**DAND Wrangle And Analyse Data Project**

**Wrangling Report**

**By Hamad Sami Al-Assafi**

Table of Contents

[**DAND Wrangle And Analyse Data Project** 1](#_Toc43474802)

[**Introduction:** 1](#_Toc43474805)

[**Gathering Data:** 2](#_Toc43474806)

[**Assessing Data:** 2](#_Toc43474807)

[Quality Issues: 3](#_Toc43474808)

[1-Missing values in (expanded\_url) column 3](#_Toc43474809)

[2-Incorrect dogs name 3](#_Toc43474810)

[3-dupliacted tweets 3](#_Toc43474811)

[4-Unwanted HTML tag should be removed in (source) column 3](#_Toc43474812)

[5-Data types problem (timestamp) should be converted to datatime, (img\_num) and (source) should be converted to category data type 3](#_Toc43474813)

[6-Fill the missing value with null values instead of (None) or other words 3](#_Toc43474814)

[Tidiness Issues: 3](#_Toc43474815)

[1-Convert the nominator rating and denominator rating to actual rating, and delete them from our dataframe 3](#_Toc43474816)

[2-Create single column that express the type of the dog instead of four columns for each type 3](#_Toc43474817)

[3-Merge the dataframes together, and create master dataframe 3](#_Toc43474818)

[**Cleaning Data:** 3](#_Toc43474819)

## **Introduction:**

**In this project I will take three steps to wrangle dataset from weRateDog account in twitter the three steps are:**

**1-Gathering Data: In this step I will gather data from multiple resources that i must have to complete this project**

**2-Assessing Data: In this step I will look through the data that i have gathered and see if there is any quality or tidiness issues in the dataset that i gathered if there are any issues, I will document it**

**3-Cleaning Data: The last step in wrangling process, in this step i will transfer the documentation that I have in the assessing step to code to clean our dataset, in this step I will use (Define, Code, Test) methodology**

**Then our dataset will become somehow clean! afterward I will go through the analysis process**

## **Gathering Data:**

In this step I have gathered datasets from three different sources:

1-From (CSV) file that contains archive tweets for weRateDog account, which was given from Udacity.

2-From (TSV) file that contains detailed data about the images in each tweet, I have extracted the file from the internet using Python and (Request) library, then I have installed the file in my local machine.

3-From (JSON) file that contains detailed data about the count of the favourite and retweets, I have extracted the file using Twitter API (tweepy), then I have installed the file in my local machine

## **Assessing Data:**

In this step, I have through the datasets to discover any quality or tidiness issues, either visually or programmatically. And I have discovered many quality or tidiness issues:

### Quality Issues:

### 1-Missing values in (expanded\_url) column

### 2-Incorrect dogs name

### 3-dupliacted tweets

### 4-Unwanted HTML tag should be removed in (source) column

### 5-Data types problem (timestamp) should be converted to datatime, (img\_num) and (source) should be converted to category data type

### 6-Fill the missing value with null values instead of (None) or other words

### Tidiness Issues:

### 1-Convert the nominator rating and denominator rating to actual rating, and delete them from our dataframe

### 2-Create single column that express the type of the dog instead of four columns for each type

### 3-Merge the dataframes together, and create master dataframe

## **Cleaning Data:**

In the cleaning process I have cleaned the missing values firstly, because maybe some of the other issues will disappear when I deal with the missing values, then I have cleaned the other quality issues that I have mentioned them above. In this process I have used the (Define, Code, Test) methodology for each issues to make the cleaning process more organized.

In the other report I will go through the analysis process. Thanks !