Checklist for Initial Data Analysis (IDA) – cross-sectional studies

Reference: Heinze G, Baillie M, Lusa L, Sauerbrei W, Schmidt CO, Harrell FE, Huebner M. Regression without regrets – initial data analysis is an essential prerequisite to multivariable regression. Under revision. https://stratosida.github.io/

Topic	Item	Features
Prerequisites		
Research aim	PRE1	Define the research aim
Analysis strategy	PRE2	Check specification of models and roles of variables in the models
Data dictionary	PRE3	For variables identified in PRE1, and any additional structural variables, check variable labels, definitions, values, units of measurement, data type
IDA screening domain: Missing values (predictor and outcome variables)		
Participant (unit) missingness	M1	Describe the numbers of participants that were potentially eligible but not assessed for eligibility, those who were assessed for eligibility but not recruited and those who were recruited but did not contribute any data.
Variable (item) missingness	M2	Provide number and proportion of missing values for each predictor and for the outcome variable; distinguish by reason of missingness, if applicable.
Complete cases	M3	Describe number of complete observations when considering outcome and predictors for any model in PRE2.
Patterns	M4	Investigate patterns of missing values across all variables, either as tables or appropriately visualized. Can be structured by structural variables.
Missing values – Optional extensions		
Predictors	ME1	Investigate predictors of missingness (complete vs incomplete cases).
IDA screening domain: Univariate descriptions (structural variables, predictors, and outcome)		
Categorical variables	U1	Summarize frequency and proportion for each category or with appropriate plots. If it is considered to collapse rare categories, also summarize frequencies of collapsed categories.
Continuous variables	U2	Inspect distributions with high-resolution histogram, summary of main quantiles (e.g. 1st, 5th, 25th, 50th, 75th, 90th, 99th) and extremes (e.g. 5 highest and 5 lowest values), further measures of location (e.g., the mean) and spread (e.g. Gini mean difference, standard deviation, interquartile range), number of distinct values. Describe the mode of the data and its frequency. Inspect distributions of transformed variables, if applicable.
Univariate analyses – Optional extensions		
Sparsity	UE1	Create distributional plots to identify observations with extreme values
IDA screening domain: Multivariate descriptions (structural variables and predictors)		
Association	V1	Visualize and summarize the association of each predictor with the structural variables
Correlation	V2	Quantify pairwise correlations between all key predictors
Interactions, if applicable	V3	Evaluate bivariate distributions of the predictors specified in interactions. Include appropriate graphical displays.
Multivariate analyses – Optional extensions		
Correlation	VE1	Compare matrix of Spearman and Pearson correlations coefficients
Clustering	VE2	Visualize clustering of predictors using a dendrogram to show closely associated predictors
Redundancy	VE3	Compute Variance Inflation Factors or fit parametric additive models to determine how well each predictor can be predicted from the remaining predictors