ETC1010 S1 Midterm Exam

Your name: SOLUTION

You should attempt all of the questions.

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.)

[5 marks]

f	y	w	r	a	n	g	l	e	s	t	a	t	
a	m		u			a			g	e	t	s	
c	d		g	m	u	t	a	t	e			p	
e						h			o			r	
t			p	i	p	e			m			e	
g	r	a	m	m	a	r		n	a	n	i	a	r
			m	a	p	p	i	n	g			d	

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

naniar 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]

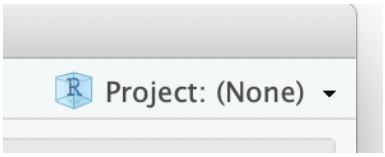
— END OF QUESTION 1 —

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.

species	location	class	count
emu	wilsons prom	aves	40
emu	grampians	aves	20
wombat	wilsons prom	mammalia	30
wombat	grampians	mammalia	10
wallaby	wilsons prom	mammalia	50
wallaby	grampians	mammalia	50

(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]

```
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]

stata sas **spss** eviews csv excel

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

(b) What type of data is this?

[1 marks]

csv html **json** wav xls xlsx

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

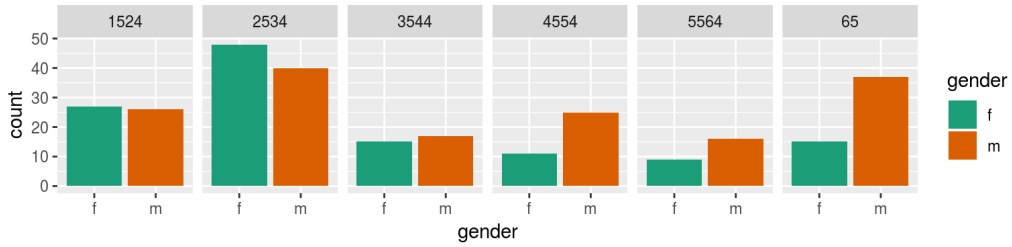
[Total: 2 marks]

— END OF QUESTION 5 —

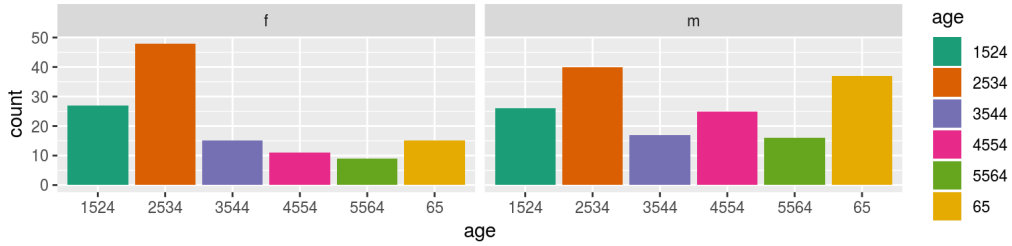
QUESTION 6

This question is about plotting data

A.



B.



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

[1 marks]

age **gender** count

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

[2 marks]

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?

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

[2 marks]

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:

[2 marks]

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

```
## Observations: 153
## Variables: 12
## $ Ozone      <int> 41, 36, 12, 18, NA, 28, 23, 19, 8, NA, 7, 16, 11, 14,...
## $ Solar.R    <int> 190, 118, 149, 313, NA, NA, 299, 99, 19, 194, NA, 256...
## $ Wind       <dbl> 7.4, 8.0, 12.6, 11.5, 14.3, 14.9, 8.6, 13.8, 20.1, 8....
## $ Temp       <int> 67, 72, 74, 62, 56, 66, 65, 59, 61, 69, 74, 69, 66, 6...
## $ Month      <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,...
## $ Day        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ Ozone_NA   <fct> !NA, !NA, !NA, !NA, NA, !NA, !NA, !NA, !NA, NA, !NA, ...
## $ Solar.R_NA <fct> !NA, !NA, !NA, !NA, NA, NA, !NA, !NA, !NA, !NA, NA, !...
## $ Wind_NA    <fct> !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA...
## $ Temp_NA    <fct> !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA...
## $ Month_NA   <fct> !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA...
## $ Day_NA     <fct> !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA, !NA...
```

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:

[2 marks]

```
miss_case_table(airquality)
```

```
## # A tibble: 3 x 3
##   n_miss_in_case n_cases pct_cases
##            <int>   <int>     <dbl>
## 1              0     111     72.5
## 2              1      40     26.1
## 3              2       2      1.31
```

How many observations have no missing values? 111

[Total: 4 marks]

— END OF QUESTION 7 —