Wrangle_report

August 17, 2020

1 Wrangle Report: By Hiten Naran

1.1 Brief Overview:

The purpose of this task is to gather, assess and clean WeRateDogs Twitter data in order to enable us to create interesting and trustworthy analyses and visualisations.

The task was broken down into three segments:

- Gathering data
- Assessing data
- Cleaning data

I will go into more detail below as to what I did during each of the three segments

1.1.1 1. Gathering Data:

Data was gathered from three different sources:

Tweet_archive:

- Which was already provided and manually downloaded and uploaded. It contained basic info such as 'tweet_id', 'timestamp', 'text' etc.

image_predictions.tsv:

- This was downloaded this programmatically using the requests library.

tweet_json.txt:

- This was one tricky as it required me to capture this data via the Twitter API.
- I first loaded the Twitter API in the Jupyter Notebook using my credentials.
- The API was then queried using the tweet ids from the 'Tweet_archive' table. Of the tweets that were available, we extracted the json code from the queried data which enabled the creation of a Pandas DataFrame. The DataFrame was then saved as a .txt file.

1.1.2 2. Assessing Data

I used the following commonly used methods to assess the data and identify 'Quality' and 'Tidyness' issues.

- .duplicated()
- .info()
- .shape
- .describe()
- .value_counts()

Through using the methods listed above I was able to identify some of the following 'Quality' and 'Tidyness' issues which will serve useful for analysing the data later:

1.1.3 Quality Issues:

Tweet_archive set - timestamp column should be reformatted to DateTime - invalid denominator_ratings values with some values having a value that is not 10 - rating_numerator goes up to 1776 - rating columns need to be reformated to floats as some of the ratings should be floats but are not being picked up properly from the tweet copy. - tweet_id is formated to interger and needs to be changed to string. - Want to remove retweets, these can be identified by seeing if there are values within retweeted_status_id

Image_prediction - Some issues with the names itself i.e. 'shopping_cart' which isn't a name of a dog

tweet_json - tweet_id is formated to interger and needs to be changed to string. - created_at column should be reformatted to DateTime

1.1.4 Tidyness Issues:

- All 3 tables should be merged on tweet_id as each table is themed around dogs

Tweet_archive set - The last four columns all relate to the same Dog variable and should have these columns melted

Image_prediction - We are only interested in understanding the True predicted outcome of Dogtype. Create additional column which only shows the correctly predicted dogtype alongside the acompanying image. Also include an additional column with the confidence interval

1.1.5 3. Cleaning Data

The cleaning process was broken down using the following logic in order to address the 'Quality' and 'Tidyness' issues identified during the assessment:

- Define and Code
- Test

Before cleaning we created copies of the original DataFrames and ensured that cleaning was only done on the copied versions.

The following methods were utilised to help with the cleaning process:

- pd.merge()

- .head()
- .shape
- .drop()
- pd.to_datetime()
- .astype()
- .info()
- .loc[]
- .sort_values()
- .isnull()
- .notnull()
- .apply()
- .value_counts()
- .columns

The final DataFrame was saved to the following csv file:

- ’twitter_archive_master.csv’

```
In [ ]:
```