Data wrangling

Benjamin Bukombe

Introduction

The main objective of this project was to put into practice lessons covered in course 4 as part of the Udacity Data Analyst Nanodegree program. In this project, I wrangled the Twitter data from @dog_rates, also known as WeRateDogs, and created analyses and visualizations that can be used to draw reliable conclusions. The project is divided into three sections namely: Data gathering, Data assessment, and Data cleaning. In the following sections, I will briefly describe my effort to complete each of these tasks.

Data gathering

On the first hand, I was provided with the WeRateDogs Twitter archive that contains more than 5000 basic tweets data, but not everything I would need to complete the analysis. Therefore, I needed to gather additional data, to answer specific questions for this project. The second file, "image prediction" data is hosted on Udacity's server, I downloaded it programmatically through the following link. This file contains breeds of dogs resulting from the classification of a neural network. The last file is the tweet_jason data and was accessed through Twitter API using my Twitter development account and credentials. This dataset includes each tweet's retweet count and likes and much more information essential for further analysis. The three files are all available in my Udacity workspace.

Data assessment

After gathering and storing the three files in my working space, the next step was to assess the data. I assessed data using visual and programmatic approaches. The visual assessment involved looking at all observations of a dataset using a spreadsheet such as Microsoft excel. One example is the Twitter archive data, a "CSV" file. So it was possible to look at the columns and rows quickly using excel. For the second option, I printed all observations of the three tables using the "print()" function available in python (see the jupyter notebook for details).
To assess the data programmatically, I used a few methods from the pandas' library such as head, info, sample, etc.(see the jupyter notebook). For example, to look at the header and first

4 rows of each table, I used the “head()” function. To look at the dimension of the data I used the “shape” method and so on.

Data cleaning

After gathering data and assessing the data, the next step was to clean up potential issues observed or identified during the assessment. The first step was to create copies so that I can keep the original datasets. There were many issues associated with these datasets but for the simplicity and purpose of this project, I focused on eight quality and two tidiness issues (Wickham, 2014) depending on the dataset. For example in the three tables, (1) some columns are not important for the analysis, or to answer my questions, these were dropped. (2)There were also incorrect data types like rating and dog stages, these were corrected depending on the values observed for each variable, (3) for the tidiness, there were columns that should be values of a categorical variable, this is the case of dog stages. Finally, I combined the three tables to have one master file that can be used for my analysis and answer the main questions for this project. For a detailed description refer to the jupyter notebook (wrangle_act.ipynb).

Conclusion

Much of the programming work in data analysis is spent on data preparation: loading, cleaning, transforming, and rearranging. Sometimes the way that data is stored in files or databases is not the way we need it for data analysis and application (McKinney, 2013). There is, therefore, a need to develop these skills and the only way was to complete the course and work through this project.

In summary, my objective was to gather the data programmatically and identify potential issues in the data that I can focus on during the cleaning process. These included inconsistency in the variable names, erroneous values, and tidiness issues. This part of the project helped me to review and practice techniques learned during the program and left me equipped with new skills of reasoning about data.

References

McKinney, W. (2013). *Python for Data Analysis* (1st ed.). O’Reilly Media.

pandas development team. (2021). *Pandas User Guide*. Pandas. Retrieved 3 14, 2021, from https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html

Wickham, H. (2014, 9 12). Tidy Data. *Journal of Statistical Software*, *59*(10), 24. 10.18637/jss.v059.i10