Week 4: Quiz ④

Q1. I can just impute the mean for any missing data. It won't affect results and improves power.
False

Q2. You're cleaning up a dataset with 1000 observations. You notice that one categorical column contain 532 missing values. What strategy should you employ to deal with these missing values?
iii. remove the column entirely

Q3. Given the observed data, the missingness mechanism does not depend on the unobserved data. This is missing at random.

Q4. Which of the following statements regarding the split-apply-combine strategy are true?
a. The split-apply-combine strategy is similar to the map-reduce strategy for processing large data.
d. The R package plyr supports the split-apply-combine strategy.

Q5. Data wrangling deals with the following tasks:
a. data transformation
b. data visualisation
e. data cleaning

Q6. Valid strategies for dealing with missing values in a column containing numerical data, assuming that we can't afford to lose any data.
c. Predict the missing values based on the other variables in the data
e. Replace missing values with the column median.

Q7. What are possible problems of skipping features with missing value (i.e. skipping columns of the data) to handle missing values?
c. If an input at prediction time has a feature missing that was always present during, this approach is not applicable.
d. So many features are skipped that prediction accuracy can degrade.

Q8. It is always better to remove data points with the missing values (i.e. rows) as opposed to removing missing features (i.e. columns)

Answer: False

Q9. Even accounting for all the available observed information, the reason by observations being missing still depends on the unseen observations themselves. This is the case of _____

Missing not at random

Q10. Wide format is preferred for recording data for data visualisation & data analysis

Answer: False