Data wrangling_reports

Project Objectives

Step 1: Gathering Data

Step 2: Assessing Data

Step 3: Cleaning Data

Analysis and Visualization

Project objectives

The objectives of this work is to Perform data wrangling (gathering, assessing and cleaning) on provided data from the following sources twitter_archive_enhanced.csv., image_predictions.tsv, tweet_json.txt.

The data are expected to be Store, analyze, and visualize.

Step 1: Gathering Data

The Data gathered via downloading it from the Udacity classroom, my twitter developer account is yet to be approved, this make me use the tweet_json.txt data provided in the classroom.

Step 2: Assessing Data

The following columns will be elaborated, namely:

tweet_id (int64): unique identifier of a tweet
timestamp (datetime): UTC time when this Tweet was created
source (str): Device used to post the Tweet, like iPhone etc.
text (str): The actual UTF-8 text of the status update.
rating_numerator (int64) dog rating numerator.
rating_denominator (int64) dog rating denominator.
Name: dog's name that appears frequently
Stages of dogs growth: doggo, floofer, pupper, puppo are kind of a dog growth stages.

Step 3: Cleaning Data

The data was cleaned by using different coding to edit, remove and add necessary columns and rows, some of the cleaning done are highlighted below

tweetdata_cleans
From twitter_archive_enhanced.csv, some columns are removed which includes columns:
- that are a replay or a retweet
- in_reply_to_status_id
- in_reply_to_user_id

- retweeted_status_id
- retweeted_status_user_id not NaN
- retweeted_status_timestamp.
- expanded_urls
- img_num

The first letter in the name column is being capitalize.
For easier method to use dropna, name ‘None’ is replace by ‘NaN’ in the dataset
The dog developmental stages are merged together using the name dog_style while the following columns “doggo, floofer, pupper, puppo” are remove and change 'None' to NaN in the column.
The column timestamp datatype is change from object to datetime64[ns, UTC].

The column “source” which has html link for each platform used to tweet was renamed with platform only, removing the html link to avoid ambiguity in understanding its visual representation.

Analysis and Visualization

Most dogs names in the tweets

- Besides cases, where a letter 'a' and an article 'the' are provided, the most popular names are: Lucy, Charlie, Cooper, Oliver, Tucker, Penny, Winston, Lola, Sadie, Daisy Tolby, Bailey and Koda.

Most common dogs' stages in the tweets

- The common stage of the dog is pupper with 221 counts, followed by doggo 83 counts, puppo 23 counts and lastly flooter with 9counts

Prediction algorithm of images

- golden_retriever with 150 tweets has the highest prediction in the first prediction algorithm followed by Labrador_retriever with 100 prediction, while Labrador_retriever is the highest for both second and third prediction algorithm respectively.

Number of tweets

- There was high tweets of 18 at the beginning of the post in january 2016, the tweet gradually reduces 2 by may and increases sharply to 4 around July 2016 which later decreases to 2 by sept 2016. From september 2016 to july 2017, the tweets oscilate between 2 and 2.5 which shows many people are not tweeting the post again.

Sources of Tweet

- 93.7% of the tweet is from iPhone, followed by Vine, Twitter and Tweetdect with 4.3%, 1.5% and 0.5% repectively

Tweet rates

- There are 2097 dogs' ratings in the dataset. The worst rate is 0, the best is 177.6 (177.6%). The ratings are not normally distributed, their distribution is right-skewed (long-tailed). Most of the ratings are below 150%, only 6 (0.29%) of them are above 150%. With Atticus being the most rated of all. 75% of ratings are 120% or below and only 25% or less are below 100%. Most dogs are very well rated, with the mean rating of 117%

Dog image prediction with rating_numerators and rating_denominators

- The analysis shows that there are two dog image prediction tweet with numerators less than or equal to zero.also, further analysis shows that there is only one dog image prediction with rating_denominator less than or equal to zero.