QUESTION 1

This question is about definitions. It is a word search. Among the letters, find the appropriate term/word for each of the descriptions, and circle it. (All words are left to right, or top to bottom. No diagonals or back to front.)

<table>
<thead>
<tr>
<th>wrangle</th>
<th>a</th>
<th>m</th>
<th>u</th>
<th>a</th>
<th>g</th>
<th>e</th>
<th>t</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>y</td>
<td>w</td>
<td>r</td>
<td>a</td>
<td>n</td>
<td>g</td>
<td>l</td>
<td>e</td>
</tr>
<tr>
<td>a</td>
<td>m</td>
<td>u</td>
<td>a</td>
<td>g</td>
<td>e</td>
<td>t</td>
<td>s</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>d</td>
<td>g</td>
<td>m</td>
<td>u</td>
<td>t</td>
<td>a</td>
<td>t</td>
<td>e</td>
</tr>
<tr>
<td>e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>p</td>
<td>i</td>
<td>p</td>
<td>e</td>
<td>m</td>
<td>e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>r</td>
<td>a</td>
<td>m</td>
<td>m</td>
<td>a</td>
<td>r</td>
<td>n</td>
<td>a</td>
</tr>
<tr>
<td>m</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>i</td>
<td>n</td>
<td>g</td>
<td>d</td>
<td></td>
</tr>
</tbody>
</table>

wrangle a term that means data transformation
pipe the word for %>%
gets the word for <-
gather the verb that helps you make tidy long formatted data
spread the verb that helps you make tidy wide formatted data
nanian package that helps work with missing values
plotting data is best done with a grammar
when you want to plot subsets of the data you use facet_wrap or facet_grid
the part of the grammar of graphics that describes the mapping of variables to plot elements
geom the part of the grammar of graphics that defines the plot elements
stat the part of the grammar of graphics that makes transformations like "identity", count the number of elements in the bins or computes the five number summary to make a boxplot. Not often directly used, but we have used it to make barcharts of categorical variables that have already been tabulated.
rug the name of the plot where ticks are put in the margins indicating data values, or another name for a carpet
ymd the lubridate function to use when the date comes as a character in this format "2019-04-11"
when you want to create a new variable, or change and existing variable you use the verb mutate

[Total: 5 marks]
QUESTION 2

This question is about software, workflow, and reproducibility.

(a) Which of these is analogous to seeing this in the RStudio window mean? (Pick one) [1 marks]

(i) All your clean clothes are hanging neatly in your cupboard, and sorted by type.
(ii) All your clean clothes are piled disorderly on the floor.
(iii) All of your freshly washed clothes are drying on the clothes rack.
(iv) Your next project is to neatly fold your washing.
(v) You have no new household projects to do.

(b) (1) TRUE or FALSE. RStudio is the same as R. FALSE [1 marks]

(c) (1) In your own words, explain why writing scripts for data analysis is useful. [2 marks]

Scripting is useful for showing all the steps conducted in the data analysis. It allows one to see how the data was processed from the raw form to final results. You can pass the script to someone else for them to reproduce your analysis.

[Total: 4 marks]

— END OF QUESTION 2 —
QUESTION 3

This question is about tidy data.

<table>
<thead>
<tr>
<th>species</th>
<th>location</th>
<th>class</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>emu</td>
<td>wilsons prom</td>
<td>aves</td>
<td>40</td>
</tr>
<tr>
<td>emu</td>
<td>grampians</td>
<td>aves</td>
<td>20</td>
</tr>
<tr>
<td>wombat</td>
<td>wilsons prom</td>
<td>mammalia</td>
<td>30</td>
</tr>
<tr>
<td>wombat</td>
<td>grampians</td>
<td>mammalia</td>
<td>10</td>
</tr>
<tr>
<td>wallaby</td>
<td>wilsons prom</td>
<td>mammalia</td>
<td>50</td>
</tr>
<tr>
<td>wallaby</td>
<td>grampians</td>
<td>mammalia</td>
<td>50</td>
</tr>
</tbody>
</table>

(a) TRUE or FALSE. The data is in tidy form.  **TRUE**  [1 marks]

(b) How many variables?  **3**  [1 marks]

(c) How many observations?  **200**  [2 marks]

(d) What is the proportion of wombats?  **40/200 = 0.2**  [1 marks]

(e) What is the proportion wombats are mammalia?  **1.0**  [1 marks]

(f) What is the proportion of the wildlife were spotted at Wilsons Prom?  **120/200 = 0.6**  [1 marks]

[Total: 7 marks]

— END OF QUESTION 3 —
QUESTION 4

This question is about wrangling data

(a) This summary of the french fries data was created using which wrangling verb, do you think? [1 marks]

filter select mutate summarise arrange **count**

# A tibble: 5 x 2
  type    n
  <fct> <int>
1 buttery 696
2 grassy  696
3 painty  696
4 potato  696
5 rancid  696

(b) Explain in your own words what `starts_with` does here? [2 marks]

```r
tb <- read_csv("data/TB_notifications_2018-03-18.csv") %>%
  select(country, year, starts_with("new_sp"))
```

This selects only columns that have names starting with `starts_with`.

[Total: 3 marks]

— END OF QUESTION 4 —
QUESTION 5

This question is about data formats

(a) What type of data format is being read by the following code? [1 marks]

```r
library(haven)
pisa_2015 <- read_sav(file.choose())
```

(b) What type of data is this? [1 marks]

```r
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  }
}
```

[Total: 2 marks]

— END OF QUESTION 5 —
QUESTION 6

This question is about plotting data

A.

B.

(a) In plot A, what variable is mapped to colour?  

\[ \text{age, gender, count} \]  

[1 marks]

(b) Which of these statements is easier to read from plot B?  

- TB incidence in 35-44 year olds is higher in males than females?  
- TB incidence in 35-44 year old females is higher than 45-54 year old females?  

[2 marks]

(c) What graphical principle is being used in deciding which of the two displays is best?  

- layering  
- variable mapping  
- faceting  
- pre-attentive proximity  
- appropriate colour  
- palette  
- coordinate system  
- hierarchy of mappings  

[Total: 5 marks]

— END OF QUESTION 6 —
QUESTION 7

This question is about missing data

(a) Explain what the NA for the variable names means in the summary of data below made using this code:

```r
library(naniar)
aq_shadow <- bind_shadow(airquality)
glimpse(aq_shadow)
```

These columns correspond to the shadow matrix. They are binary variables indicating whether a value is missing or not.

(b) A numerical summary of missings is as follows and is made with the code:

```r
miss_case_table(airquality)
```

How many observations have no missing values? 111

[Total: 4 marks]

— END OF QUESTION 7 —