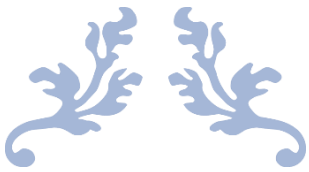

WRANGLE REPORT

Wrangle and Analyze Data

Udacity Project

Submitted By: Manal Alzeer

Introduction

The dataset that has been wrangled in this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.
Our goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations , that included:

- Data wrangling, which consists of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing your wrangled data.
- Reporting on the data wrangling efforts and data analyses and visualizations.

Gathering Data :

Data was gathered from 3 different sources:

1. **Enhanced Twitter Archive**

   The WeRateDogs Twitter archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets.

2. **Image Predictions File**

   The tweet image predictions file was provided to Udacity students.

3. **Tweet Json file**

   Additional data, including favorite count and retweet count, were gathered using the Twitter API.

   All the three files were provided from Udacity to student.

Assessing Data :

After gathering had assessed them visually and programmatically for quality and tidiness issues in file wrangle_act.ipynb.

Tidiness Issues that were cleaned:

- In twitter_archive:(dogger, floofer, pupper and puppo)columns relate to the same variable dog "stage".
- In teitter_archive calulate 'rating_numerator' and 'rating_denominator' and save it in one column.
- The tweet.json and image_predictions should be joined to twitter_archive DataFrame since they are having same columns. And drop unneeded columns.

Quality Issues that were cleaned:

- **twitter_archive:**

1. Missing data in columns(expanded_urls).
2. Delete unwonted columns (in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp).
3. The timestamp should be a datetime.
4. Some dogs that have 'a' , 'the' as a name.
5. There are some doges have names that start with a lowercase letter.
6. There is a 745 missing values in name column referred as 'None'.
7. The source column data should be extracted from the html code, like (iPhone , vine , twitterWeb ,TweetDeck), Also it's the same source as the tweet.json.
8. 1976 are not on any dog "stage", and we can note 1976 + 394 = 2370 is more than our dataset ,that mean there is some dogs in many "stages" at the same time.
9. rating_numerator and rating_denominator should be a float.

- **image_predictions:**

1. In p1, p2, and p3 columns it's contained underscores instead of spaces.
2. Delete unnecessary column (img_num).

- **Tweet json:**

1. There is a lot of the variables have missing data should be deleted.

Cleaning Data :

After the assessment, I fixed and cleaned the data the following means: Define , Code , Test.

In first, copies of the DataFrames were created before cleaning. Then, the steps of cleaning were applied iteratively on all issues. And in last step After cleaning all issues ,marge three DataFrames in one DataFrame to Analysis & Visualization