Report on Gathering, Assessing and Cleaning Exercises for The Udacity Data Analyst Nanodegree: Project Two.

Data Wrangling Process

This report covered the project's initial phase, which involved collecting data from WeRateDogs' (@dogrates) Tweets. The first step in data wrangling was to collect three datasets. The first dataset, which Udacity made easily accessible for manual download, illustrates situations in which data subject to analysis for a given project is provided by an organization or a program manager. The last two datasets show instances of data analysts or data scientists scraping data from the internet, a technique that is best for reproducibility and manageability. They were programmatically downloaded using the Requests and Tweepy packages.

These datasets were then evaluated for quality and organization. The assessments described here might not be the only problems that need to be solved. If comments on the fundamental structure and content of the data were also recorded, the assessment of the data would be even more thorough. Furthermore, keep in mind that assessing data entails both understanding the data and identifying problems.

Effective analyses of the quality and order issues facilitated the subsequent step of data wrangling and cleaning. For each assessment, a thorough, actionable plan was made, which was later turned into codes to deal with the problem. After the codes had been executed, the operation's success in solving the noted issue was assessed. If the test results showed a new issue that might affect the data analysis and visualisation, the three steps of defining, coding, and testing were repeated to address it.

Clean datasets were saved as separate files at the conclusion of the data wrangling stage and made available for this project's Part 2: Exploratory Data Analysis (EDA) and Data Visualization. During the EDA phase, we loaded datasets into Pandas dataframes, investigated dataset content using the Pandas groupby and apply functions, computed some fundamental summary statistics, grouped data into "clusters," and produced a pivot table summarising cluster membership.

Regards,

Agunoweh Timiebi