UIT
Trường Đại học
Công nghệ Thông tin

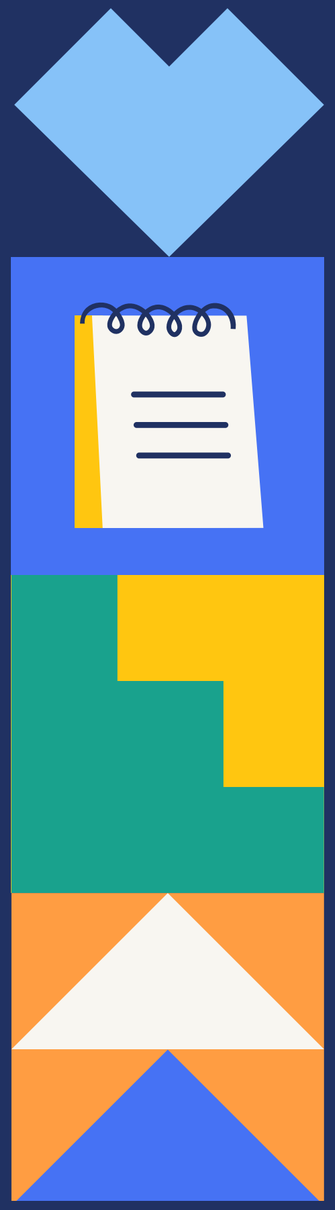
Khoa Khoa học
và Kỹ thuật Thông tin

Phân tích tai nạn giao thông tại thành phố Chicago, Mỹ từ tháng 1.2022 đến tháng 10.2024

Nhóm 03:

1. Lê Xuân Bình - 22520131
2. Trần Đại Hiển - 22520426
3. Phạm Ngọc Trí - 22521526

Tháng 12/2024



Nội dung

Giới thiệu

- Tổng quan đề tài
- Cam kết minh bạch

Phân tích chuyên sâu

- Xu hướng theo thời gian
- Loại phương tiện
- Mức độ thiệt hại và thương tích

Mô tả bộ dữ liệu

- Quá trình làm sạch dữ liệu
- Thăm dò các biến quan trọng

Mô hình dự đoán

- Dự đoán mức độ nghiêm trọng của tai nạn

1. Giới thiệu

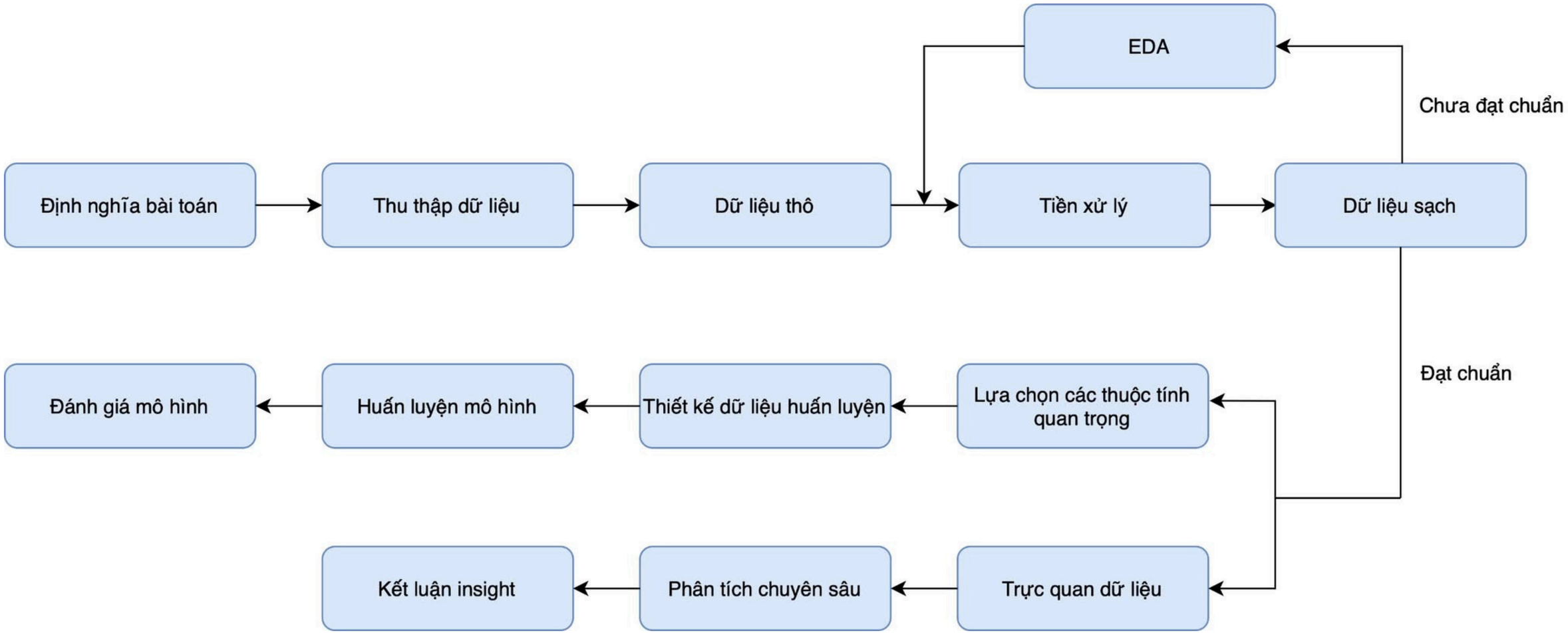
Mục tiêu: Thu thập và phân tích dữ liệu tai nạn giao thông từ nguồn mở của thành phố **Chicago** để đánh giá các **xu hướng chính** và xây dựng **mô hình dự đoán** mức độ nghiêm trọng của tai nạn**.**

Kết quả:

- Phân tích về các khía cạnh tai nạn: xu hướng thời gian, mức độ thiệt hại, loại phương tiện, địa lý và nhân khẩu học.
- Xây dựng mô hình phân loại mức độ nghiêm trọng của tai nạn.

Cam kết:
Nhóm tự thu thập và xử lý dữ liệu, cung cấp đầy đủ mã nguồn và tài liệu minh chứng, đảm bảo tính trung thực và minh bạch trong quá trình thực hiện.

PHƯƠNG PHÁP PHÂN TÍCH



Làm sạch dữ liệu

Initial Cleaning

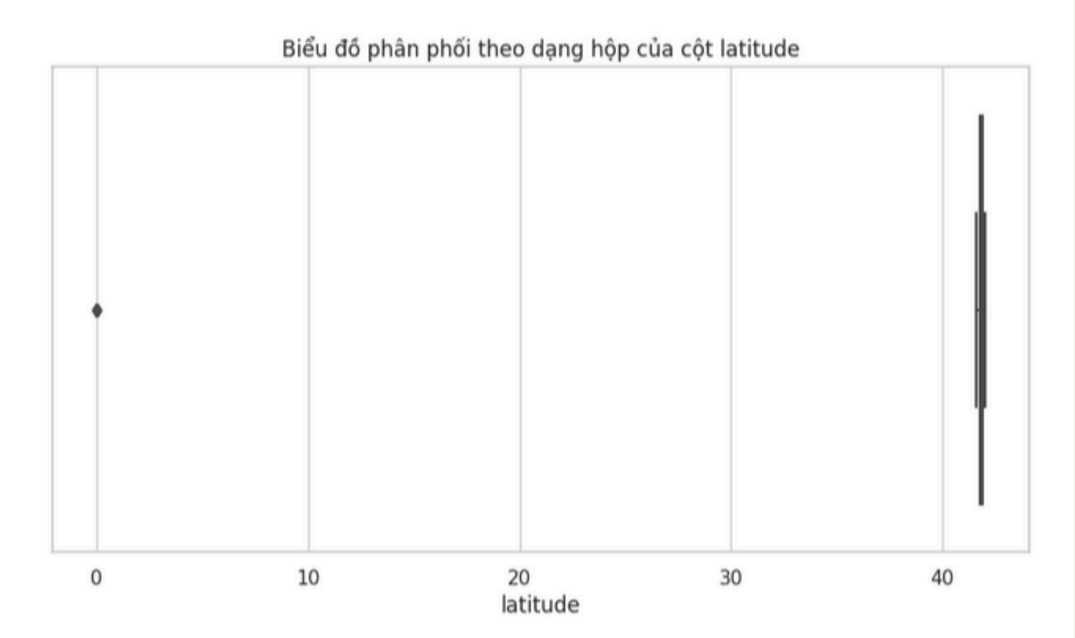
- Drop columns whose null percent > 60%
- Remove duplicates
- Remove rows without value in critical columns
- Correct column datatype
- Standadize column name

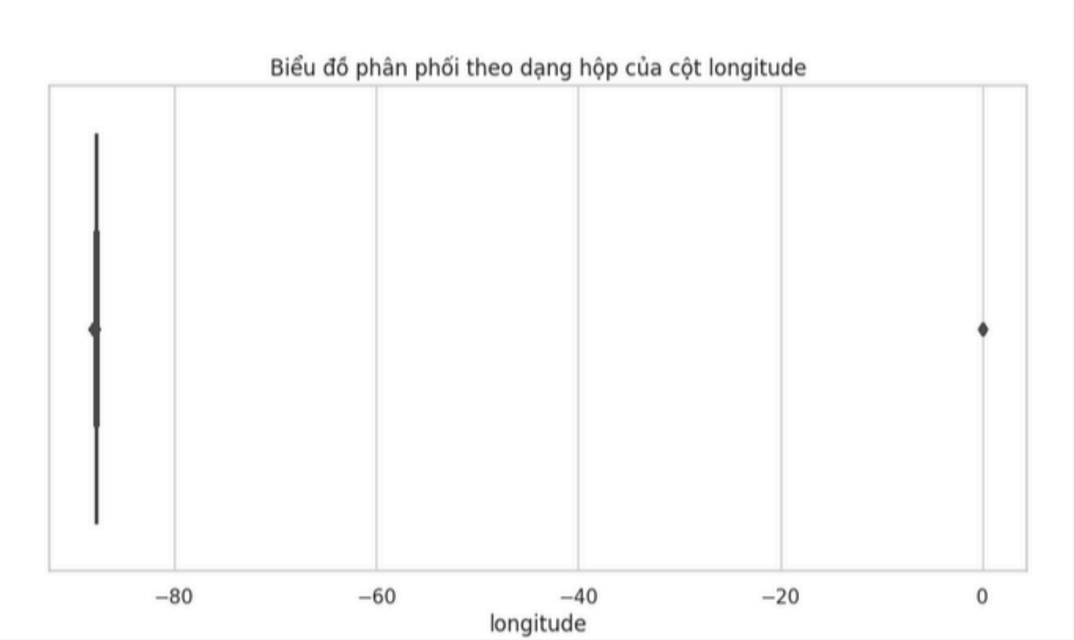
EDA

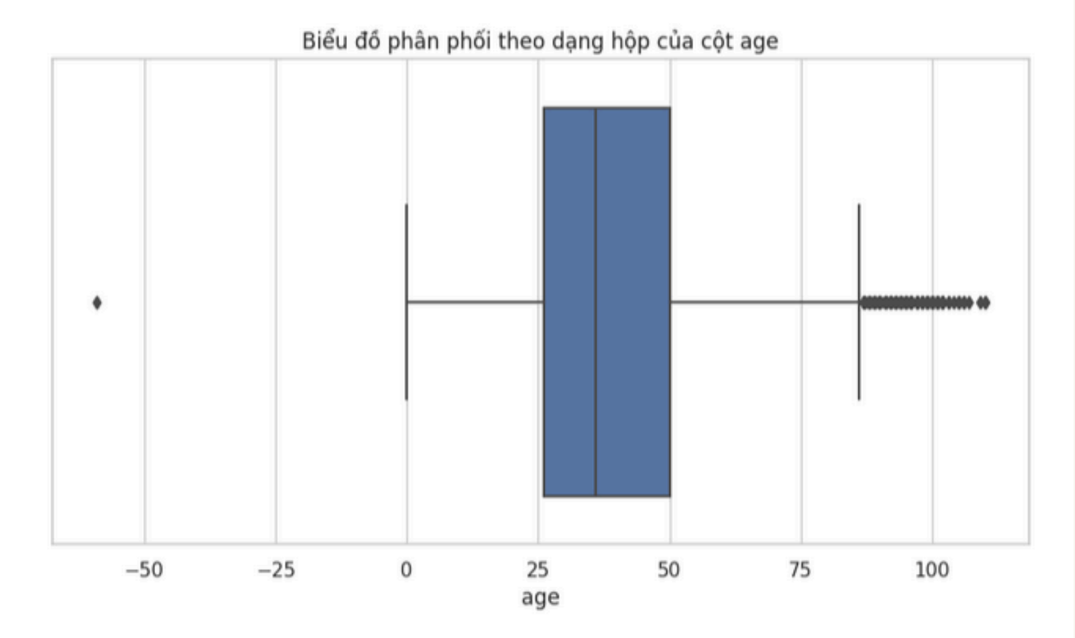
- Display Numerical Attribute Distribution
- Display Categorical Attribute Distribution
- Correlation Matrix
- Visualize Missing Values with Heatmap

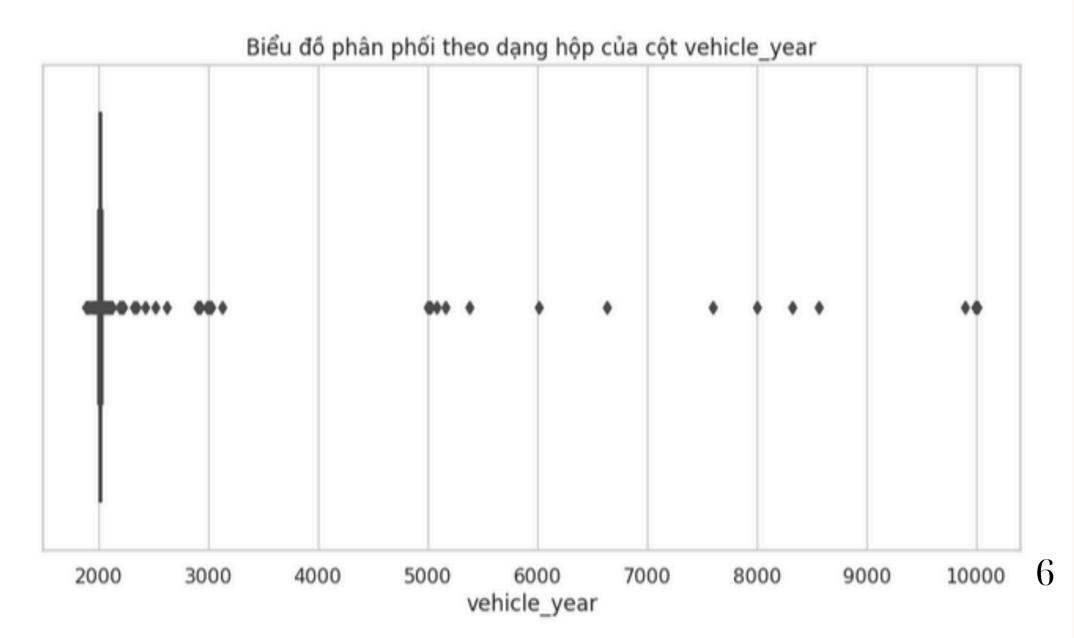
FInal Cleaning

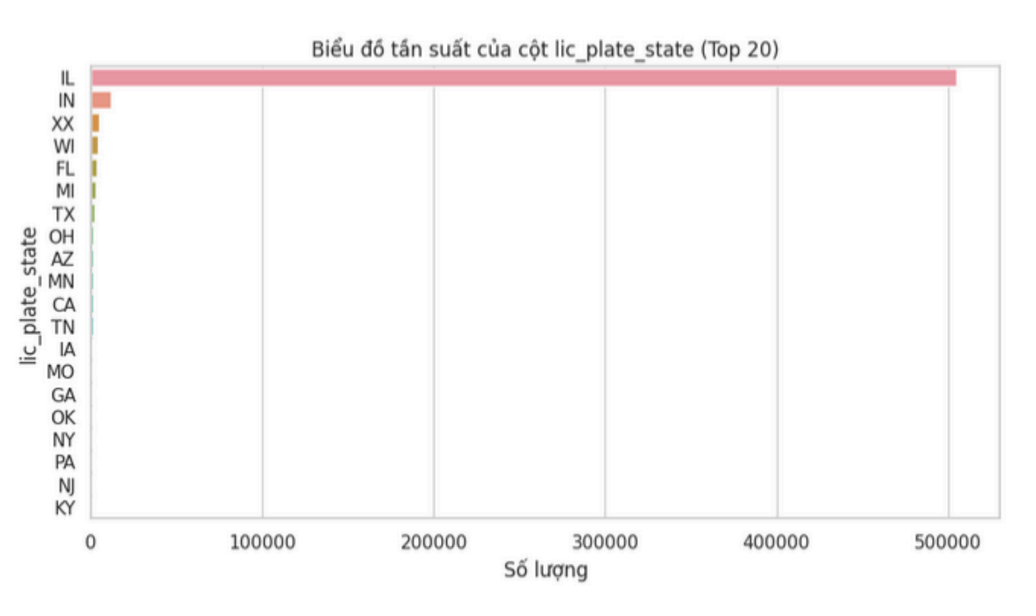
- windownize age: [0,100]
- windownize longtitude,latitude
- Drop columns with >99% dominant values.
- Remove columns with irrelevant values.
- Fill default values: UNKNOWN’,0,..
- Fill age,sex: KNN impute

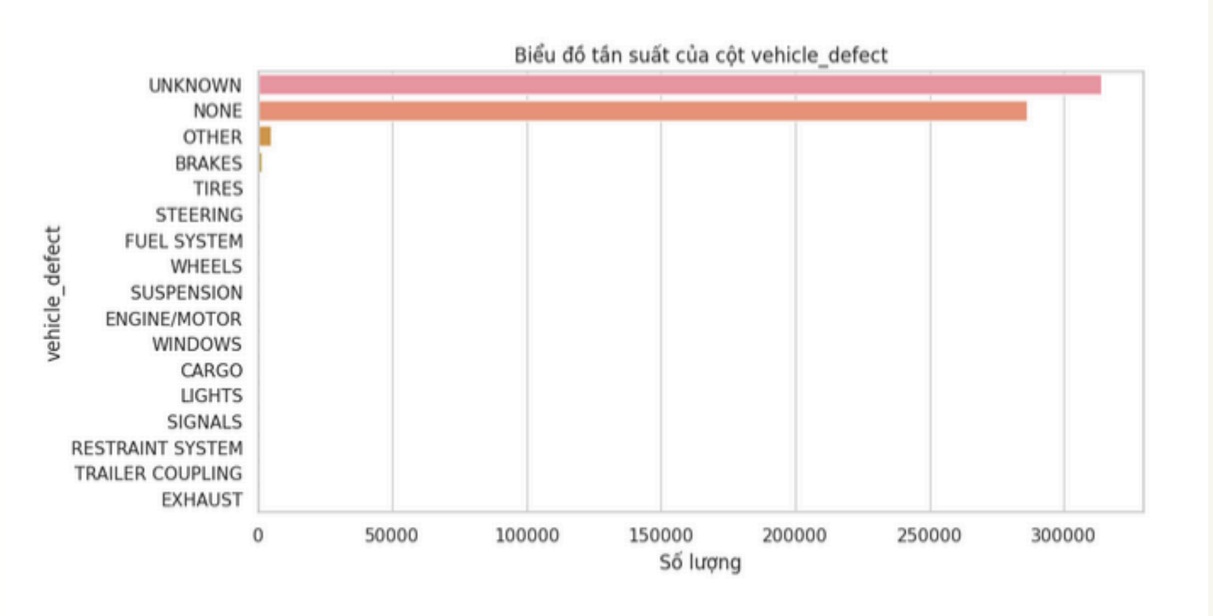

Biểu đồ phân phối theo dạng hộp của cột latitude
0
10
20
30
40
latitude

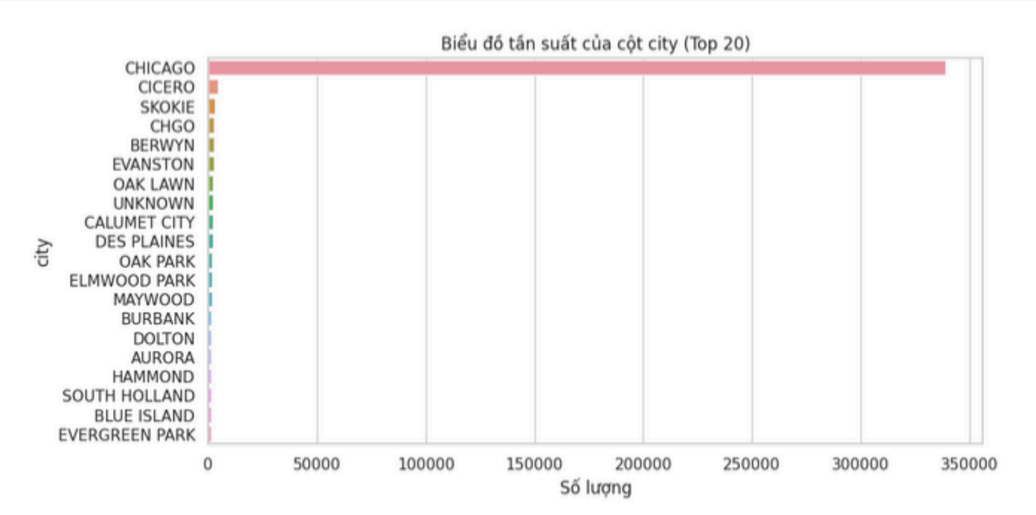

Biểu đồ phân phối theo dạng hộp của cột longitude
−80
−60
−40
−20
0
longitude

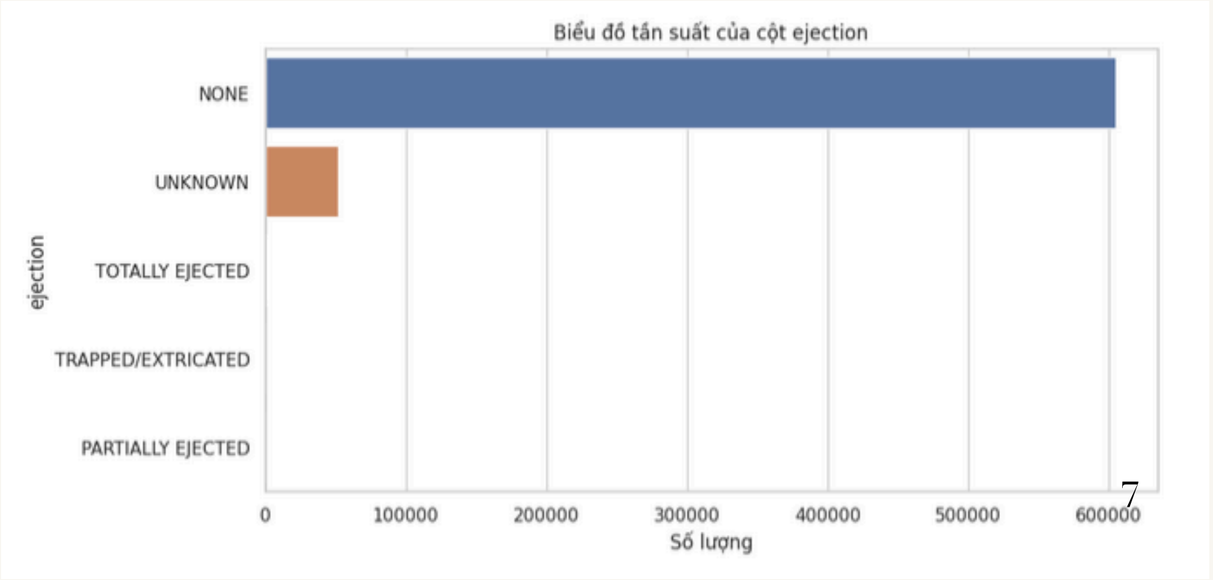

Biểu đồ phân phối theo dạng hộp của cột age
−50
−25
0
25
50
75
100
age

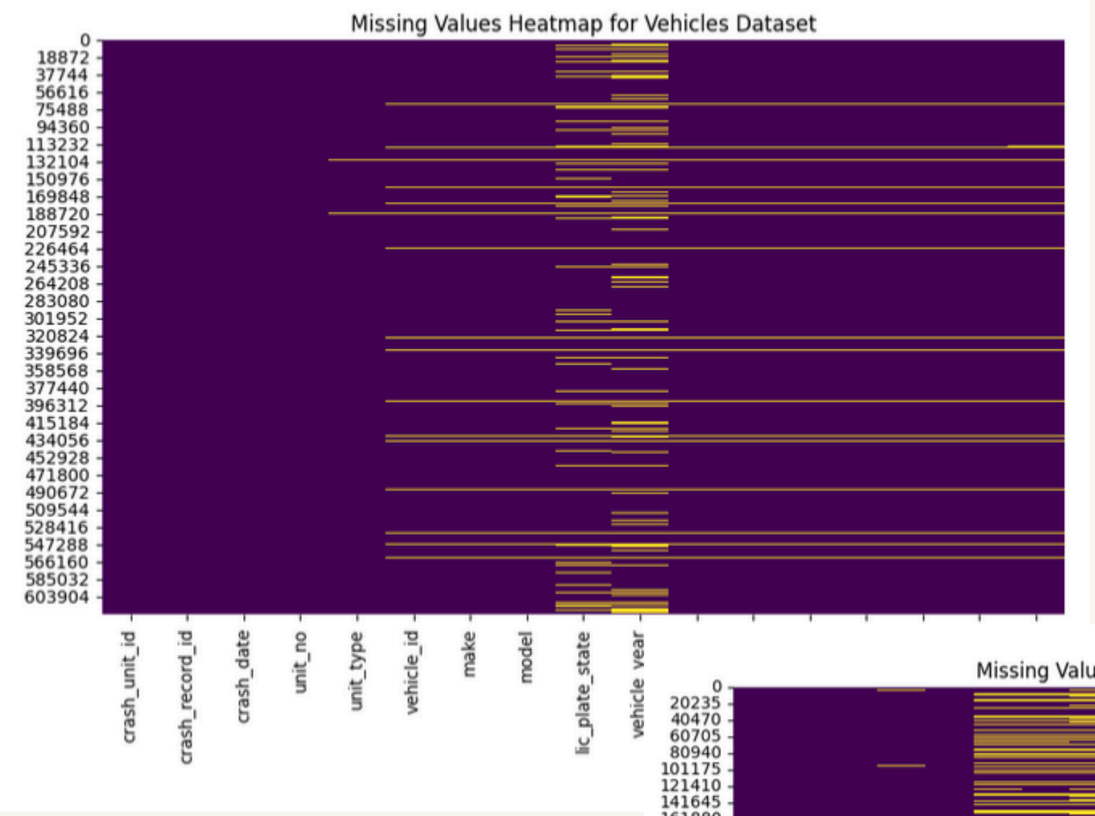

Biểu đồ phân phối theo dạng hộp của cột vehicle_year
2000
3000
4000
5000
6000
7000
8000
9000
10000
vehicle_year

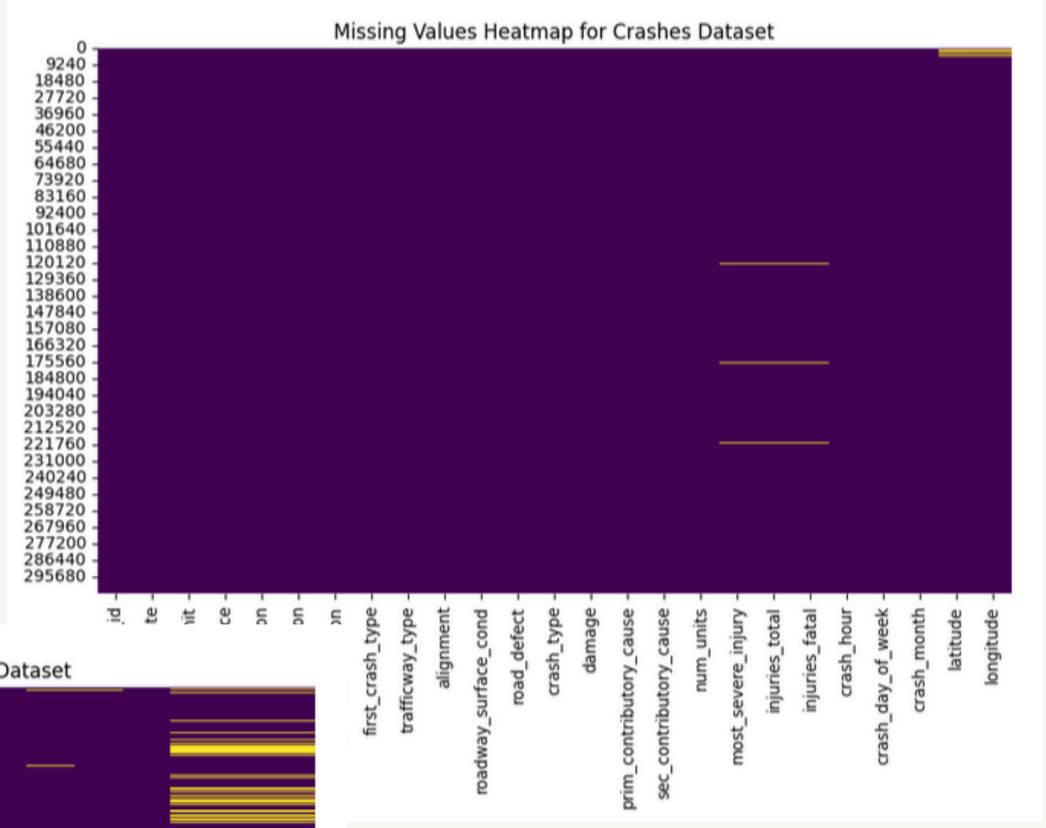

Biểu đồ tần suất của cột lic_plate_state (Top 20)
lic_plate_state
IL
IN
XX
WI
FL
MI
TX
OH
AZ
MN
CA
TN
IA
MO
GA
OK
NY
PA
NJ
KY
0
100000
200000
300000
400000
500000
Số lượng

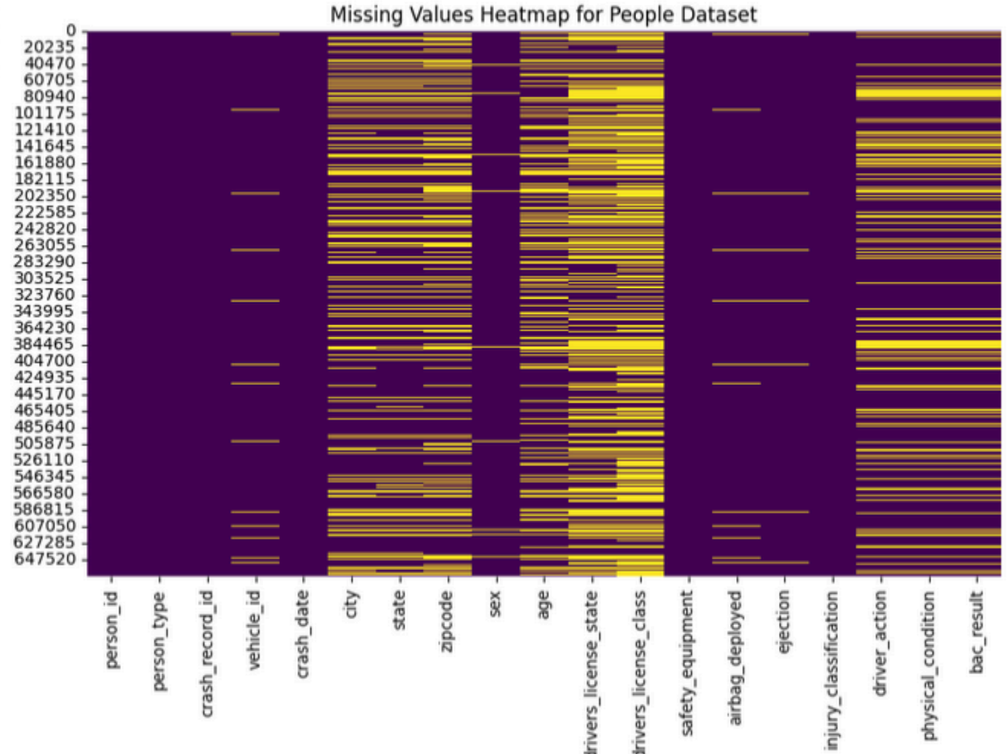

Biểu đồ tần suất của cột vehicle_defect
vehicle_defect
UNKNOWN
NONE
OTHER
BRAKES
TIRES
STEERING
FUEL SYSTEM
WHEELS
SUSPENSION
ENGINE/MOTOR
WINDOWS
CARGO
LIGHTS
SIGNALS
RESTRAINT SYSTEM
TRAILER COUPLING
EXHAUST
0
50000
100000
150000
200000
250000
300000
Số lượng


Biểu đồ tần suất của cột city (Top 20)
city
CHICAGO
CICERO
SKOKIE
CHGO
BERWYN
EVANSTON
OAK LAWN
UNKNOWN
CALUMET CITY
DES PLAINES
OAK PARK
ELMWOOD PARK
MAYWOOD
BURBANK
DOLTON
AURORA
HAMMOND
SOUTH HOLLAND
BLUE ISLAND
EVERGREEN PARK
0
50000
100000
150000
200000
250000
300000
350000
Số lượng


Biểu đồ tần suất của cột ejection
ejection
NONE
UNKNOWN
TOTALLY EJECTED
TRAPPED/EXTRICATED
PARTIALLY EJECTED
0
100000
200000
300000
400000
500000
600000
Số lượng


Missing Values Heatmap for Vehicles Dataset
0
18872
37744
56616
75488
94360
113232
132104
150976
169848
188720
207592
226464
245336
264208
283080
301952
320824
339696
358568
377440
396312
415184
434056
452928
471800
490672
509544
528416
547288
566160
585032
603904
crash_unit_id
crash_record_id
crash_date
unit_no
unit_type
vehicle_id
make
model
lic_plate_state
vehicle_year


Missing Values Heatmap for Crashes Dataset
0
9240
18480
27720
36960
46200
55440
64680
73920
83160
92400
101640
110880
120120
129360
138600
147840
157080
166320
175560
184800
194040
203280
212520
221760
231000
240240
249480
258720
267960
277200
286440
295680
first_crash_type
trafficway_type
alignment
roadway_surface_cond
road_defect
crash_type
damage
prim_contributory_cause
sec_contributory_cause
num_units
most_severe_injury
injuries_total
injuries_fatal
crash_hour
crash_day_of_week
crash_month
latitude
longitude


Missing Values Heatmap for People Dataset
0
20235
40470
60705
80940
101175
121410
141645
161880
182115
202350
222585
242820
263055
283290
303525
323760
343995
364230
384465
404700
424935
445170
465405
485640
505875
526110
546345
566580
586815
607050
627285
647520
person_id
person_type
crash_record_id
vehicle_id
crash_date
city
state
zipcode
sex
age
safety_equipment
airbag_deployed
ejection
injury_classification
driver_action
physical_condition
bac_result

2. Mô tả dữ liệu

crashes

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

Tên cột	Kiểu dữ liệu	Mô tả
crash_date	Number	Ngày xảy ra tại nạn
first_crash_type	Text	Loại tai nạn đầu tiên xảy ra
trafficway_type	Text	Loại đường giao thông
road_defect	Text	Lỗi trên bề mặt đường
damage	Text	Thiệt hại
prim_contributory_cause	Text	Nguyên nhân chính gây tai nạn
most_severe_injury	Text	Mức độ thương tích nghiêm trọng nhất
injuries_total	Number	Tổng số người bị thương trong vụ tai nạn
injuries_fatal	Number	Tổng số người tử vong trong vụ tai nạn
latitude	Number	Vĩ độ xảy ra tai nạn
longitude	Number	Kinh độ xảy ra tai nạn

2. Mô tả dữ liệu

vehicles

Tên cột	Kiểu dữ liệu	Mô tả
crash_unit_id	Number	Mã định danh cho mỗi đơn vị tham gia tai nạn
crash_record_id	Text	Mã định danh cho mỗi vụ tai nạn
crash_date	Number	Ngày xảy ra tại nạn
make	Text	Hãng xe
model	Text	Mẫu xe
occupant_cnt	Number	Số người trên phương tiện

people

Tên cột	Kiểu dữ liệu	Mô tả
person_id	Text	Mã định danh mỗi người
crash_record_id	Text	Mã định danh cho mỗi vụ tai nạn
vehicle_id	Number	Mã số phương tiện
sex	Text	Giới tính
age	Number	Tuổi
injury_classification	Text	Phân loại mức độ thương tích
physical_condition	Text	Tình trạng thể chất

3.1 Phân tich theo xu hướng thời gian và phương tiện

THỜI GIAN

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

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, từ 20 giờ trở đi** (đặc biệt là **ngày cuối tuần**).

PHƯƠNG TIỆN

Tình trạng cảm xúc (EMOTIONAL), ảnh hưởng bởi rượu bia (IMPAIRED - ALCOHOL) và **mệt mỏi (FATIGUED/ASLEEP)** là 3 nguyên nhân phổ biến nhất do tình trạng thể chất

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.

3.2 Phân tich theo mức độ thiệt hại, thương tích và vị trí địa lý

THỜI GIAN

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).**

PHƯƠNG TIỆN

Điểm nóng 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.**

Mô hình phân loại

Model pipeline

Feature Engineering

- Weight of Evidence & Information Value
- ANOVA test

Data Preparation

- train/test : 80/20
- Target: most_severe_injury
- StandardScaler
- OrdinalEncoder

Model Training

- Random Forest
- SGD
- LightGBM
- XgBoost
- CatBoost
- Cross Validation

Evaluation

- Accuracy
- F1-score
- Recall
- Precision

Mô hình phân loại

Kết quả thực nghiệm

	Model	Accuracy	Precision	Recall	F1-score	CV Mean Accuracy	CV Std Accuracy
0	Random Forest	0.934309	0.924408	0.934309	0.915942	0.934143	0.000377
1	SGDClassifier	0.853851	0.913749	0.853851	0.875081	0.893604	0.042025
2	LightGBM	0.934756	0.927319	0.934756	0.923880	0.934081	0.000518
3	XGBoost	0.934706	0.930863	0.934706	0.922415	0.934706	0.000269
4	CatBoost	0.934226	0.923052	0.934226	0.917163	0.934156	0.000163

Kết luận

Kết quả đạt được

- Các **nhận xét** dựa trên các phân tích chuyên sâu về tai nạn giao thông ở thành phố Chicago, Mỹ.
- Xây dựng **mô hình phân loại** mức độ nghiêm trọng của tai nạn.

Hạn chế

- Chưa có **hệ thống quản lý dữ liệu** và phân tích tự động trên **thời gian thực**.

Thank you