Data Wrangling Report

Ahmed Nabil Awaad

Data Gathering

First of all, I started with data gathering from three different sources:

1- From given flat file ‘twitter-archive-enhanced.csv’ using pandas library
   Using pandas.read_csv I read this file into data frame
2- From twitter API using tweets id using tweepy and json libraries
   Using tweepy I downloaded a json file from twitter API then I used Json library with loads method and opened file handling to read file lines to extract needed data in a data frame called tweet_counts
3- From downloading file using request library
   Using requests with its method get I read file then I opend file handling to write in it the page content

Output:

Then I get three uncleaned data frames

1- twitter_archive_enhanced
2- image_predictions
3- tweet_counts

Data Assessment:

After that I started to assess the data using:

Visual assessment:

I used Microsoft Excel for visual assessment

programmatic assessment:

and used, (jupyter notebook and pandas functions) for programmatic assessment

Data Cleaning:

I started cleaning with coping data frames

- I used Define, code, test structure in my cleaning process
- Also I used pandas, numpy, requests, matplotlib.pyplot, json, regular expression and tweepy libraries
- For pandas library I used a lot of methods like merge, str, astype, to_datetime, read_csv, to_csv, and others
- In request library get method had been used.
- In the following table I will give you a summary about data quality and Tidiness issues

Table Name	Quality Issues	Solution
twitter_archive_enhanced	timestamp column type is a string not a datetime tweet_id column type is a int not a string	use pd.to_datetime to convert timestamp type astype(str)
	remove +0000 in timestamp	strip +0000 from timestamp
	rating_numerator has a error numbers.	extract rating_numerator from text useing regular expression
	value 1776.00 in rating_numerator	replace this value with the max value in numertor
	rating_denominator has a lower numbers.	replace numerator at index 1950 with 9 and denominator with 10 replace numerator at index 382 with non and denominator with non replace numerator at index 1696 with non and denominator with non
	squad of dogs rating	get the average rate
	names have errors sometimes	replace 'a' and 'an with none
	delete non original tweets ,	Drop rows that have values in [in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,, retweeted_status_user_id]
	not all tweets with images	merge image prediction tweet id to remove records without images
	non-null object in doggo floofer pupper puppo	replace None from string to null
	delete un used columns (in_reply_to_status_id, in_reply_to_user_id, source,retweeted_status_id ,retweeted_status_user_id, retweeted_status_timestamp)	drop column [in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp]
image_predictions table	tweet_id column type is a int not a string	astype(str)
	number of records does not equal twitter_archive_enhanced records	merge tweet id from twitter_archive_clean to drop unused records
	column names does not express the meaning	rename columns
	merge p_dog with p in one column to remove the silly names in p type with non values in false	merge two columns and remove p_dog

tweet_counts table	number of records does not equal twitter_archive_enhanced records	Merged with first table

Table Name	**Tidiness Issues**	**Solution**
twitter_archive_enhanced	**- four variables dog stages in one column (doggo floofer pupper puppo)**	concat 4 columns in one column and drop them
	remove urls from text	split string using comma and take the first
tweet_counts	**should be part of the twitter_archive_enhanced table**	Merge weet_counts with twitter_archive on tweet_id

Output:

Finally I stored two tidy data frames:

1- twitter_archive_master.csv
2- image_predictions_master.csv