wrangle_report.md

Data Wrangling Report

Introduction

This project focuses on data wrangling on pre-gathered datasets of a Twitter user @dog_rates.

Datasets

- twitter_archive_enhanced.csv. It contains 2356 tweets with 17 columns of features about that tweet.
- image_predition.tsv. This file contains 3 possible predictions of the breed type of an image.
- tweet_json.jsonl. contains raw tweets taken from Twitter's API

Assessing Data

Examining the dataset revealed some issues.

Quality Issues

- Data type issues. timestamp is not datetime formatted, Id could be string, float Ids could be used as string id.
- Missing values not represented uniformly. Some rows have np.nan, some python None
- The source has HTML tags it could be cleaned and made categorical
- Date parsed as rating numerator
- Dog names have a large number of incorrect values (like "a", "an" the", "my")
- Unused columns could be dropped
- Rating denominator increases with the number of dogs in the picture. (like 120 if there are 12 dogs).
- Image predictions column names not clear

Tidiness Issues

- Dog types could be in one categorical column
- 3 dataset could be merged into one
- Expanded URLs of a row might have an array of URLs. if it does all the URLs are the same. they should have been "...photo/2, ...photo/3". We can either fix it or remove this - column and keep only the number of extended URLs. We could reconstruct the URL from tweet id
- Image prediction could return only one breed. the most confident one.

Iterating on data gathering

Iterating on data gathering

I decided to iterate on the data-gathering phase from the beginning. This means gathering up-to-date data from Twitter and making the preprocessing steps that Udacity already provided in the other datasets.

I followed these steps.

- Download twitter data using the twint project
- Use spacy NER, regex rules, and dog name dictionary to parse dog names from the tweet.
- Parse dog stage and ratings from text

- Create a dog breed classifier model and prepare predictions
- Examine for quality issues and clean the datasets
- Merge new datasets into a table and analyse

During this iteration, there were new Issues,

- null values are parsed as empty strings, arrays
- multiple representations of date and user field
- all True / False columns
- unused columns and when they exist their rows (quote and retweets)

Result

- resulting table has 3273 rows and 11 columns
- contains data from 2016 to 2021

	id	tweet	replies_count	retweets_count	likes_count	datetime	dog_name	rating_numerator	rating_denominator	dog_stage	breed_prediction
1089	1050816834549755904	This is Gunner. During a routine check-pup, a ...	265	6084	39807	2018-10-12 21:33:51-03:00	Gunner	13	10	NaN	Labrador_retriever
5165	667090893657276420	This is Clybe. He is an Anemone Valdez. One ea...	0	112	305	2015-11-19 00:23:57-03:00	Clybe	7	10	NaN	Chihuahua
3379	754120377874386944	When you hear your owner say they need to hatc...	43	2265	7655	2016-07-16 04:08:03-03:00	NaN	10	10	NaN	cocker_spaniel
3161	778748913645780993	This is Mya (pronounced "mmmyah?"). Her head i...	45	1268	6735	2016-09-22 03:13:04-03:00	Mya	11	10	NaN	Labrador_retriever
3430	750086836815486976	This is Spanky. He was a member of the 2002 US...	13	514	2108	2016-07-05 01:00:12-03:00	Spanky	12	10	NaN	pug
668	1151898540370608128	This is Smoltz and Maddux. They often match th...	857	16209	119088	2019-07-18 19:56:27-03:00	Smoltz	14	10	NaN	golden_retriever
4797	673355879178194945	This is Koda. She's a boss. Helps shift gears....	7	527	1405	2015-12-06 07:18:46-03:00	Koda	11	10	NaN	Border_collie
4982	670411370698022913	Meet Scooter. He's ready for his first day of ...	18	816	1873	2015-11-28 04:18:21-03:00	Scooter	12	10	NaN	Maltese_dog
797	1115650683267469312	This is Gwendy. She has lupus, so she needs to...	631	15964	130848	2019-04-09 19:20:25-03:00	Gwendy	14	10	NaN	Shiba_Dog
2240	887101392804085760	This... is a Jubilant Antarctic House Bear. We...	138	5209	27886	2017-07-18 03:07:08-03:00	NaN	12	10	NaN	Samoyed