Data Analysis Skills: Practice Class Test Marking Scheme

Task 1. Report on Data Analysis

- Appropriate Title and Student Number **1 MARK**

Please Note: the code chunks and the mathematical LaTeX code (`$` and `$$`) have been included in Task 1 to show you how the output included in the report was generated. In the final `.pdf` file the code chunks and the code betwen `$$` SHOULD NOT BE SHOWN for Task 1 (but should be shown for Task 2).

```
library(tidyverse)
library(moderndive)
library(skimr)
library(kableExtra)
library(gridExtra)
```

```
cats <- read.csv("cats.csv")
```

Introduction

- Introduction to the data being analysed and to the question of interest. Marks deducted for copying the data description as given. **3 MARKS**

Exploratory Data Analysis

- Summary statistics on heart weight by sex with appropriate comments. One mark removed if the output is simply 'copy-pasted' from R.

```
cats %>%
  group_by(Sex) %>%
  summarise(n=n(),Mean=round(mean(Hwt),digits=1), St.Dev=round(sd(Hwt),digits=1),
            Min=min(Hwt), Q1 = quantile(Hwt,0.25), Median=median(Hwt),
            Q3 = quantile(Hwt,0.75), Max=max(Hwt)) %>%
  kable(caption = '\\label{tab:summaries} Summary statistics on
        heart weight by sex of 144 adult cats.') %>%
  kable_styling(latex_options = "hold_position")
```

Table 1: Summary statistics on heart weight by sex of 144 adult cats.

Sex	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
F	47	9.2	1.4	6.3	8.35	9.1	10.1	13.0
M	97	11.3	2.5	6.5	9.40	11.4	12.8	20.5

2 MARKS

- Comments on the summary statistics related to the question of interest. **1 MARK**
- Boxplot of heart weight by sex. One mark removed if the plot is not appropriately labelled, and axis labels not adjusted accordingly.

```
```{r boxplot, out.width = '68%', fig.align = "center",
fig.cap = "\\label{fig:box} Heart weight by Sex.", fig.pos = 'H'}
ggplot(cats, aes(x = Sex, y = Hwt)) +
  geom_boxplot() +
  labs(x = "Sex", y = "Heart weight (grams)",
       title = "Heart weights of 144 adult cats")
```
```

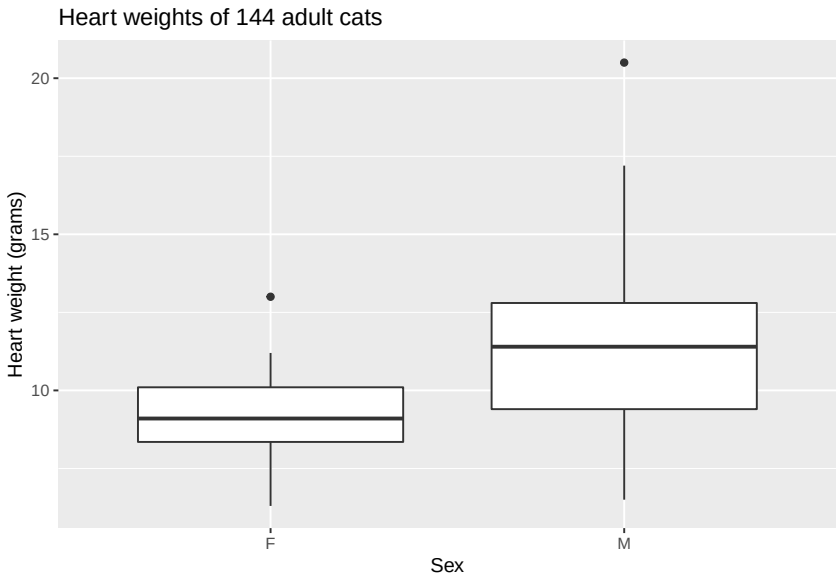

Figure 1: Heart weight by Sex.

2 MARKS

- Comments on the boxplot related to the question of interest. **2 MARKS**

Formal Data Analysis

- State the linear regression model being fitted, i.e.

$$\widehat{\text{Hwt}} = \widehat{\alpha} + \widehat{\beta}_{\text{Male}} \cdot \mathbb{I}_{\text{Male}}(x)$$

```
$$\widehat{\mbox{Hwt}} = \widehat{\alpha} +
\widehat{\beta}_{\mbox{Male}} \cdot \mathbb{I}_{\mbox{Male}}(x) $$
```

where

- the intercept $\widehat{\alpha}$ `$\widehat{\alpha}$` is the mean heart weight for the baseline category of Females;
- $\widehat{\beta}_{\text{Male}}$ `$\widehat{\beta}_{\mbox{Male}}$` is the difference in the mean heart weight of a Males relative to the baseline category Females; and
- $\mathbb{I}_{\text{Male}}(x)$ `$\mathbb{I}_{\mbox{Male}}(x)$` is an indicator function such that

$$\mathbb{I}_{\text{Male}}(x) = \begin{cases} 1 & \text{if Sex of } x\text{th observation is Male,} \\ 0 & \text{Otherwise.} \end{cases}$$

```
$$\mathbb{I}_{\mbox{Male}}(x)=\left\{
            \begin{array}{ll}
              1 ~~~ \mbox{if Sex of} ~ x \mbox{th observation is Male},\\
              0 ~~~ \mbox{Otherwise}.\\
            \end{array}
          \right.$$
```

3 MARKS

- Report the estimated model coeffecients. One mark removed if the regression output is simply 'copy-pasted' from R.

```
model <- lm(Hwt ~ Sex, data = cats)
```

```
get_regression_table(model) %>%
  dplyr::select(term,estimate) %>%  #Note that it seems necessary to include dplyr:: here!!
  kable(caption = '\\label{tab:reg} Estimates of the parameters from the fitted linear
    regression model.') %>%
  kable_styling(latex_options = 'HOLD_position')
```

Table 2: Estimates of the parameters from the fitted linear regression model.

term	estimate
intercept	9.202
SexM	2.121

3 MARKS

- Appropriate comments on the regression coefficients and the difference between males and females. **4 MARKS**

NB: THE DIAGNOSTICS IN THE REMAINDER OF THIS ANALYSIS SECTION ARE NOT REQUIRED FOR THE CLASS TEST

- Plots for checking model assumptions. One mark removed if not properly labelled.

```
```{r residplots, echo=FALSE, fig.width = 13, fig.align = "center",
fig.cap = "\\label{fig:resids} Scatterplots of the residuals by Sex (left)
and a histogram of the residuals (right).", fig.pos = 'H', message = FALSE}
regression.points <- get_regression_points(model)
p1 <- ggplot(regression.points, aes(x = Sex, y = residual)) +
      geom_jitter(width = 0.1) +
      labs(x = "Sex", y = "Residual") +
      geom_hline(yintercept = 0, col = "blue")

p2 <- ggplot(regression.points, aes(x = residual)) +
      geom_histogram(color = "white") +
      labs(x = "Residual")

grid.arrange(p1, p2, ncol = 2)
```
```

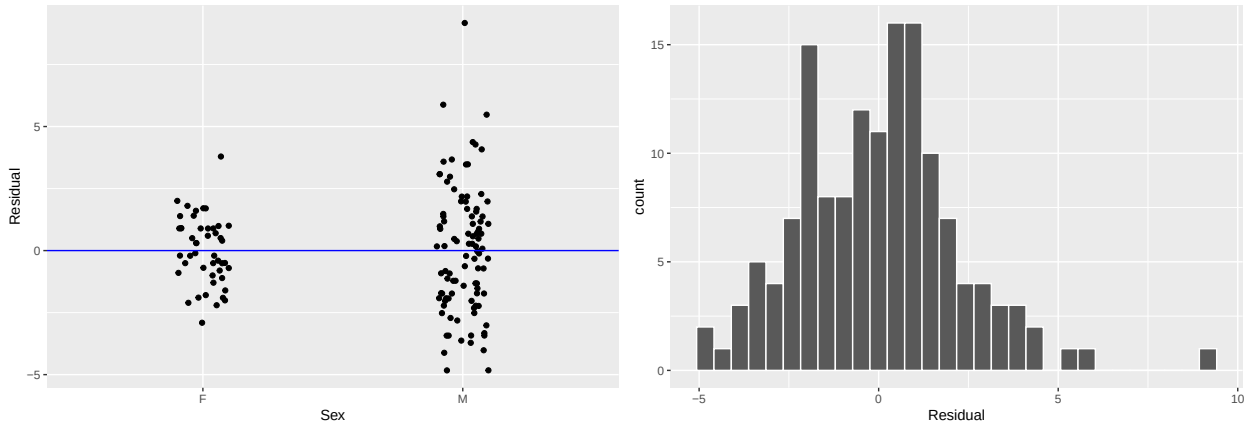

Figure 2: Scatterplots of the residuals by Sex (left) and a histogram of the residuals (right).

- Appropriate comments on the model assumptions.

Conclusions

- Overall conclusions with an answer to the question of interest.

2 MARKS

- General report layout. This include figure and table captions, labeling, positioning and quality of English.

2 MARKS

Total: 25 MARKS

Task 2. Further Task

Further Task Part a.

```
Glasgow_Ed_SIMD2020 <- read_csv("Glasgow_Edinburgh_SIMD2020.csv")
```

This data is not in `tidy` format since the measurement of interest is 'rank' of which there are 8 types (i.e. SIMD, Income, Employment, Health, Education, Access, Crime and Housing) spread over 8 columns. In `tidy` format, the 'rank' measurements should be in a single column, with a separate column indicating the type of 'rank'.

2 MARKS

To convert the data into a `tidy` format, use

```
Glasgow_Ed_SIMD2020_tidy1 <- gather(data = Glasgow_Ed_SIMD2020,
                     key = Type_of_Rank,
                     value = Rank,
                     SIMD_Rank:Housing_Rank)

Glasgow_Ed_SIMD2020_tidy1$Type_of_Rank <-
  str_replace(Glasgow_Ed_SIMD2020_tidy1$Type_of_Rank, "_Rank", "")
```

or

```
Glasgow_Ed_SIMD2020_tidy2 <- gather(data = Glasgow_Ed_SIMD2020,
                     key = Type_of_Rank,
                     value = Rank,
                     -(Data_Zone:Working_Age_population))

Glasgow_Ed_SIMD2020_tidy2$Type_of_Rank <-
  str_replace(Glasgow_Ed_SIMD2020_tidy2$Type_of_Rank, "_Rank", "")
```

3 MARKS

Further Task Part b.

```
  Gla_Ed_SIMD2020 <- Glasgow_Ed_SIMD2020_tidy2 %>%
                       filter(Type_of_Rank == "SIMD") %>%   #2 MARKS
                       mutate(Perc_Working = 100 * Working_Age_population/Total_population)  #1 MARK
```

```
ggplot(Gla_Ed_SIMD2020)+
  geom_point(mapping=aes(x=Perc_Working,y=Rank,group=Council_area,color=Council_area))+  #3 MARKS
  labs(x="Employment Rate of Working Age Population",y="SIMD2020 Rank")  #1 MARK
```

Figure 3: SIMD Rank against Percentage of working age population working for Glasgow and Edinburgh Data Zones

1 MARK (for including the plot in `.pdf`)

Total: 13 MARKS

Task 3. File Uploads

2 MARKS for uploading appropriate .pdf and .Rmd files

- Deduct these two marks if R code appeared in the Report (Task 1) or if relevant R code DID NOT appear in Task 2