Stephanie Nelson

April 6th, 2023

Professor Bemley

**Title**: “Working With Messy Data” and “Tidy Data”

**Author**: Tim Jones and Hadley Wickham

**I. Introduction**

In the articles “Working with messy data” by Tim Jones and Tidy Data by Hadley Wickham we are introduced to how to look at data from a data scientist's perspective. Jones walks us through what messy data is and how to navigate it. It is essential to use clean data for data analysis purposes or your results will not be correct. Wickham takes us through why “tidying data” is so important. This takes the concepts from Jones and further expands on them. This article takes time to discuss data structure, several ways to clean data, and what it looks like while data is being cleaned.

**II. Working With Messy Data**

This article by Jones dives into the reality of working with data. Data needs to be analyzed and sorted due to missing data, incorrect data, and duplicates. Jones takes us through the process of processing data which includes data cleansing, machine learning then data visualization. We learn there are multiple names for the data processing process which could be data extraction and other names. Jones walks us through combining datasets, detecting errors, and a chart that details the process of data cleaning. The diagrams in this article were very useful to understand how to clean data and how it is not a straight forward step. Cleaning data takes most of the process in the data science pipeline as well as multiple eyes.

**III. Tidy Data**

This article by Wickham explains what data preparation looks like through both visuals and text explanations. Wickham begins by introducing data preparation and why it is an essential part of the data science pipeline. The audience also learns that it is hard to find articles that focus on tidying data. Wickham goes on to define tidying data, data structures, and data semantics. The audience learns the correct terminology for datasets and how to look at them from a professional stance. The process of tidying data requires the user to look at columns and rows and modify information to make the data easier to understand.

**V. Conclusion**

In all, both articles played a key role in explaining how to prepare data for us. The use of texts, diagrams, and visuals allows future data scientists to understand what the process entails. I also left with an understanding that data is messy and will take time, patience, and understanding for it to be ready for data analysis.

**Vi. References**

Jones, Tim (2017). Working With Messy Data, https://bsuonline.blackboard.com/bbcswebdav/pid-17985976-dt-content-rid-28398456\_1/xid-28398456\_1

Wickham, Hadley (2023). Tidy Data, http://www.jstatsoft.org/