Data Analyst Nano degree

April 19, 2020

Wrangle Report

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. Has over 8.7 million followers as of now and almost 5 years of history.
One particular aspect is the rating system which is almost always greater than 1, e.g. 13/10.

This report describes in a technical way, the efforts required to wrangle tweets from #WeRateDogs account from a base of around 2500 tweets.

Gather

A history of tweets from WeRateDogs account were provided, however that information was not enough as it does not contain retweet and favorite counts among other interesting fields.
In order to get extra information, it was necessary to surf Twitter API.

Twitter requires us to fill a form so they can know what our purposes are.
I filled form accordingly but unfortunately have not received any response (Case# 0150166583 Twitter developer account application). COVID 19 might be a factor, so I decided to take text file provided by Udacity to forward my research. It would have been interesting to interact with API as I was interested in pulling out extra fields such as *reply_count* and *quote_count* for further engagement analysis or if certain tweets are withheld in some countries.

I also researched there is a way to get bulked tweets in packages of 100 size using this endpoint https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-statuses-lookup.

One thing I noticed is much of the gather information through API is nested, so we needed to pull fields out for further treatment.

Assessment

Some of the most notorious issue with provided tweet history is the wrong handling of dog names and rating numerators/denominator as there are some tweets that don't use the same template "This is {dog_name}. ….". Indeed, more work would be required to extract dog names with a high level of confidence, a Natural Language Processor tool or a more complex regular expression. Same applies for ratings.

One of the first things I considered was to try to find any duplicated tweets. Fortunately I did not find any on tweet history neither in tweets extracted through API.

Pandas has a bunch of util methods to get the big picture of a data set. Some of the most valuable methods I found were value_counts(), nunique(), isna() and info().
I also found data structure issues, the most notable ones to me were the doggo, floofer, pupper and puppo fields which contained redundant information and most of the cases a wrong representation of “None” values. I also considered that tweets from archive and tweets from api shoaled be merged into a single data set for simplicity.

Cleaning

After assessing data, I decided to start with tidiness issues so hopefully data cleaning would go smoother.
Right after tackling tidiness issues, I decided to get rid of tweets that don’t fulfill requirements (retweets). This saves us extra work as we don’t care to clean dirty tweets that don’t even make sense for research.
Among the remarkable tidiness processes were to extract favorite_count, retweet_count and extended_media from api tweets and attach them to archived tweets. From the extended_media I considered appropriate to set all media assets related to tweets in a different data frame. This could benefit other researchers to perform analyses on media like predict more dog breeds.
One of the toughest tasks was to create the dog_stage column. Originally the plan was to use some sort of pandas melting, however I realized certain tweets tag their dogs with multiple dog_stages. Melt is powerful when there is just one value to be assigned. So string concatenation was the simplest way to solve it.

At the end, cleaned data sets were exported as csv files and also as a SQL database.

Summary

This project depicts how important is data wrangling as part of data analysis efforts.
With a reliable methodology, there is much more certainty of data quality and therefore, more chances are to perform accurate analysis and achieve realistic conclusions.
I would like to thank Udacity team for their excellent work at content development. Being exposed to this kind of projects makes one confident about data analysis skills.