Report

On Wrangle and analyze data

The steps that I have follow in this project are :

Gathering

Assessing

Cleaning

Sorting and visualizing

And here I will talk about the wrangling effort.

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

I have Gathered the data from three files :

Twitter archive_enhanced.csv

From the link that i found it in Udacity

From Tweet-json (which I take it without an account developer).

Then I have assessed the data and clean it and visualized some graphs about the cleaned data.

The quality issues and tidiness that I have figured out are the following:

Quality

1-Twitter archive

- Some names seems not appropriate such as (a , just , etc..) : so we need to replace it with Nan.

- tweet_id is int it should be string : convert it with astype(str).

- timestamp should be in datatime type : convert it to datetime using pandas library.

- doggo floofer pupper puppo retweeted_status_id in_reply_to_status_id in_reply_to_user_id retweeted_status_id retweeted_status_user_id retweeted_status_timestamp are columns that not needed in my analysis : so delete it using drop function.

- name has None instead of Nan thats why its not appeared in arch.info() that name has non-value : So replace it using Numpy library (np.nan).

- rating_denominator has values other than 10 i think it sould be 10 for all : so filter it and take just the ratings that equal to 10.

- doggo floofer pupper puppo has None for missing we should replace it with Nan : Convert the None with Nan using np.nan

2- Image-prediction

- tweet_id should be in string datatype : convert it to string using astype(str).

- Img_num we can delete it its not necessary : delete it using drop.

3- tweet_json

- id should be a string : convert it to string using astype(str).

- the name of the id is different compared to pred and arch so we should change it : Change the name of the column using rename

- most of the columns should be deleted cuz i do not need them : drop it.

Tidiness

1. all the three tables (arch,pred,tw-json) should be in one table
2. doggo floofer pupper puppo should be in one column (its categorical)
3. some columns in pred should be just about breed and confidence such as p1 p2 p3 p1_conf...

After cleaning this issues I did some visulas and insights about the cleaned data .