# Data Exploration Project

Xianlin Feng

April 8, 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 Victorial too. There were 58 lives lost because of the road accident every day in 2018 just in Victorial. This number increased to 88 since 7 April 2019. In the last five years, the least worse situation happed 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. To achieve this goal, I found CrashStats data on Victorial government website for open data. The dataset was provide by VicRoad. The dataset includes the crash data of time, location, conditions and so on since 2000. In this report, 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. in the last part, I will provide some suggestions base on the result of the analysis. 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.

# 2 Data Wrangling

## 2.1 the Dataset

## 2.2 Delete outdate data

The crash happed 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 number include the date information, so I could just delete the instances before 1/1/2009 and after 31/12/2018 in each subset.

# 3 Data Checking

# 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

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

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 $\mathcal{I} \subseteq \mathcal{N} = \{1, 2, \ldots, n\}$. 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*

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

**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 $\mathcal{P}^1$

**begin**
  **if** *$s^0$ is feasible to $\mathcal{P}^1$* **then**
    **return** $s^0$
  **else**
    $\mathcal{I} \longleftarrow$ index set of integer or binary variables in $\mathcal{P}^1$
    **for** *$i$ in $\mathcal{I}$* **do**
      $\mathcal{P}^2 \longleftarrow$ create new variables $y_i$ and add it to $\mathcal{P}^1$:  `// add new variables`
      $\mathcal{P}^2 \longleftarrow$ add new constraints to $\mathcal{P}^2$ : $y_i \geq x_i - s_i^0$  `// add new constraints`
      $\mathcal{P}^2 \longleftarrow$ add new constraints to $\mathcal{P}^2$ : $y_i \geq s_i^0 - x_i$
    **end**
    **if** *the sense of $\mathcal{P}^2$ is not minimize* **then**
      change the sense of $\mathcal{P}^2$ to minimize
    **end**
    stop gap $\longleftarrow 0.5$
    **for** *$l$ in $\{k, k-1, \cdots, 0\}$* **do**
      $\alpha \longleftarrow \alpha \times l$
      the coefficients of variables $y$ in $\mathcal{P}^2 \longleftarrow \alpha$
      **if** *$l = 0$* **then**
        stop gap $\longleftarrow 0.0$
      **end**
      $s^* \longleftarrow$ solving the sub-MIP problem to stop gap with reoptimization
    **end**
    **return** $s^*$
  **end**
**end**