Overview of Data Preparation Procedures

Noushin Nabavi, PhD

2018-10-16

To first achieve the goals and meet expectations of how data assets can benefit us, we need to first determine how we can execute a data analytics workflow successfully. As good practice, a data scientist needs to first understand the overall goal of a project and what each member or stakeholder is looking to achieve from the dataset. A good way to start would be to have all interested parties who are involved start throwing ideas around and brainstorm on possibilities of what datasets can provide. Following a brainstorming session, the manual part of the data cleaning process starts and can be achieved by various softwares. In general, data cleaning is the process of detecting or removing corrupt and inaccurate records from a record set, table, or database and replacing, modifying, or deleting the dirty, missing, or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. Regardless of which software is used, the procedures need to be closely documented, monitored, and inconsistencies reviewed, and as such a protocol for data cleaning is imperative. This is important because incorrect or inconsistent data can lead to false conclusions and misdirected investments on both public and private scales. For instance, the government may want to analyze population income figures to decide which regions require further spending and investment on infrastructure and services. In this case, it will be important to have access to reliable data to avoid erroneous fiscal decisions.
Below are eight broad steps and examples for preparing the data with various softwares so that it is repeatable on different platforms and that the data and results are accurate.

Data Cleaning Steps

1. Install softwares, dependencies, and necessary libraries:

Softwares could include installing new packages or loading existing libraries to be able to **read** the data, **transform** or **reshape** the data, find **errors** in the data, and finally **analyze** and **tabulate** the data. This includes 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.

```
#install.packages()
#library()
```

2. Import the data and read into chosen softwares:

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 format to be compatible with other software. One could also check the top 5 rows of the table to ensure proper input. An instance is shown below:

```
library(ElemStatLearn)
data(prostate)
head(prostate, 5)
```

```
##       lcavol  lweight age      lbph svi       lcp gleason pgg45       lpsa
## 1 -0.5798185 2.769459  50 -1.386294   0 -1.386294       6     0 -0.4307829
## 2 -0.9942523 3.319626  58 -1.386294   0 -1.386294       6     0 -0.1625189
## 3 -0.5108256 2.691243  74 -1.386294   0 -1.386294       7    20 -0.1625189
## 4 -1.2039728 3.282789  58 -1.386294   0 -1.386294       6     0 -0.1625189
## 5  0.7514161 3.432373  62 -1.386294   0 -1.386294       6     0  0.3715636
```

```
##   train
## 1  TRUE
## 2  TRUE
## 3  TRUE
## 4  TRUE
## 5  TRUE
```

3. Monitor errors:

Inspect the dataset thoroughly and get to understand the fields. This includes looking at trends in data, outliers, minimum and maximum values, and missing data 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. An example of finding outliers, means, and margins of error include deciphering the summary of dataset like belo:

```
summary(prostate)
```

```
##      lcavol           lweight           age             lbph
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001
##  Mean   : 1.3500   Mean   :3.629   Mean   :63.87   Mean   : 0.1004
##  3rd Qu.: 2.1270   3rd Qu.:3.876   3rd Qu.:68.00   3rd Qu.: 1.5581
##  Max.   : 3.8210   Max.   :4.780   Max.   :79.00   Max.   : 2.3263
##       svi              lcp             gleason          pgg45
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38
##  3rd Qu.:0.0000   3rd Qu.: 1.1787   3rd Qu.:7.000   3rd Qu.: 40.00
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00
##       lpsa           train
##  Min.   :-0.4308   Mode :logical
##  1st Qu.: 1.7317   FALSE:30
##  Median : 2.5915   TRUE :67
##  Mean   : 2.4784
##  3rd Qu.: 3.0564
##  Max.   : 5.5829
```

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 accuracy and reproducibility of data linkage if this has been performed. This is done by cross checking with the files prior to input and ensuring the number of rows and columns match the metadata before deterministic or probabilistic linkage. This could also include testing the individual column, e.g. for unexpected values like NULL values, non-numeric values that should be numeric, out of range values, as well as data linkage discrepancies. 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. For instance, in RStudio, quality assurance can be performed with *dplyr* and *tidyverse* packages among others.

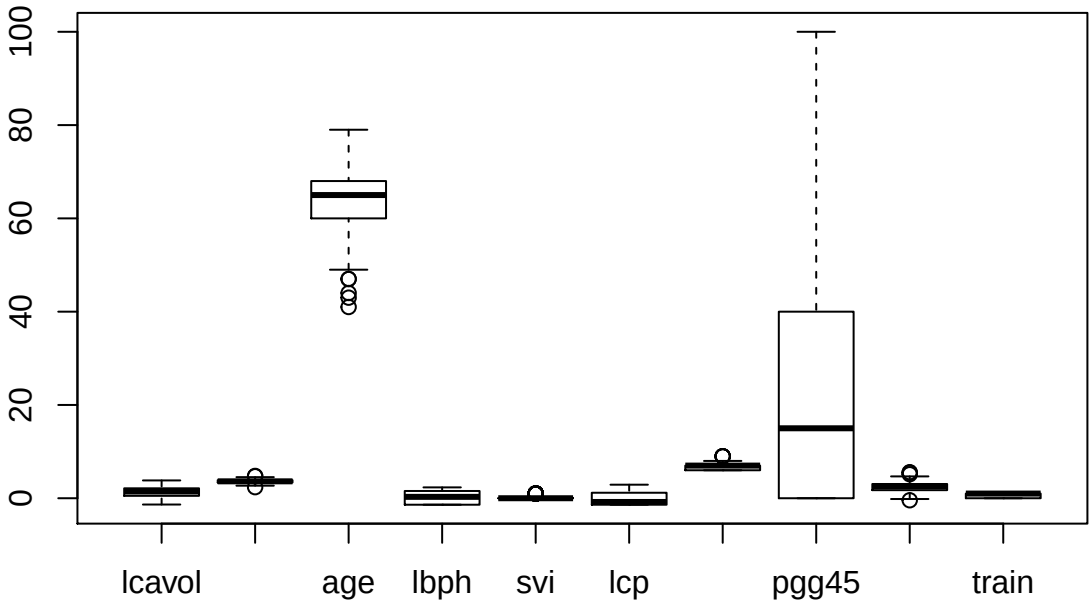
5. Standardize and validate the accuracy of both codes and data:

As reiteration of steps 3-4, it is also important to (re)validate the accuracy of the data through scanning for anomalis and contradictions once it has been cleaned. For example, fix the column names by checking with metadata information, transform the dates or genders, etc. as needed and parsing the data to detect all syntax errors by checking with metadata, or transform the dates or genders, etc. as needed according to the metadata and parsing the data to 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. Note: keeping the codes simple and functional is key. This step should help with reproducibility of codes and data tables so that analysis is repeatable and reputable. Note: keep the codes simple and functional.

6. Analyze the data:

After the data has been standardized, validated, and scrubbed for duplicates, use reliable fields to analyze the data for high-level descriptive statistical analyses such as values of mean, standard deviation, range, or clustering algorithms. These results could be visualized using simple frequency tables to more sophisticated ggplots and graphs. Compile the data plots to provide more complete information for business intelligence and analytics for operational insights. For instance:

```
boxplot(prostate)
```



7. 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 folders so that they can be accessible to anyone else who would want to replicate the data or produce new results from them.

8. Communicate the results with the team:

Communicate the new standardized cleaning process and preliminary results of analysis to the team and receive feedback because fresh set of eyes and people's insights are almost 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.