Wrangle Report

Contents

Introduction

In this project, I will take three steps to wrangle the dataset for this project

1- **Gathering the Data:** I will gather the data from multiple resources

2- **Asserting Data:** I will review the dataset that I was gathering by using two types of assessment:
   - Visual assessment and Programmatic assessment to check any quality and tidiness issues and I will document it

3- **Cleaning Data:** Last step I will clean the data by using (Define, Code, Test) methodology

Gathering Data:

I gather the data from three different sources

- I have downloaded and uploaded (twitter_archive_enhanced.csv) and read it into a Pandas DataFrame as CSV file
- I have downloaded (image_predictions.tsv) from the provided URL using the Request library as TSV file
- I have read the tweet_json.txt line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count from JSON file
-

Asserting Data:

I go through the three datasets to discover any quality or tidiness issues, and I have discovered many quality and tidiness issues as list it below:

Quality Issues

Archive Twitter

1- drop columns not needed for our analysis that contains the replay and the retweet.

2- twiteer_id should be string

3- timestampe sholud be converted to datatime

4- Correct naming issues

5- There are 5 tweets with rating_numerator greater or ugual to 15 and also there is numerator with decimal so conver it to float then drop we will drop un need it information

6- All rating_denominator should be "10" float as the numerator and some rating_numerators are extreme values.

7- change source column to categoery data type

Image Prediction

8- Change img_num column to category data type

Tidiness Issues

Twitter Archive

- Dog stage in 4 different columns (doggo, floofer, pupper, and puppo) will be in one column
- Create a new column called rating, and calculate the value with new, standardized rating

Image prediction

- Image prediction data should be combined with the archive table.

Twitter data

- tweet data should be combined with the archive table

Cleaning Data:

I have cleaned the issues I mentioned above by using (Define, Code, Test) to make the cleaning process more organized.

In the other report, I will go through the analysis process**, Thanks!**