Wrangle Report

Introduction

In this Wrangling Report I will be discussing my efforts in Gathering, Assessing, Cleaning, Analyzing, and Visualizing Data

Data Gathering

In this project I have gathered data from various sources and using different techniques, the main data was given by Udacity `twitter-archive-enhanced.csv` this data was downloaded manually, `image-predictions.tsv` was downloaded programmatically using requests library, lastly I queried Twitter API to get the retweets and likes counts and stored them in `tweet_json.txt`.

Assessing & Cleaning Data

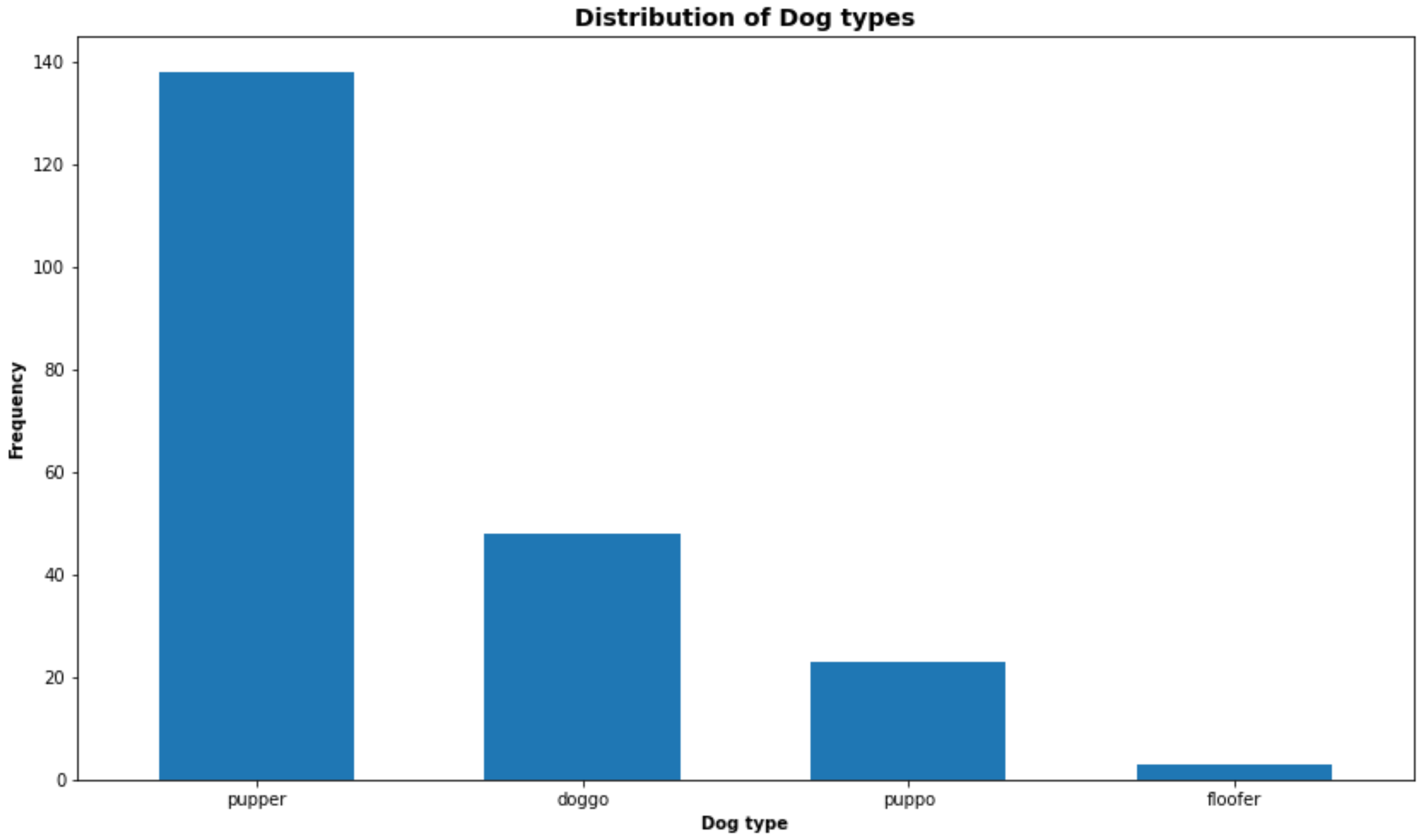
The Seconed step after gathering data is Assessing & Cleaning it, In this project I detected and documented eight (8) quality issues and two (2) tidiness issue, and I used both visual assessment and programmatic assessement to assess the data.

Issue	Type	Cleaning Actions
After inspecting Twitter API for retweets and likes counts, I discovered that some tweets in twitter_archive were	Quality	Drop rows that contain deleted tweets
There are rows in twitter_archive with rating denominator less than 10	Quality	Drop rows with denominator less than 10
There are rows in twitter_archive with numerator less than 10.	Quality	Drop rows with numerator less than 10
There are images in image_predictions with p1_dog equals False, which means that they are not dogs	Quality	Drop rows with p1_dog equals False
Duplicated images in image_predictions	Quality	Drop duplicated images rows in image_predictions_clean
Unclear column names in twitter_archive, such as ("text" > "tweet_txt", "name" > "dog_name")	Quality	Rename columns ("text" > "tweet_txt", "name" > "dog_name")
There are about 180 Retweets in twitter_archive and about 65 tweets that are replies, and the columns ("retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp","in_reply_to_status_id", "in_reply_to_user_id") are not helpful	Quality	Drop retweeted tweets in twitter_archive
"timestamp" column in twitter_archive is of type string, shall be datetime	Quality	Change "timestamp" column datatype to datetime
Dog type instead of (doggo, floofer, pupper, puppo)	Tidiness	Merge the 4 types of dogs in one column called 'dog_type'
df_twt contains retweet and like counts, to be merged with twitter_archive	Tidiness	Merge retweet and like counts from df_twt_clean and image predictions from image_predictions to twitter_archive_clean

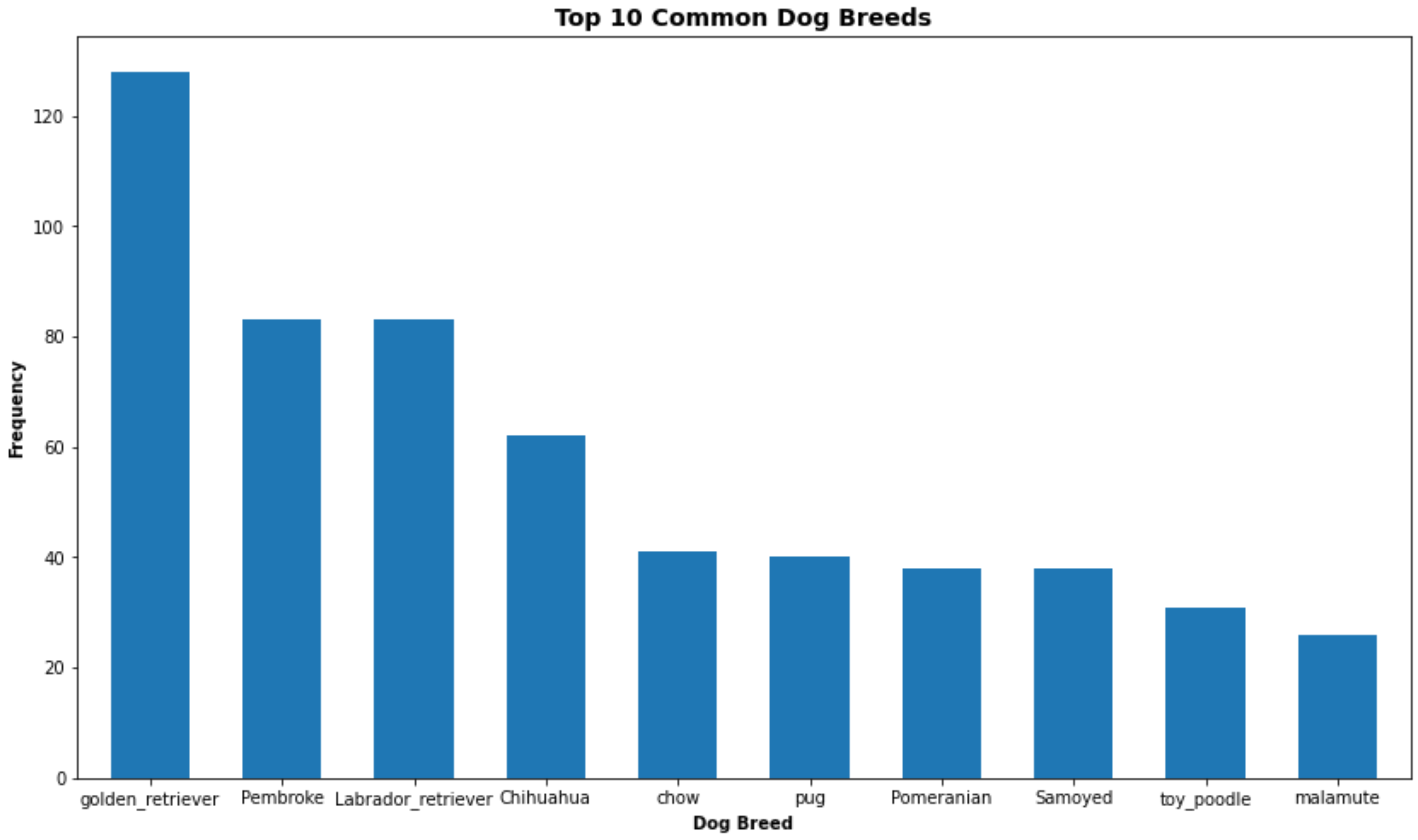
Analyzing & Visualizing Data

After assessing and cleaning our data, we can now explore our data and create amazing visualizations.

We can see below the distribution of dogs' types (stages) in our dataset, showing Pupper as the most common type followed by Doggo, Puppo, and lastly Floofer



The below visualization shows the top 10 common dog breeds in our dataset:



We can see from the above bar chart that the most common breed is Golden Retriever followed by Pembroke and Labrador Retriever

We also found that the dog with most likes is *Labrador retriever*