

# Data Wrangling Project Report — WeRateDogs®

As part of Udacity Data Analyst Nanodegree program, I will be working on a dog's dataset that is split in three different data-source formats.

For one of its wrangling projects, I was given an opportunity to go through the whole data analysis process — collecting the data, cleaning the data, analyzing the data and finally visualizing the data.

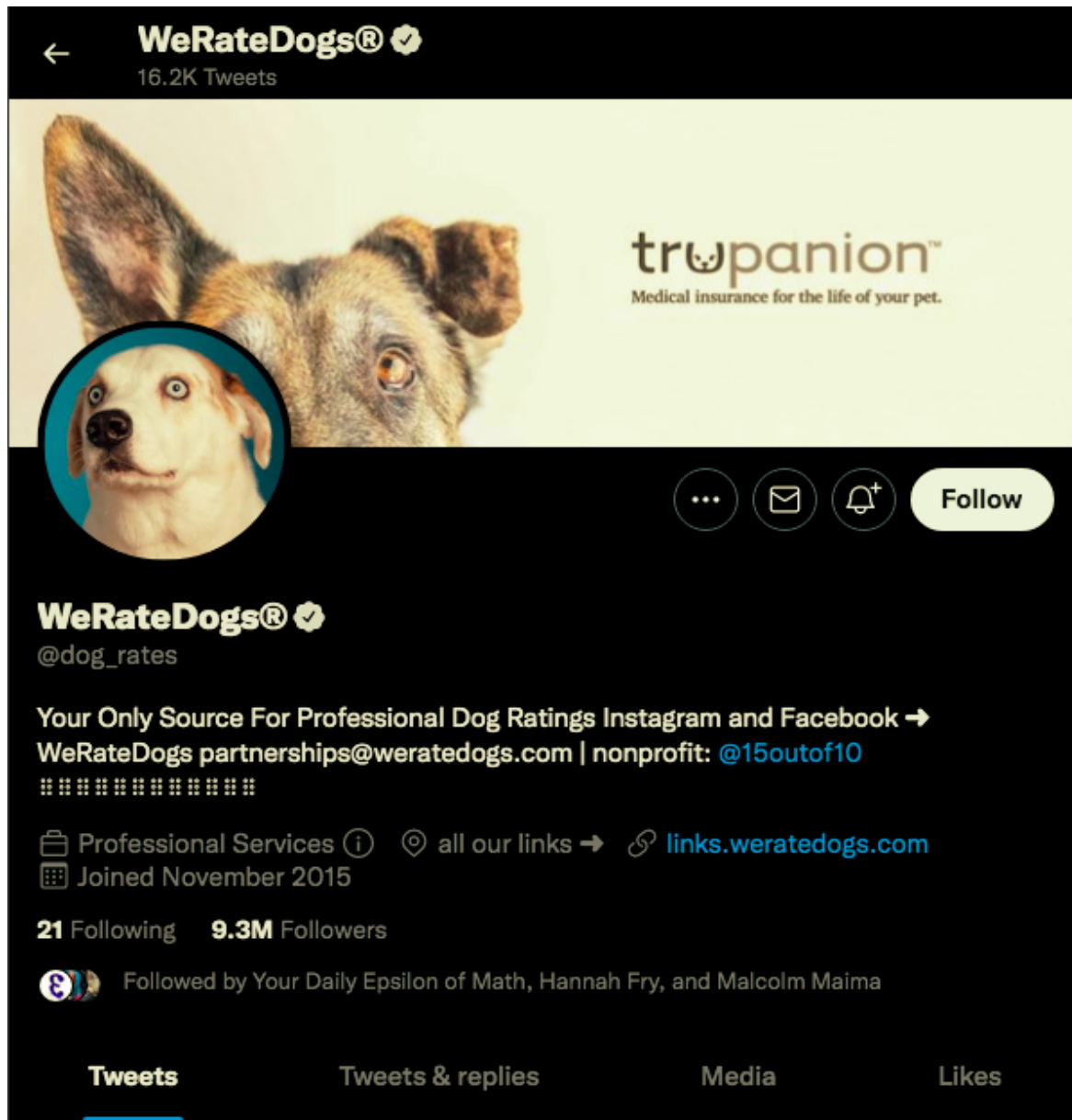
Working on this project will require gathering data from the three different data sources. After this, then go ahead and merge the three datasets for better and more comprehensive analysing of the data to answer a couple of questions later in the project.

The process is as follows:

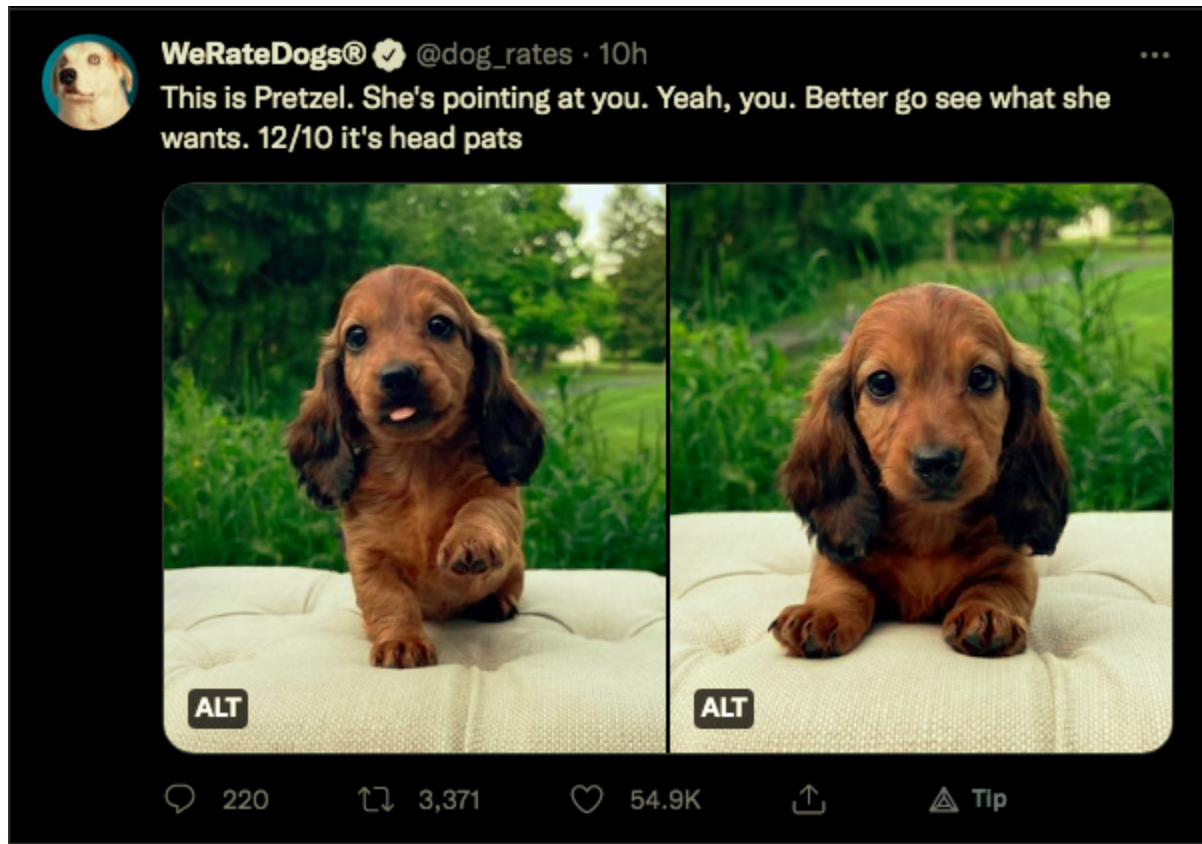
1. Data Gathering: The goal is to gather all the three pieces of datasets. The provided dataset in csv format, twitter dataset through querying Twitter API and finally pragmatically through requests of the image prediction dataset.
2. Assess: After gathering the data, assess it to check its quality and tidiness. This will be done both visually and programmatically for replicability. Any data that falls into the low quality and messy category is highly unwanted and effort must be put to clean it.
3. Clean: After noting what needs to be cleaned in the assessment phase, effort is applied in this phase to remove the unwanted characteristics whether by transformation or dropping data points
4. Analyze: After cleaning, the data is analyzed with the goal of answering posed questions or observations through charting or calculations.
5. Understand the limitations
6. Conclusions: Report findings found after analyzing the cleaned data.

Please note, the data will be assessed visually and programmatic to gain a better understanding of how the data looks.

On page 2 and 3 respectively, is a pictorial view of WeRateDogs Twitter profile and sample post on the same account.



WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dogs. WeRateDogs has over 9 million followers and has received international media coverage as at the time of writing this.



This is a sample of a tweet from the WeRateDogs Twitter account. As seen in the tweet, a few data points can be picked up. Here are a few of the data points:

1. Name of the dog: Pretzel
2. Rating: 12/10 with 12 as the numerator and 10 as the denominator
3. Image of the dog
4. Number of Retweets
5. Number of Likes
6. Number of Replies
7. Time of posting of the tweet

### Gathering Data:

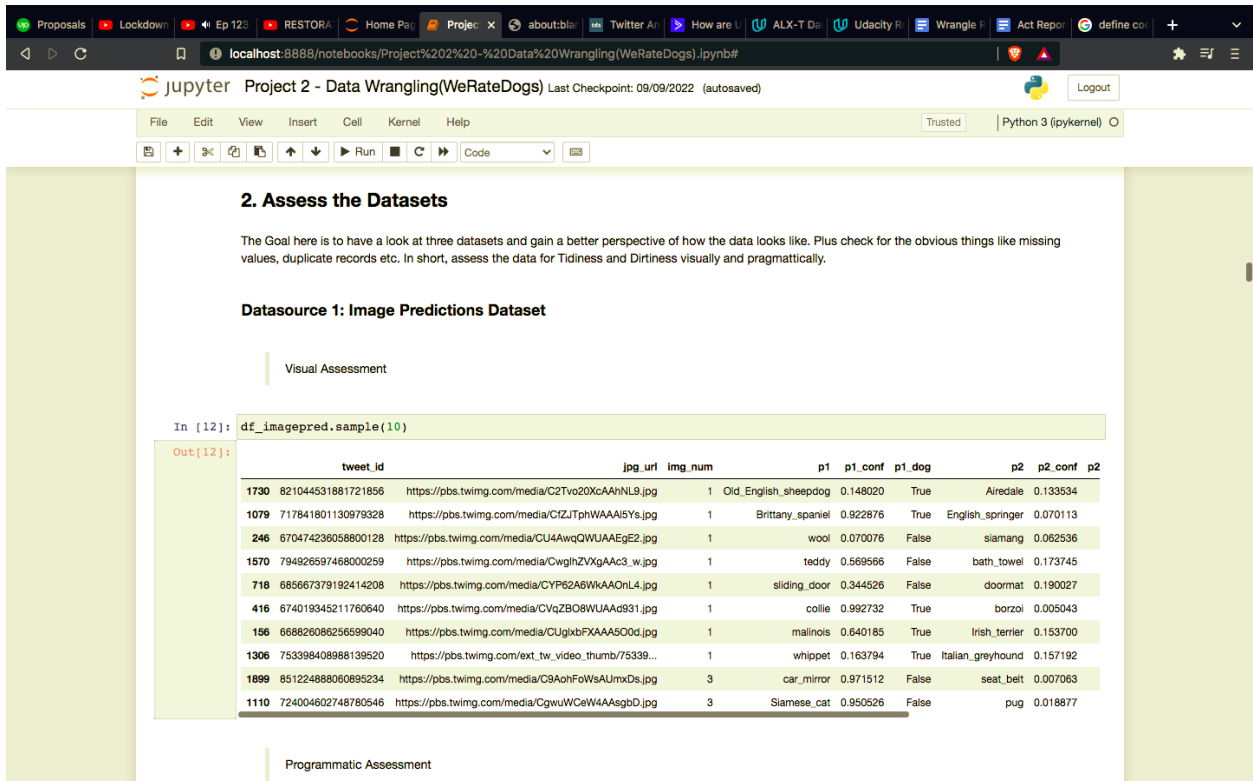
The dataset is split in three different data-source formats.

1. Twitter Archive File - WeRateDogs: Dataset provided by Udacity, manually downloaded from their servers.
2. Image Prediction File: Dataset provided by Udacity, programmatically downloaded from their servers.

- Twitter API - JSON File: Dataset extracted from Twitter account, WeRateDogs, by querying the Twitter API.

## Assessing Data:

After gathering data, assess it visually and programmatically as observed in the following images respectively.



**2. Assess the Datasets**

The Goal here is to have a look at three datasets and gain a better perspective of how the data looks like. Plus check for the obvious things like missing values, duplicate records etc. In short, assess the data for Tidiness and Dirtiness visually and pragmatically.

**Datasource 1: Image Predictions Dataset**

Visual Assessment

```
In [12]: df_imagepred.sample(10)
```

```
Out[12]:
```

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2
1730	821044531881721856	https://pbs.twimg.com/media/C2Tvo20XcAAhNL9.jpg	1	Old_English_sheepdog	0.148020	True	Airedale	0.133534	
1079	717841801130979328	https://pbs.twimg.com/media/CIZJTpHwAAAI5Ya.jpg	1	Brittany_spaniel	0.922876	True	English_springer	0.070113	
246	670474236058800128	https://pbs.twimg.com/media/CU4AwqQWUAAEgE2.jpg	1	wool	0.070076	False	siamang	0.062536	
1570	794926597468000259	https://pbs.twimg.com/media/CwglhZVgAc3_w.jpg	1	teddy	0.569566	False	bath_towel	0.173745	
718	685667379192414208	https://pbs.twimg.com/media/CYP62A6WkAAOnL4.jpg	1	sliding_door	0.344526	False	doormat	0.190027	
416	674019345211760640	https://pbs.twimg.com/media/CVqZB08WUAd931.jpg	1	collie	0.992732	True	borzoi	0.005043	
156	668626086256599040	https://pbs.twimg.com/media/CUgIxbFXAAAS00d.jpg	1	malinois	0.640185	True	Irish_terrier	0.153700	
1306	753398408988139520	https://pbs.twimg.com/ext_tw_video_thumb/75339...	1	whippet	0.163794	True	Italian_greyhound	0.157192	
1899	851224888060895234	https://pbs.twimg.com/media/C9AohFoWsAUmxDs.jpg	3	car_mirror	0.971512	False	seat_belt	0.007063	
1110	724004602748780546	https://pbs.twimg.com/media/CgwuWCeW4AAsgbD.jpg	3	Siamese_cat	0.950526	False	pug	0.018877	

Programmatic Assessment

#### Programmatic Assessment

```
In [13]: # Rows and Columns
df_imagepred.shape

Out[13]: (2075, 12)

In [14]: # Duplicate Records
df_imagepred.duplicated().sum()

Out[14]: 0

In [15]: # Missing Values/NaN
df_imagepred.isna().sum()

Out[15]: tweet_id    0
         jpg_url    0
         img_num    0
         p1         0
         p1_conf    0
         p1_dog     0
         p2         0
         p2_conf    0
         p2_dog     0
         p3         0
         p3_conf    0
         p3_dog     0
         dtype: int64

In [16]: # Assess Datatypes
df_imagepred.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
 #   Column      Non-Null Count  Dtype
```

The two main themes we are looking to assess is the quality and tidiness of data.

1. **Quality:** Commonly referred to as dirty data. Dirty data has issues with its content and dimensions to look at are Completeness, Validity, Accuracy and Consistency.
2. **Tidiness:** Untidy data is commonly referred to as “messy” data. Messy data has issues with its structure. Tidy data is where: each variable forms a column, each observation forms a row and each type of observational unit forms a table.

## Cleaning Data:

To improve the quality and tidiness, cleaning the data was necessary. Some of the issues that needed cleaning can be categorized as follows:

1. Dirty Data:
  - Incorrect Data Types
  - Some of the Columns have too many Nulls
  - Extra columns not useful for this particular Analysis
  - Completeness issues with Dog Names
  - Rating denominators lower than 10
  - Tweets that didn't match extracted tweets from API

- Missing Data/Incorrect Data
- Weird ratings

## 2. Messy Data

- The dog stage columns in twitter\_archive dataset can be arranged into a single column
- The three datasets(tables) should be merged into one dataset(table).