Tutorial 9: To what extent do you think we should let the data speak for themselves?

Harrison Huang

2024-03-12

Table of contents

Introduction

In the world full data we live in today, there are countless data sets for us to explore. Endless data sets are released every day, every second. However, the dataset in its raw form may vary. Some data that includes many unnecessary "noises", while some data are very simple and include just the right amount of information we need. In most cases, the raw data sets are not "clean". The raw data sets may have missing values, incorrect values, and even incorrect formats. These traits often cause issues when we try to manipulate the data or create visuals with the data. Therefore, before we start creating visuals and manipulations, we often need to "clean" the data. This ensure that the data sets only include the values we need and would not causes any erros later down the road. However, how much data should be cleaned? To what extend should we manipulate the original data? This short essay will discuss the topic of "To what extent do you think we should let the data speak for themselves"?

Why clean the data?

As mentioned in the article (AU 2020), Randy mentioned that the goal of cleaning data is: 1. Fix things that will make your analysis algorithm choke, 2. Reduce Unwanted Variation, and 3. Eliminate Bias (where you can). I personally think that these points are very valid and in order to obtain our goal, these three methods and reason should be applied to data sets. A good example that Randy brought up is the lagest data generation/collection system. The Larger Hadron Collider collects so much data that it is not even possible to store in the raw form, and for that reason physicists spend countless hours trying to break it down and even just to decide which data to keep which data to drop. Sometimes it is just impossible to keep and have everything.

Counter example

However, Randy also mentioned that if we are doing an linguistic study? The examples that he mentioned is the amount of misspellings in the data, if we cleaned it all up, a linguistic study wouldn't be able to discover all the different variations of how people might have spelled a word. This brings up a good point that mentioned in class countless times and we are often reminded to practice. Reproducibility! It is important to make sure that even if we cleaned a data set for our own use, the original remains untouched. This allows the next person to take the same data set and manipulate it in ways that would be beneficial for their own goal.

Conclusions

Finally, my personal thoughts on this topic is that we should allow the data to speak as much as it could. Even if we need to collect and do our own work, we must respect the data set and allow reproducibility for the next person. The original data should remain untouched as much as possible.

AU, RANDY. 2020. "Data Cleaning IS Analysis, Not Grunt Work." https://counting.substack.com/p/data-cleaning-is-analysis-not-grunt.