UDACITY NANODEGREE DATA ANALYTICS

WRANGLE REPORT (PROJECT 2)

We Rate Dogs Twitter Data Analysis

By Kehinde Salami

Introduction

The Udacity Nanodegree Data Analytics course offers students a comprehensive education in the field of data analysis. As part of the course, students are required to complete a capstone project where they apply the skills and knowledge they have acquired throughout the program to a real-world problem.

One such project is the WeRateDogs Twitter Data Wrangle Analysis. WeRateDogs is a popular Twitter account that rates photos of dogs sent in by its followers. The project involves gathering, assessing, and cleaning data from WeRateDogs' Twitter archive, as well as additional data obtained through the Twitter API. The cleaned data is then analyzed using various data analysis techniques and visualizations.

Project Details

The project provides an opportunity to practice data wrangling techniques, including gathering, assessing, and cleaning data using a variety of tools, including Python and Jupyter Notebooks, to explore and manipulate the data. Not less than eight quality issues and two tidiness issues where fully analyzed and cleaned.

The project tasks included;

- Gathering data (programmatically).
- Assessing data (Identifying quality and tidiness issues).
- Cleaning the assessed data and resolving identified issues.
- Storing the cleaned data.
- Creating Insights and visualizations.
- Writing a comprehensive report to detail the project activities.

Gathering Datasets

This project involved gathering three datasets which are:

1. WeRateDogs Twitter Archive File: Udacity programmatically extracted this data and made Twitter archive enhanced.csv available for direct usage. I downloaded this file from the Udacity platform.

2. Image Predictions File: This file is present in each tweet according to a neural network. It is hosted on Udacity's servers and downloaded programmatically using the Requests library and link below;
   URL:
   https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Tweet JSON File & Twitter API: I also downloaded the tweet json file programmatically through the Udacity platform.

Assessing Datasets

The datasets were assessed virtually and programmatically. The programmatic assessment was very important as it showed majority of the issues in my data sets. The identified issues were classified as quality issues and tidiness issues. For every dataset, I performed consistent, correct, valid and complete checks on them.

Cleaning Datasets

The three data sets that made up data wrangling exercise were further divided into the systemic cleaning stages namely; Define, Code and Test. Another important task here was creating a copy of the original file through '.copy()'. Then all the data sets were cleaned appropriately. So that instead of using the original frames, I could experiment with the copyied dataframes.

Some of the issues that were cleaned are;

- retweeted_status_timestamp, timestamp should be datetime instead of object.
- in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id are expected to be integers/strings instead of float.
- Only original tweets are needed, no reply and retweets.
- posts that are not dogs.
- Refinning the source name within the twitter archive dataset, so that they are easily readable.
- Invalid values in the numerator and denominator rating columns.
- In the name and stages columns some names are missing. They are showing as 'None'.
- Refining p1, p2 and p3 columns and confidence associated with them by combining. This is because confident prediction is mostly needed for this analysis.
- Inconsistent capitalization in the prediction column.
- Twitter_archive: doggo, floofer, pupper, puppo are all stages of dog, should be in one column.
- All the three datasets should be combined into one.

Conclusion

After the cleaning exercise, I merged the three data sets into a single file called the Twitter_archive_master.csv. I then created visualizations and analysis of important meaningful deductions from the final datasets.