WRANGLE REPORT

We Rate Dogs Twitter Data Analysis

By Tal Davidson

Introduction

Introduction The Data Analytics Nanodegree offered by Udacity provides students with a comprehensive education in the field of data analysis. As a culmination of the course, students are tasked with completing a capstone project that allows them to apply the skills and knowledge gained throughout the program to a real-world problem.

One such captivating project is the WeRateDogs Twitter Data Wrangling and Analysis. WeRateDogs is a highly popular Twitter account known for rating user-submitted photos of dogs. This project involves the gathering, assessment, and cleaning of data from WeRateDogs' Twitter archive, as well as additional data obtained through the Twitter API. The cleaned and processed data is then analyzed to derive meaningful insights and answer important questions about the WeRateDogs phenomenon.

By undertaking this project, students not only enhance their data wrangling and analysis skills but also gain hands-on experience working with real-world data and applying various techniques to extract valuable insights. The WeRateDogs Twitter Data Wrangling and Analysis project showcases the practical application of data analytics techniques and serves as a testament to the knowledge and proficiency gained through the Udacity Data Analytics Nanodegree program.

Project Details

The WeRateDogs Twitter Data Wrangle Analysis project offers an excellent opportunity to practice and refine data wrangling techniques. The project involves gathering, assessing, and cleaning data using various tools such as Python and Jupyter Notebooks, allowing for comprehensive exploration and manipulation of the dataset.

Throughout the project, a meticulous assessment of the data was conducted, focusing on identifying and addressing a minimum of eight quality issues and two tidiness issues. This meticulous analysis ensures the dataset's integrity and improves the reliability of the subsequent analysis.

The project tasks encompassed several key steps, including:

1. Gathering data programmatically: Employing Python code to retrieve the required data from diverse sources and APIs.
2. Assessing data: Carefully examining the dataset to identify quality issues (inaccurate or missing data, inconsistencies, etc.) and tidiness issues (structural problems affecting analysis).
3. Cleaning the assessed data: Implementing systematic data cleaning procedures to address the identified quality and tidiness issues, resulting in a refined and more reliable dataset.

4. Storing the cleaned data: Saving the cleaned dataset for future analysis and further exploration.
5. Creating insights and visualizations: Utilizing the cleaned dataset to derive meaningful insights and generate visualizations that effectively communicate the findings.
6. Writing a comprehensive report: Documenting the project activities, including the steps undertaken, data wrangling processes, analysis methodologies, and the derived insights, in a detailed and coherent report.

By completing these project tasks, students gain valuable hands-on experience in all stages of the data wrangling process, from data gathering and assessment to cleaning, analysis, and reporting.

Gathering Datasets

The WeRateDogs Twitter Data Wrangle Analysis project involved gathering three datasets, each obtained through different methods. These datasets are:

1. WeRateDogs Twitter Archive File: The initial dataset, named "Twitter archive enhanced.csv," was made available by Udacity. The data was programmatically extracted by Udacity and provided for direct usage. To obtain this file, I downloaded it from the Udacity platform.
2. Image Predictions File: This dataset contains predictions generated by a neural network for each tweet's accompanying image. The file, named "image-predictions.tsv," is hosted on Udacity's servers. I programmatically downloaded this file using the Requests library, using the following URL:
   https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image
   predictions/image-predictions.tsv
3. Tweet JSON File & Twitter API: The third dataset was obtained through the Twitter API. I programmatically downloaded the tweet JSON file using the provided code provided by Udacity. This allowed me to access additional data associated with the WeRateDogs Twitter account.

By gathering these three datasets, I had access to a comprehensive collection of data that formed the basis of the WeRateDogs Twitter Data Wrangle Analysis project.

Assessing Datasets

The assessment of the datasets involved a combination of virtual and programmatic approaches. The programmatic assessment played a crucial role in identifying a majority of the issues present in the datasets.

During the programmatic assessment, I conducted systematic checks on the datasets to identify quality issues and tidiness issues. Quality issues refer to problems related to the content and integrity of the data, such as missing values, incorrect data types, inconsistencies, and inaccuracies. Tidiness issues, on the other hand, pertain to the structure and organization of the data, including problems like excessive columns, redundant information, and the need for reshaping or reformatting.

For each dataset, I performed consistent, correct, valid, and complete checks to identify any anomalies or irregularities that could affect the integrity and reliability of the data. The programmatic assessment allowed me to automate these checks and efficiently identify issues across the datasets.

By conducting a thorough assessment, both virtually and programmatically, I ensured that any quality and tidiness issues were recognized and documented. This assessment laid the foundation for the subsequent data cleaning process, enabling the datasets to be refined and prepared for further analysis.

Cleaning Datasets

The data wrangling process involved cleaning the three datasets in a systematic manner, following the stages of Define, Code, and Test. To ensure the integrity of the original datasets, I created copies of the files using the `.copy()` method. This allowed me to work with the copied dataframes while preserving the original data for reference and comparison.

Several cleaning tasks were performed on the datasets, addressing various issues. Some of the key cleaning actions included:

- Converting the `retweeted_status_timestamp` and `timestamp` columns to datetime format, as they were initially identified as objects.
- Modifying the data types of `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, and `retweeted_status_user_id` to integers/strings from float values.
- Filtering out reply and retweet posts, focusing only on original tweets.
- Removing non-dog posts from the dataset.
- Refining the source names within the Twitter archive dataset for improved readability.
- Handling invalid values in the numerator and denominator rating columns.
- Addressing missing values in the `name` and `stages` columns, replacing "None" entries.
- Combining the `p1`, `p2`, and `p3` columns along with their associated confidence levels to prioritize confident predictions.
- Standardizing the capitalization in the `prediction` column for consistency.
- Consolidating the `doggo`, `floofer`, `pupper`, and `puppo` stages into a single column within the Twitter archive dataset, as they represent different stages of dogs.
- Combining the three datasets into a single cohesive dataset for further analysis.

By performing these cleaning tasks, the datasets were refined, ensuring consistency, accuracy, and compatibility for subsequent analysis.

Conclusion

Following the completion of the cleaning process, I merged the three datasets into a consolidated file called "Twitter_archive_master.csv." This unified dataset served as the foundation for generating insightful visualizations and conducting analysis.

By leveraging the cleaned and combined dataset, I derived meaningful deductions and generated visualizations that shed light on important aspects of the data. These visualizations and analyses provide valuable insights into various patterns, trends, and relationships within the dataset.

Through the data cleaning and subsequent analysis, this project has demonstrated the power of effective data wrangling techniques in extracting valuable information from complex and messy datasets. The consolidation of the datasets and the derived insights contribute to a comprehensive understanding of the WeRateDogs Twitter data.

The Twitter_archive_master.csv file, along with the visualizations and analyses, serves as an invaluable resource for further exploration and examination of the WeRateDogs dataset. The project has not only enhanced my data wrangling skills but also provided actionable insights and a deeper appreciation for the fascinating world of data analysis.