weighted Risk Score Formula:**

infra\_risk\_score = (Junction × 3 + Traffic\_Signal × 2 + Crossing × 2 + Stop × 1 + Amenity × 1) / total\_accidents

Trọng số dựa trên mức độ nguy hiểm: Junction (ngã tư) nguy hiểm nhất → weight = 3.

### Agg\_weather\_deep\_dive

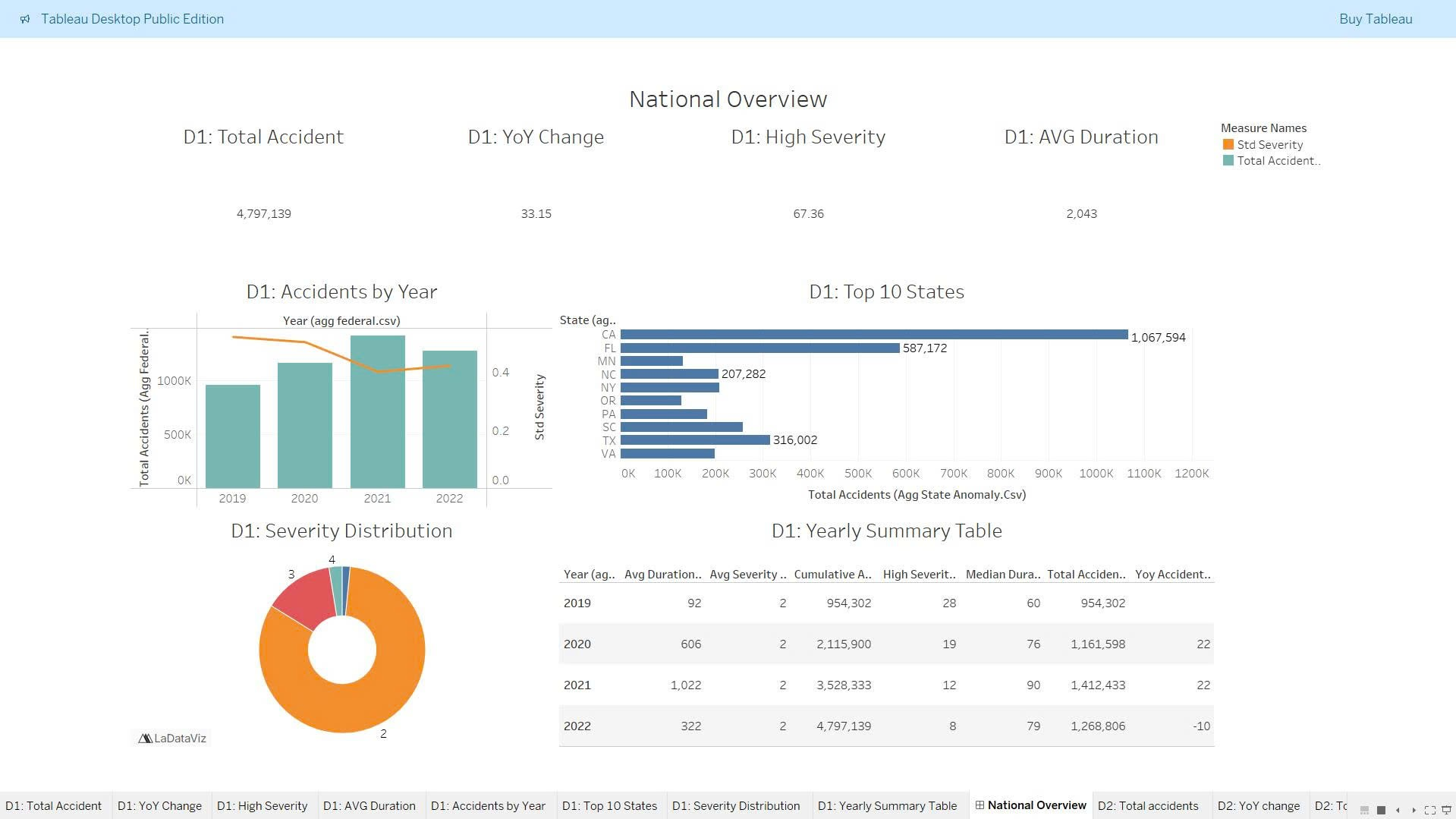
**Stakeholder:** Nạn nhân chịu ảnh hưởng

**Grain:** Weather\_Category + State + Urban/Rural

|  |  |  |
| --- | --- | --- |
| **Metric** | **Công thức** | **Mô tả** |
| severity\_impact\_pct | (weather\_avg - clear\_baseline) / clear\_baseline × 100 | % tác động severity |
| duration\_impact\_pct | (weather\_dur - clear\_dur) / clear\_dur × 100 | % tác động duration |
| weather\_risk\_score | Composite (0-8) | Điểm rủi ro tổng hợp |
| risk\_category | Based on risk\_score | Extreme/High/Moderate/Low |

## Dashboard Thiết kế

### Dashboard 1: National Overview



Hình 1.1.Dashboard 1

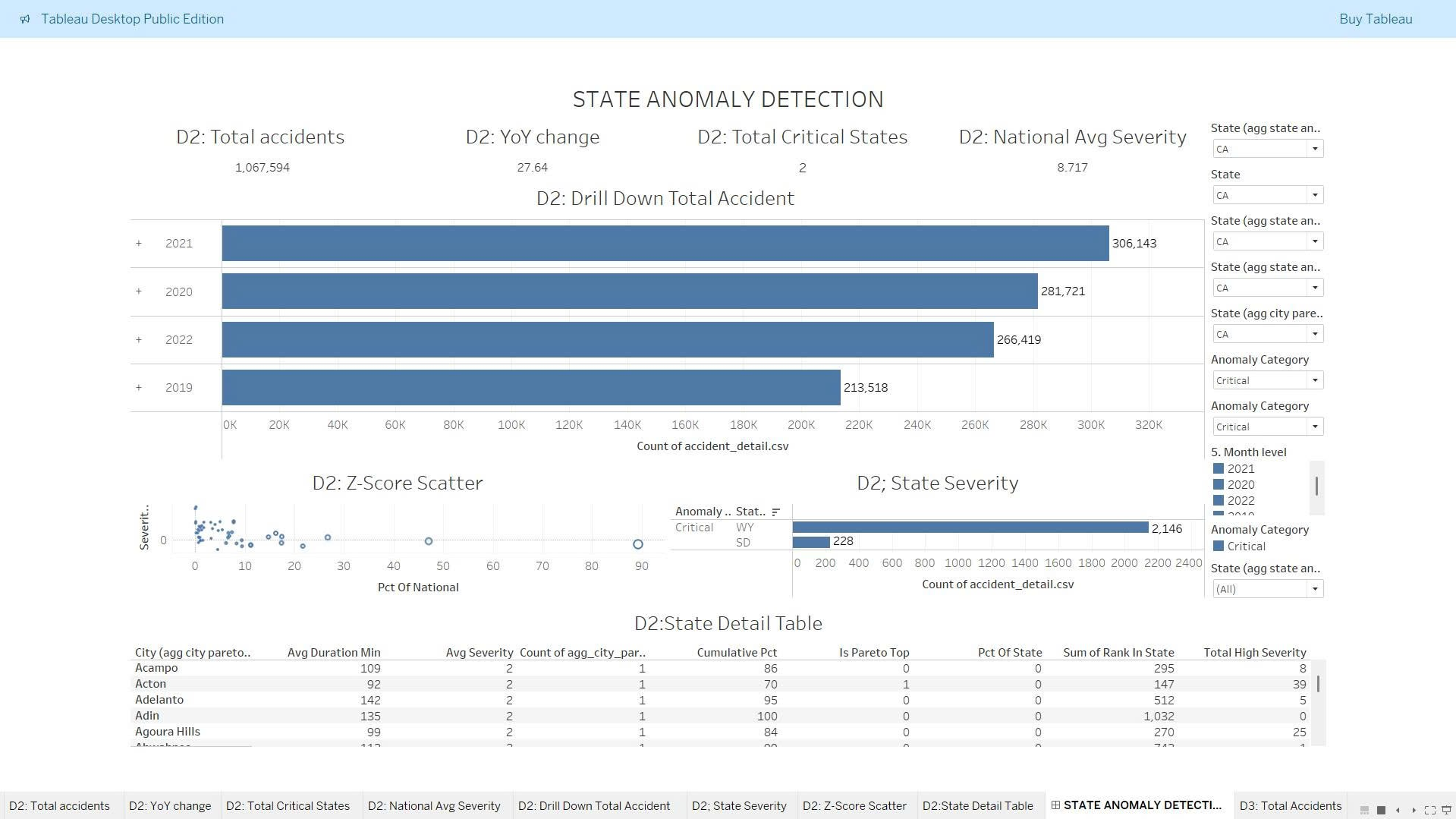
**Đối tượng:** Chính phủ liên bang (NHTSA, DOT)

**Mục đích:** Hiểu xu hướng tai nạn quốc gia, phát hiện bang cần quan tâm từ đó mở rộng phân tích

**Tần suất xem:** Yearly

|  |  |
| --- | --- |
| **#** | **Insight** |
| 1 | Tổng số tai nạn giao thông tại Hoa Kỳ trong giai đoạn 2019-2022 là bao nhiêu? |
| 2 | Xu hướng tai nạn thay đổi như thế nào qua các năm? Năm nào tăng/giảm đột biến? |
| 3 | Top 10 bang nào có số lượng tai nạn cao nhất và chiếm bao nhiêu % tổng quốc gia? |
| 4 | Phân bố mức độ nghiêm trọng (Severity 1-4) như thế nào? Tỷ lệ tai nạn nghiêm trọng (Severity ≥ 3) là bao nhiêu? |
| 5 | Thời gian ảnh hưởng giao thông trung bình của mỗi vụ tai nạn là bao lâu? Có xu hướng tăng hay giảm? |

### Dashboard 2: State Anomaly Detection

****

Hình 1.2. Dashboard 2

**Đối tượng:** Chính phủ tại 1 bang cụ thể

**Mục đích:** Thống kê các chỉ số quan trọng trong 1 bang cụ thểvà so sánh

**Tần suất xem:** Monthly

|  |  |
| --- | --- |
| **#** | **Insight** |
| 1 | Có bao nhiêu bang có mức độ nghiêm trọng trung bình cao hơn mức trung bình quốc gia (Anomaly States)? |
| 2 | Bang nào đang ở mức "Critical" (Z-score > 2) cần can thiệp khẩn cấp? |
| 3 | So sánh Avg Severity của từng bang với National Average như thế nào? Khoảng cách bao xa? |
| 4 | Bang nào vừa có Z-score cao (nguy hiểm) vừa chiếm tỷ trọng lớn (% of National) - tức impact lớn nhất? |
| 5 | Xu hướng YoY của từng bang như thế nào? Bang nào đang cải thiện, bang nào đang xấu đi? |

### Dashboard 3: Infrastructure Statistics



Hình 1.3. Dashboard 3

**Đối tượng:** Cơ quan an ninh và cơ sở hạ tầng (CISA)

**Mục đích:** Xác định điểm nóng, triển khai tuần tra

**Tần suất xem:** Weekly

|  |  |
| --- | --- |
| **#** | **Insight** |
| 1 | Có bao nhiêu location được xếp loại High Risk cần tăng cường tuần tra? |
| 2 | So sánh Urban vs Rural: Khu vực nào có nhiều tai nạn hơn? Khu vực nào có severity cao hơn? |
| 3 | Loại hạ tầng nào (Junction, Traffic Signal, Crossing, Stop, Amenity) có tương quan cao nhất với tai nạn nghiêm trọng? |
| 4 | Phân bố Risk Score như thế nào? Bao nhiêu % location thuộc từng category (Critical/High/Medium/Low)? |
| 5 | Tại mỗi State, nên ưu tiên triển khai lực lượng ở Urban hay Rural để giảm severity hiệu quả nhất? |

### Dashboard 4: Weather Deep Dive



Hình 1.4.Dashboard

**Đối tượng:** Người tham gia giao thông

**Mục đích:** Phân tích tác động thời tiết

**Tần suất xem:** Monthly

|  |  |
| --- | --- |
| **#** | **Insight** |
| 1 | Điều kiện thời tiết nào làm tăng mức độ nghiêm trọng (Severity) nhiều nhất so với trời quang (Clear)? |
| 2 | Điều kiện thời tiết nào làm tăng thời gian ảnh hưởng giao thông (Duration) nhiều nhất? |
| 3 | Có bao nhiêu tổ hợp Weather × Location được xếp loại Extreme Risk? Đó là những tổ hợp nào? |
| 4 | Tại mỗi State, loại thời tiết nào gây rủi ro cao nhất? (để điều chỉnh phí bảo hiểm theo vùng) |
| 5 | Dựa trên phân tích, những khuyến nghị cụ thể nào cho từng tổ hợp Weather + Location để giảm thiểu rủi ro? |

### Thiết kế Principles

1. **Consistency**: Cùng một metric dùng cùng màu trên tất cả dashboards
2. **Hierarchy**: KPIs ở trên, charts ở giữa, tables ở dưới
3. **White Space**: Đủ khoảng trống giữa các elements
4. **Tooltips**: Mọi chart có tooltip giải thích chi tiết
5. **Filters**: Đặt ở góc phải trên, áp dụng cho toàn dashboard

# Chương 5: Kết quả và đánh giá

## 5.1. Kết quả xử lý dữ liệu

### 5.1.1. Data Pipeline Output

|  |  |  |  |
| --- | --- | --- | --- |
| **Bước** | **Input** | **Output** | **Thời gian** |
| Load Raw | CSV 1.5GB | 7,728,394 rows | ~2 min |
| Filter | 7.7M rows | 4,797,139 rows | ~30 sec |
| Clean | 4.8M rows | 4,797,139 rows | ~5 min |
| Star Schema | Cleaned | 4 tables | ~3 min |
| Aggregates | Schema | 6 tables | ~2 min |
| **Total** | - | - | **~12 min** |

### 5.1.2. Star Schema Output

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Records** | **Size** | **PK** |
| dim\_time | 1,461 | ~50 KB | Date |
| dim\_location | ~500,000 | ~50 MB | Location\_id |
| dim\_weather | ~150 | ~5 KB | weather\_id |
| accident\_detail | 4,797,139 | ~800 MB | ID |

### 5.1.3. Aggregate Output

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Records** | **Grain** | **Dashboard** |
| agg\_federal | 4 | Year | National Overview |
| agg\_state\_anomaly | ~200 | State × Year | State Anomaly |
| agg\_infrastructure | ~100 | State × Urban/Rural | Infrastructure |
| agg\_weather\_deep | ~600 | Weather × State × Type | Weather |
| agg\_city\_pareto | ~12,000 | State × City | Supp |
| agg\_time\_pattern | 168 | Hour × DayOfWeek | Supp |

## 5.2 So sánh Kế hoạch ban đầu vs Thực tế

|  |  |  |  |
| --- | --- | --- | --- |
| **Hạng mục** | **Kế hoạch ban đầu** | **Thực tế đạt được** | **Trạng thái** |
| **Data Processing** | Xử lý 4.7M+ records | 4,797,139 records | Đạt |
| **Data Cleaning** | Xử lý missing, outliers | 100% làm sạch | Đạt |
| **Star Schema** | 3 dimensions + 1 fact | 3 dims + 1 fact table | Đạt |
| **Aggregates** | Tối thiểu 4 bảng | 6 bảng aggregate | Vượt |
| **Dashboards** | 4 dashboard | 4 dashboard đầy đủ | Đạt |
| **Insights** | 10+ insights | 10+ insights | Đạt |
| **Machine Learning** | Dự đoán Severity, Accident Probability | Chưa triển khai | Chưa đạt |
| **Real-time Processing** | Streaming pipeline | Batch processing | Chưa đạt |

#### Lý do chưa triển khai Machine Learning

|  |  |
| --- | --- |
| **Yếu tố** | **Chi tiết** |
| **Kiến thức** | Chưa đủ kiến thức chuyên sâu về ML algorithms (Random Forest, XGBoost, Neural Networks) |
| **Feature Engineering** | Cần thêm domain knowledge để tạo features có ý nghĩa cho prediction |
| **Validation** | Chưa nắm vững các phương pháp cross-validation, hyperparameter tuning |
| **Thời gian** | Ưu tiên hoàn thiện phần Data Engineering, Data Analyze và BI trước |

## 5.3. Đánh giá hiệu quả

### 5.3.1. So sánh trước/sau xử lý

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Trước xử lý** | **Sau xử lý** | **Cải thiện sau khi xử lý** |
| Missing values | 2.1M cells | 0 | 100% |
| Outliers | 15,304 | 0 | 100% |
| Độ hoàn thiện data | 73% | 100% | 27% |
| Tables | 1 | 10 | 9 |

### 5.3.2. So với mục tiêu đề ra

|  |  |  |  |
| --- | --- | --- | --- |
| **Mục tiêu** | **Tiêu chí** | **Kết quả** | **Đánh giá** |
| Data Engineering | Xử lý 4.7M+ records, 100% missing | 4.8M records, 100% clean | Đạt |
| Data Modeling | Star Schema | 3 dims + 1 fact | Đạt |
| Data Aggregation | 4+ bảng, 5+ metrics/bảng | 6 bảng, 8-15 metrics/bảng | Vượt |
| Visualization | 4 dashboards với drilldown | 4 Thông số kỹ thuật đầy đủ | Đạt |
| Business Insights | 10+ insights | 15+ insights, 4 stakeholder groups | Vượt |

### 5.3.3. Performance

|  |  |  |
| --- | --- | --- |
| **Metric** | **Giá trị** | **Đánh giá** |
| Tổng thời gian pipeline | ~12 phút | Đạt |
| Output size | ~900 MB | Tối ưu cho Tableau |

## 5.4. Recommendations

* Tai nạn tăng 33% từ 2019 đến 2021, giảm 10% năm 2022
* Top 3 bang (CA, FL, TX) chiếm 41% tổng tai nạn quốc gia
* Pareto: 20% cities chiếm 80% accidents
* Dù Urban có 45 vùng so với Rural 49 vùng nhưng tổng số tai nạn của Urban vẫn cao hơn Rual: Urban (50.13%) vs Rural (49.87%)
* Junction có weighted risk cao nhất
* Urban có risk score cao hơn (96.64%) so với Rural (3.36%)
* Rural có độ biến động (std) cao hơn về mức độ nghiêm trọng — nghĩa là tai nạn ở Rural không đồng đều, có thể dao động mạnh giữa các vụ nhẹ và rất nghiêm trọng.
* Urban tuy risk score cao nhưng std thấp hơn — cho thấy tai nạn Urban có xu hướng đồng đều hơn về mức độ nghiêm trọng (consistently severe hoặc consistently moderate).
* Kết hợp lại: Rural ít vụ nghiêm trọng hơn (risk score thấp) nhưng khi xảy ra thì unpredictable hơn — có thể liên quan đến điều kiện đường xá, thời gian cứu hộ, hoặc thiếu cơ sở hạ tầng y tế.

# CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## 6.1. Tổng kết nội dung thực hiện

Dự án đã hoàn thành các công việc chính:

|  |  |  |
| --- | --- | --- |
| **STT** | **Nội dung** | **Sản phẩm** |
| 1 | Thu thập và lọc dữ liệu | 4.8M records (2019-2022) |
| 2 | Làm sạch và xử lý dữ liệu | accidents\_cleaned.csv |
| 3 | Thiết kế Star Schema | 4 tables (3 dims + 1 fact) |
| 4 | Tạo Aggregate tables | 6 CSV files |
| 5 | Thiết kế Dashboard | 4 Dashboard với các chỉ số kỹ thuật |
| 6 | Phân tích và đề xuất | 15 insight |

## 6.2. Đánh giá tổng quan

#### Điểm mạnh

|  |  |
| --- | --- |
| **Điểm mạnh** | **Mô tả** |
| Pipeline Mở rộng Được | Xử lý 4.8 triệu bản ghi trong 12 phút, có thể mở rộng dễ dàng |
| Code Sạch sẽ | Các module Python được tách biệt, dễ bảo trì |
| BI-Optimized | Aggregate tables tối ưu cho Tableau |
| Tập trung vào Đối tượng | 4 dashboards cho 4 nhóm đối tượng khác nhau |
| Actionable Insights | Recommendations cụ thể, có thể triển khai |

#### Hạn chế

|  |  |
| --- | --- |
| **Hạn chế** | **Nguyên nhân** |
| Chỉ có Yếu tố Ngoại cảnh | Dataset không có thông tin tài xế/phương tiện |
| Xử lý Hàng loạt | Chưa có streaming thời gian thực |
| Chưa có ML | Thiếu kiến thức chuyên sâu |

## 6.3. Hướng phát triển

### 6.3.1. Ngắn hạn (1-3 tháng)

|  |  |  |
| --- | --- | --- |
| **Công việc** | **Mức ưu tiên** | **Mô tả** |
| Deploy to Tableau Public | Cao | Publish để stakeholders truy cập |
| Unit Tests | Trung bình | Thêm pytest cho pipeline |
| Performance Tuning | Trung bình | Parallel processing với multiprocessing |

### 6.3.2. Trung hạn (3-6 tháng)

|  |  |  |
| --- | --- | --- |
| **Công việc** | **Mức ưu tiên** | **Mô tả** |
| Machine Learning | Cao | Mô hình dự đoán mức độ nguy hiểm của vụ tai nạn |
| Integrate FARS Data | Cao | Thêm fatality, vehicle, driver info |
| Real-time Pipeline | Trung bình | Streaming với Kafka/Spark |
| REST API | Trung bình | API service cho data access |

#### 6.3.3. Dài hạn (6-12 tháng)

|  |  |  |
| --- | --- | --- |
| **Công việc** | **Mức ưu tiên** | **Mô tả** |
| Mobile App | Trung bình | Ứng dụng cảnh báo cho tài xế |
| Alert System | Cao | Real-time danger alerts |
| Traffic Integration | Cao | Tích hợp hệ thống quản lý giao thông |
| International Expansion | Cao | Mở rộng sang dữ liệu của Việt Nam hoặc các nước khác |

#### 6.3.4. Machine Learning Roadmap (Khi đủ kiến thức)

|  |  |  |
| --- | --- | --- |
| **Model** | **Mục đích** | **Input Features** |
| Classification | Dự đoán Severity (1-4) | Weather, Time, Location, Infrastructure |
| Regression | Dự đoán Duration | Severity, Weather, Urban/Rural |
| Clustering | Phân nhóm accident patterns | Tất cả feature |
| Time Series | Dự báo số tai nạn theo tháng | Historical trends |

## 6.5. Lời kết

Dự án "Phân tích Tai nạn Giao thông Hoa Kỳ 2019-2022" đã hoàn thành các mục tiêu đề ra, từ xây dựng ETL pipeline xử lý gần 7.7 triệu bản ghi, thiết kế Star Schema chuẩn, đến tạo hệ thống dashboard specification phục vụ 4 nhóm stakeholder.

Kết quả phân tích cho thấy các yếu tố external (thời tiết, hạ tầng, thời gian) có ảnh hưởng đáng kể đến tai nạn giao thông.

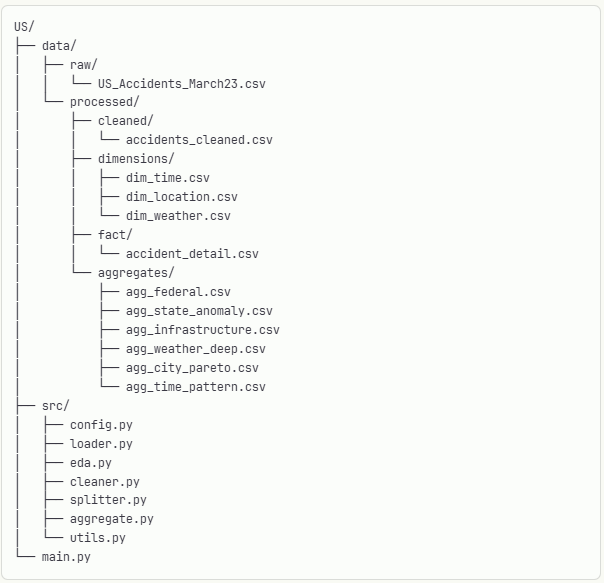
Với các khuyến nghị cụ thể cho từng nhóm stakeholder, dự án hy vọng đóng góp vào mục tiêu chung: giảm thiểu tai nạn giao thông và bảo vệ tính mạng con người trên đường.

### Tài liệu tham khảo

|  |  |  |
| --- | --- | --- |
| **STT** | **Tài liệu** | **Link/Nguồn** |
| **1** | US Accidents Dataset | <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents> |
| **2** | NHTSA Traffic Safety Facts | <https://www.nhtsa.gov/research-data> |
| **3** | Pandas Documentation | https://pandas.pydata.org/docs/ |
| **4** | Tableau Desktop Help | <https://help.tableau.com/current/pro/desktop/> |
| **5** | Star Schema Design - Kimball | The Data Warehouse Toolkit, 3rd Edition |

### Phụ lục

#### Phụ lục A: Cấu trúc thư mục dự án



**Link sản phẩm:** https://github.com/Vincentnpd/US-Accidents

#### Phụ lục B: Data Dictionary

**accident\_detail (Fact Table)**

|  |  |  |
| --- | --- | --- |
| **Column** | **Type** | **Description** |
| ID | String | Tên sự kiện tai nạn |
| Location\_id | String | FK → dim\_location |
| weather\_id | String | FK → dim\_weather |
| full\_date | Date | FK → dim\_time.Date |
| Start\_Time | Datetime | Thời điểm bắt đầu |
| End\_Time | Datetime | Thời điểm kết thúc |
| Severity | Integer | Mức độ nghiêm trọng (1-4) |
| Duration\_min | Float | Thời gian ảnh hưởng (phút) |
| Description | Text | Mô tả tai nạn |

#### Phụ lục C: Tableau Calculated Fields

**YoY Change**

(SUM([Total Accidents]) - LOOKUP(SUM([Total Accidents]), -1)) / ABS(LOOKUP(SUM([Total Accidents]), -1)) \* 100

**Anomaly Flag**

IF [Avg Severity] > [National Avg Severity] THEN 1 ELSE 0 END

**Risk Category**

CASE WHEN [Severity Zscore] > 2 THEN 'Critical' WHEN [Severity Zscore] > 1 THEN 'High' WHEN [Severity Zscore] > 0 THEN 'Elevated' ELSE 'Normal' END

**City Rank (Table Calc)**

RANK\_UNIQUE(SUM([Total Accidents]), 'desc')

#### Phụ lục D: Link sản phẩm

|  |  |
| --- | --- |
| **Sản phẩm** | **Link** |
| GitHub Repository | https://github.com/Vincentnpd/US-Accidents |
| Tableau Public Dashboard | https://public.tableau.com/views/Visualize\_17651699112650/WeatherDeepDive?:language=en-US&:sid=&:redirect=auth&:display\_count=n&:origin=viz\_share\_link |
| Dataset gốc | <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents> |

A qr code on a white background

AI-generated content may be incorrect.