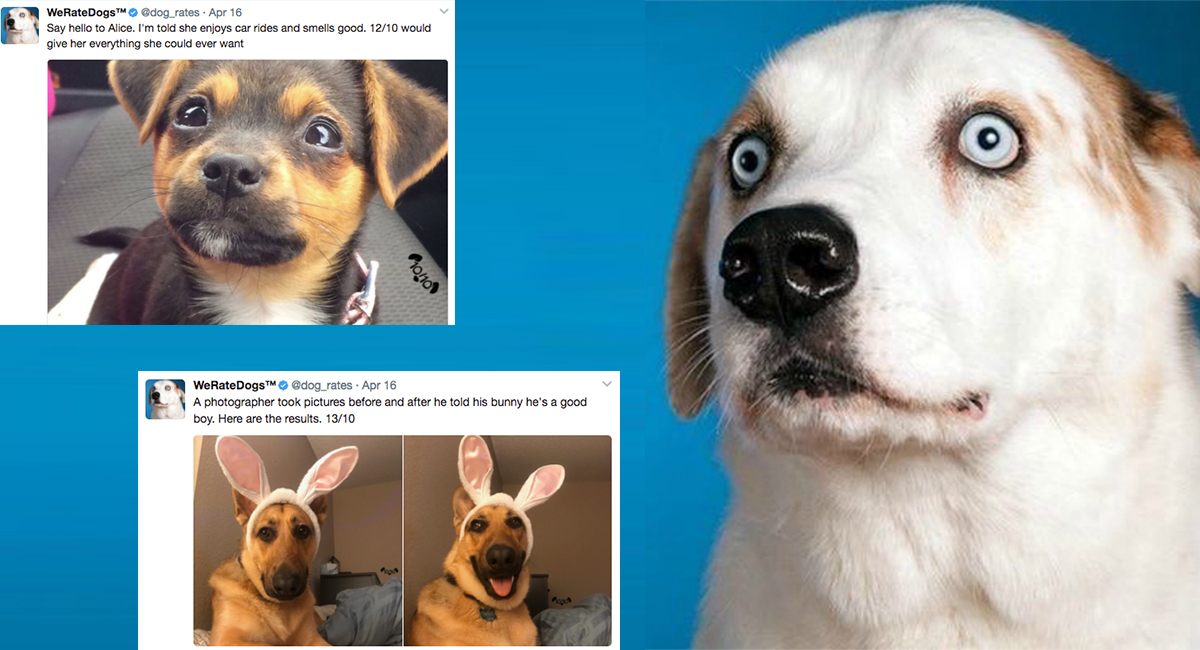
**WeRateDogs Data Wrangle Report**



**Introduction**

Data Wrangling is the process of cleaning and unifying the messy and complex data

for easy access and analysis. It is predominantly performed to bring consistency in

data which is sourced from various platforms. This process contains three main

divisions in it - data gathering, data accessing and data cleaning. This reports talks

about [WeRateDogs](https://twitter.com/dog_rates) twitter data wrangling process.

**Gathering**

For this project, data is collected from three different sources -

● Twitter\_enhanced\_archieve.csv - It contains data about each tweet like

ratings, text posted, url to the tweet, retweet info, dog’s stage, etc, which is

manually downloaded from the given link in the project specs.

● Image predictions data - This contains data about what kind of breed the dog

is in the images. This is retrieved using neural networks and so the data also

contains the confidence levels of how appropriate the results are from the

logic used to predict the right kind of breed. This data is hosted on Udacity

servers and is retrieved programmatically using python requests library.

● Favorite, Retweet count data - Number of likes and retweets of each tweet are

retrieved using Twitter API. Using the tweet\_id in twitter\_enhanced\_archieve,

an API call is made to fetch the data and is stored in a json file. Further, this

json file is read to extract the required data.

**Accessing**

Accessing data is followed by gathering process. After each piece of data is

collected, it must be accessed to find the inconsistencies between them.

Inconsistencies include wrong data types, missing values, inappropriate data, etc.

Data can be accessed visually or programmatically to define the issues and the

issues are categorized as Quality, Tidiness issues

Quality issues checks for completeness, accuracy, validity, consistency i.e., content

issues - includes missing values, wrong data types, invalid entries in the data, spell

checks, etc. Few quality issues in this project comprises of -

1. - Link in source column has some suffix and prefix in it

2. - Names of the dogs are wrong (a, an, the, this, etc)

3. - retweeted\_status\_timestamp and timestamp are not datetime objects

4. - retweets are included.

5. - number of observations(rows) is not consistent - (images\_df: 2075 instead of tweets\_df: 2356)

6. - null represented as 'None' in columns 'name', 'doggo', 'floofer', 'pupper','puppo' instead of NaN.

7. - lang(language) should be a categorical variable, read as string.

8. - in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id,

retweeted\_status\_user\_id should be str object

Tidiness issues checks for the structural patterns i.e., same number of observations

across each piece of data, merging two or more data sets if necessary, etc. Few

tidiness issues in this project are -

1. - stage (doggo, floofer, pupper, puppo) in 4 different columns

2. - All data is not present in one dataset.

**Cleaning**

After accessing data and defining the issues, cleaning comes into picture. It is where

we fix the quality and tidiness issues defined in the previous step. This can be done

manually or programmatically however, manual process is not recommended. Once

data cleaning is done, test the data to ensure the consistency. In this project, each

issue is taken one at a time, cleaned and tested.

**Conclusion**

It is important to perform data wrangling before analyzing and drawing conclusions

and one must be aware of the process because most of the world’s data isn’t clean.

If the data is analyzed without wrangling, the insights from the analysis might be

inappropriate and sometimes lead to disasters. Therefore, data wrangling is always a

good practice before analyzing the data.