MAE4000 Data Science autumn 2024

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 (you receive it ca. 2 weeks in advance) 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. *ITEMS-YOUREXAMNUMBER.txt*
    - dataset containing 3 variables representing responses on 30 survey items by parents of 5–6-year-old kids
    - Person unique person identifier
    - Item number corresponding to order of the item in the survey: 1 to 30
    - Y survey response with 1 = agree, 2 = undecided, 3 = disagree
2. *STUDENTS-YOUREXAMNUMBER.txt*
    - dataset containing 4 variables with additional information on the persons taking the survey.
    - P unique person identifier equivalent to Person in the former dataset
    - G gender of the child, with 1 = female, 2 = male
    - S self-reported perceived travel time (SRTT) of the family to school, with 4 ordered categories: 1 = very short, 2 = short, 3 = long, 4= very long
    - A age of the child in years

The desired dataset:

- A single dataset in wide format with Person as the research unit
- Variable order: Person Gender SRTT Age Item01-Item30 Undecided1 Undecided2 Use
- Rows sorted on increasing Person identifier number.
- The survey responses on the uneven items have been reverse-coded.
- The two new variables Undecided1 and Undecided2 show the number of undecided responses given by the person on the first and second half of the survey, respectively;

- The new variable Use records whether or not (TRUE/FALSE) the person has used each of the three response categories (1-2-3) across all survey items;
- ***TO ACE this component:*** Provide an extra variable Match which contains for each person, the Person ID of one other person that has a response pattern across all other variables listed above (from Gender to Use) that most closely matches with the given person [count a variable with a dissimilar value as one mismatch].

Deliverable I.1:

- a single clean RDS datafile *DATA-YOUREXAMNUMBER.RDS*
    - Your dataset's dimensions would be approximately (depending on data auditing decisions) n research units x (4+30+3+1) variables.

Deliverable I.2:

- he 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 required 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.
- 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. 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

[Portfolio] Component III: Statistical Inference

Beall, A. T., & Tracy, J. L. (2013). Women Are More Likely to Wear Red or Pink at Peak Fertility. *Psychological Science*, *24*(9), 1837-1841. https://doi-org.ezproxy.uio.no/10.1177/0956797613476045

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

Structure your paper according to the following three subsections

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

*Five terms that need to be included in your paper: *sample, effect size, replication, comparison, 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*

Grading guide MAE4000

Portfolio exam autumn 2024

[Portfolio] Component I: Data wrangling & Auditing

	EXAM NR	
	ID	52800
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 *needs to be exact*	1	1
DATA AUDITING	**7**	**4.5**
*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 = ,75 point per: duplicates ITEMS, out of range survey response not 123, duplicates STUDENTS, out of range Age=6+) & ,5 points each for justification*	5	3.5
DATA STRUCTURING	**4**	**3**
*Wide format with person as research unit	1	1
*succesful merger STUDENTS & ITEMS	1	1
*exact variable names & order as specified (Person Gender SRTT Age Item01-Item30 Undecided1 Undecided2 Use [wild card for Match can be skipped or at start or end])	1	1
*rows increasing by Person	1	0
DATA COMPUTATION	**9**	**6**
* uneven items reverse coded	1	1
* **Undecided1/2**: two new variables with # of 0 responses per person 1st/2nd half of survey	2	2
* **Use**: new variable coding for whether or not all response categories were used across the survey per person	2	2
* **Match:** for each person, the Person ID of one other person that has a response pattern across all other variables listed above (from Gender to Use) that most closely matches with the given person [count a variable with a dissimilar value as one mismatch].	2	0
* Up to date computation AFTER rescoring/anomaly handling (.5 pointy per newly computed score)	2	1

CODE SYNTAX STRUCTURE	**3**	**3**
* code runs from start to end	2	2
* readability: structure & comments	1	1
TOTAL	**26**	**19.5**
	Grade	**A**	**C**
	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	52800
Graphic Inquisition subscore	**11**	**8**
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	2
4- Graphical data integrity & lie factor	2	1
5- Annotation & stand-alone readability	2	1

arguments on point & covers most prominent features

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
APA compliance (Bold Figure next line italic title; actual figure, & below an optional note)	1	0
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***

Total	**22**	**15**
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	**52800**
What do they claim?	**5**	**3.75**
females increase own attractiveness by wearing reddish colors	0.5	0.5
especially during peak fertility	0.5	0

red and pink adornment in women is reliably associated with fertility and that female ovulation, long assumed to be hidden, is associated with a salient visual cue	3	3
high conception risk vs low conception risk 3 more likely to wear pink/red	0.5	0.25
77% of women wearing red/pink were at high rather than low risk	0.5	0
What did they actually do?	**7**	**4.5**
2 samples: A) n= 100 American women [Age 18-47;M=26] via Amazon Turk and for monetary compensation; B) n = 24 female undergraduates at University of Britisgh Columbia [Age: 17-31; M=19] for course credit	2	1.5
excluded 40+, birth control users, smokers, pregnant, atypical cycle, etc.	0.5	0
all measures online survey self report	2	2
excluded according to confidence in estimate	0.5	0
wear red/pink for high vs low risk: Sample A: 26% vs. 8%, $\chi2(1, N = 100) = 5.32$, $p = .02$, odds ratio = 3.85; Sample B: 40% vs. 7%, $\chi2(1, N = 24) = 3.82$, $p = .051$, odds ratio = 8.67.	1	1
of those wearing red/pink (n=A:17/B:5), 76% of women in Sample A and 80% of women in Sample B were at high risk; for other color (n=A:83/B:19), 46/32 % were art high risk.	1	0
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. Capped at 8 arguments, so no need to count beyond).	**8**	**7**
Check source (journal + author)	1	1
mismatch: Theory = within-subject / Study = between-subject		
mismatch: Theory = Behavior / Study = self-report survey		
Sample: small & unrepresentative	1	1
unvalidated and unreliable measurement procedures: "fertility"	1	1
unvalidated and unreliable measurement procedures: "red/pink color wearing"	1	

ignoring proxy status of operationalized measure to draw overly general conclusions	1	
ignoring proxy status of limited sample to draw overly general conclusions	1	1
causal language not warranted by study design	1	
not considering alternative explanations (cf. confounding, artefacts, …)(e.g., context, color-habits, weather, culture, social desirability, ...)	1	1
non-equivalent comparison group (e.g., two colors vs all other)	1	
staring blindly at statistical significance without considering effect size / ignoring statistical significance	1	
unclear what data were actually used / effective sample size for different parts (e.g., different restrictions in sample size and collapsing of groups)	1	1
unclear method section hiding relevant operational details & obstructing reproducibility	1	
no replication although different samples are readily available	1	
selectively cherry-picking results (and literature) and ignoring others that are not in-line with desired conclusion	1	1
other	1	
MAGIC criteria Abelson	**7**	**3.25**
Magnitude (size of "effect" in light of size of "cause", effect size)	1.5	0.75
Articulation (specific and precise formulation)	1.5	0
Generality (breadth in support and in applicability)	1.5	1
Interestingness (change, impact, surprise, value)	1	1
Credibility (believable, theoretically coherent, sound methods)	1.5	0.5
Words (included in text + proper use)	**2.5**	**2.5**
sample	1	1
effect size	1	1
replication	1	1
comparison	1	1
measurement	1	1
Total	**29.5**	**21**
Grade	**A**	**C**
Comment		

Grade Transformation	Score
F	<15
E	15
D	17
C	20
B	23
A	25

Overview

EXAMNR	52800
I	19.5
	C
	4
II	15
	C
	4
III	21
	C
	4
GRADE	C
AVE	4
ROUND(AVE)	4
sumscore	55.5

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