Wrangle Report:  
by Salma Hasan Abokamar

1. Introduction

This paper will describe the efforts which made for the Udacity Data Analyst Nanodegree Program of

project “WeRateDogs”

The report will have to following structure:

- Gathering Data

- Assessing Data

- Cleaning Data

2. Gathering Data

The data for this project came from three different sources:

- Original twitter archive data: downloaded from the udacity project details site and uploaded

into the udacity project workspace

- Predictions data: programmatically downloaded from the udacity server

- Twitter data: obtained from the Twitter API using Tweepy

3. Assessing

After gathering the data from the individual sources, the next step was to assess this data visually

and programmatically. Following quality and tidiness issues were detected:

**Quality**:

1.The timestamp column has dates in string form(object).

2.the columns doggo, floofer, pupper and puppo could be collapsed into a column

3.archive: column "expanded URLs" has missing values, duplicates and unnecessary links

4.not all dog names are filled

5.Some of the rows from the tail() output above have invalid strings in the name column "a"

6.Missing data in name and stages columns showing as 'None'

 7.Nondescriptive column headers ("p1" and "p1\_conf")

8.non-numeric values for the "tweet\_id" inputs which will need to be removed after merging data

**tidiness:**  
 1.day,month,year are in one column

2.there are unwanted columns (retweeted\_status\_timestamp, retweeted\_status\_user\_id, retweeted\_status\_id )

3.The column "jpg\_url" will be removed since url data is already contained in the twitter enhanced data

4. Cleaning

Cleaning the data is the third step in data wrangling. Following quality and tidiness issues were

cleaned:

1. The "p1" and "p1\_conf" columns will be renamed with more explanatory titles.

2. Finding non-numeric values for "tweet\_id"

3.merging the 3 dataframes into one. And then we will have only 1 dataframe to clean and process.

4.Merge the four dog type columns "puppo", "pupper", "floofer" and "doggo" into one column

5.Clean data by dropping duplicates row and not meaningful columns.

# Delete the retweets

# Delete duplicated tweet\_id

# Delete tweets with no pictures

#Delete dog\_stage duplicates

6.remove unwanted columns (retweeted\_status\_timestamp, retweeted\_status\_user\_id, retweeted\_status\_id,... )  
  
7.correct all datatypes by changing the datatypes of the columns.

8.Separate timestamp into day - month - year (3 columns)

9.drop timestamp column

10.Remove those missing values in expanded\_urls column by .dropna

11.Turn p1, p2, p3: dog breed names are not all in lowercase

12. column has some missing values and some of the names are not real dog names but articles or adjectives.

#Replace those not-real-dog-names with None

5. Conclusion

Since there is a lot of unclean data around the world. Data wrangling is a skill every data analyst

should be familiar with. After the gathering, assessing and cleaning part of the data, there is one last

step to come. The results need to be analysed and visualized to create better insights about the data.