Assignment 1 ETC1010 - 5510

New South Wales Crime Incidents Report

Your name

Friday, March 12 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",
                              "Redfern",
```

```
                                "Zetland",
                                "Claymore",
                                "Congo",
                                "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(---, 10)    # 1pt
```

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(---)    # 1pt
```

The number of variables are `dim(crime_data)---` (1pt) and the number of rows are `dim(crime_data)---` (1pt)

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

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

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`)  # 1pt

names(crime)[---] #1pt
```

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 = ---:---,  # 2pt
               names_to = "---",              # 1pt
               values_to = "---")       # 1pt

head(---)     # 1pt
```

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 = ---,                          # 1pt
           into = c("---", "---"), " " )   # 2pt

  head(---)                         # 1pt
```

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(---))   # 1pt

head(---)                # 1pt
```

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

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

```
length(unique(---))     # 1pt
n_distinct(---) # 1pt
```

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

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

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

```
crime_long_new %>%
  dplyr::filter(Year == "---") %>%    #  1pt
  count(---, wt = ---, sort = ---) # 1pt
```

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

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

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( ---== "---",                          # 2pt
                 Offence_category == "---") %>%      #1pt
          group_by(---) %>%                              # 1pt
          summarise(Number_of_incidents = sum(---)) %>%    # 1pt
          arrange(---)    # 1pt

head(---)                              # 1pt
```

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 == ---,                # 1pt
                 Subcategory == ---) %>%   # 1pt
          group_by(Year) %>%                                                  # 1pt
          summarise(Number_of_incidents = --- %>%              #1pt
  mutate(Year = as.numeric(---))    # 1pt


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

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 = ---, y = ---)) + # 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("---", ---),          # 1pt
                 Subcategory == "---") %>%   # 1pt
          group_by(Year, Suburb) %>%                                  # 1pt
          summarise(Number_of_incidents = --- %>%     # 1pt
  mutate(Year = as.numeric(---),                                    # 1pt
         Suburb = as.factor(---))                                  # 1pt


ggplot(both_cannabis, aes( x = ---,       # 1pt
                           y = ---, # 1pt
                           color = ---)) +   # 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( ---,   # 1pt
                 ---) %>%  # 1pt
```

```
          ---(Year) %>%    # 1pt
          summarise(Number_of_incidents = --- %>%  # 1pt
  mutate(Year = as.numeric(---)) %>%    # 1pt
  ggplot(aes(x = ---, y = --- )) +   # 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(---,    # 1pt
                ---) %>% # 1pt
          group_by(---) %>%   # 1pt
          summarise(--- = ---) %>%  # 1pt
  mutate(Year = as.numeric(---)) %>%  # 1pt
  ggplot(aes(x = ---, y = --- )) +  # 1pt
  geom_line() +  # 1pt
  theme_---  # 1pt
```

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

```
crime_long_new %>%
  dplyr::select(---,  # 1pt
                ---) %>%  # 1pt
          group_---(Year) %>%  # 1pt
         ---(--- = sum(---)) %>%  # 1pt
  mutate(--- = as.numeric(---)) %>%  # 1pt
  ggplot(aes(x = ---, y = --- )) +  # 1pt
  geom_line(linetype = ---, color =---)  # 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(---,  # 1pt
                ---,    # 1pt
                ---) %>%  # 1pt
  dplyr::filter( --- %in% c("Redfern", "Coogee", "Zetland")) %>%  # 1pt
          group_by(---, ---) %>% # 1pt
          summarise(Number_of_incidents = ---(---)) # 1pt

ggplot(comparison_data, aes(x = ---,    # 1pt
                            --- = ---,  # 1pt
                            fill = ---)) +  # 1pt
```

```
  geom_bar(--- = "identity",    # 1pt
           position = "---") +  # 1pt
  ---(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 = ---,     # 1pt
                            y = ---,  # 1pt
                            fill = ---)) +   # 1pt
  geom_bar(--- = "identity",    # 1pt
           --- = "dodge") +  # 1pt
  ---_bw() +  # 1pt
---(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +  # 1pt
  xlab("---") +  # 1pt
  ylab("---")  # 1pt
```

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

```
ggplot(comparison_data, aes(x = ---,    # 1pt
                            y = ---,  # 1pt
                            fill = ---)) +   # 1pt
  geom_bar(--- = "---",   # 1pt
           position = "---") +  # 1pt
  theme_---() +   # 1pt
  ---(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +   # 1pt
  ---("Years") +   # 1pt
 ---("Incidents") +   # 1pt
  ggtitle("---")  # 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(--- = Year,   # 1pt
                            --- = Number_of_incidents,  # 1pt
                            --- =Suburb)) +  # 1pt
  geom_---() +   # 1pt
  facet_wrap(~---) +  # 1pt
  ---() +  # 1pt
  ---(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 = ---,   # 1pt
              names_from = ---,  # 1pt
              values_from = ---)  # 1pt
```