WeRateDogs - Project guide

June-6, 2020

Udacity Project In Data Analyst Nanodegree

Data Analytics: Welcome to the Investigate a Dataset project! Most people can understand the visualizations, as 40% of the people can answer basic questions about the information provided on the record visualizations. Therefore, when providing information in the form of linear charts, people show a good understanding of the plotsand provide accurate forecasts in this project..

Content :

Introduction

Over thousands of years, dogs helped humans hunt and manage livestock, guarded home and farm, and played critical roles in major wars. The contrast of talent and apparent patterns along with emotional contact between dogs and humans created more than 350 distinct breeds, each of which is A closed reproductive clan that reflects a set of specific characteristics. In this project, three sets of data are provided from this data that I needed to answer the questions like to find the most common breed among dog breeds and what are the best species and what is the public opinion to distinguish the breed by society? Are there dogs that are distinguished because of their breeds, etc.

- In this research, it is necessary to focus on rare assets to preserve them and can be cross-breeded to find a strain that has the desired genetic traits for many species.
- Neglect of dogs with aggressive and non-coexisting characteristics with members of society who are of low rank
- Determine which dogs are the most admired of all years? And the best breed of dogs that achieved admiration for every year? Until an investment project is made in husbandry
- Establishing media programs that determine the most important advantages and characteristics of each type of dog for the purpose of profit and the public's buying
- We identify unwanted dogs that cannot compete in the market for demand and supply
- Taking a random sample that determines the ability of dogs to change while providing special meals
- The necessity of analyzing the rapid learning ability of dog samples under constant observation and observation

The data I needed to answer these questions is spread out in three different data sets as the data is not arranged very well and needs to be arranged and cleaned:

- **Enhanced Twitter Archive** The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything.
- **Data via the Twitter API** retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API.
- **Image Predictions File programmatically from a URL** I download every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs

Project details The tasks of this project are as follows:

- Gathering data
- Assessing data
- Cleaning data

Gathering data

We must collect three sets of data to fill in the missing values and format them in one set, so we try to download the data in the file and import it into the work environment Jupyter notebook, also we will download the data set from a website (to predict images via the neural network) and also get the data from the API Of Twitter where the API enables us to access data programmatically through applications where instead of allowing Twitter to access the site's database, it provides programmers the API can access some of the data that we will need in the process of data analysis.

Assessing data

At this stage, we will not explore the data set, but we will make sure that the data makes the data analysis process easy later as we search for the quality of the data and arrange them. All low-quality data is deleted and chaotic data is arranged.

- Dirty data, also known as low quality data. Low quality data has content issues.
- Messy data, also known as untidy data. Untidy data has

data quality issues include:

Quality issues are issues with content, like inaccurate or duplicate data.

- Missing value.(completeness)
- drop many variables.(completeness)
- Change the non-descriptive columns.(validity)
- combine 'doggo', 'floofer', 'pupper', 'puppo' columns.(consistency)
- data duplicates(validity)
- The numerator and denominator columns have unusual values must be deleted,and The numerator must be divided by the denominator, add a column that contains the two columns, and then delete the two columns.(completeness)

data Tidiness issues include:

Tidiness issues are structural issues, specifically: each variable must be a column, each observation must be a row, and each type of observational unit must be a table.

- timestamp column must be from datatype date.
- Correct name for twitter archive clean
- Add the 'year', 'month' columns for timestamp column.
- merge columns in image_prediction table in twitter_archive and tweet_additional tables

Cleaning data

An Assessement of the data was necessary to identify and fix all these problems Cleaning means working on the assessments that we conducted to improve the quality and arrangement in order to correct the data and remove unnecessary data and remove everything that is not important or wrong data or replace or merge