ANALYSIS AND VISUALIZATION REPORT

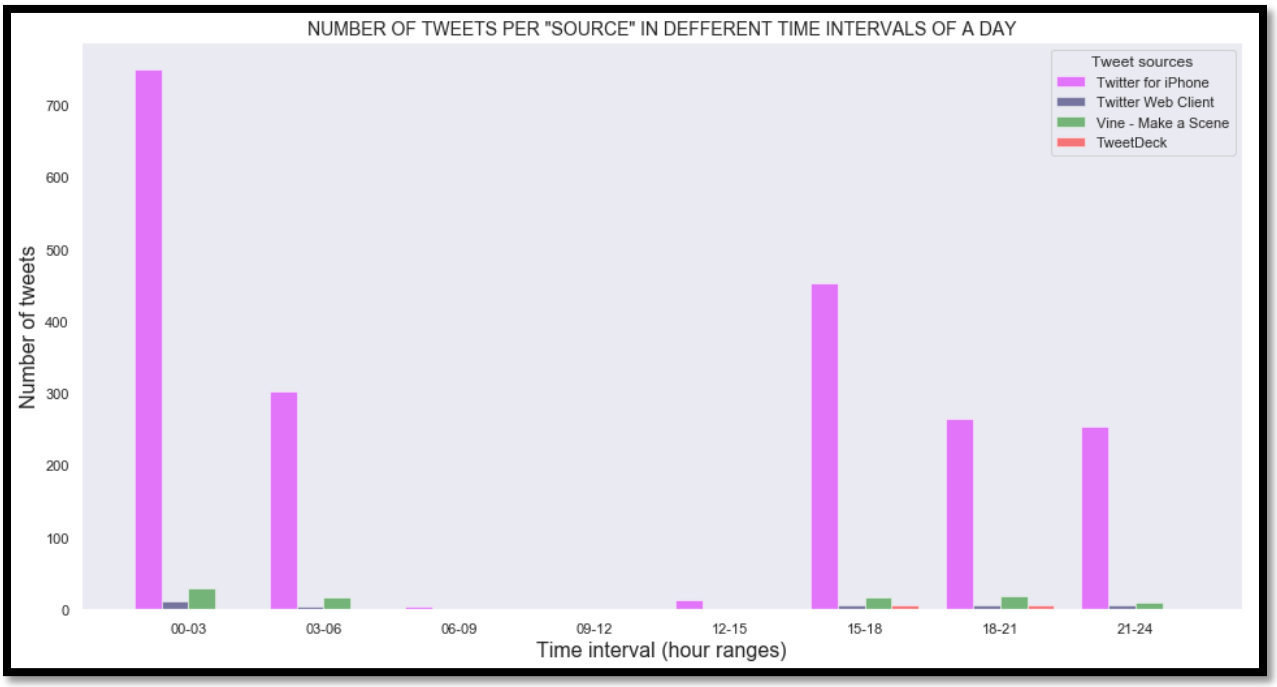
I. INTRODUCTION

After gathering, assessing and cleaning data as detailed in the ***wrangle_report*** pdf document submitted along with this one, we proceeded to analysis and visualization of three insights. For each insight, we started by setting a question, answer the question using the clean dataset and finally stated the insight that we got.

II. ANALYSIS AND VIDUALISATION

QUESTION # 1: What is the most used "source" of tweets and which part of the day different sources are most used?

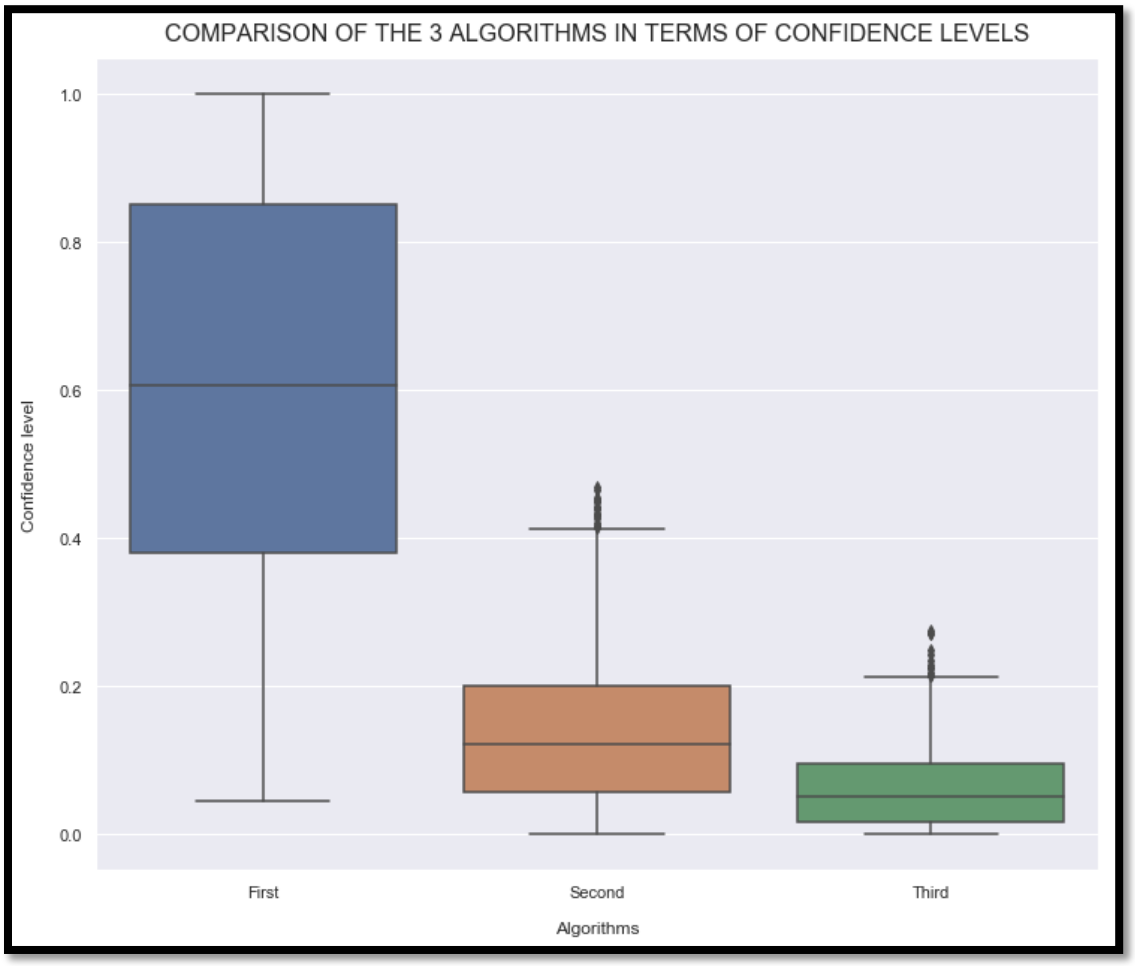
To answer this question we used the ***source*** and ***timestamp*** columns. From the source column we got all unique sources of tweets in the dataset. From the timestamp column, we extracted the hour field cut it into 8 ranges of 3 hours each and looked at the number of tweets made during each hours range as shown in figure below:



It looks like most of tweets (93%) are of '***Twitter for iPhone***' source. Also the above plot shows that most of tweets are made during night hours. It is noticeable that from 6 to 12 hours in the morning there are almost not tweets written. This suggests that maybe most of users are at their respective work places and have not time to tweet in the morning and the afternoon.

QUESTION # 2: Among the tree algorithms used to predict dog's images, which one is doing better than others (i.e. the algorithm that it is accurately predicting more images of dogs with higher level of confidence)?

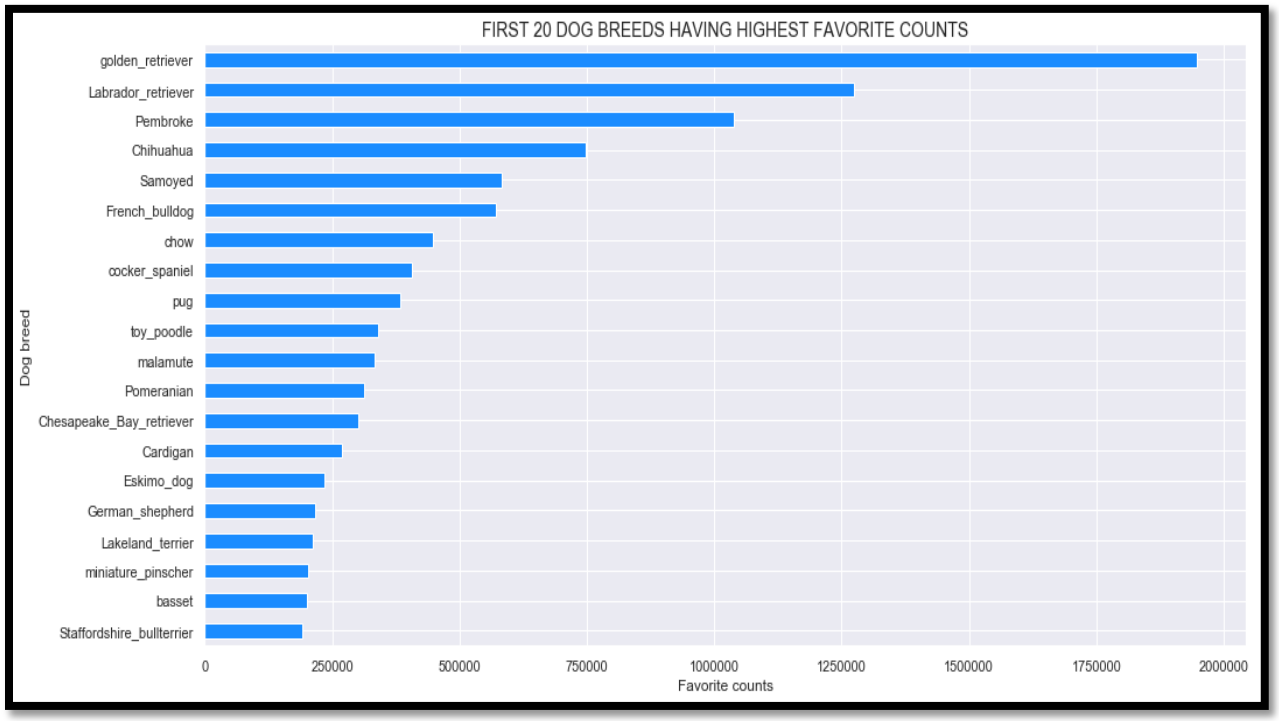
This question was answered using prediction related columns (predicted a dog, prediction confidence level and predicted breed). We counted how many dog's images were correctly predicted by each algorithm. Them we melt confidence level columns into ***algorithm*** and ***confidence_level*** columns. Using box plots, we compared performance of algorithms based on their respective confidence levels as shown below:



We see that the first, second and third algorithms correctly predicted almost equal number of dog images (1477, 1495 and 1446 images respectively). But when comparing their respective performance in terms of confidence level, we see that the first algorithm is doing better than the other two. The distribution confidence level for first algorithm varies much more than the distribution for the two other algorithms. There's a significant difference in the median confidence level for first algorithm form the two other algorithms.

QUESTION # 3: What is the most favorite breed of dogs? And what is the actual dog's picture most favored?

Columns ***dog_breed***, ***favorite_count***, ***jpg_url*** came in handy to answer this question. First, dog_breed was selected based on best algorithm. By 'best algorithm' we mean the algorithm that predicted correctly the image as a dog with highest confidence level. The first algorithm was given highest priority, meaning that it was first used to select the breed first. If it fails or the second does better, only then the second algorithm is used. The same process between second and third algorithms was applied. The favorite_count column help to sum up favorite counts for a whole breed as well as to known the most favored individual dog's image. The following figure shows most favored dog's breed:



The jpg_url column was used to look at the dog's image. The following is the most favored dog's picture in the dataset:

We notice that ***golden_retriever*** is the MOST favored breed of dogs in the dataset while ***japanese_spaniel*** is the LEAST favored breed. However the INDIVIDUAL MOST favored dog is of ***Lakeland_terrier*** breed while the INDIVIDUAL LEAST favored dog is of ***English_setter*** breed.