MAE4000 Data Science autumn 2025

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.

[Portfolio] Component I: Data wrangling & Auditing

Note down your exam number 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.

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

1. *STUDENTS-YOURID.txt*
   containing responses of 6000 students on 21 binary items (correct = 1, 0 = incorrect) of a national examination: BP1 to BP21. The students are located in 8 regions (REGION: 1 to 8) and have a regional student number (ID).
2. *ITEMS-YOURID.csv*
   containing binary codes indicating for each of the 21 items whether or not (i.e., 1 or 0) they belong to one of 6 subject-domains. 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 6000 students
    - (2) the variable REGION with information on the student's region
    - (3) a variable MSTUDENT containing the national test score for the student computed as the sum across the 21 items
    - (4) a variable MREGION containing the average national test score for a student in the region the student belongs to
    - (5) an ITEM variable uniquely identifying each of the 21 items
    - (6-11) the variables D1 to D6 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).

- (12) a variable Y containing the binary response of the student on the item
- (13) a variable DOMAINSCORE-x. Before you start computing this new variable, let your code first sample at random one of the 6 domains and substitute the x in the variable name by this domain number, then to compute this variable, calculate the average item response across the items belonging to that domain.

Extra modification needed to ace this component

- The binary responses need to be first rescored such that if the student has ended 5 subsequent items incorrect, then all responses on items after these sequence of items are set to incorrect. Thus, for instance if item responses BP7 to BP11 are all 0, then also item responses BP12 and onwards are set to 0, even if the student had correct responses on some later items!

Deliverable I.1:

- a single clean RDS datafile *DATA-YOUREXAMNUMBER.RDS*
  - Your dataset's dimensions would be approximately (depending on data auditing decisions) 126000 research units x 13 variables.

Deliverable I.2:

- The corresponding annotated Rscript *COMPI-YOUREXAMNUMBER.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
  - you cannot use functions from packages that are not included in base R (i.e., anything that requires an install.packages call)!

[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.

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.
- Discuss on maximum one page the figure's flaws (and potential merits) according to the 5 criteria of good graphical design for data representation and using relevant terminology.
    - explicitly structure your page per criterium: <criterium X>: two arguments (repeat for the other 4 criteria)
    - 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 yet auto-zero when Figure does not comply with requirements*.

Deliverable II.1

- 1 PDF file Visualization-YOUREXAMNUMBER.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.

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)
            - *caveat: Having a figure with three or more panels, where only a single variable is shown per panel = auto zero on this part of the component.*
        - 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.
    - Summarize on on maximum 1-page (i) what the figure intends to show/clarify and (ii) for each <criterium X>, a justification for a design choice 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. Yet, auto-zero when the Figure does not comply with requirements OR when it is an not-customized copy of an R-example found on the web.*

Deliverable II.2. 1 R file *Visualization-YOUREXAMNUMBER.R* with code for recreating the submitted figure

<u>[Portfolio] Component III: Statistical Inference</u>

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

Explicitly structure your answer according to the following three sections:

1. What do they claim?
2. What did they actually do? This includes study design and obtained results.

3. What can you conclude based upon what they did? Divide this section in 5 subsections that align with **Abelson's MAGIC criteria.**

The first two sections should remain factual but complete; reserve your critique for the third section.

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

Deliverable III.

- one report *COMPIII-YOUREXAMNUMBER.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*
    - *A paper that does not adher to the requested structure receives auto-zero.*

Grading guide MAE4000

[Portfolio] Component I: Data wrangling & Auditing

	EXAM NR	1234
	MAX	
DATA-IM/EXPORT (score binary 1/0)	**3**	**3**
*proper import STUDENTS-YOURID.txt	1	1
*proper import ITEMS-YOURID.csv	1	1
*proper export to DATA-YOURID.RDS of final dataset *needs to be exact*	1	1
DATA AUDITING	**7**	**5,5**
*Intentional search for anomalies (systematic search and reporting even on parts where no problems occurred)	2	1
*Finding & Handling anomalies with justification! *(4 anomalies = ,75 point per: duplicate ITEMS, non-binary item response, duplicates within-region student IDs, multiplied REGION IDs) & ,5 points each for justification*	5	4,5
DATA STRUCTURING	**3**	**2**
*Long format with response by a student on an item as research unit	1	1
*succesful merger STUDENTS & ITEMS	1	1
*exact variable names & order as specified (ID REGION MSTUDENT MREGION ITEM D1-D6 Y DOMAINSCORE-x) [x replaced by their sampled domain nr]	1	0
DATA COMPUTATION	**10**	**10**
* unique ID across all students	1	1
* **MSTUDENT** sum across all 21 items for student	1	1
* **MREGION** average MSTUDENT score in region student belongs to	2	2
* **DOMAINSCORE-x:** let your code first sample at random one of the 6 domains and substitute the x in the variable name by this domain number, then to compute this variable, calculate the average item response across the items belonging to that domain.	2	2
* The binary responses need to be first rescored such that if the student has 5 subsequent items incorrect, then all responses on items after this sequence of items are set to incorrect. Thus, for instance if item responses BP7 to BP11 are all 0, then also item responses BP12 and onwards are set to 0, even if the student had correct responses on some later items!	3	3
* Up to date computation AFTER rescoring/anomaly handling (.333 point per newly computed variable)	1	1

CODE SYNTAX STRUCTURE	3	3
* code runs from start to end	2	2
* readability: structure & comments	1	1
TOTAL	26	23,5
Grade	A	A
COMMENT		

They cannot use functions from packages that are not included in base R.

Grade Transformation	Score	
F	13	
E	15	
D	17	
C	19	
B	21	
A	23	

INSTRUCTIONS
Go to
https://cemo.shinyapps.io/datagenerator/
fill in exam number of the student for data file import
download the two files & put them in Component I folder
shorten source path in import to filename -> run syntax

[Portfolio] Component II: Data Visualization

	EXAM NR	1234
	MAX	
Graphic Inquisition subscore	11	9
APA style (Including source of figure)	1	1
1- Gestalt principles & visual structure	2	1
2- Keep it simple: Decoding & Operations	2	2
3- Less is more: Chartjunk & data-ink ratio	2	1
4- Graphical data integrity & lie factor	2	2
5- Annotation & stand-alone readability	2	2
2 arguments per criterion, no repetition of same design feature under more than one criterium, priority issues addressed, 1 page total		
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	11	7

1- Gestalt principles & visual structure	2	1
2- Keep it simple: Decoding & Operations	2	1
3- Less is more: Chartjunk & data-ink ratio	2	1
4- Graphical data integrity & lie factor	2	2
5- Annotation & stand-alone readability	2	2
correct implementation & justification per criterion		
APA compliance (Bold Figure next line italic title; actual figure, & below an optional note)	1	0
Reproducible R script	1	1
Min Req> 3 variables+2 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***

Total	**22**	**16**
Grade	**A**	**C**
COMMENT		

Grade Transformation	Score
F	10
E	11
D	13
C	15
B	17
A	20

[Portfolio] Component III: Statistical Inference

	EXAM NR	1234
	MAX	
What do they claim? (factual complete intro/abstract/discussion no critique yet)	**7**	**5**
COVID transmitted by droplets (sneezing coughing)→ talking or breathing is sufficient to produce droplets	1	0
support Inouye (2003) aspirated consonants produce more droplets in comparison to unaspirated consonants	2	2
countries whose dominant language contains aspirated consonants have more cases of COVID-19 than countries whose dominant language does not have aspirated consonants	3	3
"epidemiological implications" for COVID transmission	1	0
What did they actually do? (factual complete study design+results no critique yet)	**7**	**6**

Sample: 26 countries mostly infected (more than 1000 cases) by the COVID-19 as of 23rd of March; + excluded Switzerland since there are many official languages in its territory & countries with many cases (e.g., Italy) or very few cases (e.g., Japan) per million of population; effective sample size = 20	2	2
Measures: number of cases per million of population at country level + language group (unclear which/how many countries belong to what group)	2	1
Analysis: Descriptives (boxplot) & independent t-test; no significant differences in cases of "inflected" individuals between the languages with aspiration (M = 254.9, SD = 159.5) and the languages without aspiration (M = 206, SD = 121.9), [t(18) = 0.73, p > .05]	3	3
What can you conclude based upon what they did? Critiques in subsections according to Abelson's MAGIC criteria	**18**	**14**
Magnitude (size of "effect" in light of size of "cause", effect size)	3	2
epidemiological implications sounds big + actual effect size = small given variation + not stat significant so within sampling variation+other		
Articulation (specific and precise formulation)	4	4
not very specific or precise in how causal mechanism works + how measures are defined + contradicting use of inferential results + limited justification for choices of data + deceiving sample size 20 instead of 26 + unclear who is what group + other		
Generality (breadth in support and in applicability)	4	3
ecological fallacy (country = research unit vs individual = infected) + selection bias sample + small sample + basis of covid numbers unreliable/variable +unclear/ambiguous definition of language groups+other		
Interestingness (change, impact, surprise, value)	3	2
timely: paper during pandemic + interdisciplinary: connecting non-obvious element + no practical utility value: country cannot change language + other		
Credibility (believable, theoretically coherent, sound methods)	4	3
study design does not support causal inference+problems with all other criteria + authors & journal + causal mechanism lacking + confounders + other		
Words (included in text + proper use)	**2**	**1,6**
population	1	1
effect size	1	1
sample size	1	0
research unit	1	1
measurement	1	1

INLINE WITH REQUIREMENTS: apa 12pt font double-spaced 1inch margins + explicit sectioning as in guide	1	1
Total	**34**	**26,6**
Grade	**A**	**C**
Comment		

max6 pages anything beyond is not read

Grade Transformation	Score
F	18
E	19
D	21
C	24
B	27
A	30

Overview

	EXAM NR	**1234**
	I	23,5
		A
		6
	II	16
		C
		4
	III	26,6
		C
		4
FINAL	**GRADE**	**B**
	AVE	4,666667
	ROUND(AVE)	5
	sumscore	66,1
	Rounded average of all three components is the final grade, with the exception that failure of one component results in an overall failure	