MATHS 7107 Data Taming Assignment One Questions

Chang Dong(a1897402)

13/Feb/2023

In this assignment, we will review some mathematical concepts that you will need for later in the course and also assess what you have learnt in week one and two.

Some rules about your submissions:

- Include your working when providing solutions;
- Provide all numerical answers to **3 decimal places**;
- Make sure you include all of your R output and plots in your answers (where required);
- Make sure any tables or plots included have captions;
- Do not write directly on the question sheet;
- Each question will stipulate if the solutions can handwritten or typed;
- Assignments must be submitted as **pdf only** on MyUni;
- You can submit more than once if you find errors and your latest submission will be marked;
- Make sure you only upload one document for your final submission. If you submit multiple pages (i.e.one per question) you will be deducted 10% per page submitted;
- Penalties for late submission - within 24 hours 40% of final mark. After 24 hours, assignment is not marked and you get zero; and
- Finally, make sure you check your submitted assignment is the correct one, as we cannot accept other submissions after the due date.

Question 1 - Revision of Mathematical concepts

This question may be handwritten. Remember to scan and attach to rest of your assignment before submitting.

(a) Calculate y for each of the following:

(i) $y = 10x + 7$, where $x = 4.3$ [2 Marks]

Answer: $y = 4.3 \times 10 + 7 = 50.000$

(ii) $y = ax^2 + bx$, where $a = \frac{1}{3}, b = -3, x = 5$ [2 Marks]

Answer: $y = \frac{1}{3} \times 5 \times 5 - 3 \times 5 = -6.667$

(iii) $y = 3 + x(4s - 5z)^2$, where $x = 5, s = -4, z = -2$ [2 Marks]

Answer: $y = 3 + 5 \times (4 \times (-4) - 5 \times (-2))^2 = 183.000$

(iv) $y = \frac{x^2-a^2}{b} + \frac{b^2-x^2}{a}$, where $x = 3, a = 4, b = 5$ [2 Marks]

Answer: $y = \frac{3^2-4^2}{5} + \frac{5^2-3^2}{4} = 2.600$

(b) You are booking a birthday party and need to hire a venue. There is a flat booking fee of \$ 365 for the room hire and it costs \$ 42 for each platter of food you purchase.

(i) If the total cost is represented by y and the number of platters of food purchased is represented by x, write an equation of the form $y = ax + b$ that can be used to calculate cost. [1 Marks]

Answer: $y = 42x + 365$

(ii) You decide to order 6 platters of food. What is the total cost of the party? [2 Marks]

Answer: $y = 42 \times 6 + 365 = 617.000$

(iii) Your friend Jaimee booked her birthday party at the same venue (on a different night) and ordered 6 platters of food. As she booked a weeknight, she received a 30 % discount on the booking fee. What is her total cost? [2 Marks]

Answer: $y = 42 \times 6 + 365 \times 0.7 = 507.500$

(iv) Jaimee decided to hire a clown for her party at the last minute. She was charged an additional \$ 18 per hour for hiring the clown. The clown was at the party for 2.5 hours. What is Jaimee's total cost now?

Answer: $y = 42 \times 6 + 365 \times 0.7 + 18 \times 2.5 = 552.500$ [Total: 15 Marks][2 Marks]

Question Two: Variable types

This question may be handwritten. Remember to scan and attach to rest of your assignment before submitting.

For each of the following state whether the variable is categorical nominal, categorical ordinal, quantitative discrete, or quantitative continuous.

(a) A survey asks people to report which age bracket they belong in from the following choices: 0-19 years old, 20-39 years old, 40-59 years old, 60+ years old. [2 Marks]

Answer: categorical ordinal,becasue it has 4 age categories, and different age periods can be compared, so it is categorical ordinal.

(b) Smoking status coded as “Yes”=1 and “No”=0 [2 Marks]

Answer: categorical nominal,becasue it has 2 categories, but yes and no can not be compared, which is bigger or not, so it is categorical nominal.

(c) The tax file numbers of a set of employees. [2 Marks]

Answer: categorical nominal, because the tax file numbers is just the identification of each person, they can not be compared.

(d) Number of cars parked in a parking lot [2 Marks]

Answer: quantitative discrete, because the parking numbers is numerical data, but must be integer, so it is quantitative discrete.

(e) Average monthly temperature in Adelaide city. [2 Marks]

Answer: quantitative continuous, because temperature is numerical data, and can be any float [Total: 10 Marks] numbers if possible in Adelaide such as 37, 37.1, 37,99,37.999… , so it is quantitative continuous.

Question Three: Subjects and variables

This question may be handwritten. Remember to scan and attach to rest of your assignment before submitting.

For In each of the following scenarios, identify the subject and the variables:

(a) Suppose a study first asked 100 students whether they meditate regularly and then measured their blood pressures. The idea would be to see if those who meditate have lower blood pressure than who do not do so.

Answer: Subject is the students,and variables are meditate and blood pressure [2 Marks]

(b) Suppose a researcher would like to determine whether one grade of gasoline produces better gas mileage than other grade. Twenty cars are randomly divided into two groups, with 10 cars receiving one grade and 10 receiving the other. After many trips, average mileage is computed for each car. [2 Marks]

Answer: Subject is the cars, and variables are gasoline grade and gas mileage.

(c) In an investigation into theft at petrol stations, CCTV footage from 18 Top Gas petrol stations are viewed and details recorded over a 12 week period. The records contain the time the thief stole the petrol, the type of vehicle the thief drove, the location of the petrol station, and the type of fuel stolen. [2 Marks]

Answer: Subject is the theft cases from the 18 Top Gas petrol stations, and variable are the time the thief stole the petrol, the type of vehicle the thief drove, the location of the petrol station, and the type of fuel stolen.

(d) One study investigated the effects of stress on teenagers. For the sample three private schools and five public schools in Adelaide are chosen. There were a total of 500 students. For each participating student their gender and age were recorded together with stress level given in a likert scale from 0-10 . [2 Marks]

Answer: Subject is the teenagers, and variable are gender, age and stress level. [Total: 8 Marks]

Question Four: Using data in R

This question must be typed in Word and saved as pdf . Remember to scan and attach to rest of your assignment before submitting.

You must include your code and output for ALL of your below solutions.

The data in movies.csv gives details about a set of movies . Following are the details of each column of the dataset.

- name - name of the movie.
- genre - genre of the movie.
- runtime - running time of the movie.
- score -IMDB score

(a) Import the dataset into R. [1 Marks]

Answer:

```
pacman::p_load("tidyverse", "inspectdf")
movies = read.csv("movies.csv")
head(movies)
```

```
##                      name     genre runtime score
## 1      Assassin's Creed    Action     115   5.9
## 2            La La Land    Comedy     128   8.2
## 3         Suicide Squad    Action     123   6.2
## 4                  Sing Animation     108   7.1
## 5                 Moana Animation     107   7.6
## 6 The Magnificent Seven    Action     132   6.9
```

(b) What are the subjects in this dataset? How many subjects are there? [2 Marks]

Answer:

```
colnames(movies)
```

```
## [1] "name"    "genre"   "runtime" "score"
```

```
length(movies$name)
```

```
## [1] 142
```

Subject is the **movies**, and there are **142** subjects.

(c) What are the variables in the dataset? Write down the types of these variables? [4 Marks]

Answer: Variables and corresponding types are **“genre, categorical nominal”**, **“runtime, quantitative discrete”**, **“score, quantitative continuous”**.

(d) Show the fourth column from this dataset. [1 Marks]

Answer:

```
answer_d = movies[4]
head(answer_d)
```

```
##   score
## 1   5.9
## 2   8.2
## 3   6.2
## 4   7.1
## 5   7.6
## 6   6.9
```

To avoid this answer occupy too many pages, I only show the first 6 rows of the Fourth column. If you want to see the whole 4th column, we can use this code: **`movies[4]`**

(e) Provide the information for the tenth movie. [1 Marks]

Answer:

```
movies[10,]
```

```
##          name  genre runtime score
## 10 Rogue One Action     133   7.9
```

The 10th movie is **Rogue One**, which belongs to **Action** genre, and the duration is **133** minutes, with **7.9** score.

(f) What is the the genre of the thirtieth movie? [1 Marks]

Answer:

```
movies[30,]$genre
```

```
## [1] "Biography"
```

The Genre of the 30th movie is **Biography**.

(g) Convert the relevant columns in to categorical variables [2 Marks]

Answer:

```
movies$genre <- as.factor(movies$genre)
levels(movies$genre)
```

```
## [1] "Action"    "Animation" "Biography" "Comedy"
```

I converted movie genre into categorical variables, with 4 levels, which are **Action, Animation, Biography, Comedy**.

(h) Count the number of movies in each genre [1 Marks]

Answer:

```
table(movies$genre)
```

```
## 
##    Action Animation Biography    Comedy 
##        61        18        23        40
```

(i) Produce the five-number summary for the relevant variable/variables? [1 Marks]

Answer: Only **`movies$score`** and **`movies$runtime`** are numerical.

```
inspect_num(movies)
```

```
## # A tibble: 2 × 10
##   col_name   min    q1 median   mean    q3   max     sd pcnt_na hist        
##   <chr>    <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl>   <dbl> <named list>
## 1 runtime   76      98  108    110.    118   162   15.7       0 <tibble>    
## 2 score      4.6     6    6.5    6.59    7.1   8.7  0.798    1.41 <tibble>    
```

(j) How much data is missing in the movies data set? [1 Marks]

Answer: Here are the missing value in each column [Total: 15 Marks]

```
colSums(is.na(movies))
```

```
##    name   genre runtime   score 
##       0       0       0       2
```

```
sum(is.na(movies))
```

```
## [1] 2
```

The total of missing values in the dataset are **2**.