MAE4000 Data Science Exam

Autumn 2023

3-components Portfolio Exam

Each component counts for one-third of the final grade and you need to pass each component to pass the exam. You have to pass all three components in the same semester for the exam result to be valid.

Submission deadline: October 09 at 14:00 (through INSPERA)

[Portfolio] Component I: Data wrangling & Auditing

Go to the people list here in CANVAS and click through on your own name, your **ID** is the number in the webaddress on the top that comes after users. Note it down and fill it in to get your own custom datasets for the portfolio component here: https://cemo.shinyapps.io/datagenerator/ .

This will allow you to download two individualized datasets to wrangle & clean.

1. *STUDENTS-YOURID.txt*
   containing responses of 5000 students on 31 binary items (correct = 1, 0 = incorrect) of a national examination: BP1 to BP31. The students are located in 7 regions (REGION: 1 to 7) and have a regional student number (ID).

2. *ITEMS-YOURID.txt*
   containing binary codes indicating for each of the 31 items whether or not (i.e., 1 or 0) they belong to one of 4 subject-domains (i.e., 4 variables = 4 columns). Notice that an item can target more than one domain.

The desired dataset:

- Long format with a response by a student on an item as research unit of the study
- VARIABLES in order of the columns in the final dataset
    - (1) a new ID variable uniquely identifying each of the 5000 students
    - (2) the variable REGION with information on the student's region
    - (3) a variable MREGION containing the average score for a student in the region the student belongs to
    - (4) a variable MSTUDENT containing the average score for the student across the 31 items
    - (5) an ITEM variable uniquely identifying each of the 31 items

- (6-9) the variables D1 to D4 coding for the subject domain the item belongs to. Thus for instance D2 codes for whether or not the item belongs to subject domain 2 (as given in the ITEMS dataset by the corresponding Domain.2 variable.
- (10) a variable Y containing the binary response of the student on the item
- (11) a variable SCORE (see below)

Extra modification needed

1. The binary responses need to be rescored such that if the student has ended the test with a string of wrong responses (i.e., at least three zeros), these are all scored as missing (e.g., if responses on items BP27 to BP31 are all zeros then recode as NA, but not if BP31 would be correct).

- A miniature example: If this is the string of responses 101101000 then it becomes 101101NANANA whereas 101101001 remains untouched.
- You will likely need a loop to implement this (cf. probability module) [an approach by means of regular expressions is also feasible, but more involved]. It is probably easier to rescore the responses before you turn the dataset in a long format.

2. Let your code sample at random one of the 4 domains D1-D4 and compute a new variable SCORE that contains for each student the average item response across the items belonging to that domain.

Deliverables:

- a single clean RDS datafile *DATA-YOURID.RDS*
    - Your dataset's dimensions would be approximately (depending on data auditing decisions) 155000 research units x 11 variables.
- the corresponding annotated Rscript *COMPI-YOURID.R* to generate this datafile
    - remember to include explicit statements to import the two starting data files and export the final dataset to RDS.
    - you cannot use functions from packages that are not included in base R (i.e, this implies among others no library or require calls)
    - remember to reason for data management and/or auditing decisions in the syntax comments

Disclaimer: Scenario+data are obviously fiction, yet data issues and operations are real-life inspired.

[Portfolio] Component II: Data Visualization

The data visualization component consists of two parts, a critical evaluation of a data graphic and your own design data graphic.

Deliverables

- 1 PDF file Visualization-YOURID.pdf containing 4 pages (see below) and layout in APA style (i.e., font size 12, Times New Roman, 1 inch margins, Double-spaced lines).
    - Any page outside the stated page limits and requirements will not be read.
- 1 R file Visualization-YOURID.R (see below) for part 2

Five Criteria

1- Gestalt principles & visual structure
2- Keep it simple: Decoding & Operations
3- Less is more: Chartjunk & data-ink ratio
4- Graphical data integrity & lie factor
5- Annotation & stand-alone readability

Part 1. Graphic inquisition

-
    - Include the target figure of your critique on a separate page in APA style and with mentioning of its source.
        - Make sure you choose a figure that has sufficient meat to it, that there is something to critique. *It needs to be a data graphic, not an infographic.* When in doubt, verify with me!
        - It cannot be a figure that we discussed already in class.
        - *Grading: Auto-zero when it does not comply with requirements.*
    - Discuss on maximum one page the figure's potential merits and/or flaws according to the 5 criteria of good graphical design for data representation and using relevant terminology.
        - You have only one page so do not focus on non-crucial details, but clearly show that you understand priorities.
        - *Grading: Equal weight across criteria.*

Part 2. Graphic design

-
    - Include your own designed figure on a separate page in APA-style
    - **Figure requirements**

        - represents at the minimum 3 variables
        - consists of a minimum of two panels (either by means of facets or using the layout grid system)
        - You can use one of the pre-assigned datasets in class or use one of your own choice, but it should be freely accessible to us within R.
        - The corresponding R-code used to construct the figure needs to be ggplot-based, annotated, and stand-alone reproducible.
        - *Grading: Auto-zero when it does not comply with requirements OR when it is an uncustomized copy of an R-example found on the web.*
    - Explain on maximum 1-page what the figure intends to show/clarify including motivation for design choices you made
        - *Grading: Rated according to the same 5 criteria, with equal weight across criteria, and taking into account your own argumentation for the chosen design*

[Portfolio] Component III: Statistical Inference

Georgiou, G. P., & Kilani, A. (2020). The use of aspirated consonants during speech may increase the transmission of COVID-19. *Medical hypotheses, 144*, 109937. https://doi.org/10.1016/j.mehy.2020.109937

Critique this paper in terms of Abelson's MAGIC criteria and explicitly answer the following guiding questions in the process.

1. What do they claim?
2. What did they actually do?
3. What can you conclude based upon what they did?

Rewrite the paper's main result in line with recommended reporting practices

- if you are missing certain information/specifics, you can use a placeholder instead (e.g., <<Insert statistic x here>>)

*5 terms that need to be included in your paper: *population, research unit, sample size, effect size, and measurement*

Deliverables

- one report *COMPIII-YOURID.pdf, APA style (1inch margins, double-spaced, 12pt font size), max. 6 pages*
    - *Any page outside the stated page limits & requirements will not be read*

MAE4000 Data Science Grading Guide

Autumn 2023

Component I: Data wrangling & Auditing

	STUDENTNR	0	
	YOURID	77789	
	MAXSCORE		
DATA-IM/EXPORT (score binary 1/0)	**3**	**3**	
*proper import STUDENTS-YOURID.txt	1	1	
*proper import ITEMS-YOURID.txt	1	1	
*proper export to DATA-YOURID.RDS of final dataset	1	1	
DATA AUDITING	**6**	**4**	
*Intentional search for anomalies (Score = Grade: systematic search and reporting even on parts where no problems occurred)	2	1	
*Finding & Handling anomalies with justification! (4 anomalies 1 point per: duplicate Ids in region, region labels, binary response having non-binary values, ITEMS contains a double)	4	3	
DATA STRUCTURING	**6**	**5.5**	
*long format with response as research unit	2	2	
*Unique student ID based on original ID (max. 0,5 if just disconnected serial row numbers)	1	0.5	
*succesful merger STUDENTS & ITEMS: D1-D4	2	2	
*exact variable names & order as specified (Score binary: "ID REGION MREGION MSTUDENT ITEM D1 D2 D3 D4 Y SCORE")	1	1	
DATA COMPUTATION	**9.5**	**5**	
* MSTUDENT (mean bystudent)	2	2	
* MREGION (mean of above by region! Max 1, if merely direct average across region of item responses)	2	1	
* RESCORING FINALSTRINGOFZEROS (at least three then all) to NA)	2	0	
SCORE (same random sampled domain D1-D4 for all students and then average across items from that domain)	2	2	
* Up to date computation AFTER rescoring/anomaly handling (.5 pointy per newly computed score)	1.5	0	
CODE SYNTAX STRUCTURE	**4.5**	**4**	
* code runs from start to end	3	3	
* readability: structure & comments	1.5	1	

TOTAL	29	21.5	
Grade	A	C	

Cannot use functions from packages that are not included in base R (i.e, this implies among others no library or require calls).

Grade Transformation	Score	
F	<15	
E	15	
D	18	
C	21	
B	24	
A	27	

INSTRUCTIONS

Go to
https://cemo.shinyapps.io/datagenerator/
fill in number the student used for data file import
download the two files & put them in Component I folder
shorten source path in import to filename
you can now run their syntax!

Component II: Data Visualization

	STUDENT STUDENTNR YOURID	Example 1
	MAXSCORE	
Graphic Inquisition subscore	14	11.454545
source of figure	1	1
1- Gestalt principles & visual structure	2	2
2- Keep it simple: Decoding & Operations	2	1
3- Less is more: Chartjunk & data-ink ratio	2	1.5
4- Graphical data integrity & lie factor	2	2
5- Annotation & stand-alone readability	2	1.5
2,5 points per main theme that is explicitly discussed with proper supporting evidence.		

Figure should be a data graphic not an infographic (and also not a figure discussed in class); if they ignored this basic requirement, fill in a 0, or a 1 otherwise.	1	1

Graphic Design subscore	**14**	**8**
1- Gestalt principles & visual structure	2	1.5
2- Keep it simple: Decoding & Operations	2	1
3- Less is more: Chartjunk & data-ink ratio	2	0.5
4- Graphical data integrity & lie factor	2	2
5- Annotation & stand-alone readability	2	2
APA compliance (Bold Figure next line italic title; actual figure, & below an optional note)	2	0
Match intended message - design choices	1	0
Readable R script	1	1
Reproducible R script	1	1
Req: 3 variables+panels/facets+not mere 4 data points	1	1
NOT Potential plagiarism/mere internet copypaste: fill in a 1, or a 0 otherwise.	***1***	***1***
COMMENTS		
Total	**28**	**19**
	A	**D**

Grade Transformation	Score
F	14
E	16
D	18
C	20
B	23
A	26

Component III: Statistical Inference

	STUDENT STUDENTNR YOURID	Example 1
	MAXSCORE	
What do they claim? Aspirated consonants->droplets->COVID	**3**	**3**
COVID transmitted by droplets (sneezing coughing) --> talking or breathing is sufficient to produce droplets --> (Inouye, 2003) aspirated consonants produce more droplets in comparison to unaspirated consonants--> countries whose dominant language contains aspirated consonants had more cases of COVID-19 than countries whose dominant language does not have aspirated consonants --> has "epidemiological implications" for COVID transmission in each country		
What did they actually do? Comparison of COVID ranks of 2 groups of countries	**3**	**2**
sample description n= 26 countries who were at top of unclear COVID ranking of certain date excluded "outliers" for weakly justified reasons comparison of two groups of countries but size and categorization unclear/unknown (M = 254.9, SD = 159.5) vs (M = 206, SD = 121.9); independent group t-test t(18) = 0.73, p > .05 (group difference statistically not significantly different from zero) mixing Fisher (p value, null hypothesis) & Neyman-Pearson (alternative hypothesis) boxplot with different label variable than terminology used for test		
What can you conclude based upon what they did? (list of seen objections - can loosely interpret/classify student's text, yet no double scoring of same argument and each objection can only be checked once so 1 point max. Total capped at 7 points, so no need to count beyond).	**7**	**7**
Check source (journal + author)	1	
Sample issues (e.g., small convenience, strange exclusion of cases, mismatch reported sample size and effective sample size in analyses …)	2	2
Mismatch operationalization - theory - inferred generalization	1	1
ecological fallacy issue: conclusions at level different than research unit level in data	1	
measurement issues (droplets, aspirated consonants, COVID cases, …)	2	
causal language not warranted by study design	1	1
not considering alternative explanations (cf. Confounding, artefacts, omitted variable bias…)	1	1
ignoring limitations of study in conclusion	1	

unclear what data were actually used (source, computation of variables, …)	1	
selectively cherry-picking results (and literature) and ignoring others that are not in-line with desired conclusion (e.g., ignoring that observed difference is statistically non-significant)	1	1
other that do not roughly categorize in the above	1	1
MAGIC criteria Abelson in Critique	**10**	**6**
Magnitude (size of "effect" in light of size of "cause", effect size)	2	1
Articulation (specific and precise formulation)	2	1
Generality (breadth in support and in applicability)	2	2
Interestingness (change, impact, surprise, value)	2	1
Credibility (believable, theoretically coherent, sound methods)	2	1
Rewrite Results / Words included and properly used	**7**	**5**
Descriptives: group M's and SD's and sample size	1	1
Inferential part for group difference: par (SE), test stat, p OR CI (NOT both as that would be redundant)	2	1
Effect size	1	1
population (OK if elsewhere in critique)	1	0
research unit (OK if elsewhere in critique)	1	1
measurement (OK if elsewhere in critique)	1	1
Total	**30**	**23**
Grade	**A**	**C**

Comment

Grade Transformation	Score	
F	15	
E	16	
D	18	
C	21	
B	24	
A	27	

Overall Grade

	STUDENT	Example	
	STUDENTNR		1
	YOURID		
	I	**21.5**	
		C	
		4	
	II	**19**	

III	**D** **3** **23** **C** **4**	
GRADE	**C**	
AVE	3.66666667	
ROUND(AVE)	4	
sumscore		63.5

AVE: rounded average of all three components is the final grade, with the exception that failure of one component results in an overall failure