ST344: Professional Practice of Data Analysis 2023-24

Report 1 - Part 2: Speed Data

Example Solution

	Male	Female	All
Number			
Number (%)	277 (50.3)	274 (49.7)	551 (100)
Age			
Mean (SD)	26.6 (3.5)	26.1 (4.0)	26.4 (3.8)
Median (IQR)	27 (24-29)	26 (23-28)	26 (24-28)
Missing (%)	3 (1.0)	5 (1.8)	8 (1.5)
Average Interest in Activities (Scale 1-10, 10=high)			
Sports	7.0 (2.4)	5.7 (2.7)	6.4 (2.6)
TV-sports	5.0 (2.9)	4.1 (2.6)	4.5 (2.8)
Exercise	6.2 (2.4)	6.3 (2.5)	6.3 (2.5)
Dining	7.4 (1.8)	8.2 (1.6)	7.8 (1.8)
Museums	6.5 (2.1)	7.5 (1.9)	7.0 (2.1)
Art	6.2 (2.4)	7.2 (2.0)	6.7 (2.3)
Hiking	5.5 (2.6)	6.0 (2.6)	5.7 (2.6)
Gaming	4.4 (2.6)	3.3 (2.5)	3.9 (2.6)
Clubbing	5.6 (2.5)	5.9 (2.5)	5.8 (2.5)
Reading	7.4 (2.1)	7.9 (2.0)	7.6 (2.0)
TV	4.9 (2.5)	5.7 (2.5)	5.3 (2.5)
Theatre	6.0 (2.2)	7.5 (2.1)	6.8 (2.3)
Movies	7.6 (1.8)	8.1 (1.7)	7.9 (1.8)
Concerts	6.5 (2.3)	7.1 (2.0)	6.8 (2.2)
Music	7.7 (1.9)	8.0 (1.7)	7.9 (1.8)
Shopping	4.8 (2.5)	6.5 (2.5)	5.6 (2.6)
Yoga	3.8 (2.7)	5.0 (2.8)	4.4 (2.8)
Missing (%)	5 (1.8)	2 (0.7)	7 (1.3)
Shared Interest with partner (Scale 1-10)			
Mean (SD)	5.2 (2.3)	5.7 (2.0)	5.4 (2.2)
Missing (%)	35 (6.4)	34 (6.2)	69 (12.5)

Table 1: **Speed Data study: Summary of gender, age, activities, and shared interests for 551 participants**

Example of a paragraph:

Table 1 shows a good gender balance (50%). Participants were young (median 26 years, IQR 24-28). On average, participants' highest interest was in music, movies and dining, and lowest in gaming, TV-sports and yoga. Compared to the opposite gender, male participants had more interest in sport, females had more interest in Theatre and Shopping. Both laid medium importance (5.4) on shared interests (12.5 % missing data). Questions can be raised about representativeness of the sample (young graduate students at prestigious Universities) and the range of activities covered (Yoga and gaming, but not travelling and reading).

(95 words)

Marking Scheme

30 marks for Part 2: 15 each for the table and the paragraph.

Table:

+3 marks if the table looks clean with formatted text (something autogenerated from an R command or other means). 4-8 marks if the student has created a more customised table, depending on effort/detail.

+2 marks if decimal places are sensible and consistent across values (if presented)

+2 marks for having a title or caption

+1 to +3 marks if the student has included the number of missing values

-3 to -5 marks if the student has only given means but not included variability measures such as IQR or SD in brackets.

Paragraph:

up to 5 marks if the student has only written about numbers in their table. For example "Of the 551 respondents 277 were male and 274 were female ..." but with no further interpretations.

+1 to +5 marks if the student has discussed activities of highest and lowest interest and differences between male/female in the interest in activities

+1 to +5 marks if the student has discussed differences between age groups in the interest in activities

+1 to +3 marks if student has discussed what population the sample might or might not represent: graduate students, only young individuals, prestigious US universities, there is not much variability in age as only young people at a similar stage of their career are considered

+2 marks if the student has mentioned most speed dating participants attached medium importance to their partner having the same interests

The table below is one example. Other sensible variations of the table are acceptable. Students might include a breakdown of importance of activity by age category - this requires making some age categories, the importance of interest score might be grouped, etc.

Feedback:

Most students made a good attempt at this question and presented a useful summary of the specified variables. There were some very detailed tables.

For the choice of the variable related to "shared interest", some choices and assumptions could have been made. The study includes: "shar1_1", "shar", "shar1", "shar7_2", "shar1_2", "shar1", "shar7_3", "int_corr". You did not have to include all of them, at least one, specifying its meaning, perhaps focusing on those that had some variations across gender/age groups.

Overall, marks were not awarded in full in the following most common cases:

Table:

Very few students discussed or included in their table the number of missing observations for each variable. While the number of missing observations was small in this, it is important to let your client/reader know this. If there were large amounts of missing data this could lead to significant bias in the findings. Few students split the data by creating age groups, largely used a categorization by gender, which didn't let observing much variation across age groups, e.g. individuals in their 20s versus those in their 30s. Among those who created age groups, many created very uneven groups, in which the younger groups included about 150 individuals while older ones included only 20 or even less; this doesn't allow to have a better idea of differences among the young. Very few students included a measure of variability of the data (e.g. standard deviation) to describe the spread of the data.

Aspects related to table formatting (minor impact):

- Many missing titles in the tables
- Labels not formatted to be more readable, e.g. "shar1_1" should have been transformed into "shared interests" or similar
- Some tables were not "clean", either typed in R with # symbol in front, screenshots of the R command View()
- In some cases too many decimals were included. There is no point in having decimals for a count of observations.

  Aspects related to table formatting (medium-major impact):

- Some tables could have been reshaped to improve readability: if the dimensions are gender and the 17 activities, a better view is obtained if gender is placed column-wise and the activities row-wise.

- Columns/rows containing too many words, all squeezed, which made the entire table not readable
- Some tables were not tables but output of command summary() for all variables, these aren't technically tables

Discussion paragraph:

Many comments are "operational", that is instructions regarding how the table was built and to what measure the values correspond to, instead of briefly describing some findings (using numbers too) and some intuitions you may have (e.g. differences among young-old in activities etc.) If the table includes mean and standard deviation, this information should be written in the title/subtitle of the table.

Full marks were awarded when:

- Table is readable in its labels and numbers
- Presence of title (subtitle or subscript if needed)
- Both a measure of location and variability are included
- Presence of age/sex grouping, you would have realized that only sex grouping would not have led to any big differences
- In those cases, in which the "most/least preferred activity" is chosen, the proportion/count of individual is included
- Avoiding "operational" comments but rather discussing and interpreting the findings (f.ex. where you see the largest differences, some unexpected results).