UDACITY PROJECT- WRANGLE AND ANALYZE DATA

DATA ANALYST NANODGREE

WRANGLE REPORT BY NEHA PATIL

INTORDUCTION

In this project we have to wrangle (gather, assess and clean) dataset. The dataset is a twitter archive of twitter user @dog_rates also known as WeRateDogs. This twitter account rates people's dogs with ratings almost always greater than 10. After gathering the data, assessing it and then cleaning it we have to analyze and record our findings and insight in a separate document 'act_report'.

GATHER:

Data was gathered from three different sources:

1. The enhanced twitter archive file was provided and downloaded manually. This file includes various variables for each tweet including tweet id, timestamp, text, rating numerator and denominator, name, etc.
2. The image prediction file was downloaded programmatically from a url using the requests() function. This file contains the predictions of breed of dogs based on the images provided.
3. The json tweet file was acquired with the help of Twitter API and tweepy package in python. This file contains columns such as retweet count and favourites count.

All these files were then opened in DataFrame.

ASSESS:

Assessing of data was done in two ways:

1. Visual: In this datasets were examined manually to catch any discrepancies that stand out visually. Large datasets can be viewed in python but it collapses rows and columns of the dataset. Dataset can also be viewed using Google Sheets.
2. Programmatic: In this step python inbuilt functions were used to find out dirty and untidy data.
    - info()
    - describe()
    - value_counts()

- duplicated()
- isnull()

All the finds were documented within two categories Quality and Tidiniess, which were then subdivided by the datasets they belong to. These are listed below

Quality

`twitter_archive table`

- rating_denominator has values more than 10
- for tweet_id 887517139158093824 name is 'such'
- name column has 'O' instead of O'Malley
- Erroneous datatpes (timestamp, retweeted_status_timestamp)
- Decimal values such 13.5 are interpreted as 5.
- name column has names such as 'a' ,'the' ,'an'
- html tags in source column
- missing values have None instead of Nan

`image_predictions table`

- smallcase names in p1,p2 and p3
- names have '_' or '-' instead of space between them
- missing data (2075 instead of 2356)
- ambigous column names (p1,p1_conf,p1_dog)

`json_tweet table`

- missing values in favourites_count,retweet_count,day_of_the_week and month

Tidiness

`twitter_archive table`

- instead of seperate column for each dog stage(doggo,floofer,pupper,puppo) single column called dog_stage

`json_tweet table`

- should be joined with twitter archive table as it contains the information regarding the tweets.

`image_predictions table`

- can be joined with twitter archive table as single column for dog_breed

CLEAN:

Cleaning process as not performed on all the points listed above. Each issue was given a definition which then coded and tested to check if it worked correctly. The following functions were used in the process of cleaning.

- join()
- extract()
- drop()
- np.nan
- append()
- to_datetime()
- islower()
- replace()
- rename()
- is_null()
- shape
- value_counts()
- info()
- head()
- Loops

CONCLUSION

Data in the real world is messy and analysis of this data is next to impossible. That's why data wrangling is an important step in the data analysis process. Performing all these steps diligently can make sure that our model works. Data wrangling is an iterative process.