

# Data Exploration Project

Xianlin Feng

April 15, 2019

## 1 Introduction

Road traffic accident is a threat to all people in their daily life. According to the WHO's statistics in 2018, road traffic accidents are the eighth of the top 10 causes of death, which is the only reason of injuries, and all the remaining reason are diseases. In the worldwide, road injuries took 140 million lives in 2016, in which 74% are were men and boys. (<https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>). Serious situation happened in Victoria too. There were 58 lives lost because of the road accident every day in 2018 just in Victoria. This number increased to 88 since 7 April 2019. In the last five years, the least worse situation happened in 2017, the daily lives lost was 64. According to the TAC report, the lives lost of drivers takes nearly half (48.2%) of the daily lives lost. The age of death is concentrated in 30-69 years old. The most lives lost in rural roads(58%).(<http://www.tac.vic.gov.au/road-safety/statistics/lives-lost-year-to-date>). The analysis of past traffic accidents can provide a basis for future road construction, accident prevention, and accident rescue. The following three question will be answered in this report:

1. What is the main cause of road crashes in Victoria.
2. What is the trend in the last 10 years?
3. What suggestion we can provide for the people in different areas.

To answer those three questions, I will try to analysis and explore the dataset to find the main cause of road crashes in Victoria, as well as the trend in the last 10 years. I will try to find one or more datasets, then perform data wrangling, data cleaning and data checking before data exploration. During the data exploration, I will perform different statistic tests and visualisation to explore the data set and obtain insight from the data. At last, I will provide some suggestions base on the result of the data exploration. The main structure of this report is as follow: the data wrangling will be processed in section 2, then the data will be checking in section 3. The data exploration will be carried out in section 4, followed by the section of conclusion. In the last section, there will be a reflection of this assessment.

{ draft }

## 2 Data Wrangling

### 2.1 the Dataset

I found the data set named "CrashStats data" on Victoria government website for open data. The dataset was provide by VicRoad for educational purposes, which includes the crash data of time, location, conditions and so on since 2000. It could be download from the website <https://www.data.vic.gov.au> and the dataset is consist with 12 tables in the following list:

1. **accident**: contain the basic information about the accident, such as date, time, location, environment codition and severity.
2. **vehicle**: vehicle information, such as make, body type, year of manufacture, fuel type, vechicle capacity and so on.
3. **person**: person details, such as age, sex, sitting position, passengers or driver, license state etc.
4. **accident\_event**: the sequence of events during the accident, such as ran off carriageway, collision, fell from vehicle, and so on.
5. **accident\_location**: the location information of the accident.
6. **road\_surface\_cond**: the codition of the road: wet, dry, or icy.
7. **atmospheric\_cond**: weather condition: clear, wind, dust and so on.
8. **sub\_dca**: describing the crash detail with code.
9. **accident\_node**: more detailed location about the crash.
10. **accident\_chainage**: chainage information of the node.
11. **node\_id\_complex\_int\_id**: if the node locate in a complex intersection or not.
12. **statistic checks**: the statistic information of the crashes in this dataset.

Before the data wrangling and data cleaning, we need to understand the characteristics of each data table, for example, in the table of "accident", each accident has exact one record. But, as one accident may involved two or more people, one accident may have two or more records in the table "person". Same situation happens in table "vehicle", "accident\_event" and other tables too, this is another reason why this dataset separate to 12 data tables. Due to this reason, when we process data wrangling, data cleaning, or data exploration, extra careful should be paid with the multiple records for one accident.

## 2.2 Delete expired data

The crash happened more ten years ago is not considered in this report, so they will be delete form the dataset in the first step. Fortunately, every crash in the dataset have a accident number, the accident number consist with the date information and the index number, so the instances before 1/1/2009 and after 31/12/2018 should be delete first in each subset.

**data process pipline**

## 3 Data Checking

### 3.1 Checking with Table3au Public

After data wrangling, the data should be check to eliminate outlier. It is the process to detect and correct the inaccurate, incomplete, incorrect, or irrelevant part from the dataset. After The data exploration process can only be implemented after the data check, otherwise, the wrong data source will definitely lead the wrong results and conclusions. Data

## 4 Data Exploration

## 5 Conclusion

## 6 Reflection

In this section, we define some important definitions and algorithms....

**Definition 6.1.** (Mixed Integer Programming)

In this report we consider a generic mixed-integer programming problem (MIP) in the following form

$$\begin{aligned} \text{(MIP)} \quad & \min c^T x \\ \text{s.t.} \quad & Ax \geq b \\ & x_j \in \mathbb{Z} \quad \forall j \in \mathcal{I} \\ & x_j \in \mathbb{R} \quad \forall j \in \mathcal{N} \setminus \mathcal{I} \end{aligned}$$

where the vector  $b \in \mathbb{R}^m$  and the vector  $c \in \mathbb{R}^n$  are input vectors.  $A$  is a input matrix of size  $m \times n$ , the variable input set  $\mathcal{I} \subseteq \mathcal{N} = \{1, 2, \dots, n\}$ . We denote  $\mathcal{P}$  for this problem, which called a mixed-integer programming problem (MIP) with minimize objective function  $c^T x$  subject to the

constraints  $Ax \geq b$ . Besides, some variables are restricted to integer values while the else of are restricted to real value.  $S$  is a set of feasible solution if  $S$  satisfy all the constraints in the problem. A vector  $s^*$  with  $s^* \in S$  is called *optimal solution* when  $c^T x_{s^*} \leq c^T x_s$  for  $\forall s \in S$ . When all of the variables are restricted to integer, the problem is called *pureintegerlinearprogram*(IP) for  $\mathcal{I} = \mathcal{N}$ . If there is no integrality constraint, the program is called *linearprogram*

$$\begin{aligned} \text{(MIP)} \quad & \min \quad c^T x \\ & s.t. \quad Ax \geq b \\ & \quad \quad x_j \in \mathbb{R} \quad \forall j \in \mathcal{N} \end{aligned}$$

**Definition 6.2.** (LP-relaxation)

Lp *relaxation* is obtained by removing all integrity constraints  $\mathcal{I} \leftarrow \emptyset$ . LP-*relaxation* is the foundation of LP-based branch-and-bound technology. As the searching space is increase by removing integrity restrictions, the optimal solution in MIP problem could not better than LP-*relaxation*, which is  $s_{MIP}^* \geq s_{LP}^*$ . This means the optimal solution found in LP problem could provide a lower or prime bound for MIP problem.

## 7 Input

- A MIP problem  $\mathcal{P}^0$  with  $n$  variables  $x$ , constraint set  $C^0$  with an optimal solution  $s^0$ , where  $s^0$  is a  $n$ -vector.
- A MIP problem  $\mathcal{P}^1$  with  $n$  variables  $x$ , constraint set  $C^1$ , such that  $C^0 \subsetneq C^1$ .

## 8 Output

- An optimal solution  $s^1$  to  $\mathcal{P}^1$ , where  $s^1$  is a  $n$ -vector too.

## 9 Pseudo Code

---

### Algorithm 1: Solving Problem with Reoptimization

---

**Input:**  $\mathcal{P}^1$  where  $C^0 \subsetneq C^1$  and  $s^0, k$

**Output:** optimal solution  $s^*$  to  $\mathcal{P}^1$

---

```

1 begin
2   if  $s^0$  is feasible to  $\mathcal{P}^1$  then
3     return  $s^0$ 
4   else
5      $\mathcal{I} \leftarrow$  index set of integer or binary variables in  $\mathcal{P}^1$ 
6     for  $i$  in  $\mathcal{I}$  do
7        $\mathcal{P}^2 \leftarrow$  create new variables  $y_i$  and add it to  $\mathcal{P}^1$ :           // add new variables
8        $\mathcal{P}^2 \leftarrow$  add new constraints to  $\mathcal{P}^2 : y_i \geq x_i - s_i^0$        // add new constraints
9        $\mathcal{P}^2 \leftarrow$  add new constraints to  $\mathcal{P}^2 : y_i \geq s_i^0 - x_i$ 
10    end
11    if the sense of  $\mathcal{P}^2$  is not minimize then
12      change the sense of  $\mathcal{P}^2$  to minimize
13    end
14    stop gap  $\leftarrow 0.5$ 
15    for  $l$  in  $\{k, k-1, \dots, 0\}$  do
16       $\alpha \leftarrow \alpha \times l$ 
17      the coefficients of variables  $y$  in  $\mathcal{P}^2 \leftarrow \alpha$ 
18      if  $l = 0$  then
19        stop gap  $\leftarrow 0.0$ 
20      end
21       $s^* \leftarrow$  solving the sub-MIP problem to stop gap with reoptimization
22    end
23    return  $s^*$ 
24  end
25 end

```

---