A PROJECT REPORT
ON

PREPARATION ET ANALYSE DES DONNEES

Submitted to
UDACITY

In Partial Fulfilment of the Requirement for the Award of

DATA ANALYST CERTIFICATE
ALX-UDACITY

BY

GoldCat

AUGUST-2022

Contents

1 Introduction

In this project called Data Preparation and Analysis, we had to collect data from various sources and various formats, evaluate their quality and order, and then clean them using Python and its libraries. The dataset we processed is the archive of tweets from the Twitter user dogrates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. After the wrangling process, we carried out our analysis through observations and visualisations to find some useful insights such as the most retweeted tweet, the most liked tweet, the most used source of tweets and the prediction algorithm confidence on the 3 pictures.

2 The most retweeted tweet

We choosed a bar chart to visualise the top most retweeted tweets because, it is easy to grap the insight and the difference between different values. From this observation as shown on the figure below, it can be seen hat the most retweeted tweet was the tweet with id 744234799360020481 having 70456 retweets. This tweet was about a bred doggo with a rating numerator of 13. Here is the tweet

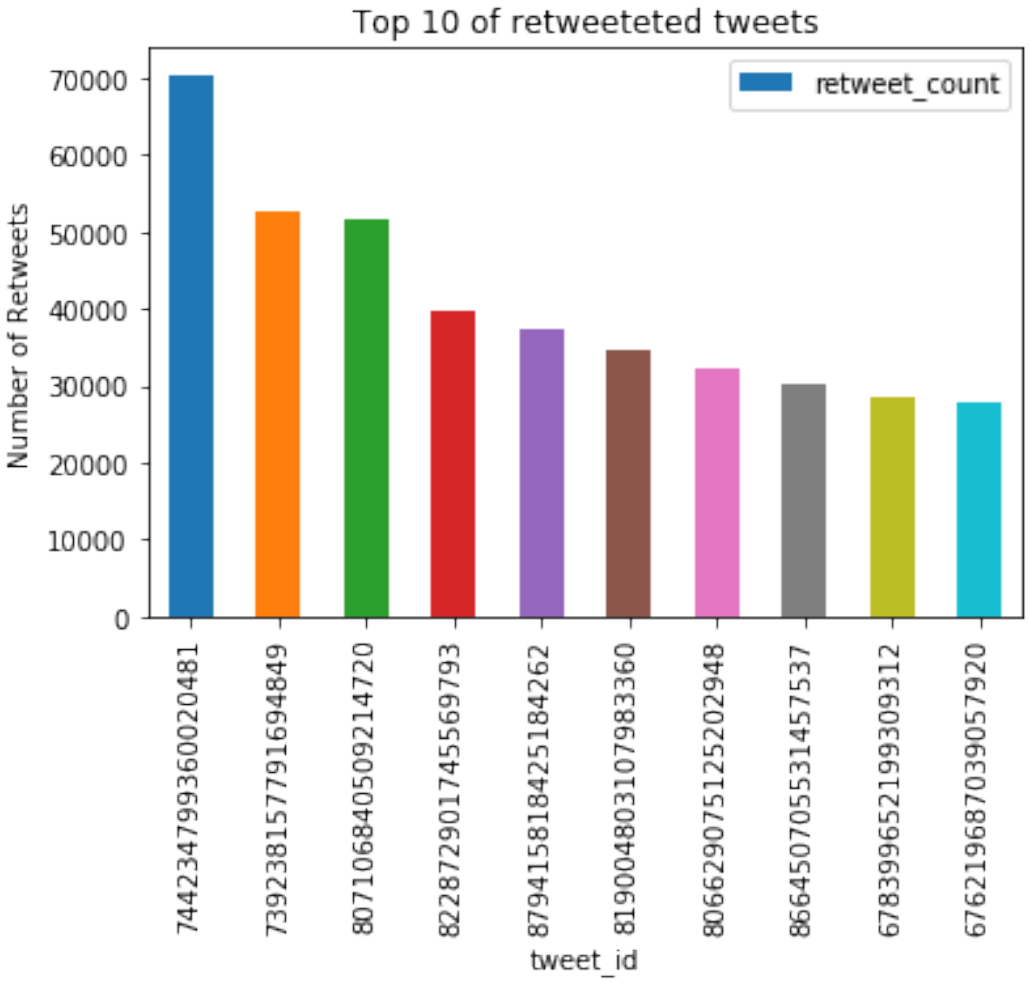

Figure 1: Top 10 of retweeteted tweets

3 The most liked tweet

For this, we also used a bar chart for the same reason. We find out that the most retweeted tweet was also the most liked tweet accumulating 144461 likes.

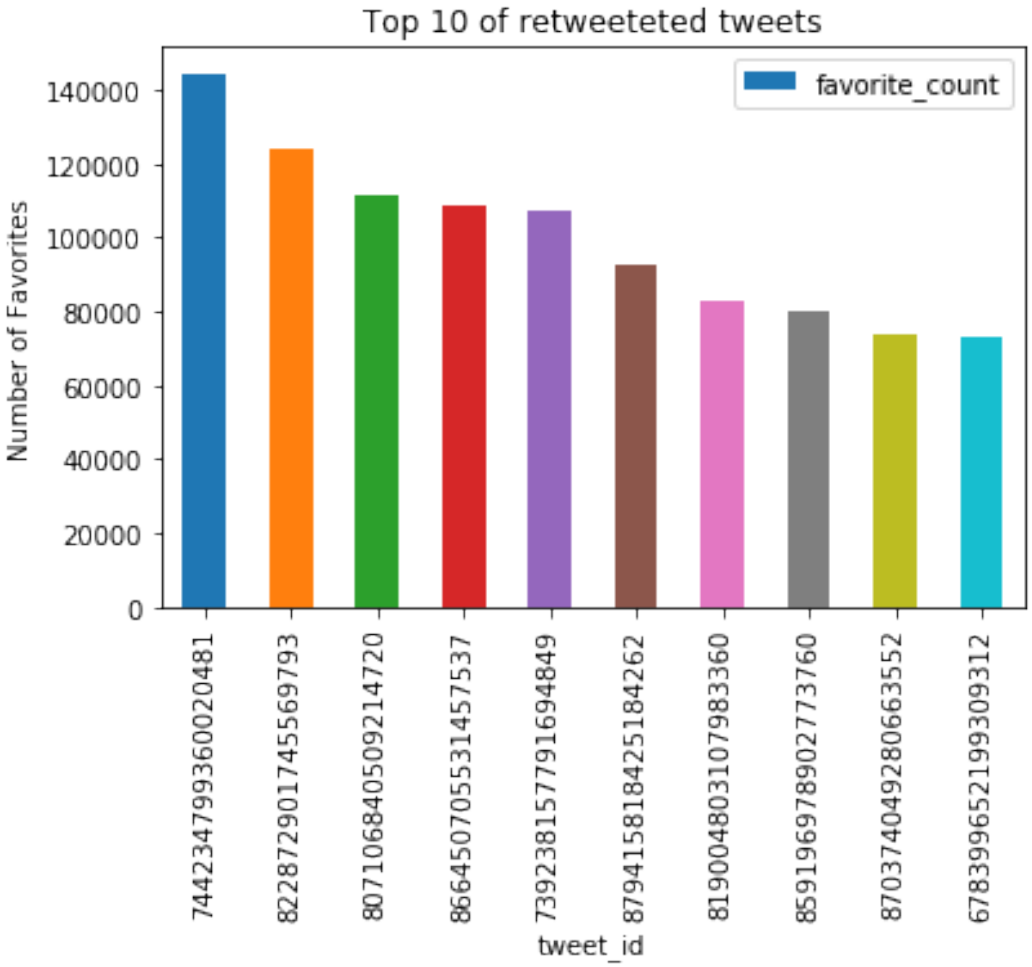

Figure 2: Top 10 of liked tweets

4 The most used source of tweets

We find out that Twitter for iPhone was the most used source for tweet accumulating 98 % of the total tweets.

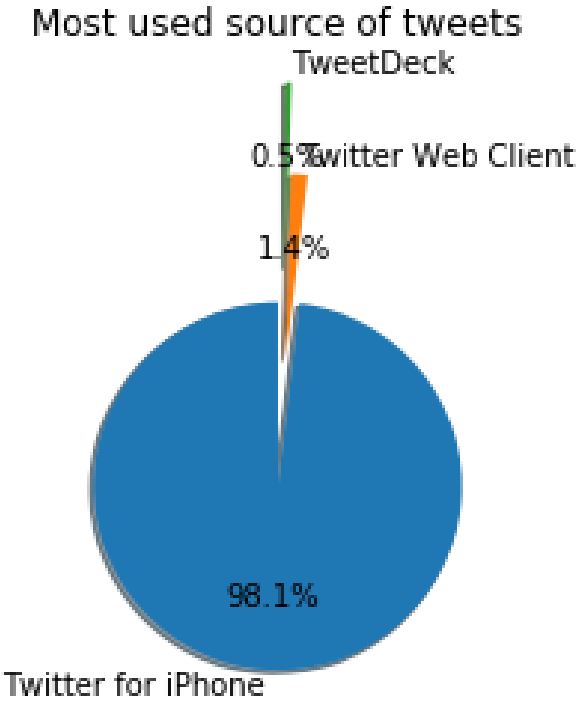

Figure 3: Most used source of tweets

5 Dog rating distribution

We find out that 23 % of dogs were rated with a 12, 21 % with 10 and 20 % with 11

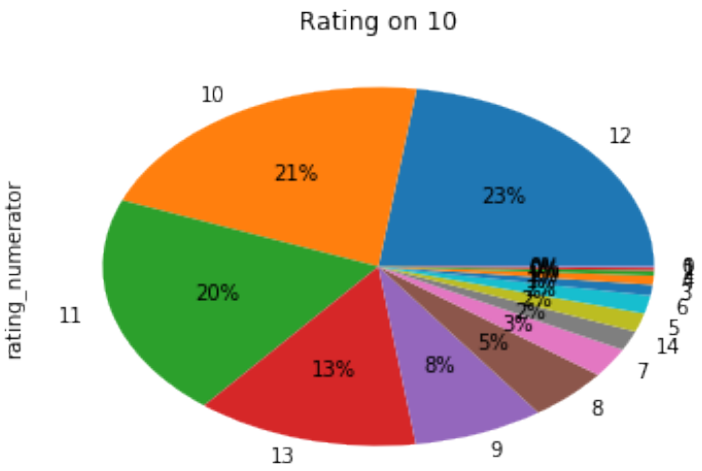

Figure 4: Rating distribution on 10

6 Conclusion

We were able to complete all of these steps in accordance with what we learned in the classroom in terms of data quality and tidiness. The most difficult step was querying the tweeter API using the Tweepy library. This took about 30 minutes to complete. In the future, we would like to explore other data collection techniques, such as web scraping.