Project: Wrangle and Analyze Data

Introduction

This project is about wrangling data for the Wrangle and Analyze Project The document is divided in 3 sections, each for one step of the data wrangling process that we saw in the lessons

Context

Our goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

Data Wrangling

Gather first we need to gather the data from at least 3 different sources. Some data were provides and other were obtained

- first load all the libraries needed
- Now load the twitter_archive_enhanced.csv file into the dataframe
- Now, we retrieve image-predictions.tsv file
- Now we retrieve the image_predictions.tsv file into its dataframe

Assess

Assessing our data is the second step in data wrangling. When assessing, we are like a detective at work, inspecting our dataset for two things: data quality issues (i.e. content issues) and lack of tidiness (i.e. structural issues).

Quality issues are issues with content, like inaccurate or duplicate data. Tidiness issues are structural issues, specifically: each variable must be a column, each observation must be a row, and each type of observational unit must be a table.

We will use two steps: Visual Assessment and Programmatic Assessment

Summary of Quality and Tidiness Issues

Quality

tweet_data dataframe

1. Name column, have "None" as name. This can be confusing
2. There is missing data in "name" columns and is written as the string "None"
3. Missing "expanded URLS", there are 2297 instead of 2356
4. retweeted_status_id need to be "int64"
5. timestamp is better to be "date_time" objet type
6. there are retweet data, it shouldn´t
7. There is a minimum value of 0 in the rating_denominator column
8. The maximun value is higher for both numerator and denominator

prediction_data dataframe

3. In p1, p2, and p3 the type of dogs uses underscores that are diffucult to read
4. There are 2075 entries and in the tweet_data dataframe there are 2376 rows, so there is missing data

<u>Tidiness</u>

tweet_data dataframe

1. The columns "doggo", "floofer", "pupper" and "puppo" represent the same variable "type"

prediction_data dataframe

2. Correct predictions of dogs should be combined with tweet_data

3. Clean

Cleaning our data is the third step in data wrangling. It is where we fix the quality and tidiness issues that we identified in the assess step

We will use the data cleaning process: defining, coding, and testing

- We will address the missing data first

- We will tackle the tidiness issues next

- And finally, we will clean up the quality issues

The very first thing to do before any cleaning occurs is to make a copy of each piece of data. All of the cleaning operations will be conducted on this copy so we can still view the original dirty and/or messy dataset later

Some of the cleanings are:

Quality;

5. timestamp is better to be "date_time" objet type
   Define Convert the timestamp column into datetime

7. there are retweet data, it shouldn´t

   Define Remove all rows where retweeted_status_id. Afterwards, reset the index.

1. Name column, have "None" as name. This can be confusing

Define We will replace the string "None" with "NaN"

3. Storing and Acting on Wrangling Data

Storing

# save our dataframe in the file specified

# save our dataframe in the file specified

Analysis and Insights 3 insights and 1 labeled visualization

Question: Which are the most common scores on the tweets

Insight: The most common scores are around 10/10 +1 or -1