382.1.2601

Introduction to Data Science

מבוא לניתוח נתונים

סמסטר אביב, 2023-2024

Final project

Pick an interesting research question and answer it using data

The project is 75% of the class final grade[1].

The goal of the project is for you to demonstrate proficiency in the techniques we have covered in class (and beyond, if you like) and apply them to data in meaningful ways. The goal IS NOT to do an exhaustive analysis of every possible statistic in the data or use every method you have learned or will learn about in class.

Deliverables

1. ~~Proposal~~ (10%) - Done.
2. Presentation (20%) - submit on Moodle and deliver in class.
3. Peer evaluations (8%) - link will be provided later on.
4. Final report (62%) that meets the requirements below and includes a link to a code repository and data (if you collected your own data). Submit on Moodle.

ALL of the above deliverables except for peer evaluations **ARE REQUIRED** in order to get a passing grade in the project and subsequently in class.

[1] As specified in the syllabus, submission of all project parts is a prerequisite for getting a passing grade in the project, and a passing grade in the project is required to get a passing final grade in class.

Presentation

(20% of final project grade)

Deliver a five minutes presentation (four minutes of presentation + one minute for questions) in week #9, #10 or #11. The presentation should motivate the research question(s) tackled in the project, cover the state of current knowledge and what your project contributed to it, briefly describe the data collected and methods used to process it, and highlight your key findings. Each team member should say something substantial. Your presentation should not cover everything you have done ("we did this, then we did this, etc."), instead, it should convey key decisions you've made, why, and what you found.

Time will strictly be enforced. Make sure to allocate sufficient time to the things you want to cover the most and help your project stand out. It is recommended to practice the presentation a few times before giving it in class.

<u>Grading and evaluation:</u>

10 points	Is the research question clearly articulated and the problem well-motivated?
20 points	Breadth and depth of familiarity with prior work and/or theory.
10 points	Novelty: What is the contribution and/or innovation of this analysis?
10 points	Data overview
20 points	Methods appropriateness and quality of execution.
10 points	Findings
10 points	Professionalism: (1) How well did the team present? (2) Does the presentation appear to be well-practiced? (3) Did the team present a unified story, or did it seem like independent pieces of work patched together? (4) Are the slides well organized, readable, not full of text, featuring figures with legible labels, legends, etc.?
5 points	Teamwork: Did everyone get a chance to say something meaningful about the project?
5 points	Time management: Did the team spend time well during the presentation? Finished on time or got cut off?

– 100 points total.

Peer evaluations

(8% of final project grade)

Each student will randomly and anonymously be assigned to evaluate five other projects. Submitting all of your assigned assessments will earn you 25 points. The remaining 75 points will be based on the assessments you receive from your peers.

Final report requirements

(62% of final project grade)

You need to submit a report as a single pdf per team on Moodle with up to 3 pages of the main content (including figures and tables) and an unlimited number of pages for appendices. Note that the appendices may not be evaluated at all. The main content of the report needs to include the following information:

- Link to a code repository: Provide a link to your clean and neatly organized code repository with clear instructions in the README.md file for replicating your analysis end-to-end. Any part that is not reproducible should clearly be indicated. DO NOT include data in your code repository, but DO provide instructions on where to place the data your code relies on.
- Link to a folder with your data: This only applies to projects that use public data or data you collected, not staff-provided data, which we already have access to.
- Section 1 - Introduction: Introduce your general and specific research question(s). What is the general problem area that this analysis contributes to? What specific problem are you trying to solve? Why is this important? Why is this hard? What does theory / prior work tell us about this problem and how are you extending it? What is your approach?
- Section 2 - Data overview: Describe your data at a high level, answering questions such as what are the entities in the data, how many entities are there, and what are the features or feature families relevant to the problem you're tackling?
- Section 3 - Methods and results: Describe the *key results* of your analysis and the model(s) used to obtain them. Explain why you chose to model the data in this particular way and what you've learned from it. Make sure to check how well the model fits the data and verify the model results are well-supported by the data. You are free to use figures and tables in this section (as well as elsewhere in the main body), but these count towards your 3-page limit. Think carefully about how you're using your 3 pages.
- Section 4 - Limitations and Future Work: What are the limitations of your approach and /or findings in generalizing beyond your sample? In addition, briefly describe future directions you would follow if you had an additional month or additional three months to work on this project.

The report can be written in Hebrew or English, and it does NOT have to be generated from Rmd notebook. DO check the readability of the text of the report and that the figures' information is readable.

Grading and evaluation:

The final report will be graded as follows:

5 points	Is the research question clearly articulated and the problem well-motivated?
10 points	Breadth and depth of familiarity with prior work and/or theory.
10 points	Novelty: What is the contribution and/or innovation of this analysis?

30 points	Methods appropriateness and quality of execution of the analysis.
10 points	Creativity and critical thinking.
5 points	Future work potential for innovation and feasibility.
15 points	Report quality.
5 points	Is the amount of work conducted reasonable for the team size?
10 points	Code organization, cleanliness, and reproducibility.

– 100 points total.

FAQ

Can I use models not officially covered in class?
Yes, if you're feeling adventurous, but you don't have to.

Can I use models implemented in other languages than R?
You can, but you need to get the staff's approval for that in the proposal phase, provide a good reason for doing that, and do all other parts of the project in R and RStudio.

Can I use other R packages for data loading and manipulation outside of tidyverse?
Generally no, with one exception being the *data.table* package, which is allowed.

How will you evaluate my visualization(s)?
Subjectively based on our best judgment. You do not need to visualize all of the data at once. A single high quality visualization will receive a higher grade than a large number of poor quality visualizations.

I still have some unanswered questions, who do I go about getting answers for them?
Your best and first option should be the project forum on Moodle. Other students probably have similar questions or already tackled a similar issue. Asking questions on Moodle will help you get answers faster and allow other students to benefit from them. If you did not get an answer within 24h, you are welcome to email the staff. Your other option is to come to office hours, just email us ahead of time to coordinate.