# ANALYSIS AND VISUALIZATION

Project 4 Report

Dalal Bin Gheshiyan

## INTRODUCTION

Real-world data rarely come clean. Using Python and its libraries, we can do so. the aim of this project is to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. then communicates the insights and displays the visualizations produced from wrangled data.
Through this project, I used WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations, WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

## STEP 1: GATHERING DATA

to complete this project, I need to gather data from multiple sources with different data formats, which required different methods.

1. The WeRateDogs Twitter archive. The file was provided by the project and can be downloaded directly from the Udacity website, This archive contains basic tweet data (tweet ID, timestamp, text, etc.)
2. The tweet image predictions. The file is hosted on Udacity's servers. I downloaded this file programmatically by using the Requests library in Python, which is a table full of image predictions alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction.
3. Get retweets count and favorite count information missing from the Twitter archive from another file. I chose to download the tweet JSON file programmatically by using the Requests library, as Twitter decline my request for using their Tweepy - Twitter API-.

## STEP 2: ASSESSING DATA

After gathering all three pieces of data, I assessed data visually and programmatically for quality and tidiness issues.

### QUALITY:

issues with content. Low-quality data is also known as dirty data.

1. Completeness
2. Validity
3. Accuracy
4. Consistency

### TIDINESS:

issues with the structure that prevent easy analysis. Untidy data is also known as messy data.

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

## STEP 3: CLEANING DATA

Cleaning data is the third step in data wrangling. where I fixed the quality and tidiness issues that were identified in the assessment step, I used programmatic data cleaning methods.

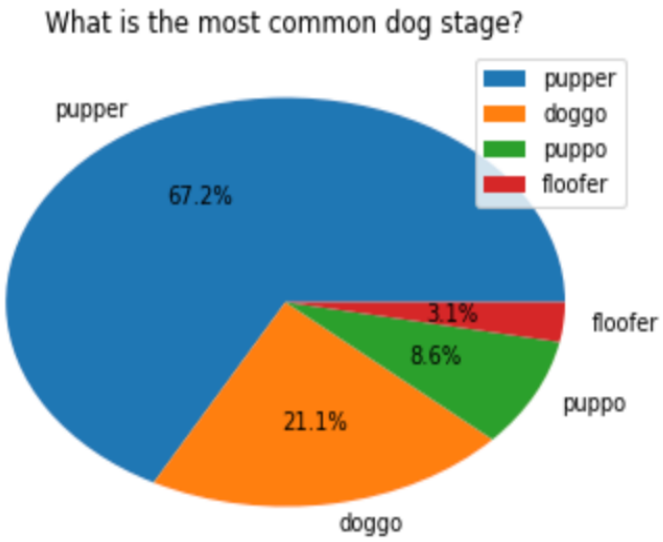
### THE PROGRAMMATIC DATA CLEANING PROCESS:

1. Define: convert our assessments into defined cleaning tasks.
2. Code: convert those definitions to code and run that code.
3. Test: test the dataset, visually or with code, to make sure the cleaning operations worked.

## STEP 3: ANALYSES AND VISUALIZATIONS

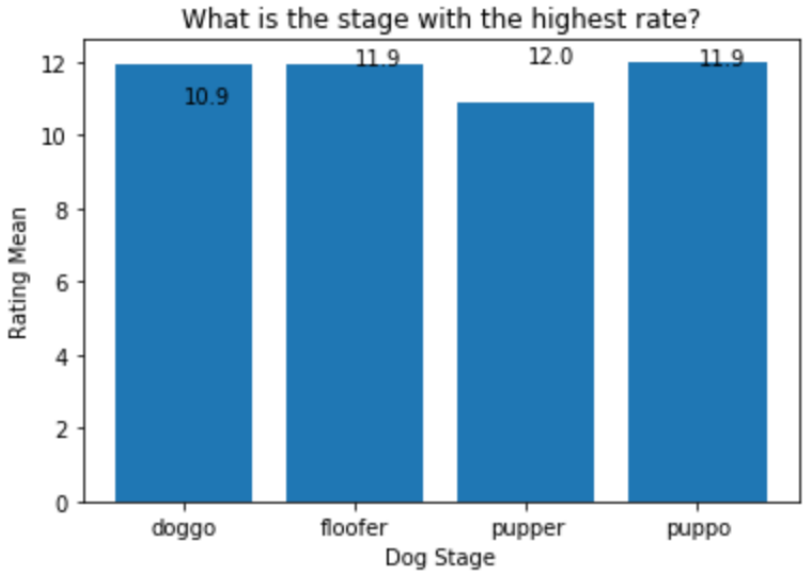
### INSIGHTS:

1. What is the most common dog stage?



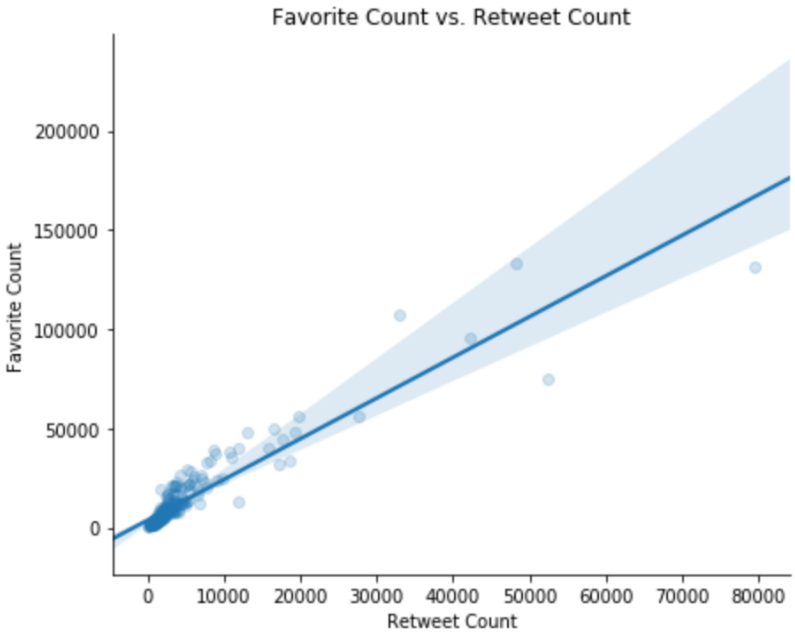

As illustrated in the above pie chart, pupper dogs take the first place in popularity on the dataset with a percentage of 67.2%, while doggo took the second place with 21.1%, and then puppo with 8.6%, and the least popularity for floofer with 3.1%.

2. What is the stage with the highest rate?




As illustrated in the above bar chart, pupper dogs take the first place in rating on the dataset with a mean of 12, while floofer and puppo took the second place with a mean of 11.9, and the least rating for doogoe with a mean of 10.9.

3. retweet counts, and favorite counts comparison over time.




Favorite counts are correlated with retweet counts - this is a positive correlation.