Data Science - Practice 5 (Data Preparation I)

Make sure you not only just “write down” the R code but also “explain the answer with your own language”. All answers without explanation will not be accepted.

Problem

Import ‘UserConsump_ed.RData’ and ‘daily.attend.RData’. The prior data is about user’s OTT consumption in 2019. Here, the users are classified in one of three groups, Low (0), Medium (1) and High Consumption (2). The latter data is about daily listing (counts) of US students registered, present, absent.

< Question 1 – UserConsump data>

Explore the data set and answer the following question. 1) What are the number of rows and columns? 2) What does each column stand for? Explain.

< Question 2a – UserConsump data>

We want to check the frequency of Google and Youtube’s data consumption. Here we want to see four categories of low (0%<= x <= 25%), mid-low (25% < x <= 50%), mid-high (50% < x <= 75%), and high (75% < x <=100%). Create new variable called UserConsump_ed_ev (which is the same to UserConsump_ed) and write down a R code that produces the new variables called “YouTube_data_d” and “Google_data_d”.

Expected Result

```
> head(UserConsump_ed_vec[,c("src_ip_numeric","YouTube_data_d","Google_data_d")])
  src_ip_numeric YouTube_data_d Google_data_d
1     3232266497         YT_low        GG_low
2     3232266498     YT_mid-low   GG_mid-high
3     3232266499        YT_high   GG_mid-high
4     3232266500        YT_high    GG_mid-low
5     3232266501    YT_mid-high       GG_high
6     3232266502        YT_high    GG_mid-low
```

< Question 2b – UserConsump data>

Below is the frequency table of “YouTube_data_d” and “Google_data_d”. Write down a R code that produces this table and explain the meaning of the numbers and what you can notice from this.W

Expected Result

	GG_low	GG_mid-low	GG_mid-high	GG_high
YT_low	138	33	38	35
YT_mid-low	50	74	64	55
YT_mid-high	32	77	69	65
YT_high	24	59	72	88

< Question 3 – UserConsump data>

As you noticed, the current data format is “wide”. Convert this into long format and create a variable called “UserConsump.long” as shown below.

Expected Result

```
> head(UserConsump.long)
  src_ip_numeric cluster                 variable     value
1     3232266497       0 Amazon_time_occupation     0.000
2     3232266498       0 Amazon_time_occupation  3335.362
3     3232266499       1 Amazon_time_occupation 26998.860
4     3232266500       1 Amazon_time_occupation 12373.206
5     3232266501       0 Amazon_time_occupation 10672.897
6     3232266502       1 Amazon_time_occupation 25134.147
```

< Question 4 – UserConsump data>

In UserConsump.long, there are many elements included in “variable”. Since these elements contain both the name of OTT service and type of occupation, it is difficult for us to separate them. In this respect, do the following task. 1) create a variable called “type”, which indicates whether it is about “time” or “data”. 2) Convert the values of “variable” to have only the name of OTT. 3) Rename “variable” to “OTT”. Below is the updated UserConsump.long.

Expected Result

```
> head(UserConsump.long)
  src_ip_numeric cluster    OTT     value type
1     3232266497       0 Amazon     0.000 time
2     3232266498       0 Amazon  3335.362 time
3     3232266499       1 Amazon 26998.860 time
4     3232266500       1 Amazon 12373.206 time
5     3232266501       0 Amazon 10672.897 time
6     3232266502       1 Amazon 25134.147 time
```

< Question 5 – UserConsump data>

Write down a R code that answers the following questions: 1) What are Top 10 OTT services that users consume the time mostly? 2) What are Top 10 OTT services that users consume the data mostly?

Expected Result

(1)

	OTT	value
13	Google	92796481
19	GoogleServices	12082105
20	HTTP	10826017
1	Amazon	9177739
55	YouTube	5758410
21	HTTP_Proxy	5165123
12	GMail	4546011
9	Dropbox	4056039
49	WhatsApp	3346472
45	Twitter	3202591

(2)

	OTT	value
55	YouTube	515811148
14	GoogleDocs	514056042
20	HTTP	347797438
12	GMail	195844839
15	GoogleDrive	135879144
2	AmazonVideo	122659385
1	Amazon	114423505
13	Google	96398399
49	WhatsApp	52488599
6	AppleStore	52217035

< Question 6 – UserConsump data>

Compare the average time consumption between “low” consumption group and “high” consumption group. To do so, (1) Create a table that compares average time consumption between “low” and “high” group (name it “UserConsump.long.lh”). (2) Calculate average overall average time consumption between two group. Knowing the fact the time is measured in seconds, compare those two with the unit of minutes.

Expected Result

(1)

```
> head(UserConsump.long.lh)
          OTT         Low       High
1      Amazon 3614.513789 8225.39878
2 AmazonVideo  270.677217  506.11787
3       Apple   49.161219  857.78082
4 AppleiCloud    8.339664  607.13028
5 AppleiTunes    1.772032   93.35627
6  AppleStore    1.897551   71.94670
```

(2)

```
      Low      High
18.06463 38.30120
```

< Question 7 – daily.attend data>

‘Date’ variable contains month/day/year information. Use ‘Date’ variable to create ‘month’, ‘day’, and ‘year’ variables.

Expected Result

```
> head(daily.attend)
  School month day year SchoolYear Enrolled Present Absent Released
1 01M015    01  04 2016   20152016      168     157     11        0
2 01M015    01  05 2016   20152016      168     153     15        0
3 01M015    01  06 2016   20152016      168     163      5        0
4 01M015    01  07 2016   20152016      168     154     14        0
5 01M015    01  08 2016   20152016      168     158     10        0
6 01M015    01  11 2016   20152016      167     160      7        0
```

< Question 8 – daily.attend data>

The current form of ‘SchoolYear’ variable may confuse people. Add “-“ between two years as shown below (Hint: use substr() function).

Expected Result

```
> head(daily.attend)
  School month day year SchoolYear Enrolled Present Absent Released
1 01M015    01  04 2016  2015-2016      168     157     11        0
2 01M015    01  05 2016  2015-2016      168     153     15        0
3 01M015    01  06 2016  2015-2016      168     163      5        0
4 01M015    01  07 2016  2015-2016      168     154     14        0
5 01M015    01  08 2016  2015-2016      168     158     10        0
6 01M015    01  11 2016  2015-2016      167     160      7        0
```

< Question 9 – daily.attend data>

It seems like ‘Enrolled’ variable should be the one that sums up ‘Present’, ‘Absent’, ‘Released’ variables. Write down a R code that checks whether there is any case that the summation of ‘Present’, ‘Absent’, and ‘Released’ does not equal to ‘Enrolled’.

< Question 10 – daily.attend data>

In which month do students absent mostly on average? Write down a R code that can provide an answer for this question as shown below. Also, provide logical explanation for your answer.

Expected Result

	month	Absent
1	01	54.44190
2	02	51.42469
3	03	54.22450
4	04	49.14762
5	05	50.79031
6	06	75.69964
7	09	39.21051
8	10	38.69605
9	11	44.32010
10	12	50.19072