UDACITY

Name YAZEED ALARIFI

The Data Wrangling Report

Introduction

Data wrangling is It is the most important skill for everyone who works with the data known to it because most of the data is not clean

data wrangling phase

- Gathering data.
- Assessing data.
- Cleaning data.

Gathering Data

Upgraded Twitter Archive

• The Twitter Archive #WeRateDogs, a csv document that contains the tweet, tweet id, timestamp, text, rating numerator and denominator, dogs name, and so forth.

Image forecast File

• Image forecast file, based variety of canine is in each tweet, I downloaded

the Image forecast file automatically from Udacity's workers utilizing the requests library.

Extra Data by means of the Twitter API

• Additional data assortment including Retweet check & Most Favorite check.

utilizing python's tweepy library.

Assessing Data

If we finish collecting data and storing it in DataFrames , we evaluate the data, its quality and accuracy.

data were evaluated dependent on quality and accuracy.

Low quality is the data that has problems due to poor quality and accuracy، Because of that, we're scanning the unimportant columns and saying "Data Transform" and Clear outliers

• Untidy is data has basic issues. So I doing gathering dogs stages from different columns into one, making a "forecast" column (dogs, not dogs, perhaps dogs), and consolidating the three datasets into one.

Cleaning Data

In this step we will solve the quality and cleaning problems that we identified in the evaluation stage

• Define
• Code
• Test

cleaning data steps

• remove missing value

• remove superfluous columns

• exchange empty value with a space

• Create new columns "dog_stage" and combine 4 columns

• exchange space with NaNs

• Change the data type from timestamp to datetime

• Create new columns (year, month, day, time)

• Create new columns (WeekDay)

• Change datatype for (p1_dog, p2_dog, p3_dog) to number

• Create new column 'Dogs_Predictions' and insert into the column number of True and False

• Change the number with a text and the text based on (0, 'Not Dog'), (1 or 2 'Maybe Dog'), (3, 'Dog')

Conclusion

After these steps, the data is clean and ready for analysis ،and created initial visuals using Matplotlib in Python