Wrangling Report For The We Rate Dogs Analysis

Table of content

Introduction

This report is a documentation of the wrangling efforts done during the analysis of We rate dogs twitter page (@dog_rates), the wrangling effort includes gathering, assessing and cleaning data which were done in order.

Gathering data

There are three different datasets given for the purpose of this project

1. **The tweets archive dataset:**

   This was provided as a file by Udacity, it was downloaded and loaded into the Jupyter notebook using the read_csv function of the panda library.

2. **The image predictions dataset:**

   This data is hosted on udacity's server and was downloaded programmatically using the r library, it was saved as image_predictions.tsv and was loaded into the jupyter notebook using pandas read_csv function.

3. **The tweets dataset:**

   The tweets archive dataset did not include some information that would be important during analysis such as retweet and favorite count. This additional data had to be gathered by accessing twitter's API. A personal twitter API key and token was gotten from twitter and the tweepy library was used to extract the tweets using the tweet Id in the tweets archive dataset. The tweets were saved in a JSON file and a pandas data frame was created from it.

Assessing data

Two types of data assessment were used:

- Visual assessment:

  The visual assessment involved viewing the data with my eyes and finding issues or mistakes within the dataset. For the tweets archive dataset this assessment was mainly carried out using excel, jupyter notebook was used for the other datasets.

- Programmatic assessment:

  This was done using the help of pandas library in python, functions like info(), describe(), sample() etc.

During the assessment any problem or error encountered were documented and separated into the quality and tidiness issues. I found about 7 quality issues and 2 tidiness isssues but during the cleaning process some other issues were found and added, after cleaning a total of 14 quality and 3 tidiness issues were found.

Cleaning data

Before cleaning copies of the original datasets were made. The issues listed under quality and tidiness were cleaned programmatically using functions from the pandas library. The steps involved includes writing the issue to be cleaned, defining how to clean the issue, writing the code/codes to fix the issue and testing to find out if the cleaning was successful.

The 3 datasets were merged into 1 to ensure tidiness, this dataset called master_df was saved as a separate file in the system. The cleaning act reduced the number of observations in our data from about 2356 to 1605 but the data was still sufficient for our analysis.