Wrangling Report

To be able to explore and analyze data properly and ensure there are no missing or wrong data, we have to wrangle the data. But before the wrangling process starts, we have to gather the data first, and it must be gathered from reliable sources.

Our wrangling process begins with assessing the data, visually and programmatically, mainly we look for low quality data, and messy data, look for data that requires correction from us and document it, assessing the data requires the usage of the python module pandas; it has many functions that can help with that. Examples for quality issues are: wrong data types, missing values, a value that makes no sense, even just capitalized names that should be lowercase. Tidiness issues like, more than an observation stored in one column, multiple columns for one variable, and different datasets that should be joined together.

After documenting our assessment, we begin with the final step, cleaning. Solving the issues requires sound programmatic knowledge, since they are solved programmatically, sometimes we have to think of an algorithm, or implement a regex statement even, the pandas library especially is incredibly useful.

Data wrangling is an iterative process, there might be issues that present themselves after cleaning, or new data might come in. It's not a one-time process, and we have to remain vigilant, since our entire analysis and models depend on this.

Gather

For our datasets, we had to gather them from different sources in different formats. Our first dataset was a csv file called `twitter-archive.csv`, and it's our main dataset, luckily reading it is pretty straightforward using pandas `read_csv` function.

The second dataset we got was from a url that was provided, using the well known requests module, we were able to retrieve the file and keep it in storage, and like before we used pandas `read_csv` to access it programmatically.

Our third dataset, was a JSON file extracted using the tweepy module to access the twitter api, here we needed to match the id of our twitter archive dataset with the tweets extracted from the api, and write the extracted data to a text file in JSON format. Lastly, using python tools to open files directly and the json module we were able to iterate over the text file line by line, and store it in a python list which is then stored in a pandas dataframe.

Assess

In our assessment, we first looked at our entire datasets, and their summary statistics and general information. From there we identified some errors and called on more specific functions that showed us more inconsistencies and redundancies.

Every step of the way we documented any quality or tidiness issues we found, so we have a clear mission, as to what needs to be done.

Clean

When cleaning the data, it's important that we see things clearly, which is why when going over our data issues, we first define the way to solve the issues, and being specific is very encouraged, and that's what we did.

We defined the problem and solved it programmatically as we saw was fit, after that we have to re-assess and make sure the matter is resolved.

We first started with the quality issues, where you have to access specific rows and columns and implement our correction. On to tidiness issues we condensed columns that shouldn't have been separated, and joined our three dataset together since they were all connected by the tweet it..

By the end we had clear tables that encouraged analysis and exploration.