STAT7123 Assignment 2

Student ID: 47541164

Aditya Tanaji Sagave

Question 1

Answer 1

```
#Importing data
penalty_data <- read.csv("./penalty_notice.csv")

# Use dplyr to group by LGA and calculate the total number of offenses
offenses_by_LGA <- penalty_data %>%
  group_by(Offence_LGA) %>%
  summarise(Total_Offenses = n())

# View the resulting data frame
print(offenses_by_LGA)
```

```
## # A tibble: 54 x 2
##    Offence_LGA          Total_Offenses
##    <chr>                         <int>
##  1 Bayside                          14
##  2 Bega Valley                       2
##  3 Blacktown                       107
##  4 Blue Mountains                   19
##  5 Burwood                         127
##  6 Byron                             6
##  7 Camden                           24
##  8 Campbelltown                     35
##  9 Canterbury-Bankstown            349
## 10 Central Coast                     3
## # i 44 more rows
```

Answer 1 a)

1. Large Number of LGAs: With 54 different LGAs in the dataset, visualizing them all on a single plot can lead to overcrowding and difficulty in distinguishing between LGAs. The sheer number of categories can make the plot cluttered and less interpretable.
2. Varying Offense Counts: The total offenses vary significantly across LGAs, with some having a high number of offenses and others having very few. This wide range can result in a skewed distribution, making it challenging to choose an appropriate scale for the y-axis in a single plot.
3. Categorical Data: LGAs are categorical data, and creating meaningful visualizations for categorical data can be challenging. Traditional plot types like bar charts may not be the most effective way to represent this data, especially when there are many categories.

4. Label Overlapping: When displaying labels for each LGA on a plot, there might be issues with label overlapping, especially if the LGAs have long names. This can lead to readability problems in the visualization.

Answer 1 b)

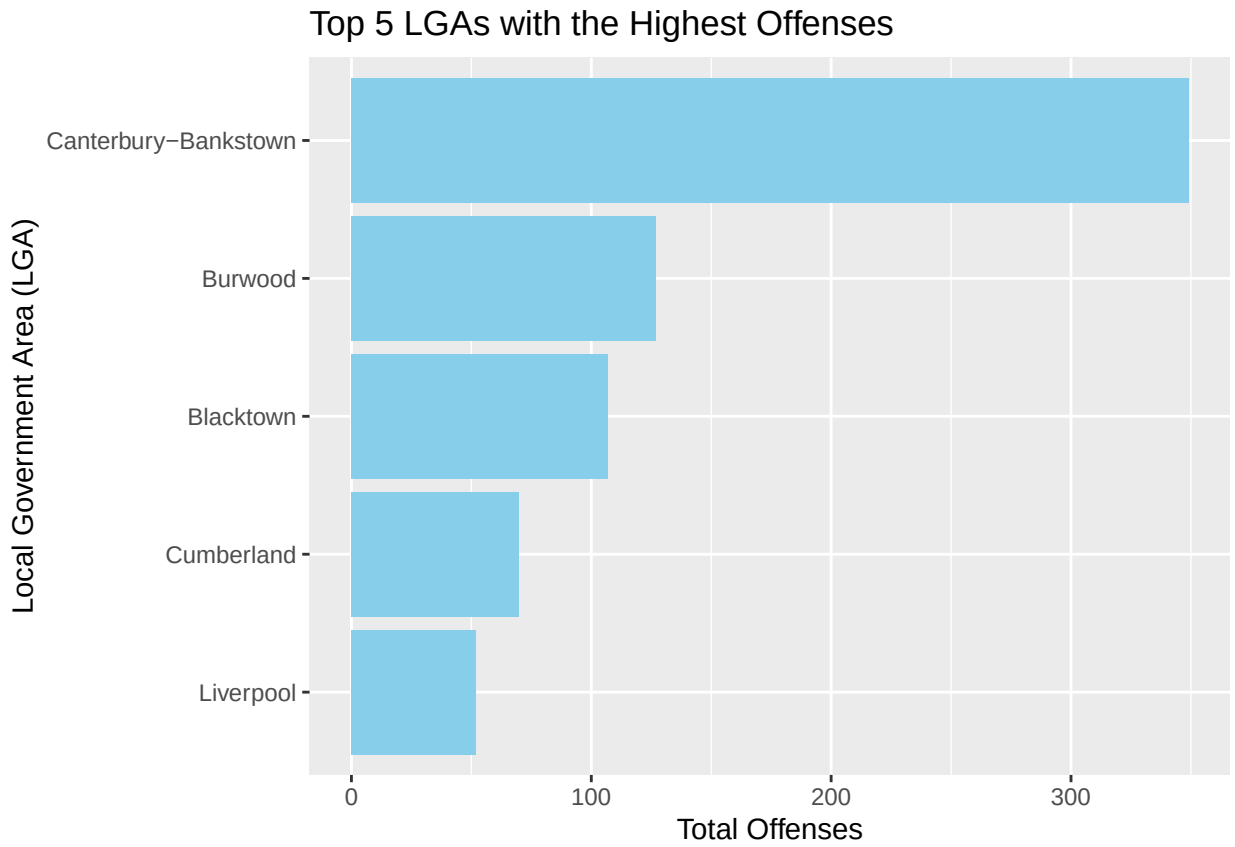
```
# Sort the offenses_by_LGA data frame by Total_Offenses in descending order
sorted_offenses <- offenses_by_LGA %>%
  arrange(desc(Total_Offenses))

# Select the LGA with the most offenses (the first row)
lga_with_most_offenses <- sorted_offenses[1, ]

# Print the LGA with the most offenses
print(lga_with_most_offenses)
```

```
## # A tibble: 1 x 2
##   Offence_LGA          Total_Offenses
##   <chr>                         <int>
## 1 Canterbury-Bankstown            349
```

The result obtained indicates that the LGA with the most offenses is "Canterbury-Bankstown," which has a total of 349 offenses.

Answer 1 c)

```
top_n_lgas <- 5

top_lgas <- sorted_offenses %>%
  head(top_n_lgas)


# Create a bar chart
ggplot(top_lgas, aes(x = reorder(Offence_LGA, Total_Offenses), y = Total_Offenses)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +  # Horizontal bars
  labs(x = "Local Government Area (LGA)", y = "Total Offenses") +
  ggtitle(paste("Top", top_n_lgas, "LGAs with the Highest Offenses"))
```



The plot displays top 5 areas withing Sydney area having biggest offences. The area with most offenses is Canterbury-Bankstown followed by Burwoord, Blacktown, Cumberland and Liverpool.

Question 2

Answer 2

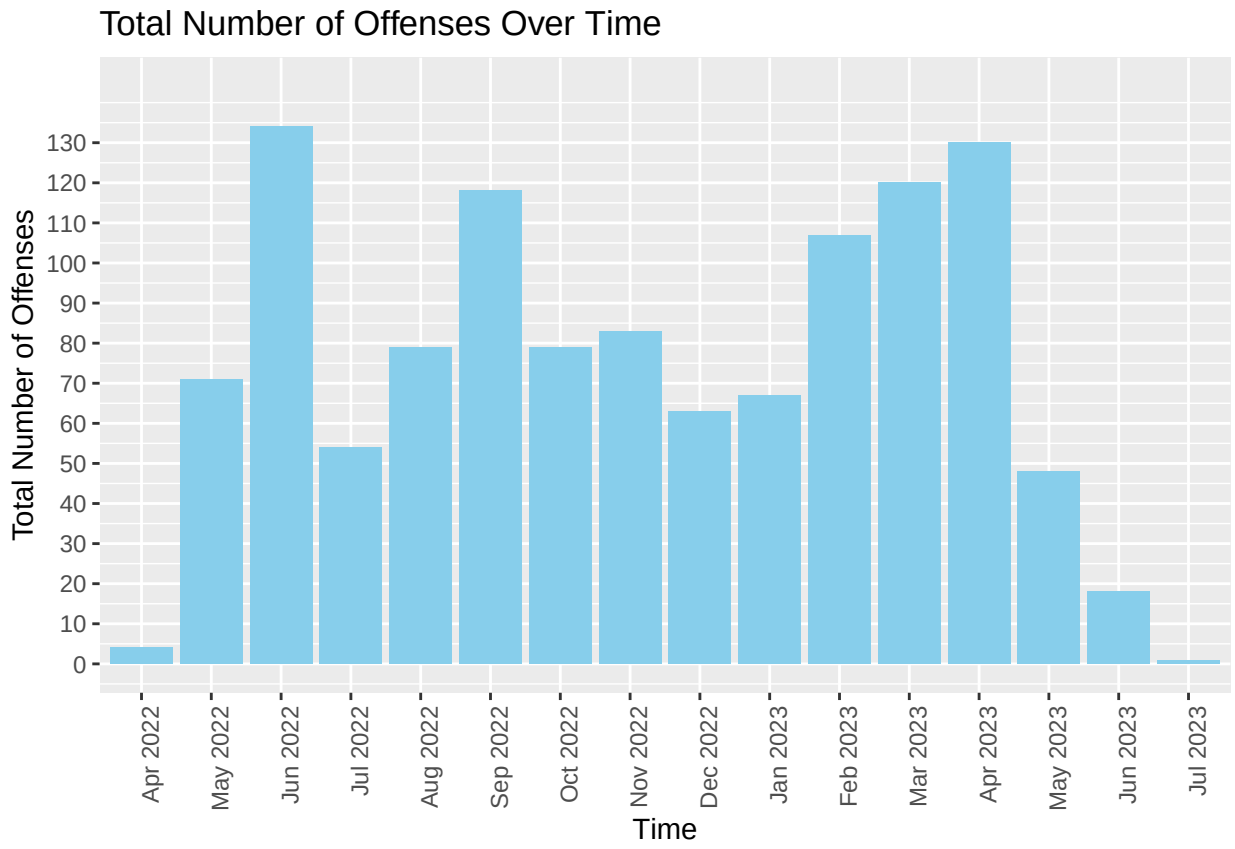
```
# Convert the Date_Issued column to a Date format with the correct format
penalty_data$Date_Issued <- dmy(penalty_data$Date_Issued)

# Extract the Year and Month from the Date_Issued column
penalty_data <- penalty_data %>%
  mutate(Year = year(Date_Issued),
         Month = month(Date_Issued))

# Group the data by Year, Month, and Offence_Code, and calculate the number of offenses and average amo
offenses_by_month_year <- penalty_data %>%
  group_by(Year, Month, Offence_Code) %>%
  summarise(Number_of_Offenses = n(),
            Average_Amount_Payable = mean(Amount_Payable, na.rm = TRUE))

# View the resulting data frame
print(offenses_by_month_year)
```

```
## # A tibble: 89 x 5
## # Groups:   Year, Month [16]
```

```
##     Year Month Offence_Code Number_of_Offenses Average_Amount_Payable
##    <dbl> <dbl>        <int>              <int>                   <dbl>
##  1  2022     4        11339                  4                     880
##  2  2022     5        11322                  1                     660
##  3  2022     5        11338                 13                     440
##  4  2022     5        11339                 52                     880
##  5  2022     5        11341                  2                     880
##  6  2022     5        11343                  2                     880
##  7  2022     5        23086                  1                     660
##  8  2022     6        11338                 23                     440
##  9  2022     6        11339                 94                     880
## 10  2022     6        11340                  1                     440
## # i 79 more rows
```

Answer 2 a)

```r
# Create a time series variable
offenses_by_month_year$time <- as.yearmon(paste(offenses_by_month_year$Year, offenses_by_month_year$Mon

# Group the data by time
grouped_data <- offenses_by_month_year %>%
  group_by(time) %>%
  summarise(Total_Offenses = sum(Number_of_Offenses))


# Plot the number of offenses by time with rotated x-axis labels
ggplot(data = grouped_data, aes( fill = origin, x = factor(time), y = Total_Offenses)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  ylab("Total Number of Offenses") + xlab("Time") +
  ggtitle("Total Number of Offenses Over Time") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 1)) +
  scale_y_continuous(breaks = seq(0, max(grouped_data$Total_Offenses), by = 10), limits = c(0, max(grou
```



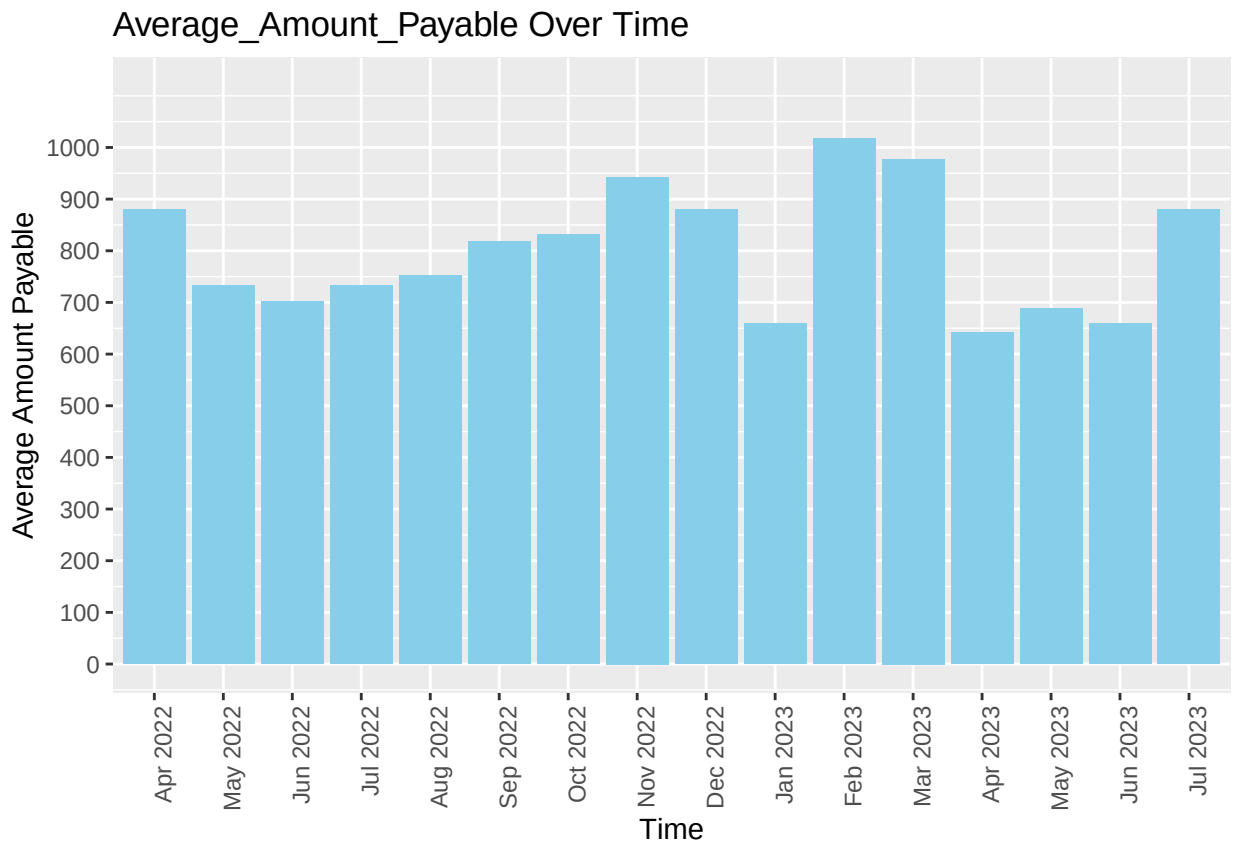
Answer 2 b)

- Trend: From the plot, it appears that there is an increasing trend in the number of offenses over time. The data shows a general upward movement, especially from around February 2023 onwards.
- Seasonality: There may be some seasonality in the data, as there are noticeable peaks and troughs at regular intervals. However, it's not clear whether these patterns are consistent enough to establish a clear seasonality.
- Patterns: There are fluctuations in the number of offenses each month, but it's not clear if these fluctuations follow a specific pattern or are influenced by external factors.

Answer 2 c)

```
# Create a time series variable
offenses_by_month_year$time <- as.yearmon(paste(offenses_by_month_year$Year, offenses_by_month_year$Mon


# Group the data by time
grouped_data <- offenses_by_month_year %>%
  group_by(time) %>%
  summarise(Total_Average_Amount_Payable = mean(Average_Amount_Payable))


# Plot the number of offenses by time with rotated x-axis labels
ggplot(data = grouped_data, aes( fill = origin, x = factor(time), y = Total_Average_Amount_Payable)) +
  geom_bar(stat = "identity", fill = "skyblue") +
```

```
  ylab("Average Amount Payable") + xlab("Time") +
  ggtitle("Average_Amount_Payable Over Time") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 1)) +
  scale_y_continuous(breaks = seq(0, max(grouped_data$Total_Average_Amount_Payable), by = 100), limits =
```



Answer 2 d) The data raises some questions because in April 2022, there were only 4 offenses, but the average amount payable was close to $900. This suggests either heavy fines for those offenses or a potential calculation issue.

When examining the plot of average amounts payable over time, there isn't a clear pattern, but there is a noticeable increasing trend from June 2022 to November 2022.

Comparing it to the number of offenses from April 2022 to July 2023, we see significant fluctuations in offenses by food businesses. However, the average amount fined to these businesses tends to hover between $602 and $1,000.

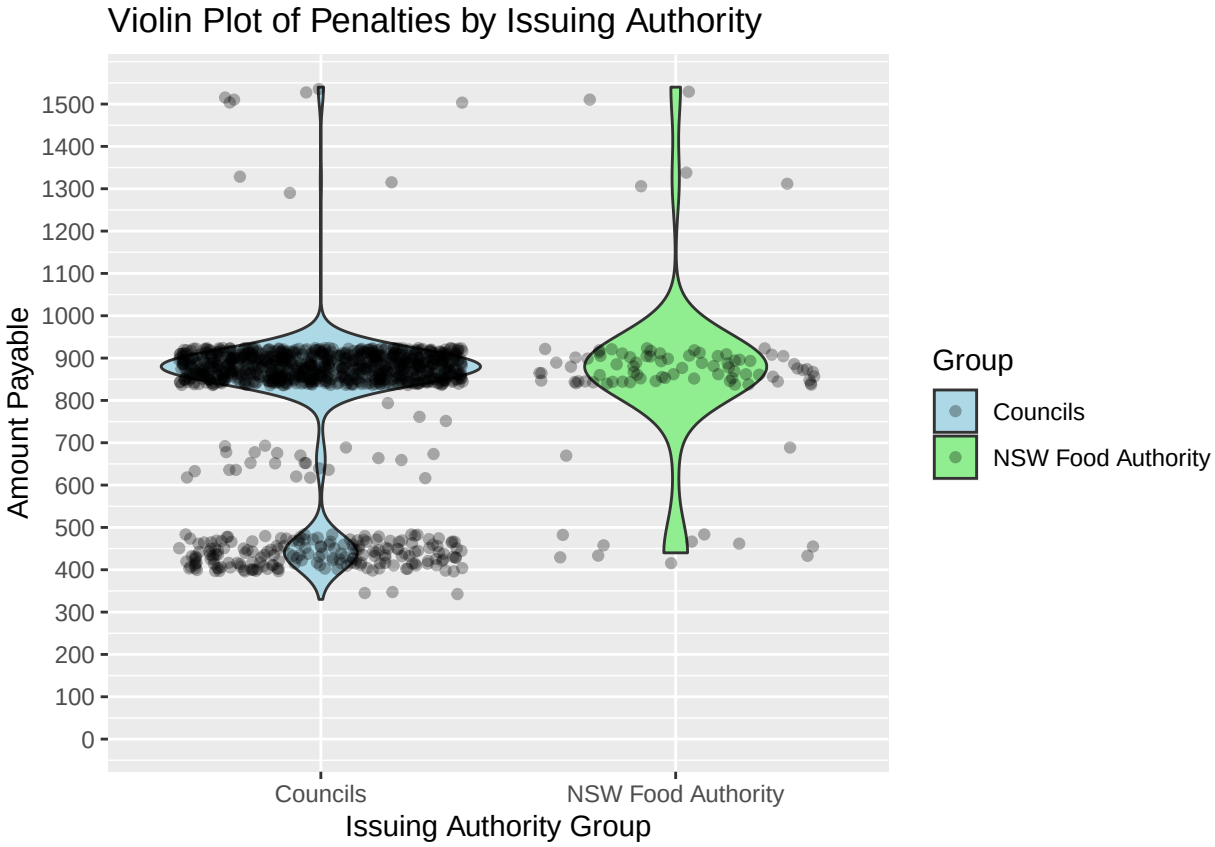
Question 3

Answer 3

Answer 3 a)

```
# Create a new grouping variable to distinguish between Councils and NSW Food Authority
penalty_data$Group <- ifelse(grepl("Council", penalty_data$Issuing_Authority), "Councils", "NSW Food Au

# Create a violin plot with data points
```

```
ggplot(data = penalty_data, aes(x = Group, y = Amount_Payable, fill = Group)) +
  geom_violin() +
  geom_jitter(alpha = 0.3) +
  xlab("Issuing Authority Group") + ylab("Amount Payable") +
  ggtitle("Violin Plot of Penalties by Issuing Authority") +
  scale_fill_manual(values = c("NSW Food Authority" = "lightgreen", "Councils" = "lightblue")) +
  scale_y_continuous(breaks = seq(0, max(penalty_data$Amount_Payable), by = 100), limits = c(0, max(pena
```



Answer 3 b) Upon examining the violin plot comparing Councils and the NSW Food Authority, it is evident that Councils have issued penalties primarily in the ranges of $400-$500 and $800-$950. Additionally, there is a subtle trend where Councils tend to issue penalties in the $600-$750 range. Conversely, the NSW Food Authority also issues penalties in the $850-$950 range, and there is a slight pattern of them issuing penalties in the $400-$500 range as well.

Question 4

Answer 4

Answer 4 a)

```
lgas <- st_read("./LGA_2021_AUST_GDA2020_SHP/LGA_2021_AUST_GDA2020.shp")
```

```
## Reading layer `LGA_2021_AUST_GDA2020' from data source
##   `E:\Uni_Practicals\STAT8123_Statistical_Graphics\Assignment_2\Resources\LGA_2021_AUST_GDA2020_SHP\
##   using driver `ESRI Shapefile'
## Simple feature collection with 566 features and 10 fields (with 19 geometries empty)
```

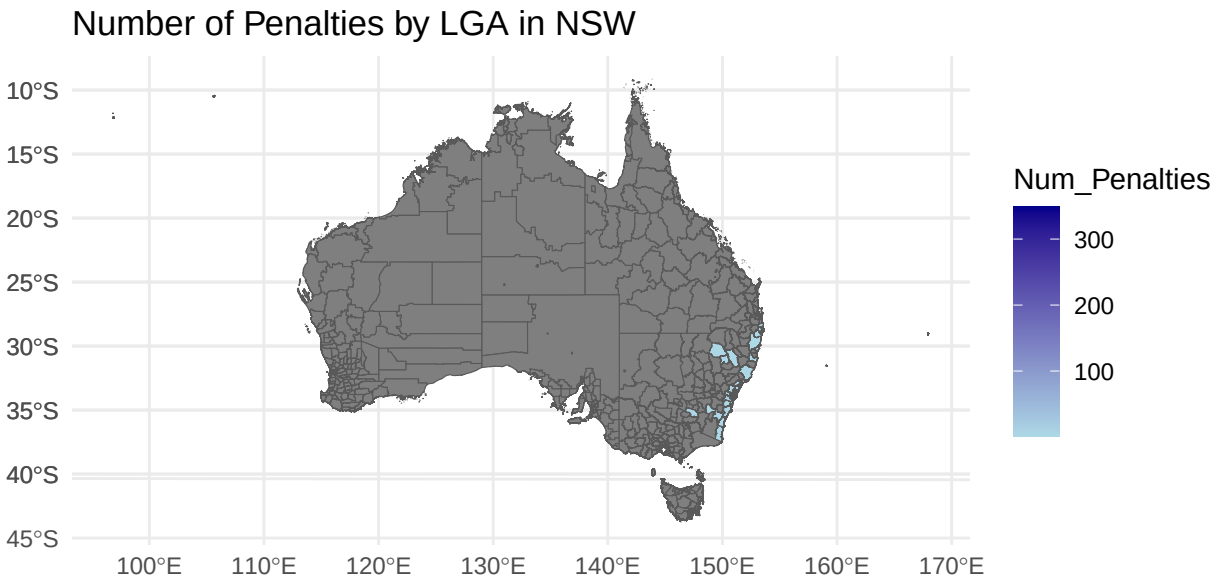
```
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 96.81695 ymin: -43.7405 xmax: 167.998 ymax: -9.142163
## Geodetic CRS:  GDA2020
```

```
library(dplyr)

# Create a mapping table
lga_mapping <- data.frame(
  Dataset_LGA = c(
    "City of Canada Bay",
    "Central Coast",
    "Campbelltown",
    "City of Parramatta",
    "Tamworth",
    "Queanbeyan-Palerang",
    "City of Sydney",
    "Sutherland",
    "The Hills",
    "Bayside",
    "Ku-Ring-Gai"
  ),
  Shapefile_LGA = c(
    "Canada Bay",
    "Central Coast (NSW)",
    "Campbelltown (NSW)",
    "Parramatta",
    "Tamworth Regional",
    "Queanbeyan-Palerang Regional",
    "Sydney",
    "Sutherland Shire",
    "The Hills Shire",
    "Bayside (NSW)",
    "Ku-ring-gai"
  )
)

# Update LGA names in the penalty dataset based on the mapping table
penalty_data <- penalty_data %>%
  left_join(lga_mapping, by = c("Offence_LGA" = "Dataset_LGA")) %>%
  mutate(Offence_LGA = ifelse(is.na(Shapefile_LGA), Offence_LGA, Shapefile_LGA)) %>%
  select(-Shapefile_LGA)
```

```
penalty_counts <- penalty_data %>%
  group_by(Offence_LGA) %>%
  summarize(Num_Penalties = n())
```

```
lgas <- lgas %>%
  left_join(penalty_counts, by = c("LGA_NAME21" = "Offence_LGA"))
```

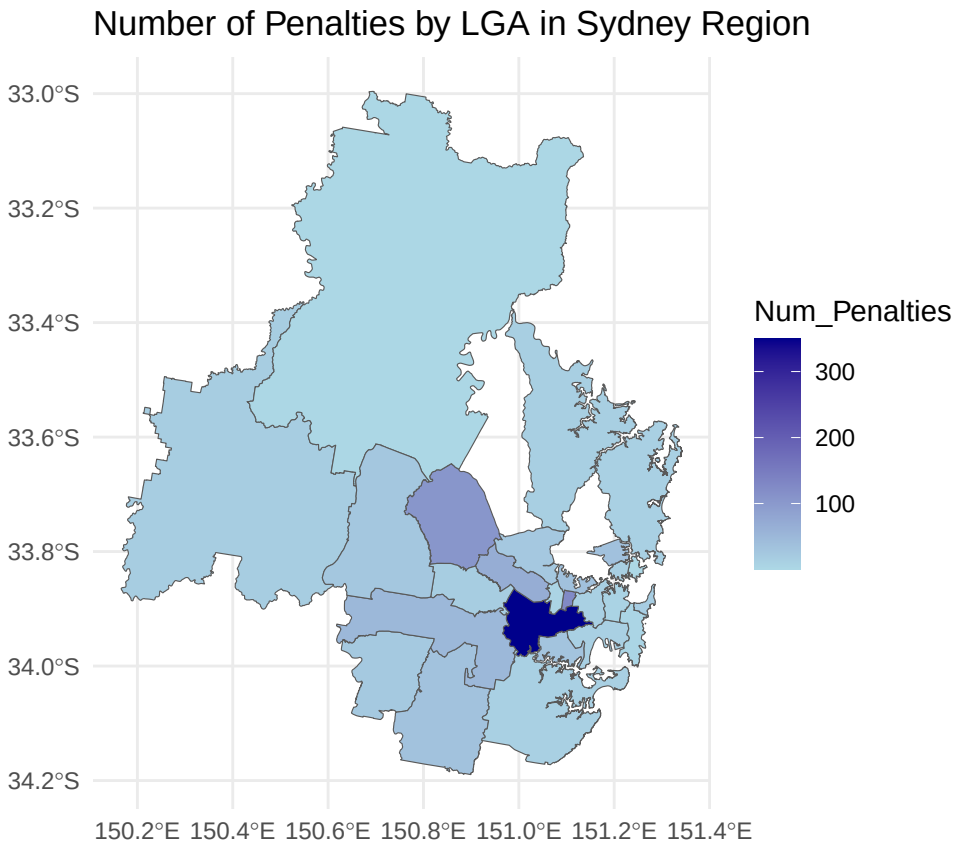
```
ggplot(data = lgas) +
geom_sf(aes(fill = Num_Penalties)) +
scale_fill_gradient(low = "lightblue", high = "darkblue") +
labs(title = "Number of Penalties by LGA in NSW") +
theme_minimal()
```



Answer 4 b)

```
# Define a vector of LGA names for the Sydney region
sydney_region_lgas <- c(
  "Bayside (NSW)",
  "Blacktown",
  "Blue Mountains",
  "Burwood",
  "Camden",
  "Campbelltown (NSW)",
  "Canada Bay",
  "Canterbury-Bankstown",
  "Cumberland",
  "Fairfield",
  "Georges River",
  "Hawkesbury",
  "Hills Shire",
  "Hornsby",
  "Inner West",
  "Liverpool",
  "Mosman",
  "North Sydney",
  "Northern Beaches",
  "Parramatta",
  "Penrith",
```

```
  "Randwick",
  "Strathfield",
  "Sutherland Shire",
  "Sydney",
  "Waverley",
  "Willoughby"
)

# Filter the data to include only LGAs in the Sydney region
sydney_lgas <- lgas %>%
  filter(LGA_NAME21 %in% sydney_region_lgas)

# Create a choropleth map for the Sydney region
ggplot(data = sydney_lgas) +
 geom_sf(aes(fill = Num_Penalties)) +
 scale_fill_gradient(low = "lightblue", high = "darkblue") +
 labs(title = "Number of Penalties by LGA in Sydney Region") +
 theme_minimal()
```



Answer 4 c)

The reason for focusing on the Sydney region becomes evident when we observe that the top five areas with the highest number of issued offenses are all located within the Sydney region. Canterbury-Bankstown, in particular, stands out as the highest with over 350 offenses issued, highlighting the significance of this region in terms of regulatory violations.

Answer 4 d)

The map shows that the number of penalties applied to food businesses in Sydney varies by LGA. The LGAs with the most penalties are Canterbury-Bankstown, Burwood Blacktown, Cumberland and Liverpool. This suggests that there is a correlation between the number of penalties applied to food businesses and the socioeconomic status of the LGA in which they are located.

Question 5

Answer 5

Answer 5 a)

Graphic 1: Word Cloud - Most Frequent Offense Keywords

A word cloud that visually represents the most frequent words or phrases in the offense descriptions. The size of each word in the cloud is proportional to its frequency. This graphic provides an intuitive view of the most common terms associated with offenses.

Graphic 2: Bar Chart - Top Offense Keywords

In this graphic, we create a bar chart that displays the top offense keywords X-axis: Offense keywords Y-axis: Number of Offenses The bars represent different offense keywords, and their heights represent the number of offenses in each category. This graphic provides an overview of the most common types of offenses.

Answer 5 b)

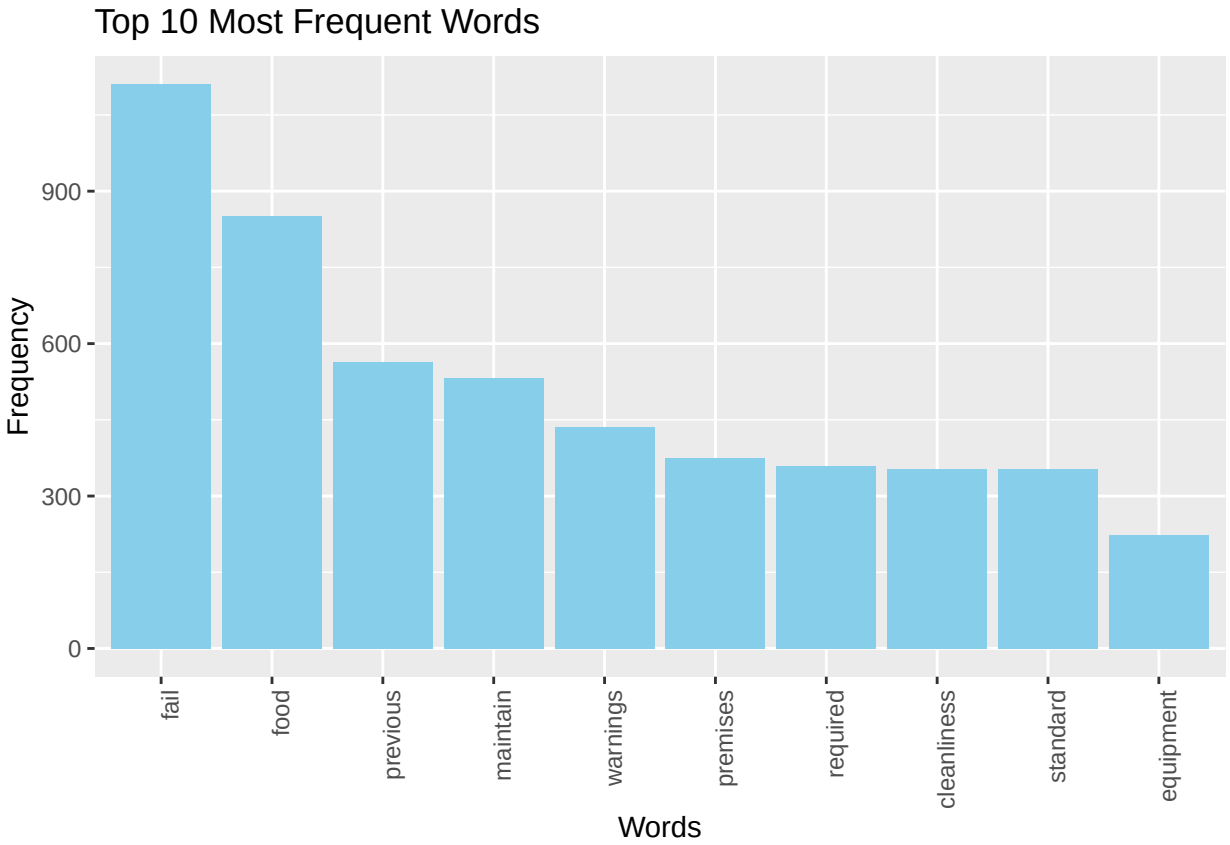
GRAPHIC #1

```
text <- penalty_data$Nature_of_Offence_Full

# Create a corpus
docs <- Corpus(VectorSource(text))

# Define a function to clean and preprocess text
clean_and_preprocess <- function(text) {
  text <- tolower(text)
  text <- removeNumbers(text)
  text <- removePunctuation(text)
  text <- removeWords(text, c("given",stopwords("english")))  # Remove common stopwords
  text <- stripWhitespace(text)
  return(text)
}

# Apply the cleaning and preprocessing function to the corpus
docs_clean <- tm_map(docs, clean_and_preprocess)

dtm = TermDocumentMatrix(docs_clean)
m = as.matrix(dtm)
v = sort(rowSums(m), decreasing = TRUE)
d = data.frame(word = names(v), freq = v)


wordcloud(words = d$word,
          freq = d$freq,
          min.freq = 1,
          max.words = 200,
```

```
          random.order = FALSE,
          rot.per = 0.35,
          colors = brewer.pal(8, "Dark2"))
```

GRAPHIC 2

```
# Sort the word frequency dataframe by frequency in descending order
d <- d[order(-d$freq), ]

# Select the top 10 words
top_10_words <- head(d, 10)


# Create a bar plot using ggplot2
ggplot(data = top_10_words, aes(x = reorder(word, -freq), y = freq)) +
  geom_bar(stat = "identity", fill = 'skyblue') +
  labs(title = "Top 10 Most Frequent Words", x = "Words", y = "Frequency") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels for better visibilit
```


Top 10 Most Frequent Words
Frequency
900
600
300
0
fail
food
previous
maintain
warnings
premises
required
cleanliness
standard
equipment
Words