**Decision-making and the Business Research Process & Tidy Data**

Louise Bassog

Dr. Bemley

CTEC298

Bowie state university

11/06/2022

**Decision-making and the Business Research Process**

Tim Jones in his article Working with Messy Data emphasizes that for any operation to be carried out accurately and provide efficient results, there is need for information to be used. Such fields include business management and decision-making. Decisions such as marketing, where to invest and what to improve are all based on the information the organization obtains from customers and the market in general. For a business to make the right decisions, the information needs to be as accurate as possible. However, information is not always readily available. Obtaining relevant information for decision is a whole process. Data has to first go through processes before becoming useful information. After raw data is transformed into useful information, the information is stored in data warehouses, where it can be accessed by any party in need. Preparing data involves data cleansing, data blending and data profiling.

A data set is a collection of raw data gathered during research mostly in numeric form. A dataset is mostly represented in tabular form though not always. The components of the dataset are rows and columns which represent members of the dataset and variables respectively. The variables explain the properties of the members using unknowns such as weight, height and age among others. They have a specific data type. Data sets that are not in tabular form can take the form of marked up strings of character. An example is XML file.

Statistics is the study of how data is collected, analyzed, interpreted, presented and organized. Statistics can be considered the grammar of data science because of its role in ensuring that data is as accurate and meaningful as possible. This is clearly illustrated in the article by Jones. Learning statistics enables one to ensure that the data they are working on is valid, consistent and free of errors. It is in statistics that one learns the process of data cleansing, data blending and data profiling, all of which play a major role in transforming raw data into accurate useful information, either manually or automatically (Jones, 2017). All the aspects of statistics above are responsible for creating information that is reliable to other organization.

First, under cleansing as part of the process of cleaning messy data, Tim Jones insists that one has to use cleansing to detect things like errors, missing data and additional data. These short comings have to be dealt with in order to proceed with handling the data. A cleansing tool checks the data through parsing, where each observation is taken from its data file and each independent element extracted. Through parsing, all the accurate data is validated through schema, which checks for mistakes such as duplication, omission, correct entry into all fields and other errors (Jones, 2017). After validation of data, the data moves to a section of data splitting and summarization. The validated data can either be sent to a test file or training file according to the user-defined probability.

On the other hand, data that is not validated is classified as erroneous and passed through an area where correction is done manually before beginning the process again until it is validated and deposited in the splitting and summarization area for use as cleansed test data, cleansed training data or data profile. This means that apart from being able to run automatic cleansing tools, one has to know how to manually detect errors and correct them as this is part of the data cleansing process. According to the article, cleansing tools can be classified as either automatic or manual. Some tools have a self-describing nature which makes operation much simpler to manipulate. There are also some cleansing tools that can be found free as open source. All the discussed aspects make statistics the grammar of data science because of the way it corrects data and gives a more presentable output.

In conclusion, the article emphasizes that data science is key in processing of data into information. This aspect needs to be appreciated since raw data lacking any kind of order or orientation is as meaningless as no data. Moreover, the effectiveness of the decisions made in a business or any other organization depends highly on the accuracy and correct presentation of data. Wrong information interpretation leads to faulty decision making hence the organization suffers terrible consequences as a result of unstructured or wrongly structured data. Moreover, it is quite evident that just as grammar constraints ensure accuracy and effectiveness of language, statistics plays a major role in ensuring that data is accurate and meaningful. Processes such as parsing, profiling and data cleansing whether automatic or manual, ensure that the initial unstructured, messy and faulty data is transformed into more meaningful information that can be used by organizations to make critical decisions that affect the organization. Lastly, it is important to learn how to create data cleansing tools as it is quite easy to do so with just a little training. However, there are various open-source data cleansing tools available on websites that can be used for data cleansing (Jones, 2017).

**Tidy Data**

The article by Hadley Wickham mainly addresses data tidying. Data tidying is known to be a very hectic process that requires a lot of time to complete. Transforming messy data into tidy data takes a lot of skill, keenness and time. No one has attempted to find an easier way of tidying u data for analysis. However, various ways of tidying up data by dealing with issues such as an instance where instead of headers storing variable names, they store values, where a single column contains more than one variable and where variables are scattered in both columns and rows have been realized. Tidy data has been defined as data that is easy to manipulate, modify and visualize. There are also several tools for tidying up data that make the task easier. Implementing these techniques is guaranteed to make data cleaning much easier.

A data set is a collection of raw data gathered during research mostly in numeric form. A dataset is mostly represented in tabular form though not always. The components of the dataset are rows and columns which represent members of the dataset and variables respectively. The variables explain the properties of the members using unknowns such as weight, height and age among others. They have a specific data type. Data sets that are not in tabular form can take the form of marked up strings of character. An example is XML file (Wickham, 2014).

In seeing how statistics can be used to tidy up data for analysis and manipulation, it is evident from Wickham’s article that statistics is the grammar of data science. The same way grammar defines specific constraints to be followed, statistics also defines laws that need to be followed for data to be considered tidy. Data is considered tidy if all variables are stored in columns with each variable stored inside a column, each entity is stored in a distinct row and if different types of related information are stored in different tables. If all this is adhered to, data is considered tidy and is suitable for analysis. However, an instance of getting data that is ready is an exception and hence one has to learn data tidying techniques.

The first instance of messy data pinpointed in Wickham’s article is where a column header has values instead of the name of a variable. Although this might be considered messy, it is sometimes quite useful in saving storage and desired computation. However, where it is not desired, melting or stacking is advisable. This is where rows are turned into columns. The second instance is where more than one variable is stored in one column. In this case, more careful string process would be the solution. The messiest instance, however, is where variables are present in both rows and columns. This case is solved by first melting the table and re-arranging it to solve the error (Wickham, 2014). Another instance is where data concerning different observations are found in the same table. This is where unrelated information is stored in the same table. First, to tidy up and remove inconsistencies, normalization should be the first step. However, there are tools available for solving the error. The tools involve denormalization or merging of data sets. The last type is quite the opposite of the latter, where data related to different observations are found in one table (Wickham, 2014). The problem is easy to fix as it only involves reading the components into an arrangement of tables, adding columns for variables to all the tables formed and lastly combining all tables into one.

In his article, Wickham defines tidy tools are instruments that take in tidy data sets and give out tidy data sets. They are convenient to use in the sense that one tool’s output can be used as another’s input (Wickham, 2014). However, some inputs are inconvenient as they sometimes take in messy data sets or give out messy data sets. This is frustrating as it slows down data analysis. There are tidy and messy tools for manipulation, visualization, and modelling. All these processes run smoother in the presence of tidy data tools and tidy data. Hence tools that either take in or give out messy data should be avoided as they make the data analysis process complex and lengthy.

In conclusion, the article insists that data cleaning can be an easy and less hectic if the data scientist is able to distinguish which one among the mentioned errors is the problem. Once the main error is determined, following the guidelines given can make transformation of messy data into tidy data set. Tidy data is manageable when it comes to manipulation, modelling and visualization. Lastly, using the correct tidy tools saves one the time lost in using faulty tools that either take in or give out messy data. Tidying data, just like grammar, has some rules that need to be followed in order to be achieved.

**References**

Jones, T. M. (2017). *Working With Messy Data*. IBM developer. Retrieved November 6, 2022, from https://developer.ibm.com/tutorials/ba-cleanse-process-visualize-data-set-1/

Wickham, H. (2014). *Tidy Data*. Journal of Statistical Software. Retrieved November 6, 2022, from https://www.jstatsoft.org/article/view/v059i10