STAT7123/STAT8123 Assessment Task 2: Visualising Statistical Data

Submission details

Weight: 35%
Due Date/Time: 11:55pm Friday 29th September 2023
Submission: Two versions of your work will be submitted in iLearn. The first will be a **PDF created from your rmarkdown file** which must be submitted to the **turnitin** link, and the second (separate) version will be the original rmarkdown template file in the form surname_IDnumber.rmd which must be submitted to the **rmarkdown file submission link**.

Formatting details

- You must use the provided rmarkdown assignment template to create your submission
- One version must be in **PDF format**
- Your rmarkdown template file must be submitted in the form **surname_IDnumber.rmd**

Late Penalties: Standard Late Penalty applies (see the unit guides/Assessments block of iLearn for details).

Purpose

This assessment allows you to demonstrate your ability to create and interpret statistical graphics that can be used to answer research questions. This is a key capability in all professions whose role may entail creating and explaining such graphics to non-experts.

Outcomes addressed

This assessment addresses the following unit outcome/s:
ULO3: use the computer to generate appropriate graphics using particular packages or languages and be able to develop the ability to do so in others.
ULO4: be familiar with a range of modern multivariate graphical techniques and know when it is appropriate to use them.
ULO5: use statistical graphics to investigate and analyse data, check statistical model assumptions and effectively present the results of statistical investigations graphically to a range of audiences.

Skills assessed

Using the methods and techniques described in Lectures and SGTAs, this task allows you to demonstrate:

1. Your written communication skills
2. Your ability to answer research questions using suitable statistical graphics
3. Your ability to create different statistical graphics
4. Your proficiency in using rmarkdown to present results.

Task overview

For this assessment, you will answer a series of research questions using R statistical software with the tidyverse suite of commands. You will present the results via a PDF generated by compiling an rmarkdown template file. You will also submit the original file with descriptions of code intent (annotations) within each code chunk. Note that you must use tidyverse to summarise the data where necessary and ggplot to create the graphics. Marks will be deducted where this is not done. The questions begin on the page 3.

Quality Criteria

A high-quality submission will:

1. Use the tidyverse suite of commands to summarise and manipulate data
2. Present professional-quality graphics using ggplot
3. Provide accurate interpretations of each graph where required
4. Provide detailed and clear written responses where required
5. Deliver an rmarkdown file that can be compiled to produce the submitted PDF and which contains brief descriptions of code intent (annotations) within each code chunk.

Task details/specific instructions

Context

In Australia, the food industry is big business! Encompassing restaurant businesses and businesses that grow and /or process food to sell to supermarkets and restaurants, the food industry is a significant contributor to the economy. In 2018, the food industry contributed more than $187 billion towards Australia's GDP.[1]

Given the size and importance of the food industry, it is important to ensure that food is safe to eat. As such, the food industry is carefully regulated with rules and procedures that provide minimum standards that businesses must meet.

In New South Wales, the NSW Food Authority regulates and manages food safety (https://www.foodauthority.nsw.gov.au/media/3196). Despite clear processes and procedures that inform businesses on mandated standards (https://www.foodstandards.gov.au/code/Pages/default.aspx), some businesses fail to abide by the rules. The NSW Food Authority publishes a list of the businesses that breach the law in their Name and Shame list (https://www.foodauthority.nsw.gov.au/offences).

Files

In the assignment 2 folder you can find data (called "`penalty_notice.csv`") supplied by the NSW Food Authority from the Name and Shame site showing the following the following information:

Variable	Description
`Infringement_Number`	`Unique number (ID) of infringement`
`Trading_Name`	`Name of business`
`Issuing_Authority`	`NSW food authority or council issuing penalty`
`Date_Issued`	`Date penalty notice issued`
`Nature_of_Offence_Full`	`Detailed description of offence`
`Offence_Code`	`Code for the offence`
`Offence_Description`	`Offence code description`
`Offence_Date`	`Date of offence`
`Year`	`Year of offence`
`Month`	`Month of offence`
`Published_Address`	`Address of business`
`Postcode`	`Postcode of business`
`Amount_Payable`	`Penalty amount ($)`
`Offence_LGA`	`Local government area (LGA) of offence`

The assignment folder also contains a zip file containing LGA shapefile information that you may need to answer one or more questions.

Notes

When analysing data, it is important to explore the data graphically before or in conjunction with statistical tests. In this assignment you **must not undertake any statistical tests**. The purpose of this assignment is to **create graphics that would assist in answering research questions**. The questions that need to be answered begin on the following page.

[1] https://research.csiro.au/foodag/economy/

Question 1 [6 marks]

Use tidyverse commands to calculate the total number of offences by local government area (LGA).

a) Using 3-5 sentences, describe what features of this summary make the data challenging to plot.

b) Which LGA has the most offences?

c) Use an appropriate plot to display a subset of the LGAs (do not use a map here) that could be used to answer the research question: "*Which LGAs have the highest number of offences?*" Hint: the `filter()` command could be useful here.

Question 2 [10 marks]

Use tidyverse commands to calculate the number of offences and average amounts payable by month per year.

a) Plot the number of offences by time

Hint: you may need the zoo library with the following command, where df is the relevant data frame:

```
df$time <- as.yearmon(paste(df$year, df$month), "%Y %m")
```

b) In 3-5 sentences, interpret your plot to answer the research question "*Are there any trends, patterns or seasonality in the number of events over time?*"

c) Plot the average amount payable by time

d) In 3-5 sentences, interpret your plot to answer the research question "*Are there any trends, patterns or seasonality in average amounts payable over time?*"

Question 3 [5 marks]

Using all of the data:

a) Create a violin plot that includes the individual data points to determine if there is a difference in the penalties (amounts payable) issued by the NSW Food Authority versus Councils.

Hint: the commands given in this page may be useful here:
https://www.sfu.ca/~mjbrydon/tutorials/BAinR/recode.html

b) Interpret the plot to answer the research question: "*Is there a difference in the amounts payable issued by the NSW Food Authority compared to Councils?*"

Question 4 [7 marks]

Using all of the data,

a) Create and present a map that shows the number of penalties by LGA in NSW (this could be a choropleth map or a dot density map).

Hint: The code to remove the additional characters at the end of the LGA names is:

```
mutate(NAME = str_remove(NAME, " \\(.+\\)"))
```

Alternately (and maybe better!) you can use the shapefile provided by unzipping the contents of the included zip file and read the shapefile in with the command (from the sf library):

```
st_read("LGA_2021_AUST_GDA2020.shp")
```

b) Focus the map on the Sydney region

c) In one sentence, explain why we may want to focus on the Sydney region.

d) In about 3 sentences, summarise the main results of the plot presented in part (c).

Question 5 [18 marks]

The data contains a column showing a detailed description of the offence (`Nature_of_Offence_Full`). The description is quite long, and so it cannot be used directly in plots. However, it is of interest to summarise. It may be necessary to simplify the information contained in this column *without* losing important information about the nature of the offence.

a) Describe in detail, using a series of bullet points, how you could handle (simplify) the data in this column so that it can be graphically presented.

b) [5 marks per graphic = total of 15 marks available]. Complete your data handling (you can ask for help from your conveners with this), and create and present **three** statistical graphics. The graphics should display the nature of the offence, along with other relevant variables that you think provide insight into the data. For each plot presented, explain in detail (about 100 words) what the plot shows so that a non-expert can understand the pattern/trend.

Note: Higher marks will be awarded for:

- Using more complex graphics to represent relationships (e.g., plots where you have used more data manipulation, added more layers, judiciously used faceting)
- Using a variety of graphics (e.g., three different graphics, rather than all in the same format)
- Using multiple variables in one plot to provide greater insight into the data (e.g., comparing trends)