Data Wrangling and Exploration on Toronto's Pedestrian collisions and Injured Persons Data

Dr. Eric Asare

2023-01-08

1. Background

A bad data set which has errors is more likely to produce unreliable insights. This exercise is grouped into two main sections, which are data wrangling and data exploration. The goal is to produce a reliable dataset which can be used to produce reliable insights for the Minitry of Transportation. SmartEDA is the main R package that will be used to wrangle and explore the raw data.

1.1 Data Wrangling

In this section, a) I will familiarize myself with the data (data discovery), b) clean the data by removing potential outliers, drop missing/empty information or replace them using imputation, and identify row and column duplicates, c) extract relevant columns or create new variables from existing columns, and d) validate the data by ensuring that columns contain consistent, accurate and expected range of values.

A. Data Discovery

The number of rows and columns of the collisions data are 504,069 and 12, respectively. Eight of the variables are non-numeric as expected. The involved persons data, which focuses pedestrian collisions, has 16,045 rows and 15 columns or variables. Most of the columns both data have significant amounts of missing (NA) or empty information (Table 1).

Table 1. Number of missing and empty information in the collision data

```
##             variable sum_missing sum_empty
## 1       collision_no           0         0
## 2            accdate           0        NA
## 3           district           0     39738
## 4           latitude          13        NA
## 5          longitude          13        NA
## 6        impact_type           0       276
## 7              light           0       381
## 8     location_class           0    216939
## 9  road_surface_cond           0       326
## 10        road_class           0     34711
## 11   traffic_control           0      1564
## 12        visibility           0      3693
```

Table 1. Number of missing and empty information in the involved persons data

```
##                      variable sum_missing sum_empty
## 1                collision_no           0         0
## 2                      rec_id           0         0
## 3                     accdate           0        NA
## 4                actual_speed       14830        NA
## 5            driver_condition           0     15870
## 6             impact_location           0      4192
## 7       involved_injury_class           0         6
## 8                involved_age         384        NA
## 9              involved_class           0         0
## 10                  manoeuver           0     15506
## 11          pedestrian_action           0      1126
## 12       pedestrian_condition           0       169
## 13 pedestrian_collision_type           0      1455
## 14               posted_speed       14110        NA
## 15              vehicle_class           0     12372
```

Summary For the involved persons data, the columns, actual_speed, driver_condition, posted_speed, impact_location, manoeuver, and vehicle_class, have significant missing or empty information problem. Similarly, the columns, district, location_class, road_class, in the collisions data have significant missing or empty information. I believe it will not be appropriate to fill the missing or empty information with imputation.

B. Data Cleaning

I generated a preprocess function (in the function script, R/functions.R) which is used to clean the data. First, I dropped all the columns suggested in the data discovery step above. For the remaining columns (with some amount of missing or empty information) I dropped them.
Subsequently, I checked for duplicate rows in both data. There were no duplicate rows in the collisions data. However, there where significant duplicates in the involved persons data. For each duplicate(s) in the involved persons data, I retained only one row.
Also, at in this step, I created the outcome variable (serious_fatal) from the involved persons data (specifically, the involved_injury_class column). The serious_fatal was converted to a factor variable which equals 1 if the involved_injury_class is fatal or major.

Summary

C. Data Transformation

First, I performed a left join operation the two data sets, using the collision_no column as the common id. The number of rows for the new data is 1,595. Also, I created 10 binary variables from the existing columns. The variables are:
1. age_65_plus: equals 1 if pedestrian's age is 65 or more and 0 otherwise. 2. impaired_pedestrian_condition: equals 1 if pedestrian has been drinking or have impaired ability from drugs or alcohol or have been drinking or have ability impaired with alchol over .80, and 0 otherwise.
3. pedestrian_inattentive: equals 1 if pedestrian condition is inattentive.
4. collision_midblock: equals 1 if pedestrian_collision_type is pedestrian hit at mid-block.
5. collision_inter_wrow: equals 1 if pedestrian_collision_type is vehicle going through intersection while pedestrian cross without row or vehicle turns left while pedestrian crosses without row at intersection or

vehicle turns right while pedestrian crosses without row at intersection.
6. collision_inter_row: equals 1 if pedestrian_collision_type is vehicle going through intersection while pedestrian cross with row or vehicle turns left while pedestrian crosses with row at intersection or vehicle turns right while pedestrian crosses with row at intersection.
7. collision_transit: equals 1 if pedestrian_collision_type is pedestrian involved in a collision with transit vehicle anywhere along roadway.
8. weather_bad: equals 1 if visibility is “fog, mist, smoke or dust” or strong wind or drifting snow or freesing rain.
9. no_traffic_control: equals 1 if traffic control is no control.
10. dark: equals 1 if light is dark artificial or dark or dusk or dusk artificial or dawn artificial or dawn.
Inaddition, dropped the other variables expect the outcome variable (serious_fatal), latitude and longitude.These variables will be used to estimate the probability that an accident will be fatal or major.

D. Data Validation

All the selected variables for model estimation have expected values, accurate and consistent. The problem is that the level, pedestrians with major or fatal collisions, is under represented in the data (11% of the total sample, Table 2). Model estimation with this data will produce bias and unreliable inference.
Table 2. Distribution of outcome variable

To solve this problem, I used the employed the performed a SMOTE on the data set to produce a relatively balanced outcome distribution. The new distribution (Table 3) is better than the previous one.

Table 3. Distribution of outcome variable post SMOTE

Summary s