ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH TAI NẠN GIAO THÔNG TẠI THÀNH PHỐ CHICAGO, MỸ TỪ THÁNG 1/2022 ĐẾN THÁNG 10/2024

Nhóm 03			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Lê Xuân Bình	22520131	KHDL
2	Trần Đại Hiển	22520426	KHDL
3	Phạm Ngọc Trí	22521526	KHDL

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Đề tài này tập trung vào việc thu thập và phân tích dữ liệu tai nạn giao thông tại thành phố Chicago, Mỹ từ tháng 1/2022 đến tháng 10/2024. Nhóm đã xây dựng các kịch bản phân tích cụ thể và sử dụng công cụ Power BI để trực quan hóa dữ liệu, làm nổi bật những xu hướng quan trọng như thời gian xảy ra tai nạn, mức độ thiệt hại và thương tích so với điều kiện tai nạn, loại phương tiện tham gia, cũng như phân tích địa lý và nhân khẩu học. Kết quả của đề tài không chỉ cung cấp bức tranh tổng thể về nhiều khía cạnh của các vụ tai nạn giao thông tại Chicago mà còn xây dựng mô hình phân loại mức độ nghiêm trọng của tai nạn dựa trên các điều kiện được chọn lọc, góp phần đề xuất các giải pháp kiểm soát, cảnh báo và giảm thiểu tai nạn giao thông trong tương lai.

Trong đề tài này, chúng tôi cam kết minh bạch về các thông tin liên quan đến bộ dữ liệu và phương pháp phân tích. Bộ dữ liệu được nhóm tự thu thập từ nguồn dữ liệu mở của thành phố Chicago, bao gồm các tập dữ liệu: Traffic Crashes – Crashes [1], Traffic Crashes – People [2], và Traffic Crashes – Vehicles [3]. Nhóm đã trực tiếp làm sạch, xử lý và tổ chức lại các dữ liệu này để phù hợp với mục tiêu nghiên cứu và phân tích của đề tài. Đề tài và bộ dữ liệu hoàn toàn do nhóm tự thiết kế và thực hiện, không sao chép, tham khảo hoặc dựa trên bất kỳ đồ án hoặc dự án mẫu nào khác. Nhóm cam kết cung cấp đầy đủ mã nguồn, tài liệu xử lý dữ liệu và các thông tin liên quan để minh chứng cho quá trình thực hiện, đảm bảo tính trung thực và minh bạch trong quá trình đánh giá.

2. MÔ TẢ BỘ DỮ LỆU

Bộ dữ liệu về tai nạn giao thông bao gồm ba tập dữ liệu: dữ liệu về vụ tai nạn (crashes) , dữ liệu về người liên quan trong vụ tai nạn (people), dữ liệu về phương tiện gặp tai nạn (vehicles) được chúng tôi tự thu thập tại [1], [2], [3], cập nhật lần cuối ngày 06/10/2024.

Mô tả cụ thể về ba tập dữ liệu và thông tin liên kết giữa các bảng như sau:

- crashes: Lưu thông tin về các vụ tai nạn, sử dụng trường khóa chính là *crash_record_id*.
- vehicles: Lưu thông tin về các phương tiện liên quan đến vụ tai nạn, sử dụng trường khóa chính là *vehicle_id*. Bảng này liên kết với bảng crashes thông qua trường *crash_record_id*, cho thấy một vụ tai nạn có thể có nhiều phương tiện tham gia.
- people: Lưu thông tin về những người liên quan đến vụ tai nạn, sử dụng trường khóa chính là *person_id*. Bảng này liên kết với bảng crashes thông qua trường *crash_record_id* và với bảng vehicles thông qua trường *vehicle_id*. Điều này thể hiện một phương tiện có thể chở nhiều người, và một vụ tai nạn có thể liên quan đến nhiều người khác nhau.

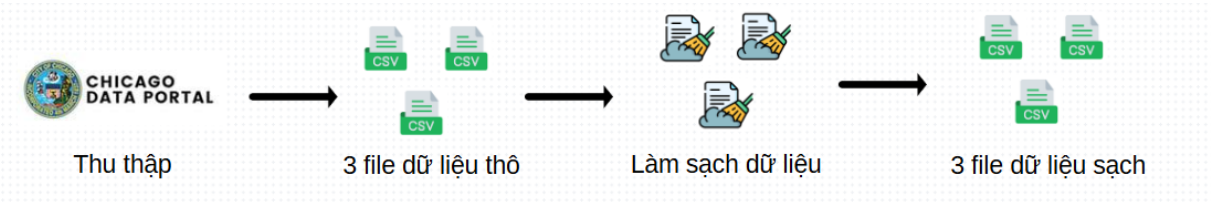
Thống kê về bộ dữ liệu crashes, vehicles, people được chúng tôi trình bày tại *Bảng 1*.

Bảng 1. Thống kê về bộ dữ liệu

Dataset	Số dòng	Số cột	Số biến phân loại	Số biến số
crashes	302,020	25	15	9
vehicles	614,494	14	9	4
people	667,725	13	10	2

Các thuộc tính quan trọng trên từng bảng dữ liệu được trình bày ở Phụ lục 1 của báo cáo này.

Hình 1 biểu diễn quá trình thu thập và xử lý để có bộ dữ liệu mà chúng tôi phân tích.

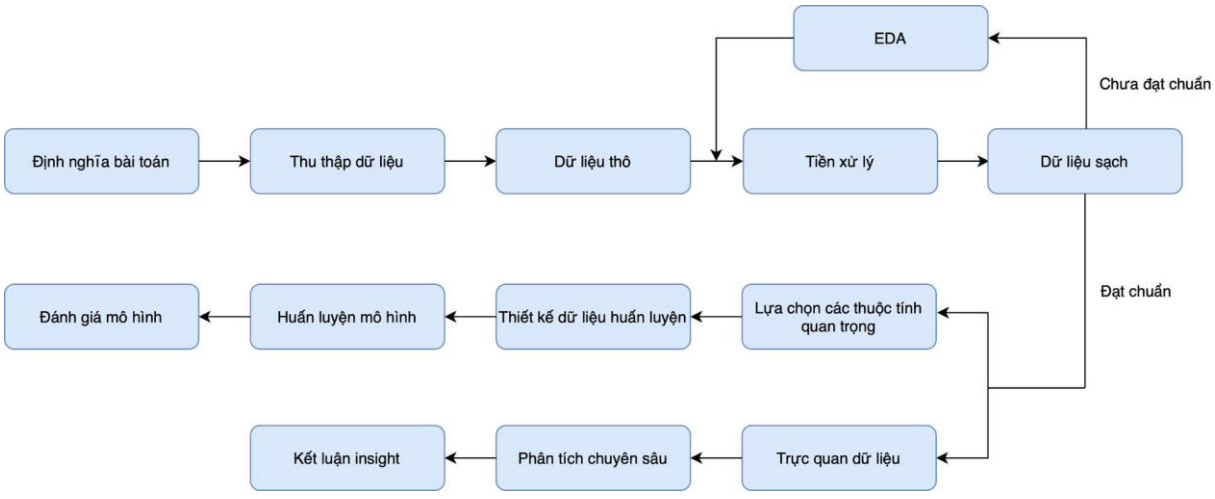


Hình 1: Quy trình thu thập dữ liệu

3. PHƯƠNG PHÁP PHÂN TÍCH

3.1. Quy trình thực hiện

Hình 2 là quy trình thực hiện tiền xử lí, phân tích và xây dựng mô hình đánh giá bộ dữ liệu.



Hình 2: Quy trình thực hiện

3.2. Tiền xử lý và làm sạch dữ liệu

Trong đó, quy trình tiền xử lý và làm sạch dữ liệu được chia thành 3 giai đoạn chính:

- Làm sạch trước EDA (Initial Cleaning):
  + Giới hạn ngày từ 2022-01-01 đến 2024-10-06.

+ Loại bỏ cột có >60% giá trị rỗng và các dòng trùng lặp.
+ Xóa dòng thiếu giá trị ở cột quan trọng (*latitude, longitude*).
+ Hiệu chỉnh kiểu dữ liệu và chuẩn hóa tên cột.

– Phân tích Dữ liệu Khám phá (EDA):
+ Boxplot, bar chart để phát hiện outliers và xu hướng dữ liệu.
+ Ma trận tương quan xác định mối quan hệ tuyến tính giữa các cột số.
+ Heatmap trực quan hóa giá trị rỗng.

– Làm sạch sau cùng (Final Cleaning):
+ Giới hạn giá trị hợp lệ (tuổi, tọa độ).
+ Loại bỏ cột có >99% giá trị trùng lặp hoặc không liên quan.
+ Điền giá trị rỗng:

Phân loại: thay bằng 'UNKNOWN'/'OTHER'.

Số liệu: đặt 0 hoặc giá trị phù hợp.

+ Sử dụng KNN Imputation điền khuyết để giữ nguyên phân phối đặc trưng.

Hình vẽ quy trình được thể hiện cụ thể ở Phụ lục 2 của báo cáo này.

Các hình ảnh trực quan EDA được chúng tôi thể hiện ở Phụ lục 3.

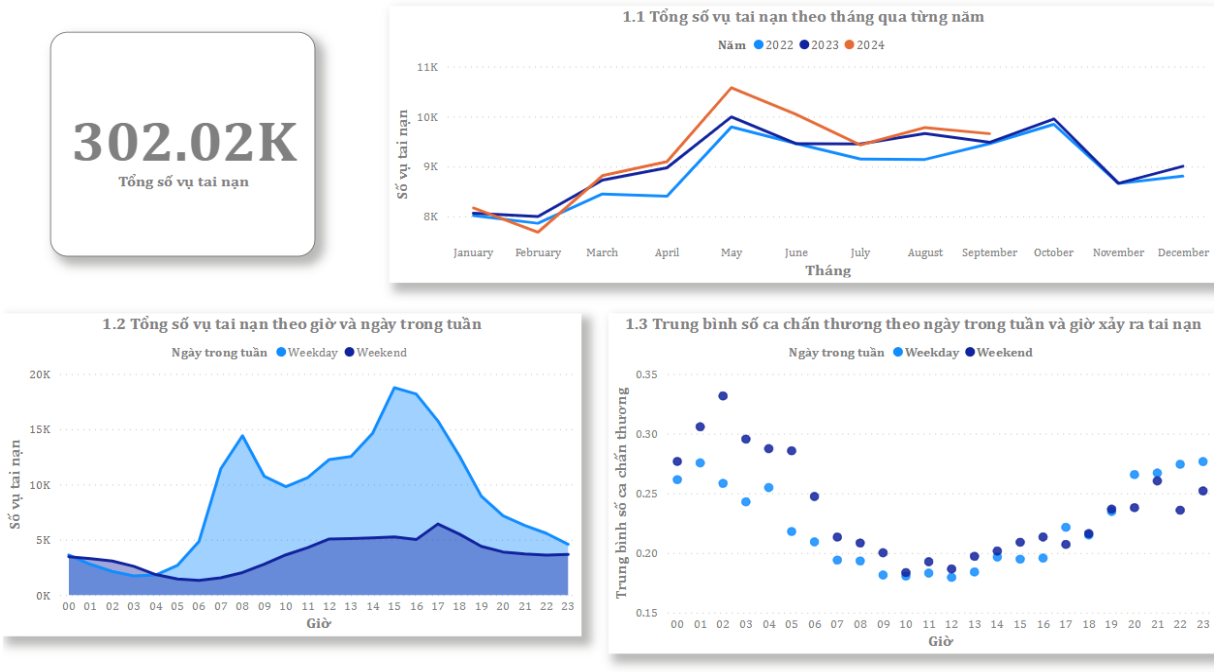
4. PHÂN TÍCH THĂM DÒ CHUYÊN SÂU

Chúng tôi đã tiến hành phân tích và trực quan hóa bộ dữ liệu tai nạn giao thông bằng phần mềm Power BI. Chúng tôi xác định các hướng phân tích chính như sau:

– **Phân tích xu hướng theo thời gian:** Đánh giá số lượng tai nạn giao thông, mức độ nghiêm trọng, hoặc các nguyên nhân phổ biến theo thời gian.
– **Phân tích theo loại phương tiện:** Đánh giá sự ảnh hưởng của các loại phương tiện và hành vi của tài xế trong tai nạn.
– **Phân tích mức độ thiệt hại và thương tích:** Tìm hiểu mức độ thương tích và thiệt hại liên quan đến các yếu tố như điều kiện đường xá, phương tiện và hành vi người tham gia giao thông.
– **Phân tích theo vị trí địa lý:** Phân tích tỉ lệ tử vong xảy ra theo các khu vực địa lý cụ thể giúp xác định các điểm nóng tai nạn và khu vực cần ưu tiên cải thiện tình trạng này.

4.1. Phân tích xu hướng theo thời gian

Phân tích trực quan hóa xu hướng theo thời gian được chúng tôi thể hiện ở *Hình 3*.



Hình 3: Dashboard trực quan phân tích xu hướng theo thời gian

4.1.1. Tổng số vụ tai nạn theo tháng qua từng năm (Biểu đồ 1.1)

Số vụ tai nạn tăng dần từ đầu năm, đạt đỉnh vào tháng 5 và giảm vào các tháng cuối năm (năm 2024 chúng tôi chỉ thu thập đến thời điểm 6/10). So sánh giữa các năm cho thấy tháng 1 và 2 ổn định, trong khi các tháng sau trở đi có xu hướng tăng cao hơn, đặc biệt là tháng 5 do hoạt động giao thông mùa hè.

4.1.2. Tổng số vụ tai nạn theo giờ và ngày trong tuần (biểu đồ 1.2)

Xu hướng theo giờ:

- Số vụ tai nạn tăng đột biến từ khoảng 6 giờ sáng đến 9 giờ sáng và từ 15 giờ đến 18 giờ. Đây là các khung giờ cao điểm khi người dân đi làm và tan tầm.
- Số vụ tai nạn giảm dần vào ban đêm (từ 0 giờ đến 5 giờ sáng) và ổn định ở mức thấp.

So sánh giữa ngày trong tuần:

- Trong ngày thường (Weekday), số vụ tai nạn cao hơn rõ rệt, đặc biệt trong khung giờ 6h-9h và 15h-18h.
- Trong ngày cuối tuần (Weekend), số vụ tai nạn giảm, tuy nhiên vẫn tập trung nhiều vào buổi chiều và tối (từ 16 giờ trở đi).

4.1.3. Trung bình số ca chấn thương theo ngày và giờ (biểu đồ 1.3)

Xu hướng chung: Trung bình số ca chấn thương cũng thể hiện xu hướng tương tự với số vụ tai nạn, cao hơn vào các giờ cao điểm.

So sánh giữa ngày thường và cuối tuần: Ngày cuối tuần (Weekend) và ngày thường (Weekday) có xu hướng giống nhau

- Có xu hướng chấn thương cao hơn vào các giờ tối muộn (sau 20 giờ), có thể liên quan đến các hoạt động giải trí và tình trạng lái xe kém an toàn vào cuối ngày.
- Có xu hướng tăng cao vào hai khung giờ cao điểm
  + Buổi sáng: Từ 6 giờ đến 9 giờ, khi mật độ giao thông tăng mạnh do người dân di chuyển đến nơi làm việc và trường học.
  + Buổi chiều: Từ 15 giờ đến 18 giờ, khi lượng phương tiện gia tăng đáng kể trong khoảng thời gian tan tầm.

4.1.4. Kết luận chung:

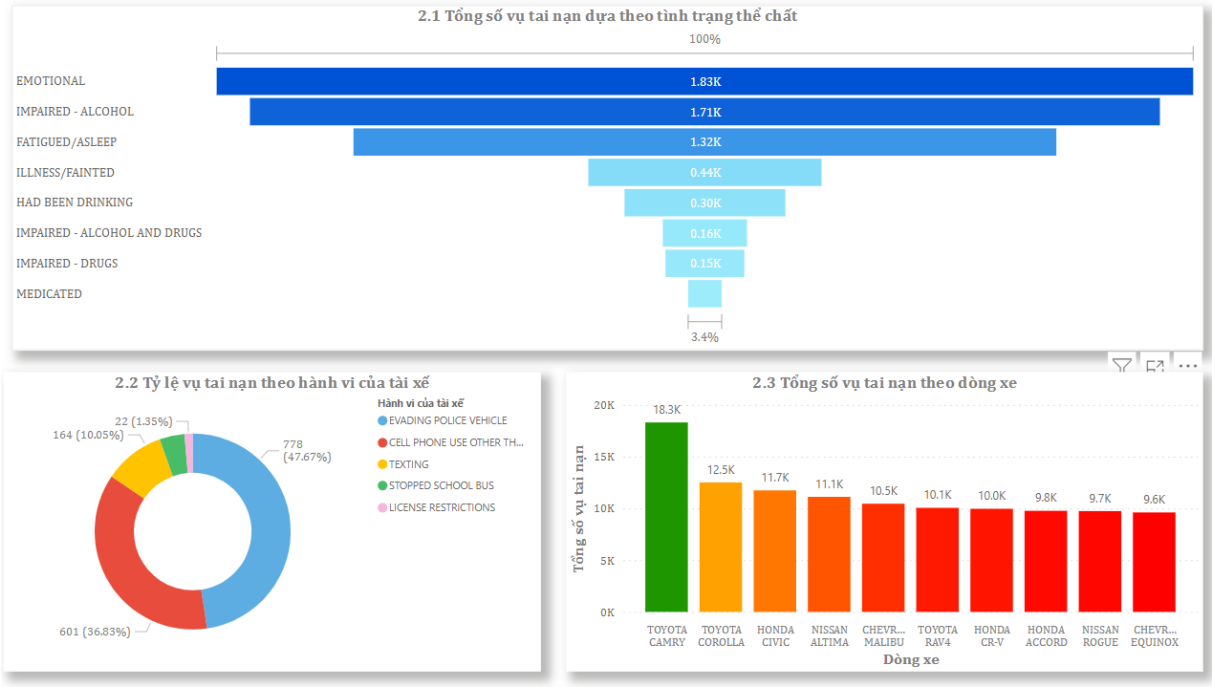
Tháng 5 là tháng cao điểm ghi nhận số vụ tai nạn cao nhất.

Khung giờ rủi ro cao: 6-9h sáng và 15-18h chiều là các khung giờ tai nạn tập trung nhiều nhất trong ngày thường.

Tai nạn và số ca chấn thương tăng vào buổi tối muộn (đặc biệt là ngày cuối tuần).

4.2. Phân tích theo loại phương tiện

Phân tích trực quan hóa xu hướng theo loại phương tiện được chúng tôi thể hiện ở *Hình 4*.



Hình 4: Dashboard trực quan phân tích theo loại phương tiện

4.2.1. Tổng số vụ tai nạn dựa theo tình trạng thể chất (biểu đồ 2.1)

Tình trạng cảm xúc (EMOTIONAL), ảnh hưởng bởi rượu bia (IMPAIRED - ALCOHOL) và mệt mỏi (FATIGUED/ASLEEP) là 3 nguyên nhân phổ biến nhất, chiếm phần lớn số vụ tai nạn.

Các tình trạng sức khỏe yếu (hoặc bị ảnh hưởng bởi chất kích thích tuy ít hơn nhưng cần đặc biệt quan tâm vì mức độ nghiêm trọng của các vụ tai nạn này có thể cao hơn.

4.2.2. Tỷ lệ tai nạn theo hành vi của tài xế (biểu đồ 2.2)

Evading Police Vehicle (Trốn tránh cảnh sát) chiếm tỷ lệ cao nhất (47.67%) với 778 vụ. Đây là hành vi nguy hiểm và có thể dẫn đến tai nạn nghiêm trọng do tốc độ cao hoặc thiếu kiểm soát khi lái xe.

Tiếp đó, Cell Phone Use chiếm tỷ lệ 36.83% với 601 vụ. Hành vi này cho thấy vấn đề mất tập trung trong quá trình lái xe vẫn là nguyên nhân quan trọng dẫn đến các vụ tai nạn.

Texting (10.05%) và các lỗi như License Restrictions hay Stopped School Bus có tỷ lệ thấp hơn nhưng vẫn cần lưu ý.

4.2.3. Tổng số vụ tai nạn theo dòng xe (biểu đồ 2.3)

Phương tiện liên quan đến nhiều vụ tai nạn nhất:

- Toyota Camry đứng đầu với khoảng 19K vụ tai nạn.
- Tiếp theo là Toyota Corolla và Honda Civic, với số vụ tai nạn dao động từ 12K đến 14K.
- Các phương tiện như Nissan Altima, Chevrolet Malibu, Toyota RAV4 và Honda CR-V đều có số vụ tai nạn tương đối cao.

4.2.4. Kết luận chung:

Tình trạng thể chất ảnh hưởng đến khả năng lái xe và là nguyên nhân dẫn đến tai nạn giao thông. Việc tập trung giảm thiểu các yếu tố hàng đầu như cảm xúc tiêu cực, sử dụng rượu bia, và tình trạng mệt mỏi sẽ giúp giảm thiểu đáng kể số vụ tai nạn trong tương lai.

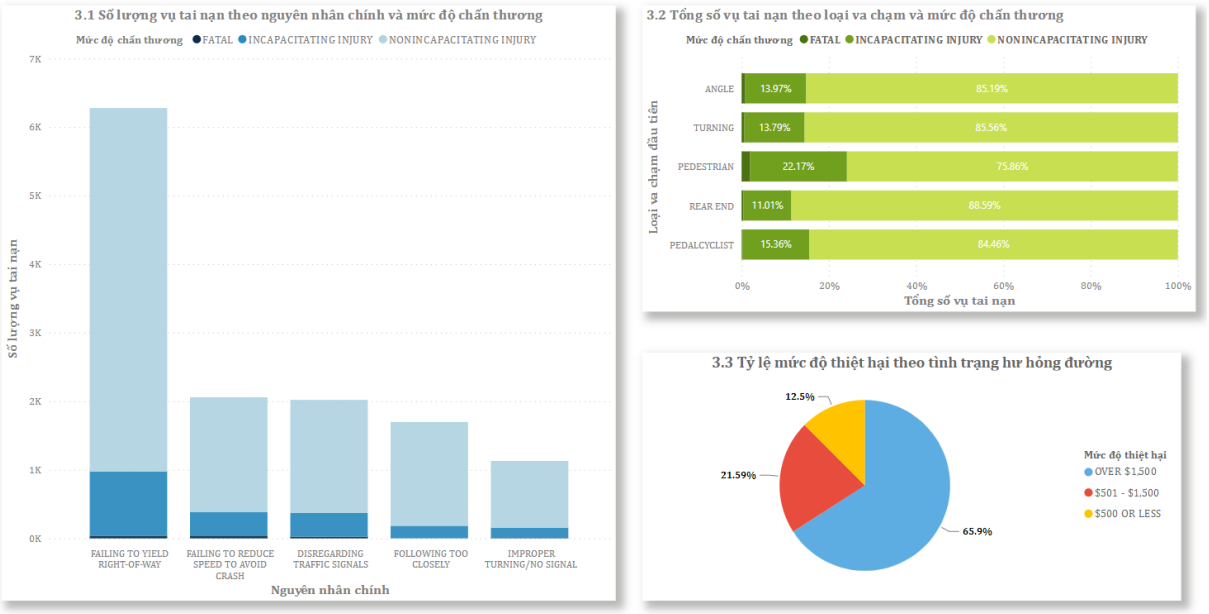
Hành vi nguy hiểm của tài xế, đặc biệt là trốn tránh cảnh sát và sử dụng điện thoại khi lái xe, là nguyên nhân chính dẫn đến tai nạn.

Các dòng xe phổ biến như Toyota Camry, Corolla, và Honda Civic có tỷ lệ tai nạn cao.

Việc kết hợp phân tích hành vi tài xế và loại phương tiện sẽ giúp cơ quan chức năng triển khai các biện pháp phù hợp nhằm nâng cao an toàn giao thông.

4.3. Phân tích mức độ thiệt hại và thương tích

Phân tích trực quan hóa xu hướng theo mức độ thiệt hại và thương tích được chúng tôi thể hiện ở *Hình 5*.



Hình 5: Dashboard trực quan phân tích mức độ thiệt hại và thương tích.

4.3.1. Số lượng vụ tai nạn theo nguyên nhân chính và mức độ chấn thương (biểu đồ 3.1)

Không nhường đường (Failing to Yield Right-of-Way) là nguyên nhân chính, chiếm số lượng lớn nhất với hơn 6K vụ tai nạn. Chấn thương không nghiêm trọng (Nonincapacitating Injury) chiếm phần lớn, với một lượng nhỏ là chấn thương nghiêm trọng và tử vong.

Hành vi của người lái xe, đặc biệt là không nhường đường và vi phạm tín hiệu giao thông là nguyên nhân chính dẫn đến tai nạn với thương tích ở mức trung bình.

4.3.2. Tổng số vụ tai nạn theo loại va chạm và mức độ chấn thương (biểu đồ 3.2)

Va chạm góc (Angle) và chuyển hướng (Turning) có tỷ lệ cao nhất về tai nạn (trên 85%). Trong đó, tỷ lệ chấn thương không nghiêm trọng (Nonincapacitating Injury) chiếm ưu thế.

Va chạm phía sau (Rear End) và người đi bộ (Pedestrian) có tỷ lệ chấn thương nghiêm trọng và tử vong cao hơn các loại va chạm khác.

Tai nạn liên quan đến người đi bộ và xe đạp thường dẫn đến thương tích nặng hơn do không có bảo hộ. Va chạm phía sau và chuyển hướng sai thường xảy ra do không giữ khoảng cách an toàn và tín hiệu sai lệch.

4.3.3. Tỷ lệ mức độ thiệt hại theo tình trạng hư hỏng đường (biểu đồ 3.3)

Biểu đồ tròn phía trên cho thấy:

- 65.9% vụ tai nạn gây thiệt hại lớn hơn $1,500.
- 21.59% vụ tai nạn có mức thiệt hại từ $501 - $1,500.
- Chỉ 12.5% vụ tai nạn gây thiệt hại nhỏ hơn $500.

Phần lớn tai nạn gây thiệt hại nặng nề về kinh tế.

4.3.4. Kết luận chung:

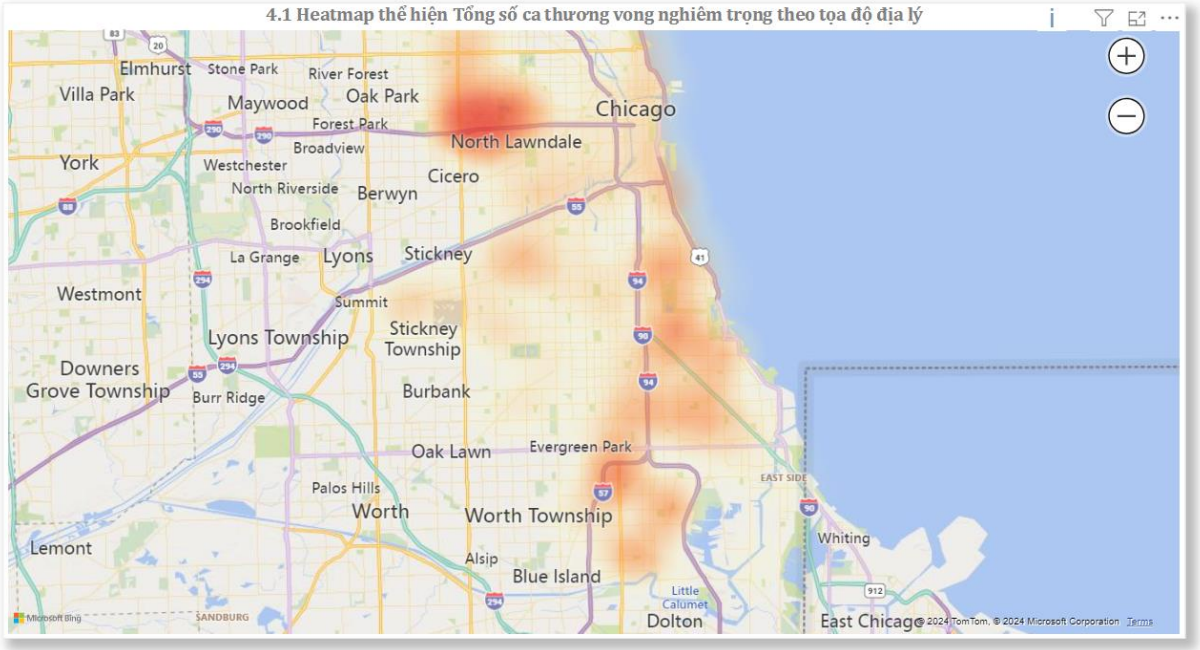
Hành vi lái xe như không nhường đường và vi phạm tín hiệu giao thông là nguyên nhân chủ yếu dẫn đến tai nạn, nhưng thương tích chủ yếu ở mức không nghiêm trọng.

Loại va chạm như góc (Angle) và chuyển hướng (Turning) chiếm tỷ lệ lớn nhất về số vụ tai nạn. Tuy nhiên, tai nạn liên quan đến người đi bộ và va chạm sau có nguy cơ gây thương tích nghiêm trọng và tử vong cao hơn.

Thiệt hại kinh tế chủ yếu ở mức cao (> $1,500), cho thấy mức độ nghiêm trọng của va chạm.

4.4. Phân tích theo vị trí địa lý

Phân tích trực quan hóa theo vị trí địa lý được chúng tôi thể hiện ở *Hình 6.*



Hình 6: Dashboard trực quan phân tích theo vị trí địa lý

Tập trung tai nạn: trung tâm thành phố và các khu vực gần bờ hồ Michigan.

Khu vực nguy hiểm cao: Các trục đường băng qua các vùng như "North Lawndale" và "Stickney" xuất hiện với màu đậm, thể hiện số lượng thương vong nghiêm trọng cao.

5. XÂY DỰNG MÔ HÌNH DỰ ĐOÁN VÀ ĐÁNH GIÁ

Chúng tôi thực hiện xây dựng mô hình máy học để dự đoán mức độ nghiêm trọng của tai nạn giao thông dựa trên các điều kiện xảy ra tai nạn, sử dụng dữ liệu từ bộ crashes. Mức độ nghiêm trọng của tai nạn được biểu diễn bằng biến mục tiêu *most_severe_injury*, là một biến phân loại với các giá trị:

- NO INDICATION OF INJURY: Không có dấu hiệu bị thương.
- NONINCAPACITATING INJURY: Bị thương nhưng không nghiêm trọng.
- REPORTED, NOT EVIDENT: Có báo cáo bị thương nhưng không rõ ràng.
- INCAPACITATING INJURY: Bị thương nghiêm trọng.
- FATAL: Tử vong.

Mục tiêu của mô hình này là hỗ trợ tự động hóa quá trình đánh giá, từ đó giúp đưa ra các cảnh báo kịp thời cho các khu vực hoặc điều kiện tương tự trong tương lai.

Quy trình xây dựng mô hình được thể hiện ở Phụ lục 2 của báo cáo này.

5.1. Xác định các đặc trưng quan trọng

Để xác định các đặc trưng quan trọng, chúng tôi đã áp dụng hai kỹ thuật chính: Weight of Evidence (WoE) và Information Value (IV), cùng với ANOVA Test.

Sau khi áp dụng các kỹ thuật trên, chúng tôi đã chọn ra các đặc trưng đầu vào bao gồm:

- Các đặc trưng phân loại (categorical features): *traffic_control_device, device_condition, weather_condition, lighting_condition, first_crash_type, trafficway_type, roadway_surface_cond, crash_type, damage, prim_contributory_cause, sec_contributory_cause.*
- Các đặc trưng số liên tục (numerical features): *posted_speed_limit, num_units, injuries_total, crash_hour, crash_month, latitude, longitude, injuries_no_indication.*

5.2. Chia dữ liệu huấn luyện và kiểm tra

Dữ liệu được chia thành 2 tập train-set (80%) và test-set (20%) và sử dụng kỹ thuật phân tầng (stratified sampling) theo biến mục tiêu (y) để đảm bảo phân phối đồng đều giữa các mức độ nghiêm trọng.

5.3. Lựa chọn mô hình và huấn luyện

Chúng tôi đã triển khai và đánh giá hiệu năng của nhiều thuật toán phổ biến trong học máy cho bài toán phân loại, bao gồm: Random Forest, SGDClassifier, LightGBM, XGBoost và CatBoost.

Ngoài ra, trong quá trình huấn luyện, chúng tôi cũng thực hiện kiểm chéo (cross-validation) với cv=3 để đánh giá độ ổn định và khả năng tổng quát hóa của mô hình trên các tập dữ liệu khác nhau.

5.4. Đánh giá mô hình

Kết quả đánh giá mô hình được thể hiện ở *Bảng 2*.

Bảng 2: Kết quả thực nghiệm mô hình phân loại

Model	Acc	Precision	Recall	F1	CV mean acc	CV std acc
Random Forest	93.43%	92.44%	93.43%	91.6%	93.41%	0.0003
SGDClassifier	85.38%	91.37%	85.38%	87.5%	89.36%	0.042
LightGBM	93.47%	92.73%	93.48%	92.39%	93.4%	0.0005
XGBoost	93.47%	93.1%	93.47%	92.24%	93.47%	0.0002
CatBoost	93.42%	92.3%	93.42%	91.71%	93.41%	0.0001

Dựa trên các kết quả đạt được:

- Các mô hình Random Forest, LightGBM, XGBoost và CatBoost có khả năng dự đoán chính xác không quá chênh lệch (lệch khoảng 0.04%-0.05%).
- SGDClassifier có hiệu năng thấp nhất, cho thấy thuật toán này không phù hợp để xử lý dữ liệu phức tạp với nhiều đặc trưng.
- LightGBM và XGBoost là hai mô hình có hiệu năng cao nhất, đạt Accuracy ở mức 93.47%, với độ ổn định cao trong kiểm định chéo (độ lệch chuẩn thấp).

6. KẾT LUẬN

Chúng tôi đã phân tích dữ liệu tai nạn giao thông tại Chicago (01/2022 - 10/2024) từ ba bảng chính: crashes, vehicles, và people, nhằm phân tích các xu hướng trong dữ liệu này và xây dựng mô hình dự đoán. Kết quả cho thấy, các thời điểm nguy cơ tai nạn cao gồm tháng 5 và khung giờ 6h-9h sáng, 15h-18h chiều. Hành vi nguy hiểm như không nhường đường, sử dụng điện thoại, trốn cảnh sát, và các dòng xe phổ biến như Toyota Camry là nguyên nhân chính. Tai nạn chủ yếu gây thiệt hại kinh tế lớn (> $1,500), với va chạm góc và người đi bộ dẫn đến thương tích nghiêm trọng. Các điểm nóng là trung tâm thành phố và vùng gần hồ Michigan. Mô hình LightGBM và XGBoost đạt hiệu năng dự đoán cao nhất, với độ chính xác trên 93.4%.

TÀI LIỆU THAM KHẢO

[1] Traffic Crashes – Crashes. Link: Traffic-Crashes-Crashes (06/10/2024).

[2] Traffic Crashes – People. Link: Traffic-Crashes-People (06/10/2024).

[3] Traffic Crashes – Vehicles. Link: Traffic-Crashes-Vehicles (06/10/2024).

PHỤ LỤC 1

Bảng 3: Mô tả các thuộc tính quan trọng trong bộ crashes

Tên cột	**Kiểu dữ liệu**	**Mô tả**	**Ví dụ**	**Phần trăm giá trị thiếu**
crash_date	Number	Ngày xảy ra tại nạn		0.0%
first_crash_type	Text	Loại tai nạn đầu tiên xảy ra	ANGLE, REAR END, PARKED MOTOR VEHICLE	0.0%
trafficway_type	Text	Loại đường giao thông	DIVIDED - W/MEDIAN BARRIER, DIVIDED	0.0%
road_defect	Text	Lỗi trên bề mặt đường	NO DEFECTS, UNKNOWN, DEBRIS ON ROADWAY	0.0%
damage	Text	Thiệt hại	OVER $1,500, $501 - $1,500, $500 OR LESS	0.0%
prim_contributory_cause	Text	Nguyên nhân chính gây tai nạn	UNABLE TO DETERMINE, FOLLOWING TOO CLOSELY	0.0%
most_severe_injury	Text	Mức độ thương tích nghiêm trọng nhất	INCAPACITATING INJURY, NO INDICATION OF INJURY	0.25%
injuries_total	Number	Tổng số người bị thương trong vụ tai nạn	3.0, 0.0, 1.0	0.25%
injuries_fatal	Number	Tổng số người tử vong trong vụ tai nạn	0.0, 1.0, 2.0	0.25%
latitude	Number	Vĩ độ xảy ra tai nạn	41.854120263, …	0.95%
longitude	Number	Kinh độ xảy ra tai nạn	-87.665902343, …	0.95%

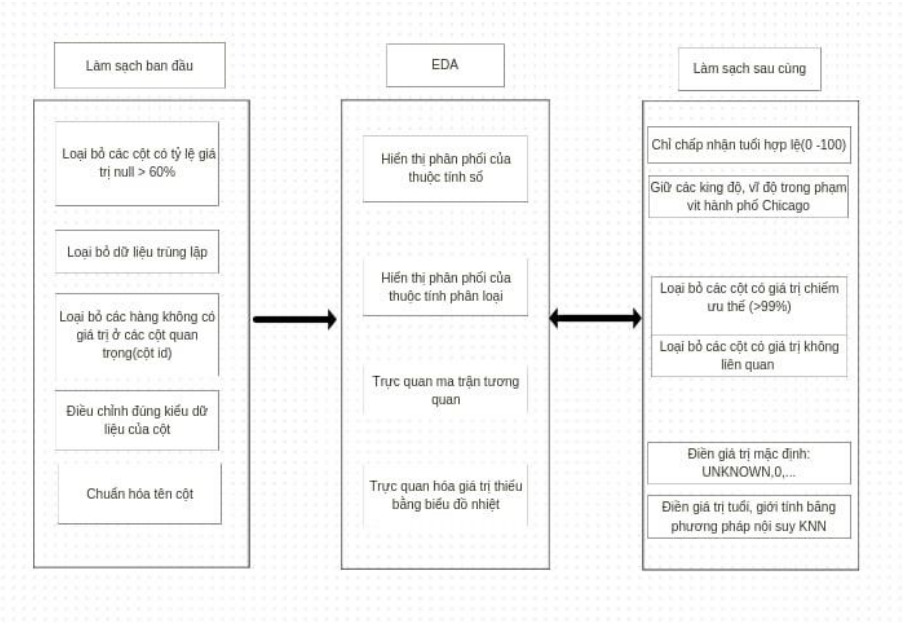
Bảng 4: Mô tả các thuộc tính quan trọng trong bộ people

Tên cột	Kiểu dữ liệu	Mô tả	Ví dụ	Phần trăm giá trị thiếu
person_id	Text	Mã định danh mỗi người		0.0%
crash_record_id	Text	Mã định danh cho mỗi vụ tai nạn		0.0%
vehicle_id	Number	Mã số phương tiện		2.18%
sex	Text	Giới tính	M, X, F	2.0%
age	Number	Tuổi	28.0, 27.0, 29.0	29.73%
injury_classification	Text	Phân loại mức độ thương tích	NO INDICATION OF INJURY, NONINCAPACITATING INJURY	0.02%
physical_condition	Text	Tình trạng thể chất	NORMAL, IMPAIRED - DRUGS	20.04%
person_id	Text	Mã định danh mỗi người		0.0%

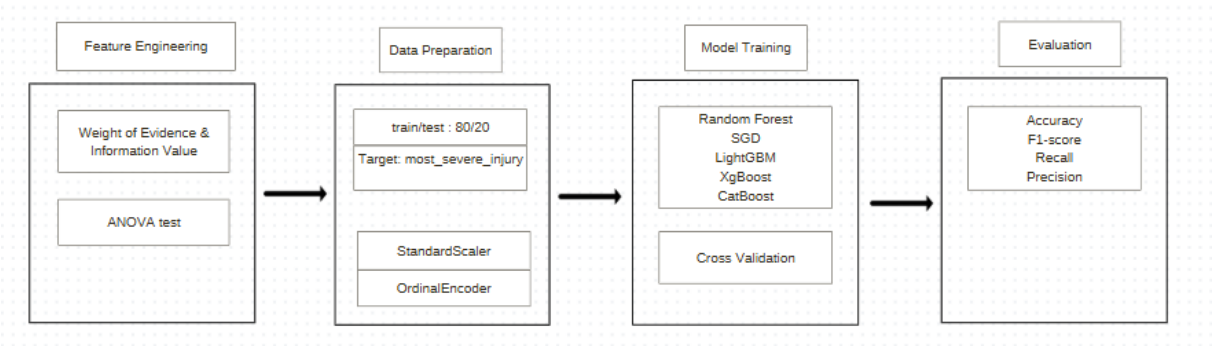
Bảng 5: Mô tả các thuộc tính quan trọng trong bộ vehicles

Tên cột	Kiểu dữ liệu	Mô tả	Ví dụ	Phần trăm giá trị thiếu
crash_unit_id	Number	Mã định danh cho mỗi đơn vị tham gia tai nạn		0.0%
crash_record_id	Text	Mã định danh cho mỗi vụ tai nạn		0.0%
crash_date	Number	Ngày xảy ra tại nạn		0.0%
make	Text	Hãng xe	NISSAN, CHRYSLER, SUBARU	2.42%
model	Text	Mẫu xe	SENTRA, SEBRING, OUTBACK	2.42%
occupant_cnt	Number	Số người trên phương tiện	1, 2, 3	2.42%

PHỤ LỤC 2

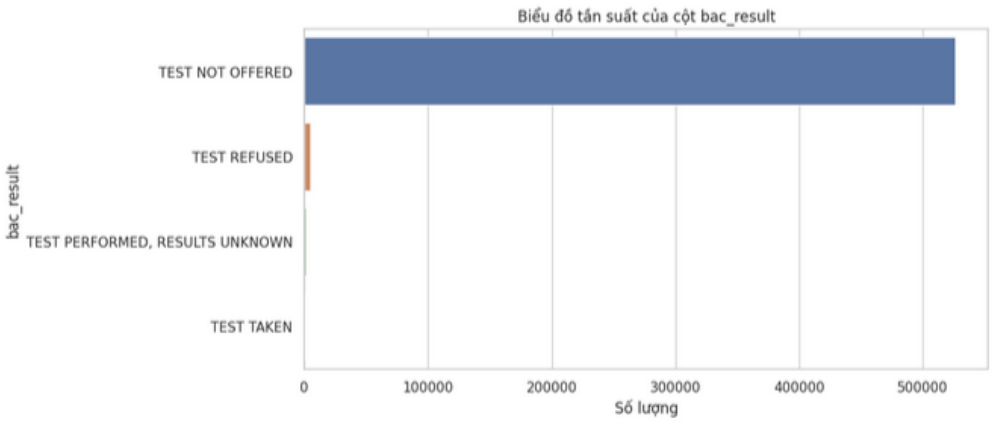


Hình 7: Quy trình làm sạch dữ liệu

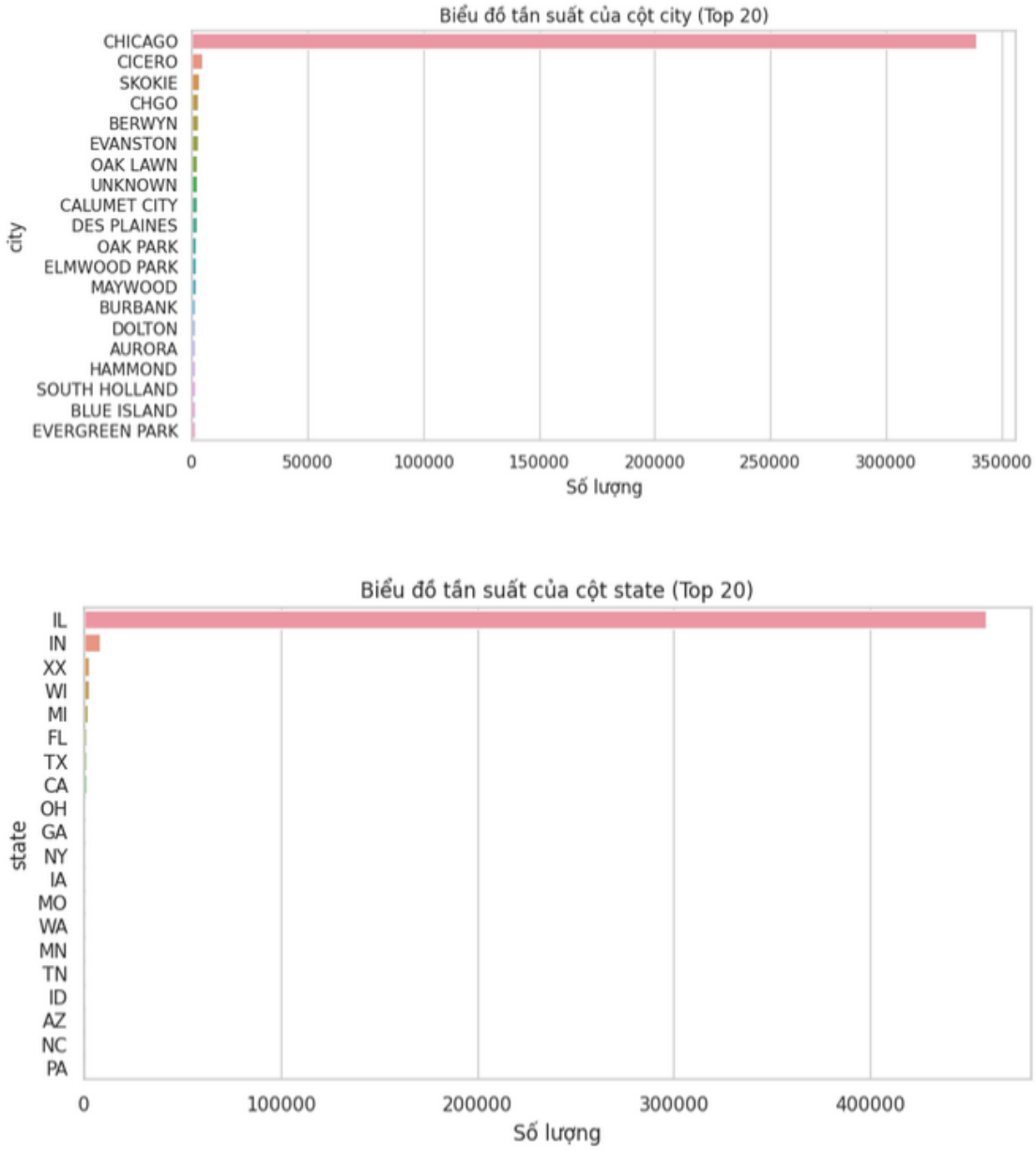


Hình 8: Quy trình xây dựng mô hình

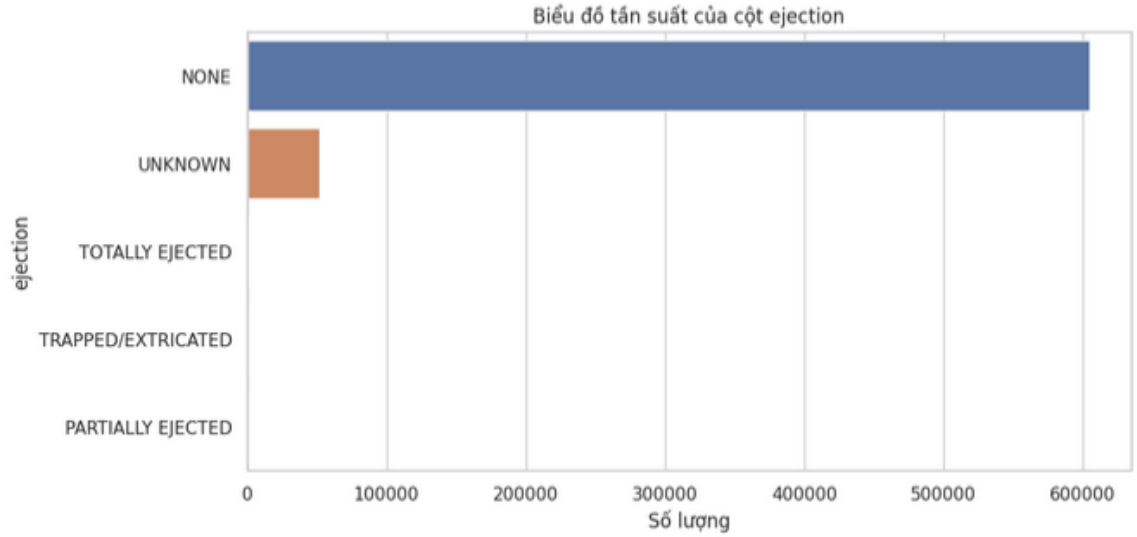
PHỤ LỤC 3



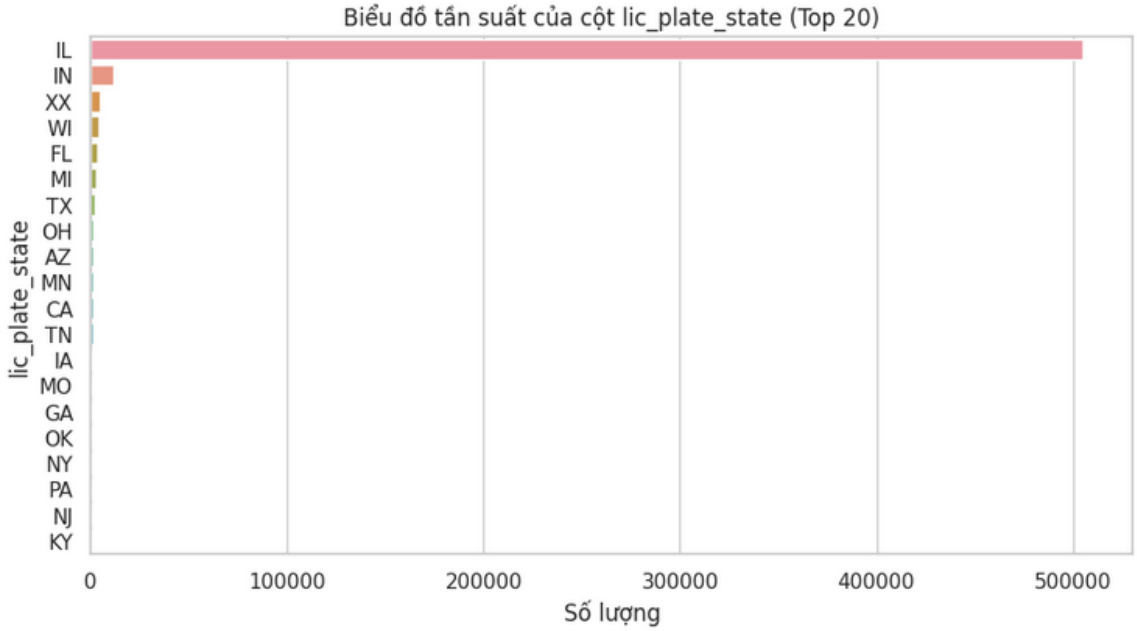
Hình 9: Biểu đồ cột thể hiện tần suất của cột bac_result.



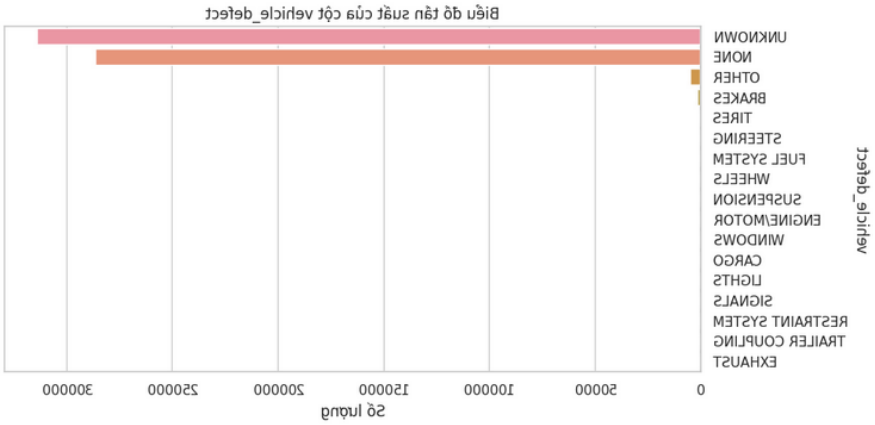
Hình 10: Biểu đồ cột thể hiện phân bố giá trị của cột city và state.



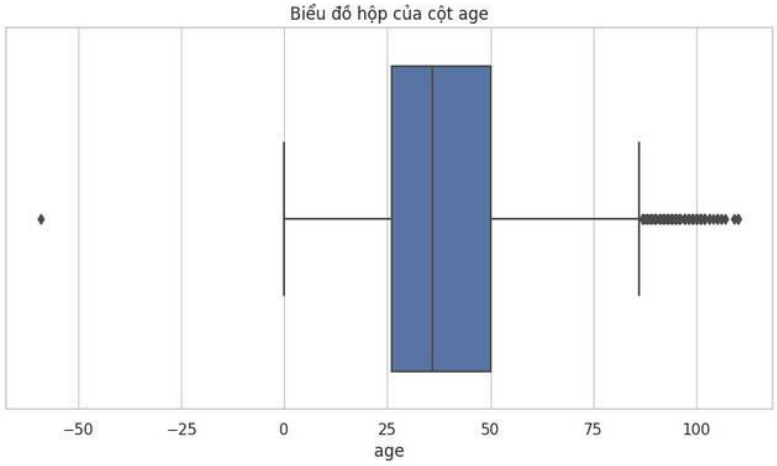
Hình 11: Biểu đồ cột thể hiện phân bố giá trị của cột ejection.



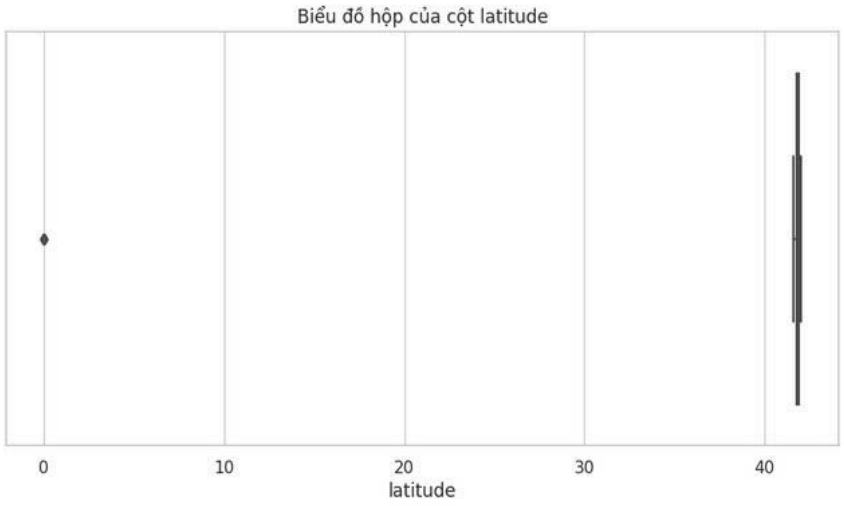
Hình 12: Biểu đồ cột thể hiện phân bố giá trị của cột place_state (top 20).



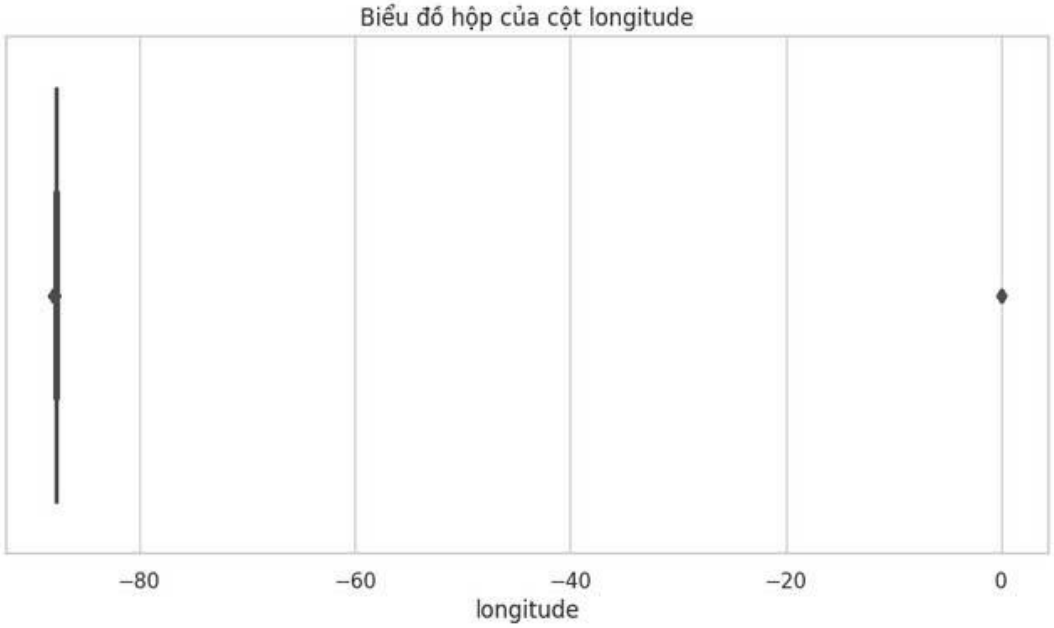
Hình 13: Biểu đồ cột thể hiện phân bố giá trị của cột vehicle_defect.



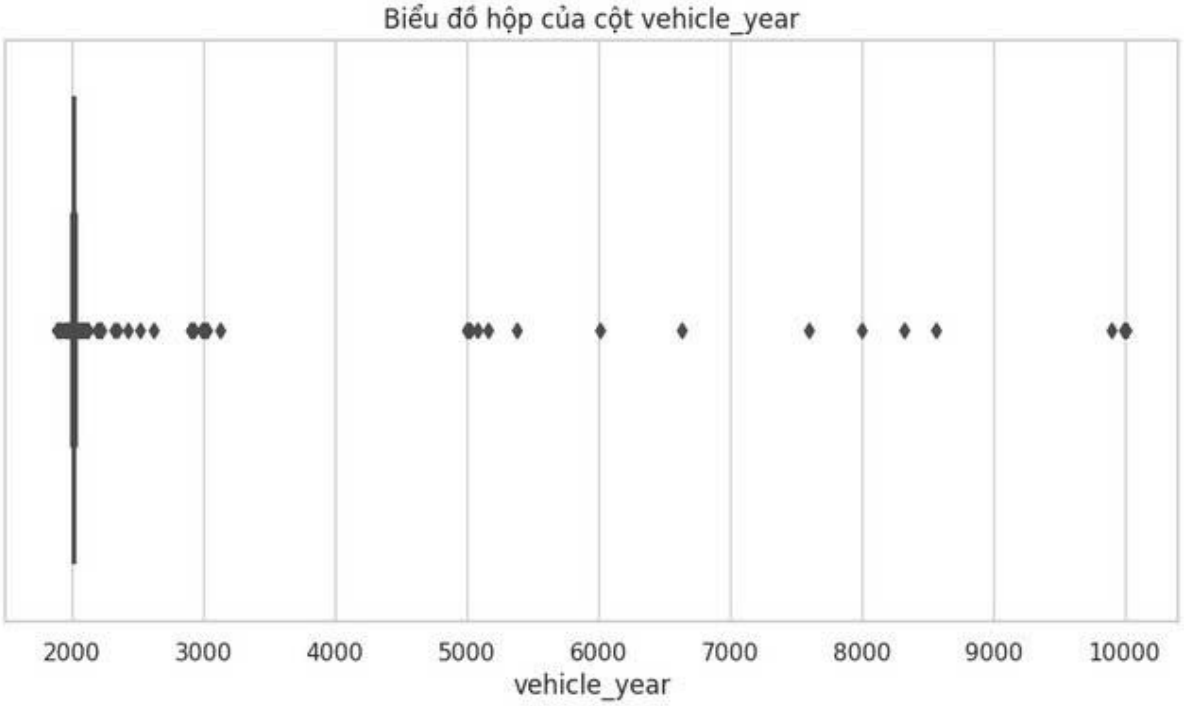
Hình 14: Biểu đồ hộp của cột age.



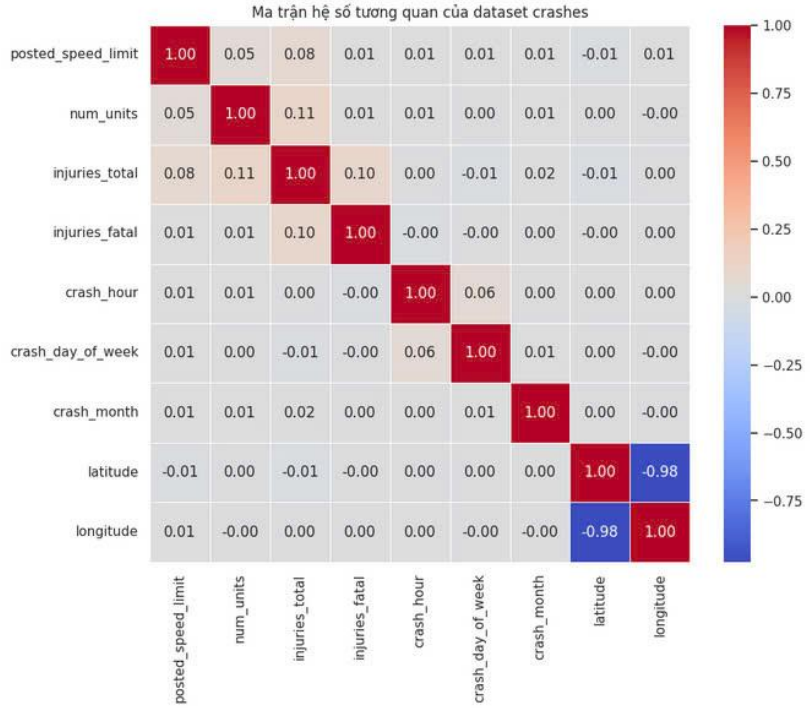
Hình 15: Biểu đồ hộp của cột latitude.



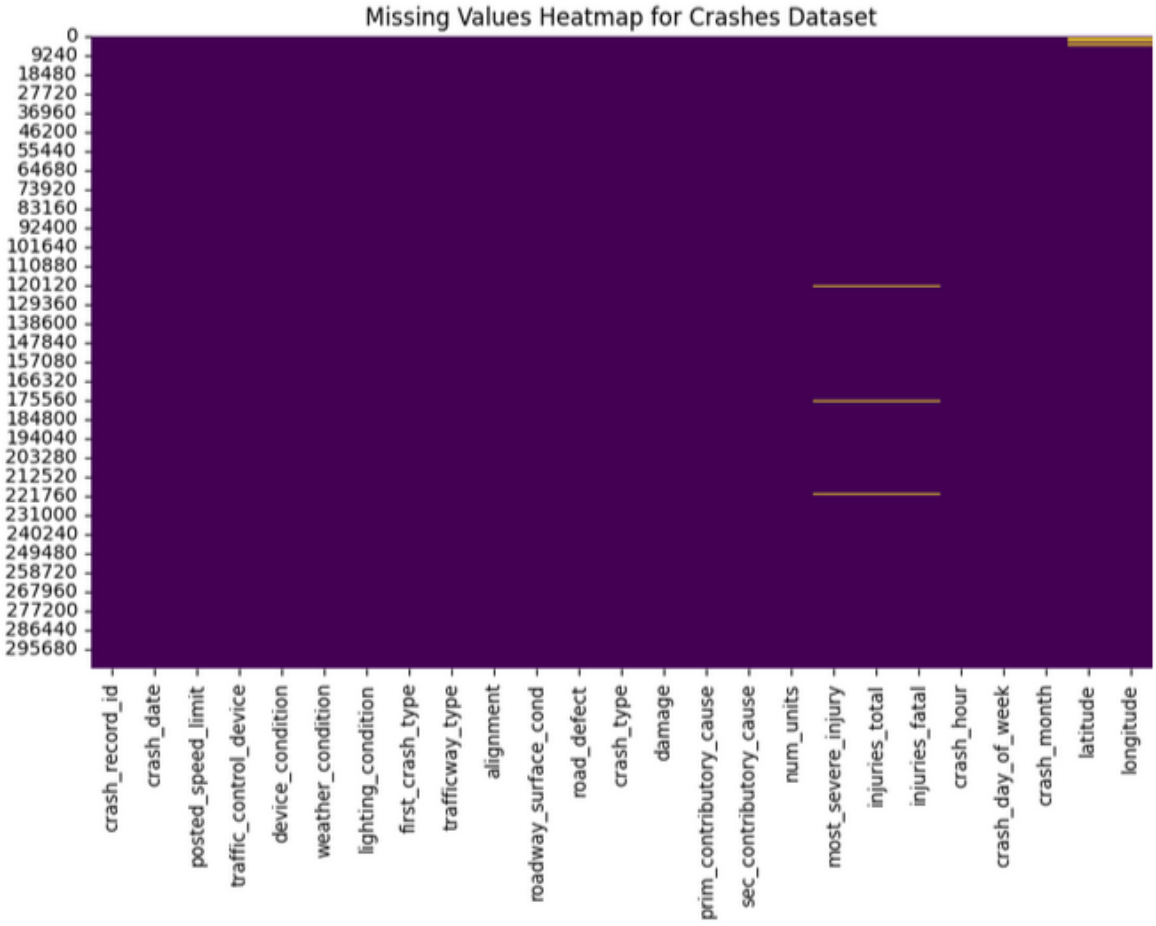
Hình 16: Biểu đồ hộp của cột long_latitude.



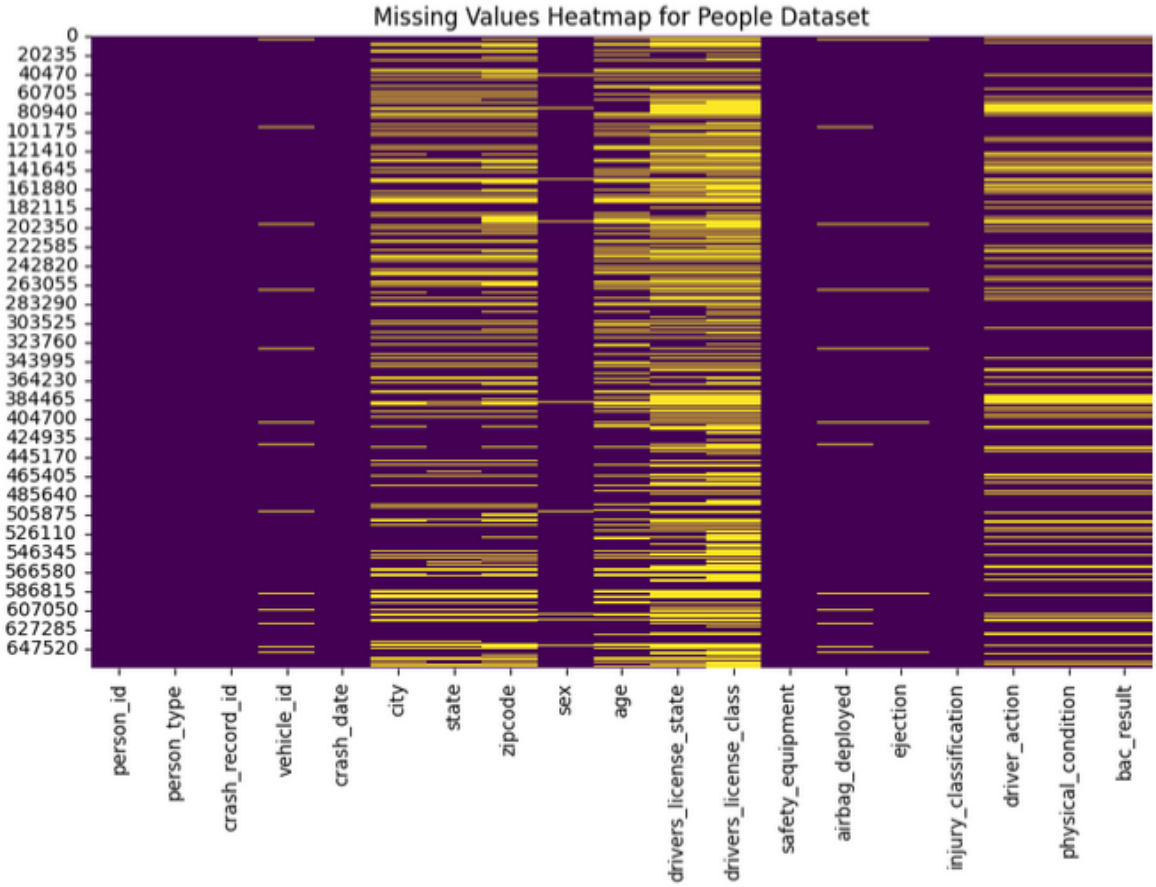
Hình 17: Biểu đồ hộp của cột vehicle_year.



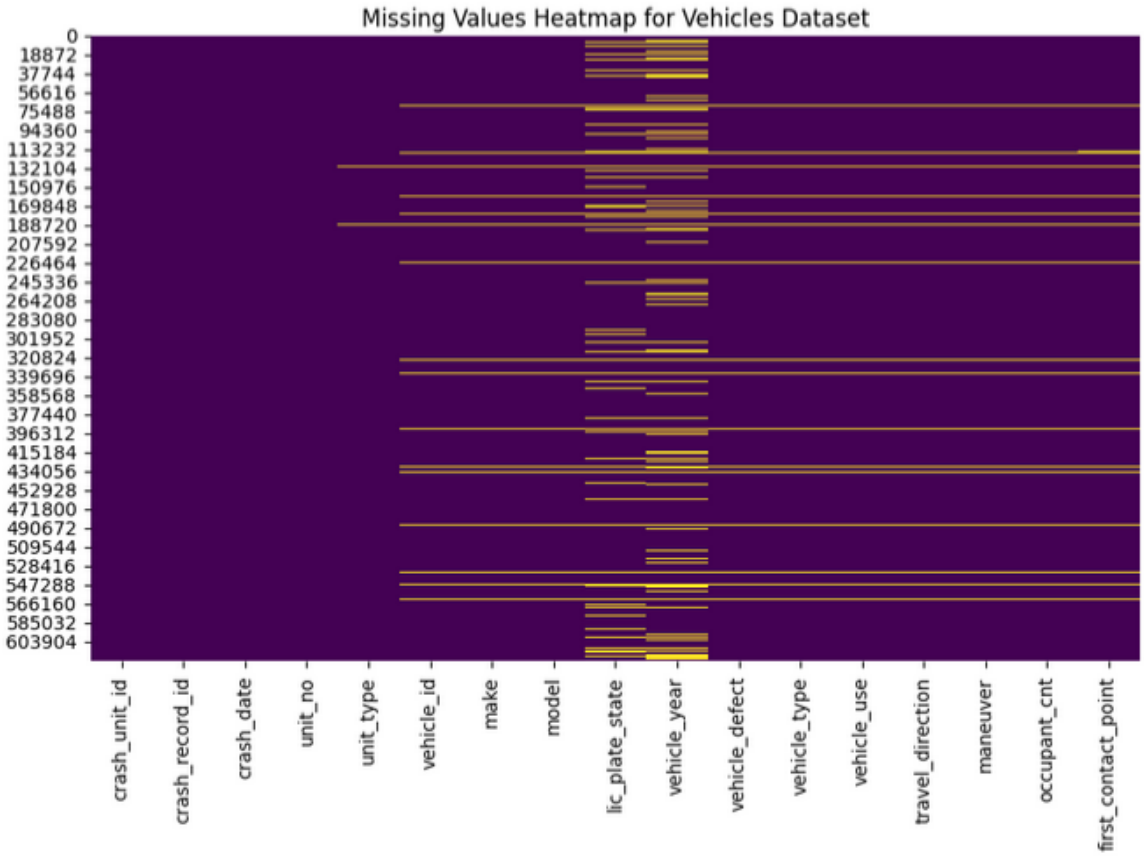
Hình 18: Ma trận hệ số tương quan của crashes.



Hình 19: Missing Values Heatmap của crashes.



Hình 20: Missing Values Heatmap của people.



Hình 21: Missing Values Heatmap của vehicles.

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Lê Xuân Bình	− Thu thập dữ liệu − Xây dựng mô hình phân loại − Viết báo cáo − Chuẩn bị slide thuyết trình
2	Trần Đại Hiển	− Thu thập dữ liệu − Làm sạch dữ liệu − Mô tả bộ dữ liệu − Phân tích tổng quan
3	Phạm Ngọc Trí	− Thu thập dữ liệu − Đề xuất phương pháp phân tích − Phân tích chuyên sâu − Chuẩn bị slide thuyết trình