UDACITY

Project Report

By Shikhar Sharma

Data Analysis and Visualization

Libraries Used- Seaborn, matplotlib, requests, pandas, numpy, os, PIL, Byteslo

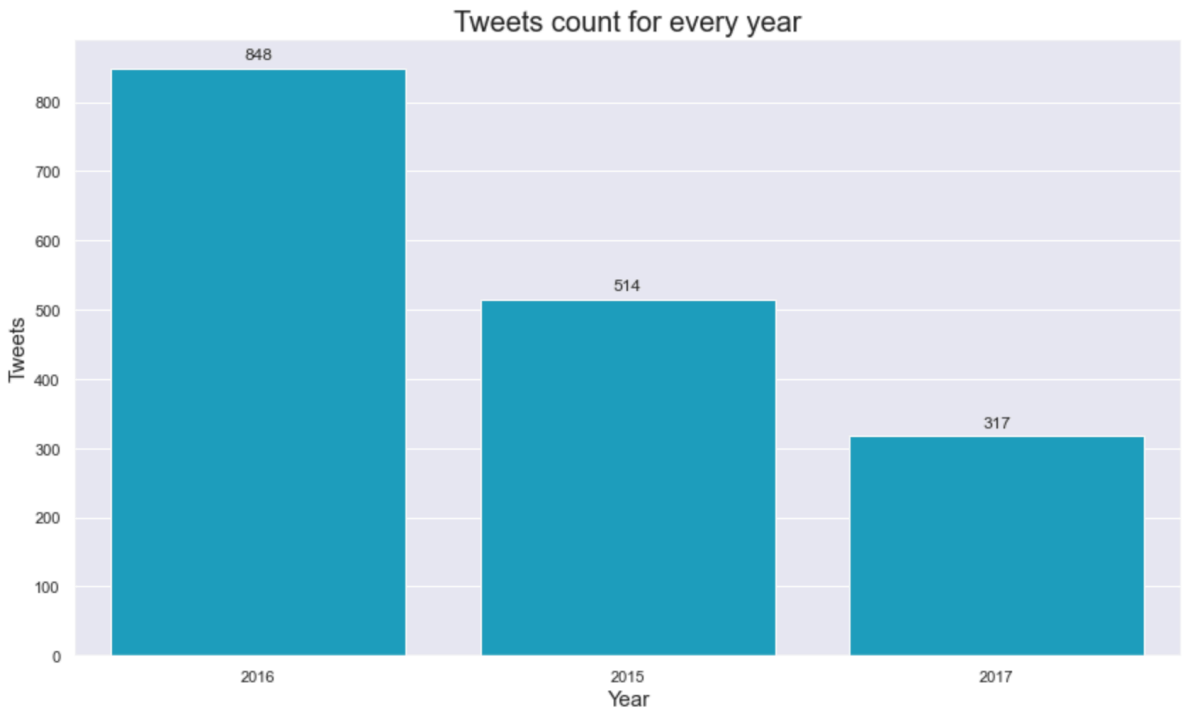
Tools Used- Jupyter Notebook.

The data analysis and visualization part was done on the cleaned data obtained from data wrangling in which I some data quality and data tidiness issues using three datasets which were based on tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The final cleaned dataset included one table master_dataset which was used for the data analysis and visualization part. So, my analysis included four insights which have been discussed below in the subsequent points:

1- In which year, month, day and date tweets were the highest and the lowest?

First, I plotted a bar plot which showed the total tweets count for year 2015, 2016 and 2017. Below is the image of following barplot.

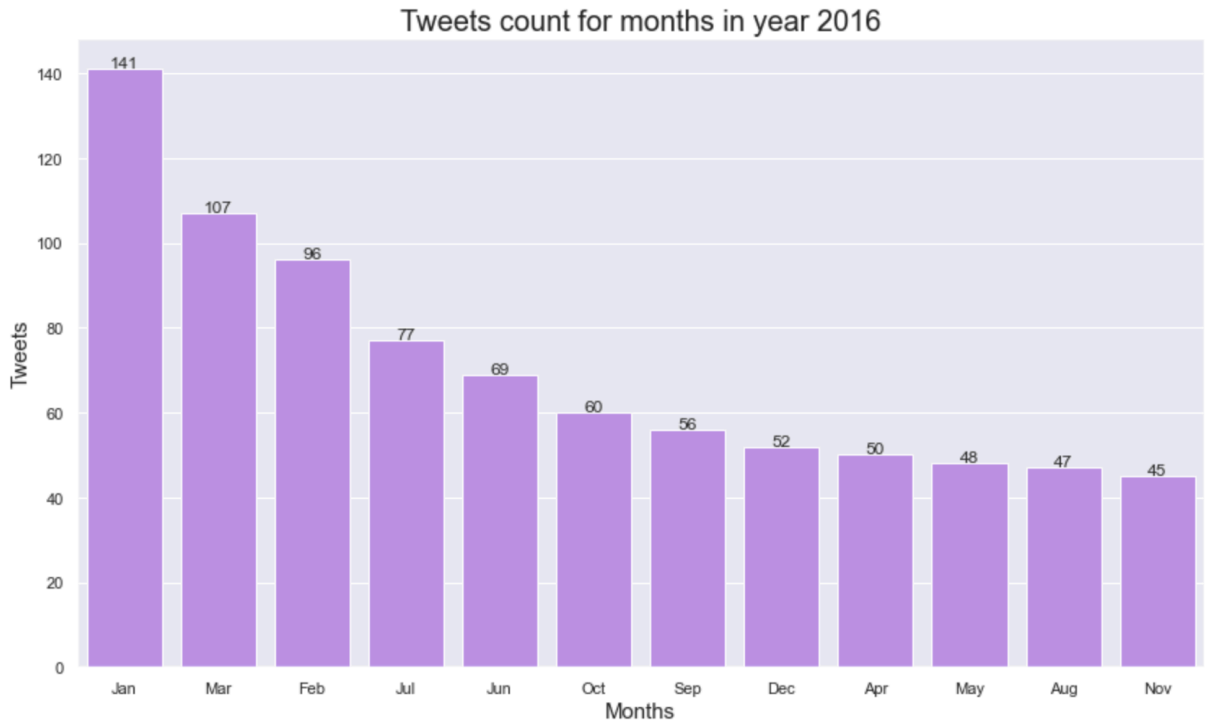
PLOT-1



The above plot displays that the total number of tweets were highest in 2016 and lowest in 2017.

Secondly, I plotted a barplot to show the month/s with highest and lowest tweet count in 2016. Below is the screenshot of the plot.

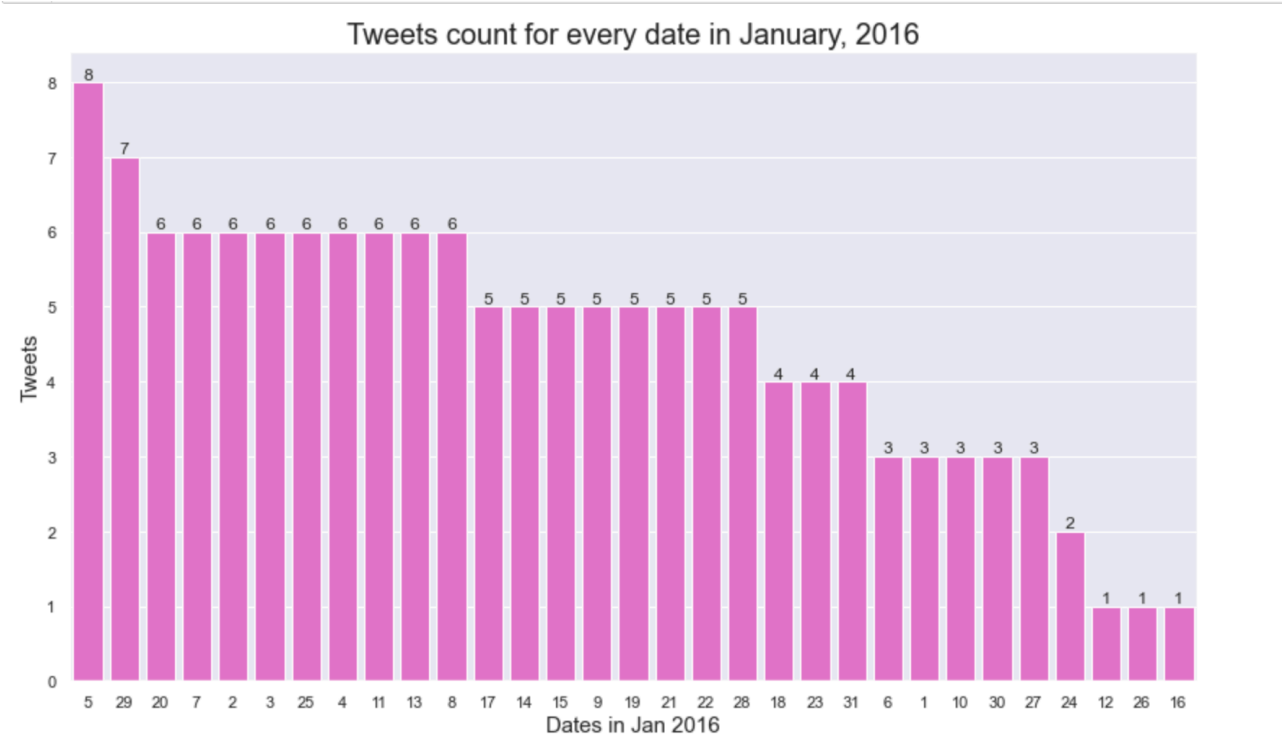
PLOT-2



From this I concluded that in January maximum tweets were made with count 141. Whereas, minimum number of tweets (45) were made in November

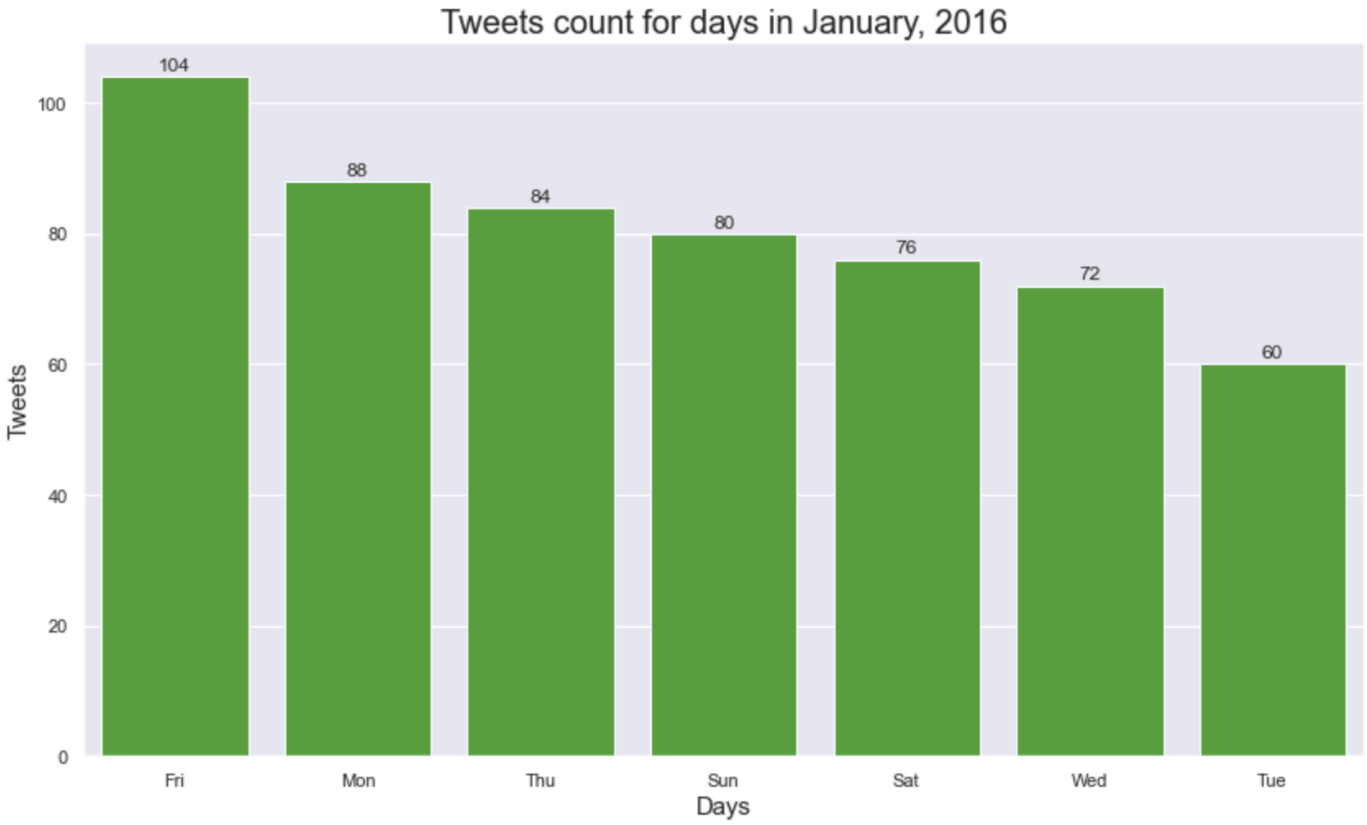
Third, I again used the same barplot to show the date and day on which maximum tweets were made in January, 2016. Below is the plot shown.

PLOT-3



We can see that on 5th of January maximum tweets were made with a count of 32 Whereas, lowest counts of tweets were on 16th January 2016.

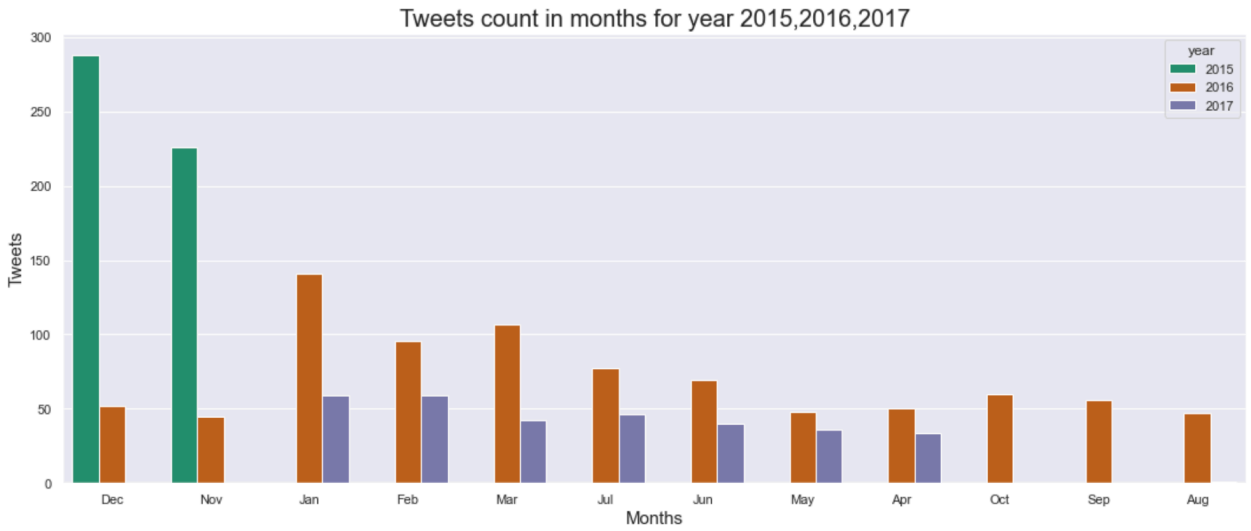
PLOT-4



From the above plot I concluded that in January 2016, maximum number of times tweets were made on ‘Friday’.

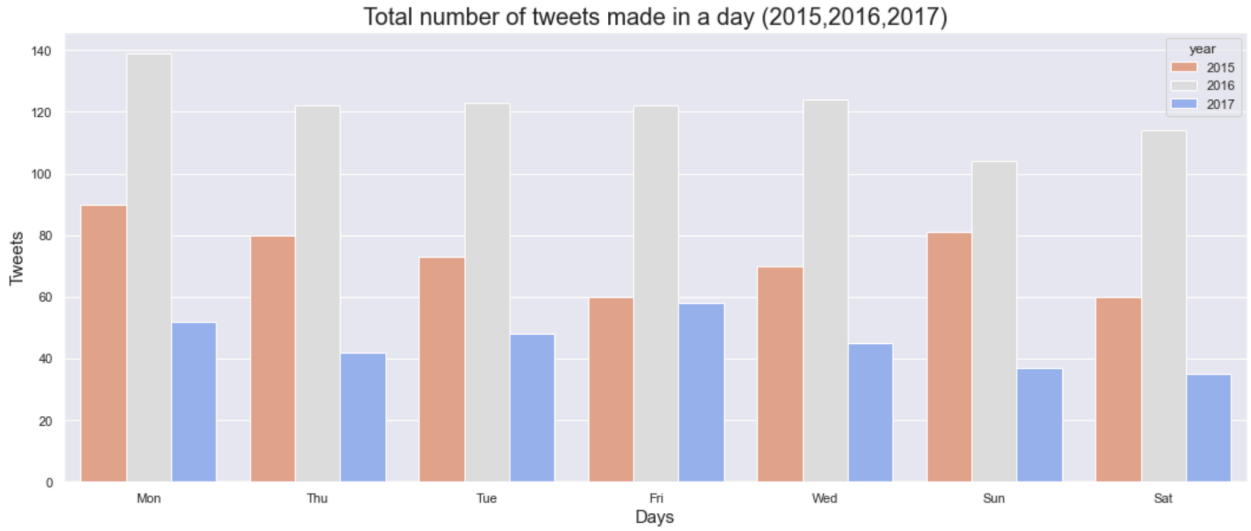
Last but not the least, I plotted two stacked bar plots and a line plot to show total number of tweets made in months and days and relation between favourite count with month for year 2015,2016, and 2017.Screenshot of the plot has been provided below.

PLOT-5



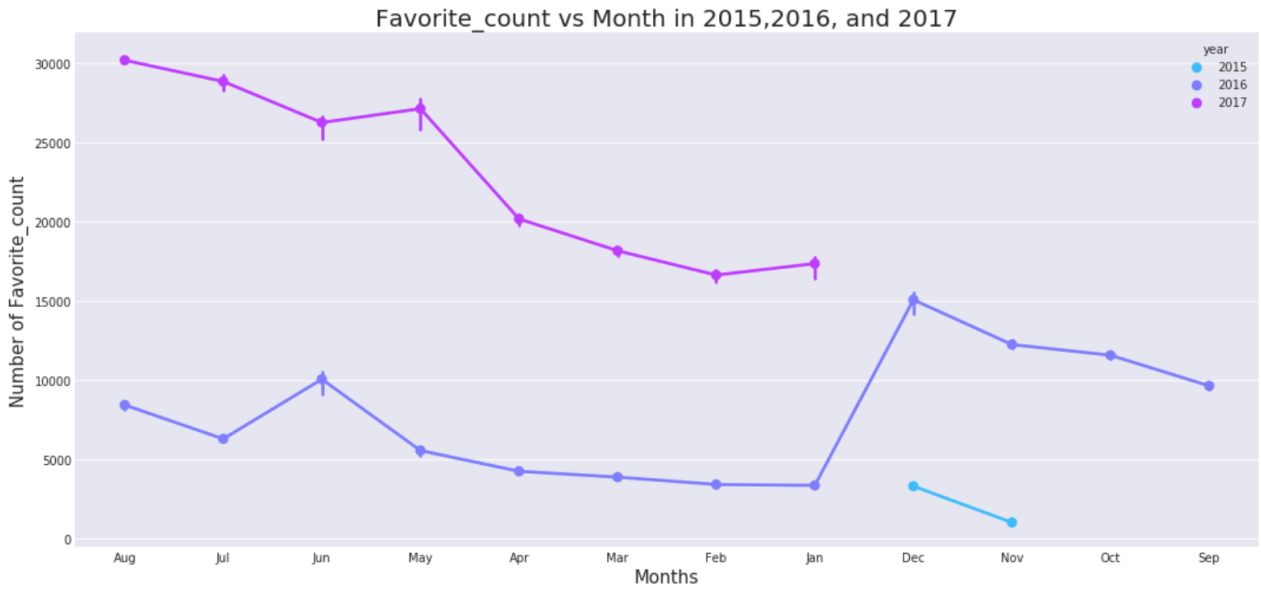
From the above plot, I observed that, in 2015 maximum tweets were made in November, and December. Also, November and December was the only month in two years 2015, 2016 which had maximum number of tweets made. For year 2017, ‘February’ had the highest counts for tweets.

PLOT-6



Out of all the tweets made in 2015 and 2016, 'Monday' has the longest bars for both 2015 and 2016. Whereas in 2017, maximum tweets were made on ‘Friday'.

PLOT-7

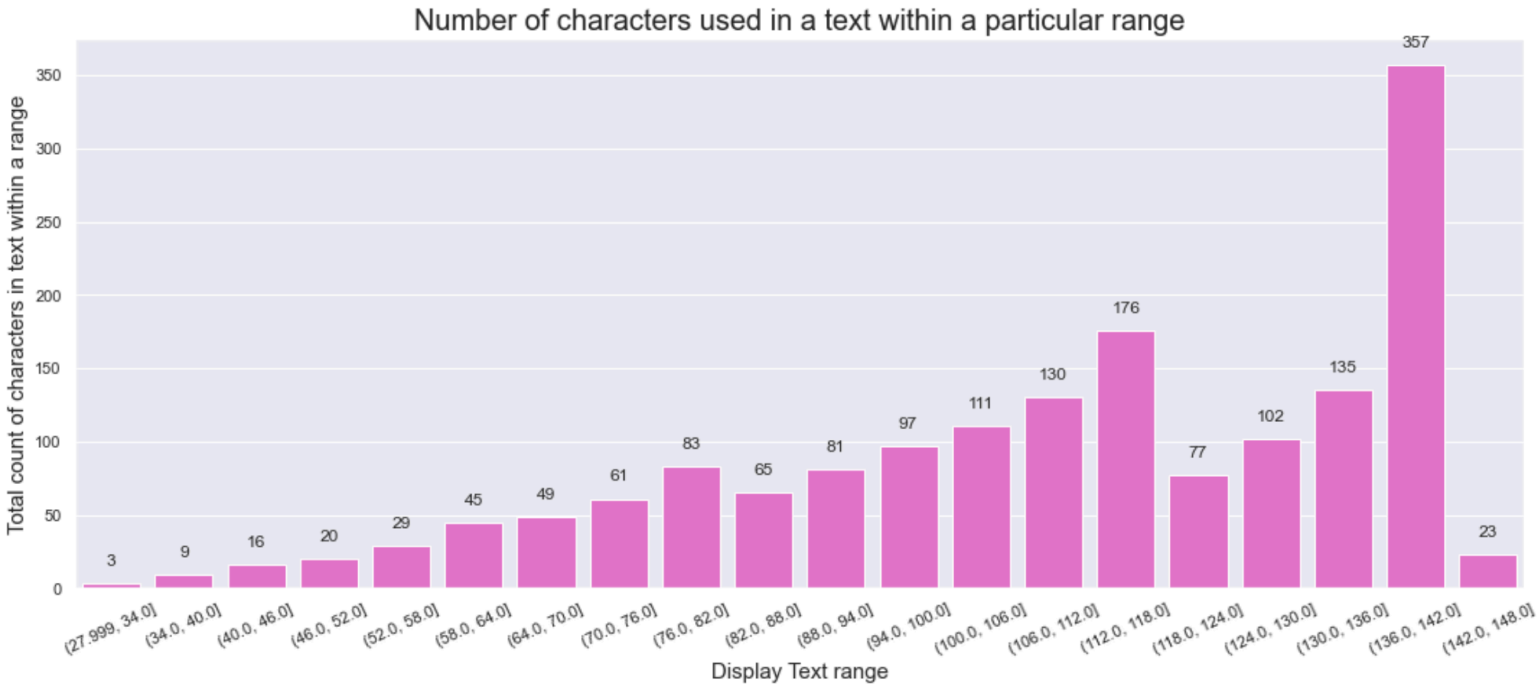


From the above plot I observed that for month 'august' favourite count was maximum in 2017 around 30000. Whereas in 2016, favourite count in month 'Aug' was somewhere around 8000. Lowest number of favourite count was in month 'Nov' for 2015 overall.

2- What was the average count of display_text_range within a particular interval?

In order to answer this question first I made intervals separated by 6 numbers in order to calculate and plot the total occurrence of display_text_range within that particular intervals. Below I have provided a screenshot of the barplot.

PLOT-8

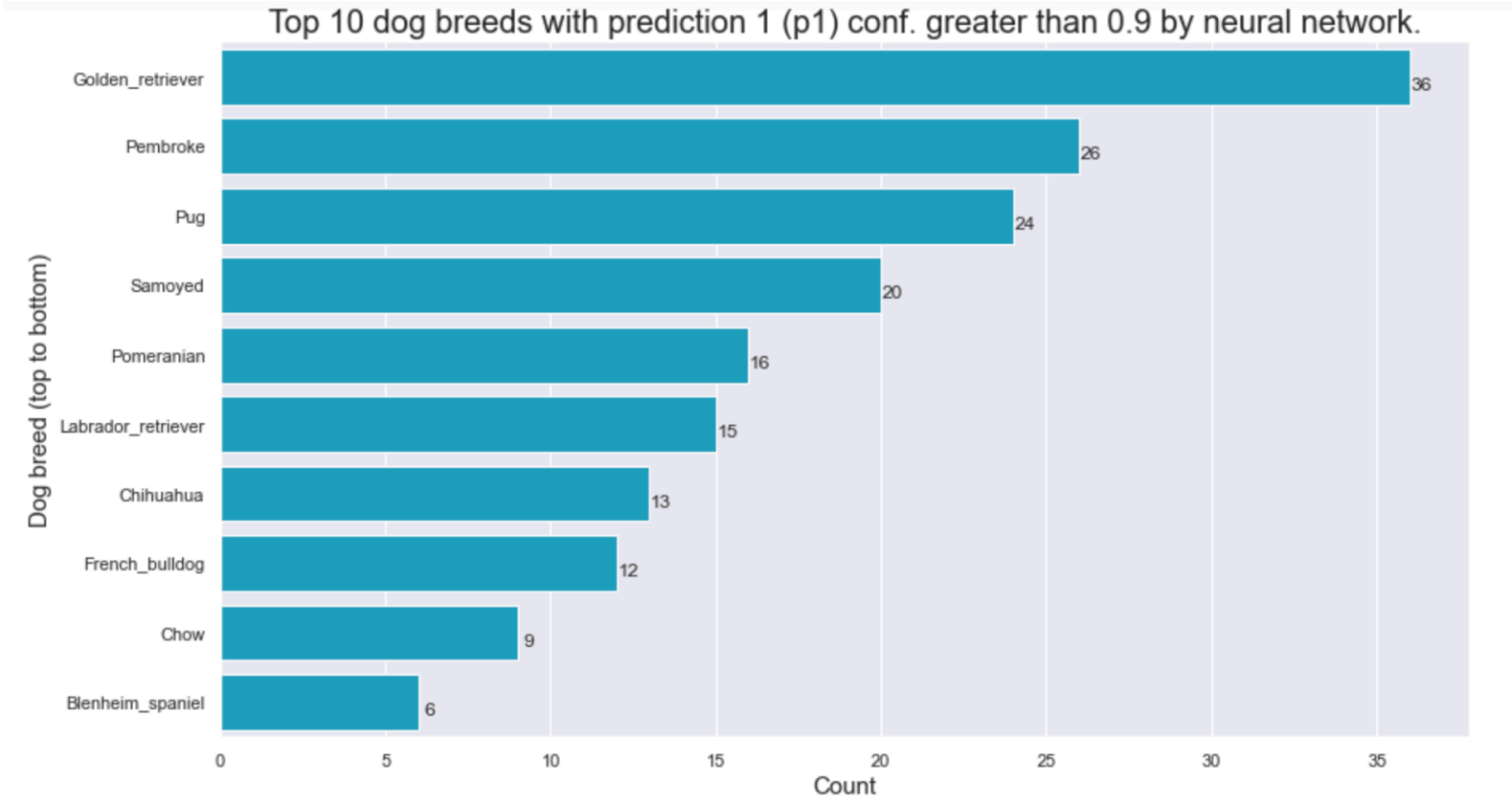


From the above plot, I observed that tweets with display text range within the interval of 136-142 were highest (357 tweets). I think people were engaging with ‘we rate dogs’ pretty well.

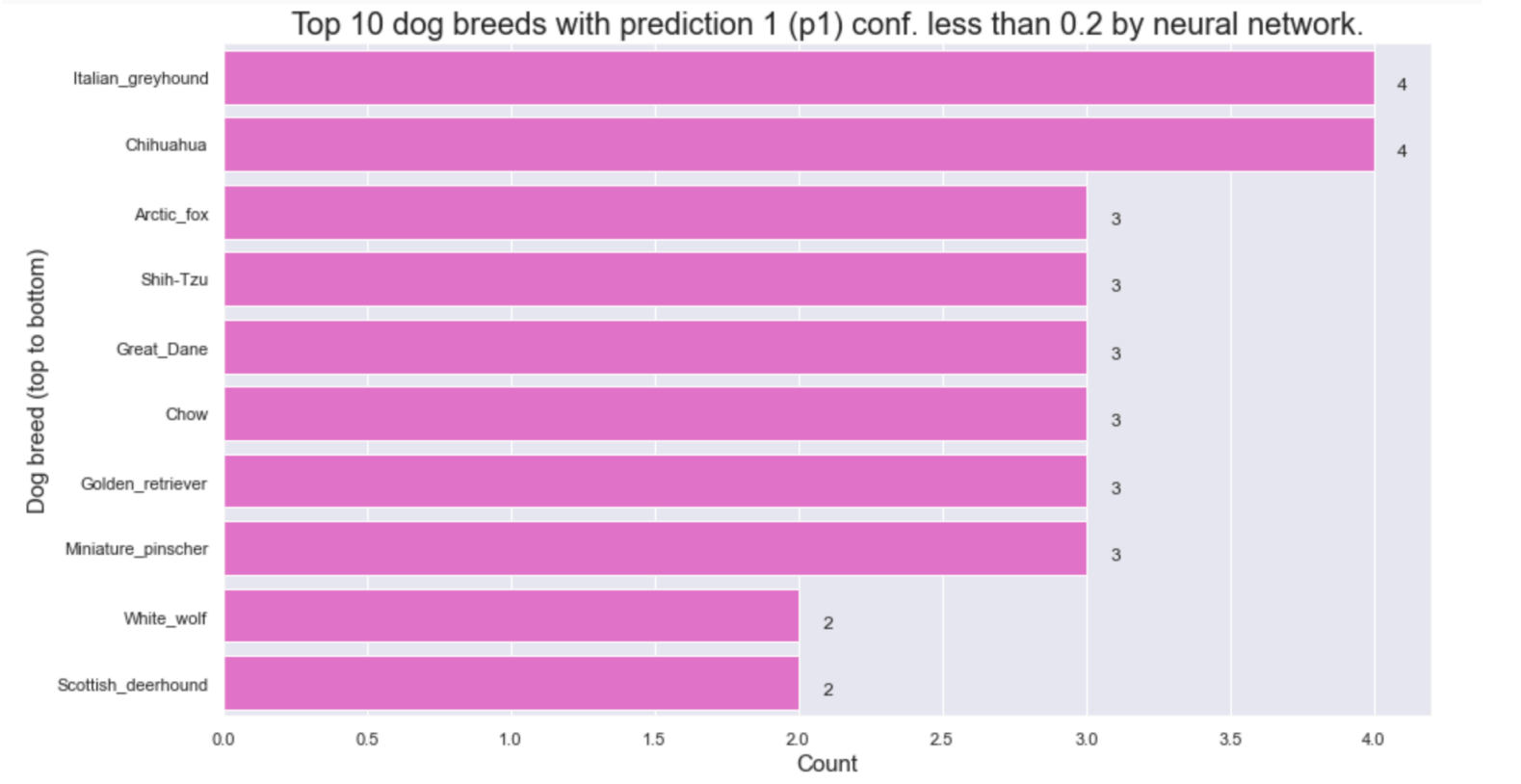
3- For which dog breed was the p1_conf greater than 0.9 and less than 0.2?

In this part of analysis, I first filtered predictions_dog dataset into two parts. First one include data for p1_conf greater than 0.9 and another one for p1_conf less that 0.2. Below are the screenshots of two plots.

PLOT-9



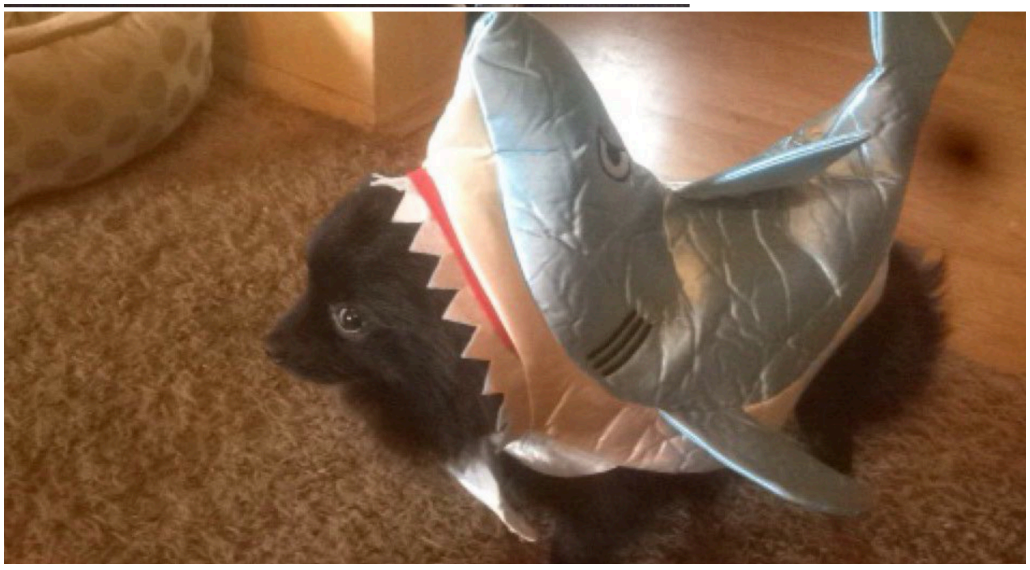
PLOT-10



I made several observations from the above two plots:
Golden retriever was the dog breed with p1_ conf greater than 0.9 and have 144 counts which is the highest. Also, for p1_conf less than 0.2, 'Italian greyhound' and 'Chihuahua' were the dog breeds with maximum counts i.e 4. Moreover, we can see that gloden retriever had a p1 coeff. less than 0.2.

4-What could be the reason behind predicting dog breeds with low and high p1 conf.?

This was my last insight which I made using cleaned datasets, twitter_merge and predictions_dog. So, for this part, I used images posted by people on We Rate Dogs page and used the predictions_dog dataset to filtered the images on the basis of p1_conf greater than 0.9 and less than 0.2. I used requests library to collect the data related to each image using image links in jpg_url feature and then used PIL library to read the images. After gathering image data and reading it, I displayed some images for both p1_conf > 0.9 and p1_conf < 0.2. The main difference which I found was that images with p1_conf > 0.9 were clear and had little to no noise whereas images with p1_conf < 0.2 had lots of noise and also in some of the images dogs were not visible clear enough for neural network to predict them as Dog. So, that's why for some images neural network predicted invalid objects like towel, shopping cart etc. Two images have been provide below having p1_conf < 0.2 and > 0.9.



Neural network predicted this image as 'ox' which is wrong.
(p1_conf < 0.2).

Neural network predicted this image as ‘Blenheim_spaniel’ which is correct.
(p1_conf > 0.9)

END.