WRANGLE_REPORT

PROJECT: WRANGLE AND ANALYZE DATA

Abstract

This is a report of wrangling (analyzing and visualizing) dataset which is a 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.

UMAR FARUQ ZUBAIRU

umarfaruqzubairu@gmail.com

INTRODUCTION

A good data wrangler knows how to integrate information from multiple data sources, solve common transformation problems, and resolve data cleansing and quality issues. Though I am not yet good but I will keep trying to be the best among equals.

In this project, I used Python and its libraries to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. In the meantime, it will be hard to do all the necessary wrangling acts on the dataset. However, I tried to see I meet the minimum requirement of the project. I document the wrangling efforts and showcase them through analyses and visualizations using Python (and its libraries).

The dataset that I wranglinged 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.

Project Steps Overview

Steps that I follow in this project are:

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 5: Analyzing, and visualizing data

Step 6: Reporting

GATHERING DATA

I follow different methods as required to gather different data of different formats from different sources, viz:

i. **The WeRateDogs Twitter archieve**

   This data is given to me to download manually. I downloaded, upload and read the data into a pandas DataFrame. ***twitter-archive-enhanced.csv***

ii. **The tweet image predictions**

   This data is hosted on Udacity's servers, I downloaded it programmatically using the Requests Library and the url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv provided to me. ***image-predictions.tsv***

iii. **Additional data from the Twitter API**

   My First time using tweepy Library. This piece of data I perform some processes / steps to get my hands on the data. I query the twitter API for each tweet's JSON data using Python's Tweepy Library and store the set of JSON data in a file ***tweet_json.txt*** I then read the file line by line into a pandas DataFrame.

ASSESSING DATA

This is the second stage of Data Wrangling, where I both visually and programmatically assessed the data gathered for quality and tidiness issues. I detect and document about 11 issues (8 quality and 3 tidiness issues).

CLEANING DATA

This is the last stage of data wrangling, where all the detected issues are cleaned (resolved) for analysis and visualization. Before taking steps to clean the issues, I make a copy of the data using the **.copy()** pandas method, should in case I might need the original data later on.

Furthermore, I follow the **define → code → test** framework and clearly document it then finally merge the different data pieces into one single dataframe (table), this is according to the rules of **tidy data.**