PROJECT 4: Wrangle report

briefly description of the wrangling efforts.

> This project consisted of elaborating in an orderly manner the different steps of the data wrangling: collecting, assessment, cleaning Data collection is an essential and inevitable part of data analysis processes. It is the foundation stone that can define the success or failure of an analysis project. The goal is to wrangle WeRateDogs twitter data to create interesting and trustworthy analyzes and visualizations. To develop this project we have collected data from three different sources.

Gathering data

In [1]:

```
import pandas as pd
import numpy as np
```

> • Firstly, we were provided with a .csv file: twitter_archive with 5000+ records. This was easily incorporated into the workspace using a PD. Read_csv

In [2]:

```
twitter_archive = pd.read_csv(r'recursos/twitter-archive-enhanced.csv')
twitter_archive.head()
```

Out[2]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	retwee
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" r...	This is Phineas. He's a mystical boy. Only eve...	
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you....	
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" r...	This is Archie. He is a rare Norwegian Pouncin...	
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http://twitter.com/download/iphone" r...	This is Darla. She commenced a snooze mid meal...	
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="http://twitter.com/download/iphone" r...	This is Franklin. He would like you to stop ca...	

> • Second, a url address was supplied containing the results of the predictive model that identifies the breed of the dog that appears in each .jpg. To access this file, a 'requests' was requested and it was saved in the local repository. As it

> is a file in .tsv format, when entering it into the workspace, its type of separator must have been specified (i.e. sep = ’\t’). the resulting file was registered as image_predict.tsv

In [3]:

```
image_pred = pd.read_csv(r'tweet_image/image-predictions.tsv', sep='\t')
image_pred.head()
```

Out[3]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074	True	
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826	True	miniature_
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461	True	
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143	True	
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311	True	F

> • Third, the tweepy function was used to make an api call to the twitter database. After generating the tokens and keys, because you have the WeRateDogs Twitter archive and specifically the tweet IDs within it, we use the values in the twitter_archive.csv to get the count of retweets and bookmarks that those tweets got.

In [4]:

```
tweet_info = pd.read_csv('tweet_info.csv')
tweet_info.head()
```

Out[4]:

	tweet_id	fav_count	retweet_count
0	666020888022790149	2424	464
1	666029285002620928	121	42
2	666033412701032449	112	41
3	666044226329800704	273	132
4	666049248165822465	96	40

Assessing Data

> Assessing and cleaning the entire dataset completely would require a lot of time, and is not necessary to provide a complete process in data wrangling. Therefore, the assessment and cleaning steps are not complete. In the second step, the data was inspected for two things: quality issues (i.e. content issues) and lack of tidiness (i.e. structural issues).
>
> - Data qualities dimentions:
>   - Completeness: do we have missing record or not?
>   - Validity: records that do not conform to a defined schema (i.e content out
>   - Accuracy: wrong data that is value (i.e. wrong dog race names)
>   - Consistency: is both valid and accurate but there are multiple corrects ways to referring (i.e. Capitalizations cases)
> - Data tidiness dimentions:
>   - Each variable forms a column.
>   - Each observation forms a row.
>   - Each observation unit forms a table.

The resulting assessment

Quality

`twiter_archive` table:

- Missing values in *in_reply_to_status_id* and *in_reply_to_user_id*.
- Erroneous Datatypes(*timestamp*)
- Invalid data: values prior to August 1st, 2017 can be removed, no image_pred related.
- invalid data: retweets do not have to be included in the prediction algorithm.

`image_pred` table:

- Lowercase and Uppercase mixed in *p1, p2, p2* columns.
- Invalid value *web_site*, *teddy* in *p1* column.
- Invalid links in *jpg_url*
- Invalid value *doormat* in p1
- jpg_url duplicaded values

`tweet_info` table:

Tidines

`twiter_archive` table:

- Unsustantial columns
- 1 variable in 4 columns (*floofer, pupper, puppo, doggo*)
- *fav_count* and *retweet_count* should be in twiter_archive
- retweet info columns

Cleaning Data

To do the cleaning process, several numpy and panda tools will be used. Changing the structure of the table by removing invalid and inappropriate values. Below is the result of this process. Contained in two files:

- `'twitter_archive_clean.csv'`
- `'image_pred_clean.csv'`

In [5]:

```
df1 = pd.read_csv('twitter_archive_clean.csv')
df1.head()
```

Out[5]:

	tweet_id	timestamp	source	text	rating_numerator	rating_denominator	nam
0	891815181378084864	2017-07-31 00:18:03+00:00	<a href="http://twitter.com/download/iphone" r...	This is Archie. He is a rare Norwegian Pouncin...	12.0	10.0	Archi
1	891689557279858688	2017-07-30 15:58:51+00:00	<a href="http://twitter.com/download/iphone" r...	This is Darla. She commenced a snooze mid meal...	13.0	10.0	Darl
2	891327558926688256	2017-07-29 16:00:24+00:00	<a href="http://twitter.com/download/iphone" r...	This is Franklin. He would like you to stop ca...	12.0	10.0	Frankli
3	891087950875897856	2017-07-29 00:08:17+00:00	<a href="http://twitter.com/download/iphone" r...	Here we have a majestic great white breaching ...	13.0	10.0	Non

	tweet_id	timestamp	source	text	rating_numerator	rating_denominator	nam
4	890971913173991426	2017-07-28 16:27:12+00:00	<a href="http://twitter.com/download/iphone" r...	Meet Jax. He enjoys ice cream so much he gets ...	13.0	10.0	Ja

```
In [6]:
```

```
df2 = pd.read_csv('image_pred_clean.csv')
df2.head()
```

```
Out[6]:
```

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	666020888022790149	0.465074	True	666020888
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	666029285002620928	0.506826	True	666029285
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	666033412701032449	0.596461	True	666033412
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	666044226329800704	0.408143	True	666044226
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	666049248165822465	0.560311	True	666049248

```
In [ ]:
```