# Instructions
# Capstone project IDS (MA)

This project is supposed to simulate the Data Science Cascade (answering questions by using data and code), thereby yielding actionable insights and elucidating meaning.

## Specific steps:

1) Form a small team of 2-5 people (this can in principle also be a one-person solo team, if you strongly prefer that, but that will be a lot of work, as this is designed as a team project) and give your team a suitable name.
2) Find a dataset in a subject area/domain that you're interested in. The richer (lots of rows and columns) this dataset is, the better.
3) Ask questions of this data and answer them using the topics we covered in this class. The general workflow should be that you ask a question, then write code that – together with the data – yields a numerical result (and a figure). You then need to interpret the figure and the numerical result to provide a qualitative answer to the question you started with.
4) The deliverable (to be uploaded to the Brightspace portal) is a zip file that consists of 3 parts: The file with the project report (PDF, please), the data file (if the data is too large to upload, please provide a link to an online repository in the report) and the code file. The zip file should be named *"dsga1001_capstoneProject_groupname.zip"*
5) The project report should be 5 pages (4-6, if you must), with the following structure:
    a. Page 1: Introduction page. Should have a header that lists the group name, then all full names of all team members along with their netID. After the header, please describe the dataset in detail (what do the rows represent, what do the columns represent, where did you get it from) and your overarching interest in this dataset. Also outline your general approach as to how you will handle missing data and extreme values in this dataset. In case there is any pre-processing of the dataset done, e.g. dimensionality reduction, explain that here as well (usually, there will be a need for that). Make sure to include a *brief* introduction to the domain or topic (e.g. mortgage markets, medical diagnostics, etc.) without jargon, aimed at non-experts and why you find this issue particularly important or interesting.
    b. Page 2: Inference question. Open this page with a qualitative (narrative, but specific) question. Then, explain how you are – in principle – answering the question (e.g. "we're comparing the means of these two columns, implementing an A/B test"). After that, state the numerical results and include a figure that illustrates the result. Note that no code is necessary here. Put the code in the code appendix (which ideally consists of 3 sections that ideally correspond to the analysis code for pages 2-4). End this page with a qualitative (narrative) answer that interprets the quantitative results you obtained. To answer your question, you can use any inference method we discussed in class (parametric tests, nonparametric tests, Bayesian inference or resampling methods), but pick one. Make sure to comment on the usual metrics of statistical inference (e.g. p-values, confidence, effect size, etc. – as suitable for the approach you pick – e.g., if the approach is Bayesian inference, you need to comment on Bayes factors).
    c. Page 3: Prediction question, otherwise structure like b). This will become more clear when we do the ML part of the class, but here, use a supervised learning method that makes predictions of some sort (multiple regression, regularized regression, some kind of NN, etc.). Basically ask (and answer) a question of the form: "Does feature x predict outcomes y, while controlling for confounds $c_1$, $c_2$ and $c_3$"? Make sure to comment on the accuracy of the prediction by talking about $R^2$, residuals or RMSE.

d. Page 4: Classification question, same structure as in b). Again, this will become more clear once we do ML, but we would like you to ask a classification question, and then use this data and a classification algorithm (e.g. trees or forests or boosting) to answer it. Make sure to also do include some kind of clustering algorithm before doing the classification. Also make sure to include some kind of assessment tool to judge the quality of the classification (e.g. ROC or P/R curves).
e. Page 5: Overall summary and conclusions. Make sure to touch on the following: What did you learn about the data from doing these analyses, all in all (integrating the results from the 3 answers to the 3 questions)? What overarching conclusions can you draw? What assumptions did you make? What limitations does your analysis have? How could the questions be answered better, if better data were available? Anything else you noticed or realized while doing this project?

## Grading rubric (GP = grade points):

**Part a**:

1) Is the domain/topic introduced properly? [*0.5 GP*]
2) Is the dataset explained clearly? (What do rows and columns represent and how did it come about/where did it come from) [*0.5 GP*]
3) Is the overall approach to handling missing data/extreme values suitable and explained well? [*0.5 GP*]
4) Is the preprocessing reasonable (e.g. reasonable factors and number of factors in dimensionality reduction) and explained well? [*0.5 GP*]

*Each* of the questions (**parts b/c/d**) will be graded by the following rubric:

1) Is the question clearly stated? [*0.5 GP*]
2) Is the analysis approach that will answer the question with data described clearly? [*0.5 GP*]
3) Are the computations done properly and is the right answer (e.g. the correct p-value, within numerical precision) arrived at? [*0.25 GP* - the code appendix will be particularly valuable to check this part]
4) Does the analysis contain a suitable figure (e.g. a bar graph with confidence intervals that estimate the effect or one that compares mean differences with SEM, for the interference question, some best fit line for prediction and maybe a decision boundary for classification)? [*0.25 GP*]
5) Is the question answered adequately by properly interpreting your own results? (e.g. adopting some kind of decision criterion for inference, talking about $R^2$ or RMSE in prediction and ROC or P/R in classification) [*0.5 GP*]

The rubric for **part e**:

1) Overall conclusion/takeaway about this dataset, across all specific questions? [*1 GP*]
2) What limitations threaten this conclusion (assumptions you had to make, biases that you are aware of, scope and generality of the study, etc.) [Note: This is completely ok. *Every* study has limitations – the important part is to be explicit about that, so we can tell that you are aware of them] [*0.5 GP*]
3) In an ideal world, what would a dataset look like that doesn't have these limitations and how could one go about gathering such a dataset? [*0.5 GP*]
4) Anything else you noticed / find remarkable about the dataset that is not already contained in any of the specific questions above [up to 0.5 GP extra credit]

Thus, a maximum of 10.5 GP are attainable. We grade out of 10 grade points for a perfect score.

That is how grade points are **gained**. Points can also be **lost**, if best practices are violated. So please make sure to use "best practices" (as introduced in the class) throughout.
For instance, for the inference question, make sure to do a power analysis before doing anything else (to make sure you have adequate statistical power to even answer your question). For the prediction question, make sure to statistically control for confounds. For the classification question, make sure to use a suitable train/test split to avoid overfitting. Among many other things (e.g. proper handling of missing data, proper handling of extreme values, etc.)

## Finally, three logistical stipulations:

1) To prevent cheating (e.g. you downloading the whole thing from somewhere [please don't do this – it is easily detected], it is very important that you – at the beginning of the code file – seed the random number generator with the N-number of one of the team members. That way, the correct answers will be keyed to your own solution (as this matters, e.g. for the specific train/test split). As N-numbers are unique, this will also protect your work from plagiarism.
   Failure to seed the RNG in this way will also result in the loss of grade points.
2) To prevent chaos, please make sure that only ONE (1) of the authors of the report (the one whose N-number is used to seed the RNG above) uploads the report to their Brightspace portal. All authors who are listed on page 1 of the document will get full credit. This is important to avoid duplicates – all of the logistics of the grading will be handled algorithmically, and doing otherwise can introduce complications, as Brightspace can be a bit wonky. So please trust us and designate one student from the team who will supply their N-number and upload the report to the portal before the due date.
3) When identifying the team and listing the team members in the report, if the team is larger than 2 people, please also outline "author contributions" – specify how each team member contributed to the project, specifically. The point of this stipulation is to prevent freeloading, which is increasingly likely, the larger the group is. Importantly, by being listed here, all authors confirm that they approve of the final version (the one that is uploaded to Brightspace) of the report and understand that they will receive the same grade score for this project, for purposes of the course grade.

In conclusion, we do wish you all the best in executing on this guidance. We aimed at an optimal balance between specific and instructions that yet give you maximal leeway to implement in whichever way you see fit, while still being able to be graded in a consistent and fair manner.
Everything should be doable from what you learned in this course.
If you take this project seriously and do a quality job, you can easily use it as an item in your DS portfolio. Former students told us that they secured internships and even jobs by well executed capstone projects that impressed recruiters and interviewers.