**Toronto KSI Collisions: Comprehensive Project Report**

Group 6

Group Members:

* Carlos De La Cruz
* Manav
* Harsh
* Rishi

**1. Dataset Description**

This dataset contains information on Killed or Seriously Injured (KSI) collisions that occurred in Toronto, sourced from the Toronto Police Service Public Safety Data Portal. It provides detailed records of traffic accidents, including location, time, contributing factors, and involved parties, with a focus on incidents resulting in fatalities or serious injuries.

**2. Data Fields**

This section describes the various fields present in the dataset used for this project, including information on data types and the count of missing values.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Field | Field Name | Description | Categorical | Categories (if applicable) | Missing/Null Count | Data Type |
| 0 | OBJECTID | Unique Identifier (auto generated) | No | N/A | 0 | int64 |
| 1 | INDEX\_ | Unique Identifier | No | N/A | 0 | int64 |
| 2 | ACCNUM | Accident Number | No | N/A | 4930 | float64 |
| 3 | DATE | Date Collision Occurred (time is displayed in UTC format when downloaded as a CSV) | No | N/A | 0 | object |
| 4 | TIME | Time Collision Occurred | No | N/A | 0 | int64 |
| 5 | STREET1 | Street Collision Occurred | Yes | Varies (Street Names) | 0 | object |
| 6 | STREET2 | Street Collision Occurred | Yes | Varies (Street Names) | 1706 | object |
| 7 | OFFSET | Distance and direction of the Collision | No | N/A | 15137 | object |
| 8 | ROAD\_CLASS | Road Classification | Yes | e.g., Major Arterial, Collector, Local, Expressway | 486 | object |
| 9 | DISTRICT | City District | Yes | e.g., Etobicoke, North York, Scarborough, Toronto, York | 229 | object |
| 10 | LATITUDE | Latitude | No | N/A | 0 | float64 |
| 11 | LONGITUDE | Longitude | No | N/A | 0 | float64 |
| 12 | ACCLOC | Collision Location | Yes | e.g., Intersection, Mid-Block | 5456 | object |
| 13 | TRAFFCTL | Traffic Control Type | Yes | e.g., Traffic Signal, Stop Sign, No Control, Pedestrian Crossover | 75 | object |
| 14 | VISIBILITY | Environment Condition | Yes | e.g., Clear, Rain, Snow, Fog | 24 | object |
| 15 | LIGHT | Light Condition | Yes | e.g., Day, Dusk, Dawn, Dark (no street lights), Dark (street lights) | 4 | object |
| 16 | RDSFCOND | Road Surface Condition | Yes | e.g., Dry, Wet, Snow, Ice | 29 | object |
| 17 | ACCLASS | Classification of Accident | Yes | e.g., Fatal, Non-Fatal Injury, Property Damage Only | 1 | object |
| 18 | IMPACTYPE | Initial Impact Type | Yes | e.g., Rear-End, Side-Impact, Head-On, Single Vehicle | 27 | object |
| 19 | INVTYPE | Involvement Type | Yes | e.g., Driver, Pedestrian, Cyclist, Passenger | 16 | object |
| 20 | INVAGE | Age of Involved Party | No | N/A (though often grouped into categories for analysis) | 0 | object |
| 21 | INJURY | Severity of Injury | Yes | e.g., Fatal, Serious, Minor, No Injury | 8897 | object |
| 22 | FATAL\_NO | Sequential Number | No | N/A | 18087 | float64 |
| 23 | INITDIR | Initial Direction of Travel | Yes | e.g., North, South, East, West | 5277 | object |
| 24 | VEHTYPE | Type of Vehicle | Yes | e.g., Passenger Car, Truck, Motorcycle, Bus, Bicycle | 3487 | object |
| 25 | MANOEUVER | Vehicle Manoeuver | Yes | e.g., Going Ahead, Turning Left, Turning Right, Changing Lanes | 7953 | object |
| 26 | DRIVACT | Apparent Driver Action | Yes | e.g., Impaired, Disobeyed Traffic Sign, Speeding, Distracted | 9289 | object |
| 27 | DRIVCOND | Driver Condition | Yes | e.g., Normal, Fatigued, Impaired, Medical Condition | 9291 | object |
| 28 | PEDTYPE | Pedestrian Crash Type - detail | Yes | Varies (Specific pedestrian involvement scenarios) | 15728 | object |
| 29 | PEDACT | Pedestrian Action | Yes | e.g., Crossing with right-of-way, Crossing against signal, Jaywalking | 15730 | object |
| 30 | PEDCOND | Condition of Pedestrian | Yes | e.g., Normal, Impaired, Medical Condition | 15711 | object |
| 31 | CYCLISTYPE | Cyclist Crash Type - detail | Yes | Varies (Specific cyclist involvement scenarios) | 18152 | object |
| 32 | CYCACT | Cyclist Action | Yes | e.g., Going Ahead, Turning, Changing Lanes, Unknown | 18155 | object |
| 33 | CYCCOND | Cyclist Condition | Yes | e.g., Normal, Impaired, Medical Condition | 18157 | object |
| 34 | PEDESTRIAN | Pedestrian Involved In Collision | Yes | Boolean (Yes/No, 1/0) | 11269 | object |
| 35 | CYCLIST | Cyclists Involved in Collision | Yes | Boolean (Yes/No, 1/0) | 16971 | object |
| 36 | AUTOMOBILE | Driver Involved in Collision | Yes | Boolean (Yes/No, 1/0) | 1727 | object |
| 37 | MOTORCYCLE | Motorcyclist Involved in Collision | Yes | Boolean (Yes/No, 1/0) | 17273 | object |
| 38 | TRUCK | Truck Driver Involved in Collision | Yes | Boolean (Yes/No, 1/0) | 17788 | object |
| 39 | TRSN\_CITY\_VEH | Transit or City Vehicle Involved in Collision | Yes | Boolean (Yes/No, 1/0) | 17809 | object |
| 40 | EMERG\_VEH | Emergency Vehicle Involved in Collision | Yes | Boolean (Yes/No, 1/0) | 18908 | object |
| 41 | PASSENGER | Passenger Involved in Collision | Yes | Boolean (Yes/No, 1/0) | 11774 | object |
| 42 | SPEEDING | Speeding Related Collision | Yes | Boolean (Yes/No, 1/0) | 16263 | object |
| 43 | AG\_DRIV | Aggressive and Distracted Driving Collision | Yes | Boolean (Yes/No, 1/0) | 9121 | object |
| 44 | REDLIGHT | Red Light Related Collision | Yes | Boolean (Yes/No, 1/0) | 17380 | object |
| 45 | ALCOHOL | Alcohol Related Collision | Yes | Boolean (Yes/No, 1/0) | 18149 | object |
| 46 | DISABILITY | Medical or Physical Disability Related Collision | Yes | Boolean (Yes/No, 1/0) | 18464 | object |
| 47 | HOOD\_158 | Unique ID for City of Toronto Neighbourhood (new) | Yes | Numeric IDs, representing specific neighborhoods | 0 | object |
| 48 | NEIGHBOURHOOD\_158 | City of Toronto Neighbourhood name (new) | Yes | Varies (Specific neighborhood names) | 0 | object |
| 49 | HOOD\_140 | Unique ID for City of Toronto Neighbourhood (old) | Yes | Numeric IDs, representing specific neighborhoods | 0 | object |
| 50 | NEIGHBOURHOOD\_140 | City of Toronto Neighbourhood name (old) | Yes | Varies (Specific neighborhood names) | 0 | object |
| 51 | DIVISION | Toronto Police Service Division | Yes | e.g., 51 Division, 52 Division, 53 Division, etc. | 0 | object |
| 52 | x | X Coordinate (likely related to Latitude/Longitude) | No | N/A | 0 | float64 |
| 53 | y | Y Coordinate (likely related to Latitude/Longitude) | No | N/A | 0 | float64 |

This section describes the various fields present in the dataset used for this project.

**3. Data Stats**

**Numerical Features**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature Name | Count | Mean | Std Dev | Min Value | 25th Percentile | 50th Percentile (Median) | 75th Percentile | Max Value | Notes |
| ACCNUM | 14027.0 | 5.58e+08 | 1.18e+09 | 2.53e+04 | 1.03e+06 | 1.22e+06 | 1.39e+06 | 4.01e+09 | Accident number; significant missing values (4930). |
| TIME | 18957.0 | 1364.96 | 631.31 | 0.0 | 924.0 | 1450.0 | 1852.0 | 2359.0 | Represents time in HHMM format. Distribution suggests peaks around midday/afternoon. |
| LATITUDE | 18957.0 | 43.71 | 0.056 | 43.59 | 43.66 | 43.70 | 43.76 | 43.86 | Geographic coordinate within Toronto. |
| LONGITUDE | 18957.0 | -79.396 | 0.104 | -79.64 | -79.47 | -79.397 | -79.318 | -79.123 | Geographic coordinate within Toronto. |
| FATAL\_NO | 870.0 | 28.75 | 17.66 | 1.0 | 14.0 | 27.5 | 42.0 | 78.0 | Sequential number for fatalities; highly sparse (18087 missing). Can be used as a flag. |
| x | 18957.0 | 629181.57 | 8364.34 | 609625.70 | 623177.00 | 629199.08 | 635424.04 | 651024.09 | Cartesian coordinate; likely derived from Lat/Long. |
| y | 18957.0 | 4.84e+06 | 6.32e+03 | 4.83e+06 | 4.84e+06 | 4.84e+06 | 4.85e+06 | 4.86e+06 | Cartesian coordinate; likely derived from Lat/Long. |

**Categorical Features**

**Time and Location Descriptors**

* **DATE** (Top 5 Dates with most collisions):
  + 8/17/2014 8:00:00 AM: 35
  + 9/1/2007 8:00:00 AM: 24
  + 7/20/2012 8:00:00 AM: 23
  + 3/20/2016 8:00:00 AM: 22
  + 4/17/2007 8:00:00 AM: 22
  + *Note: Specific dates with high counts might indicate local events or data anomalies.*
* **STREET1** (Top 5 primary streets):
  + YONGE ST: 403
  + BATHURST ST: 338
  + DUNDAS ST W: 304
  + DUFFERIN ST: 294
  + EGLINTON AVE E: 288
  + *Note: Identifies frequently occurring collision locations.*
* **STREET2** (Top 5 secondary streets for collisions):
  + BATHURST ST: 156
  + LAWRENCE AVE E: 153
  + YONGE ST: 133
  + FINCH AVE E: 126
  + EGLINTON AVE E: 123
  + *Note: Complements STREET1 for intersection analysis.*
* **OFFSET** (Top 5 offsets from a reference point):
  + 10 m West of: 61
  + 10 m North o: 60
  + 5 m South of: 60
  + 10 m South o: 59
  + 5 m East of: 55
  + *Note: Highly granular, may be challenging to use directly without aggregation or conversion to numerical distance.*
* **ROAD\_CLASS** (Road classification where collision occurred):
  + Major Arterial: 13376
  + Minor Arterial: 2958
  + Collector: 1032
  + Local: 865
  + Expressway: 164
  + *Note: Major arterials account for the vast majority of collisions.*
* **DISTRICT** (City district where collision occurred):
  + Toronto and East York: 6328
  + Etobicoke York: 4342
  + Scarborough: 4270
  + North York: 3788
  + *Note: Provides geographical segmentation within Toronto.*
* **ACCLOC** (Collision location type):
  + At Intersection: 8774
  + Non Intersection: 2660
  + Intersection Related: 1604
  + At/Near Private Drive: 407
  + Overpass or Bridge: 14
  + *Note: Intersections are the most common collision location. Significant missing values present.*
* **HOOD\_158** (Unique ID for new neighbourhood definition - Top 5 IDs):
  + 1: 597
  + 170: 376
  + 119: 361
  + 70: 353
  + 85: 304
* **NEIGHBOURHOOD\_158** (City of Toronto Neighbourhood name (new) - Top 5 names):
  + West Humber-Clairville: 597
  + Yonge-Bay Corridor: 376
  + Wexford/Maryvale: 361
  + South Riverdale: 353
  + South Parkdale: 304
  + *Note: Provides finer-grained location information than District.*
* **HOOD\_140** (Unique ID for old neighbourhood definition - Top 5 IDs):
  + 77: 740
  + 1: 592
  + 76: 445
  + 137: 409
  + 119: 361
* **NEIGHBOURHOOD\_140** (City of Toronto Neighbourhood name (old) - Top 5 names):
  + Waterfront Communities-The Island (77): 740
  + West Humber-Clairville (1): 592
  + Bay Street Corridor (76): 445
  + Woburn (137): 409
  + Wexford/Maryvale (119): 361
  + *Note: Provides alternative neighbourhood definitions; check for overlap/redundancy with \_158 versions.*
* **DIVISION** (Toronto Police Service Division - Top 5 divisions):
  + D42: 1813
  + D55: 1530
  + D41: 1435
  + D22: 1413
  + D32: 1328
  + *Note: Another geographical indicator, potentially useful for resource allocation or local policy analysis.*

**Environmental Conditions**

* **TRAFFCTL** (Traffic Control Type):
  + No Control: 9021
  + Traffic Signal: 8035
  + Stop Sign: 1464
  + Pedestrian Crossover: 208
  + Traffic Controller: 108
  + *Note: 'No Control' and 'Traffic Signal' are the dominant categories.*
* **VISIBILITY** (Environment Condition):
  + Clear: 16373
  + Rain: 1976
  + Snow: 356
  + Other: 98
  + Fog, Mist, Smoke, Dust: 52
  + *Note: Most collisions occur under clear visibility.*
* **LIGHT** (Light Condition):
  + Daylight: 10779
  + Dark: 3746
  + Dark, artificial: 3552
  + Dusk: 253
  + Dusk, artificial: 253
  + *Note: Majority occur in daylight, but a significant portion in dark conditions, with or without artificial light.*
* **RDSFCOND** (Road Surface Condition):
  + Dry: 15231
  + Wet: 3140
  + Loose Snow: 174
  + Other: 147
  + Slush: 102
  + *Note: Most collisions on dry surfaces, followed by wet surfaces.*

**Collision Characteristics**

* **ACCLASS** (Classification of Accident - **Your Target Variable**):
  + Non-Fatal Injury: 16268
  + Fatal: 2670
  + Property Damage O: 18
  + *Note: This is the primary target for your prediction. Observe the significant class imbalance between Non-Fatal Injury and Fatal collisions.*
* **IMPACTYPE** (Initial Impact Type):
  + Pedestrian Collisions: 7684
  + Turning Movement: 2934
  + Cyclist Collisions: 1861
  + Rear End: 1804
  + SMV Other: 1465
  + *Note: Pedestrian collisions are the most frequent impact type.*

**Involved Party Details**

* **INVTYPE** (Involvement Type):
  + Driver: 8651
  + Pedestrian: 3275
  + Passenger: 2889
  + Vehicle Owner: 1638
  + Cyclist: 822
  + *Note: Drivers are the most frequently involved party.*
* **INVAGE** (Age of Involved Party - Top 5 age ranges):
  + unknown: 2625
  + 20 to 24: 1800
  + 25 to 29: 1723
  + 30 to 34: 1450
  + 35 to 39: 1382
  + *Note: Contains 'unknown' values which need handling. This is a categorical representation of age.*
* **INJURY** (Severity of Injury):
  + Major: 6445
  + Minor: 1479
  + Minimal: 1160
  + Fatal: 976
  + *Note: This is closely related to target variable ACCLASS. The count for 'Fatal' here (976) differs from ACCLASS 'Fatal' (2670), which suggests careful reconciliation is needed for your target definition.*
* **INITDIR** (Initial Direction of Travel):
  + East: 3388
  + West: 3339
  + South: 3226
  + North: 3201
  + Unknown: 526
  + *Note: Relatively even distribution across cardinal directions, with some 'Unknown' values.*
* **VEHTYPE** (Type of Vehicle):
  + Automobile, Station Wagon: 7805
  + Other: 4753
  + Bicycle: 819
  + Motorcycle: 747
  + Municipal Transit Bus (TTC): 284
  + *Note: Automobiles are the most common vehicle type involved.*
* **MANOEUVER** (Vehicle Manoeuver):
  + Going Ahead: 6542
  + Turning Left: 1877
  + Stopped: 634
  + Turning Right: 505
  + Slowing or Stopping: 294
  + *Note: 'Going Ahead' is by far the most frequent maneuver. Significant missing values present.*
* **DRIVACT** (Apparent Driver Action):
  + Driving Properly: 4425
  + Failed to Yield Right of Way: 1603
  + Lost control: 1007
  + Improper Turn: 614
  + Other: 529
  + *Note: 'Driving Properly' is the top action, which might seem counter-intuitive for an accident; this category could include situations where the other party was at fault, or where the 'action' leading to the collision isn't well captured.*
* **DRIVCOND** (Driver Condition):
  + Normal: 6158
  + Inattentive: 1603
  + Unknown: 1129
  + Medical or Physical Disability: 181
  + Had Been Drinking: 166
  + *Note: Majority are 'Normal', but 'Inattentive' is a significant factor. Also contains 'Unknown'.*
* **PEDTYPE** (Pedestrian Crash Type - detail - Top 5):
  + Pedestrian hit at mid-block: 810
  + Vehicle turns left while ped crosses with ROW at inter.: 646
  + Vehicle is going straight thru inter.while ped cross without ROW: 543
  + Pedestrian hit on sidewalk or shoulder: 257
  + Vehicle is going straight thru inter.while ped cross with ROW: 189
  + *Note: Highly detailed, applies only when a pedestrian is involved. Very high missing value count for records without pedestrians.*
* **PEDACT** (Pedestrian Action - Top 5):
  + Crossing with right of way: 1019
  + Crossing, no Traffic Control: 730
  + Crossing without right of way: 453
  + On Sidewalk or Shoulder: 284
  + Other: 240
  + *Note: Similar to PEDTYPE, applicable only to pedestrian-involved collisions.*
* **PEDCOND** (Condition of Pedestrian - Top 5):
  + Normal: 1842
  + Inattentive: 560
  + Unknown: 399
  + Had Been Drinking: 222
  + Other: 84
  + *Note: Condition of pedestrian at the time of collision.*
* **CYCLISTYPE** (Cyclist Crash Type - detail - Top 5):
  + Motorist turned left across cyclists path.: 140
  + Cyclist without ROW rides into path of motorist at inter, lnwy, dwy-Cyclist not turn.: 117
  + Cyclist and Driver travelling in same direction. One vehicle sideswipes the other.: 114
  + Cyclist and Driver travelling in same direction. One vehicle rear-ended the other.: 59
  + Motorist without ROW drives into path of cyclist at inter, lnwy, dwy-Driver not turn.: 58
  + *Note: Highly detailed, applies only when a cyclist is involved. Very high missing value count.*
* **CYCACT** (Cyclist Action - Top 5):
  + Driving Properly: 444
  + Disobeyed Traffic Control: 85
  + Other: 79
  + Failed to Yield Right of Way: 65
  + Lost control: 42
  + *Note: Actions of the cyclist at the time of collision.*
* **CYCCOND** (Condition of Cyclist - Top 5):
  + Normal: 557
  + Inattentive: 110
  + Unknown: 79
  + Had Been Drinking: 30
  + Other: 12
  + *Note: Condition of the cyclist at the time of collision.*

**Involvement Flags (Binary 'Yes' Features)**

* **PEDESTRIAN**: Yes: 7688
* **CYCLIST**: Yes: 1986
* **AUTOMOBILE**: Yes: 17230
* **MOTORCYCLE**: Yes: 1684
* **TRUCK**: Yes: 1169
* **TRSN\_CITY\_VEH**: Yes: 1148
* **EMERG\_VEH**: Yes: 49
  + *Note: Very low count for emergency vehicles, indicating rarity. Consider if this feature is useful or too sparse.*
* **PASSENGER**: Yes: 7183

**Contributing Factors (Binary 'Yes' Features)**

* **SPEEDING**: Yes: 2694
* **AG\_DRIV** (Aggressive and Distracted Driving Collision): Yes: 9836
* **REDLIGHT**: Yes: 1577
* **ALCOHOL**: Yes: 808
* **DISABILITY**: Yes: 493

**4. Data Preprocessing and Feature Engineering**

The primary goal was to convert person-level data into an accident-level, while engineering features from existing information.

**Handling Missing Values**

* **Imputing Missing Accident Numbers (ACCNUM)**:
  + The ACCNUM (renamed to accident\_number) column initially had missing values, and multiple rows could pertain to the same accident.
  + A function handle\_missing\_accident\_numbers was developed to identify unique accidents using a composite key (date, time, street1, latitude, longitude, light, accident\_class). This allowed the team to salvage over 4,900 orphan samples and generate over 1,900 valid collisions.
* **Standardizing Boolean Columns**:
  + Columns like pedestrian, cyclist, automobile, speeding, alcohol, aggressive\_driving, redlight, and disability contained 'Yes', 'No', or NaN values.
  + These were transformed into a numerical format: 'Yes' values were mapped to 1, and all other values (including 'No' and NaN) were mapped to 0. This treats the absence of a 'Yes' indicator as 0.

**Data Type Conversions**

* **Date Conversion**:
  + The DATE column, initially an object type, was converted into a standardized datetime format (pd.to\_datetime). This conversion is essential for accurate temporal analysis and feature derivation.

**Feature Creation**

New features were engineered to provide more insightful information from the existing data, and existing features were transformed to be more suitable for analysis at the accident level.

* **Year Extraction**:
  + From the converted date column, a new year feature was extracted, enabling analysis of accident trends over different years.
* **Injury Severity Scoring**:
  + The INJURY column, describing injury severity ('Fatal', 'Major', 'Minor', 'Minimal', 'None'), was mapped to a new numerical feature, injury\_severity\_score. A numerical mapping was applied: 'Fatal' = 4, 'Major' = 3, 'Minor' = 2, 'Minimal' = 1, and 'None' = 0. This ordinal encoding provides a quantitative measure of accident severity.
* **Aggregation and Feature Consolidation (Accident-Level Features)**:
  + The team transformed the dataset from an individual-level records to a single record per unique accident\_number through a groupby().agg() operation. This provided an accident focused view with features:
    - **first aggregation**: For accident specific attributes (e.g., date, time, street1, coordinates, road\_class, district, traffic control, visibility, light conditions, road surface conditions, neighborhood information, accident\_class, and transformed boolean involvement flags like pedestrian, cyclist, automobile, speeding, alcohol), the first observed value for that accident\_number was taken.
    - **get\_most\_frequent aggregation**: For features like impact\_type and initial\_direction, the most frequently occurring value within each accident group was used.
    - **get\_all\_unique aggregation**: For descriptive features that vary among individuals within an accident (e.g., involvement\_type, vehicle\_type, manoeuver, driver\_action, driver\_condition, and all pedestrian/cyclist-related action/condition columns, involvement\_age), all unique non-null values were collected into lists. This allows the team to select the most optimal encoding strategy as the next feature engineering step.
    - **Fatal Injury Flag**: The injury column was transformed into a boolean (True/False) indicating whether *any* fatal injury occurred within that specific accident.
    - **Maximum Injury Severity Score**: For the injury\_severity\_score, the maximum score among all individuals in an accident was taken, representing the most severe outcome for that accident.

**Feature Selection and Removal**

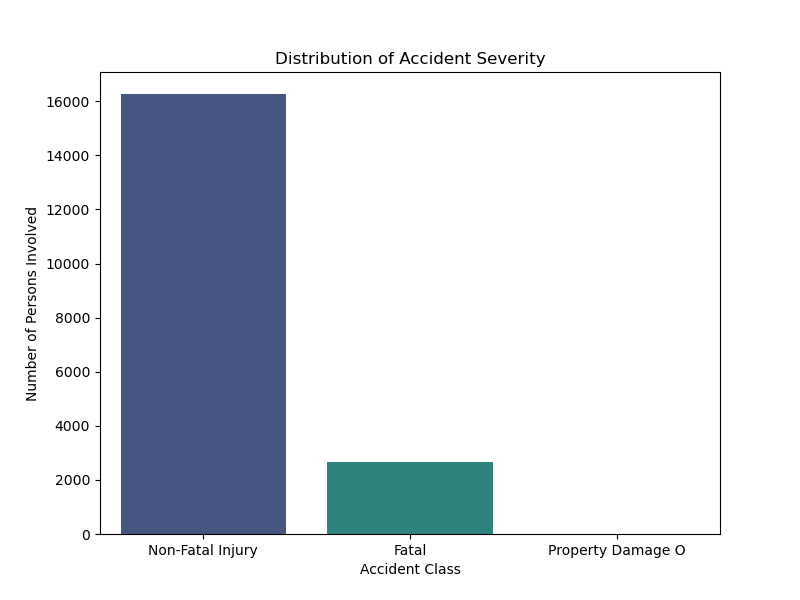
* **Implicit Feature Selection via Aggregation**: By grouping data by accident\_number and applying specific aggregation functions, the process effectively selected and transformed features from individual-level detail to accident-level summaries. For instance, instead of multiple rows for involvement\_type, a single row now contains a list of all unique involvement types for that accident.
* The OBJECTID column from the original dataset was implicitly dropped during the aggregation process, as it was not included in the aggregation\_dict.
* The INDEX columns from the original dataset was also implicitly dropped during aggregation.

**Future Data Preprocessing and Feature Engineering Considerations**

* The team has identified several areas for refinement in data preprocessing and feature engineering to enhance the dataset.
* Addressing Redundancy in Involvement Features:
  + Observation: A redundancy exists between the involvement\_type column (which captures all unique involvement types as a list for each accident, e.g., ['Driver', 'Pedestrian']) and the individual boolean flags (e.g., pedestrian, automobile, motorcycle, etc.). These boolean flags act as a form of one-hot encoding for specific involvement categories at the accident level.
  + **Proposed Action**: To improve feature efficiency, the team is considering dropping the existing individual binary features (such as pedestrian, cyclist, automobile, motorcycle, truck, trsn\_city\_veh, emerg\_veh, passenger, speeding, aggressive\_driving, redlight, alcohol, and disability). Instead, we will do a one-hot encoding approach to the involvement\_type and vehicle\_type lists. This would generate new, explicit one-hot encoded features like involvement\_type\_pedestrian or vehicle\_type\_truck.

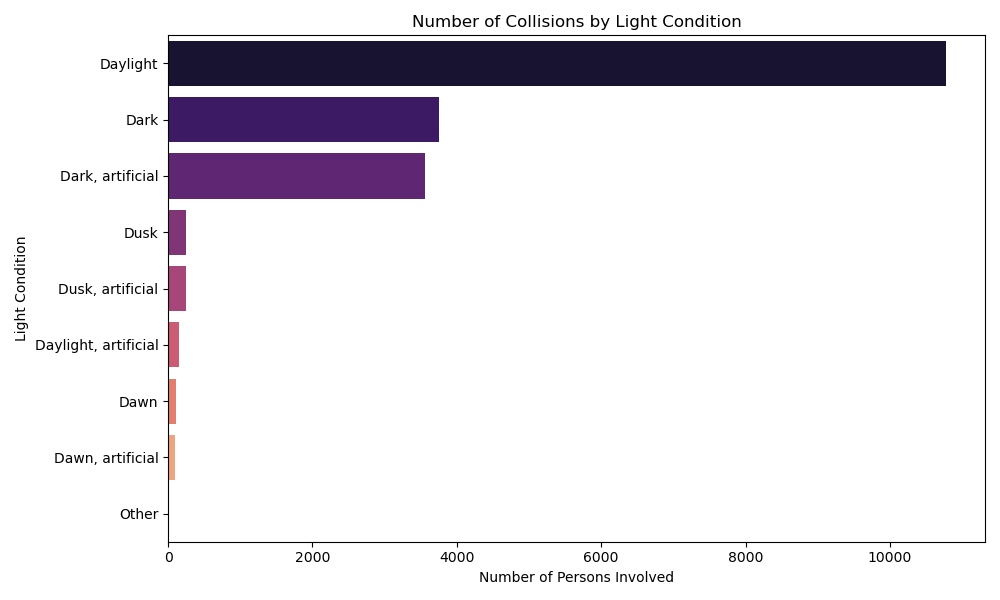
**Data Visualization and Exploratory Data Analysis**

**1. Distribution of Accident Severity**

****

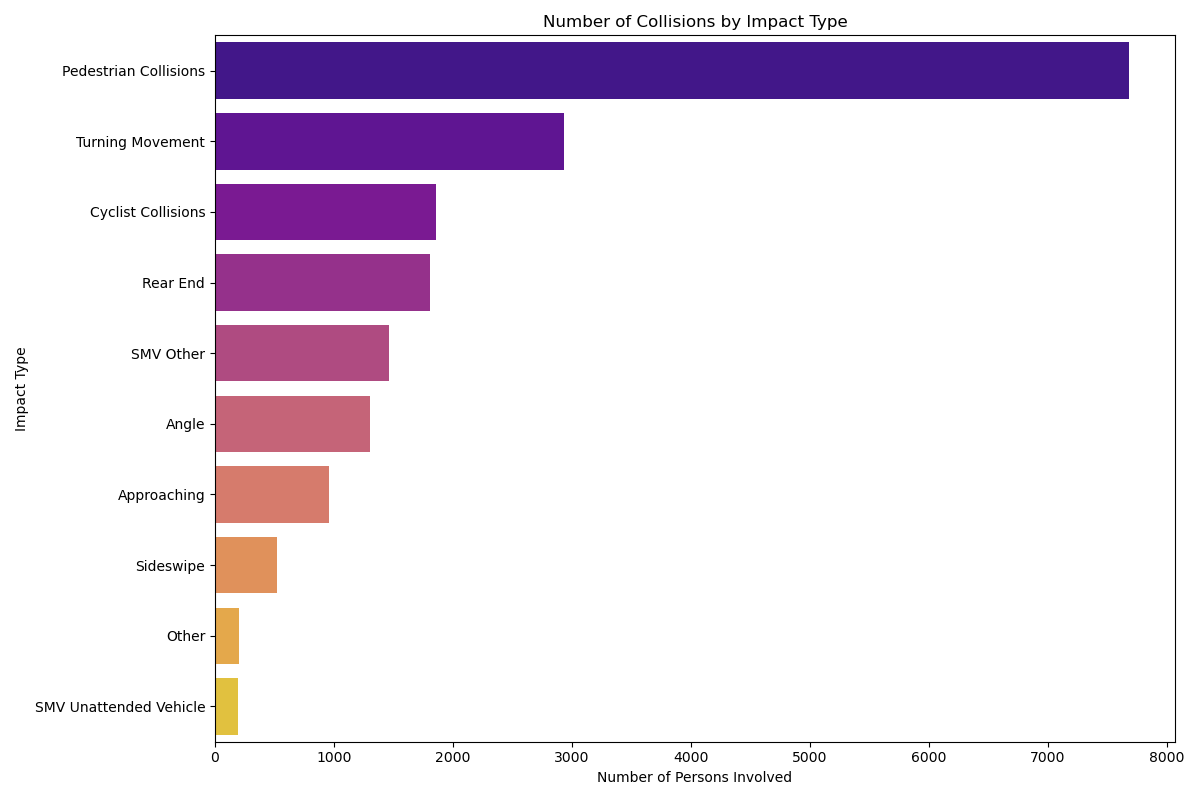
This bar chart illustrates the overall distribution of accident severities within the Toronto KSI (Killed or Seriously Injured) collisions dataset. It highlights the significant class imbalance, with **"Non-Fatal Injury"** collisions being by far the most frequent, followed by **"Fatal"** collisions, and a very small proportion of **"Property Damage Only"** incidents. This imbalance is critical for subsequent modeling efforts, particularly when predicting accident class.

**2. Number of Collisions by Light Condition**

****

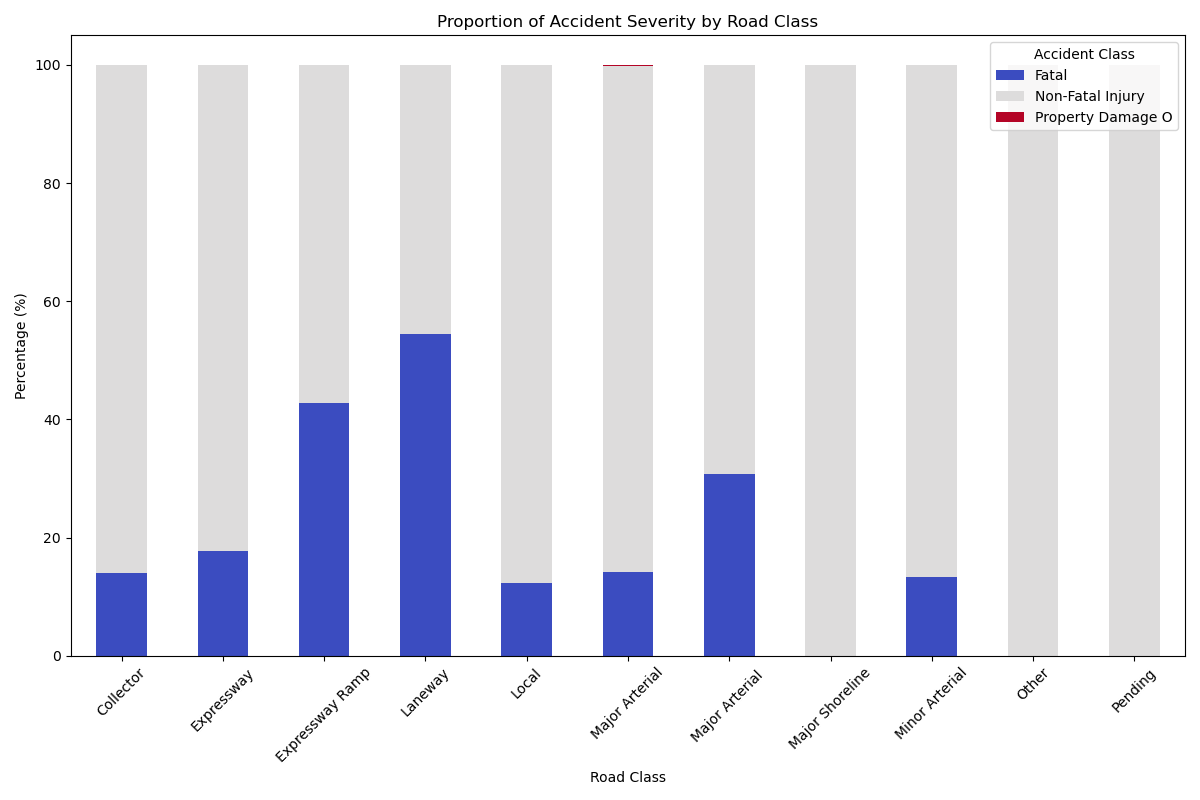
This horizontal bar chart displays the frequency of collisions under different light conditions. As expected, the majority of collisions occur during **"Daylight"**. However, a substantial number of incidents also happen in **"Dark"** conditions, both with and without artificial lighting, indicating that visibility due to ambient light is a significant factor in collision occurrences.

**3. Number of Collisions by Impact Type**

****

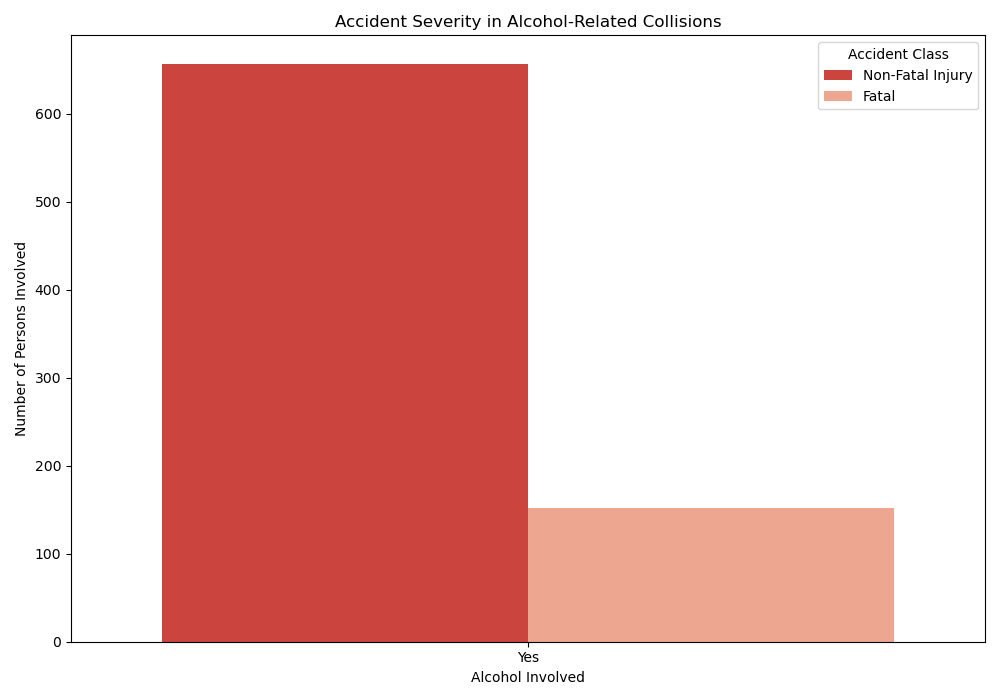
This chart categorizes collisions based on their initial impact type. It clearly shows that **"Pedestrian Collisions"** are the most common type of incident, followed by "Turning Movement" and "Cyclist Collisions." This suggests that interactions between vehicles and vulnerable road users, as well as complex maneuvers, are major contributors to KSI events in Toronto.

**4. Proportion of Accident Severity by Road Class**

****

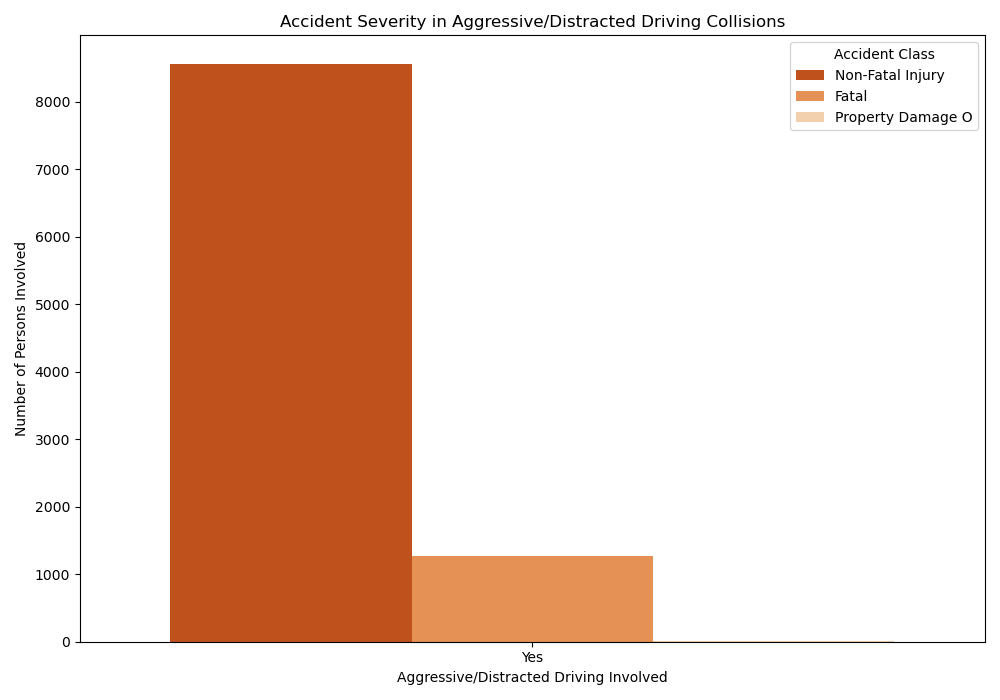
This stacked bar chart breaks down the proportion of accident severities (Fatal, Non-Fatal Injury, Property Damage Only) across different road classifications. While "Non-Fatal Injury" dominates across all road classes, the chart allows for an examination of the relative proportion of "Fatal" collisions on different road types. For example, **"Major Arterial"** roads account for a significant absolute number of collisions, and this chart helps understand the *proportion* of severe outcomes on these and other road classes.

**5. Accident Severity in Alcohol-Related Collisions**

****

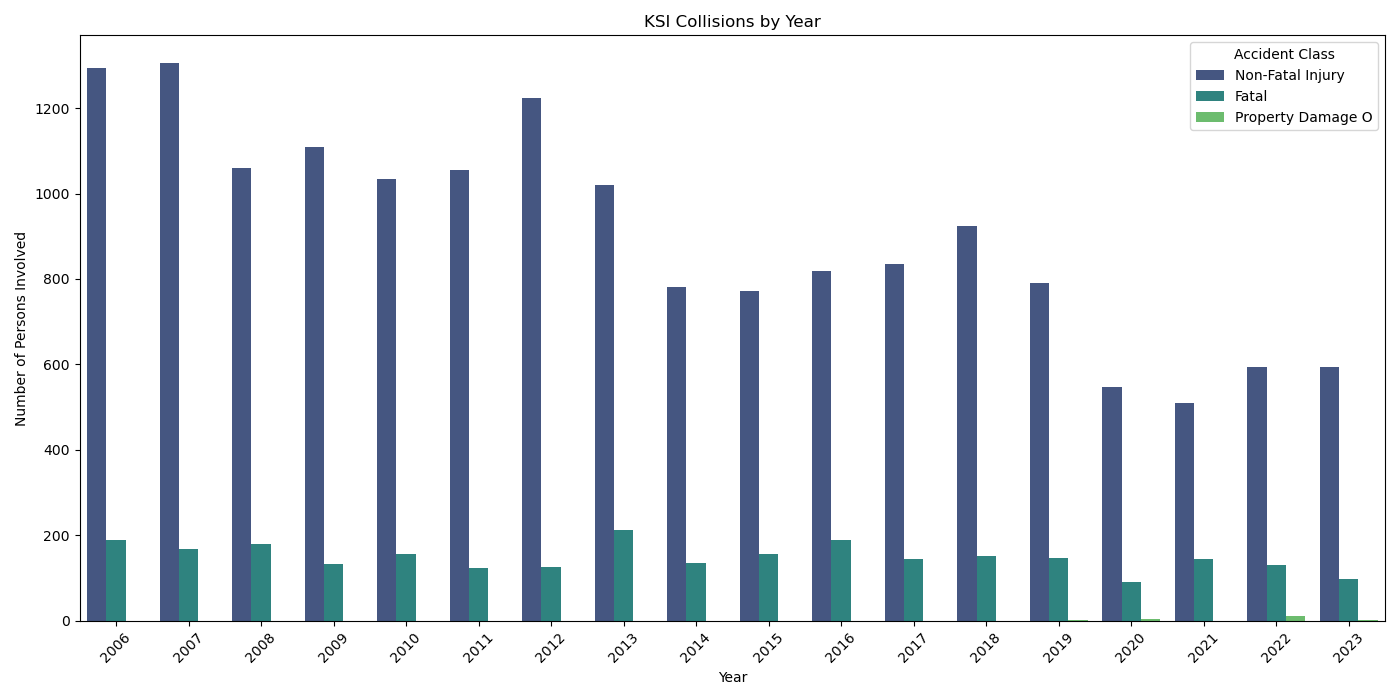
This grouped bar chart specifically examines the distribution of accident severity (Non-Fatal Injury vs. Fatal) for collisions where alcohol was a contributing factor. It demonstrates that while "Non-Fatal Injury" collisions are more prevalent even in alcohol-related incidents, a notable number of **fatalities** are associated with alcohol involvement, underscoring its severe impact on road safety.

**6. Accident Severity in Aggressive/Distracted Driving Collisions**

****

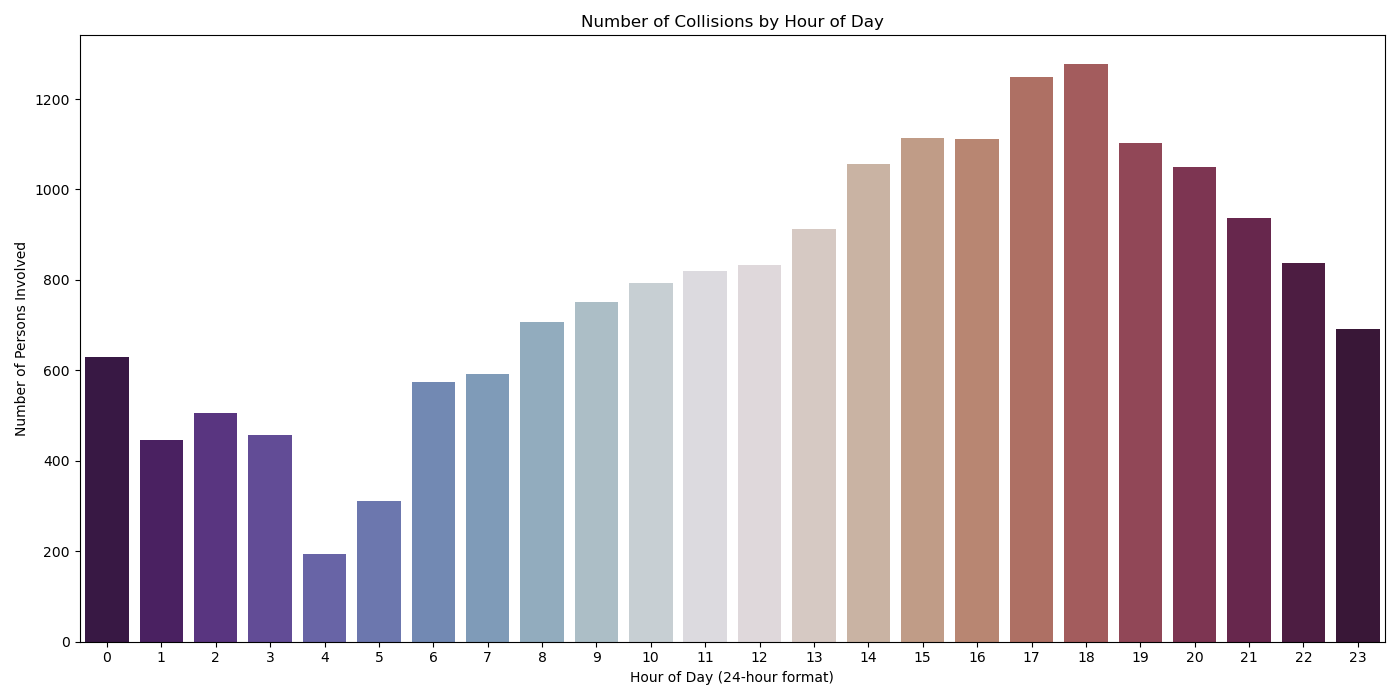
Similar to alcohol-related collisions, this chart illustrates the breakdown of accident severity for incidents involving aggressive or distracted driving. It highlights that a very large proportion of collisions, including a significant number of **fatal** ones, are linked to aggressive/distracted driving behaviors. This indicates that such behaviors are a major contributing factor to severe outcomes.

**7. KSI Collisions by Year**

****

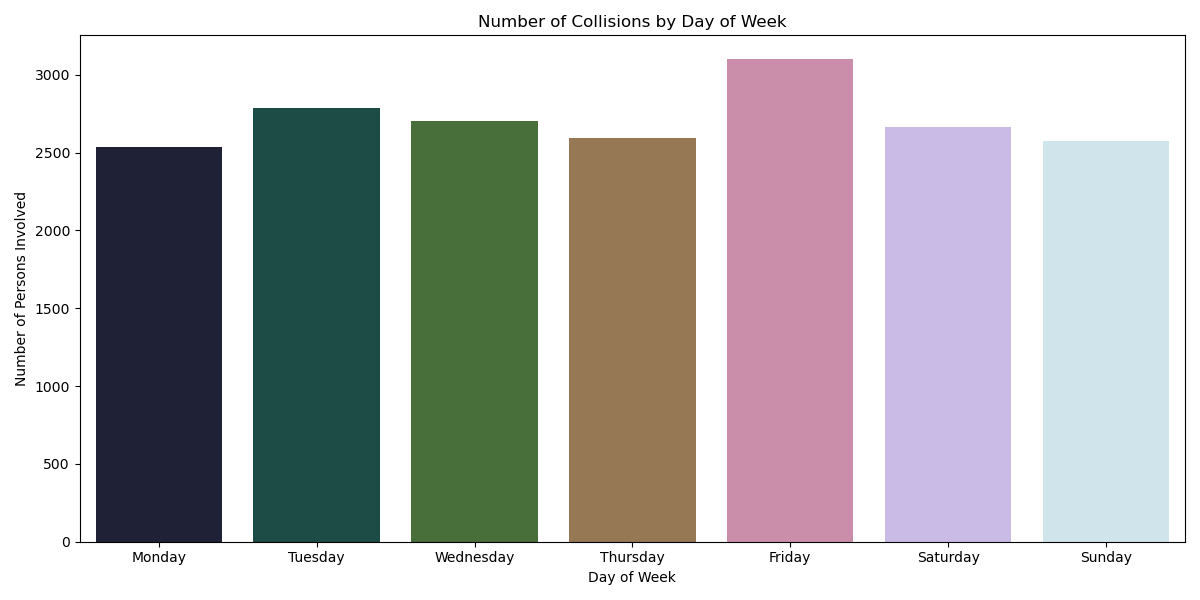
This line chart shows the trend of Killed or Seriously Injured (KSI) collisions in Toronto from 2006 to 2023, broken down by accident class. It provides a historical perspective on collision frequency and severity over time, allowing for the identification of periods with higher or lower incident rates. Trends in "Non-Fatal Injury" and "Fatal" collisions can be observed, revealing changes in road safety over nearly two decades.

**8. Number of Collisions by Hour of Day**

****

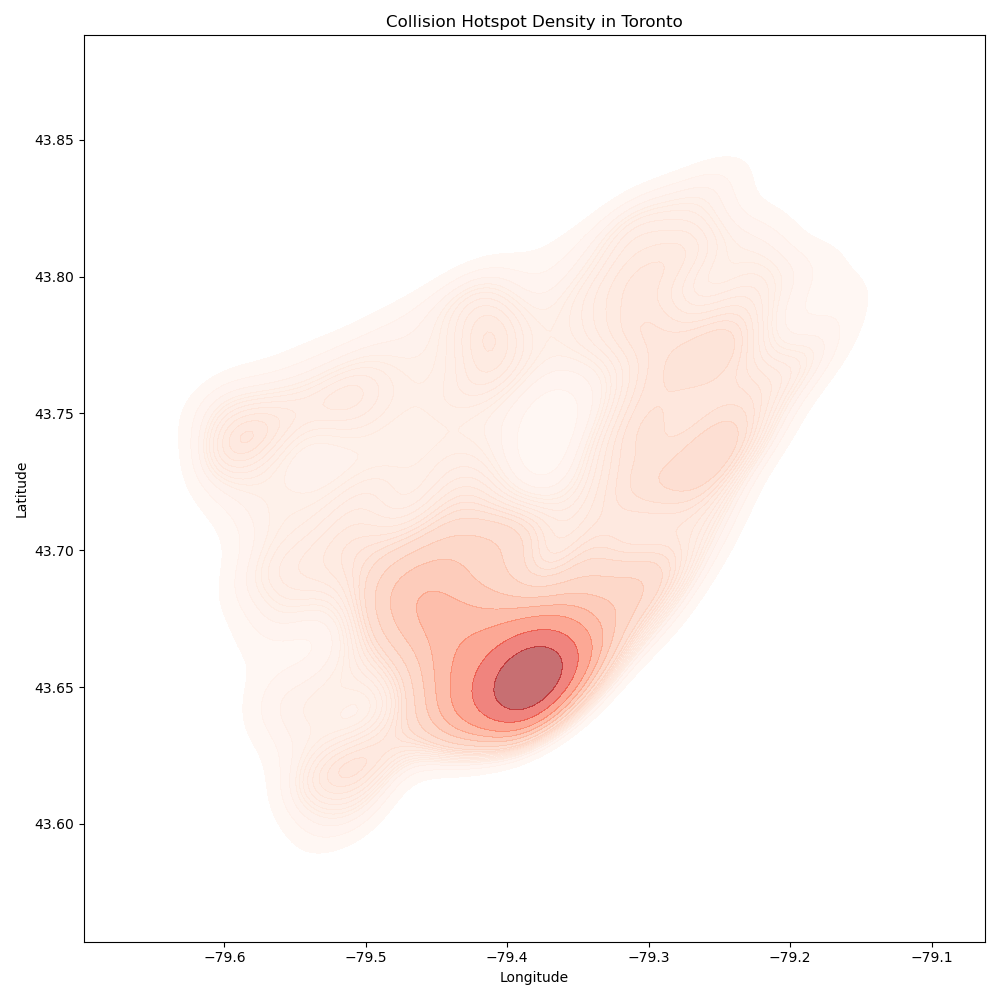
This bar chart illustrates the hourly distribution of collisions throughout a 24-hour day. It reveals clear peaks in collision frequency during typical rush hour periods, particularly in the **afternoon/evening (e.g., 4 PM to 7 PM)**, corresponding to increased traffic volume during commutes. There is also a smaller peak in the morning.

**9. Number of Collisions by Day of Week**

****

This bar chart displays the total number of collisions for each day of the week. It shows that collisions are relatively consistent across weekdays but often see a slight increase towards the end of the week, with **Friday** frequently exhibiting the highest number of incidents. Weekends generally show slightly fewer collisions compared to weekdays, though still substantial.

**10. Collision Hotspot Density in Toronto**

****

This Kernel Density Estimate (KDE) plot visually represents the geographical areas in Toronto with the highest concentration of collisions, indicating "hotspots." The darker, more intense red areas correspond to locations where collisions occur most frequently. This map is crucial for identifying specific high-risk zones within the city, which can inform targeted safety interventions and urban planning.

**Phase 2**

**Date: August 7, 2025**

**1. Executive Summary**

This report details the end-to-end development of a machine learning model to predict the severity of traffic accidents in Toronto, using the Killed or Seriously Injured (KSI) dataset. The primary objective was to build a classifier that could effectively identify 'Fatal' accidents. This presented a significant challenge due to the severe class imbalance, where fatal accidents represent a very small fraction of the total incidents.

Through a rigorous, multi-stage process, the team successfully developed a **weighted voting ensemble model**. This final model combines a fine-tuned Logistic Regression classifier and a Random Forest classifier, leveraging the distinct strengths of each.

On unseen test data, the model achieved the project's primary goal: it correctly identified **57% of all fatal accidents (Recall = 0.57)**. This is a significant improvement over baseline models which struggled to identify even 15% of fatal cases, demonstrating the success of the project's data-driven approach to feature engineering and its sophisticated methods for handling class imbalance.

**2. Data Preparation and Feature Engineering (data\_processing.py)**

The foundation of this project's success was a meticulous data preparation pipeline that transformed the raw, complex dataset into a clean, feature-rich format suitable for machine learning.

**2.1. Data Aggregation: From People to Accidents**

The raw data contained one row for every person involved, which is not suitable for predicting the outcome of an accident as a whole. The first critical step was to change the unit of analysis from individual people to unique accidents.

* **Action Taken**: The data was grouped by a unique accident\_number.
* **Implementation**: A comprehensive aggregation\_dict was used to define how each column should be handled.
  + Static information (like location, time, and road conditions) was kept by taking the 'first' value in each group.
  + Dynamic, person-level information (like vehicle\_type, driver\_action) was aggregated into lists using a get\_all\_unique function. This correctly captured all actions and vehicle types involved in a single accident without duplication.
* **Outcome**: This produced a clean dataset where each row represents one unique accident, which is the correct structure for this prediction task.

**2.2. Feature Creation: Adding New Insights**

Several new, high-value features were engineered to provide the model with more direct and powerful signals.

* **Time-Based Features**: is\_weekend, season, and time\_of\_day were extracted from the original date and time columns. This allows the model to learn patterns related to weekly cycles, weather, and daily traffic patterns like rush hour.
* **Age-Based Counts**: The involvement\_age list was transformed into a set of numerical count features (age\_0\_14, age\_25\_59, etc.) and a total\_involved count. This is a much more effective way for the model to understand the demographic makeup of an accident.
* **Advanced Spatial & Interaction Features**:
  + A accident\_hotspot\_cluster feature was created using the DBSCAN clustering algorithm on the x and y coordinates. This proactively identifies high-risk geographical zones for the model.
  + Interaction features like bad\_weather\_and\_dark and reckless\_at\_intersection were created to explicitly combine different risk factors, providing a stronger signal to the model.

**2.3. Handling Categorical Data: From Text to Numbers**

A sophisticated, multi-step process was used to convert all categorical text data into a numerical format.

* **Feature Consolidation (Bundling)**: This was a standout part of the process. Instead of keeping dozens of granular categories, similar values were grouped together using mapping dictionaries. For example, all truck types (Pick Up Truck, Truck - Dump, etc.) were mapped to a single truck category. This crucial step reduced dimensionality, eliminated noise from rare categories, and created more robust features.
* **Multi-Label Binarization**: For features aggregated into lists, scikit-learn's MultiLabelBinarizer was used. This correctly created a separate binary column for each possible category, allowing a single accident to have multiple "true" values (e.g., involving both a vehicle\_type\_truck and a vehicle\_type\_passenger\_vehicle).
* **One-Hot Encoding**: For standard, single-label categorical features (like road\_class), pandas.get\_dummies was used to create binary columns.

**2.4. Feature Selection**

A logical feature selection process was applied to remove redundant or non-predictive columns:

* Redundant IDs (accident\_number, old neighborhood codes) were dropped.
* Features with high rates of missing data and low utility (offset) were removed.
* Data that could cause data leakage (injury\_severity\_score) was correctly removed before training.
* Redundant location data (latitude, longitude, street names) were dropped in favor of the more useful x and y coordinates.

**3. Modeling and Tuning Strategy (tune\_\*.py, build\_voting\_ensemble.py)**

A professional, modular workflow was adopted, separating data processing from model training and evaluation into distinct scripts.

* **Individual Model Tuning:** Separate scripts (tune\_logistic\_regression.py, tune\_random\_forest.py) were created to fine-tune each base model. This modular approach is efficient, organized, and a best practice.
* **Handling Imbalance:** The core challenge of the project was the severe class imbalance. This was expertly handled by:
  1. Using class\_weight='balanced' for the Logistic Regression model. This technique adjusts the model's loss function to penalize misclassifications of the minority class more heavily.
  2. Using a SMOTE (Synthetic Minority Over-sampling Technique) pipeline for the Random Forest model. This technique creates synthetic examples of the 'Fatal' class in the training data to create a more balanced dataset.
* **Rigorous Hyperparameter Tuning:** GridSearchCV and RandomizedSearchCV were used to find the optimal settings for each model. Critically, the search was configured to optimize for **recall of the 'Fatal' class**, ensuring the tuning process was aligned with the project's primary goal.
* **Reliable Evaluation:** StratifiedKFold was used for all cross-validation. This guarantees that each fold has a representative sample of the imbalanced classes, leading to reliable and unbiased performance metrics.

**4. Final Model Performance**

The final model is a VotingClassifier that combines the predictions of the two fine-tuned base models. To maximize performance, the ensemble was weighted to give more influence to the high-recall Logistic Regression model (weights=[0.7, 0.3]).

**Results of individual model:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Logistic Regression Model | Random Forest Model |  | Ensemble Model |
| **Recall (Fatal)** |  |  |  | 0.57 |
| **Precision (Fatal)** |  |  |  | 0.31 |
| **F1-Score (Fatal)** |  |  |  | 0.41 |
| **Recall (Non-Fatal Injury)** |  |  |  | 0.80 |
| **Precision (Non-Fatal Injury)** |  |  |  | 0.92 |
| **F1-Score (Non-Fatal Injury)** |  |  |  | 0.86 |

**Final Results on Unseen Test Data:**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Fatal** | **Non-Fatal Injury** |
| **Recall** | **0.57** | 0.80 |
| **Precision** | 0.31 | 0.92 |
| **F1-Score** | 0.41 | 0.86 |

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted: Fatal | Predicted: Non-Fatal |
| **Actual: Fatal** | 136 | 103 |
| **Actual: Non-Fatal** | 296 | 1181 |

Interpretation:

The final model correctly identifies 57% of all fatal accidents. This is a strong result that proves the viability of the approach. The confusion matrix shows that to achieve this, the model correctly flagged 136 fatal accidents while missing 103. The trade-off is a lower precision, resulting in 296 false alarms, which is an expected and often acceptable outcome for a high-recall system where missing a critical event is the worst-case scenario.

**5. Conclusion & Recommendations**

This project successfully demonstrates a complete and professional machine learning workflow. The team effectively navigated the challenges of a complex, imbalanced dataset to produce a model that provides real predictive value.

**Key Achievements:**

* Transformed raw, messy data into a clean, feature-rich dataset of over 200 features.
* Correctly identified and addressed the critical issue of class imbalance using multiple techniques.
* Systematically tuned and evaluated multiple models to find the best performers.
* Successfully built a weighted ensemble model that achieved a **Fatal Recall of 57%**.

**Recommendations for Future Work:**

1. **Improve the Weaker Ensemble Member:** The ensemble is still limited by the overfitted Random Forest model. Further tuning (e.g., reducing max\_depth, increasing min\_samples\_leaf) could improve its generalization and boost the ensemble's overall performance.
2. **Add more, Diverse Model:** Consider tuning and adding a LightGBM or XGBoost classifier to the ensemble. A third "voter" from a different algorithm family (boosting) could further improve the model's stability and accuracy.
3. **Refine Feature Selection:** After tuning, a feature importance analysis could be run to remove the least important features, potentially simplifying the model and reducing noise without sacrificing performance.