wrangle_report

October 11, 2022

1 Project: Wrangle and Analyse Data

1.1 Wrangling Report

1.1.1 Introduction

Real-world data rarely comes clean. In this project, I used Python and its libraries, to gather data from three sources and in three different formats. I assessed its quality and tidiness, then cleaned it.

I followed three steps in the wrangling process which are: 1. Gather data 2. Assess data 3. Clean data

I would now explain these steps in detail.

1.1.2 1.Gather Data

The data used for this project was gathered from three sources as follows:

a. Twitter archive enhanced: This was a csv file that was provided by the Udacity Instructors. I downloaded this file and read it using the pandas library.

b Image Predictions: This was a tsv file hosted on a webpage. I got this data using the request python package and the url provided.

c. Twitter API & JSON: The expectation was to access this data from the twitter database using the tweepy library however, I had a challenge getting access from Twitter. An alternative was provided. I used the JSON file provided and the json library to gather this data.

1.1.3 2.Assess Data

After gathering the data (twitter=archive-enhanced, images data and tweets data), I accessed the data both visually and programmatically. I detected some quality and tidiness issues as described below:

QUALITY ISSUES I identified the following quality issues in the data

1. Wrong datatype (integer to strings)
2. Wrong datatype (timestamp)
3. Irrelevant rows
4. Irrelevant columns
5. Incorrect dog names
6. Incorrect ratings

7. Some images do not belong to dogs
8. Inconsistent dog breed names

TIDINESS ISSUES 1.Doggo, floofer, pupper, puppo columns should be in one column (dog type)

2.Twitter archive and tweet json tables should be one

1.1.4 3.Clean Data

In this step, I cleaned the quality and tidiness issues I encountered in the assess step.

The cleaning process had three steps for each issue: a. Define b. Code c. Test

Firstly, I duplicated each table to avoid losing or tempering with the data during my cleaning process.

Some cleaning actions I took included changing datatypes of some columns, dropping unnecessary columns, changing text case to have consistent data e.t.c

For the tidiness issues, I melted four columns describing dog types into one and merged the three tables into one main dataframe.

The detailed cleaning process is available in the wrangle_act jupyter notebook.

1.1.5 Conclusion

Data wrangling is an iterative process. After identifying and cleaning 10 issues in total, it does not certify that the data is free of issues.

This data can still be worked on to obtain an even cleaner data and be used for analysis.