Wrangle Report

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL. “Reference Udacity-project-Overview”

The dataset that is to be wrangled (and analyzed and visualized) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage. “Reference Udacity-project-Overview”

Data Wrangling Process

Importing Libraries

Some libraries were needed for the sake of the project processing:

{ pandas: ,
numpy: ,
requests: to get data from urls,
matplotlib: for graphs,
tweepy: to query twitter API,
Json: to create json file containing the data gathered from Twitter API,
timeit: to know how much it takes to query Twitter API,
Ipython: to display images,
urllib: for url requests
}

Data Gathering

I consider this step is the most vital steps as by gathering the data you’re building the infrastructure of your project, any mistakes while gathering your data would directly cause data analysis failure somehow.

1- Enhanced Twitter Archive.csv: Given
2- Image Predictions.tsv (neural network image prediction) : Imported programmatically from url using request

3- tweet_json.txt (Twitter API Data): File Constructed via query twitter API to get favorite counts and retweet counts

Data Assessment

Assessing the data visually and programmatically to indicate tidiness and quality issues

Tidiness issues:

1- Dog stage data are on 4 columns
2- Data are splitted into multiple data frames

Quality issues:

1- In Twitter archive Retweet data are not required
2- In Twitter archive some dog names are mistaken
3- In Twitter archive Tweet_id type should be string instead of int
4- In Twitter archive Timestamp type should be datetime instead of string
5- In Twitter archive Some rating numerator is less than 10
6- In Twitter archive record 313 denominator is == 0
7- In Twitter archive some rating dominator are not equal 10
8- In image prediction some IDs photos are missing
9- In image prediction underscores are used instead of spaces
10- In image prediction all name should be started by uppercase letter
11- Time stamp format is not proper
12- two columns for rating
13- Unnecessary columns are existing
14- Columns are not ordered properly

Data Cleaning

In this step we are cleaning the tidiness and quality issues we have Assessed

1- Dog stage data are on 4 columns
Define
Merge the 4 columns into 1 coulumn

2- Data are splitted into multiple data frames¶
Define
Merge all dataframes into one

1- In Twitter archive Retweet data are not required¶
Define
Delete all retweet data

2- In Twitter archive some dog names are mistaken

Define

Extract the correct names from the text column for these records

3- In Twitter archive Tweet_id type should be string instead of int

Define

Convert tweet_id to str

4- In Twitter archive Timestamp type should be datetime instead of string

Define

Convert timestamp to datetime

8- In image prediction some IDs photos are missing

Define

Keeping only records with photos

9- In image prediction underscores are used instead of spaces¶

Define

replace any underscores by space

10- In image prediction all name should be started by upppercase letter

Define

Make all names proper

11- Time stamp format is not proper

Define

Make 2 columns one for date abd the other for time

5- In Twitter archive Some rating nomerator is less than 10

Define¶

Replace any numerator less than 10 with forward fill

6- In Twitter archive record 313 denominator is == 0

7- In Twitter archive some rating dominator are not equal 10

Define

Replace any dominator not equal to 10 with 10

12- two columns for rating

Define

Merge the two columns of ratings together

13- Unnecessary columns are existing

Define

Delete unnecessary columns

14- Columns are not ordered properly

Define

Reorder columns