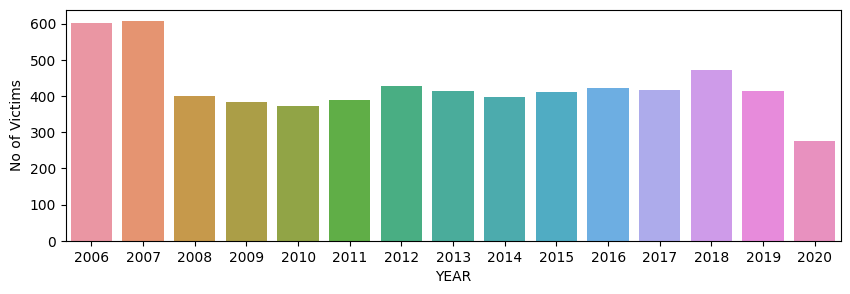
**DATA DESCRIPTION AND DATA PRE-PROCESSING**

In this paper, we analyze and forecast all police-reported vehicle collisions on Canadian public roads between 2006 and 2020. The open-source database and the data dictionary are both provided by Transport Canada at the Government of Canada's National Collision Database (NCDB). The NCDB dataset contains 57 columns (excluding the ID columns) and 16,860 rows of observations, each of which represents a person involved in a collision. Temporal attributes include the year, month, day of the week, and collision hour; spatial attributes include collision configuration, roadway configuration, number of vehicles involved in the collision, weather condition, road surface, and road alignment; and collision severity and traffic control attributes. Personal-related elements include the person's visibility, alcohol, position in the collision, road user class, safety devices used, and accident class. "ACCLASS," which stands for the accident class, is our dependent variable. This is divided into three categories: no fatal injury (14561), injury (2297), and only property damage (2). In our preliminary analysis of 14 years of data, from 2006 to 2020, we discovered that 86.36% of collisions resulted in no fatalities, 13.62% resulted in fatalities, and two resulted in property damage only. We see a general decreasing trend in the number of collisions from 2006 to 2020, reaching a record low in 2020. Then we shifted our focus from collisions to victims of these collisions, as shown in Figure 1, which shows a decrease in the number of victims over time.



*Figure 1: Number of victims over 14 years*

For example, over the 14-year period, 47.56% of those involved in collisions had no injuries, while 33.62% had major injuries, 7.78% had minor injuries, 6.13% had minimal injuries, and 4.97% died before receiving medical attention. It is worth noting that the number of non-injuries and injuries has reached a new low in 2020. The number of collisions gradually increased from Monday to Friday, peaking on Fridays. We discovered that most collisions occurred between 3 and 6 p.m.

Surprisingly, the highest frequency was observed in August, a summer month, and the lowest in February. This is consistent with our findings that the majority of collisions occur under clear and sunny skies, on dry and normal road surfaces, and on straight and leveled roads, outnumbering collisions under harsher weather and road conditions.

We also discovered that more than half of the collisions involved two vehicles. Furthermore, we can conclude that the majority of collisions involved light-duty vehicles such as passenger cars, vans, and pick-up trucks. The majority of the collisions have been reported to have occurred at intersections such as mid-block or at the intersection of at least two public roadways. The majority of collisions occurred in areas with no traffic control, accounting for nearly 60% of all collisions.

Furthermore, despite having a lower death ratio than people aged 60 and up, a significant number of fatalities have occurred among people aged 20 to 24. However, despite being involved in less than 4% of all collisions, in-line skaters, trailer owners, and cyclist passengers account for more than 10% of all fatalities, indicating their vulnerability. Identifying inconsistencies within personal and vehicle data elements enabled us to address more missing values correctly. For example, we imputed the unknown "road class user" attributes by using the known positions of individuals in a vehicle. We also substituted the maximum vehicle ID involved in the collision for the missing values under the "number of vehicles" attribute. We excluded 6.51% of the "individual fatality" attribute because it either had missing severity values or represented hypothetical passengers in parked cars. Our final pre-processed dataset contains 16860 collisions and 8842 individuals involved in collisions.

| **VARIABLES** | **Data Samples** |
| --- | --- |
| *YEAR* | 2006 |
| *DATE* | 2006-03-11 05:00:00+00:00 |
| *TIME* | 852 |
| *HOUR* | 8 |
| *STREET1* | BLOOR ST W |
| *STREET2* | DUNDAS ST W |
| *ROAD\_CLASS* | Major Arterial |
| *DISTRICT* | Toronto and East York |
| *LOCCOORD* | Intersection |
| *ACCLOC* | At Intersection |
| *TRAFFCTL* | Traffic Signal |
| *VISIBILITY* | Clear |
| *LIGHT* | Daylight |
| *RDSFCOND* | Dry |
| *ACCLASS* | Fatal |
| *IMPACTYPE* | Pedestrian Collisions |
| *INVTYPE* | Driver |
| *INVAGE* | unknown |
| *INJURY* | None |
| *INITDIR* | South |
| *VEHTYPE* | Automobile, Station Wagon |
| *MANOEUVER* | Turning Left |
| *DRIVACT* | Failed to Yield Right of Way |
| *DRIVCOND* | Unknown |
| *PEDTYPE* | <Null> |
| *PEDACT* | <Null> |
| *PEDCOND* | <Null> |
| *PEDESTRIAN* | Yes |
| *CYCLIST* | <Null> |
| *AUTOMOBILE* | Yes |
| *MOTORCYCLE* | <Null> |
| *TRUCK* | <Null> |
| *TRSN\_CITY\_VEH* | <Null> |
| *EMERG\_VEH* | <Null> |
| *PASSENGER* | <Null> |
| *SPEEDING* | <Null> |
| *AG\_DRIV* | Yes |
| *REDLIGHT* | <Null> |
| *ALCOHOL* | <Null> |
| *DISABILITY* | <Null> |
| *POLICE\_DIVISION* | D11 |
| *NEIGHBOURHOOD* | High Park North (88) |
| *DAY* | Saturday |