Data Wrangling Report

In this report we will wrangle WeRateDogs Twitter data.

Data wrangling processes:

▪Gathering data

▪ Assessing data

▪ Cleaning data

Gathering Data:

Gathering Data is always the first step in data wrangling, We will gather the data from three different sources.

1 The first source - file on hand

- The WeRateDogs Twitter archive

2 The second source - extracted programmatically

- The tweet image predictions, I downloaded this file programmatically

3 The third source - Twitter API

- I downloaded the tweets data (based on the tweet_id from the WeRateDogs Twitter archive), saved them to a text file called tweet_json.txt

Assessing Data:

In Assessing data step, I used the two types of assessment Programmatic assessment and Visual assessment for Detect quality and tidiness issues

Quality issues:

1. There are missing values in (in_reply_to_status_id, in_reply_to_user_id) columns and because these columns have many missings I will delete the whole columns. Then delete retweeted tweets

2.tweet_id is an integer which is a data type issue, we need to convert it to string.
3.The data in columns (p1, p2, and p3) had uppercase
4.Non-descriptive names of columns.
5.incorrect some of dogs name like (''one'', "a", "an", "by","very") ,replaced with no name
6.Change Timestamp datatype from Object/String to Datetime
7.Change Text range each tweet to be integer because it always starts from 0
8.rating's numerator and denominator should be calculated in one column then change the type to be float

Tidiness issues:

1.FCreate one column for dog stags(doggo, floofer, pupper, puppo)
2.Merging all dataframes into one data frame

Cleaning Data:

In Cleaning data step, I Cleaned all the issues (quality and Tidiness) detected while assessing.

1. There are missing values in (in_reply_to_status_id, in_reply_to_user_id) columns and because these columns have many missings I will delete the whole columns. Then delete retweeted tweets

2.tweet_id is an integer which is a data type issue, we need to convert it to string.
3.The data in columns (p1, p2, and p3) had uppercase
4.Non-descriptive names of columns.
5.incorrect some of dogs name like (''one'', "a", "an", "by","very") ,replaced with no name
6.Change Timestamp datatype from Object/String to Datetime
7.Change Text range each tweet to be integer because it always starts from 0
8.rating's numerator and denominator should be calculated in one column then change the type to be float
9.FCreate one column for dog stags(doggo, floofer, pupper, puppo)
10.Merging all dataframes into one data frame

The End Thank You

by: Amjad Almutairi