CHEAT SHEET – METHODOLOGY DOs AND DON’Ts

DO	DON’T
Do describe your project in **clear and simple** terms and in a **logical order.**	**Don’t** try to impress (or bamboozle) the reviewer with excessive statistical complexity and jargon.
Do commit to research questions before seeing the data, and account for multiple comparisons.	**Don’t** do anything that looks like “data dredging” — i.e., cherry-picking the results from a large group.
Do have “falsifiable” hypotheses, and be open to any possible conclusion from your data.	**Don’t** say anything that “pre-empts your conclusion”, or suggests that you want a certain conclusion to be true (e.g., getting a “significant” effect).
Do be clear and precise about whether or not you are interested in cause-and-effect, and use precise and consistent language for this.	**Don’t** equivocate between statistical association and causality, or treat predictive and causal terms as if they were interchangeable.
Do consider confounders, mediators, etc., and use appropriate experimental/statistical protocols to deal with these.	**Don’t** write in a way that might make the reviewer think that you don’t understand that “correlation is not cause”.

CHEAT SHEET – DATA AND MODELLING DOs AND DON’Ts

DO	DON’T
Do ensure that your data are obtained from an appropriate sampling method that allows you to make sound statistical inferences.	**Don’t** gloss over the sampling method as if it were unimportant, or omit it and hope no-one notices.
Do ensure that your sample size is appropriate to the constraints and trade-offs of the project, and justify your sample size.	**Don’t** leave available data unused unless there is good reason to, or make arbitrary or unjustified choices on trade-offs.
Do give clear heuristic explanations of statistical methods and models, backed up with references.	**Don’t** get into the weeds with all the technical details of the statistical methods and models.
Do formulate a proposed model that you think will be appropriate to your data.	**Don’t** lock yourself into specific model choices that might turn out to be inappropriate.
Do consider common statistical issues, and frankly acknowledge these. State clearly how you will deal with these issues.	**Don’t** “gloss over” tricky statistical issues, or try to bamboozle the reviewer with jargon to avoid dealing with them properly.

CHEAT SHEET - DESCRIBING YOUR METHODOLOGY

Description	What you should explain
Methodology and research goal (2-6 sentences)	What is your **population** of interest? What **characteristic(s)** are you interested in, and how are these **measured**? Give an clear and logical overview of how you will use data to answer your research question. Distinguish clearly between the variable you are interested in, and any “proxy” or operational measurement being used as a stand-in for this.
Experimental protocols (if needed) (4-8 sentences)	Are you doing **predictive inference** or **causal inference** — i.e., do you need to know the *causal effects* of variables or not? If you are interested in making causal inferences, how will you go from statistical associations to cause-and-effect? What (if any) protocols have you imposed — e.g., randomisation/blinding. How are you dealing with confounding variables and mediator variables?
Pre-registered research questions (1 sentence)	Have you publicly pre-registered your research questions? If so, where are these registered? (This is something you should consider – pre-registration of research questions in a public repository prevents *post hoc* analysis and adds credibility.)

CHEAT SHEET - DESCRIBING YOUR DATA

Description	What you should explain
Source of your data (1-2 sentences)	What is your **sampling frame**, and how did you get your **sample** from this sampling frame — i.e., what is your **sampling method**? If necessary, reiterate the target population in your research and its difference to the sampling frame.
Joining data from multiple sources (1-3 sentences)	If you are joining data from multiple sources, specify the sampling frame and sampling method for each source. Describe, in general terms, how you join these data sets (i.e., how you match people/items in the different sets). Don't go into technical detail — just give the reviewer a clear idea of the final data.
Variables and types (1-3 sentences)	List the variables in the final data and their types (e.g., numerical, categorical, ordinal, etc.). If there are too many variables to list these easily, at least give an overview of contents of categories. Specify which variables are to be predicted, and which are being used as predictors.
Overview and structure (1-3 sentences)	What do your "data frame(s)" look like? How many variables? Don't be ambiguous – are numbers from long-form or wide-form of data? How many data points will you have (sample-size calculation – discussed soon)?

CHEAT SHEET - DESCRIBING YOUR STATISTICAL ANALYSIS/MODEL

Description	What you should explain
Describe the proposed model (1-2 sentences)	You should generally specify a reasonably broad class of model (e.g., linear regression model, negative binomial model, etc.) Describe the model type, and state the output variable and input variables. Don't use model terminology that is not widely known; if you refer to a model form that the reviewer might not know, give a rough explanation to go with it. Have you given enough information to allow the reviewer to write out your model form?
Always specify some wiggle-room (1-3 sentences)	You propose to use a particular class of model, but your specific model will be chosen *after* seeing the data, based on statistical considerations and diagnostic testing. It might involve transformation of variables, changing model form, etc. Reviewer wants to know that you won't cheat to get a pre-conceived conclusion, so make sure you say that your choices will be based on *statistical considerations*.
Outputs, tests and comparisons (1-3 sentences)	What outputs will you get from your model, and what tests and comparisons will you look at? How will you deal with multiple comparisons (if they arise)? Do your outputs answer your research questions? Have you established this clearly?

CHEAT SHEET - DESCRIBING YOUR SAMPLE SIZE

Description	What you should explain
Describe the basis for the calculation (1-2 sentences)	Sample size is determined by reference to a statistical test or inference (e.g., test/confidence interval). What is used as the basis of your calculation? Is data collection done in one go, or is it done sequentially?
Specify desired level of accuracy (1-2 sentences)	There is no such thing as the "right sample size" without specifying the desired level of accuracy for a statistical inference/prediction. This accuracy will be relative to the standard deviation of your —yet to be collected— data. For a **hypothesis test**, specify significance level, power, and minimum detectable effect size. For a **confidence interval**, specify confidence level and interval length.
Describe expected missing data (1-2 sentences)	Anticipate non-response/missing data and account for this in your calculation. If your data collection is done sequentially then this might not be an issue, since you can keep collecting data until you get the amount you want.
Justify your specifications (1-4 sentences)	Show trade-off between accuracy and number of data points, and use this to choose your sample size (calculations in supplementary materials). Make sure you specify your sample size in an appropriate unit of measurement.