DATA WRANGLING REPORT

INTRODUCTION

This project mainly focuses on data wrangling to fix the data quality and tidiness issues using python.

DATA GATHERING

1. 'Twitter_archive_df' : The WeRateDogs Twitter archive, which is provided by Udacity and pd.read_csv() to import them into dataframe.
2. 'Image_df' : The tweet image predictions, i.e., what breed of dog (or other objects, animal, etc) is present in each tweet according to a neural network. This file('image_predictions.tsv') is hosted on Udacity's servers and downloaded programmatically using the requests library and the provided url.
3. 'Twitter_json_df' : Using the tweet IDs in the WeRateDogs Twitter archive, query the twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called twitter_json_df file. Each tweet's JSON data is written to its own line.

DATA ASSESSMENT

The dataset is assessed for two types of issue : tidiness and quality.

Tidiness issues:

1. Dog stages are separately present in four stages - 'doggo', 'floofer', 'pupper' and 'puppo'.
2. The data is present in three different dataframes..
3. Numerator and denominator are present as separate column.

Quality issues:

1. Only original tweets should be considered but retweeted ones are also present in the dataframe.

2. Datatype of timestamp field is incorrect.
3. Since image data is present only for tweets till August 1,2017, so tweets beyond that date should be removed.
4. Datatype of tweet_id is incorrect.
5. Some dog names are invalid such as "A","An","The".
6. Some tweets have no images.
7. Denominator has strange values.
8. Incorrect 'rating_numerator' values are present.
9. Source column should be optimized.
10. Many columns have incorrect datatypes.
11. Many unwanted columns are present in the dataframe.

DATA CLEANING

1. Dog stages are separately present in four stages - ‘doggo’, ‘floofer’, ‘pupper’ and ‘puppo’ but they can be combined together into a single categorical column.
2. The data is present in three different dataframes, so three tables has to be combined to a single master dataframe ‘twitter_df’.
3. Numerator and denominator should be combined to a single column.
4. Remove re-tweeted ones and save only original tweets.
5. Change the datatype of timestamp field from string to datetime.
6. Remove tweets that are beyond Aug 1, 2017.
7. Convert tweet_id from int to string.
8. Some dog names are invalid such as "A","An","The".
9. Drop tweets that has no images.
10. Clean the denominator that has strange values.
11. Correct the 'rating_numerator' values and extract from the text information.
12. Optimize the source content by ‘Twitter for iphone’, ‘Twitter Web Client’, and ‘TweetDeck’.
13. Change datatypes of columns to their appropriate ones.
14. Unwanted columns should be dropped.