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

QUESTION 1

This question is about tidy data, principles and practice.

(a) Fill in the blanks. [6 marks]

A ________________ is a quantity, quality, or property that you can measure. For tabular (tidy) data, these would be all the column headers.

An ________________ is a set of measurements made under similar conditions (you usually make all of the measurements in an observation at the same time and on the same object).

The ________________ of a variable typically changes from observation to observation.

Word bank: value, name, observation, object, variable, variance, variability, variation, case, subject, data point, measure, type, model, regression.

(b) From the following summary,

(i) Which variables would be considered to be quantitative? Circle them. [2 marks]

(ii) How many observations? _________________________ [2 marks]

(iii) How many variables? _________________________ [2 marks]
(c) The following data is showing tuberculosis incidence for Australia, in messy format.

Observations: 16
Variables: 22

`iso3` <chr> "AUS", "AUS", "AUS", "AUS", "AUS", "AUS", "AUS", ...  
$m04$ <int> NA, NA, NA, NA, NA, 1, 0, NA, 0, 0, 2, NA, NA, NA, NA  
$m514$ <int> NA, NA, NA, NA, NA, 1, 3, NA, 3, 2, 2, 1, NA, NA, NA, NA  
$m014$ <int> 1, 1, 0, 0, 1, 3, 2, 3, 2, 2, 3, NA, NA, NA, NA  
$m1524$ <int> 23, 15, 14, 18, 32, 33, 30, 46, 30, 42, 38, 26, NA, NA,...  
$m2534$ <int> 20, 20, 10, 16, 27, 35, 33, 33, 37, 33, 44, 40, NA, NA,...  
$m3544$ <int> 18, 26, 2, 17, 23, 23, 20, 20, 16, 22, 26, 17, NA,...  
$m4554$ <int> 18, 19, 11, 15, 11, 21, 15, 27, 24, 25, 19, 25, NA, NA,...  
$m5564$ <int> 13, 13, 5, 11, 12, 16, 14, 23, 12, 9, 12, 16, NA, NA,...  
$m65$ <int> 35, 34, 30, 32, 30, 43, 37, 42, 34, 27, 37, 37, NA, NA,...  
$mu$ <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, 0, 0, 0, NA, NA,...  
$f04$ <int> NA, NA, NA, NA, NA, 1, 0, NA, 1, 1, 2, 0, NA, NA, NA  
$f514$ <int> NA, NA, NA, NA, NA, 1, 4, NA, 3, 3, 1, 1, NA, NA, NA  
$f014$ <int> 1, 0, 0, 0, 2, 2, 4, 3, 4, 4, 3, 1, NA, NA, NA  
$f1524$ <int> 21, 15, 9, 6, 18, 18, 26, 27, 31, 36, 26, 27, NA, NA,...  
$f2534$ <int> 27, 21, 13, 17, 26, 27, 37, 32, 27, 43, 40, 48, NA, NA,...  
$f3544$ <int> 16, 15, 3, 5, 11, 14, 20, 14, 14, 12, 23, 15, NA, NA,...  
$f4554$ <int> 7, 6, 5, 7, 10, 7, 12, 6, 12, 2, 7, 11, NA, NA, NA  
$f5564$ <int> 8, 4, 3, 6, 9, 7, 11, 11, 5, 7, 9, NA, NA, NA, NA  
$f65$ <int> 20, 23, 7, 19, 14, 21, 23, 10, 12, 12, 17, 15, NA, NA,...  
$fu$ <int> NA, NA, NA, NA, NA, NA, NA, NA, 0, 0, 0, NA, NA, NA...

(i) How many observations? ________________ [2 marks]

(ii) How many variables? ________________ [2 marks]

(iii) Which variables would be considered to be categorical? ________________ [2 marks]

(iv) What value indicates missing information? ________________ [2 marks]

(v) Map out the steps that you would need to take to get it into tidy format. [4 marks]
(d) TRUE or FALSE: The variables used to connect multiple tables are called keys. A key is a variable (or set of variables) that uniquely identifies a measured value. [2 marks]

(e) Identify the possible key(s) in the following data: [2 marks]

```r
> library(babynames)
> babynames
# A tibble: 1,858,689 x 5
   year sex name      n  prop
   <dbl> <chr> <chr> <int> <dbl>
1  1880 F  Mary  7065 0.0724
2  1880 F  Anna  2604 0.0267
3  1880 F  Emma  2003 0.0205
4  1880 F  Elizabeth  1939 0.0199
5  1880 F  Minnie  1746 0.0179
6  1880 F  Margaret 1578 0.0162
7  1880 F   Ida  1472 0.0151
8  1880 F     Alice  1414 0.0145
9  1880 F Bertha  1320 0.0135
10 1880 F     Sarah  1288 0.0132
# ... with 1,858,679 more rows
```

[Total: 28 marks]
QUESTION 2

This question is about wrangling data, verbs, definitions and usage.

(a) Match the verb to its usage by drawing lines to connect the verb and usage: [6 marks]

<table>
<thead>
<tr>
<th>verb</th>
<th>usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>filter</td>
<td>create new, or change, a variable</td>
</tr>
<tr>
<td>select</td>
<td>order a table by values in one column</td>
</tr>
<tr>
<td>mutate</td>
<td>operate on subsets specified by a categorical variable</td>
</tr>
<tr>
<td>summarise</td>
<td>subset variables</td>
</tr>
<tr>
<td>arrange</td>
<td>subset cases</td>
</tr>
<tr>
<td>group_by</td>
<td>compute a single number from a collection</td>
</tr>
</tbody>
</table>

(b) For each of these questions, write down the verbs that you would need to use to do the calculations to answer the question about the nycflights13 data. [10 marks]

> glimpse(flights)
Observations: 336,776
Variables: 15
$ month <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ day <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ dep_time <int> 517, 533, 542, 544, 554, 554, 555, 555, 557, 557, 558, 55...
$ dep_delay <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -2...
$ arr_time <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 8...
$ arr_delay <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, ...
$ carrier <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6"...
$ flight <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301...
$ tailnum <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N...
$ origin <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA"...
$ dest <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IA...
$ air_time <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149...
$ distance <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 73...
$ hour <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6...

(i) Find all flights that were operated by United, American, or Delta.

(ii) What hour of day should you fly if you want to avoid delays as much as possible?

(iii) Find all destinations that are flown to by at least two carriers.

(iv) Find the busiest airports.

(v) Find the plane (identified by their tail number) that has the highest average delays.

Full verb list: filter, select, mutate, summarise, arrange, group_by, near, desc, starts_with, ends_with, contains, matches, rename, top_n, first, min_rank, lag, cumsum, count, tally

[Total: 16 marks]
QUESTION 3

This question is about making good plots of data.

(a) How would you improve the following plot? [2 marks]

(b) Each of the following two plots was produced to answer the question how does taking ETC1010 as core or elective vary by year in school?

(i) Which variable is the explanatory variable? [2 marks]

(ii) Which display makes it easier to answer the question? Explain your answer. [4 marks]

(iii) Give one suggestion to improve the plot. [2 marks]
(c) Below are three plots (A, B, C) made to examine how enjoyment of the unit, and whether you are struggling, depends on having prior coding experience or not. Write a paragraph describing the pros and cons of each display, in addressing the purpose.

(d) John Tukey said "The greatest value of a picture is when it forces us to notice what we never saw to see." What is the missing word? [2 marks]

[ ] wanted  [ ] tried  [ ] expected  [ ] wanted

(e) Which of the following are true about the grammar of graphics? [3 marks]

[ ] the variables are directly mapped to an element in the plot
[ ] it is possible to see how one display is similar or different from another, rather than thinking of plots like animals in a zoo, specific beasts (pie chart, barchart, scatterplot, ?)
[ ] themes are one of the seven components

[Total: 21 marks]
QUESTION 4

This question is about multiple regression modelling.

(a) For the model \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \), what is the correct model formula in R? [3 marks]

\( y \sim x_1 + x_2 \)  
\( y \sim x_1 \times x_2 \)  
\( y \sim x_1 + x_2 - 1 \)  
\( y \sim b_1 \times x_1 + b_2 \times x_2 \)

(b) If you wanted to force the intercept to go through 0 what the formula be? (write it out here) [2 marks]

(c) It’s also useful to see what the model doesn’t capture, the so-called 5cm which are left after subtracting the predictions from the data. [2 marks]

\( \square \) fitted values  \( \square \) residuals  \( \square \) predictions  \( \square \) coefficients

(d) The following plot shows the residuals from two different model fits (mod1, mod2). Which model (or both, or neither) best captures the original structure in the data? Explain your answer. [3 marks]

![Residual plots](image)
The following plot shows the fitted values from two different model fits.

(i) Which model (or both, or neither) contains an interaction term? Explain your answer.

(ii) Is the interaction with a categorical or quantitative variable?

(f) TRUE or FALSE: Adding an interaction between two quantitative variables to the model enables a nonlinear relationship to be fitted.

(g) From the following model summary, for science score,

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.633345e+02</td>
<td>43.50881</td>
</tr>
<tr>
<td>log_science_time</td>
<td>-3.995844e-03</td>
<td>18.60566</td>
</tr>
<tr>
<td>science_fun_cDisagree</td>
<td>-1.518829e+02</td>
<td>55.46131</td>
</tr>
<tr>
<td>science_fun_cAgree</td>
<td>-1.667660e+02</td>
<td>48.13726</td>
</tr>
<tr>
<td>science_fun_cStrongly agree</td>
<td>-1.224884e+02</td>
<td>53.60320</td>
</tr>
<tr>
<td>log_science_time:science_fun_cDisagree</td>
<td>7.864348e+01</td>
<td>23.75214</td>
</tr>
<tr>
<td>log_science_time:science_fun_cAgree</td>
<td>9.840520e+01</td>
<td>20.56815</td>
</tr>
<tr>
<td>log_science_time:science_fun_cStrongly agree</td>
<td>9.693076e+01</td>
<td>22.78145</td>
</tr>
</tbody>
</table>

(i) Write out the fitted model equation(s).
(ii) For a new observation where science_time=1000, and science_fun is Agree, predict the average science score.  

[3 marks]

(iii) Would be the predicted average score for a student who answered Strongly agree for science_fun be higher? Why? 

[2 marks]

(iv) What was the purpose of using the log transformed values for science_time, do you think? 

[3 marks]

(h) From the model summary, this is the equation describing the fitted model. TRUE or FALSE 

\[ \log(Price) = 5.873 + 0.080\text{Rooms} + 0.051\text{Bathroom} - 0.016\text{Distance} \] 

[2 marks]

(i) When imputing missing values in preparation for fitting a multiple linear model, we will use a separate regression model for the variable with missing values. The variable containing missing values, will be regressed on other explanatory variables, using the complete cases. Explain why it is not a good idea to use the response variable to do the imputation. 

[3 marks]

(j) Of the two models (mod3 or mod4), based on the fit statistics below, which is the best? Explain your answer. 

[3 marks]

```
> glance(mod3)
  r.squared adj.r.squared sigma statistic p.value df
1 0.3545374 0.354395 0.1802053 2489.321 0 7
  logLik AIC BIC deviance df.residual
1 8019.584 -16023.17 -15957.48 883.032 27192
> glance(mod4)
  r.squared adj.r.squared sigma statistic p.value df
1 0.5315746 0.5314196 0.1535252 3428.14 0 10
  logLik AIC BIC deviance df.residual
1 12378.79 -24735.57 -24645.26 640.8208 27188
```

[Total: 35 marks]
Summary statistics

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

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

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: \( b_1 = r_{xy} \frac{s_y}{s_x}, b_0 = \bar{Y} - b_1 \bar{X} \)
- \( R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \)
- \( MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - 2} \)
- \( RMSE = \sqrt{MSE} \)
- \( MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n - 2} \)