Wrangling action taken

WeRateDogs dataframe

The project was done in 3 sections gathering, assessing, cleanig, storing and lastly analyzed and visualised data.

Gathering

I used Anaconda from computer which has pandas and Numpy pre-installed how ever what was installed was request and tweepy using pip. I started gathering data process by directly downloading twitter_archive.csv from udacity classroom by clicking on it, saved it and imported it to jupyter.The image prediction tsv file was downloaded programmatically using requests. The last one I used tweepy library to sercure additional data via the twitter API. Function .value_counts() was used to see total of

Assessing data

Assessing data included visualising and prgrammatic assessment. Quality issues and tidiness issues were also written down. A few markdowns and some code explanation was done. After importing and displaying the data by using the name it was imported by the data showed missing data and some tidiness issues which were written down at the end of this section.

Data consistsency was check only at the beginning of the data and end. The summary of the dataframe was seen using .info which shows that there are 2355 rows and 17 columns, different data types, size if the data and headers for each column in the twitter_archive.csv dataframe which I renamed it to enhanced_df.

The image file sran it using image_df.info() code to see its data which showed 12 columns and 2075 rows and different data types. Furethuremore the tweet-jason.txt has 31 columns and 2353 rows with different datatypes.

In the assessment and throught out the project some quality,visualisation and programmatic assessment will be done where needed throughout the project.

Cleaning data

In this section a copy of the data which was gathered and assessed was copied before it could be cleaned just incase a mistake was to be made or the original data need to be looked at again the data will be available.

Previous assessments are being looked at and issues seen from assessment section is used to clean data.

Code was defined, written and tested to explain each step and the purpose of it. All three dataframes were merged so cleaning and anlysing will be easier and the merged data was named combined_df. Missing data was removed since it has been showing to be a common problem. Null values was replaced by 0.Data types was changed. Check if there were any duplicates which showed 0 meaning no there are no duplicates, this is important because some data might be count twice when it’s the same thing, giving an imoressin of the wrong data. Some columns were dropped.

Data was stored and later used for visualisation.

WeRateDogs™ @dog_rates · Feb 1

This is Ellie. It's her first time walking with a leash. Rather skeptical of these new frolicking limitations. 12/10 it'll be ok Ellie

THE Katie Freeman