Wrangle Report

Introduction :

This document is for documenting the data wrangling process which is not an easy one but thanks to the course and this project now it’s no problem

What are the steps of data wrangling?

The steps of data wrangling are :

1-Gathering data

2-Assessing the gathered data

3-Cleaning the data according to the notes taken

Gathering

There are 3 files to be gathered

1-twitter archive file: a csv file provided just downloaded it from the project page this file contains id of all tweets needed for the project time of the tweet and ratings and names for the dogs.

2-Image prediction file: a tsv file format and it had to be downloaded via a url downloaded programmatically this file contains photos url of dogs prediction of the dog types and confidence levels for these predictions.

3-tweet json file: a txt file format but it’s contents are json this file is acquired via scraping the twitter api this file contains a lot of info but the most important is that tweets id , the tweet is original or a retweet, number of likes and retweets

Assessing

Assessing the data is not that hard there are two ways to assess the data and after assessing take notes to change what you noticed

1-visually:
Simply taking a good look at the data by displaying with the head and tail notice if there is anything weird.

2-programmatically:
Assessing the data with code is as simple as using the (info, value_counts, duplicated, query, groupby)and many more but these are the ones i used

Cleaning

Quality issues:
-First need to make a copy of the data set before cleaning it to prevent any problems if anything wrong happened while cleaning not having a copy of the data may result in permanent data corruption

-In this stage after the data is assessed all we have to do is clean the data it can be removing unnecessary columns or rows like retweets id and many more

-Dropping empty or wrong values or other values that are corrupted as i dropped the retweets in the csv and json tables

-Correcting wrong values that we know their actual value through assessing like the wrong denominator and numerator values or the ones that were taken by mistake as a rating

Tidiness issues:

Tidiness issues are not that hard to deal with it’s only about

1-each variable is a column.

2-each observation is a row.

3-each type of observational unit forms a table.

So it’s all about merging tables and forming a tidy dataset