# 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, and c) 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).

#### A.1. Collisions Data

```
##     Index     Variable_Name Variable_Type Sample_n Missing_Count Per_of_Missing
##  1:     1      collision_no       integer   504069             0          0.000
##  2:     2           accdate    IDate:Date   504069             0          0.000
##  3:     3          district     character   464331         39738          0.079
##  4:     4          latitude       numeric   504056            13          0.000
##  5:     5         longitude       numeric   504056            13          0.000
##  6:     6       impact_type     character   503793           276          0.001
##  7:     7             light     character   503688           381          0.001
##  8:     8    location_class     character   287130        216939          0.430
##  9:     9 road_surface_cond     character   503743           326          0.001
## 10:    10        road_class     character   469358         34711          0.069
## 11:    11   traffic_control     character   502505          1564          0.003
## 12:    12        visibility     character   500376          3693          0.007
##     No_of_distinct_values
```

```
##  1:                504069
##  2:                  3653
##  3:                     6
##  4:                118094
##  5:                125454
##  6:                    11
##  7:                    10
##  8:                    13
##  9:                    11
## 10:                    15
## 11:                    13
## 12:                     9
```

### A.2. Injured Persons Data

```
##    Index             Variable_Name  Variable_Type Sample_n Missing_Count
## 1      1              collision_no        integer    16045             0
## 2      2                    rec_id        integer    16045             0
## 3      3                   accdate POSIXct:POSIXt    16045             0
## 4      4              actual_speed        integer     1215         14830
## 5      5          driver_condition      character      175         15870
## 6      6           impact_location      character    11853          4192
## 7      7     involved_injury_class      character    16039             6
## 8      8              involved_age        integer    15661           384
## 9      9            involved_class      character    16045             0
## 10    10                 manoeuver      character      539         15506
## 11    11         pedestrian_action      character    14919          1126
## 12    12      pedestrian_condition      character    15876           169
## 13    13 pedestrian_collision_type      character    14590          1455
## 14    14              posted_speed        integer     1935         14110
## 15    15             vehicle_class      character     3673         12372
##    Per_of_Missing No_of_distinct_values
## 1           0.000                15178
## 2           0.000                16045
## 3           0.000                 3445
## 4           0.924                   36
## 5           0.989                    4
## 6           0.261                   14
## 7           0.000                    6
## 8           0.024                  100
## 9           0.000                    1
## 10          0.966                   13
## 11          0.070                   16
## 12          0.011                   11
## 13          0.091                   17
## 14          0.879                   12
## 15          0.771                   17
```

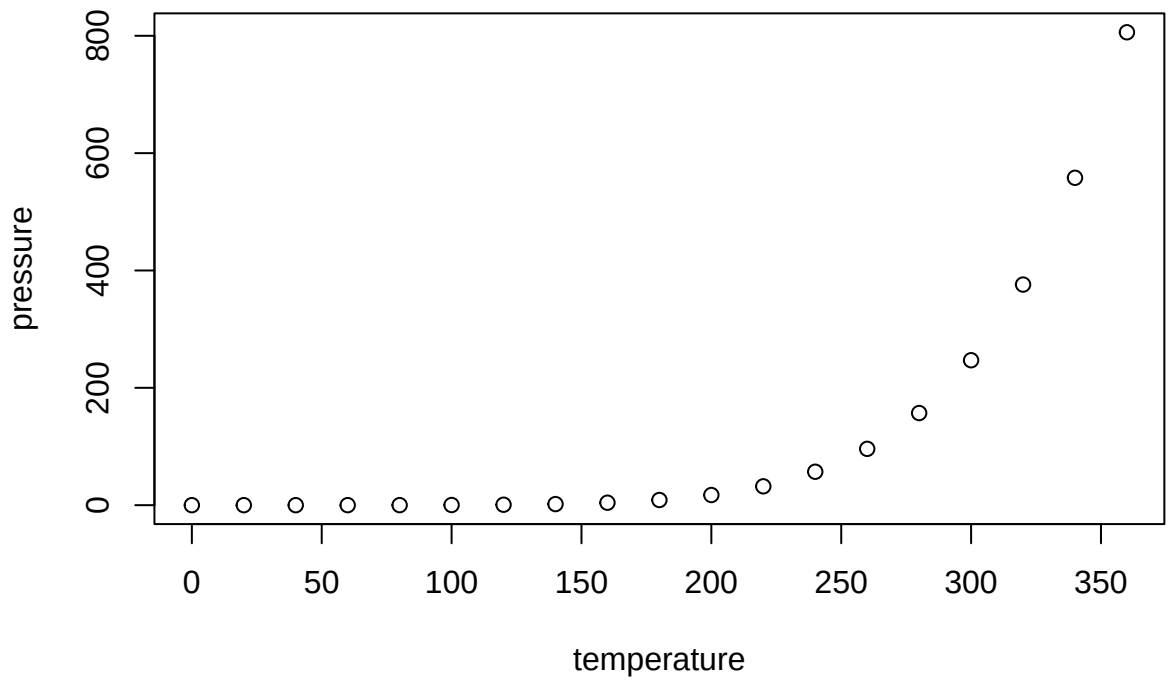
or this project, data wrangling encompasses activities such as:
* unordered list + sub-item 1 + sub-item 2 - sub-sub-item 1
* It is a problem because institutions
* hhh
who rely on such information for policy recommendations could lose lots of resources, including money,

unreliable information from data analytics Unreliable insights from This data exercise will be grouped into data wrangling and data exploration. When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

## Including Plots

You can also embed plots, for example:




Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.