PYTHON FOR DATA SCIENCE

BÁO CÁO ĐỒ ÁN CUỐI KỲ

Đề tài: Phân tích các vụ tai nạn giao thông ở Canada

Thành viên nhóm 1:

20280011 – Hoàng Hải Đăng

20280016 – Trần Tiến Đạt

20280083 – Lại Toàn Thắng

20280105 – Đào Minh Trí

Giảng viên hướng dẫn: Hà Văn Thảo

Ngành Khoa Học Dữ Liệu

Đại học Khoa học Tự nhiên – ĐHQG TP.HCM

Mục Lục

1 LỜI NÓI ĐẦU 3

2 TỔNG QUAN 5

2.1. Mô Tả Vấn Đề (Problem Overview) 5

2.2. Dataset sử dụng 5

3 PHÂN TÍCH CÁC VỤ TAI NẠN GIAO THÔNG 6

3.1. Đọc hiểu dataset 6

3.2. Xử lý dữ liệu 7

3.3. Phân tích dữ liệu 8

3.3.1. Thống kê các thông số gây ra tai nạn 8

3.3.2. Xu hướng các vụ tai nạn theo tháng và năm 10

3.3.3. Khung giờ thường xảy ra tai nạn trong ngày 12

3.3.4. Các ngày trong tuần thường xảy ra tai nạn 12

3.3.5. Thống kê độ tuổi thường gặp tai nạn 14

3.3.6. Mối quan hệ giữa độ tuổi, giới tính, và mức độ nghiêm trọng của các vụ tai nạn 15

4 CÁC TÀI LIỆU THAM KHẢO 17

1 LỜI NÓI ĐẦU

Tử vong và thương tích do tai nạn giao thông đường bộ đang ngày càng gia tăng trên thế giới. Ngày nay, số người tử vong vì giao thông đường bộ còn nhiều hơn số người tử vong do HIV/AIDS, lao hay tiêu chảy.

Hàng năm, có khoảng 1,25 triệu người tử vong vì tai nạn giao thông đường bộ trên toàn thế giới. Ngoài ra, có từ 20 đến 50 triệu người chịu thương tật, dẫn tới khuyết tật và gặp phải tình trạng kinh tế khó khăn vì thương tật do tai nạn giao thông đường bộ gây ra những thiệt hại tài chính đáng kể cho các nạn nhân, gia đình và quốc gia. Mỗi ngày, trên thế giới, gần 2.000 người thiệt mạng trong các vụ tai nạn đường bộ, trong số đó có 500 trẻ em. Trung bình cứ bốn phút có một trẻ tử vong trên đường do tai nạn giao thông. Hàng trăm trẻ bị thương, rất nhiều trẻ bị thương nghiêm trọng. 50% các ca tử vong rơi vào nhóm đối tượng dễ bị tổn thương như người đi bộ, người đi xe đạp và người đi xe máy.

Tại Việt Nam, theo ủy ban an toàn giao thông quốc gia, trong năm 2016, tai nạn giao thông đường bộ đã khiến gần 9.000 người chết và hàng chục nghìn người bị thương. Tai nạn giao thông đường bộ là nguyên nhân đứng thứ hai gây tử vong và thương tích nghiêm trọng ở trẻ em và thanh thiếu niên trong độ tuổi 0-19, chỉ đứng sau tai nạn đuối nước. Tai nạn giao thông đường bộ cũng gây ra 50% ca tử vong của thanh thiếu niên trong độ tuổi 15-19.

Tai nạn giao thông có thể ngăn chặn nếu chúng ta thận trọng hơn khi tham gia giao thông, việc đó có thể làm được nếu ta tìm hiểu kĩ càng nguyên nhân gây ra tai nạn.

2

TỔNG QUAN

2.1. Mô Tả Vấn Đề (Problem Overview)

An toàn giao thông phụ thuộc vào các yếu tố: Người tham gia giao thông, phương tiện giao thông, cơ sở hạ tầng và môi trường. Một trong những yếu tố này có sự bất bình thường đều có thể dẫn đến tai nạn giao thông hoặc mất an toàn giao thông. Do đời sống của nhân dân được nâng cao, nhu cầu đi lại gia tăng, việc sử dụng phương tiện (xe máy, ô tô con) ngày càng tăng cao, song sự bùng nổ và tiếp tục gia tăng nhu cầu tham gia giao thông ngày càng lớn, vượt năng lực đáp ứng của kết cấu hạ tầng giao thông, ý thức người tham gia giao thông còn hạn chế, thói quen, tập quán vùng, miền nên đã thường xuyên đã vi phạm an toàn giao thông (uống rượu, bia, không đội mũ bảo hiểm,…).

Để đánh giá rõ hơn các nguyên nhân dẫn đến mất an toàn giao thông, chúng ta sẽ tiến hành phân tích các yếu tố ảnh hưởng đến một vụ tai nạn giao thông.

2.2. Dataset sử dụng

https://open.canada.ca/data/en/dataset/8dd0ab9b-d45d-4526-9256-c598fbc4ff3a

Bộ dữ liệu trên bao gồm: Ngày/giờ xảy ra tai nạn, tình trạng mặt đường, tình trạng thời tiết, loại phương tiện và các thông tin khác về nạn nhân của 13839 vụ tai nạn giao thông xảy ra ở Canada. Ta sẽ sử dụng bộ dữ liệu này để phân tích các nguyên nhân gây ra tai nạn giao thông phổ biến.

3

PHÂN TÍCH CÁC VỤ TAI NẠN GIAO THÔNG

3.1. Đọc hiểu dataset

Dữ liệu được sử dụng chủ yếu từ 2 file:

afterPreprocess.csv: Chứa dữ liệu giao thông dạng bảng.

guidance.xlsx: Chứa thông tin về các trường dữ liệu trong **afterPreprocess.csv**.

Tiến hành đọc dữ liệu từ dataset.

```
In [2]: df = pd.read_csv("./afterPreprocess.csv")
        df.head()
```

In ra 5 dòng đầu tiên của dataset.

	Unnamed: 0	Accident Date	Time (24hr)	Road Surface	Weather Conditions	Type of Vehicle	Casualty Class	Casualty Severity	Sex of Casualty	Age of Casualty
0	0	2014-01-01	14:15:00	Dry	Fine without high winds	Car	Passenger	Slight	Male	28
1	1	2014-01-01	00:05:00	Dry	Fine without high winds	Car	Passenger	Slight	Male	29
2	2	2014-01-01	02:20:00	Dry	Fine without high winds	Car	Driver	Slight	Female	21
3	3	2014-01-01	01:30:00	Wet/Damp	Fine without high winds	Car	Pedestrian	Serious	Female	34
4	4	2014-01-01	14:15:00	Dry	Fine without high winds	Car	Driver	Slight	Male	34

Theo sự quan sát sơ bộ về bộ dữ liệu. Nhận thấy bộ dữ liệu có 13839 dòng và 10 cột, mỗi dòng trong Dataframe cho biết về thông tin của một vụ tai nạn. Sau đây là ý nghĩa của từng cột trong Dataframe:

- **Accident Date**: Ngày xảy ra tai nạn

- **Time (24hr)**: Giờ xảy ra tai nạn
- **Road Surface**: Tình trạng mặt đường
- **Weather Conditions**: Điều kiện thời tiết lúc đấy
- **Type of Vehicle**: Loại phương tiện gây tai nạn
- **Casualty Class**: Đối tượng có trong vụ tai nạn
- **Casualty Severity**: Mức độ nghiêm trọng của vụ tai nạn
- **Sex of Casualty**: Giới tính đối tượng
- **Age of Casualty**: Tuổi của đối tượng

3.2. Xử lý dữ liệu

Ta thấy các cột đa phần đều ở dạng object, điều này là không nên bởi vì nếu cột nào cũng ở dạng object thì rất khó để làm việc.

```
Unnamed: 0             int64
Accident Date         object
Time (24hr)           object
Road Surface          object
Weather Conditions    object
Type of Vehicle       object
Casualty Class        object
Casualty Severity     object
Sex of Casualty       object
Age of Casualty        int64
dtype: object
```

Ta nhận thấy rằng có 3 cột có thể chuyển sang dạng khác đó là cột Accident Date (datetime), Time (24hr) (datetime). Chúng mình sẽ tiến hành chuyển dạng của các cột này về định dạng datetime.

In [7]:
```
dt = []
list_year = []
list_month = []
for i in range(len(df.index)):
      dt.append(datetime.strptime(df["Accident Date"][i], '%Y-%m-%d'))
      list_month.append(dt[i].month)
      list_year.append(dt[i].year)
```

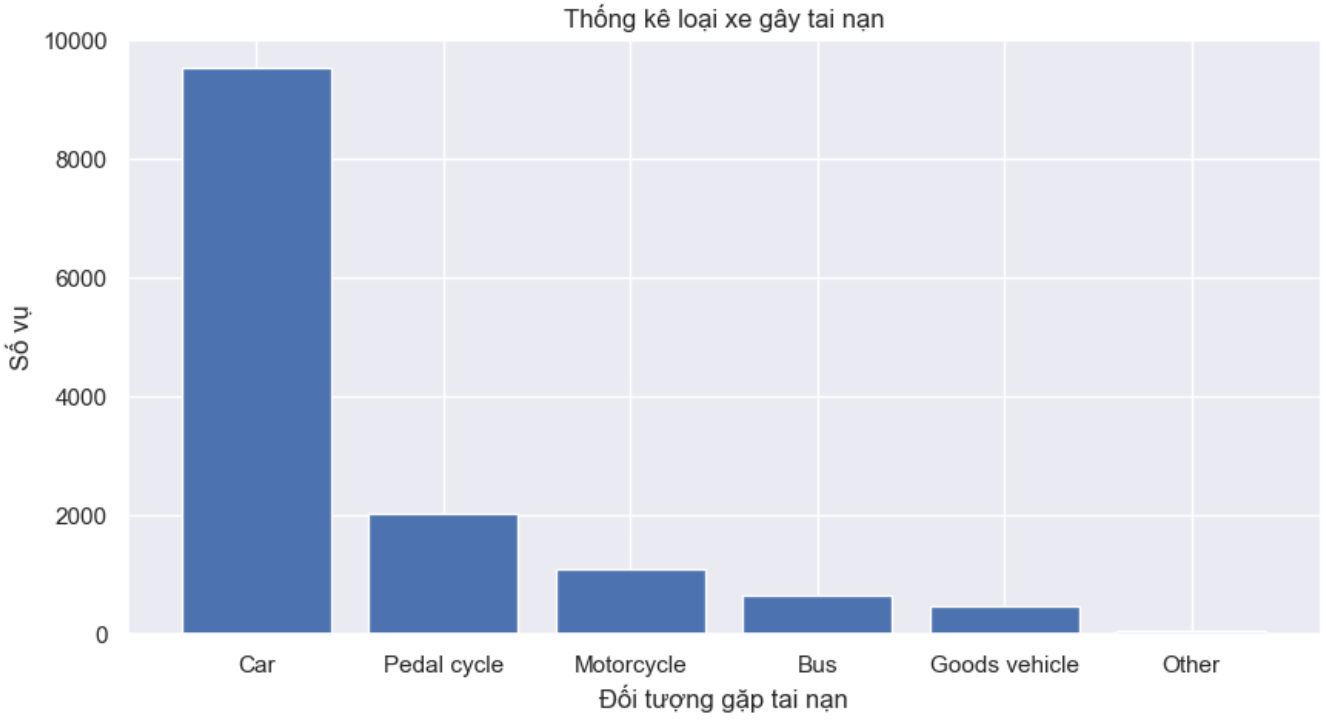
List dt dùng để chứa dữ liệu của cột Accident Date sau khi đã được chuyển đổi về dạng dữ liệu datetime. list_month và list_year dùng để chứa lần lượt tháng và năm của từng trường dữ liệu trong list dt.

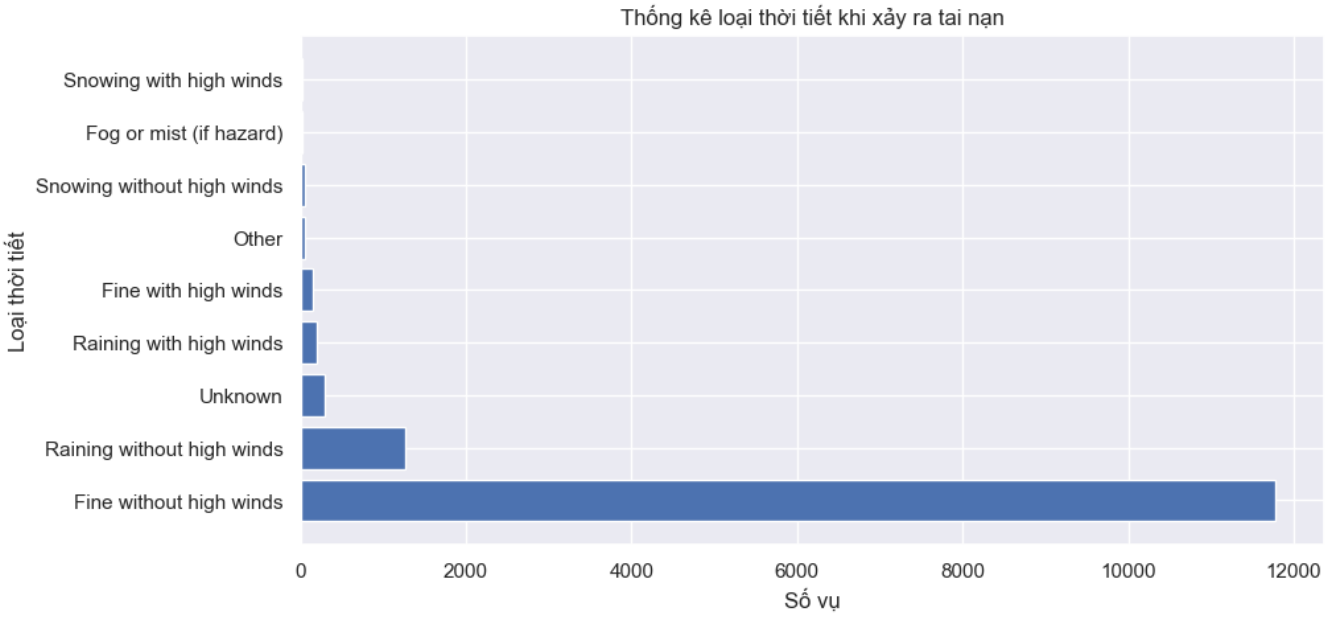
3.3. Phân tích dữ liệu

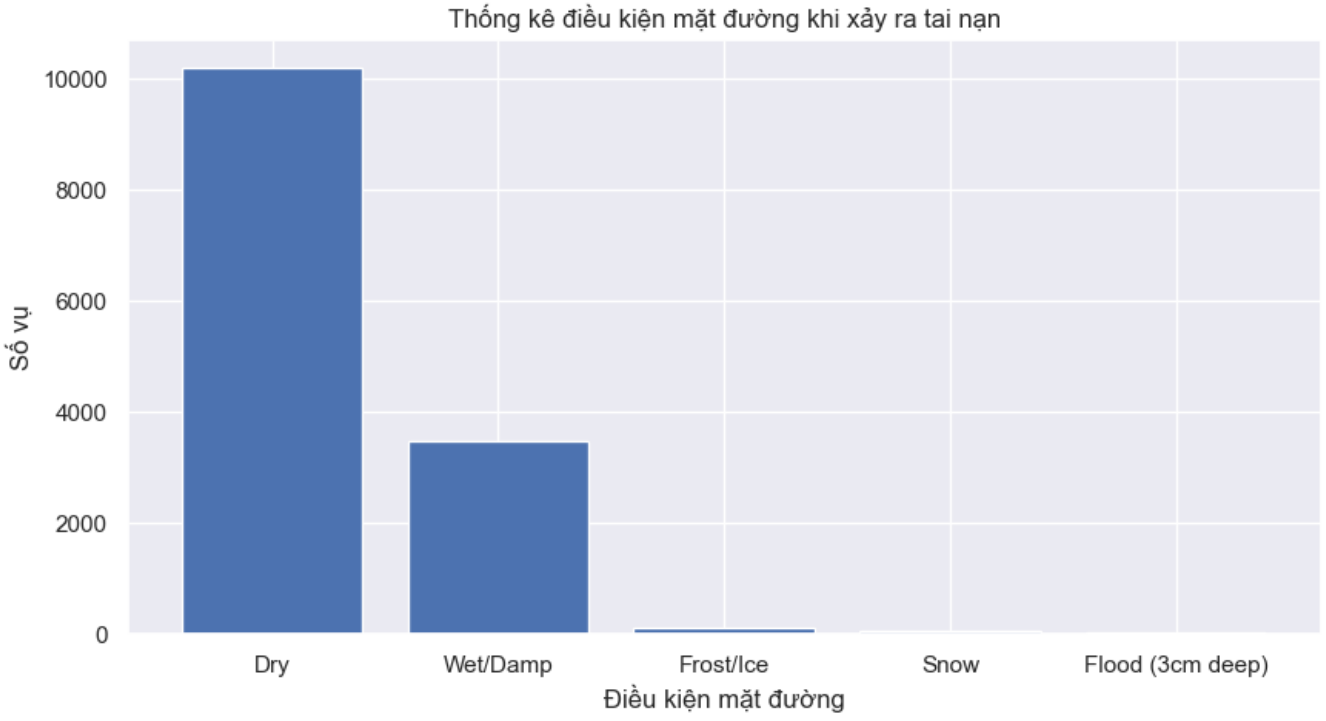
3.3.1. Thống kê các thông số gây ra tai nạn

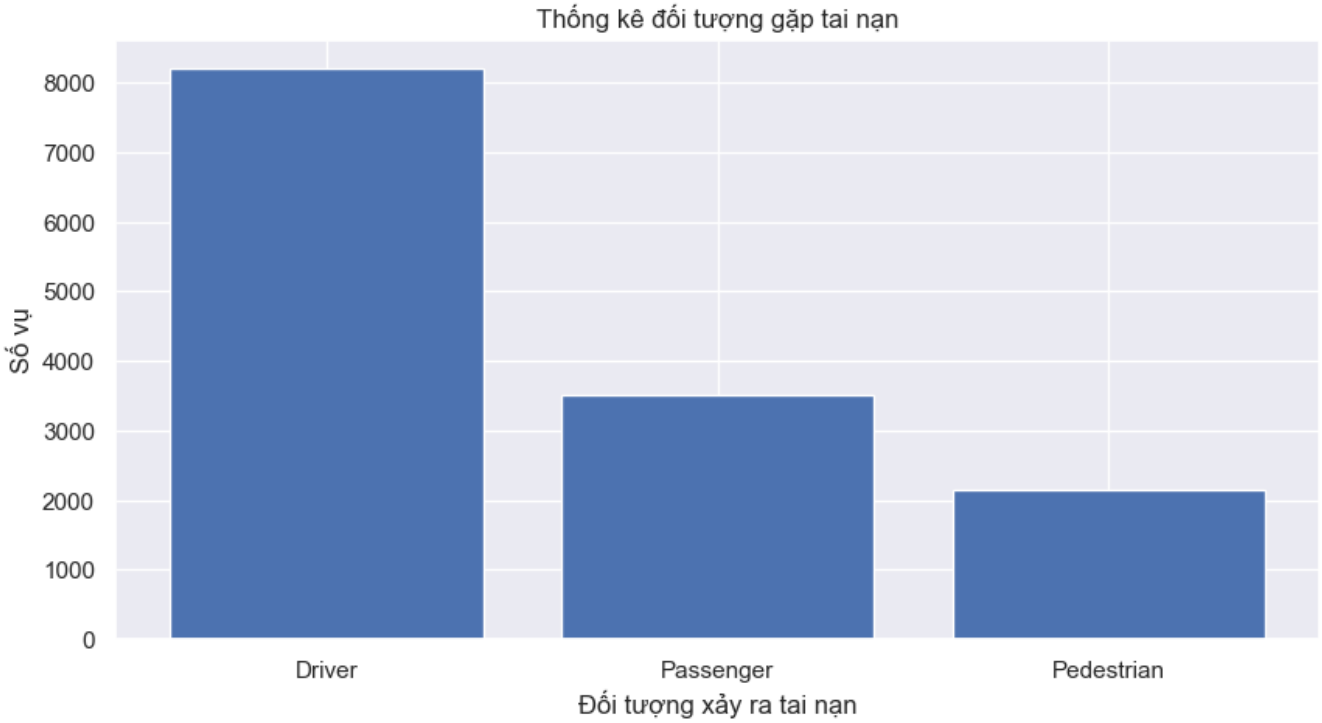
Ở phần này, chúng mình đã thử quan sát phân bố của một vài biến category.

Vì thế nên chúng mình đã vẽ 4 biểu đồ cột để thể hiện tần suất xuất hiện của các loại phương tiện/ thời tiết/ điều kiện mặt đường/ đối tượng tai nạn và sắp xếp tần số theo thứ tự giảm dần (để nhấn mạnh vào những tần số có giá trị cao).




Thống kê loại thời tiết khi xảy ra tai nạn
Snowing with high winds
Fog or mist (if hazard)
Snowing without high winds
Other
Fine with high winds
Raining with high winds
Unknown
Raining without high winds
Fine without high winds
Loại thời tiết
0
2000
4000
6000
8000
10000
12000
Số vụ


Thống kê điều kiện mặt đường khi xảy ra tai nạn
10000
8000
6000
4000
2000
0
Số vụ
Dry
Wet/Damp
Frost/Ice
Snow
Flood (3cm deep)
Điều kiện mặt đường



Nhận xét:

Loại phương tiện phổ biến là xe hơi (Car)

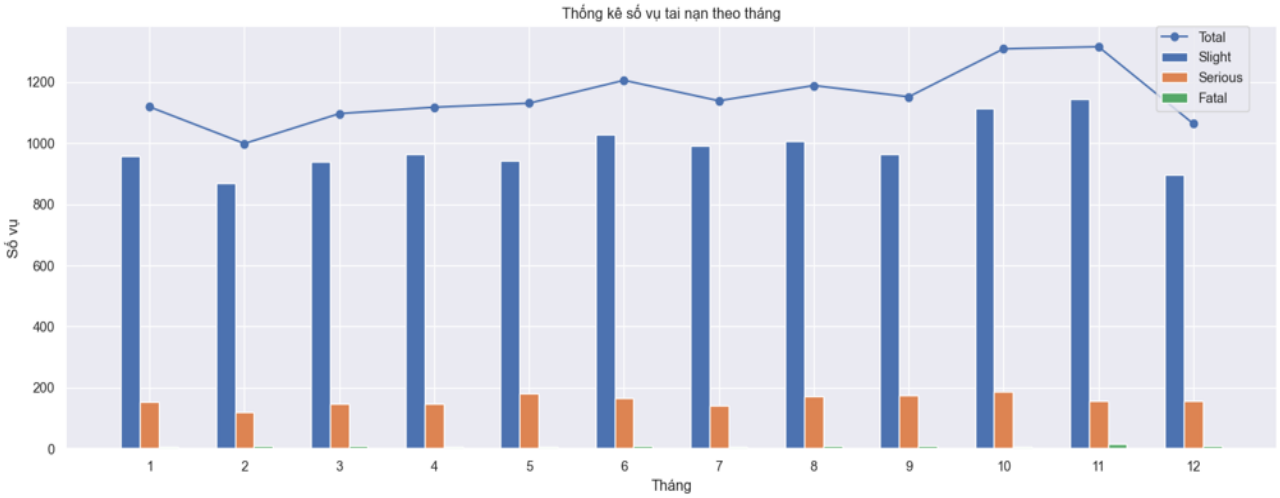
Kiểu thời tiết phổ biến là thời tiết tốt không có gió to (Fine without high winds)

Điều kiện mặt đường phổ biến là khô ráo (Dry)

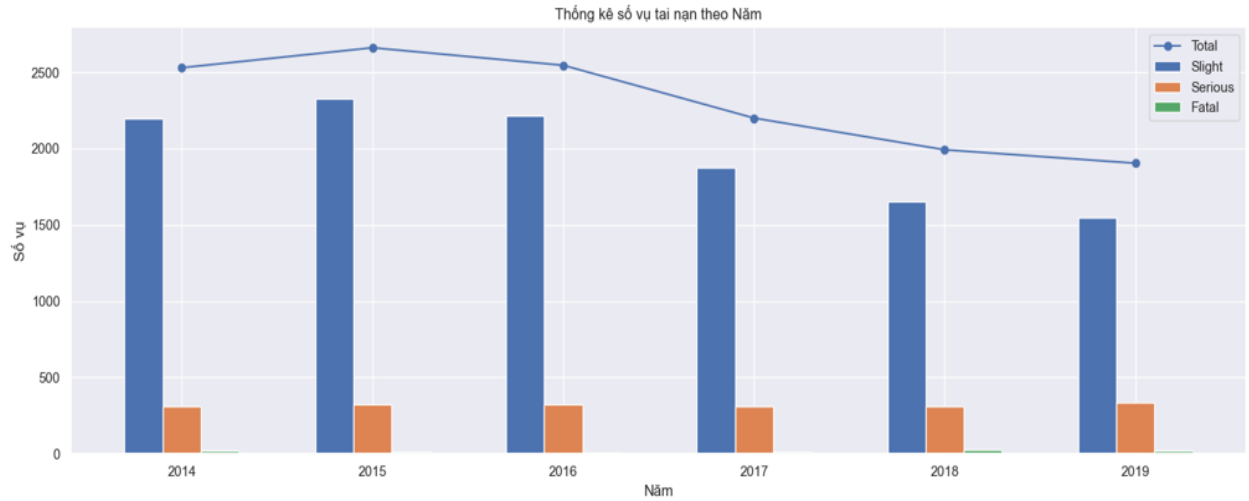
Đối tượng gặp tai nạn phổ biến là tài xế (Driver)

3.3.2. Xu hướng các vụ tai nạn theo tháng và năm

Ở phần này nhóm em vẽ 2 group bar chart (1 theo tháng và 1 theo năm). Trong mỗi group, cần thể hiện được số lượng các vụ tai nạn theo mức độ (như vậy, mỗi group sẽ có 3 cột). Cũng trong 2 biểu đồ trên, tụi mình vẽ thêm 2-line chart thể hiện tổng số ca bị tai nạn theo tháng/ năm.



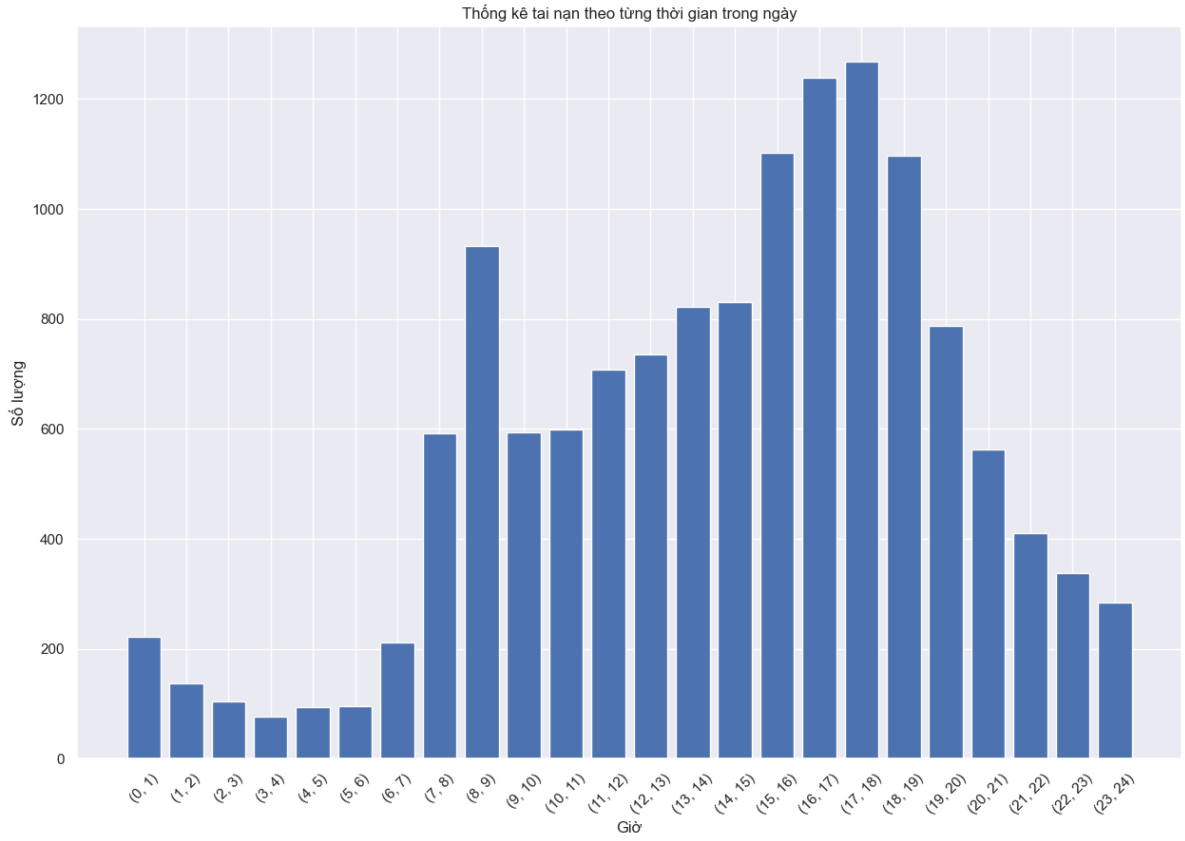
Các vụ tai nạn thường có xu hướng giảm từ tháng 12 cho đến tháng 5 và tăng mạnh từ tháng 6 đến tháng 11. Nguyên nhân có thể là do tình hình thời tiết từ tháng 12-5 thường mưa và ẩm ướt và từ tháng 6-11 thời tiết khô ráo và ít mưa hơn (dựa vào số vụ tai nạn theo tình hình thời tiết và điều kiện mặt đường). Đỉnh điểm là thời điểm tháng 10 đến tháng 11. Thấp nhất là khoảng thời gian từ tháng 1 đến tháng 2.



Các vụ tai nạn giảm dần qua các năm sau khi tăng nhẹ vào năm 2015. Nguyên nhân có thể là do mọi người dần chuyển qua di chuyển bằng các phương tiện công cộng và ít sử dụng xe hơi cá nhân để di chuyển, nhờ đó các vụ tai nạn giảm đáng kể so với các năm trước.

Đây là điều tích cực mà chúng ta có thể thấy. Được kết quả như thế thì con người đang dần có ý thức hơn trong việc tham gia giao thông. Bên cạnh đấy không thể phủ nhận những công trình, những cơ sở vật chất, đường đi đã và đang được cải thiện.

3.3.3. Khung giờ thường xảy ra tai nạn trong ngày



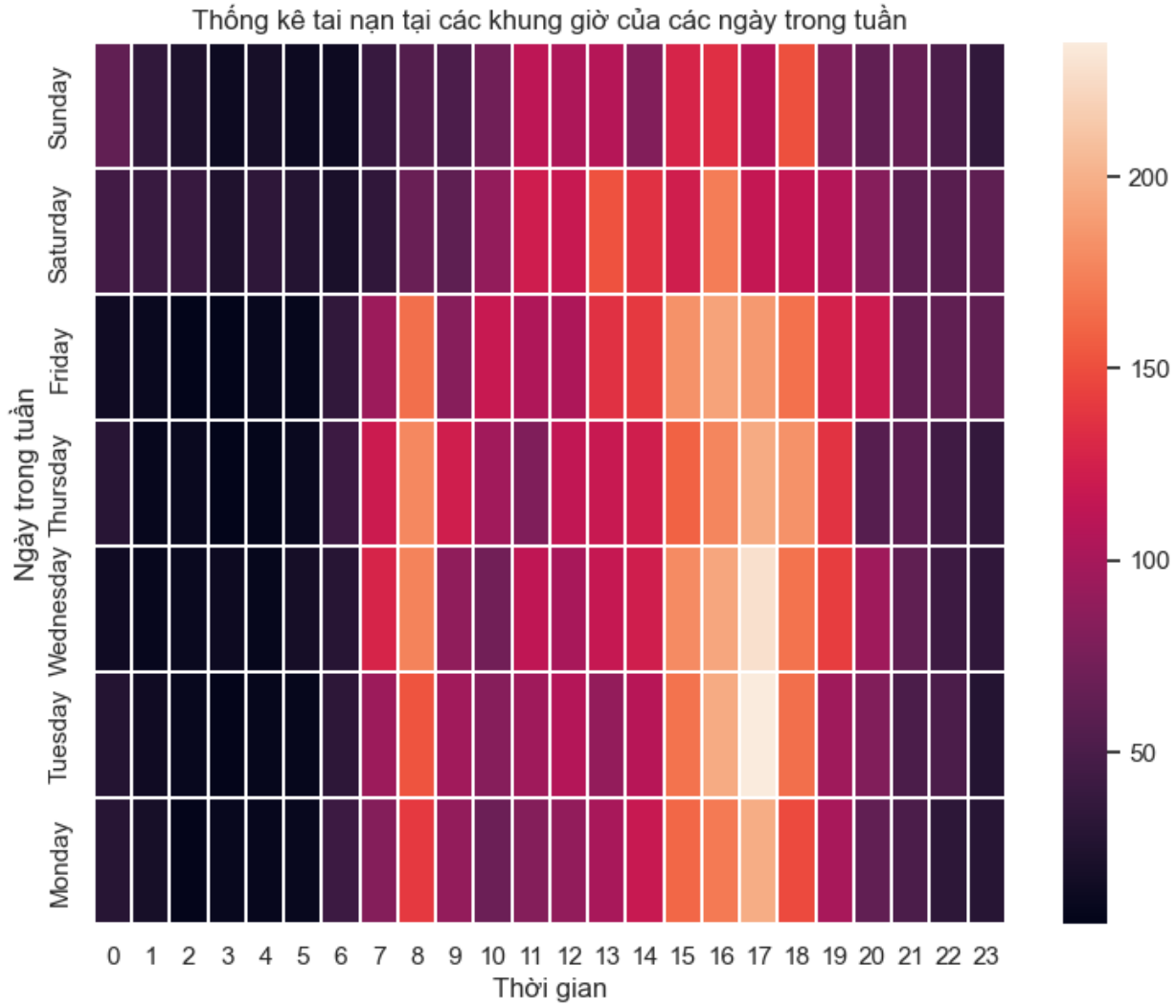
Theo quan sát ta nhận thấy tai nạn xảy ra nhiều nhất vào khoảng 8 - 9 giờ (khung giờ sáng) và khoảng 15 - 20 giờ (khung giờ tối).

3.3.4. Các ngày trong tuần thường xảy ra tai nạn

Câu trả lời cho câu hỏi trên đã giúp xác định khoảng thời gian thường xảy ra tai nạn. Tuy nhiên, liệu ngày nào trong tuần cũng xảy ra tai nạn vào các khung giờ đó? Để kiểm tra, nhóm sẽ tìm hiểu cụ thể xem ngày nào trong tuần thì hay xảy ra tai nạn.

Nhóm em đã vẽ 1 heatmap với trục tung thể hiện các ngày trong tuần (từ Thứ Hai đến Chủ Nhật) và trục hoành thể hiện giờ trong ngày (0h - 23h). Màu trong mỗi cell sẽ thể hiện số lượng tai nạn. Màu càng sáng thì tai nạn xảy ra càng nhiều.

```
In [18]: fig = plt.figure(figsize = (9, 7))
         # sns.heatmap(g)
         sns.heatmap(g4, linewidth=.25)
         plt.xlabel("Thời gian")
         plt.ylabel("Ngày trong tuần")
         plt.yticks(np.arange(0.5, 7.5), day_labels)
         plt.title("Thống kê tai nạn tại các khung giờ của các ngày trong tuần")
         plt.show()
```

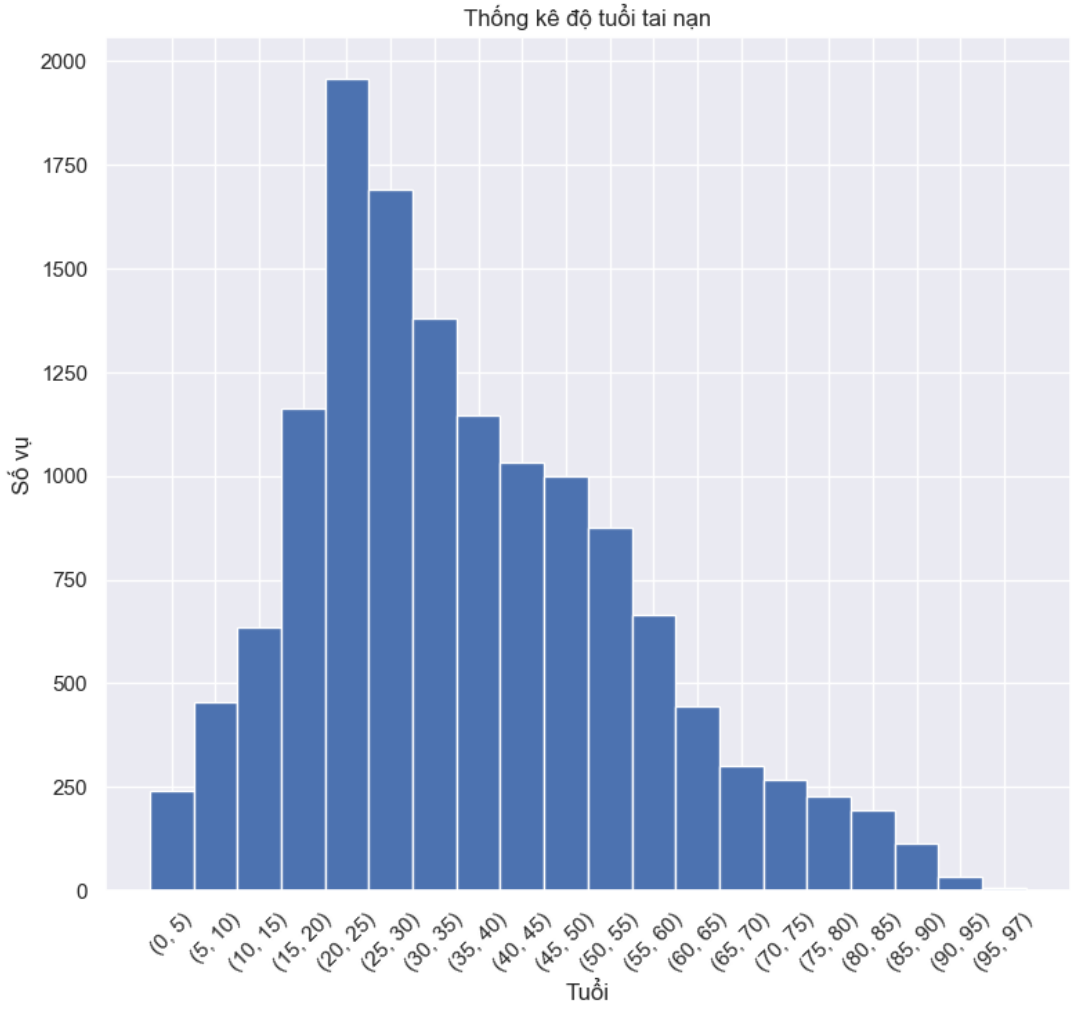


Màu càng sáng càng xảy ra nhiều vụ tai nạn. Chúng ta thấy cell sáng nhất là thời gian lúc 17h thứ Ba theo sau đó là cell ở vị trí 17h thứ Tư.

Đa phần các vụ tai nạn xảy ra ở trong khoảng 7 giờ sáng đến 7 giờ tối. Nhiều nhất là khung giờ 15 đến 18 giờ mỗi ngày. Lý do vì đây là giờ cao điểm, lượng người dân di chuyển từ nhà đến nơi làm việc và ngược lại cao nên tỷ lệ xảy ra tai nạn trong khung giờ này cao hơn so với những khung giờ khác trong ngày.

3.3.5. Thống kê độ tuổi thường gặp tai nạn

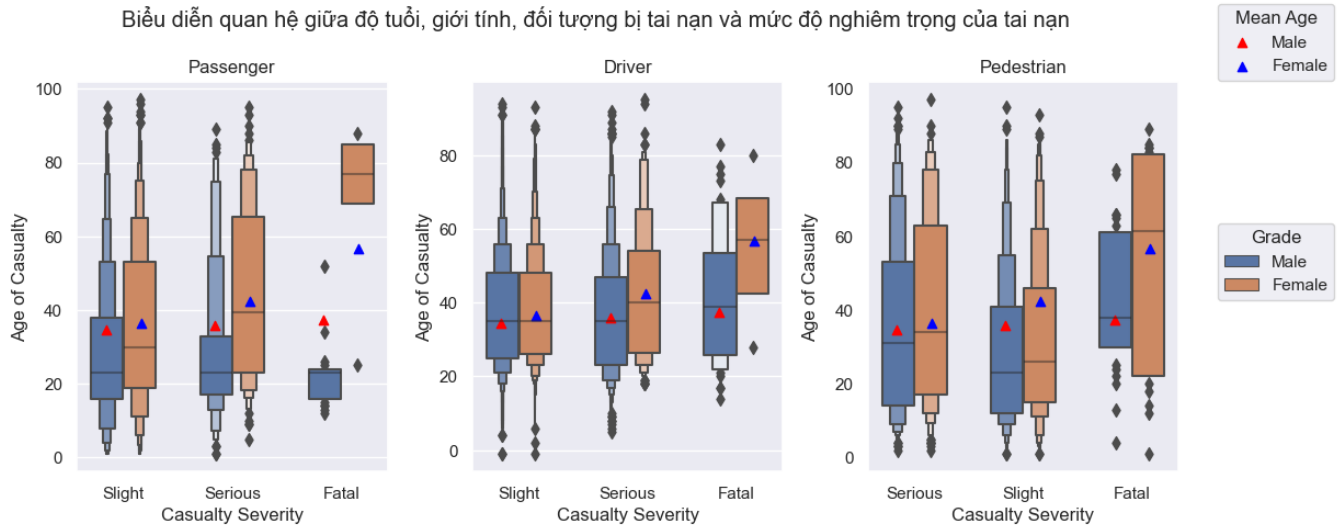
Phần này nhóm sẽ vẽ biểu đồ cột để thể hiện sự tương quan giữa các số lượng các vụ tai nạn so với các nhóm tuổi khác nhau.



Quan sát biểu đồ ta nhận thấy nhóm tuổi từ 20 đến 35 là nhóm tuổi thường gặp tai nạn. Cụ thể nhóm 20 đến 25 có gần 2000 vụ.

Lý do cũng dễ hiểu, đây là nhóm tuổi tham gia giao thông nhiều hơn so với các nhóm đối tượng khác. Một số tham gia giao thông thiếu ý thức, chủ yếu là thanh niên trẻ tuổi.

3.3.6. Mối quan hệ giữa độ tuổi, giới tính, và mức độ nghiêm trọng của các vụ tai nạn



Theo figure biểu diễn mỗi quan hệ giữa 4 đối tượng độ tuổi, giới tính, đối tượng và mức độ nghiệm trọng của tai nạn thì ta rút được những điều sau:

Với đối tượng là "Passenger" (hành khách) thì các vụ tai nạn ở nữ giới độ tuổi phân bố rộng hơn nam giới. Độ tuổi trung bình gây tai nạn của nữ cũng cao hơn nam giới. Và đặc biệt là số vụ tai nạn ở mức độ gây tử vong thì có sự chênh lệch rất lớn giữa nam và nữ giới (độ tuổi của nữ giới cao hơn nhiều so với nam giới). Ý nghĩa là nữ giới thì mức độ xảy ra tai nạn trải đều hơn so với nam giới và độ tuổi xảy ra tai nạn cũng cao hơn. Nguyên nhân có lẽ là do nam giới thì thường thích tự lái xe hơn. Hoặc cũng có lẽ là do tuổi thọ của nữ giới cao hơn.

Với đối tượng là "Driver" (tài xế) thì sự phân bố tuổi và tuổi trung bình của nam và nữ giới khá bằng nhau và sự chênh lệnh cũng giảm mạnh so với đối tượng hành khách. Ý nghĩa là đối tượng tài xế thì mức độ xảy ra tai nạn ở các độ tuổi của nam và nữ gần như là bằng nhau. Nguyên nhân có lẽ là do với đối tượng tài xế thì dù là nam hay nữ thì thường chỉ lái xe vào độ tuổi từ 20 - 50 tuổi.

Với đối tượng là "Pedestrian" (người đi bộ) thì sự phân bố tuổi và trung bình tuổi của nam và nữ giới không có sự chênh lệch quá lớn và sự phân bố độ tuổi ở cả nam và nữ cũng rộng hơn các đối tượng khác. Ý nghĩa là tai nạn xảy ra ở các độ tuổi ở cả nam và nữ dường như là như nhau. Nguyên nhân có lẽ là tai nạn xảy ra với người đi bộ ít bị ảnh hưởng bởi yếu tố độ tuổi vì người đi bộ ít khi là nguyên nhân chính dẫn đến tai nạn.

Tổng quan tất cả các quan hệ cho ta thấy, sự phân bố tuổi của nữ giới luôn cao hơn nam giới. Ý nghĩa là nữ giới ở độ tuổi cao có tỉ lệ xảy ra tai nạn cao hơn và ở nam thì độ tuổi có tỉ lệ gây ra tai nạn cao hơn là từ 20 - 50 tuổi và thường chỉ nằm ở khoảng này.

4

CÁC TÀI LIỆU THAM KHẢO

Link Github:

https://github.com/laitoanthang/KHDL-PythonForDS

Link Notion:

https://www.notion.so/thangdumbest/N-CU-I-K-2dd34951d9fe4c269adf020471cd16c7\

Link câu hỏi:
https://docs.google.com/spreadsheets/d/1UG6YbiDmWzxuseNBtZs652whpZgNG6vaiT6ynEhZRBM/edit#gid=0

Data set:

https://open.canada.ca/data/en/dataset/8dd0ab9b-d45d-4526-9256-c598fbc4ff3a