Data Wrangling Report

Udacity Project: Wrangle and Analyse Data

Introduction:

I completed this project as part of Udacity's Data Analyst Nanodegree. The project is based around wrangling data of the various tweets from the twitter user @dog_rates also known as the "WeRateDogs" Twitter page, a page which will kindly rate pictures and videos of dogs out of ten. Since dogs are all round fantastic creatures, all of WeRateDogs' ratings are above ten. They also tag each dog with a different dog category out of "doggo", "floofer", "pupper", or "puppo".I approached this project using the three steps of data wrangling: gather, assess, clean.

GATHER:

In the gather phase, data was downloaded from 3 different sources namely:

1.In the gather phase, the enhanced twitter archive csv file was available for manual download and had various columns like tweet-id, rating numerator and denominator, timestamp and other retweet columns.

2.Secondly ,additional twitter information (i.e retweet and favorite counts) was downloaded using the twitter API through a json text file.

3. Thirdly, image predictions file was downloaded programmatically using the python's requests library.

After the data was gathered, I then assessed the dataframes in order to find any quality or tidiness issues.

ASSESS:

The data was assessed using the following functions for each of the three dataframes:

1. .info()
2. .sample()
3. .value_counts()

4. .duplicated()
5. .isnull()
6. .head()
7. .tail()

Soon after assessing the data, the cleaning step subsequently involved implementing steps to fix the quality and tidiness issues.

CLEAN:

The following quality and tidiness issues were looking into:

Quality issues:

1.Finding incorrect rating_numerator values.

2. Name column had names like ‘a’,’an’,none’ etc which were all replaced to NaN values.

3.Unwanted columns were removed.

4.timestamp column was changed to datetime object.

5.rating_denominator had values which were not equal to 10, those rows were removed.

6.Changing the name and datatype of the ‘id ’ column for merging.

7.Removing all the retweet columns that had NaN values.

8.Removing all the retweets that were without images by checking the ‘jpg_url’ column.

TIDINESS:

1. Combining all three datasets into a single master dataset as they all had data which were related to each other.

2.Combining the different dog categories like doggo,puppo,floofer and pupper into a single column dog_category.

Following the data wrangling process, some exploration and analysis of the now clean and tidy data, was carried out and results were observed for the same.The project emphasized the need to gather data into a single place from various sources, inspecting the data to find out the various quality and tidiness issues and ultimately cleaning the data to give us final results.