Wrangle report

Introduction

Data are available but knowing how to collect and clean them is something else. In this work, we used three differents data collection methods in order to gather information together in one dataset and perform analysis. We therefore used python libraries like Pandas, numpy, matplotlib, request and json. The main objective of this work is to explore and clean the twitter archive dataset by using the define, code and test process and perform interesting visualizations. Our work is organized in five steps mainly data collection, evaluation, cleaning, storage and visualization. The last step is presented in another document called act_report.

Data collection

We worked on three differents datasets. The first dataset is a csv file collected using the reading file method of pandas. We use the Requests library to download the tweet image prediction corresponding to the second dataset, and we got the third dataset by reading the json file line by line.

Evaluation : Discovered Problems

We discover 8 quality problems and 2 tidness problems. There are :

1) There are unnecessary columns in the dataset : The dataset is too large with some informations that will not help us.
2) Among dog's rating some information are retweet. Therefore we are going to drop all the informations concerning the retweets.
3) Some columns names are confusing and do not give much information about the content (like p1, p2, p3…)
4) timestamp has wrong type : in our dataset, time stamp is object instead of datetime
5) There are wrong Dog's names : we observe dog's name like banana, orange
   There are Dog's without names : name column has 745 entries without dog's name
6) There are some duplicated images

7) There are so many missing values in the dataset. there are dog 1976 dogs without information on their stage
8) Columns names do not give much information about the data

Looking at the tidyness issues, we saw that :

9) In the twitter archive dataset, informations about dog's stage (flopper,doggo) are not well classify it enlarges the dataset.
10) The datasets should be merge with the twitter archive dataset.

Before starting with the cleaning step, we saved the three original datasets as copy_archive for the WeRateDogs Twitter archive data, copy_predict for the tweet image prediction and copy_api for the api dataset. We merge them together and save as

Data Cleaning

We started the clean process by merging the three dataset and saving a copy.

- Cleaning 1 : consist of Dropping retweets and unnecessary informations. To reach our goal, we first looked at the dimension of the dataset, then we checked the columns list and we finally dropped unnecessary columns ;
- Cleaning 2 : deleting retweets.
- Cleaning 3 consist of changing timestamp type from Object to datetime ;
- Cleaning 4 : we have replaced all the wrong dog names into nan
- Cleaning 5 : we identified all the duplicated images, we kept the most recent image and drop the rest.
- Cleaning 6 : we created one new column name 'phase' and store the information about dog classification (doggo, floofer,puppo,pupper) within, then we deleted the four columns in order to reduce the dimension of the dataset.
- Cleaning 7 : dilling with missing values. We checked the missing value of our dataset and observe that 1653 Dog's stage were missing, we recorded them as 'undefined' and dropped all the missing values in the dataset.
- Cleaning 8 : we created a new column named 'phase', classified all dog's stage in it and drop the four other column(doggo,floofer, puppo,pupper)
- Cleaning 9 : we rename some columns of the dataframe
- Cleaning 10 : we reordered all the columns in the data set

Conclusion

The twitter archive datasets were were downloaded. We merge the three datsets together, we identified more than 8 quality issues and 2 tidness issues. We cleaned up the dataset by removing the nan value, changing column type, we rearrange and reorganize the dataset and save it as ‘twitter_archive_master.csv’. We also observed that the first prediction is more accurate than the others. Looking at this results, can we say that the majority of dogs participating in the challenge are golden retriever? The first and second prediction show that 75% of the dog are either Labrador or golden retriever. is it not because these are the most appreciated breed dogs in the challenges?