pdf

December 25, 2020

1 WeRateDogs wrangling report

1.1 Introduction

In this project, I will Wrangling , analyze and clean data on some dogs from archive of Twitter user @dog_rates, this data contains pictures, the type of dogs and their rate, then I will Visualize it in consistent, easy-to-understand forms.

1.1.1 1. Gathering Data

Data was gathered from 3 sources:

- Download Twitter archive Manually in `twitter_archive_enhanced.csv` file
- Tweet image predictions by using URL 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599f predictions/image-predictions.tsv' Then saving on `image_predictions.tsv` file
- Each tweet's retweet count and favorite ("like") count at minimum, and any additional data Using tweepy Package for python and save it on `tweet_json.txt`

1.1.2 2. Assessing Data

The assessing data had Two types of issues Quality and Tidiness it's containing :

Quality Issues

- `Twitter_archive` : Erroneous DataType (`twitter_id` , `timestamp`)
- `Twitter_archive` : Handle Capitalize value in (`p1`, `p2` , `p3`)
- `Twitter_archive` : Non and incorrect Names
- `Twitter_archive` : Remove retweets
- `Twitter_archive` : Some Tweets has no image
- `Twitter_archive` : `numerator` and `denominator` rates must be fix

- `Twitter_archive` : `denominator` rates must be on range
- `Twitter_archive` : numerator and denominator in a Column

Tidiness Issues

- `Genrally` The three tables has same type of observation unit
- `Twitter_archive` : Drop all columns related for Retweet and reply
- `Twitter_archive` : merge four Columns (doggo , floofer , pupper , puppo)

1.1.3 3. Cleaning Data

Cleaning were fixed any issue we found in assessment it's Two Type Tidiness cleaning amd Quality Cleaning and contain of Three steps for every issue

```
Define: Explaining the problem and the approach.
Code: The complete code that run to fix the data.
Test: Assess the data again to make sure the code works and fixed the issue.
```

1.2 Conclusion

By wrangling the Twitter datasets with many Python libraries with above 3 steps. The datasets became much cleaner. The datasets are now ready to be analyzes to find meaningful insights, then build visualization to summarize the results.

```
In [ ]:
```