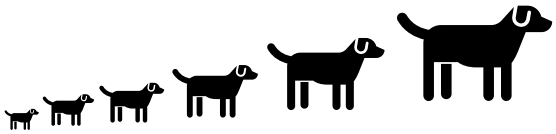
Udacity

WeRateDogs Project



In this project I will process the data received from three different sources based on the "We Rate Dogs" social media accounts, wrangle, clean and analyse it, as well as visualise and report my findings. One of the sources is a pre-processed "We Rate Dogs" Twitter archive containing over 5000 tweets. Another source is "tweet image predictions" which gives predictions about the breed of dogs and other information according to a neural network, received in the form of a URL. The remaining necessary information is obtained by querying the Twitter API.
The overall goal is to assess which breed of dog is the most popular one. While working on this project, we need to keep in mind that the data is based on humorous content and that the ratings per se are all very positive and do not follow traditional scientific logic (e.g. the dogs receive ratings higher than the presented threshold of 10, because they are all "very good boys").

Tasks of this project:

-Data wrangling
  Gathering data
  Assessing data
-Cleaning data
-Storing, analyzing, and visualizing wrangled data
-Reporting on
  1) data wrangling efforts
  2) data analyses and visualizations

Data Cleaning

Cleaning of the three obtained dataframes:

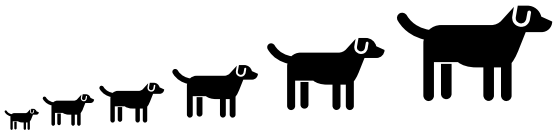
- df_archive
- df_pred
- df_api

Tidiness Issues

- dog age categories are all in separate columns, should be merged
- df_archive and df_api should be merged

Quality Issues

- there are a lot of missing values in df_archive
- there is an inconsistency in indicating missing values ("None" vs. "NaN" vs. "")
- the name list contains "a", "an", "the" and "officially" and other values which are not names
- some rows are moved to the left and needed to be dropped beforehand ("https..." in "tweet_id" row)
- the rating pattern is not consistent, as it is humorous content
(some ratings very high, falsifying data analysis results)
- the dog names include "_" as well as upper and lower case
- some dog age stages are a compound of two stages (it is unknown whether this was done on purpose in some cases or whether it is a formatting error)
- in the image prediction results, some are not dog breeds, but objects and need to be dropped

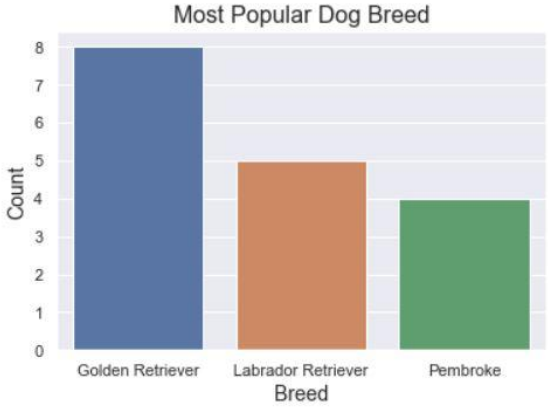


Analyzing and Visualizing

Most Popular Breed

1) Which dog breed is the most popular one?

```
Out[452]: Text(0.5, 1.0, 'Most Popular Dog Breed')
```

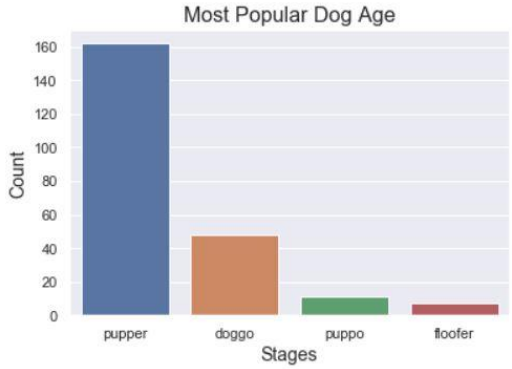



```
As we can see, Golden Retriever, Labrador Retriever and Pembroke are the most prevalent
and thus most popular dog breeds : ).
```

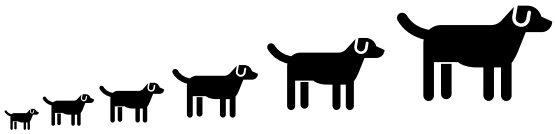
Dog Stages

2) The dogs' ages are divided into four dog stages. Which one is the most posted and therefore the most popular one?

```
Out[453]: Text(0.5, 1.0, 'Most Popular Dog Age')
```



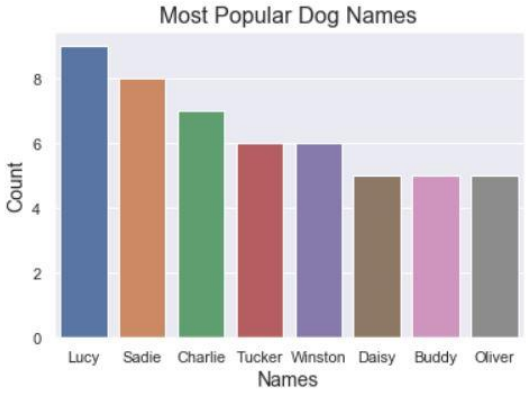

```
As we can see, "pupper" is the most posted dog type. It makes sense, since young dogs are considered the cutest : ).
```



Dog Names

3) Which dog names are the most popular ones?

```
Out[447]: Text(0.5, 1.0, 'Most Popular Dog Names')
```




```
As we can see, Lucy, Sadie and Charlie are the most popular dog names. Followed directly by Tucker,
winston, Daisy, Buddy and Oliver : ).
```

Sources:

http://www.compciv.org/guides/python/how-tos/downloading-files-with-requests/
https://stackoverflow.com/questions/9652832/how-to-load-a-tsv-file-into-a-pandas-dataframe
https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781783551668/1/ch01lvl1sec10/reading-and-writing-csv-tsv-files-with-python
https://stackoverflow.com/questions/39267614/csv-file-does-not-exist-pandas-dataframe
https://stackoverflow.com/questions/28384588/twitter-api-get-tweets-with-specific-id
https://wiki.python.org/moin/HandlingExceptions
https://stackoverflow.com/questions/19377969/combine-two-columns-of-text-in-dataframe-in-pandas-python
https://stackoverflow.com/questions/11858472/string-concatenation-of-two-pandas-columns
https://cmdlinetips.com/2018/04/how-to-drop-one-or-more-columns-in-pandas-dataframe/
https://stackoverflow.com/questions/11346283/renaming-columns-in-pandas
https://pandas.pydata.org/pandasdocs/stable/reference/api/pandas.DataFrame.sort_values.html
https://stackoverflow.com/questions/41800424/remove-rows-in-python-less-than-a-certain-value
https://medium.com/@waliamrinal/saving-a-pandas-dataframe-as-a-csv-file-4f8b74b7a1bc
https://medium.com/@waliamrinal/saving-a-pandas-dataframe-as-a-csv-file-4f8b74b7a1bc
https://stackoverflow.com/questions/28679930/how-to-drop-rows-from-pandas-data-frame-that-contains-a-particular-string-in-a-p/43399866
https://stackoverflow.com/questions/41157981/pandas-convert-float-in-scientific-notation-to-string
https://stackoverflow.com/questions/17950374/converting-a-column-within-pandas-dataframe-from-int-to-string
https://stackoverflow.com/questions/25050141/how-to-filter-in-nan-pandas/25050185
https://seaborn.pydata.org/generated/seaborn.countplot.html