The data analysis checklist

[Note: this checklist is from simplystats.org – see below for link]

This checklist provides a condensed look at the information in this book [Leek 2015]. It can be used as a guide during the process of a data analysis, as a rubric for grading data analysis projects, or as a way to evaluate the quality of a reported data analysis.

I Answering the question

1. Did you specify the type of data analytic question (e.g. exploration, association, causality) before touching the data?
2. Did you define the metric for success before beginning?
3. Did you understand the context for the question and the scientific or business application?
4. Did you record the experimental design?
5. Did you consider whether the question could be answered with the available data?

II Checking the data

1. Did you plot univariate and multivariate summaries of the data?
2. Did you check for outliers?
3. Did you identify the missing data code?

III Tidying the data

1. Is each variable one column?
2. Is each observation one row?
3. Do different data types appear in each table?
4. Did you record the recipe for moving from raw to tidy data?
5. Did you create a code book?
6. Did you record all parameters, units, and functions applied to the data?

IV Exploratory analysis

1. Did you identify missing values?
2. Did you make univariate plots (histograms, density plots, boxplots)?
3. Did you consider correlations between variables (scatterplots)?
4. Did you check the units of all data points to make sure they are in the right range?
5. Did you try to identify any errors or miscoding of variables?
6. Did you consider plotting on a log scale?
7. Would a scatterplot be more informative?

V Inference

1. Did you identify what large population you are trying to describe?
2. Did you clearly identify the quantities of interest in your model?
3. Did you consider potential confounders?
4. Did you identify and model potential sources of correlation such as measurements over time or space?
5. Did you calculate a measure of uncertainty for each estimate on the scientific scale?

VI Prediction

1. Did you identify in advance your error measure?
2. Did you immediately split your data into training and validation?

3. Did you use cross validation, resampling, or bootstrapping only on the training data?
4. Did you create features using only the training data?
5. Did you estimate parameters only on the training data?
6. Did you fix all features, parameters, and models before applying to the validation data?
7. Did you apply only one final model to the validation data and report the error rate?

VII Causality

1. Did you identify whether your study was randomized?
2. Did you identify potential reasons that causality may not be appropriate such as confounders, missing data, non-ignorable dropout, or unblinded experiments?
2. If not, did you avoid using language that would imply cause and effect?

VIII Written analyses

1. Did you describe the question of interest?
2. Did you describe the data set, experimental design, and question you are answering?
3. Did you specify the type of data analytic question you are answering?
4. Did you specify in clear notation the exact model you are fitting?
5. Did you explain on the scale of interest what each estimate and measure of uncertainty means?
6. Did you report a measure of uncertainty for each estimate on the scientific scale?

IX Figures

1. Does each figure communicate an important piece of information or address a question of interest?
2. Do all your figures include plain language axis labels?
3. Is the font size large enough to read?
4. Does every figure have a detailed caption that explains all axes, legends, and trends in the figure?

X Presentations

1. Did you lead with a brief, understandable to everyone statement of your problem?
2. Did you explain the data, measurement technology, and experimental design before you explained your model?
3. Did you explain the features you will use to model data before you explain the model?
4. Did you make sure all legends and axes were legible from the back of the room?

XI Reproducibility

1. Did you avoid doing calculations manually?
2. Did you create a script that reproduces all your analyses?
3. Did you save the raw and processed versions of your data?
4. Did you record all versions of the software you used to process the data?
5. Did you try to have someone else run your analysis code to confirm they got the same answers?

XI R packages

1. Did you make your package name "Googleable"
2. Did you write unit tests for your functions?
3. Did you write help files for all functions?

4. Did you write a vignette?
5. Did you try to reduce dependencies to actively maintained packages?
6. Have you eliminated all errors and warnings from R CMD CHECK?

References

1) simplystats http://simplystatistics.org/2015/03/03/the-elements-of-data-analytic-style-so-much-for-a-soft-launch/
2) Leek, J. 2015. The elements of data analytic style, Kindle edition. http://www.amazon.com/Elements-Data-Analytic-Style-ebook/dp/B00U6D80YY/ref=sr_1_1?ie=UTF8&qid=1425397222&sr=8-1&keywords=elements+of+data+analytic+style