We rate dog’s data quality issues

Merged Data

1. There are 3 separate tables.

* We fix this by merging our data into one table using “twitter\_id”. --Done

1. Timestamp datatype values are typed as object. --Done

* We change this to datetime.

1. Re-tweets are included in the dataset these are identified by “RT @”

* We query the text field in our merged data for “RT @” and drop those columns. --Done

1. Replies are included in the dataset. -- Done

* Identified in “reply\_to\_status\_id” Query the table if “reply\_to\_status\_id”not null then drop rows.

1. Dog classification name values are “NONE” should be NaN. -- Done

* We run a for loop to identify the instance of “NONE” and we change to NaN.

1. Redundant columns (doggo, puppo etc). -- Done

* We Place the values into a new one dimensional column called “dog\_type” and drop redundant columns.

1. Missing fields for newly created dog type field. – done maybe

* This is not worth fixing, noticing that we may build inaccurate data. We observe that the name could potentially be contained within the text field and mined. However the risks of building false data are too high.

1. Some names are missing in the name column. -- Done

* We fix this by running a for loop and populating our name column based on observable criteria.

1. Only tweets with images

* We query “jpg\_url” in our merged data and exclude the null fields. --Done

Predictions

1. Not very human readable columns – Change and make understandable

Tidiness

1. Drop unnecessary tables
2. Fix column headings to make everything more readable to a human
3. Lower case everything.

|  |  |  |
| --- | --- | --- |
|  | There are 3 separate tables. | We fix this by merging our data into one table using “twitter\_id”. |
|  | Timestamp datatype values are typed as object. | We change this to datetime. |
|  | Re-tweets are included in the dataset these are identified by “RT @” | We query the text field in our merged data for “RT @” and drop those columns. |
|  | Replies are included in the dataset. | Identified in “reply\_to\_status\_id” Query the table if “reply\_to\_status\_id”not null then drop rows. |
|  | Dog classification name values are “NONE” should be NaN. | We run a for loop to identify the instance of “NONE” and we change to NaN. |
|  | Redundant columns (doggo, puppo etc). | We Place the values into a new one dimensional column called “dog\_type” and drop redundant columns. |
|  | Missing fields for newly created dog type field. | This is not worth fixing, noticing that we may build inaccurate data. We observe that the name could potentially be contained within the text field and mined. However the risks of building false data are too high. |
|  | Some names are missing in the name column. | We fix this by running a for loop and populating our name column based on observable criteria. |
|  | Only tweets with images | We query “jpg\_url” in our merged data and exclude the null fields |

***Merged Data Quality Issues***

*After gathering the data we merge it into one dataset using outer join based on "Tweet\_id" as to not lose data.*

* First we check our separate dataset shapes.
* Then we merge using an outer join.