Instructions

There are 9 questions worth a total of 100 marks. You should attempt them all.

QUESTION 1

This question is about simple descriptive statistics, and rearranging data.

colour	koala.NSW	koala.VIC	bilby.NSW	bilby.VIC
grey	23	43	11	8
cream	56	89	22	17
white	35	72	13	6
black	28	44	19	16
taupe	25	37	21	12

(a) Compute the total number of koalas in NSW. ___

[1 marks]

(b) Compute the proportion of koalas in NSW (relative to both NSW and VIC). ___

[1 marks]

(c) Are there relatively more koalas than bilbies?

[1 marks]

(d) How many variables are in this data set? List them. ___

[2 marks]

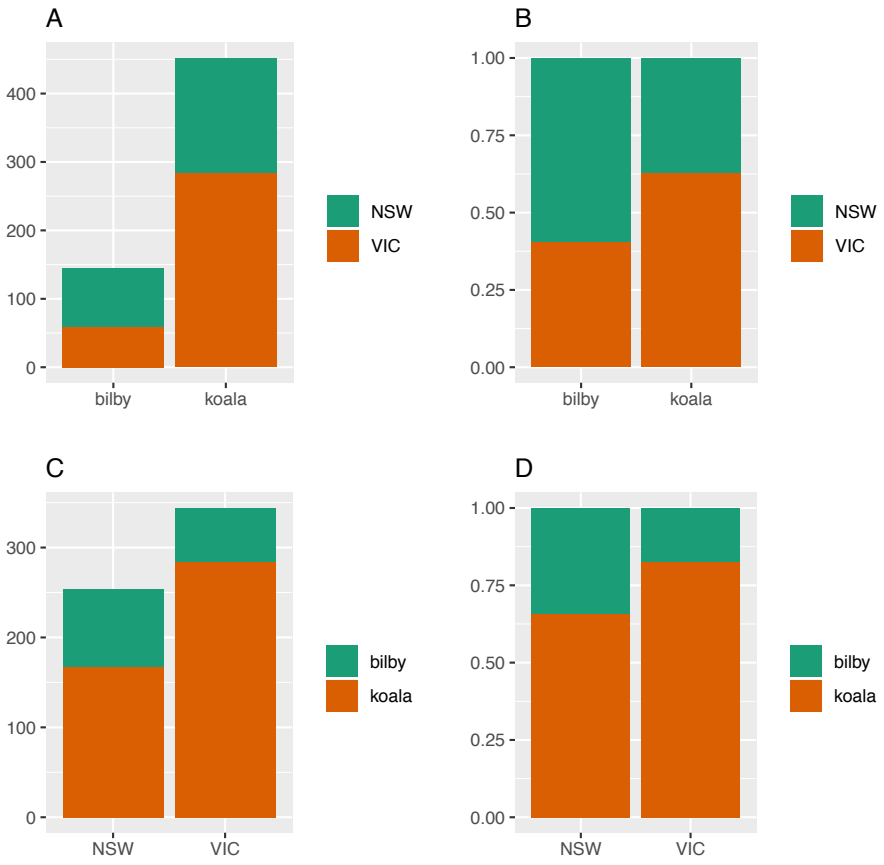
(e) Map out the steps, using the tidy verbs, that you would need to make on this data to get it into long tidy form. The end result should have these columns:

[2 marks]

colour	*marsupial*	*state*	*count*

(f) Match the plot to the question that it would best help answer?

[3 marks]



- Is the proportion of koalas higher in VIC or NSW? ___
- Are there more koalas in VIC or NSW? ___
- Does VIC have a higher proportion of bilbies or koalas? ___
- Are there more bilbies or koalas in VIC? ___

[Total: 10 marks]

— END OF QUESTION 1 —

QUESTION 2

This question is about wrangling data, verbs, definitions and usage, and working with temporal variables.

```
# A tibble: 12,144 x 3
   date       halfhour   kwh
   <date>        <dbl> <dbl>
 1 2017-11-24      0.5 0
 2 2017-11-25      0.5 0
 3 2017-11-26      0.5 0
 4 2017-11-27      0.5 0
 5 2017-11-28      0.5 0
 6 2017-11-29      0.5 0
 7 2017-11-30      0.5 1.06
 8 2017-12-01      0.5 1.19
 9 2017-12-02      0.5 0.063
10 2017-12-03      0.5 0.019
...
```

(a) What type of data object does R think this is? (Circle one)

[1 marks]

`data.frame` `tibble` `matrix` `ts`

(b) You need to convert date into several new variables, year, month and day of the week. What wrangling verb do you need to use? (Circle one)

[1 marks]

`filter` `arrange` `select` `mutate` `summarise` `group_by` `count` `tally`

(c) You want to compute the total kwh used each day. What wrangling verb do you need to use? (Circle one)

[1 marks]

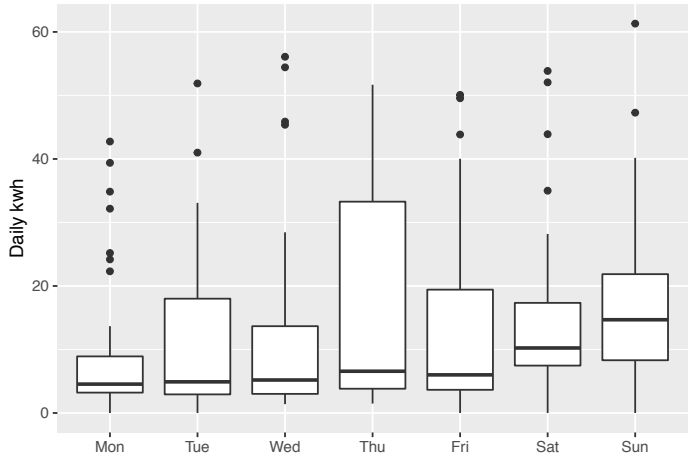
`filter` `arrange` `select` `mutate` `summarise` `group_by` `count` `tally`

(d) You want to compute the maximum daily kwh. What wrangling verb(s) do you need to use? (Circle one or more)

[1 marks]

`filter` `arrange` `select` `mutate` `summarise` `group_by` `count` `tally`

(e) In the plot below, a boxplot of daily total kwh is shown by day of the week. Data for approximately a year has been collected.



(i) What day has the highest median electricity usage? ___

[1 marks]

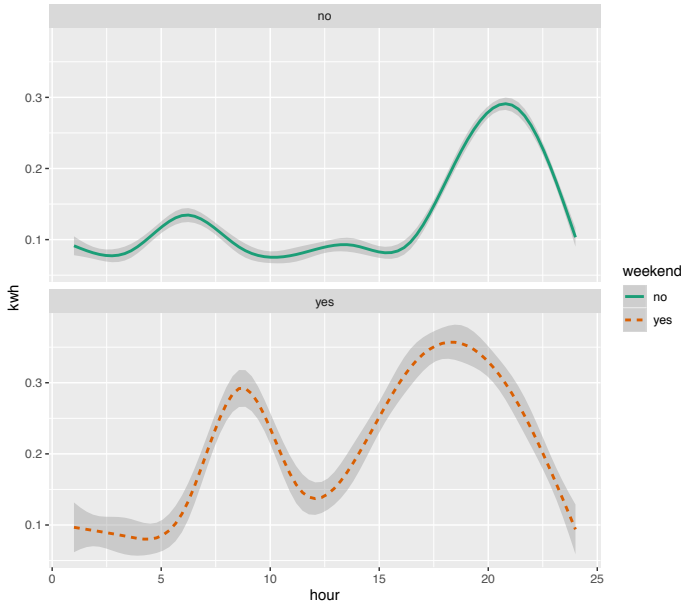
(ii) What day has the highest variability in electricity usage, as measured by the IQR? ___

[1 marks]

(iii) What day(s) of the week have symmetric distributions in energy use, as indicated by the boxplots? ___

[1 marks]

(f) The plot below shows smoothed kwh usage by hour of the day, across several years of data from a single household, coloured by weekend vs weekday.



(i) Which type of day (weekend or week day) has the highest average hourly usage? ___

[1 marks]

(ii) What time of day is typically peak usage on a weekday? and a weekend? ___

[2 marks]

(iii) Write a paragraph describing this person's electricity use, on a week day relative to the weekend.

[2 marks]

(iv) What does the grey band indicate for this data? ___

[1 marks]

(v) Why do you think that the facets of this plot were stacked, instead of placed side-by-side? (Choose one) ___

[1 marks]

A. To compare the temporal pattern between types of day.

B. Compare the overall magnitude of electricity usage between types of day.

C. For no particular reason.

[Total: 14 marks]

— END OF QUESTION 2 —

QUESTION 3

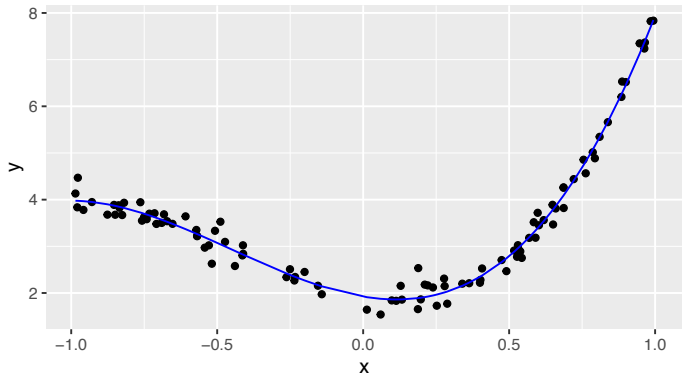
This question is about fitting models using optimisation.

We have some data, that is simulated from this model

$$y = 2 - x + 4 * x^2 + 3 * x^3 + \varepsilon$$

where $x \in (-1, 1)$ and $\varepsilon \sim N(0, 0.2^2)$.

The following plot shows the data, and a fitted model.



The model was fitted by minimizing the least square error, using this code:

```
square_err <- function(par, data) {
  sq <- sum((data$y-(par[1]+par[2]*data$x+par[3]*data$x^2+par[4]*data$x^3))^2)
  return(sq)
}
fit <- optim(c(1,1,1,1), square_err, data=df)
df <- df %>% mutate(fitted = fit$par[1] + fit$par[2]*x +
                      fit$par[3]*x^2 + fit$par[4]*x^3)
ggplot(df, aes(x=x, y=y)) + geom_point() +
  geom_line(aes(y=fitted), colour="blue")
```

(a) Write down the equation for least square error. _____

[2 marks]

(b) The fitted model has these parameter estimates:

```
$par
[1]  1.922487 -1.037443  4.038679  3.020152
```

(i) How many parameters in the model? __

[1 marks]

(ii) What is the parameter estimate for the quadratic term in the model? __

[1 marks]

(iii) If $x = 1$, what is the predicted value of y? __

[1 marks]

(iv) If $x = 0$, and $y = 3$, what is the residual from the model? __

[1 marks]

(c) Suppose you were asked to use least absolute error instead. What would need to be changed in the code to achieve the model fit?

[2 marks]

(d) If you fitted a simple linear model to this data, sketch what you think the residuals vs fitted plot would look like.

[2 marks]

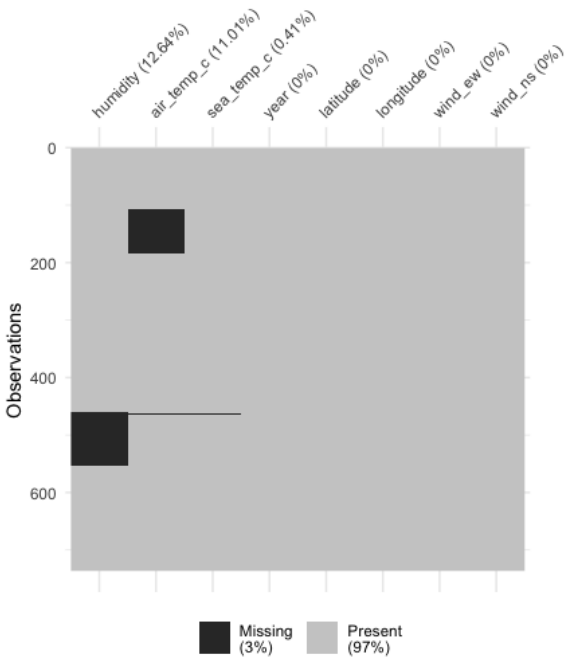
[Total: 10 marks]

— END OF QUESTION 3 —

QUESTION 4

This question is about handling missing values.

(a) The plot below shows a missing value summary of the ocean buoys data.



Which variables have missing values? (Circle one or more.)

[2 marks]

`year` `latitude` `longitude` `sea_temp_c` `air_temp_c` `humidity` `wind_ew` `wind_ns`

(b) If there are 736 observations, and 12.63587% of humidity values are missing, how many values are actually missing? __

[1 marks]

(c) Below is a missing case table, summarising the number of rows that have k missing values:

```
# A tibble: 4 x 3
  n_miss_in_case n_cases pct_cases
           <int>   <int>     <dbl>
              0     565    76.8
              1     167    22.7
              2       2     0.272
              3       2     0.272
```

(i) How many rows in the data have 3 missing values? __

[1 marks]

(ii) What percent of cases have no missing values? __

[2 marks]

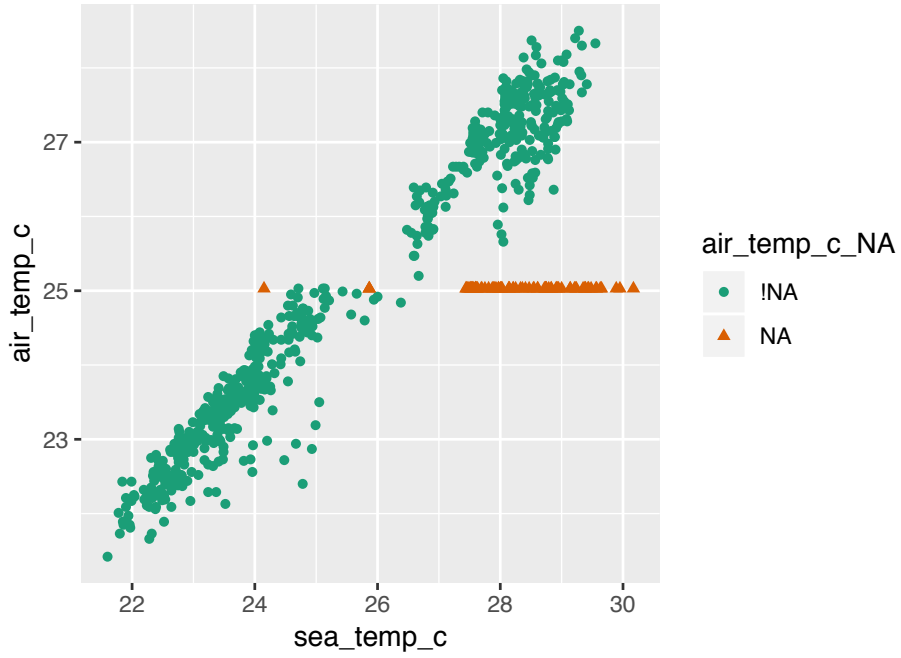
(iii) If a row has 3 missing values, which 3 variables would these be? __

[1 marks]

(d) Below is a plot of two of the variables, where missing values have been imputed (indicated by orange triangles). What imputation method do you think was used? (Circle one) __

[2 marks]

mean nearest.neighbours random regression multiple.imputation knn



[Total: 9 marks]

— END OF QUESTION 4 —

QUESTION 5

This question is about text analysis.

(a) The following code processes the text of the first edition of Darwin's Origin of the Species.

```
library(tidytext)
library(tm)
library(gutenbergr)
darwin1 <- gutenberg_download(1228)
darwin1$text <- removeNumbers(darwin1$text)
darwin1_words <- darwin1 %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(word, sort=TRUE) %>%
  mutate(len = str_length(word))
```

(i) What does the function `unnest_tokens(word, text)` do?

[2 marks]

(ii) What is an `anti_join` and how is it used here?

[2 marks]

(b) This is a summary of the number of times a word is used in the book. 75% of words are used _____ or less.

[2 marks]

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  1.000    1.000    2.000    8.855    7.000 1541.000
```

(c) In order to compare word frequency in multiple books, the statistic "term frequency, inverse document frequency (tf.idf)" is used. It is computed using these formula.

$tf_{word} = \frac{Number\ of\ times\ the\ word\ appears\ in\ a\ document}{Total\ number\ of\ words\ in\ the\ document}$

$idf_{word} = log_e \frac{number\ of\ documents}{number\ of\ documents\ word\ appears\ in}$ (natural log)

$td.df_{word} = tf \times idf$

The `janeaustenr` package has the text from 6 books. The word "elinor" occurs 623 times in Sense and Sensibility which has 12262 words, and it is not used in any other book.

(i) Compute tf_{elinor}. ___

[1 marks]

(ii) Compute idf_{elinor}. ___

[1 marks]

(iii) Compute $td.df_{elinor}$ ___

[1 marks]

(d) In your own words, explain why `tf.idf` is a useful statistic.

[2 marks]

(e) This is about sentiment analysis of Jane Austen's book Emma, using two different lexicons, "nrc" and "bing":

```
> emma_nrc %>% count(sentiment) %>% mutate(p=n/sum(n))
# A tibble: 2 x 3
  sentiment     n     p
  <chr>     <int> <dbl>
1 negative   4473 0.321
2 positive   9471 0.679
> emma_bing %>% count(sentiment) %>% mutate(p=n/sum(n))
# A tibble: 2 x 3
  sentiment     n     p
  <chr>     <int> <dbl>
1 negative   4809 0.402
2 positive   7157 0.598
```

(i) Explain in your own words what a "lexicon" is.

[2 marks]

(ii) TRUE or FALSE. The results differ between the two lexicons because both contain different word sets, and a different number of sentiment tagged words. ___

[1 marks]

(iii) TRUE or FALSE. Emma would be considered to be a quite dark (negative) book. ___

[1 marks]

[Total: 15 marks]

— END OF QUESTION 5 —

QUESTION 6

This question is about networks.

This is the association matrix for a part of the class. The rows contain people who responded, and the columns indicate people who were selected as "go-to" class members, with 1 indicating a connection. (Any rows or columns missing, eg a row for Mitch, contains all zeros.) Sketch the network, with directional arrows.

[5 marks]

	Arunabh	Christopher	Mitch	Victor	Zina
Cameron	0	0	1	0	0
Cassandra	0	0	1	0	0
Christopher	0	0	1	0	1
Naufal	0	1	0	0	0
Naomi	1	0	0	1	1
Ziyue	0	0	1	0	1

[Total: 5 marks]

— END OF QUESTION 6 —

QUESTION 7

This question is about advanced modeling.

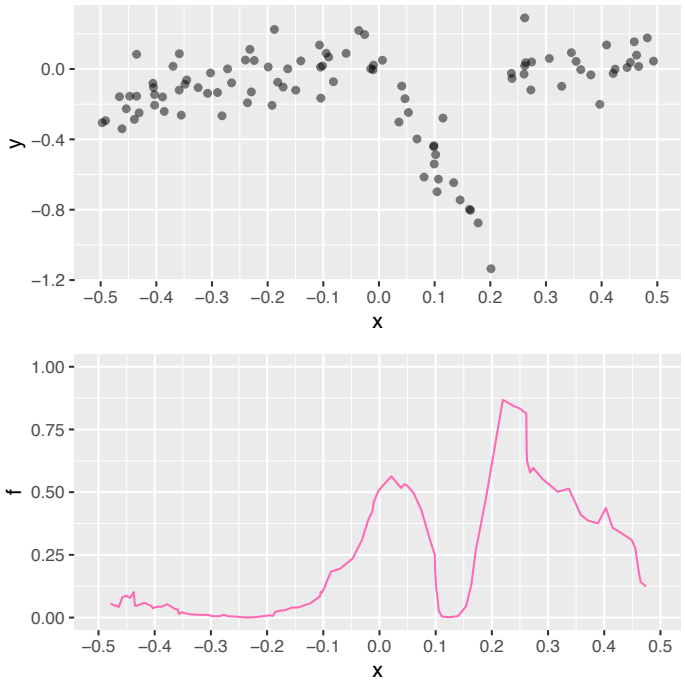
A regression tree is fitted to the data in the top plot below. This is a summary of the model fit.

```
n= 100

node), split, n, deviance, yval
      * denotes terminal node

 1) root 100 6.6425130 -0.135905800
   2) x< 0.2196274 75 5.5350070 -0.189703700
     4) x>=0.06063225 15 0.6699516 -0.635101200 *
     5) x< 0.06063225 60 1.1454510 -0.078354280
      10) x< -0.2769443 27 0.3287124 -0.152736000
        20) x< -0.437389 7 0.0325988 -0.252211800 *
        21) x>=-0.437389 20 0.2026017 -0.117919400 *
      11) x>=-0.2769443 33 0.5451360 -0.017496530
        22) x>=-0.01259873 7 0.1132896 -0.106969300 *
        23) x< -0.01259873 26 0.3607217  0.006592299 *
   3) x>=0.2196274 25 0.2392421  0.025487820 *
```

Partitions are decided by optimising the criteria,

$$SS_T - (SS_L + SS_R) \text{ where } SS_T = \sum_{i=1}^{\#before\ split} (y_i - \bar{y})^2,$$

and SS_L, SS_R are the equivalent sum of squares for the left and right partition. The bottom plot shows this function, evaluated for all possible splits in the data.



(a) How many total observations in the data set?

[2 marks]

(b) How many possible splits are there for this data set (assuming no ties)?

[2 marks]

(c) How many terminal nodes in the tree?

[2 marks]

(d) Sketch the first split on the data plot?

[2 marks]

(e) What is the deviance of the right side subset produced by the first split?

[2 marks]

(f) How many observations are in the left side of the first split?

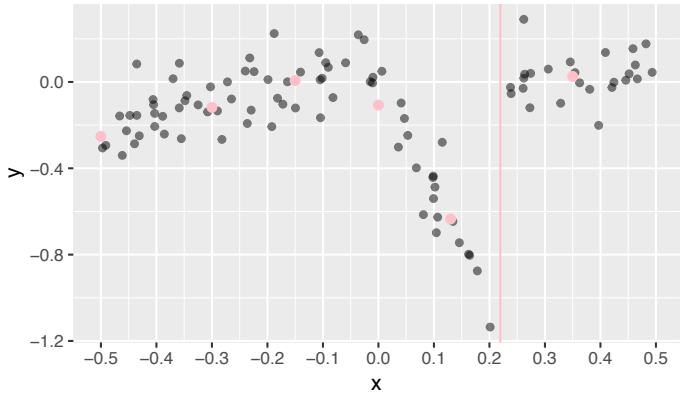
[2 marks]

(g) At what x value is the second split made?

[2 marks]

(h) Sketch the model fit on the data plot.

[2 marks]



(i) TRUE or FALSE. Stopping rules are used to prevent this model overfitting the data.

[2 marks]

[Total: 18 marks]

— END OF QUESTION 7 —

QUESTION 8

This question is about good data collection practices.

(a) Read this article, and answer the questions in relation to this material.

Really? The Claim: Excess Weight Raises the Risk of Acne
By ANAHAD O'CONNOR, January 23, 2012, New York Times

Teenagers and young adults may be able to prevent acne by stepping on the bathroom scale. Over the years, researchers have found that weight gain and moderate to severe acne is a problem that affects about one in five teenagers, and go hand in hand, particularly among young women. While it is not entirely clear why, excess hormones most likely play a role.

The most recent study highlighting a link was published this month in The Archives of Dermatology and included roughly 3,600 teenagers. The researchers looked closely at their weight and its relation to their skin, factoring in several variables that could also play a role, including age, puberty and diet. High-sugar junk foods like candy and soda are not only linked to weight gain, for example, but are also known to worsen acne. After adjusting for these and other factors that could affect acne risk, the researchers found that overweight or obese teenagers - particularly young women - were significantly more likely to develop acne than normal-weight adolescents.

(i) Is this an experiment or an observational study? Explain your answer.

[2 marks]

(ii) What is the response variable? _____

[1 marks]

(iii) Circle which ONE of the following would be considered to be the explanatory variable?

[1 marks]

Weight Sugar Candy Soda Age Puberty Diet

(iv) How were participants recruited into the study?

[1 marks]

(v) Is it possible to conclude from this study that weight causes acne to develop? Explain your thinking.

[1 marks]

(b) Read this article, and answer the questions in relation to this material.

Remedies: Tea Tree Oil for Acne
By ANAHAD O'CONNOR, January 27, 2012, New York Times

A small, randomized study published in 2007 involved 60 patients with mild to moderate cases of acne. The patients were randomly divided into two groups, one treated with a gel containing 5 percent tea tree oil and the other given placebo for 45 days. The scientists found the tea tree oil worked far better than placebo in reducing the number and severity of acne lesions.

(i) Is this an experiment or an observational study? Explain your answer.

[2 marks]

(ii) Identify the parts of the experiment:

[3 marks]

Experimental unit (subjects) _____

Response variable _____

Factor (explanatory variable) _____

(iii) Explain how replication was used in the study?

[1 marks]

(iv) How were subjects allocated to treatments?

[1 marks]

(v) Is it possible to conclude from this study that tea tree oil causes a reduction in the number and severity of acne lesions? Explain your thinking.

[1 marks]

[Total: 14 marks]

— END OF QUESTION 8 —

QUESTION 9

This question is about general knowledge.

From the following pool of statistical terms fill in the blanks in the descriptions below.

> correlation, positive, negative, strong, moderate, weak, least squares regression line, slope, intercept, extrapolation, simple random sample, stratified random sample, undercoverage, bias, long term relative frequency, association, R^2, r, $\bar{x}$, $\bar{y}$, s_x, s_y, random, double-blind, single blind, treatments, factors, levels

(i) The linear association between two variables would be considered to be _____ if $r = -0.2$.
[1 marks]

(ii) If the _____ for a linear fit is 0.90 we would say that there is the explanatory variable (x) explains 90% of the variation in the response variable (y).
[1 marks]

(iii) When neither the subjects in an experiment nor the person delivering the treatment are aware of the allocation of people to treatments, the experiment is said to be _____.
[1 marks]

(iv) When all possible samples of size n from a population have an equal chance of being chosen, we have taken a _____ of size n.
[1 marks]

(v) _____ are all possible combinations of factors and levels given to experimental units.
[1 marks]

[Total: 5 marks]

— END OF QUESTION 9 —

Formula sheet

Summary statistics

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}, \quad r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Descriptive words for univariate distributions:

- unimodal, bimodal, multimodal
- symmetric, right-skewed, left-skewed, uniform
- outliers

Descriptive words for bivariate distributions:

- shape: linear, non-linear, no relationship
- strength: weak, moderate, strong
- form: positive, negative

Tidy data

Verbs: `gather, spread, nest/unnest, separate/unite`

Wrangling data

Verbs: `filter, arrange, select, mutate, summarise, group_by/ungroup, count, tally`

Grammar of graphics

There are seven components of the grammar that define a data plot: DATA, AESTHETICS/MAPPINGS, GEOM, STAT, POSITION, COORDINATE, FACET.

Colour palettes: sequential, diverging, qualitative

Models

Simple linear: $Y = \beta_0 + \beta_1 X + \varepsilon$

Multiple linear: $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon$

- $\varepsilon \sim N(0, \sigma^2)$
- Fitted values: $\hat{Y} = b_0 + b_1 X$
- Residual: $e = Y - \hat{Y}$
- Estimates: $b_1 = r\frac{s_y}{s_x}$, $b_0 = \bar{Y} - b_1\bar{X}$
- $R^2 = 1 - \frac{\sum e^2}{\sum (Y_i - Y)^2}$
- $MeanSquaredError = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)}$
- $RMSE = \sqrt{MSE}$
- $MeanAbsoluteError = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{(n-2)}$