**Wrangle And Analyze Data Project**

**First – Importing Useful Packages.**

pandas – numpy – requests – tweepy - os - json - time - matplotlib.pyplot - warnings - IPython - re - seaborn - datetime - PIL - glob - matplotlib inline

**Second – Gathering Files And Creating Dataframes.**

1. Downloaded Twitter Achieve CSV file
2. Downloaded Image Predictions file by coding from Udacity url
3. Downloaded Twitter API file from Udacity because I can not get my Twitter developer credentials till now, that’s why I can not get a file from Twitter directly.

**Third – Assessing The Collected Data.**

**Assessment Summary in points.**

Completeness, Validity, Accuracy and consistency

**Twitter Archive Data frame (df\_twitter\_arch)**

**Quality**

1. timestamp column is str instead of datetime
2. expanded\_urls has Nan values should be replaced
3. The rating\_numerator column should of type float and also has invalid values it should be correctly extracted from text column.
4. The rating\_denominator column has invalid values should be removed.
5. There are retweets should be removed.
6. There are invalid names in name column should be removed.
7. There are row has not data for dogs should be removed.
8. Source has a link should be extracted to get only source and should be changed to category.
9. (doggo, floofer, puppe and puppo) Columns has none or missing values.
10. Removing all retweet columns.

**Tidiness**

1. doggo, floofer, puppe and puppo columns should be merged in one column dog\_stage

**Image Predictions Data frame (df\_img\_pred)**

**Quality**

1. There is 66 duplicated url in jpg\_url should be removed.
2. img\_num is usless and should be removed and all unnecessary columns.

**Tidiness**

1. p1\_conf, p2\_conf and p3\_cong columns should be merged in one column confidence and p1\_dog, p2\_dog and p3\_dog columns should be merged in one column dog\_type.

**Twitter API Data frame (df\_api)**

**Quality**

1. id column should be renamed to tweet\_it .
2. All columns are usless expet (id, favorite\_count and retweet\_count) and should be removed.

**Fourth – Cleaning The Collected Data.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Quality** | | | |
| Dataframe |  | Points | Solution |
| df\_twitter\_arch | 1 | timestamp column is str instead of datetime | Type converted using pandas to\_datetime function |
| 2 | expanded\_urls has Nan values should be replaced | Nan Values has been replaced with empty |
| 3 | The rating\_numerator column should of type float and also has invalid values it should be correctly extracted from text column. | Extracted from text column and added , Type has been changed to float |
| 4 | The rating\_denominator column has invalid values should be removed. | All rows has value more than the expected has been removed |
| 5 | There are retweets should be removed. | All Retweets has been removed |
| 6 | There are invalid names in name column should be removed. | Some names was writen wrong has been solved and the other invalid has been removed |
| 7 | There are row has not data for dogs should be removed. | All Rows exept dogs has been removed |
| 8 | Source has a link should be extracted to get only source and should be changed to category. | Source has been extract |
| 9 | (doggo, floofer, puppe and puppo) Columns has none or missing values. | Replaced with np.non |
| 10 | Removing all retweet columns. | All Retweet columns has been removed |
| df\_img\_pred | 11 | There is 66 duplicated url in jpg\_url should be removed. | Duplicated values has been removed |
| 12 | img\_num is usless and should be removed and all unnecessary columns. | All usless columns has been removed |
| df\_api | 13 | id column should be renamed to tweet\_it | Column name has been changed to the correct one |
| 14 | All columns are usless expet (id, favorite\_count and retweet\_count) and should be removed | All usless columns has been removed |
| Tidiness | | | |
| Dataframe |  | Points | Solution |
| df\_twitter\_arch | 1 | doggo, floofer, puppe and puppo columns should be merged in one column dog\_stage | Columns has been merged to dog\_stage |
| df\_img\_pred | 2 | p1\_conf, p2\_conf and p3\_cong columns should be merged in one column confidence and p1\_dog, p2\_dog and p3\_dog columns should be merged in one column dog\_type. | All Conf cloumns has been merged and all dog columns also |

**Fifth – Merging All Data Frames.**

**Sixth – Reporting And Visualization.**

**By : Mahmoud Hassan Khalil El-Tobgy**