MAE4000 Data Science autumn 2022

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 14 at 14:00 (through INSPERA)

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/Links to an external site.

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

1. *STUDENTS-YOURID.txt*
   containing responses of 4000 students on 31 binary items (correct = 1, 0 = incorrect) of a national examination: BP1 to BP31. The students are located in 5 regions (REGION: 1 to 5) 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 4000 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

Extra modification needed

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 two 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.

Deliverables:

- a single clean RDS datafile *DATA-YOURID.RDS*
  - Your dataset's dimensions would be approximately (depending on data auditing decisions) 124000 research units x 10 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.
  - remember to reason for data management and/or auditing decisions in the syntax comments

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.

<u>Deliverables</u>

- 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

<u>Five Criteria</u>

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 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!
    - *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 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*

Component III: Statistical Inference

Rao, M.E. & Rao, D.M. (2021). The Mental Health of High School Students During the COVID-19 Pandemic. Frontiers in Education. https://doi.org/10.3389/feduc.2021.719539Links to an external site.

-
    - *You can skip the subsection on Principal Component Analysis and do not worry about not getting the full technical details of the subsection on multivariate regression analysis: Study design > Analysis*

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?

*5 Words that need to be included in your paper: *population, sample size, effect size, mental health, measurement*

<u>Deliverables</u>

- 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*

Grading guide for MAE4000 Data Science autumn 2022

Component I: Data wrangling & Auditing

	STUDENTNR	XXXXX	
	YOURID	XXXXX	
DATA-IM/EXPORT	**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**	**3**	
*Intentional search for anomalies (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 coded as 3, ITEMS contains a double)	4	2	
DATA STRUCTURING (needs to account for auditing actions)	**6**	**5**	
*long format with response as research unit	2	2	
*Unique student ID based on original ID (not just serial row numbers)	1	0	
*succesful merger STUDENTS & ITEMS: D1-D4	2	2	
*exact variable names & order as specified (ID REGION MREGION MSTUDENT ITEM D1 D2 D3 D4 Y)	1	1	
DATA COMPUTATION (needs to account for auditing actions)	**8**	**6**	
* MSTUDENT (mean bystudent)	2	2	
* MREGION (mean by region)	2	2	
* RESCORING FINALSTRINGOFZEROS to NA	2	2	
* Up to date computation AFTER rescoring/anomaly handling	2	0	
CODE SYNTAX STRUCTURE	**4**	**4**	
* code runs from start to end	1	1	
* all required packages at front of syntax	1	1	
* readability: structure & comments	2	2	
TOTAL	**27**	**21**	
Grade		**C**	

Comment

Grade Transformation	Score	
F	**<13**	
E	**13**	
D	**16**	
C	**19**	
B	**22**	
A	**25**	

INSTRUCTIONS
Go to
https://cemo.shinyapps.io/datagenerator/
fill in number the student used for data file import

download the two files, these are the data that the student used--> allows to run their syntax if you adapt their file paths in import statements

Component II: Data Visualization

	STUDENTNR	XXXXXX
	YOURID	XXXXXX
Graphic Inquisition subscore	**14**	**7.5**
source of figure	1.5	1.5
1- Gestalt principles & visual structure	2.5	1.5
2- Keep it simple: Decoding & Operations	2.5	0
3- Less is more: Chartjunk & data-ink ratio	2.5	1.5
4- Graphical data integrity & lie factor	2.5	2
5- Annotation & stand-alone readability	2.5	1.00
2,5 points per main theme that is discussed with proper supporting evidence.		
Figure should be a data graphic not an infographic; if they ignored this basic requirement, fill in a 0, or a 1 otherwise.	1	1
Graphic Design subscore	**20**	**15.5**
1- Gestalt principles & visual structure	2	2
2- Keep it simple: Decoding & Operations	2	1
3- Less is more: Chartjunk & data-ink ratio	2	2
4- Graphical data integrity & lie factor	2	2
5- Annotation & stand-alone readability	2	1

APA compliance (Bold Figure next line italic title; actual figure, & below an optional note)	2	0
Argumentation for design:	2	2
Does message come across?	3	2.5
Reproducible R script	1	1
Readable 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	**34**	**23**
		C

Grade Transformation	Score	
F	**<16**	
E	**16**	
D	**19**	
C	**22**	
B	**25**	
A	**29**	

Notes
For the five criteria see the set of course slides.

Component III: Statistical Inference

	STUDENTNR	XXXXXX
	YOURID	XXXXXX
What do they claim?	7	5.95
degradation in mental health of high schoolers before --> during pandemic	4	4
identified stressors using many outcome variables & accounting for preconditions		
less for Asian students	1	1

no gender difference	0.5	0
no influence of preconditions+therapy	0.5	0
exercise time reduces degradation	1	1
online schooling worsens degradation	1	0.5
results are step from correlation to causality	1	1
results are generalizable	1	1
What did they actually do?	**7**	**5**
sample description: restricted subset of students in biomedical programme of 1 high school in Mason, Ohio, USA	1	1
web-based survey administered without incentives & anonymous	1	0
n = 107 voluntary self-selection mostly Asian & female & in-person teaching (imbalance between 2/3-3/4)	2	2
all variables are self-report: binary, likert 5point, freeform	1	0.5
pairwise t-test t = 0.636 (p ≪ 0.001) (with Shapiro-Wilk normality test (W = 0.944, p-value = 0.00025)).	1	1
cronbach alpha, PCA, correlations, mostly only reported test statistic and p-value & graphics of before-during not in line of design	1	0.5
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 10 points, so no need to count beyond).	**10**	**10**
Check source (journal + author)	1	
Mismatch operationalization - theory - inferred generalization	1	1
"longitudinal" while just question about past /recall bias / perceived metal health	1	1
small / unrepresentative sample versus population size / character	1	1

selection bias / restriction of range / convenience sampling ...	1	1
broad variable labels vs thin/narrow actual measured variable	1	1
leading questions / item formulation unvalidated and unreliable procedures	1	
ignoring proxy status of operationalized measure to draw overly general conclusions	1	1
ignoring proxy status of limited sample to draw overly general conclusions	1	1
common language use of statistical terminology	1	
causal language not warranted by study design	1	1
conclusions ignoring stated limitations	1	
not considering alternative explanations (cf. Confounding, artefacts, omitted variable bias...)	1	1
claiming intervention effect in absence of active control group	1	
non-double blind and other potential sources of researcher bias	1	
staring blindly at statistical significance without considering effect size	1	1
unclear what data were actually used to get to the final results	1	
unclear method section hiding relevant operational details & obstructing reproducibility	1	
redundant reporting and technical terminology ”name dropping”	1	
no mentioning of assumptions modelling approach	1	
zero justification for techniques used/choices made	1	
no replication although different samples are readily available	1	
widely differing effective sample size in parts of the analyses or sample size as implied by reported results differs from the one being reported	1	

selectively cherry-picking results (and literature) and ignoring others that are not in-line with desired conclusion	1	
misreporting of results	1	1
other	1	
MAGIC criteria Abelson	**7**	**5.5**
Magnitude (size of "effect" in light of size of "cause", effect size)	1.5	0
Articulation (specific and precise formulation)	1.5	1.5
Generality (breadth in support and in applicability)	1.5	1.5
Interestingness (change, impact, surprise, value)	1	1
Credibility (believable, theoretically coherent, sound methods)	1.5	1.5
Words (included in text + proper use)	**5**	**5**
population	1	1
sample size	1	1
mental health	1	1
measurement	1	1
effect size	1	1
Total	**36**	**31.5**
Grade	**A**	**A**

Comment

Grade Transformation	Score	
F	<16	
E	16	
D	19	
C	22	
B	25	
A	29	