Wrangle Report

1. Introduction

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

(WeRate Dogs)

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 collecting data from individual sources, the next step was the visual and programmatic evaluation of this data

Actual visual evaluation time in Excel. The following quality and cleaning issues have been detected:

Archive file

The archive column “expanded URLs” had empty rows without links. There are also a lot of duplicated twitter links and links of other sources.

The "name" column of the archive table isn’t always filled. A lot of those names are stored in the "text" column.

The “doggo”, “floofer”, “pupper” and “puppo” column stores the same data. These columns should be merged into one single column.

The "text" column includes more than just the photo description. It also includes the rating, dog names, IG names and short URLS.

4. Cleaning

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

Tidiness issues

- *Column headers are values, not variables (doggo - floofer - pupper - puppo)*
- *Information about one type of observational unit (tweets) is spread across three different files/dataframes. So these three dataframes should be merged as they are part of the same observational unit*

Quality issues

Change the names of columns type (p1, p2 and p3) to a more descriptive name

The P1 confidence interval might not be always convincing. I will set the acceptance threshold at P ≥ 0.5. This might mean neglecting a lot of the data (the median is .585).However, the trade off is higher confidence in the results

Null values are expressed by None

You only want original dog ratings that have images (a user can retweet their on tweet)

There are some dog names in lower case letters. I think they are not dogs

erroneou datetype (timestamp column) and tweet_id

For the column "rating_numirator" there are zero values. delete these two columns

For the column "rating_numirator" there are extreme values

Delete the cells with numerator more than 20. This will not affect the data

For the column "rating_denumirator" there are extreme values

For the column "rating_denumirator" there is a zero value. delete it

Drop unneeded columns

There are some strange values in the p1 columns such as "school_bus, pillow, cartoon" these cells must be checked first then deleted if not needed

271 rows (frequency less than 3) could be deleted because they are not that frequent and most probably not dogs

erroneou data type (tweet_id)

Remove duplicate tweets "retweat_count"

5. Storing Data

Store the clean df in CSV file with name using .to_csv('Twitter_archive_master.csv')

6. 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.