Wrangle report

This report briefly describes the data wrangling effort exerted in the project.

The data sets we working with :-

1-Archive

Which contain

-tweets id -timestamp -text(rate &dog name)

2-image prediction

It use neural network to predict Kinds of dogs

3- twitter api

To gets data about rates people's dogs with comments about the dog .these ratings always greater than 10 like 15/10 , 12/10

Now we begin to work on the project , firstly we put in our mind that

Wrangling process divides into three steps

1-Gathering Data 2- Assessing Data 3- cleaning data

The way we work on the project :

1-Loading Libraries

Importing libraries which we will use like

Pandas – Numpy – Requests – Matplotlip – Tweepy - Json

2-Gathering Data

We gather data from our tree data set resources

1. Twitter Archive:
2. Image Predictions:
3. Twitter API:

3-Assessing Data

Before cleaning, it is essential to assess data to inspect what to clean. In this process, two issues are concerned:

- data quality issue
- data tidiness issue

The assessment can be done visually and programmatically.

First: Assessing Data Archive

Second**:** Assessing Image Predictions

Third**:** Assessing Twitter API

We here discovered some problem in Tidiness and quality we work on cleaning them

Tidiness:

- Most tweets don't specify the dog_stage.
- All data is related but separated into 3 dataframes

Quality:

1. Timestamp column has dates in string form.
2. Rating_numerator & Rating_denominator columns should be float.
3. Row 313 has 0 denominator
4. Not all tweets contain photos.(2075 entries of 2356)
5. underscores are used in many names in columns p1,p2,p3 instead of spaces.
6. Not all tweets didn't include the dog's name correctly. ex(rows:570,2065,310 name = None & row:759 name = an)
7. There are 78 reply tweets
8. Not all tweets start with uppercase letters.
9. Tweet_id column has ids in int64 form which we don't need to have this type because there are no mathematical operations on it.

4-Cleaning Data

First. Copy Data frames:

Second. Clean Tidiness:

1. Most tweets don't specify the dog_stage.

Define:

Merge 4 columns to 1 columns Called Stage

2. All data is related but separated into 3 dataframes

Define:

Merge dataframes into 1 dataframe based on Tweet_id

Third. Clean Quality:

2. Timestamp column has dates in string form.

Define:

Convert invalid datatype of timestamp column to datatime

3. Rating_numerator & Rating_denominator columns should be float.

Define:

Convert invalid datatype of Rating_numerator & Rating_denominator columns that should be float

4. Row 313 has 0 denominator

Define:

Manual fix entry 313 has rating_denominator 0 replace to 10

5. Not all tweets contain photos.(2075 entries of 2356)

Define:

Drop all tweets that doesn't contain pictures of dogs

6. underscores are used in many names in columns p1,p2,p3 instead of spaces.

Define:
replace names that contain underscores to spaces

7. Not all tweets didn't include the dog's name correctly. ex(rows:570,2065,310 name = None & row:759 name = an)

Define:
Replace invalid names to none value

8. There are 78 reply tweets

Define:
Drop all rows that are replies, those that have non-null values in these columns: in_reply_to_status_id and in_reply_to_user_id.

9. Not all tweets start with uppercase letters.

Define:
Convert lowercase letter to uppercase

10. Tweet_id column has ids in int64 form which we don't need to have this type because there are no mathematical operations on it.

Define:
Convert tweet_id datatype to string

5-Store Data

We here stored the cleaned data in a file called ('`Data_Cleaned`.csv')

6-Data Visualization

We here use Matplotlib to visualize our data to get output and gain info as we will know in the act report