Instructions

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

QUESTION 1

(a) This question is about simple descriptive statistics. Here is a small data set

X_1	X_2	X_3
-1	15	-5
5	17	-5
-3	16	-5
-4	0	-5
21	4	-5
22	4	-5

(a) Compute the sample mean.

[1 marks]

(b) Is the correlation between X_1 and X_2 positive or negative?

[1 marks]

(c) Which one of these is the standard deviation of X_3?

[2 marks]

-5, 1, 0, 5, can't be computed

(d) If 22 was removed from X_1 would the **mean** increase or decrease?

[2 marks]

(e) If 5 was added to all the values of X_3 what would the standard deviation become?

[2 marks]

1, 0, 5, 25, can't be computed

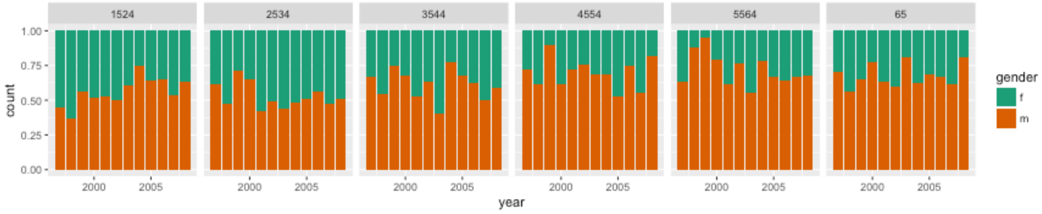
(b) For this table of numbers, what is the best number of decimal places to use for presentation purposes?

[2 marks]

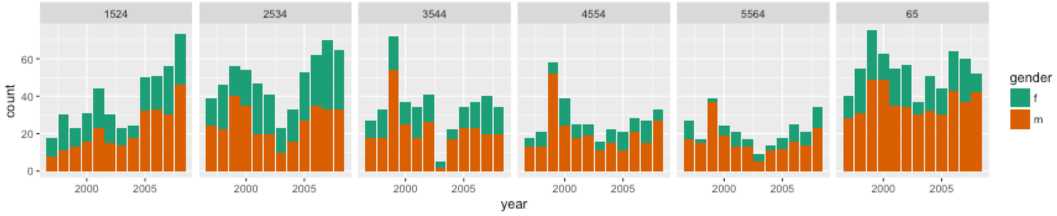
id	purchase_rate
1	0.9055024
2	0.9355204
3	0.2124656
4	0.3790097
5	0.5610250
6	0.4468279
7	0.0920617
8	0.9581790
9	0.6320397
10	0.8465992

(c) Both of the displays below show tuberculosis incidents in Australia from 1997-2008, for age groups 15-24, 25-34, 35-44, 45-54, 55-64, over 65 for males and females.

Display A:



Display B:



- Which of these two displays makes it easier to answer this question? Explain.

[3 marks]

“Has the number of tuberculosis incidences among 15-24 year olds been increasing in recent years?"

- What would be a question that could be answered better by the other display. Explain.

[2 marks]

[Total: 15 marks]

— **END OF QUESTION 1** —

QUESTION 2

Below is the first few rows of weather data collected from the station near Melbourne airport. The columns in order are station id, year, month, variable, values for each day. In the variable column there are just three possibilities, TMIN, TMAX, PRCP, indicating minimum temperature, maximum temperature. and precipitation. Remember that the temperature numbers are C^o times 10.

	V1	V2	V3	V4	V5	V9	V13	V17	V21	V25	...	V125
1	ASN00086282	1970	7	TMAX	141	124	113	123	148	149	...	115
2	ASN00086282	1970	7	TMIN	80	63	36	57	69	47	...	39
3	ASN00086282	1970	7	PRCP	3	30	0	0	36	3	...	3
4	ASN00086282	1970	8	TMAX	145	128	150	122	109	112	...	129
5	ASN00086282	1970	8	TMIN	50	61	75	67	41	51	...	39

(a) What are the observations?

[2 marks]

(b) What are the variables?

[4 marks]

(c) Is the data in tidy format? If no, sketch out what a tidy format of this data would look like.

[3 marks]

[Total: 9 marks]

— END OF QUESTION 2 —

QUESTION 3

(a) Match these data file types to their descriptions.

[5 marks]

csv	binary audio
sqlite	SPSS binary
sav	javascript mark up
wav	small database
xlsx	comma delimited text
feather	binary comma delimited
json	Excel binary

(b) If you need to calculate the standard deviation of a variable using a database, describe how you can modify the formula for standard deviation so that it is possible, using two computations on the variable. (Remember that some databases have a function to compute a mean or sum, and even a sum of squares, but do not have a command enabling standard deviation to be computed.)

[3 marks]

[Total: 8 marks]

— END OF QUESTION 3 —

QUESTION 4

This question relates to the click through hotel booking data used in one of the labs, and the explanation of the (important) variables is

$SRCH\_BEGIN\_USE\_DATE$	is the date for the hotel check in
$SRCH\_END\_USE\_DATE$	is the date for the hotel check out
$CLICK\_THRU\_DATETM$	is the date and time when the user is searching
$CLICK\_THRU\_TYP\_ID$	result of the search
IS_PROMO_FLAG	whether of not a promotion is active

(a) If we want to answer the question *“What day of the week is the most common check-in day?"* what would you need to do to the data?

[3 marks]

(b) If I want to answer this question *“What proportion of people searching, actually booked a hotel room?"* what would you need to do to the data? (The variable recording the searcher's final decision is $CLICK\_THRU\_TYP\_ID$, and the code indicating a booking is ‘3406‘).

[3 marks]

(c) There are a lot of missing values in the data, number of NAs, particularly this is true for the booking variable ($CLICK\_THRU\_TYP\_ID$). If an NA essentially means that the person searching, quit the site without doing a booking, how would you recode the missing value?

[2 marks]

(d) If I want to answer the question *“How far ahead of the check-in date do people typically search for a hotel room?"* what needs to done with the data.

[2 marks]

[Total: 10 marks]

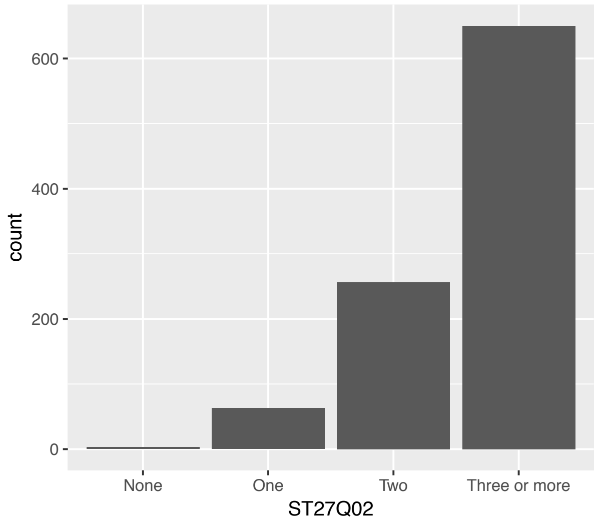
— END OF QUESTION 4 —

QUESTION 5

The grammar of graphics provides a mapping from variables in the tidy data to visual elements of a plot. For each of the following plots, specify the grammar that created it, all seven components. (The R code creating the plots is provided to help you.)

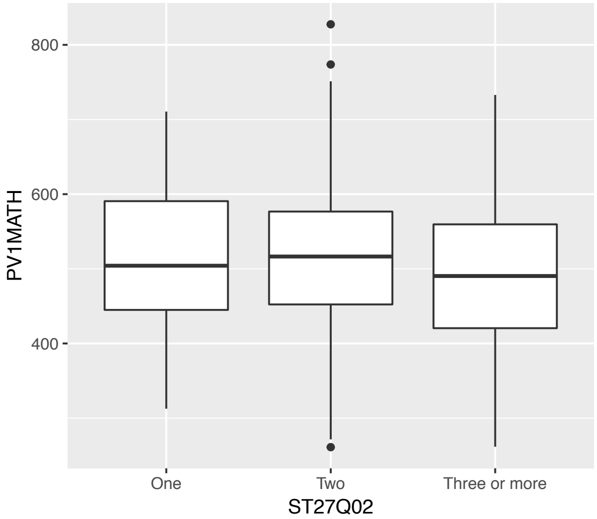
(a) `ggplot(PISA, aes(x=ST27Q02)) + geom_bar()`

[3 marks]



(b) `ggplot(PISA, aes(x=ST27Q02, y=PV1MATH)) + geom_boxplot()`

[3 marks]



[Total: 6 marks]

— END OF QUESTION 5 —

QUESTION 6

Any point on the globe can be represented by its longitude and latitude, and stated as a pair (X, Y), with X representing longitude and Y representing latitude. The latitude and longitude of several very closely related points on the globe are given below.

Table 1: Longitude and Latitude points

Point	X- Longitude	Y- Latitude
B	9	11
C	14	18
D	18	22
E	13	16
F	17	21

The goal of this problem will be to find a point A such that the overall distance between A and points B-F is minimized.

(a) Formulate the above optimization problem using the distance $d\{x, y\} = (x - y)^2$ as the model family. For full credit, you must precisely state

[3 marks]

(i) the objective function;

(ii) the full optimization problem.

(b) Using your optimization problem in part (a):

(i) state the first order conditions of the optimization problem.

[2 marks]

(ii) solve the optimization problem.

[3 marks]

(iii) Is the solution a minimum or a maximum?

[3 marks]

(iv) How would the answer in (a) change if instead of the distance $d\{x, y\} = (x - y)^2$, we used the distance $d\{x, y\} = |x - y|$ as the model family?

[2 marks]

[Total: 13 marks]

— END OF QUESTION 6 —

QUESTION 7

Rick is an unemployed scientist that makes robots for fun and then sells them for profit. Rick has asked his nephew to help him decide on the optimal number of robots to produce each month. Rick makes two types of robots: 'designer' robots and 'regular' robots. Rick's profit for a 'designer' robot is \$35 and for a 'regular' robot his profit is \$20. Rick's goal is to maximize his profit from producing robots. A 'designer' robot requires 10 hours of labor and 10 ounces of raw material. A 'regular' robot requires 20 hours of labor and 5 ounces of raw material. Both types of robots require 5 meters of internal wiring. Rick can't work more than 200 hours of work per month, else he may loose his pension. In addition, Rick can only buy 80 ounces of raw material and 50 meters of wiring per month.

(a) Formulate the optimization problem by answering the following questions

(i) What are the decision variables?

[2 marks]

(ii) What is the objective function?

[2 marks]

(iii) What are the constraints?

[2 marks]

(iv) Using (1)-(3), state the optimization problem.

[2 marks]

(b) Sketch the feasible region for the optimization problem.

[3 marks]

(c) Using the results from (a) and (b), and the idea of LP relaxation, give a potential solution to the optimization problem.

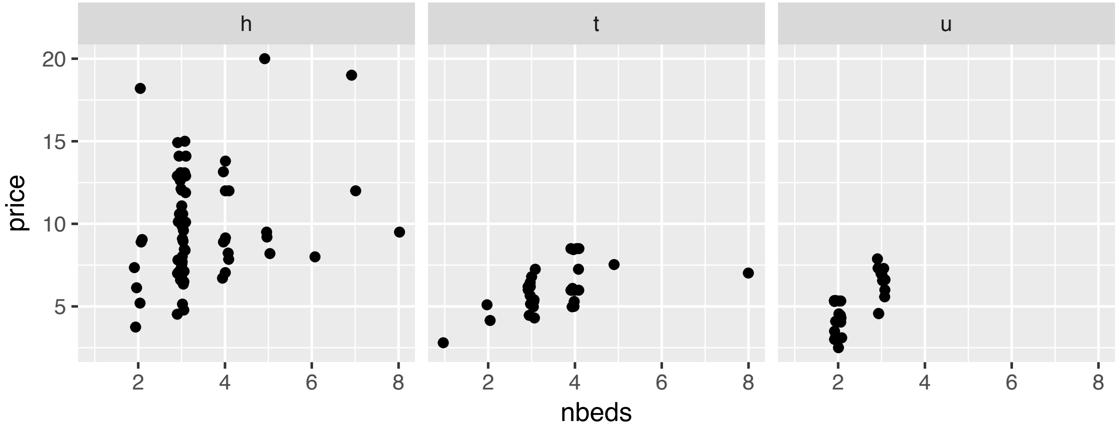
[3 marks]

[Total: 14 marks]

— END OF QUESTION 7 —

QUESTION 8

This question is about simple linear models. The plot below shows 2013-2016 auction prices ('000,000s) of properties in Clayton against number of bedrooms, for three property types, h=house, t=townhouse, u=unit. The points are spread out a little horizontally, since number of bedrooms is a discrete variable.



(a) What type of variable is property type?

[2 marks]

(b) Which of the following is false?

[2 marks]

(i) Price is positively linearly associated with nbeds, for townhouses.

(ii) Price has a nonlinear association with nbeds, for houses.

(iii) There are outliers.

(iv) We cannot assess the relationship between price and nbeds for 6 bedroom units

(c) A simple linear model is fitted for houses and townhouses separately. These are the results.

Houses:

```
lm(formula = price ~ nbeds, data = filter(clayton, property_type ==
    "h"))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0772     1.2268   5.769 2.14e-07 ***
nbeds         0.7631     0.3425   2.228   0.0292 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 3.228 on 68 degrees of freedom
Multiple R-squared:  0.06802,Adjusted R-squared:  0.05432
F-statistic: 4.963 on 1 and 68 DF,  p-value: 0.0292
```

Townhouses:

```
lm(formula = price ~ nbeds, data = filter(clayton, property_type ==
    "t"))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.8172     0.6874   5.553 4.89e-06 ***
nbeds         0.6289     0.1856   3.389  0.00198 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 1.173 on 30 degrees of freedom
Multiple R-squared:  0.2768,Adjusted R-squared:  0.2527
F-statistic: 11.48 on 1 and 30 DF,  p-value: 0.001981
```

(i) Write down the model describing the relationship between price and nbeds for houses.

[3 marks]

(ii) Which property type has the higher price when the number of bedrooms is 3? (houses or townhouses) Explain.

[2 marks]

(iii) Which model shows the strongest relationship between price and nbeds? Explain.

[2 marks]

(d) What criteria was (algebraicly) optimised to yield the fitted model in both of these cases?

[2 marks]

(e) If instead of least squares we used a least absolute deviation criteria to fit the models, describe how we could do this numerically.

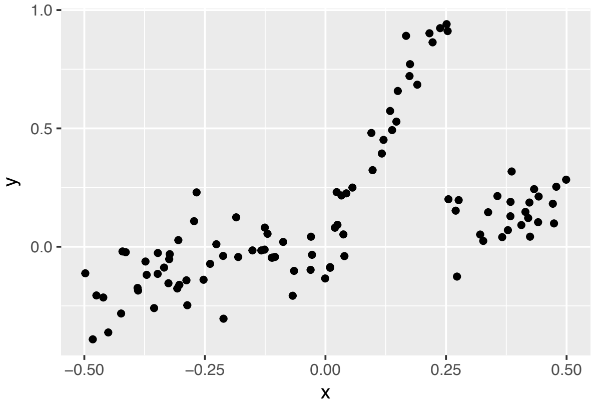
[2 marks]

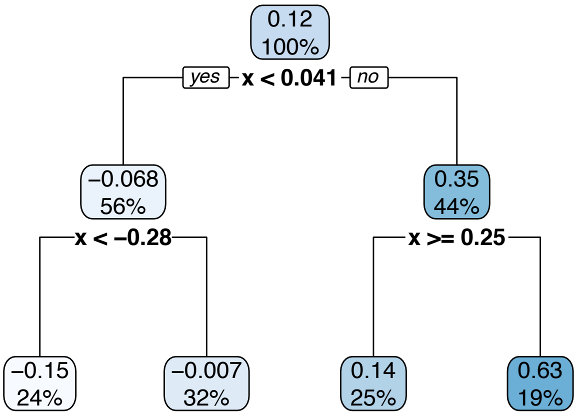
[Total: 15 marks]

— END OF QUESTION 8 —

QUESTION 9

Regression (decision) trees are fit to data, by recursively partitioning it into subsets. Below is a data set (x, y) and the resulting fitted regression tree.





(a) What value of x defines the first partition.

[2 marks]

(b) How many terminal nodes in the tree?

[2 marks]

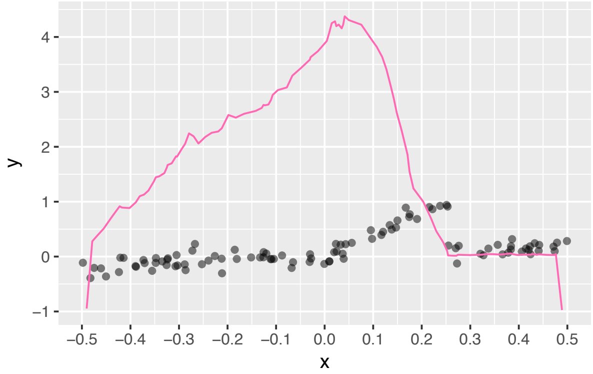
(c) Write down the decisions that would need to be followed to obtain fitted values for the model.

[4 marks]

(d) 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. This is a plot of the function, showing the partitions that were evaluated in order to decide on the best.



Which value of x corresponds to the optimal value of the function?

[2 marks]

[Total: 10 marks]

— END OF QUESTION 9 —

Formula sheet

Summary statistics

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

Types of variables: categorical, quantitative, logical, date.

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/ungroup

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

Optimization

One variable

For a single variable x and $f(x)$ a continuously differentiable function on $[a, b]$, recall that the conditions for a local optima are as follows:

$$\begin{aligned} f'(x) &= 0 \quad \text{First-order condition,} \\ f''(x) &< 0 \quad \text{Second-order condition: Max,} \\ f''(x) &> 0 \quad \text{Second-order condition: Min.} \end{aligned}$$

Two variables

For two variables x, y and $f(x, y)$ a continuously differentiable function on $[a, b] \times [a, b]$, recall that the conditions for a local optima are as follows:

$$\begin{pmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{First-order condition,}$$

$$\frac{\partial^2 f(x,y)}{\partial x^2} < 0,\ \frac{\partial^2 f(x,y)}{\partial y^2} < 0,\ \left\{ \left(\frac{\partial^2 f(x,y)}{\partial x^2} \right) \left(\frac{\partial^2 f(x,y)}{\partial y} \right) - \frac{\partial^2 f(x,y)}{\partial x \partial y} \right\} > 0 \quad \text{Second-order condition: Max,}$$

$$\frac{\partial^2 f(x,y)}{\partial x^2} > 0,\ \frac{\partial^2 f(x,y)}{\partial y^2} > 0,\ \left\{ \left(\frac{\partial^2 f(x,y)}{\partial x^2} \right) \left(\frac{\partial^2 f(x,y)}{\partial y} \right) - \frac{\partial^2 f(x,y)}{\partial x \partial y} \right\} > 0 \quad \text{Second-order condition: Min.}$$

Models

Simple linear:

$Y = \beta_0 + \beta_1 X + \varepsilon$

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

Decision trees:

ANOVA criterion: $SS_T - (SS_L + SS_R)$, $SS_T = \sum (y_i - \bar{y})^2$, and SS_L, SS_R are the equivalent values for the two subsets created by partitioning.