14-May-18

# Executive Summary

## Product Description/Objective

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog Target Audience. These ratings almost always have a denominator of 10.The numerators, almost always greater than 10. 11/10, 12/10, 13/10, etc.
The Objective is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Wrangling involves following steps.

1. GATHERING: It is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes.
2. ASSESSING: It is the process of scientifically and statistically evaluating data in order to determine whether they meet the quality required for projects or business processes and are of the right type and quantity to be able to actually support their intended use.
3. CLEANING: It is the process of detecting and correcting (or removing) quality or tidiness issues from a data set and then replacing, modifying, or deleting the dirty or coarse data.

### Gathering:

In this section, we are going to gather the data which will help us doing interesting analyses and visualizations. The data we have already doesn't have enough information like retweet count, favorite counts or what's the breed of the dog.

We will be using two approaches to gather the data,

1. To get the type of breed, we gonna a use neural network approach to get the breed of dog based on their images
2. To extract retweet and favorite counts. We gonna a query twitter API for each tweet id's json data using python's tweepy library.

### ASSESSING:

In this section, we gonna assess the gathered data both visually and programmatically to identify quality and tidiness issues.

1. Visual Assessing: we will load the gathered data in to the jupyter notebook and visually analyze the data set to identify quality and tidiness issue and document them respectively.

2. Programmatic Assessment: Python has libraries like numpy and pandas which will help us assessing the data which can’t be done through visually and document the quality and tidiness issues respectively.

Below are some of the issues identified during assessment:

Quality:

1. Rating Numerator data type should be float
2. Rating Numerator values needs to get modified
3. Timestamp datatype should get converted to datetime.date
4. Incorrect names are displayed Eg: paper_towel is not a dog breed
5. Confidence intervals columns can be round-off to 2 decimals to the right.
6. Type of dog datatype needs to be category type
7. Name column has some incorrect values like a,an,The,This.

Tidiness:

1. Tweet_api table can be merged with twitter enhanced table
2. Image predictions table can be merged with twitter enhanced table
3. Combine different types of dog stage in to a single Dog_Stage column

# CLEANING:

In this final section, we gonna clean the data set with the help of the documentation from assessing section which has the quality and tidiness issues. Cleaning process is carried out so that we can get accurate visual results during the analysis process.

# Conclusions

In this project, we took a 'We rate dogs' dataset and have used Data Wrangling processes like Gathering the data from various sources and Assessed the data to find out the Quality and Tidiness issues and then cleaning all the data issues identified using python in jupyter notebook.

After performing all the above processes, now the dataset is ready for Visualization Analysis.