Data wrangling report

Project Details

Data wrangling, which consists of:

- Gathering data (downloadable file in the Resources tab in the left most panel of your classroom and linked in step 1 below).
- Assessing data
- Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on
    - 1) your data wrangling efforts and
    - 2) your data analyses and visualizations

First :Gathering data:

- From three (3) different sources.
- In three (3) different file formats.

Each piece of data is imported into a separate pandas DataFrame at first.

Source of the data	The name of the data	The format of the data
A given file I Downloaded this file manually then uploaded it to the workspace then read it with pandas.read_csv() method.	twitter_archive_enhanced.csv	CSV
A file that hosted on Udacity's servers,I downloaded programmatically using the Requests library from a given URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv	image_predictions.tsv	TSV
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 tweet_json.txt. Then read this .txt file line by line into a pandas DataFrame after I extracted the data I need(id, favourite count, retweet count)	tweet_json.txt	TXT

Second: Assessing data:

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues. Visually by looking at the data and examine it . And programmatically using some pandas dataframe methods . and I do that for each dataframe. Then I wrote any issue I found for every file separately.
And I found:

Tidiness:

- T1- There are 4 columns for stages of dog should be values in column named stage
- T2- There is 3 data frame that related each other should be merged

Quality:

A- enhanced twitter archive data:

- Q1- There are 181 retweets and 78 replies need to delete (validity)
- Q2- Invalid tweet_id data type (integer instead of string) (validity)
- Q3- timestamp data type (object instead of datetime)(consistency)(validity)
- Q4- Some names are not dog names (a, an,quite, None...)(accuray)
- Q5- 23 rating_denominator not equal 10 & 12 of them are a multiply of 10 & 5 of them aren't the multiple of 10(3of them are retwwets and replies therefore 2 have a mistake) & 3 are a replying & 3 is a mastaken data: (accuracy)
- Q6- index 1165 the rate should be 13/10 instead of 4/20 (accuracy)
- Q7- index 2335 the rate should be 9/10 instead of 1/2 (accuracy)
- Q8- index 516 the rate should be 11/10 according to replies instead of 24/7 (accuracy)
- Q9- index 1068 the rate should be 14/10 instead of 9/11 (accuracy)
- Q10- index 1662 the rate should be 10/10 instead of 7/11 (accuracy)
- Q11- There are rating_numerator fractions that cited by mistake as a rate (accuracy)
- Q12- rating_numerator has skewed numbers (1776 index 979 & 420 index 2074)(accuracy)
- Q13- Many columns are unuseful (consistency)

B- image predictions:

- Q1- missing images for the data (2075 rows only instead of 2356)(completeness)(accuracy)
- Q2- undescriptive name columns (p1_dog,p1_conf...)(consistency)
- Q3- unnecessary columns (i.e conf1 is the highest level of confidence)(accuracy)

C- twitter API:

- Q1- missing for the data (1219 instead of 2356)(completeness)
- Q2- id column's name is different with two data frames(consistency)

Third: Cleaning data:

First I copied the three dataframes to modify on the copy.
Then I started to solve the issues by defining the issue and how can it solve
Then write the solving code then test the code and the result.

After finishing assessing and cleaning the data I save the last modified data in new file to find it and read it by pandas easily.
Then start to ***visualize*** the cleaned data and ***analize*** it.