Qasem M. Dhamad

Udacity - Data Analyst - MCIT

Wrangle Report

This report briefly describes my wrangling efforts in this project. The dataset of twitter archive of the twitter user named "@dog_rates" or called WeRateDogs. This twitter account rates people's dogs with comments about their dogs. To illustrates, originally the ratings have a dominator of 10. However, some ratings have greater than 10. For instances, 13/10, 11/10, 12/10, and so on. This is because some people think that their dogs' ratings deserve more than 10. The twitter account of @dog_rates has more than 9 million followers. In addition, he received an international medical coverage as stated.

This project was completed on the Udacity Project Workspace. As I choose to generate this report as a PDFs using Microsoft Word.

The following are the processes of generating this project:

1- Gathering data.

2- Assessing Data.

3- Cleaning Data.

Gathering data

The Data In this project that we will work on the following three datasets.

A. Enhanced Twitter Archive:

The first column contains each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo. There are 2356 tweets.

Qasem M. Dhamad

Udacity - Data Analyst - MCIT

1. Directly download the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv)

```
In [2]: twitter_archive_df=pd.read_csv('twitter-archive-enhanced.csv')

In [3]: twitter_archive_df.tail()
```

2352	666044226329800704	NaN	NaN	16 00:04:52 +0000	href="http://twitter.
2353	666033412701032449	NaN	NaN	2015-11-15 23:21:54 +0000	href="http://twitter.
2354	666029285002620928	NaN	NaN	2015-11-15 23:05:30 +0000	href="http://twitter.

B. Additional Data via the Twitter API:
Contains tweet ID, retweet count, and favorite count.":

Downlaoding tweet image prediction programmaticly using the request library:

```
url='https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_i

response = requests.get(url)

assert response.status_code==200
response.content

with open('image-predictions.tsv', mode= 'wb') as file:
    file.write(response.content)
```

```
image_prediction = pd.read_csv('image-predictions.tsv', sep='\t')
image_prediction.head()
```

	tweet_id	jpg_url	img_num	
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springe
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_r
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_

Qasem M. Dhamad

Udacity - Data Analyst - MCIT

C. Image Predictions File:
Contains tweet_id, image_url, image_num, p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog:

Downlaoding tweet image prediction programmaticly using the request library:

```
url='https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_i

response = requests.get(url)

assert response.status_code==200
response.content

with open('image-predictions.tsv', mode= 'wb') as file:
    file.write(response.content)
```

```
image_prediction = pd.read_csv('image-predictions.tsv', sep='\t')
image_prediction.head()
```

	tweet_id	jpg_url	img_num	
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springe
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_r
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_

Assessing Data

After gathering the dataset together, we asses visually and programmatically the quality and tidiness issues.

Following are the findings that were conducted:

Qasem M. Dhamad

Udacity - Data Analyst - MCIT

Quality issues

Enhanced Twitter Archive:

1. Some dog has an invalid name such as "a, an, None, etc")
2. Data type of tweet_id is invalid (int instead of object or str).
3. row number of 315 & 1016 have Zero numerator.
4. 20 rating_denominator greater than 10.
5. 23 rating_denominator not equal 10.
6. 181 retweets as indicated by "retweeted_status_id".
7. Invalid timestamp dtype which is (string not datatime).

Additional Data via the Twitter API:

1. Two missing IDs (2356 - 2354)

Image Predictions File: ¶

1. In columns p1,p2,and p3:underscores are used in some words instead of spaces.
2. Also,in columns p1,p2,and p3: some values start with uppercase letter and some start with lowecase.

Tidiness issues

1. All three data above are relevant.However, they are separated into three dataset named as "image_prediction, tweet_data_df,twitter_archive_df".

2. Dogs stage data is separated into 4 columns including:doggo, floofer, pupper, puppo.

Cleaning Data

After assessing the data, we started cleaning the dataset by making copies of the original data before cleaning. And these cleaning includes merging individual pieces of data according to the rules of tidy data. The result was conducted as a high-quality and tidy master pandas DataFrame (or DataFrames, etc).

Qasem M. Dhamad

Udacity - Data Analyst - MCIT

Creating copies from the original data before cleaning:

Making copies of original pieces of data

```
#Taking copy of the additional Data via the Twitter API
tweet_data_df_copy = tweet_data_df.copy()
```

```
# Taking copy of the Enhanced Twitter Archive
twitter_archive_df_copy = twitter_archive_df.copy()
```

```
#- Taking copy of the Image Predictions File
image_prediction_copy = image_prediction.copy()
```

Below as examples of the cleaning process for one issue:

Issue #1:

Tidiness issues

1- Dog stage data is seperated into four columns:

Define:

Merging the four columns into one column as shown below:

Code

Here, we are extracting dog stages from text columns into a new column named "dog_stages"

```
twitter_archive_df_copy['dog_stages'] = twitter_archive_df_copy['text'].
```

```
twitter_archive_df_copy.head()
```

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	href="http://twitter.com
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	href="http://twitter.com