Project 4 [Wrangle Data] Report

Twitter doesn’t agree to give the required data, so the data attached to this project has been used instead.

Data Attached:

- Twitter-archive-enhanced.csv
- Image-Predection-3.tsv
- Tweet-Json copy.txt

First: CSV File

1- Gathering Data

Placing the data in an csv file

2- Assess Data.

Tidiness: Columns where two or more variables are stored.

Quality: Removing Duplicates and Missing Values.

3- Clean Data.

3 steps: define, code and test.

I. Renaming column "tweet_id" to "ID"
II. Merging 4 columns "doggo, floofer, pupper, puppo" into one column named "Type_of_Object".
- doggo to 1.
- floofer to 2.
- pupper to 3.
- puppo to 4.
- and 0 if none of the above.

III. Replacing "Dates, Time" with column "Hour" for use in Data Analysis.
IV. Deleting all column that we don’t need which are "timestamp, rating_numerator, rating_denominator, name".

Second: TSV File

1- Gathering Data.

Placing the data in an csv file

2- Assess Data.

Quality: Removing Duplicates and Missing Values.

3- Clean Data.

3 steps: define, code and test.

I. Replacing True and False to 1 and 0 in p1_dog column.
II. Renaming column "tweet_id" to "ID" and column "p1_dog" to "Is_It_Dog?".
III. Deleting all column that we don’t need which are "jpg_url, img_num, p1, p2, p3, p1_conf, p2_conf, p3_conf, p2_dog, p3_dog".

Third: TXT File

1- Gathering Data.

Placing the data in an csv file

2- Assess Data.

Tidiness: Removing Duplicates and Missing Values.

Quality: Removing Duplicates, Missing Values and Incorrect Data.

3- Clean Data.

3 steps: define, code and test.

I. Renaming column "id" to "ID" and column "retweet_count" to "Num_of_Retweet" and column "favorite_count" to "Num_of_Favorite".
II. Replacing "Dates, Time" with "Hour" for use in Data Analysis.
III. Deleting all column that we don’t need which are "created_at, id_str, truncated, display_text_range, entities, source, is_quote_status, favorited, retweeted, lang".

Fourth: The Merge Table

Assess Data

Quality: Removing Duplicates, Missing Values and Incorrect Data.

Fifth: Data storage

We stored the data in CSV file named "Table_Copy".

Sixth: Analysis & Visualization

1- Are most of the pictures of dogs?

Yes, dogs represent seventy-four percent 74% of the pictures.

2- What type of dog is the most seen in pictures?

The pupper dogs are the most seen with a percentage of ten 10%.

3-Is there a relationship between Num_of_Retweet and Num_of_Favorite?

The relationship between them is positive and strong.

4-What is the most time you post tweets?

From the Pie graph, we conclude that most tweets are posted between 12 am and 2 am.