**Cleaning Large Data Sets**

Laurence T. Burden

Purdue Global University

IN300: Programming for Data Analysis

John Brooks

June 1, 2021

Data cleansing is the act of removing or replacing incorrect data within a data set (Data cleaning: The benefits and steps to creating and using clean data ). Data can be incorrect because it is improperly formatted, missing, or duplicated. Any of these situations will throw off the analysis and lead to incorrect results.

One of the first steps in cleaning data is to determine what information is incorrect in some way (*What is Data Cleaning?*, 2021). A review of the data must be performed. The review can be a manual review by humans, or a computer program can be used to identify anomalies.

Determining how the data will be cleaned can be done once a review is complete. The analyst can come up with the best approach to clean the incorrect data at this point. The strategy is then most likely implemented using a find and replace program for the data. By hand, replacements would be very time-consuming.

Once the data has been cleaned, then the results must be verified (Elgabry, 2019). This is very similar to the first step where the original incorrect data was found. The same review process can be used again to see if any more anomalous data. The actual analysis can now be conducted that the data is a correct as possible.

The final step is to communicate the process with your team (*6 Steps for data cleaning and why it matters*). This step ensures that others know what data was changed and why. The report can also serve as a lessons learned document for others in the future.

Removing duplicates is a common occurrence in cleaning data. As more and more sources are pulled together, the chances of duplicates happening increase. If duplicates are not removed, then the statistical analysis will be thrown off.

Another common occurrence in large data sets is a type mismatch. This is when data types such as numerical and strings are used for the same data in different sources. A data set may have been stored as a string rather than a date type. The analyst must find these instances and correct them.

Syntax errors are close to data type mismatches, but occur when strings aren’t formatted the same. This can include extra spaces and typos. Common typos can include gender or another categorical label. It may not always be a typo, though. It could be two different shortened terms for the same word. This can include Mr. and Mister, fem. and female, or doctor and Ph.D.

These steps can be applied to the IN300\_Dataset1.csv file. The first step to determine the best way to clean the data is to decide the type of analysis that needs to be done. Do we only care about completed network connections and their stats? Maybe we want to know how many connections failed compared to the successful connections.

Assuming we are wanting to replace the missing data, then we would use a process called imputation (Moses, 2016). This would result in replacing non-standard success messages to be replaced by a standard failure code for use during the analysis.

Another aspect that will need to be handled in this data set is the nonstandard time readings. Some have 6 digits after the decimals and others have as few as four. This data can be cleaned by adding trailing zeros if necessary for the desired language.

Yet another issue found in this data is the packet contents occasionally added to the info column. This includes information about the type of response from the destination, such as a SYN-ACK packet. These could be cleaned by removing unnecessary information. This could be done using a regular expression to only include the numbers and len. That would turn “62058 > 8009 [PSH, ACK] Seq=1 Ack=1 Win=2048 Len=110 TSval=895979745 TSecr=4151010 [TCP segment of a reassembled PDU]” into “62058 > 8009 len=110.”

Data cleaning is vital to ensure proper results. Data analysis on dirty data leads to incorrect models. Incorrect models lead to incorrect predictions. Incorrect predictions lead to business failures. For this reason, the data analyst must make sure they are using the most correct data that they can.

**References:**

*Data cleaning: The benefits and steps to creating and using clean data*. Tableau. (n.d.). <https://www.tableau.com/learn/articles/what-is-data-cleaning>.

*What is Data Cleaning?* Sisense. (2021, April 8). https://www.sisense.com/glossary/data-cleaning/.

Elgabry, O. (2019, March 2). *The Ultimate Guide to Data Cleaning*. Medium. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.

*6 Steps for data cleaning and why it matters*. Geotab. (n.d.). https://www.geotab.com/blog/data-cleaning/.

Moses, E. (2016). *R: Data Analysis and Visualization*. Packt Publishing.