wrangle_report

August 21, 2018

1 INTRODUCTION

I will be briefly going over the steps that I took through this project. There are 3 main steps that had to be covered:

```
1. Gathering Data
2. Assessing Data
3. Cleaning Data
```

1.1 GATHERING DATA:

There were 3 datasets from 3 different sources that had to be loaded into the Jupyter Notebook:

The twitter archive enhanced csv that was given to us at the beginning of the project. This was manually downloaded then loaded into the Notebook named `TAE_DF`.

The tweet image predications had to be downloaded programmatically from the Udacity server, that was provided to us. This was a TSV file that was loaded into the Notebook named `IMAGE_DF`.

The last dataset was using the Twitter API for each tweet's JSON data. This is where I had the most issues loading in data. I never received an API consumer key and token from Twitter, I reached out and received the txt file from a Udacity mentor then loaded the txt file into the Notebook named `TWEET_DF`.

1.2 ASSESSING DATA:

With this second step, I had to view each dataset that was loaded into the Notebook both programmatically and visually. I had to assess each for the quality and tidiness of the data. From there, found these issues to correct:

1.2.1 QUALITY FINDINGS

TAE_DF:

```
Remove Retweets
Remove columns that are not useful in the analysis
Erroneous dog names to change to None
Numerator and Denominator have ludicrous numbers or NaN, need to be converted to a float
Convert id to string
```

IMAGE_DF:

```
Remove _ between names
Capitalize Names in columns p1, p2, p3
Remove Duplicates
Convert id to string
```

TWEET_DF:

```
Convert id to string
```

TIDINESS FINDINGS

```
Move all dog breeds into a single column
Merge p1, p2, p3 & p1_conf, p2_conf, p3_conf in images_df dataset into seperate, condensed colum
Megre all 3 datasets into 1 dataset
```

1.3 CLEANING DATA:

In this last step, I first made copies of all 3 datasets. I listed out each issue in stages: Define, Code & Test. I had issues with finding all the numerators and denominators, then converting them to 10. Also, my biggest hurdle was combining predicting breeds, finding the highest confidence level, if the dog ==True, then bringing back just the breed and confidence level.

After cleaning and tiding the data, I stored the data into 1 dataset, `twitter_archive_master.csv`.