Wrangle Report

Wrangle and Analyze Data Project

5/24/2020

Suryaday Nath

INTRODUCTION

This Wrangle and Analyze Data Project is part of Udacity's Nanodegree Data Analyst program. The work involved wrangling data from the Twitter user @ WeRateDogs, collected from different sources connected with the tweets. WeRateDogs rates photos of people's dogs in a humorous way, most frequently offering scores higher than 10/10. That forms a part of their USP.

In this project, performance and tidiness problems were analyzed and then cleaned after the data had been scraped together. A total of 2 tidiness issues and 11 Quality issues including 1 additional Quality issue have been addressed. An additional part of this project involves comparing all the Dog breeds in the dataset with a master database of dog breeds to weed out incorrect Dog breeds and unnatural entries.

Finally, two visualizations were made with inferences and two word clouds were created. The act report.pdf document contains those insights and visualisations.

GATHERING DATA

The data was collected from 3 different sources:

1) The enhanced twitter archive file was provided and manually downloaded. This file contains specific variables for each tweet including tweet Id, timestamp, text, number and denominator ranking, name, etc.

2) Additional data was collected using the Twitter API including favorite count and retweet count.

3) The tweet image predictions file was programmatically downloaded from Udacity 's servers using the Requests library. This file had the breed of dog which was predicted based on the picture, using machine learning techniques.

ASSESSING DATA

After the data was gathered, assessment was performed using the following methods:

- .head()
- .info()
- .value_counts()
- .isna()
- .describe()
- .summary()

Tidiness issues that were cleaned:

- Combining 4 variables of dog type into 1 column "Doggie_Type"
- Merging all the relevant Columns into two Tables to enhance readability and creating separate tables for Tweet details and Dog Classification details

Quality issues that were cleaned:

- Dataframe contains retweets and not only original ratings
- Remove Less Useful columns from the Dataframes
- tweet_id data type is an integer, should be object
- Timestamp data type is not Datetime
- Name column contains "None" values
- Name column has Lowercase invalid values
- Name column has an invalid entry for Name
- No Separate Column for Rating
- Inaccuracies in Ratings where denominator is not equal to 10
- Inaccuracies in Ratings where numerator is less than 10
- Inconsistent ratings with Decimal Ratings present
- Quality Issue - Classification of Dogs is incorrect for some entries with some very weird values

CLEANING DATA

The issues found during the assessment process were cleaned and tested using the following methods and techniques:

- .isnan
- .astype()
- .to_datetime()
- merge()
- extract()
- .drop()
- .islower()
- .replace()
- .rename()
- .loc[]
- .value_counts()
- Loops
- Functions
- Regular expressions

VISUALISATION OF DATA

- Word Cloud
- Regplot(Seaborn)
- Plot function, Matplot

CONCLUSION

This project helped understand that real world data is usually not found in 1 single source and has to be combined across different sources before any kind of analysis can be done on the data.

The project also helped understand Tidyness and Quality issues that can be present in the data and how to handle those issues.

There was also a part of how Machine learning can be used to extract information and also showed the limitation that such a method can have.