Data Wrangling Efforts

Well, first off, I gather all the datasets needed for the project, including the Twitter archive dataset, the image predictions dataset, and the tweets dataset. I used different libraries and methods depending on the type of file and the environment it was in. Some datasets were easier to access, while others took a bit more effort to pull together.

After gathering the data, I moved on to the assessing phase. I started by displaying the data to get a good look at it and tried to understand the structure of each dataset. I looked at what every column meant, and most of it was pretty self-explanatory, to be honest. After that, I used my skills in Python, specially with the pandas library, to assess the dataset more thoroughly. I gathered information about statistics, checked for outliers, examined data types, and learned more about the overall quality of the dataset.

As I dug deeper, I figured out there was around 11 quality problems and 2 tidiness problems. The overall quality of the dataset wasn't that bad after all, but these issues still needed to be addressed. The problems included things like invalid names, missing values, HTML tags where they shouldn't be, unused columns that weren't really needed, and outliers that could mess with the results. On top of that, I also found some tidiness problems that needed cleaning, and I made sure to documented them as I went along.

When it came to the cleaning phase, I took all the issues one by one, not in any particular order. I just handled them based on importance, starting with the most important ones and working my way down. The cleaning phase was pretty easy and smooth. I tried to make my work as simple to understand as possible. After I finished cleaning all the issues, I documented everything. The dataset now looks almost perfect, after making it clean I can say that the master dataset is ready for analysis.