AKINGBENI David Data Wrangling Efforts

(Explained to a toddler)

Data wrangling can be a daunting task but no (accurate) data analysis or visualization will happen without such noble efforts.

It is important that for such a daunting task that we have some basic and fundamental map and ideas on how to deal with it. Data Wrangling Efforts can be split into three main parts, which are Data Gathering, Data Assessments and Data Cleaning.

I will take each part to define the process and ultimately reference my python notebook as a wholesome example.

Data Gathering

This involves collecting coherent data from different sources and in different formats for the purpose of answering the question that you have originally posed. Some sources you can gather your data from include websites, data bases, hand given e.t.c. In the jupyter notebook attached alongside this document, we were given a file "twitter-archive-enhanced.csv", we had to programmatically download the "image-predictions.tsv" file using the request library and the final piece of additional information where we needed to collect the number of retweets and likes that a tweet has associated with it came from accessing the twitter API using the tweepy library.

Different formats of data have different way of handling them but in this notebook attached we worked with flat files (.csv – comma separated values and .tsv files – tab separated values) and JSON (Javascript Object Notation) formatted files which is characteristics of how applications (twitter in this case) store their information because of its flexible structure and lightweight performance. The JSON format is similar to a python dictionary and thus slicing it after it has been appropriately 'dumped' and 'loaded' becomes intuitive.

Data Assessment/Assessing

Retrieving data from different sources also means that you are not sure of the quality of the files you have with you. Every source has errors that are associated with it. It could be missing values, invalid values or data types of a column or inaccurate data entry, inconsistent values or tidiness issues as described by Hadley Wickham,

This involves putting down notes for every error or mistake spotted both visually (by intentionally and carefully trying to spot issues or unintentional scrolling past errors) and programmatically (using pandas methods such as info, describe, head, tail, value_counts, unique, nunique e.t.c.)

Data Cleaning

In the words of David, “This is where the magic occurs”. In this stage you begin to transform the assessments you have made into action by as simple framework of define, code and test.

In the define framework, you state the process you want to carry out to deal with the particular assessed issue. In the code framework, you turn each definition into codes to take care of the problem. And in the test framework which is similar to assessing the data to check if the issue has been appropriately dealt with.

In order of severity and importantly the order of handing assessment issues, missing values (incomplete values) take priority followed by tidiness issues, then validity issues, accuracy issues and finally inconsistent data values.

Note that the process is iterative and you can at any point go back and forth to any point