ANALYSIS AND VISUALIZATION REPORT

UDACITY 'WERATEDOGS' PROJECT

By: Genevieve Aggrey-Ampiah

WeRateDogs is a Twitter account that rates dogs doing funny things and then, gives them hilarious ratings such as 20/10 to describe how unique they are.
In this Udacity project, we were to retrieve the relevant data from three different sources and then run the Data Analysis steps of Gathering, Assessing, Cleaning, Storing and Visualizing.

Gathering
In this step all three datasets were collected from different sources and run to get the complete dataset that would be used for the project.

Assessing
This stage involves checking the data both visually and programmatically for issues that need cleaning. Data is to be checked for both its Tidiness and its Quality.

Cleaning
After the above stage of assessing is done, then it's time to clean the data. Data is said to be Tidy if each variable forms a column, each observation forms a row and each type of observational unit forms a table.

Storing
After the initial processes of gathering, assessing and cleaning had been completed, some columns were dropped after which the remaining data was stored in a main data frame.

Insights
With the stages above being completed, a few insights were drawn after the analysis.

1. **To find out which name is given the most to dogs.**

```
# Checking the count value of the dog names to know the most used name.
Total_Data['name'].value_counts()
```

```
Charlie      12
Oliver       11
Lucy         11
Cooper       11
Penny        10
Tucker       10
Lola         10
Winston       9
Bo            9
Sadie         8
Bailey        7
Daisy         7
Buddy         7
Toby          7
Leo           6
Milo          6
Jax           6
Rusty         6
Bella         6
Dave          6
Scout         6
Koda          6
```

The above image displays the most used 20 names and their counts after the programmatic assessment of the data. It further goes to show that the name Charlie is the most used name with a count of 12. Close seconds are Oliver, Lucy and Cooper with a count of 11 each.

The second insight to be drawn is:

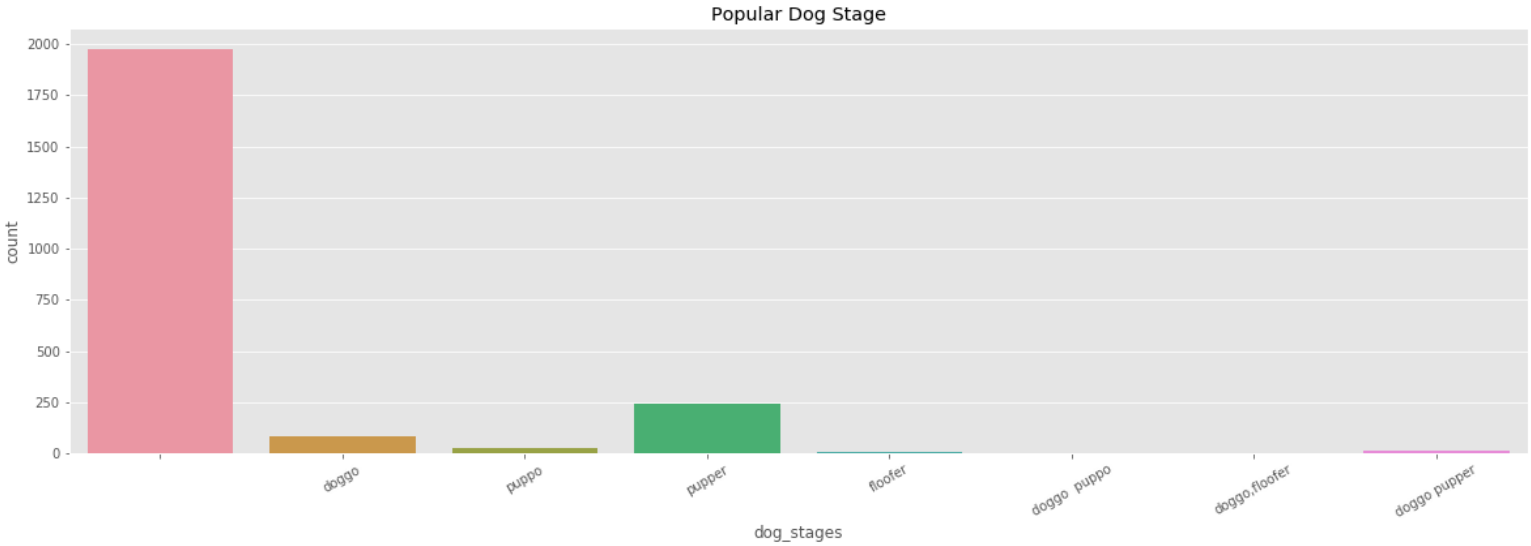
1. **2. Which dog stage is most documented.**
   To ascertain the answer to this question, columns containing all dog stages had to be merged into one column named dog stage. After the merger a count was carried out.

```
Count_dog_stages = Total_Data['dog_stages'].value_counts()
```

```
Count_dog_stages
```

```
                1976
pupper           245
doggo             83
puppo             29
doggo pupper      12
floofer            9
doggo  puppo       1
doggo,floofer      1
Name: dog_stages, dtype: int64
```

Image above displays the count of all the dog stage. Aside from the columns with no clear classification of dog stage, Pupper with a count of 245 is the stage which most dogs are.



Visualization to prove the analysis of the insight.

The third insight drawn from the analysis of the Udacity project is:

3. Which dog breed is the most popular with people.

Golden_retriever	150
Labrador_retriever	100
Pembroke	89
Chihuahua	83
Pug	57
Chow	44
Samoyed	43
Toy_poodle	39
Pomeranian	38
Cocker_spaniel	30
Malamute	30
French_bulldog	26
Chesapeake_bay_retriever	23
Miniature_pinscher	23
Seat_belt	22
Siberian_husky	20
Staffordshire_bullterrier	20
German_shepherd	20
Web_site	19
Cardigan	19

After further analysis of the entire data set, I derived that most dogs represented on the site were Golden Retrievers with a count of 150. Labrador Retriever is the second most popular dog breed with 100.

Inconclusion, further analysis would have to be run to ascertain a more conclusive outcome to the projects dataset.