Data Analysis Professional Trak

We Rate Dogs

Data Wrangling

@dog_rates · Apr 16
m told she enjoys car rides and smells good. 12/10 would
she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Name: Eslam Abdellatif Dyab

Email: es-eslam.dyab2019@alexu.edu.eg

Introduction

In this report I will give some sort of a big picture of how I tackled this challenging project.

I will talk about the data wrangling processes here,

which consists of 3 main points:

I. **Gathering the data**
II. **Assessing the data**
III. **Cleaning the data**

Gathering Data:

I gathered data from three sources:

1. Enhanced Twitter Archive (.csv)
2. Additional Data via the Twitter API
3. Image Predictions File

- I started by downloading the Enhanced Twitter Archive (.csv) then read it to a pandas data frame called archive_df
- Due to some issues with Twitter Api, I downloaded the tweet_json.txt file and read it using json & pandas library to a data frame called api_df
- Then read the image_predictions.tsv using pandas to a data frame called image_predictions_df

Assessing Data:

I used the two assessing techniques, visually and Programmatically.

Programmatically i used .info() , .describe() , value_counts() to get a sense of the problems.

And the issues i’ve found were both Quality and Tidiness issues.

Quality issues i found:

Archive Enhanced Tabel:

- wrong datatype in (tweet_id) (int >>> str)*
- (2356 - 78) Missing records in (in_reply_to_status_id & in_reply_to_user_id) columns*
- wrong datatype in (in_reply_to_status_id & in_reply_to_user_id) columns (float >>> str)*
- wrong datatype in (timestamp) column (object >> date)*
- (2356 - 181) Missing records in (retweeted_status_id & retweeted_status_user_id & retweeted_status_timestamp) columns*
- wrong datatype in (retweeted_status_id & retweeted_status_user_id) columns (float >>> str)*
- wrong datatype in (retweeted_status_timestamp) columns (object >>> date)*
- (2356 - 2297) Missing records in (expanded_urls) columns which can be dropped(data with no images)*
- rating isn't always correct (like in Bella at index 45)*
- wrong datatype in (rating_numerator) column (int >>> float)*
- missing &wrong data in name column*
- missing &wrong data in (doggo, floofer, pupper, puppo) columns*
- wrong representation of null value in (name,doggo, floofer, pupper, puppo) columns (None >> Nan)*
- retweetes & replies should be removed*
- tweet ids with no images*
- tweet id = 670842764863651840 is not a dog, numerator & denominator >> Null*

- tweet id = 749981277374128128 is a dog but with no rating, numerator & denominator >> Null*
- numerator = 24 is wrong >> null*
- (dog_stage) Dealing with (doggopupper,doggopuppo,doggofloofer)*

Image Predictions Tabel:

- wrong datatype in (tweet_id) (int >>> str)*
- bad column names*
- number of entries = 2075 (<2356 in archive) >>> some tweets without images will be deleted*
- retweets & replies should be removed*

Api Tabel:

- wrong datatype in (tweet_id) (int >>> str)*
- number of entries = 2354 >>> some tweets will be deleted*

Tidiness issues:

- (doggo, floofer, pupper, puppo) columns in Archive Enhanced Table should be combined into one column (stage)*
- (in_reply_to_status_id & in_reply_to_user_id & retweeted_status_id & retweeted_status_user_id & retweeted_status_timestamp) columns in Archive Enhanced Tabel need to be removed.*
- api tabel should be with the archive table in one table*

Cleaning Data:

Now after the Assessing is complete i go to clean those issues in the Cleaning phase.

Starting by making copies of the three data frames i have, (archive_clean, image_predictions_clean, api_clean)

I then started to solve those issues by applying the (Define - Code - Test) method. “Details in the attached code file”

Along the way i combined the archive_clean & api_clean data frames into one data frame called master_df

After finishing the cleaning process, i’ve stored the the cleaned data frames in csv files called twitter_archive_master.csv and image_predictions_clean.csv