On the philosophy of 'Clean Data'

October 2016

Abstract

As GA's instructors eloquently put it: "Good models cannot produce good predictions without good data." This paper describes how important it is to have clean data if you want to do data science.

1 What is data

According to the Wikipedia (`https://en.wikipedia.org/wiki/Data`), "Data is a set of values of qualitative or quantitative variables. An example of qualitative data would be an anthropologist's handwritten notes about her interviews with people of an Indigenous tribe."
Data is important if we want to gather information about the world to make informed decisions. For example, if we decide to eat a healthy diet and want to choose healthy foods, we need data to choose wich foods are healthier than others. If we decide to invest our saving for retirement, we need data to tell us which investments offer bigger returns given the risks that we are willing to take. If we want to drive to work, we need to know which roads to take, and if possible avoid the ones that have more traffic. All these tasks need data in one form or another.

2 Cleaning data

There are a few cleaning data tasks that we have done this week.

1. Fix data formats: sometimes data are stored as strings when they should be dates, for example. In this case, it is necessary to correctly interpret these values, for example using `.to_datetime()`. When there are missong values in columns that are supposed to be numbers, they will appear as strings and need to be converted.

2. Fill in missing values: this can be done simply by substituting missing strings with empty strings and missing numbers with zeros. However, sometimes we need to follow other approches, for example substitute missing numbers with the mean of the remaining numbers.

3. Correct erroneous values: sometimes values in a column are so far outside the expected values that they can be attributed to an error and must be corrected. For example, an age of 567 years, or a gender of 27. We can simply delete those values or try to find approximations for them, either in the same dataset or from other sources.

4. Standardise categories: sometimes, it's useful to have a pre-defined set of categories to use with our data, for example, a pre-defined set of music genres. However, if the genre is entered manually, we may find a few that are outside our categories (for example due to spelling mistakes) in this case, this needs to be corrected.

5. Eliminate superfluous rows/columns: sometimes, we have a huge dataset, but we only need some of the data to draw our conclusions. Or we may have rows full of missing values. In this case, it is useful to drop useless data and make our dataset smaller and more manageable.

3 Conclusion

There is much more to be said about this subject, but it does not fit in 500 words. These are just some of the most fundamental ideas.