Overview of Data Preparation Procedures

Noushin Nabavi, PhD

2018-10-22

The purpose of analytic data preparation is to get the data in a form and shape that can be used for proposed analyses. This document explains the work required to prepare data for analysis after project data is identified, requested, acquired and linked. The first step is to understand the project proposal, metadata, and questions that will be asked using the data. The procedure includes: data cleaning, data wrangling, error monitoring, standardizing codes, harmonizing with meta data, analyzing, and communicating results with team. To answer different research questions, data may need to be transformed into different form and shape. It is important for a data scientist to carefully document and monitor the steps taken for data preparation in relation to the analysis proposed. Therefore, data preparation and analysis is often reiterative and can be completed once everyone on team agrees on a set cohort. Below is an overview of steps for preparing the data ensuring traceability and reproducibility irrespective of analytic software platforms.

A.Data Discovery

1. Identify project questions:

Data projects must identify, acquire and link the data needed. This initial work includes activities such as (i) identifying project questions, (ii) identifying data required to answer project questions, (iii) identifying sources for project data and metadata, (iv) acquiring project data and metadata

2. Identify the best software and tools to answer project questions and install them:

This includes chosing softwares that enables reading the data, transforming or reshaping the data, finding errors in the data, and finally analyzing and tabulating the data. This could include using SQL, Python, R, SAS, etc. The reiterative procedures in this step can be avoided by researching and investing in data cleaning tools (dependencies, libraries) such that the data can be analyzed in bulk and the process is automated.

3. Import and read the data into chosen software(s):

One may need to convert the data upstream outside of a chosen software to have a ready format file in order to read into the chosen software. The steps here include (i) staging or locating the files, (ii) loading or importing the data, and (iii) recoding and transforming the data columns to match metadata/code table, (iv) saving the imported data in a format that is compatible with other softwares and accessible to other users

B.Data Cleaning

4. Monitor errors:

Inspect the dataset thoroughly and get to understand the fields. This includes looking at trends in data such as outliers, minimum and maximum values, as well as missing data, discontinuities or errors. Next step is keeping a log of where most errors are in order to identify and fix the incorrect or corrupt data types, including

typographical errors. A harmonization of names and variables is also necessary if one is integrating the data with other datasets. Examining the data quality includes inspecting the validity, accuracy, completeness, consistency, and uniformity of data columns and rows. An example of finding outliers, means, and margins of error.

4. Scrub for duplicated or erroneous data:

Similar to step #3, as part of the quality assurance, we need to identify duplicates and error schema in the dataset since this will help save time when analyzing. This step also entails scanning for completeness, validity, timeliness, consistency and integrity of data fields. Identify and quantify probabilistic or deterministic linkage of values. This could also include testing the individual columns for unexpected values like NULL values, non-numeric values that should be numeric, out of range values, as well as data linkage discrepancies.

5. Validate and standardize the processes:

(Re)validate the accuracy of the data through scanning for anomalis and contradictions once it has been cleaned. This could include: (i) fixing or correcting column names by checking with metadata, (ii) normalizing or filtering for correct outlier values, (iii) transforming the dates or genders so that they read correctly as needed, and (iv) detecting detect all syntax errors. The purpose here is to avoid lengthy and reiterative codes, write more functions instead of loops, and reduce or compact the source codes so that the procedure is more accurate and reliable. That is, keeping the codes simple and functional is key.

6. Restructure the data and prepare for extraction:

Data restructuring involves data transformations to make analysis easier. These includes (i) data aggregation, transformation, or reshape, (ii) data filtering by cohort, subject, time frame and regions, (iii) de-normalization (flattening), (iv) identifying variables to be included/excluded. Subsequently, data extraction involves deriving data structures and variables that reflect analytic concepts of interests (e.g. cohorts). For this, identify data files/views to be extracted for the project, and define extracted variables, keys, ranges, and necessary calculations.

C. Data Analysis and Reporting

7. Analyze the data:

After the data has been standardized, validated, scrubbed for duplicates, and extracted, use reliable fields to analyze the data for high-level descriptive statistical analyses such as values of mean, standard deviation, ranges, frequencies, or using clustering or classification algorithms. It is also important to consider the linkage probabilities for this step so that those can be built into the margins of error. These results could be visualized using simple frequency tables to more complex plots.

8. Report the data:

The data and the software specifications, codes used to analyze the data, as well as the results are to be closely documented, exported, and saved in related project folders so that they can be accessible to anyone else who may want to replicate the data or produce new results from them.

9. Communicate the results with the team:

Communicate the standardized cleaning process and preliminary results of analysis to the team and receive feedback because fresh set of eyes and people's insights are always beneficial to improving the analysis. This step is also needed to develop and strengthen existing research and policy questions in order to later send more targeted information to stakeholders. Meeting with the team ensures that the team is in line with data cleanup before moving to other parties.