MAS2908 - Report (Solutions)

Clement Lee

Semester 2, 2025/2026

Instructions

1. This report assesses your ability to create basic data visualisations, interpret them appropriately, and embed them in dynamic document generation using R Markdown.
2. You need to create plots using `ggplot2` to answer the questions.
3. You need to submit a PDF (maximum 3 pages) and the R Markdown file that generates the PDF. There are marks allocated for following these instructions – see the last question.
4. Submission deadline is Feb 18th (Wed) 4pm, unless you have been granted an extension.
5. This assignment has a full marks of 20, and accounts for 10% of the whole module mark.

Enabling PDF output (if you haven't done so in practical)

To generate PDF documents from R Markdown, you need a LaTeX distribution installed on your computer. The easiest way to do this is to use the `tinytex` package, which installs a lightweight TeX distribution:

```
# Run these commands ONCE in your R console (not in an Rmd file)
install.packages("tinytex")
tinytex::install_tinytex()
```

This is a one-off operation — you only need to do it once per computer. After installation, R Markdown will automatically use this distribution to compile PDF documents. If you only need HTML output, you can skip this step.

Data & Setup

You will use the built-in `iris` dataset, which contains measurements of 150 iris flowers from three species: setosa, versicolor, and virginica. The variables are:

- `Sepal.Length`: Length of the sepal, in cm
- `Sepal.Width`: Width of the sepal, in cm
- `Petal.Length`: Length of the petal, in cm
- `Petal.Width`: Width of the petal, in cm
- `Species`: The species of iris (setosa, versicolor, or virginica)

First, familiarise yourself with the dataset using the following example commands in the RStudio console. You do not need to (and should not) include them in the your submission PDF.

```
head(iris)
str(iris)
?iris
```

You might want to include the following code chunk towards top of the R Markdown (Rmd) file, just below the lines ringfenced by the triple dashes (`---`):

````
```{r setup}
#| echo: false
library(ggplot2)
knitr::opts_chunk$set(
  echo = FALSE,
  out.width = "60%",
  fig.align = "center",
  fig.asp = 0.7
)
```
````

The `echo = FALSE` is so that space is saved by hiding the code in the PDF, as I will be able to see that in the Rmd file.

Questions

1. **(4 marks)** Figure 1 attempts to visualise the distribution of `Sepal.Width` by `Species`.

   a. Comment on two different aspects that make this a poor plot.
   b. Improve this plot by using a different `geom_*()` and making necessary changes to the chunk options.

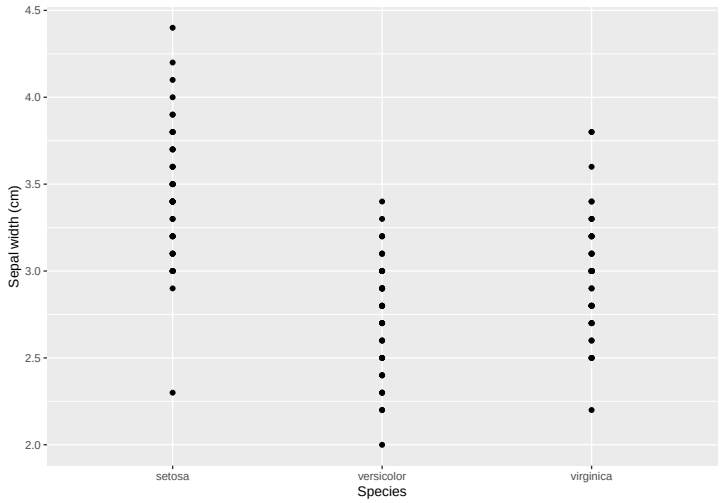

Figure 1: Scatterplot of the species and width of the sepal.

Answer:

   a. The font size is too small; for each species, there are repeated values of the sepal width, rendering the scatterplot not useful to show the distribution.
   b. To increase the font size of the plot to one that is similar to that of normal text, either increase 'out.width' or 'fig.width'. To show the distribution by species, use boxplot or stripchart (jitter plot):

```
ggplot(iris, aes(Species, Sepal.Width)) +
  geom_boxplot() +
  labs(y = "Sepal width (cm)")
```

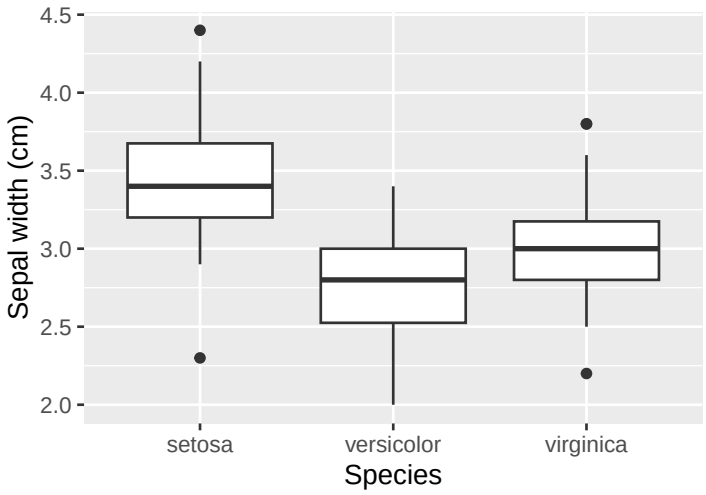

Figure 2: Boxplot of the species and width of the sepal.

2. **(4 marks)** The boxplot of `Petal.Length` is provided in Figure 3.

   a. By providing an alternative plot that only involves `Petal.Length` and not other variables, argue in words why the boxplot does not tell the whole picture of this variable's distribution.
   b. Explain, in words, the main reason that `Petal.Length`'s distribution is not unimodal and symmetric, and provide a plot (that is not the same as part a) to illustrate your argument.

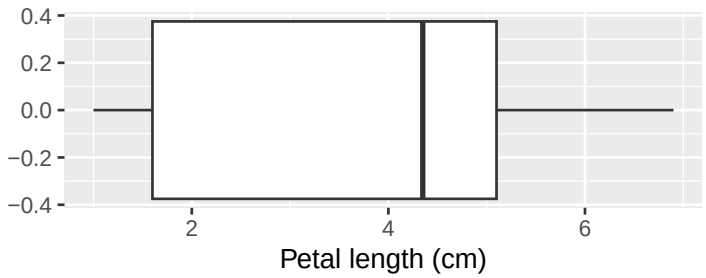

Figure 3: Boxplot of the length of the petal.

Answer: a. The histogram (or density plot) reveals that the distribution is bimodal.

```
ggplot(iris, aes(Petal.Length)) +
  geom_histogram(binwidth = 0.2) +
  labs(x = "Petal length (cm)")
```

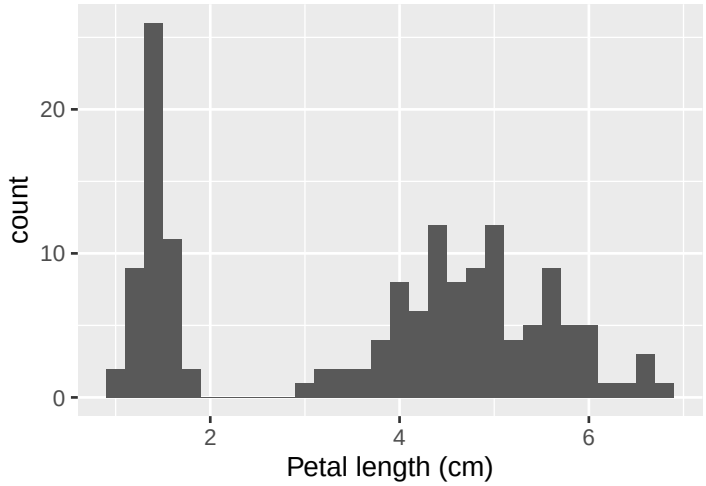

Figure 4: Histogram of the length of the petal.

Answer: b. The mean and standard deviation vary across species. A boxplot by species reveals this. A histogram or density plot by species (using e.g. colours) is also acceptable.

```
ggplot(iris, aes(Species, Petal.Length)) +
  geom_boxplot() +
  labs(y = "Petal length (cm)")
```

3. **(3 marks)** Find a pair of variables that are strongly linearly correlated, and do the following in one single plot:

   a. create a scatterplot of the two variables, and
   b. overlay the straight line of best fit.

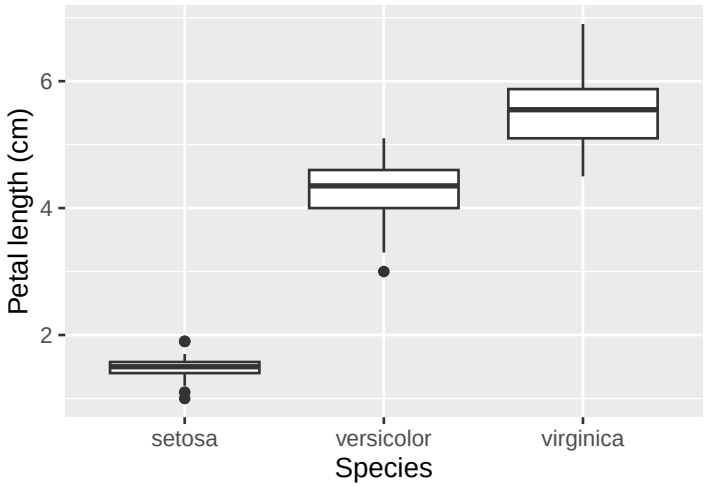

Figure 5: Boxplot of the length of the petal by species.

Answer: a. Petal length and petal width are most strongly correlated amongst all pairs. Also acceptable are petal length and sepal length, or petal width and sepal length.

```
ggplot(iris, aes(Petal.Length, Petal.Width)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(y = "Petal width (cm)", x = "Petal length (cm)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

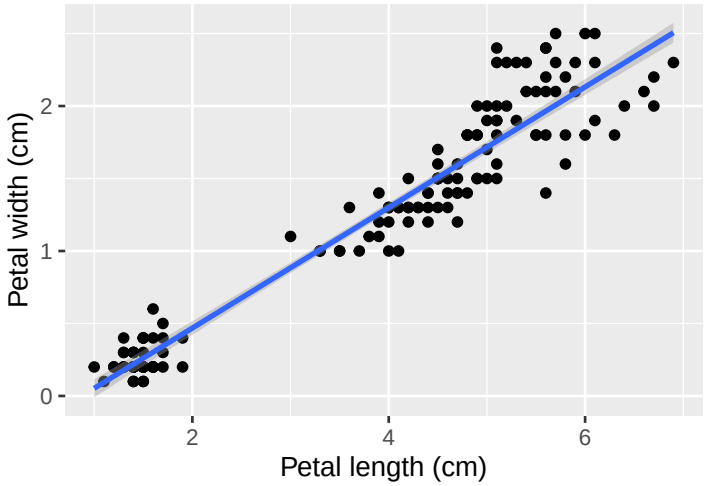

Figure 6: Scatterplot of the width against the length of the petal.

Answer: b. `geom_smooth()` without `method = 'lm'` is not accepted, but `geom_abline()` with the correct slope and intercept is accepted.

4. **(9 marks)** Overall:

   a. Your submission should contain a single, coherent report **in PDF, and the Rmd file that generates the report**.
   b. The page limit for the report is 3 pages. Any reports going over this limit will be penalised.
   c. Ensure that you have included sensible axis labels, figure captions, and font size in the plots.

Answer: Full marks will be given by following all these rules.