Wrangling and Analyze Data

Data Wrangling is the process of transforming and mapping data from raw data form into information data. Information data is the data that is ready for use in a variety every day purposes for example projecting sales.

Introduction

The dataset that I will be wrangling, analyzing and visualizing is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Step1: Data Gathering

The data is collected from three sources:

```
- The twitter archive data is provided by udacity and its manually downloaded and
updated to the         jupyter notebook.
- The image prediction data is downloaded programatically from udacity servers.
- The twitter jason data is downloaded from twitter using tweepy
```

Step 2: Assessing Data

We will use two types of analysis to assess our data:

```
1. Visual assessment, we would load our data and assess it using jupyter notebook.
For additional          visualization we would use excel.
2 Programatic assessment, we would use pandas functions to perform programtic
assesment. e.g             describe, info, sample, value_counts, duplicates
```

The issues found can be divided into two main types:

Quality Issues:- These issues are because of dirty data i.e. the data has problems with its content. Common data quality issues include missing data, invalid data, inaccurate data, and inconsistent data.

Tidiness Issues:- These are issues due to the structure of the data. It can also be referred to as messy data.

Quality

tweeter enhanced archive table

1. Sometimes the names of the dogs are incorrect like(all, my, not, a, an, the, by, such)
2. Some of the ratings are from retweets.

3. Sometimes the rating denominator is not equal to 10
4. Sometimes the rating numerator is extremely large
5. There is <a href= at the start and r-- at the end of the source column
6. Contains erronous data type of tweet_id
7. Some tweet_id do not have image predictions
8. Erronous data type timestamp, retweeted_status_timestamp are strings. They should be datetime.
9. The source column is duplicated.

image prediction

1. Erronous data type tweet_id
2. p1, p2 and p3 has mixture of lower case and upper case at their start(can't clean)

df table

1. Erronous data type tweet_id

tidiness

1. The retweet_count and favorite_count in df should be in the twitter enhanced table
2. The doggo, floofer, pupper, and puppo columns in twitter_enhanced table should be represented under a column named stage(the stage of the dog)

Before we start the cleaning process, let's creating the copies of our original datasets using pandas copy function

Step 3: Cleaning

Tidiness

We would start by solving tidiness issues then moving to data quality issues.

1.The retweet_count and favorite_count in df should be in the twitter enhanced table

solution

merge retweet_count and favorite_count from df table with the twitter table.

2.The doggo, floofer, pupper, and puppo columns in twitter table should be represented under a column named stage(the stage of the dog)

solution

we could join the doggo, floofer, pupper and puppo to a column named stage then delete the columns doggo, floofer, pupper and puppo.

Quality

Some tweet_id in twitter table do not have image predictions

solution

we could remove tweet_ids that are in twitter_clean table but not in predictions_clean table.

some of the ratings are from retweets

solution

Remove all tweets that are from retweets by cheking if the retweet_id is not empty.

sometimes the rating numerator is extremely large

solution

For consistency of our ratings we could replace the numerators which are greater than 17(where 17 is my maximum numerator) with 17.

sometimes the rating denominator is not equal to 10

solution

Replace rating_denominator for rows that have a rating_denominator that is not equal to 10 with 10

There is <a href= at the start and r-- at the end of the source column

Remove <a href= and r-- from the source column

The source column is duplicated

Drop the source column

Sometimes the names of the dogs are incorrect like(all, my, not, a, an, the, by, such)

Replace the names of the dogs that are like: all, my, not, a, an, the, by, such with "Unnamed" Also replace None in names to unnamed.

Erronous data type timestamp, retweeted_status_timestamp are strings

Convert data types in timestamp, retweeted_status_timestamp to datetime

Erronous data type in tweet id

Convert data type in tweet_id from interger to string

Step 4: Storing Data

We could use pandas to_csv function and pass the argument index=False to avoid it storing any index from our data