Assignment 1 ETC1010 - 5510

New South Wales Crime Incidents Report

Karan Garg

Monday, March 22 2021

Instructions to Students

This assignment is designed to simulate a scenario in which you are taking over someone's existing work and continuing with it to draw some further insights.

This is a real world dataset taken from the New South Wales Bureau of Crime Statistics and Research. The data can be found here at https://www.bocsar.nsw.gov.au/Documents/Datasets/SuburbData.zip. Specifically, the data file called "SuburbData2019csv" located in your data folder inside the RStudio project will be used for this assignment.

You have just joined a consulting company as a data scientist. To give you some experience and guidance, you are performing a quick summary of the data while answering a number questions that the chief business analytics leader has. This is not a formal report, but rather something you are giving to your manager that describes the data with some interesting insights.

Please make sure you read the hints throughout the assignment to help guide you on the tasks.

The points allocated for each of the elements in the assignment are marked next to the code for each question.

Marking + Grades

- This assignment will be worth **10%** of your total grade, and is marked out of 116 marks total. **Due on: Friday 26 March**.

For this assignment, you will need to upload the following into Moodle:

```
- Your Rmd file,
- The rendered html file, and
- The PDF rendered file.
```

How to find help from R functions?

Remember, you can look up the help file for functions by typing: `?function_name`. For example, `?mean`. Feel free to google questions you have about how to do other kinds of plots, and post on the "Assignment Discussion Forum" any questions you have about the assignment.

How to complete this assignment?

To complete the assignment, you will need to fill in the blanks with appropriate function names, arguments, or other names. These sections are marked with `___`. **At a minimum, your assignment should be able to be "knitted"** using the `Knit` button for your Rmarkdown document.

If you want to look at what the assignment looks like in progress with some of the R codes remaining invalid in the R code chunks, remember that you can set the R chunk options to `eval = FALSE` like so:

```
```{r this-chunk-will-not-run, eval = FALSE} `r''`
ggplot()
```
```

If you use `eval = FALSE` or `cache = TRUE`, please remember to ensure that you have set to `eval = TRUE` when you submit the assignment, to ensure all your R codes run.

There are a few tricky bits that might require you to look back into your previous R code chunks (that is intentionally done for you to understand how things work within an Rmd file!)

You will be completing this assignment **INDIVIDUALLY**.

Due Date

This assignment is due in by close of business (5pm) on Friday, 26 March 2021. You will submit the assignment via Moodle. Please make sure you add your name on the YAML part of this Rmd file.

Treatment

You work as a data scientist in the well-named consulting company, "Consulting for You".

It's your second day at the company, and you're taken to your desk. Your boss says to you:

> We have a data set with the crime statistics in New South Wales for the past years!
>
> We've got a meeting coming up soon to get insights about the crime in NSW. We want you to tell us about this data set and what we can do with it.
>
> You're in with the new hires of data scientists here. We'd like you to take a look at the data and tell me what the spreadsheet tells us. I've written some questions on the report for you to answer.
>
> Most importantly, can you get this to me by **5pm, Friday, 26 March 2021**.
>
> Please read below and answer all the questions (ensure that you can knit the file to produce an html file and a PDF file to hand them in to me via Moodle):

Load all the libraries that you need here

```
library(tidyverse)
```

Reading and preparing data

```
crime_dat <- read_csv("data/SuburbData2019.csv")
```

```
# I am selecting here only a portion of the data
# to reduce computation times.
 crime_data <-crime_dat %>%
    select(-c(`Jan 1995`:`Jan 2010`)) %>%
    dplyr::filter(Suburb %in% c("Chifley",
                                "Redfern",
                                "Clare",
                                "Paddington",
                                "Zetland",
```

```
                                "Claymore",
                                 "Congo",
                                "Coogee",
                                "Yenda",
                                 "Young",
                                 "Yarra",
                                "Woodcroft",
                                "Woodhill",
                                 "Warri",
                                 "Waterloo",
                                 "Randwick"))
```

Question 1: Display the first 10 rows of the data set

Hint: Check *?head* in your R console

```
head(crime_data, 10)   # 1pt
```

```
## # A tibble: 10 x 122
##    Suburb  `Offence categor~ Subcategory        `Feb 2010` `Mar 2010` `Apr 2010`
##    <chr>   <chr>             <chr>                   <dbl>      <dbl>      <dbl>
##  1 Chifley Homicide          Murder *                    0          0          0
##  2 Chifley Homicide          Attempted murder            0          0          0
##  3 Chifley Homicide          Murder accessory,~          0          0          0
##  4 Chifley Homicide          Manslaughter *              0          0          0
##  5 Chifley Assault           Domestic violence~          1          0          1
##  6 Chifley Assault           Non-domestic viol~          2          0          0
##  7 Chifley Assault           Assault Police              0          0          0
##  8 Chifley Sexual offences   Sexual assault              1          0          0
##  9 Chifley Sexual offences   Indecent assault,~          0          0          0
## 10 Chifley Abduction and ki~ Abduction and kid~          0          0          0
## # ... with 116 more variables: May 2010 <dbl>, Jun 2010 <dbl>, Jul 2010 <dbl>,
## #   Aug 2010 <dbl>, Sep 2010 <dbl>, Oct 2010 <dbl>, Nov 2010 <dbl>,
## #   Dec 2010 <dbl>, Jan 2011 <dbl>, Feb 2011 <dbl>, Mar 2011 <dbl>,
## #   Apr 2011 <dbl>, May 2011 <dbl>, Jun 2011 <dbl>, Jul 2011 <dbl>,
## #   Aug 2011 <dbl>, Sep 2011 <dbl>, Oct 2011 <dbl>, Nov 2011 <dbl>,
## #   Dec 2011 <dbl>, Jan 2012 <dbl>, Feb 2012 <dbl>, Mar 2012 <dbl>,
## #   Apr 2012 <dbl>, May 2012 <dbl>, Jun 2012 <dbl>, Jul 2012 <dbl>,
## #   Aug 2012 <dbl>, Sep 2012 <dbl>, Oct 2012 <dbl>, Nov 2012 <dbl>,
## #   Dec 2012 <dbl>, Jan 2013 <dbl>, Feb 2013 <dbl>, Mar 2013 <dbl>,
## #   Apr 2013 <dbl>, May 2013 <dbl>, Jun 2013 <dbl>, Jul 2013 <dbl>,
## #   Aug 2013 <dbl>, Sep 2013 <dbl>, Oct 2013 <dbl>, Nov 2013 <dbl>,
## #   Dec 2013 <dbl>, Jan 2014 <dbl>, Feb 2014 <dbl>, Mar 2014 <dbl>,
## #   Apr 2014 <dbl>, May 2014 <dbl>, Jun 2014 <dbl>, Jul 2014 <dbl>,
## #   Aug 2014 <dbl>, Sep 2014 <dbl>, Oct 2014 <dbl>, Nov 2014 <dbl>,
## #   Dec 2014 <dbl>, Jan 2015 <dbl>, Feb 2015 <dbl>, Mar 2015 <dbl>,
## #   Apr 2015 <dbl>, May 2015 <dbl>, Jun 2015 <dbl>, Jul 2015 <dbl>,
## #   Aug 2015 <dbl>, Sep 2015 <dbl>, Oct 2015 <dbl>, Nov 2015 <dbl>,
## #   Dec 2015 <dbl>, Jan 2016 <dbl>, Feb 2016 <dbl>, Mar 2016 <dbl>,
## #   Apr 2016 <dbl>, May 2016 <dbl>, Jun 2016 <dbl>, Jul 2016 <dbl>,
## #   Aug 2016 <dbl>, Sep 2016 <dbl>, Oct 2016 <dbl>, Nov 2016 <dbl>,
## #   Dec 2016 <dbl>, Jan 2017 <dbl>, Feb 2017 <dbl>, Mar 2017 <dbl>,
## #   Apr 2017 <dbl>, May 2017 <dbl>, Jun 2017 <dbl>, Jul 2017 <dbl>,
## #   Aug 2017 <dbl>, Sep 2017 <dbl>, Oct 2017 <dbl>, Nov 2017 <dbl>,
```

```
## #   Dec 2017 <dbl>, Jan 2018 <dbl>, Feb 2018 <dbl>, Mar 2018 <dbl>,
## #   Apr 2018 <dbl>, May 2018 <dbl>, Jun 2018 <dbl>, Jul 2018 <dbl>,
## #   Aug 2018 <dbl>, ...
```

Question 2: How many variables and observations do we have?

Hint: Look for help *?dim* in your R console and remember that variables are in columns and observations in rows. *dim()* returns the number of rows and the number of columns in the data set (in that order)

```
dim(crime_data)   # 1pt
```

```
## [1] 992 122
```

The number of variables are 122 (1pt) and the number of rows are 992 (1pt)

Question 3: What are the names of the first 20 variables in this data set?

```
names(crime_data)[1:20]   #1pt
```

```
##  [1] "Suburb"           "Offence category" "Subcategory"      "Feb 2010"
##  [5] "Mar 2010"         "Apr 2010"         "May 2010"         "Jun 2010"
##  [9] "Jul 2010"         "Aug 2010"         "Sep 2010"         "Oct 2010"
## [13] "Nov 2010"         "Dec 2010"         "Jan 2011"         "Feb 2011"
## [17] "Mar 2011"         "Apr 2011"         "May 2011"         "Jun 2011"
```

Question 4: Rename the variable of "Offence category" to "Offence_category" and show the names of the first 4 variables in the data set

```
crime <- crime_data %>%
  rename(Offence_category = `Offence category`)  # 1pt

names(crime)[1:4] #1pt
```

```
## [1] "Suburb"           "Offence_category" "Subcategory"      "Feb 2010"
```

Question 5: Change the "crime" data ("SuburbData2019csv") into long format so that all the years are grouped together into a variable called "year" and the corresponding incidents count into a variable called "incidents"

```
crime_long <- crime %>%
  pivot_longer(cols = 4:122,  # 2pt
               names_to = "year",             # 1pt
               values_to = "incidents")       # 1pt

head(crime_long)     # 1pt
```

```
## # A tibble: 6 x 5
##   Suburb  Offence_category Subcategory year     incidents
##   <chr>   <chr>            <chr>       <chr>        <dbl>
## 1 Chifley Homicide         Murder *    Feb 2010         0
## 2 Chifley Homicide         Murder *    Mar 2010         0
## 3 Chifley Homicide         Murder *    Apr 2010         0
## 4 Chifley Homicide         Murder *    May 2010         0
## 5 Chifley Homicide         Murder *    Jun 2010         0
## 6 Chifley Homicide         Murder *    Jul 2010         0
```

Question 6: Separate the column "year" into two columns with names "Month" and "Year". Display the first 3 lines of the data set to show the updated data set

```
crime_long_new <- crime_long %>%
  separate(col = year,                          # 1pt
           into = c("Month", "Year"), " " )    # 2pt

  head(crime_long_new,3)                         # 1pt
```

```
## # A tibble: 3 x 6
##   Suburb  Offence_category Subcategory Month Year  incidents
##   <chr>   <chr>            <chr>       <chr> <chr>     <dbl>
## 1 Chifley Homicide         Murder *    Feb   2010          0
## 2 Chifley Homicide         Murder *    Mar   2010          0
## 3 Chifley Homicide         Murder *    Apr   2010          0
```

Question 7: If you look at the data *crime_long_new*, you will notice that the variable of "Year" is coded as character. In this section, we are going to convert the variable of "Year" to a numeric variable

```
crime_long_new %>%
  mutate(Year = as.numeric(Year))   # 1pt
```

```
## # A tibble: 118,048 x 6
##    Suburb  Offence_category Subcategory Month  Year incidents
##    <chr>   <chr>            <chr>       <chr> <dbl>     <dbl>
##  1 Chifley Homicide         Murder *    Feb    2010         0
##  2 Chifley Homicide         Murder *    Mar    2010         0
##  3 Chifley Homicide         Murder *    Apr    2010         0
##  4 Chifley Homicide         Murder *    May    2010         0
##  5 Chifley Homicide         Murder *    Jun    2010         0
##  6 Chifley Homicide         Murder *    Jul    2010         0
##  7 Chifley Homicide         Murder *    Aug    2010         0
##  8 Chifley Homicide         Murder *    Sep    2010         0
##  9 Chifley Homicide         Murder *    Oct    2010         0
## 10 Chifley Homicide         Murder *    Nov    2010         0
## # ... with 118,038 more rows
```

```
head(crime_long_new)                # 1pt
```

```
## # A tibble: 6 x 6
##   Suburb  Offence_category Subcategory Month Year  incidents
##   <chr>   <chr>            <chr>       <chr> <chr>     <dbl>
## 1 Chifley Homicide         Murder *    Feb   2010          0
## 2 Chifley Homicide         Murder *    Mar   2010          0
## 3 Chifley Homicide         Murder *    Apr   2010          0
## 4 Chifley Homicide         Murder *    May   2010          0
## 5 Chifley Homicide         Murder *    Jun   2010          0
## 6 Chifley Homicide         Murder *    Jul   2010          0
```

Question 8: Display the years in the data set. How many years are included in this data set?

> Remember that you can learn more about what these functions by typing: `?unique` or `?length` into the R console.

```
unique(crime_long_new$Year)    # 1pt
```

```
##  [1] "2010" "2011" "2012" "2013" "2014" "2015" "2016" "2017" "2018" "2019"
```

```
# length tell us the length or longitude of a variable or a vector
length(unique(crime_long_new$Year))   #1pt
```

```
## [1] 10
```

Question 9: How many different suburbs are there in the data set?

```
length(unique(crime_long_new$Suburb))     # 1pt
```

```
## [1] 16
```

```
n_distinct(crime_long_new$Suburb) # 1pt
```

```
## [1] 16
```

Question 10: How many incidents do we have per "Offence_category" in total for 2019?

```
crime_long_new %>%
  dplyr::filter(Year == "2019") %>%    # 1pt
  count(Offence_category , wt = incidents)              # 1pt
```

```
## # A tibble: 21 x 2
##    Offence_category                      n
##    <chr>                             <dbl>
##  1 Abduction and kidnapping              1
##  2 Against justice procedures         1950
##  3 Arson                                60
##  4 Assault                            1396
##  5 Betting and gaming offences           1
##  6 Blackmail and extortion               2
```

```
##  7 Disorderly conduct                       429
##  8 Drug offences                           1416
##  9 Homicide                                   2
## 10 Intimidation, stalking and harassment    566
## # ... with 11 more rows
```

Question 11: Which is the "Offence_category" with highest number of incidents in 2019?

```
crime_long_new %>%
  dplyr::filter(Year == "2019") %>%    #  1pt
  count(Offence_category, wt = incidents, sort = TRUE) # 1pt
```

```
## # A tibble: 21 x 2
##    Offence_category                          n
##    <chr>                                 <dbl>
##  1 Theft                                  4061
##  2 Against justice procedures             1950
##  3 Drug offences                          1416
##  4 Assault                                1396
##  5 Malicious damage to property           1093
##  6 Intimidation, stalking and harassment   566
##  7 Transport regulatory offences           517
##  8 Disorderly conduct                      429
##  9 Liquor offences                         356
## 10 Sexual offences                         273
## # ... with 11 more rows
```

Question 12: How many offences are there in each Subcategory of the "Offence_category" of *Homicide*?

```
crime_long_new %>%
  dplyr::filter(Offence_category == "Homicide") %>%    # 1pt
  group_by(Subcategory) %>%                            # 1pt
  summarise(Number_of_incidents = sum(incidents))      # 1pt
```

```
## # A tibble: 4 x 2
##   Subcategory                  Number_of_incidents
##   <chr>                                      <dbl>
## 1 Attempted murder                               3
## 2 Manslaughter *                                 1
## 3 Murder *                                      14
## 4 Murder accessory, conspiracy                   1
```

Question 13: Select the suburb called "Paddington" and calculate the number of incidents for "Offence_category" of "Drug offences" then calculate the total number of incidents for each Subcategory. Finally, show a table arranged by "Number_of_ incidents" (high to low)

```
Paddington <- crime_long_new %>%
  dplyr::filter( Suburb == "Paddington",                          # 2pt
                 Offence_category == "Drug offences") %>%       #1pt
          group_by(Subcategory) %>%                              # 1pt
          summarise(Number_of_incidents = sum(incidents)) %>%    # 1pt
          arrange(desc(Number_of_incidents))    # 1pt

head(Paddington)                                 # 1pt
```

```
## # A tibble: 6 x 2
##   Subcategory                              Number_of_incidents
##   <chr>                                                  <dbl>
## 1 Possession and/or use of cannabis                        154
## 2 Possession and/or use of cocaine                         111
## 3 Possession and/or use of other drugs                      82
## 4 Other drug offences                                       73
## 5 Dealing, trafficking in cocaine                           68
## 6 Possession and/or use of amphetamines                     57
```

Question 14: Let's have a look at the changes over time for "Possession and/or use of cannabis" in the suburb of Paddington

To answer this question, we need to first filter the "Suburb" and the "Subcategory". Then, group incident by year and finally sum the number of incidents for each year

```
Paddington_cannabis <- crime_long_new %>%
  dplyr::filter( Suburb == "Paddington",                # 1pt
                 Subcategory == "Possession and/or use of cannabis") %>%   # 1pt
          group_by(Year) %>%                                               # 1pt
          summarise(Number_of_incidents = sum(incidents)) %>%              #1pt
  mutate(Year = as.numeric(Year))    # 1pt

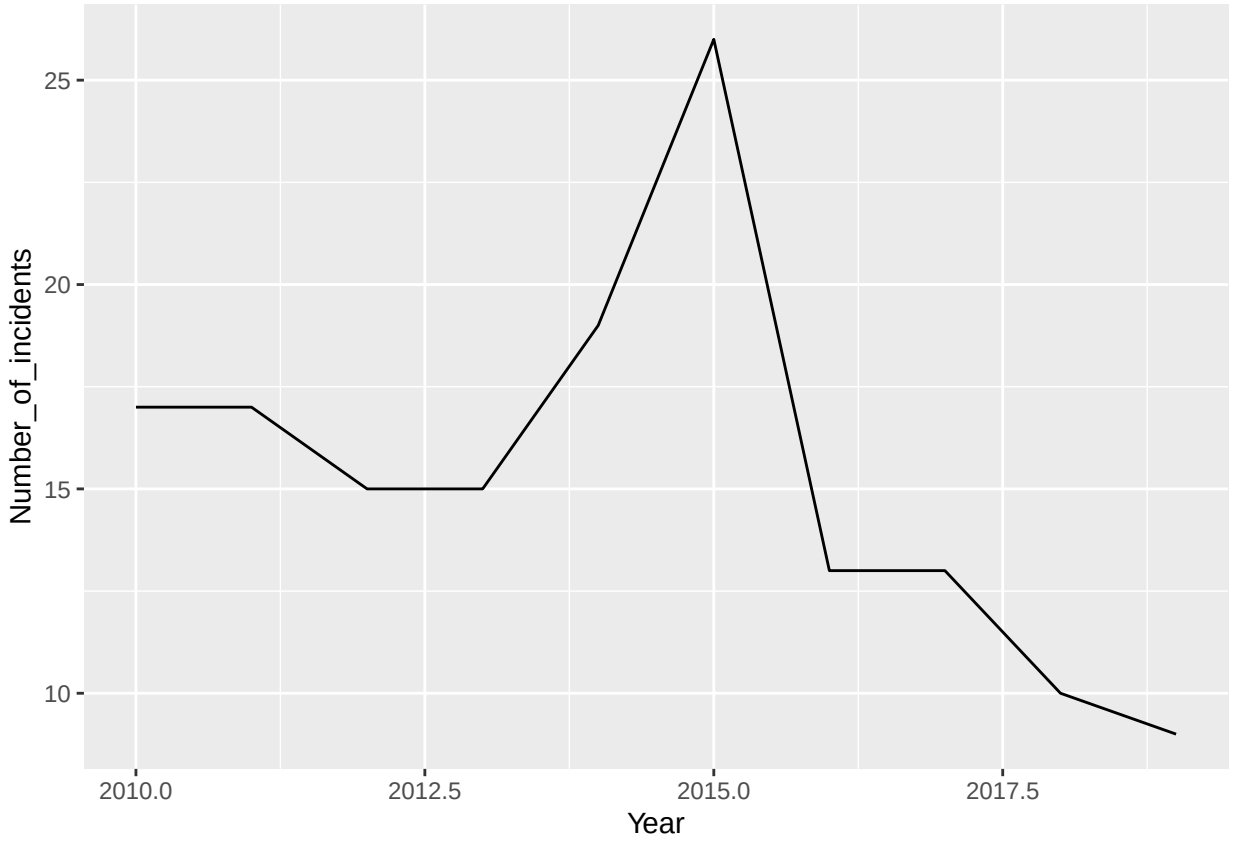

head(Paddington_cannabis,3)  # 1pt
```

```
## # A tibble: 3 x 2
##    Year Number_of_incidents
##   <dbl>               <dbl>
## 1  2010                  17
## 2  2011                  17
## 3  2012                  15
```

Question 15: Create a line plot to display the trend of the incidents that you calculated for Paddington

On the x-axis you should have "Year" and on the y-axis you should display "Number_of_incidents"

```
ggplot(Paddington_cannabis, aes( x = Year, y = Number_of_incidents)) + # 2pt
  geom_line()                      # 1pt
```



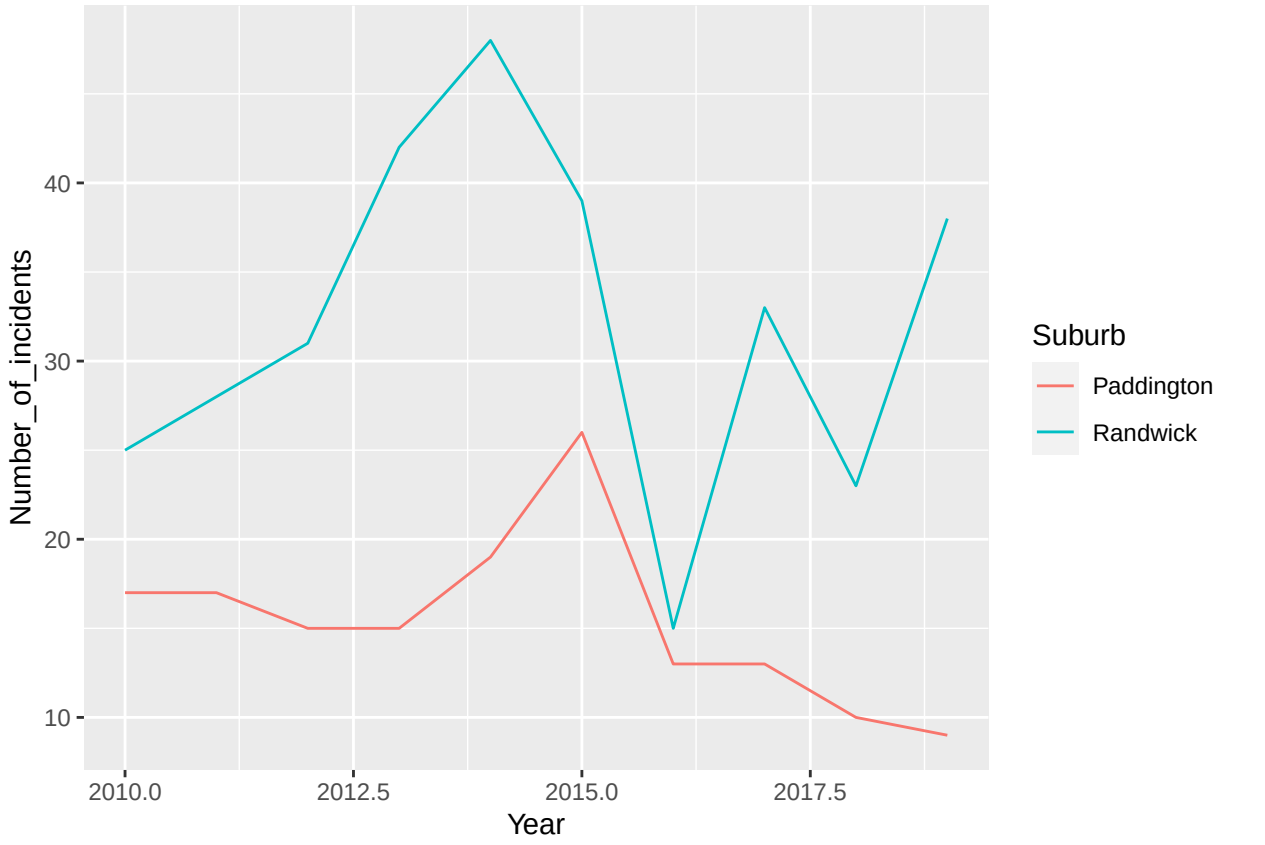
Question 16: Create the same plot as in Question 15 but now include also the suburb called "Randwick" (you will see two trends in the same plot). Make sure that the variable of "Suburb" is defined as a *factor*

```
both_cannabis <- crime_long_new %>%
  dplyr::filter(Suburb %in% c("Paddington", "Randwick"),          # 1pt
                Subcategory == "Possession and/or use of cannabis") %>%   # 1pt
          group_by(Year, Suburb) %>%                              # 1pt
          summarise(Number_of_incidents = sum(incidents)) %>%     # 1pt
  mutate(Year = as.numeric(Year),                                 # 1pt
         Suburb = as.factor(Suburb))                              # 1pt
```

```
## `summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```

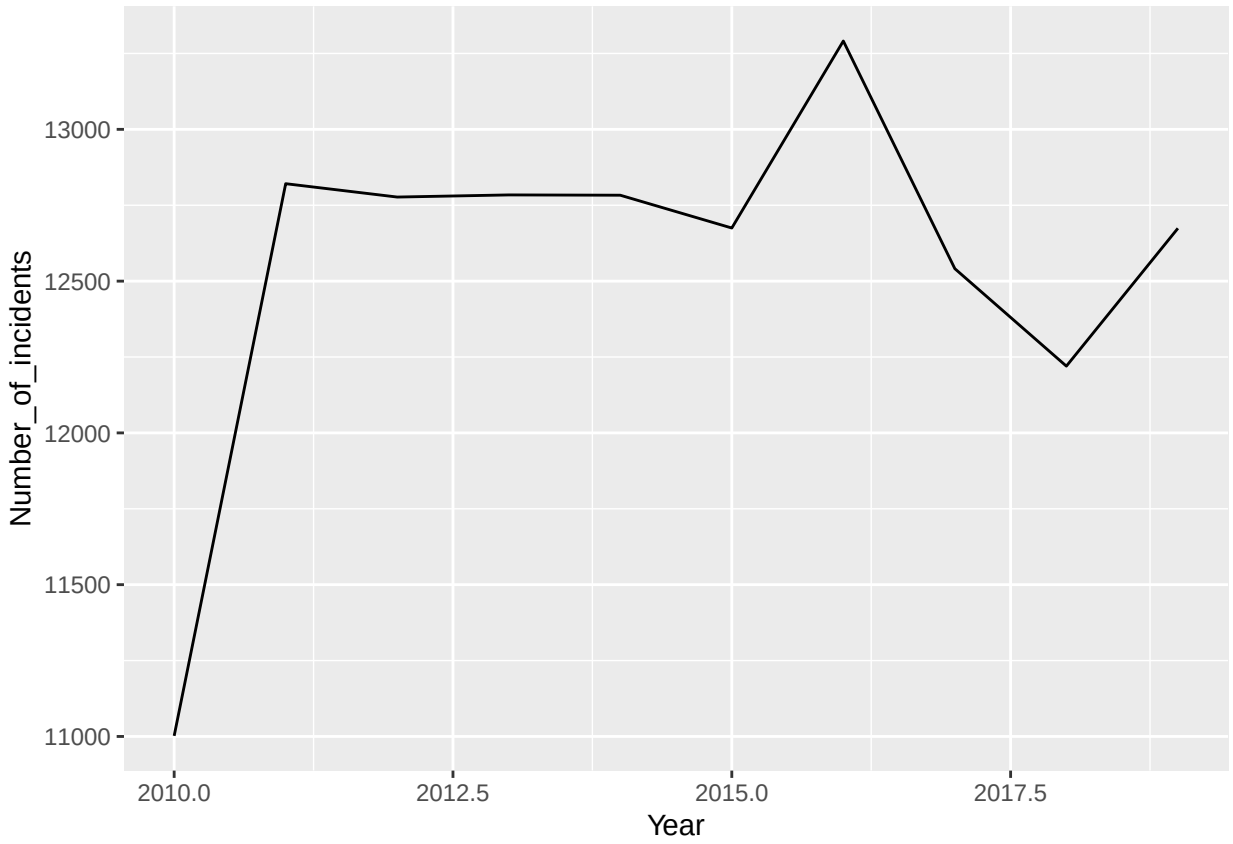
```
ggplot(both_cannabis, aes( x = Year,      # 1pt
                           y = Number_of_incidents, # 1pt
```

```
                          color = Suburb)) +    # 1pt
  geom_line()    # 1pt
```



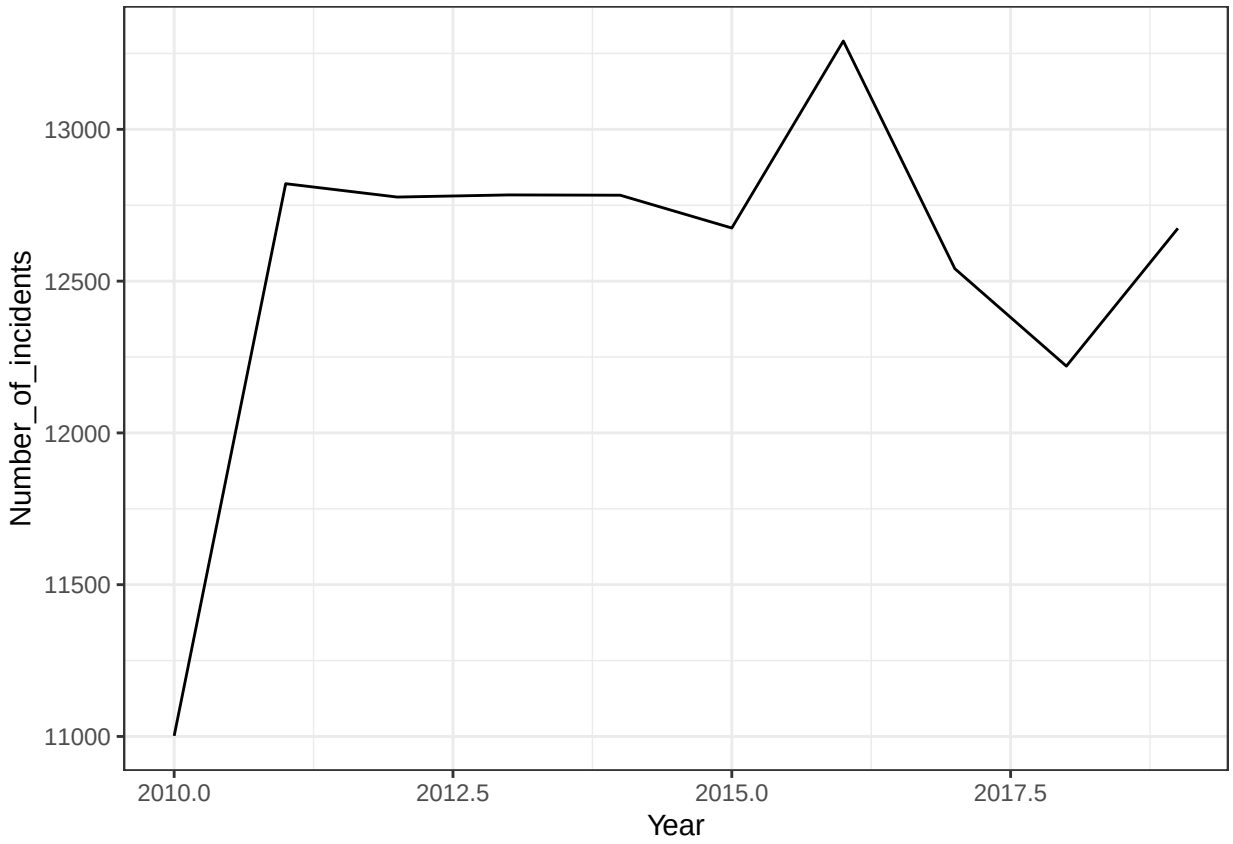
Question 17: Let's now look at the total number of crime incidents in NSW and create a plot to visualize the trend

```
crime_long_new %>%
  dplyr::select( Year,    # 1pt
                 incidents) %>%  # 1pt
          group_by(Year) %>%    # 1pt
          summarise(Number_of_incidents = sum(incidents)) %>%  # 1pt
  mutate(Year = as.numeric(Year)) %>%    # 1pt
  ggplot(aes(x = Year, y = Number_of_incidents )) +   # 1pt
  geom_line()  # 1pt
```



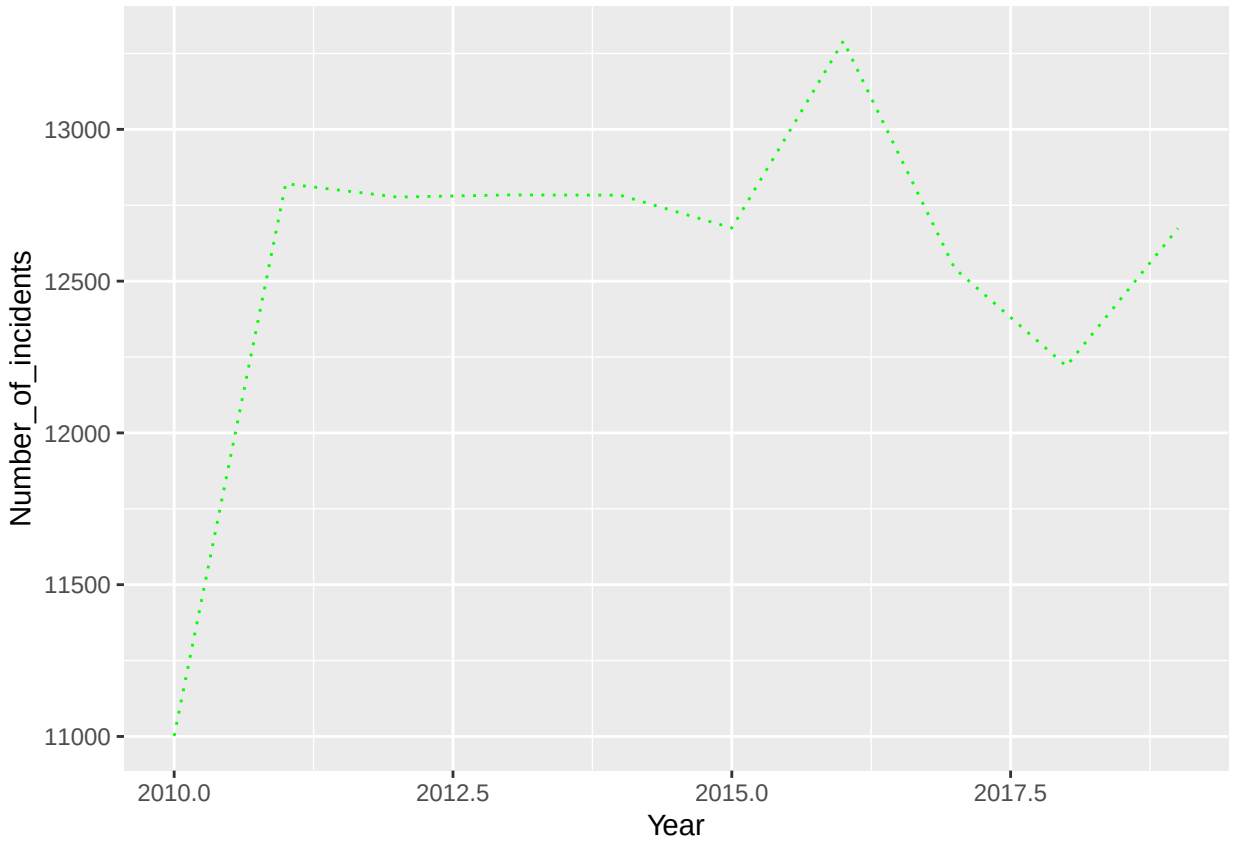
Question 18: Now, let's change the background color of the plot to white using the *theme_bw()*

```
crime_long_new %>%
  dplyr::select(Year,    # 1pt
                incidents) %>% # 1pt
          group_by(Year) %>%   # 1pt
          summarise(Number_of_incidents = sum(incidents)) %>%  # 1pt
  mutate(Year = as.numeric(Year)) %>%  # 1pt
  ggplot(aes(x = Year, y = Number_of_incidents )) +  # 1pt
  geom_line() +  # 1pt
  theme_bw()  # 1pt
```



Question 19: Let's change the line color to green and replace it with a dotted line

```
crime_long_new %>%
  dplyr::select(Year,  # 1pt
                incidents) %>%  # 1pt
          group_by(Year) %>%  # 1pt
         summarise(Number_of_incidents = sum(incidents)) %>%  # 1pt
  mutate(Year = as.numeric(Year)) %>%  # 1pt
  ggplot(aes(x = Year, y = Number_of_incidents )) +  # 1pt
  geom_line(linetype = 3, color ="green")  # 1pt
```

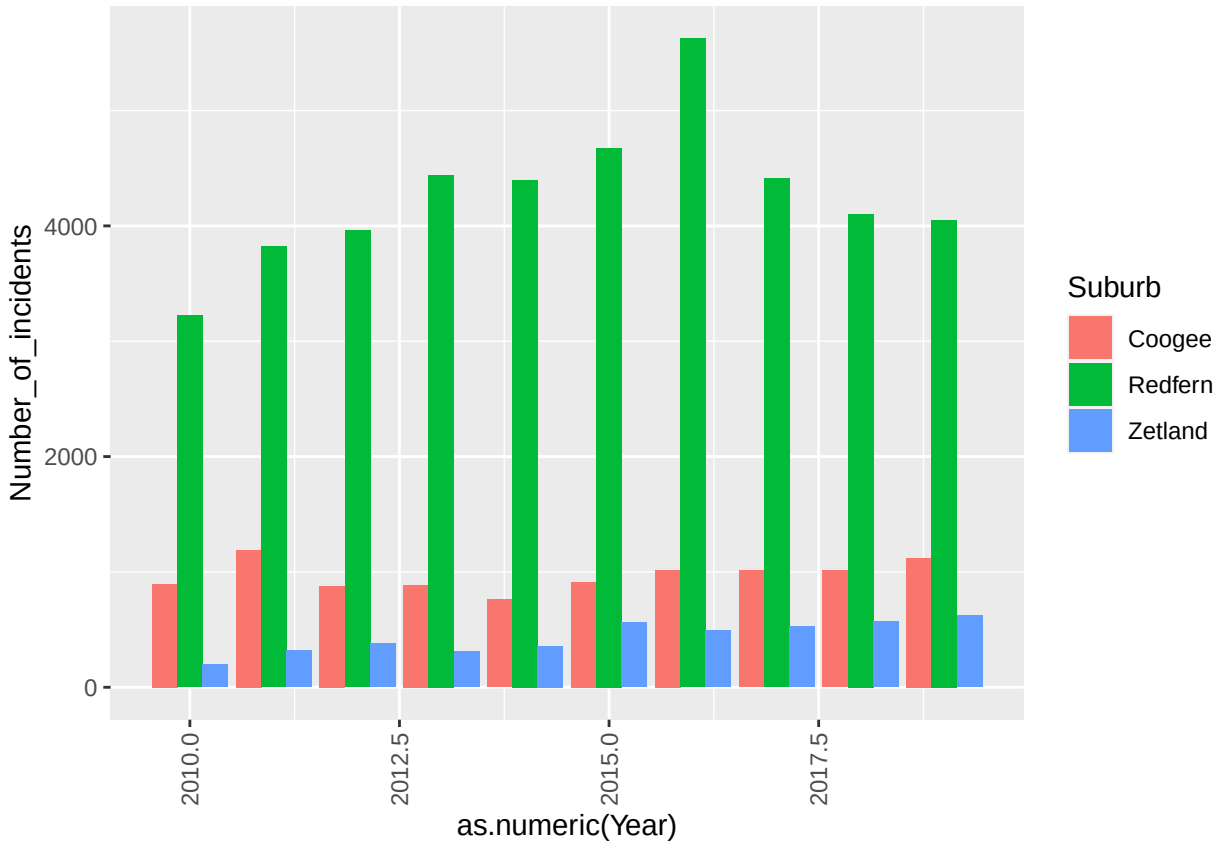


Question 20: Now, let's look at the total number of crime incidents for the suburbs of Redfern, Coogee, and Zetland by creating a bar plot where we have the incidents per suburb by year next to each other

```
comparison_data<- crime_long_new %>%
  dplyr::select(Suburb,  # 1pt
                Year,    # 1pt
                incidents) %>%  # 1pt
  dplyr::filter( Suburb %in% c("Redfern", "Coogee", "Zetland")) %>%  # 1pt
          group_by(Suburb, Year) %>% # 1pt
          summarise(Number_of_incidents = sum(incidents)) # 1pt
```

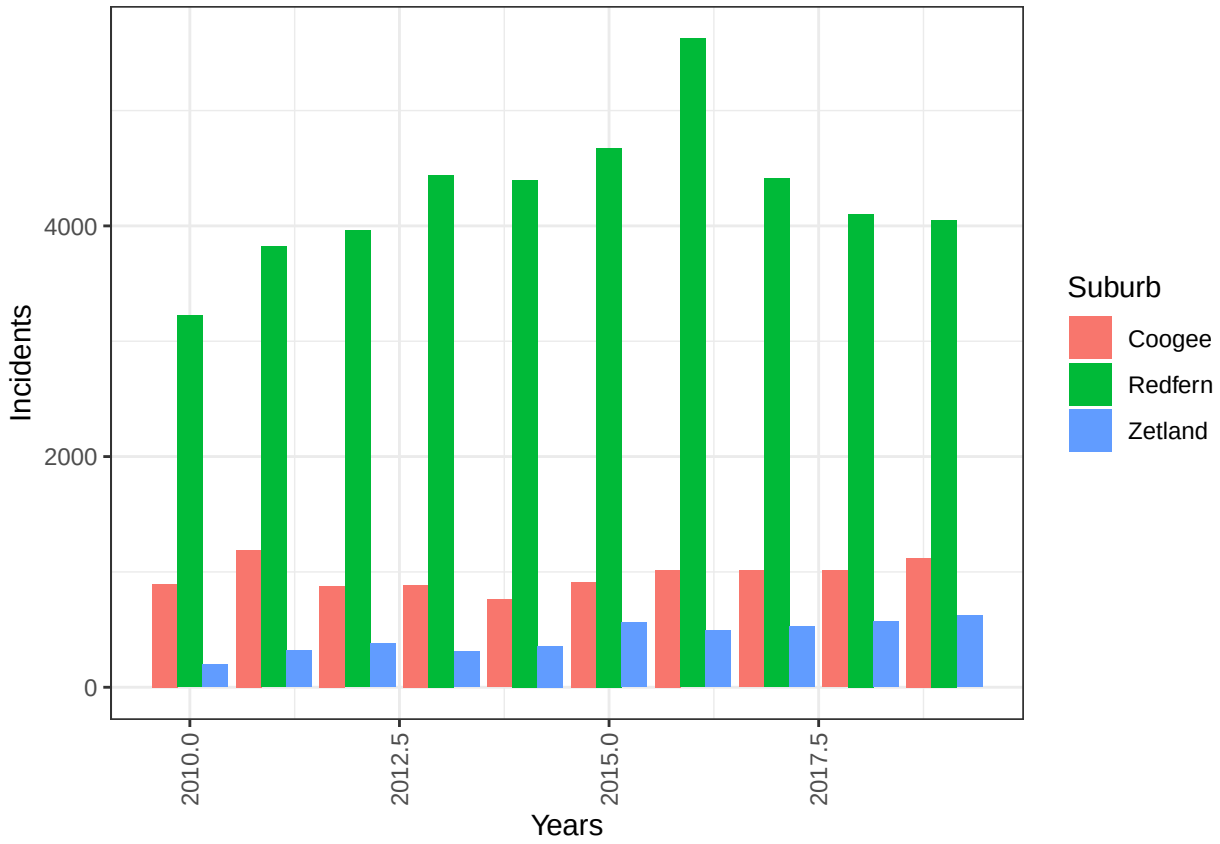
```
## `summarise()` has grouped output by 'Suburb'. You can override using the `.groups` argument.
```

```
ggplot(comparison_data, aes(x = as.numeric(Year),    # 1pt
                            y = Number_of_incidents,  # 1pt
                            fill = Suburb)) +  # 1pt
  geom_bar(stat = "identity",   # 1pt
             position = "dodge") +  # 1pt
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) # 1pt
```



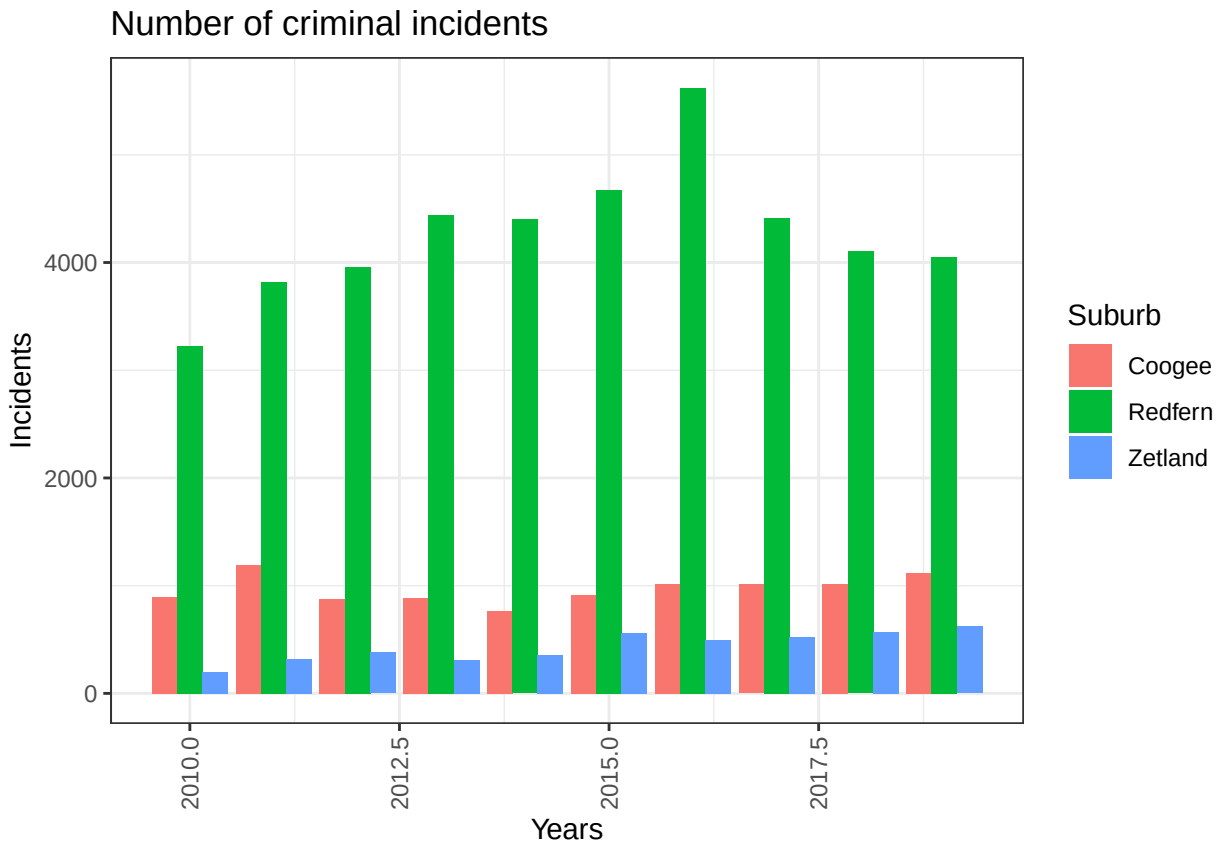
Question 21: Change the x and y-axis labels to "Years" and " Incidents", respectively, for the figure in Question 20 and use the black and white theme

```
ggplot(comparison_data, aes(x = as.numeric(Year),     # 1pt
                            y = Number_of_incidents,  # 1pt
                            fill = Suburb)) +   # 1pt
  geom_bar(stat = "identity",    # 1pt
           position = "dodge") +  # 1pt
theme_bw() +  # 1pt
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +  # 1pt
  xlab("Years") +  # 1pt
  ylab("Incidents")  # 1pt
```



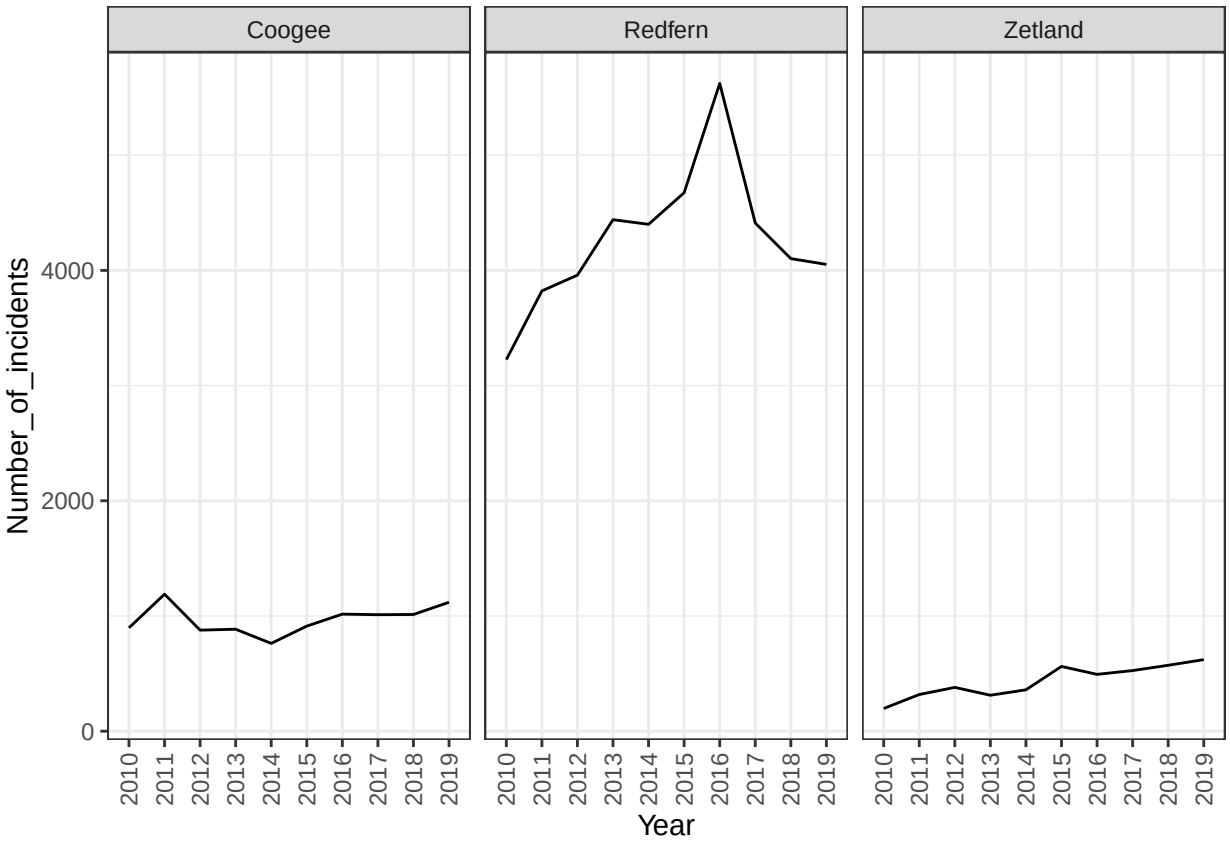
Question 22: Add the following title to the figure constructed in Question 21: "Number of criminal incidents"

```
ggplot(comparison_data, aes(x = as.numeric(Year),    # 1pt
                            y = Number_of_incidents,  # 1pt
                            fill = Suburb)) +    # 1pt
  geom_bar(stat = "identity",    # 1pt
           position = "dodge") +  # 1pt
  theme_bw() +    # 1pt
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +    # 1pt
  xlab("Years") +    # 1pt
 ylab("Incidents") +    # 1pt
  ggtitle("Number of criminal incidents")  # 1pt
```



Question 23: By using "facet_wrap", create a line plot to show the trends for "Number_of_incidents" for each of the three suburbs

```
ggplot(comparison_data, aes(x= Year,    # 1pt
                            y = Number_of_incidents,  # 1pt
                            group =Suburb)) +  # 1pt
  geom_line() +   # 1pt
  facet_wrap(~Suburb) +  # 1pt
  theme_bw() +  # 1pt
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))  # 1pt
```



Question 24: Transform the data set named *comparison_data* into a wide format where the suburbs of Coogee, Redfern, and Zetland are displayed as columns

```
comparison_data %>%
 pivot_wider(id_cols = 1:3,    # 1pt
             names_from = Suburb,  # 1pt
             values_from = Year)   # 1pt
```

```
## # A tibble: 30 x 4
##    Number_of_incidents Coogee Redfern Zetland
##                  <dbl> <chr>  <chr>   <chr>
##  1                 897 2010   <NA>    <NA>
##  2                1189 2011   <NA>    <NA>
##  3                 877 2012   <NA>    <NA>
##  4                 885 2013   <NA>    <NA>
##  5                 762 2014   <NA>    <NA>
##  6                 912 2015   <NA>    <NA>
##  7                1016 2016   <NA>    <NA>
##  8                1011 2017   <NA>    <NA>
##  9                1013 2018   <NA>    <NA>
## 10                1119 2019   <NA>    <NA>
## # ... with 20 more rows
```