The distribution of crime in Victoria

Crime Reports for Different Ages

Ruiyang Fu - s3679150

Last updated: 29 May, 2022

RPubs link information

http://rpubs.com/Ezio/907927

Introduction

The figures are compiled by the Crime Statistics Agency (CSA), which looks at the number of crimes recorded in Victoria from 2012 to 2021.The crime statistics recorded are based on data extracted by Victoria Police on the 18th day after the reference period and are subject to change between releases. Person-based counts with a value of 1 to 3 in this data are given a value of 2 to calculate the total in order to keep the data confidential. The main crime types in the data table are grouped according to the CSA crime classification.

Problem Statement

In this report, I assume that the number of crimes is distributed as follows: 10% of people aged 10-17, 30% of people aged 18-24, and 60% of people over 25 years old I will prove the positive and negative of the hypothesis by using the chi-square test after preprocessing the data

Data

I will use the Table 08 data from the Data_Tables_Alleged_Offender_Incidents_Visualisation_Year_Ending_December_202 table. The website link is : https://discover.data.vic.gov.au/dataset/data-tables-alleged-offender-incidents This data mainly includes the following variables: Year: year of record Year ending: month of record Offence Division: The description of the type of crime mainly includes the following six: Crimes against the person, Property and deception offences, Drug offences, Public order and security offences, Justice procedures offences, Other offences Outcome:The specific processing results after the crime mainly include the following four: Arrest, Not authorised, Other, Summons Age Group: age range of offenders and total Alleged Offender Incidents: Number of Alleged Offender Incidents and total

preprocess the data

read data

```
# Import the data
VictimReports <- read_xlsx("Data_Tables_Alleged_Offender_Incidents_Visualisation_Year_Ending_December_2021.xlsx", sheet =
        "Table 08")
# Checking the VictimReports data head
head(VictimReports)
```

```
# Checking the VictimReports data tail
tail(VictimReports)
```

preprocess the data Cont.

Because we don't need to analyze the total number, use the filter function to filter the total people in the Age Group variable The variable name at the beginning has special symbols and is not easy to be manipulated, so use the rename function in the plyr package to rename the variable At the same time, I use the rules in the editrules package to define my rules. Here I mainly check whether the NumberOfVictim variable has data that does not meet the rules.

```
# filter data
VictimReports <- VictimReports %>% filter(`Age Group` != "Total people")
# Count how many rows have NA values
library(plyr)
VictimReports <- rename(VictimReports, c(`Year ending` = "Month"))
VictimReports <- rename(VictimReports, c(`Offence Division` = "OffenceDivision"))
VictimReports <- rename(VictimReports, c(`Age Group` = "ageGroup"))
VictimReports <- rename(VictimReports, c(`Alleged Offender Incidents` = "NumberOfVictim"))
# The reason why plyr is canceled here is because many functions in plyr and dplyr conflict.
# If it is not canceled, the next functions will call the plyr package, which will cause code errors.
detach("package:plyr", unload = TRUE)
sum(is.na(VictimReports))
```

```
## [1] 0
```

```
VictimReports <- na.omit(VictimReports)
library(editrules)
(Rule1 <- editset(c("NumberOfVictim >= 0")))
```

```
##
## Edit set:
## num1 : 0 <= NumberOfVictim
```

```
sum(violatedEdits(Rule1, VictimReports))
```

```
## [1] 0
```

```
detach("package:editrules", unload = TRUE)
```

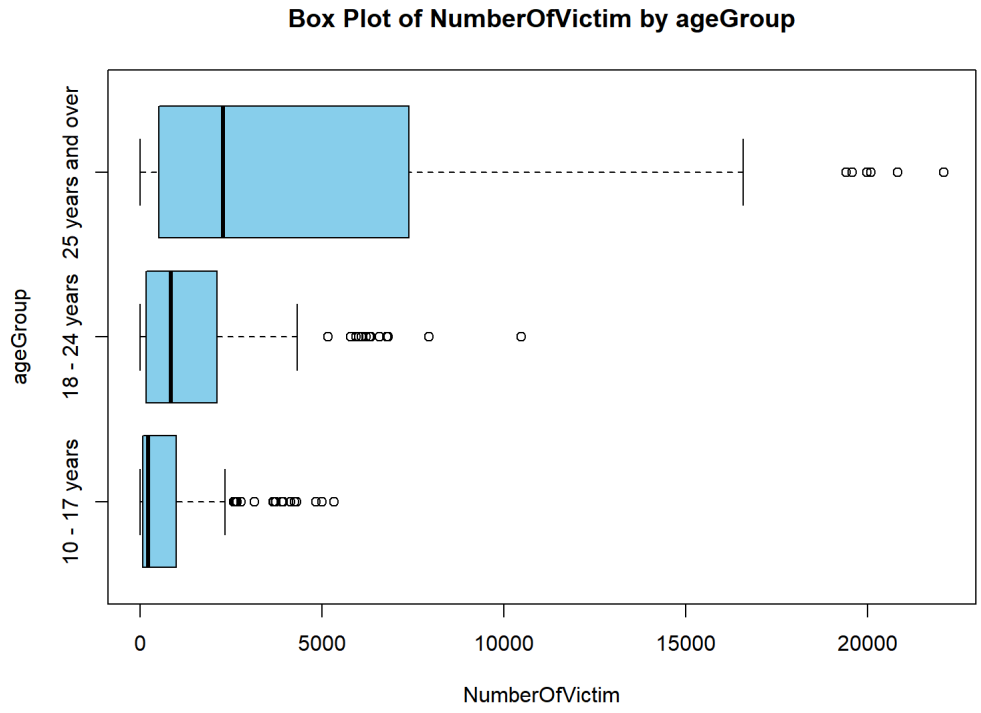
Descriptive Statistics and Visualisation

```
VictimReports %>% group_by(ageGroup) %>%
  summarise(Min = min(NumberOfVictim,na.rm = TRUE),
            Q1 = quantile(NumberOfVictim,probs = .25,na.rm = TRUE),
            Median = median(NumberOfVictim, na.rm = TRUE),
            Q3 = quantile(NumberOfVictim,probs = .75,na.rm = TRUE),
            Max = max(NumberOfVictim,na.rm = TRUE),
            Mean = mean(NumberOfVictim, na.rm = TRUE),
            SD = sd(NumberOfVictim, na.rm = TRUE),
            n = n(),
            Missing = sum(is.na(NumberOfVictim))) -> table1
knitr::kable(table1)
```

ageGroup	Min	Q1	Median	Q3	Max	Mean	SD	n	Missing
10 - 17 years	2	78.00	225.0	998.50	5340	753.000	1058.667	239	0
18 - 24 years	2	175.75	850.5	2116.25	10475	1429.442	1680.995	240	0
25 years and over	5	525.00	2274.5	7381.50	22101	4284.350	4623.273	240	0

Decsriptive Statistics Cont.

```
VictimReports %>% boxplot(NumberOfVictim ~ ageGroup, data = ., main = "Box Plot of NumberOfVictim by ageGroup",
                          ylab = "ageGroup", xlab = "NumberOfVictim", horizontal = TRUE, col = "skyblue")
```




```
Victim10 <- VictimReports %>% filter(ageGroup == "10 - 17 years")
Victim18 <- VictimReports %>% filter(ageGroup == "18 - 24 years")
Victim25 <- VictimReports %>% filter(ageGroup == "25 years and over")
summary(Victim10$NumberOfVictim)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    78.0   225.0   753.0   998.5  5340.0
```

```
summary(Victim18$NumberOfVictim)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0   175.8   850.5  1429.4  2116.2 10475.0
```

```
summary(Victim25$NumberOfVictim)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       5     525    2274    4284    7382   22101
```

```
Victim10_clean <- Victim10 %>% filter(NumberOfVictim < 998.5 + ((988.5 - 78.0) * 3))
Victim18_clean <- Victim18 %>% filter(NumberOfVictim < 2116.2 + ((2116.2 - 175.8) * 3))
Victim25_clean <- Victim25 %>% filter(NumberOfVictim < 7382 + ((7382 - 525) * 3))
```

Hypothesis Testing

The total amount of data is counted here, which is convenient for us to carry out the next chi-square test, and a data table is created according to the statistical data.

```
total <- sum(Victim10_clean$NumberOfVictim) + sum(Victim18_clean$NumberOfVictim) + sum(Victim25_clean$NumberOfVictim)
sum(Victim10_clean$NumberOfVictim)/total
```

```
## [1] 0.09409274
```

```
sum(Victim18_clean$NumberOfVictim)/total
```

```
## [1] 0.2173813
```

```
sum(Victim25_clean$NumberOfVictim)/total
```

```
## [1] 0.688526
```

```
total * 0.1
```

```
## [1] 149339.9
```

```
total * 0.3
```

```
## [1] 448019.7
```

```
total * 0.6
```

```
## [1] 896039.4
```

```
(140518 - 149340)^2 / 149340 + (324637 - 448019)^2 / 448019 + (1028224 - 896039)^2 / 896039
```

```
## [1] 54000
```

Information

NumberOfVictim	10-17 years	18-24 years	over 25 years
Observed	140518	324637	1028224
Proportion	0.094	0.217	0.689
Expected	149340	448019	896039
pi	0.1	0.3	0.6

Hypthesis Testing

H0: The crime distribution is that 10-17 years old account for 10% of crimes, 18-24 years old account for 30% of crimes, and over 25 years old account for 60% of crimes

H1: The crime distribution is not 10% of the crimes for 10-17 years old, 30% of crimes for 18-24 years old, and 60% of crimes for people over 25 years old

$$x^2 = \sum (Obs - Exp)^2 / Exp$$

$$x^2 = (140518 - 149340)^2/149340 + (324637 - 448019)^2/448019 + (1028224 -$$

Chi-square.

Here we use the chi-square test, create the data frame, and check the observed vs expected value

```
pop_prop <- c(0.1, 0.3, 0.6)
df <- data.frame("years(10-17)" = c(140518), "years(18-24)" = c(324637), "years(over25)" = c(1028224))

chil <- chisq.test(df, p = pop_prop)
chil$observed
```

```
## [1]  140518  324637 1028224
```

```
chil$expected
```

```
## [1] 149337.9 448013.7 896027.4
```

Discussion

A chi-square goodness-of-fit test was used to determine whether the distribution of crime followed a distribution of 10% from people aged 10-17, 30% from people aged 18-24, and 60% from people over 25. The test was statistically significant, $\chi 2 = 5400$, $df = 2$, $p < .001$. This indicates that the distribution of M&M colors does not follow the assumed distribution. Because some abnormal data are deleted, some data may be inaccurate, and for real situations, some data may be missing At the same time, this data is from CSA, and the data may change after the release And to keep the data confidential, people-based counts with values 1 to 3 were given a value of 2 to count the total, all of which created limitations in the data. For future investigations, changes in the amount of data should be tracked and analyzed in a timely manner.Through data analysis, we can clearly find that the entire crime distribution is not distributed according to 10%, 30%, and 60%. At the same time, it can be clearly found that people over the age of 25 will be more likely to commit crimes. We may need to Curb crime declines by giving more attention to people over 25 from a life and work perspective

References

https://discover.data.vic.gov.au/dataset/data-tables-alleged-offender-incidents