Objectives of the report:

Data Gathering
Assessing the Data
Quality Issues
Cleaning Data
Tidying the data
Storing and Cleaning Data

1. **Data Gathering**
    a. CSV Format:
        Comma Separated file with 2356 tweet data from November 2015 to August, 2017
        https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv
    b. TSV Format:
        TSV is tab separated file with data.
        https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
    c. Fetching twitter tweets using twitter tweepy API using access keys and access keys. This is stored In JSON object with tweet_json.txt file.
2. **Assessing the Data**
    a. df.head() : gives top 5 rows from the dataset
    b. df.tail() : gives bottom 5 rows from the dataset
    c. df.sample(10): random 10 rows will be given from the dataset.
    d. df.info() : returns column name, its datatype, and how many values are present so we can gain insights about missing values.
    e. df.describe(): return the the mean, std top 10%, 25%, 75% dataset values of floating point datatypes and integer type data.
3. **Quality Issues**
    a. Missing Data:
        1. many tweet_id(s) of df table are missing in img_df (image predictions). So drop the tweet_ids that are not present in img_df
        2. in_reply_to_status_id, in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp have a lot of missing values.
        3. expanded_urls also has a few missing values.
    b. In correct datatypes:
        i. Timestamp of the tweet should be in datetime format. So that data is converted to timestamp format.
    c. Removing invalid data:
        i. Dataset contains retweet data too. So while fetching retweet data we are getting same content of data. So that duplicate data should be removed.

ii. name column has values starting with lowercase characters (e.g. a, an, actually, by) which are incorrect names
iii. values of rating_numerator and rating_denominator has range between 0 to 10. But some values are higher than 10 which makes ration of numerator and denominator higher than 1.

d. Tidiness Issues:
i. doggo, floofer, pupper and puppo columns in df should be merged into one column named "stage"
ii. img_df table should be merged with df on tweet_id
iii. status_df table should be merged with df on tweet_id

e. Inconsistent data
i. Inconsistent pattern in values in p1, p2, p3 variables (first letter is in upper case and sometimes in lower case)

4. Cleaning Data:

a. Kept only those records in df_clean table whose tweet_id exists in img_df table
b. Kept only those rows in df_clean that are original tweets and NOT retweets (i.e. rows where retweeted_status_id column is null)
c. Dropped retweeted_status_id, retweeted_status_user_id , retweeted_status_timestamp from df_clean because they have no values now
d. Used to_datetime of pandas to change datetime object of the dataset
e. Dropped irreverent in_reply_to_status_id, in_reply_to_user_id columns
f. Removed all the HTML tags from the data
g. Deleted rows where rating_numerator outliers
h. Some rows have more than one dog stage

One tweet_id has both doggo and puppo
Nine tweet_id(s) have values present in both doggo column and pupper column.
One tweet_id has both doggo and floofer
Assigned value ‘Multiple’ for stage variable for the rows that have more than one dog stage.

i. Created a function to find gender based on some words in text column
j. Created breed column on the basis of p1 and p1_conf. If p1_conf > 0.95 and p1_dog is True, then breed is the value contained in p1.

5. Tidying the data

a. doggo, floofer, pupper and puppo columns in df_clean table were merged into one column named "stage". The data type of stage column was changed to category. Later doggo, floofer, pupper and puppo columns were dropped.
b. retweet_count, favorite_count,display_text_range columns from status_df table were joined with df_clean table on the basis of tweet_id by doing inner join.
c. Using pd.merge merged p1, p1_conf, p1_dog from img_df_copy with df_clean dataset on tweet_id

6. Storing the cleaned data

a. Stored the cleaned data in data_clean in a csv file named twitter_archive_master.csv