Report

Data Wrangling Steps

Created By:
Sultanah Aldossari
16 February 2022

About

The dataset provided by Udacity is the tweet archive of twitter user known as WeRateDogs or @dog_rates. WeRateDogs is a twitter account that rates people's dogs.

Project Aim

This project aims at wrangle the twitter data through;

- Gathering Data
- Assessing Data
- Cleaning Data

And then analyze and visualize the wrangled data. Finally, report on the data wrangling project and data analysis.

Data Wrangling process:

1. Gathering Data

During this step, it was required to gather data from various sources.

- WeRateDogs twitter archived. The first dataset, it was a CSV file provided by Udacity. The file contains more than 5000 tweets.
- Tweet Image Prediction, This dataset was scraped from a link provided by Udacity and then saved as a tsp file.
- Gather tweets using Twitter API and tweeps library. At this step I have created a twitter developer account. After approval, tokens were provided. Finally I have used the keys to scrape data from WeRateDogs account. The data contains the retweets count and like counts.

2. Assessing Data

After gathering data, assessing data step begins. At this phase I assessed data on both the quality and tidiness of our datasets.

1. Quality Issues:

- Remove columns that are not needed
- Change Tweet_id from int to string
- Change Timestamp into date time format
- Change Name into string type
- Delete retweets info
- Drop columns with missing values
- Feature engineering: extracting rating column from (rating_numerator, rat- ing_denominator)

2. Tidiness Issues:

- Doggo, floofer, pupper and puppo should be combined under a variable named Dog Type
- Merge the three dataframes into one

3. Cleaning Data

Using my assessment I pursued on cleaning step. At this step:

- I have created a copy of each dataframe, which in this case 3 dataframes
- Merged the copied dataframe into one.
- Created a new column called dog_type, it contains various dog types like: doggo, floofer, pupper and puppo.
- Drop unwanted columns and also columns with missing values.
- Changed Tweet_id from int to string
- Changed Name into string type
- Changed timestamp into date time format
- Extracted from the dataset a new feature which is dog ratings.