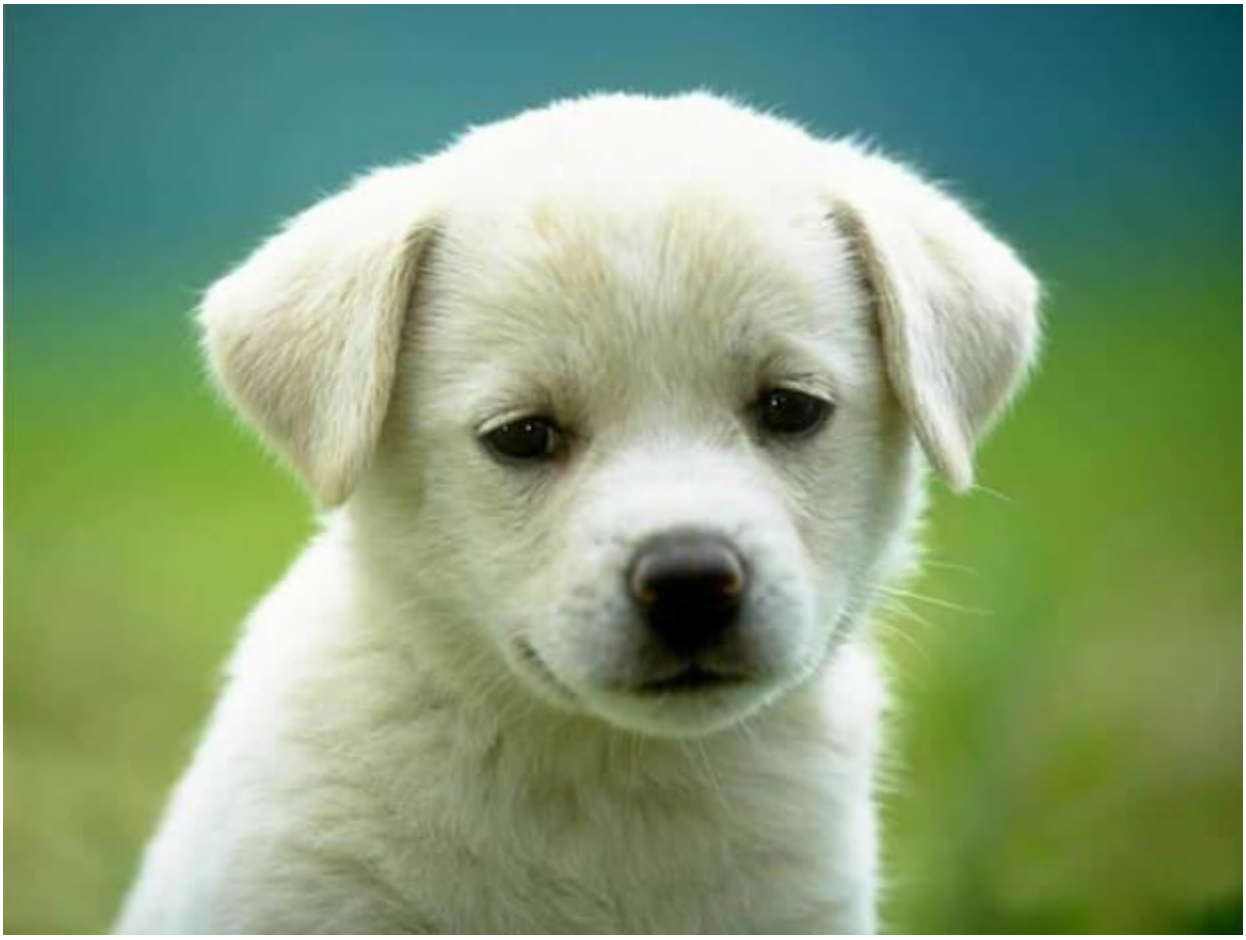
Data Wrangling Report

By Alaa Nagy

December 2020



Introduction

The main purpose of this project is wrangling data by applying the three main stages of wrangling process which are gathering, assessment and cleaning.

After that, we extract some insights and produce visualization from our cleaning data.

First stage: Gathering data

The data for this project consists of three parts as following:

- The twitter_archive_enhanced.csv file which is given and all we do is to download it manually.
- The tweet image prediction which contains the breed of each image using neural network, we are given a link to download this file which is (image_predictions.tsv) from the internet using Requests library and the given URL.
- Twitter API and JSON file, Using the tweet IDs in the WeRateDogs Twitter archive, we query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet.json.txt ,we read this .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count columns.

Second stage: Assessing Data

I use two methods for assessing data, visually and programmatically

- Visually by printing out the content of data frame using df.head() and df.sample() methods
- Programmatically by using some methods from pandas library such as df.info(), df.describe(),df.duplicated(), df.value_counts and others.

After that , I make an assessment summary consists of two parts as following:

- Quality issues which are issues with content. Low quality data is also known as dirty data.
- Tidiness issues which are issues with structure that prevent easy analysis. Untidy data is also known as messy data. Tidy data requirements:
    - Each variable forms a column.
    - Each observation forms a row.
    - Each type of observational unit forms a table.

Third stage: Cleaning Data

First of all, I make a copy of the data in order to use it instead the original data
This part consists of three main parts which are define, code, and test.

1. Define: I convert the assessments into defined cleaning tasks. These definitions also serve as an instruction list so others (or me in the future) can look at my work and reproduce it.
2. Code: I convert those definitions to code and run that code.
3. Test: I test my dataset, visually or with code, to make sure that the cleaning operations worked.

After the cleaning process, I produce some visualizations to extract insights from the cleaned data.