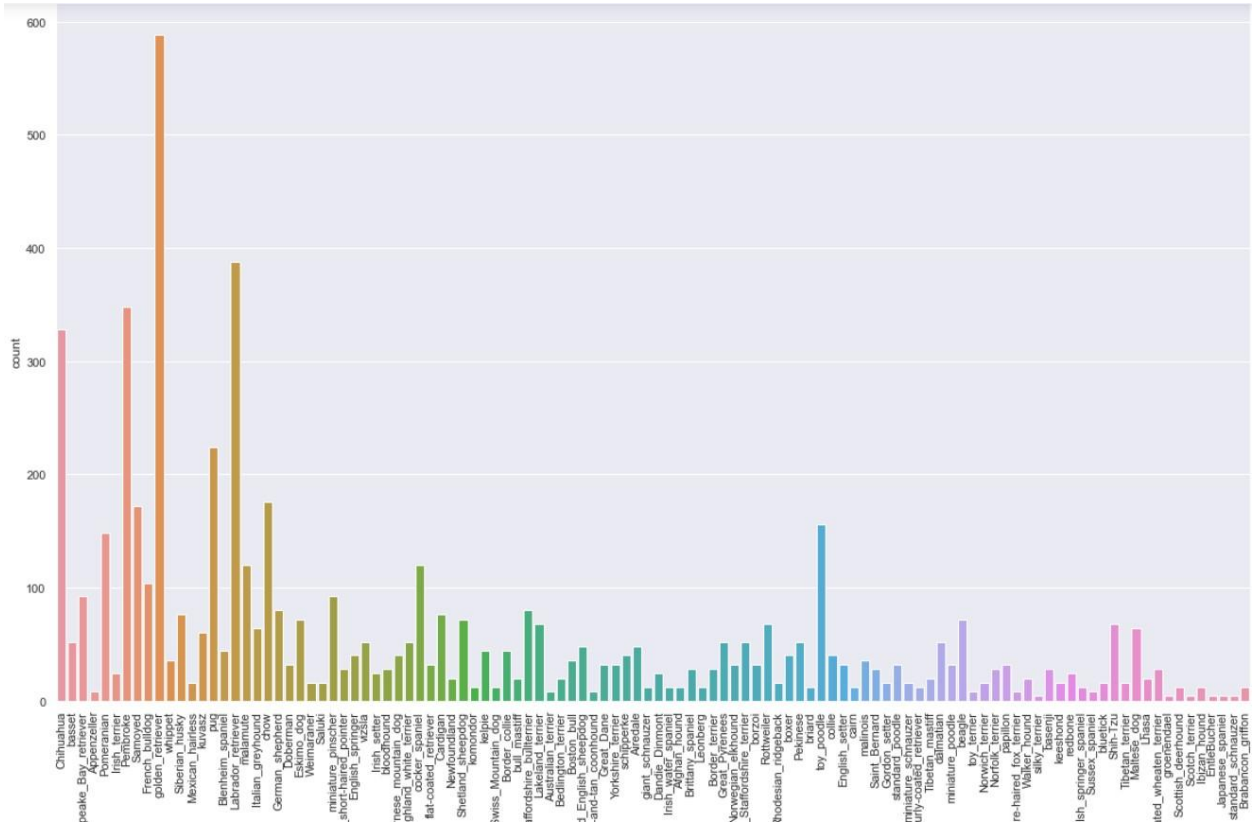
Data in real-world often comes messy and untidy, and sometimes the data we have at hand may not be sufficient for the project we are currently working on and we need to gather additional data from multiple sources, in various formats and assess and clean them which will improve on our analysis. This is known as Data wrangling.

In this project, data were gathered from three different sources namely,
Twitter archive
From a neural network that classified the breed of dogs
And from the Twitter API.

After gathering, assessin,g and cleaning the data quality issues in our datasets and storing the cleaned file into a CSV. It was now time for analyzing and visualizing the data.

Scrolling through the data frame I raised some questions which could be answered and visualized namely:

1) What p1(dog_breed) was predicted the most?
I used the seaborn library to visualize the dog breed that was predicted the most by the neural network, which signifies the most amount of dog breeds present in our dataset. From the analysis, I discovered that the Golden retriever is the most predicted dog. followed by Labrador retriever.



2) What stage was the most popular apart from 'None'?

To answer this question, I checked the value counts of all the stages of dogs in the stage column and it was discovered that apart from dog breeds whose stages wasn't documented, the most popular stage of dogs in our dataset is the pupper with 157 counts followed by doggo with 63.

3) What source did people use to upload the most?
Analyzing the sources people used to tweet and upload their good dogs, I discovered that twitter for iPhone was the most popular means!