We Rate Dogs

Report1

Introduction

In this project we will go through data wrangling process on gathering , assessing , and cleaning the data

The wrangling processes

We have a three steps in data wrangling process:

1-Gathering

2-Assessing

3-Cleaning

Step 1: Gathering:

In this process we will try to gather the data from multiple sources and in this project, we gather three dataset

1-Archive twitter : I download it the regular way and read it straight for it

2-Image predication: I download programmatically using the **Request** library

3-Tweet json: I used the file provided from Udacity because the twitter took to long to replay to me

Step 2: Assessing

Quality issue:

1-Time stamp is object instead of timestamp

2-Tweet_id is int instead of string in the three datasets

3-missing data in expanded_urls

4-some of the names is just typed wrong like 'a' or 'none' or 'this'

5-dosen't benefit our analysis like (,in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id ,etc…)

6-worng data in the rating numerator like (id tweet is 832215909146226688,..) the rating is 75 but the actual rating is 9.75

7-the rating numerator is int instead of float

8-The tweet_id (832215909146226688 and 786709082849828864) have duplicate value for the same dog

9- the names of column not clear EX(img_num,p1,p1_conf,p1_dog,p2,p2_conf,p2_dog,p3,p3_conf,p3_dog)

10- the p1,p2,p3 should start with the first letter capital

11- "jpg_ur" information that don't benefit any thing to our analysis, so we don't need it

Tidiness issue:

1-the doggo, floofer ,pupper and puppo should be in one columns

2-the three tables need to be merged

Step 3 : Cleaning

In this step after we gathered and assess the datasets, we will try to solve these problems

Quality issues:

1-convert the Tweet_id column data type from int to a string using .**astype** method

2-convert the timestamp from string to a timestamp by using **.to_datetime** method by pandas

3-the expanded url is missing and we need to check the data, after checking the missing URLs we found that most of these url either a retweet tweet or a replay to some one tweet so we will use the **.drop** method

4-change the odd name we found to nan like (none , a , this) by replacing them and using **.replace** method

5-dosen't benefit our analysis ex(in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id ,retweeted_status_user_id and retweeted_status_timestamp) so we will use the **.drop**

6- wrong rating numerator ex: for the dog Logan(id tweet is 832215909146226688) the rating is 75 but the actual rating is 9.75 so we will fix the rating numerator

7-convert the rating_numerator column data type from int to a float using **.astype** method

Continue..

8-The tweet_id (832215909146226688 and 786709082849828864) have duplicate valuesame dog
we have to tweet with a duplicate value for the same dog "logan" so we will drop the (tweet_id=786709082849828864) i choosed this id because the other id we just fixed the rating numerator value for them
9-the names of column not clear EX(img_num,p1,p1_conf,p1_dog,p2,p2_conf,p2_dog,p3,p3_conf,p3_dog)
we will change the names of the column in image prediction (img_num,p1,p1_conf,p1_dog,p2,p2_conf,p2_dog,p3,p3_conf,p3_dog) to be clearer (readable) by using the method **.rename** method
10-the (p1,p2,p3) should be start with a capital letter by using **.str.title**
11-jpg_ur" information that don't benefit any thing to our analysis, so we don't need it
the jpg_url is not necessary in our analysis so we will get rid of it by using the **.drop** method
Tidiness:
1-the last four columns should be in one column by so we will replacing the none to empty and creating a dog_stage andmerging the four column and after adding them we will use the drop 4 coulmns
2-we will merge the three tables into one on tweet_id by using the '.merge' method