UDACITY

The Data Wrangling Report

Introduction

Data wrangling is a core skill that everyone who works with data should be familiar with since so much of the world's data is not clean.

The data wrangling process

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

Gathering Data

Enhanced Twitter Archive

- The Twitter archive #WeRateDogs, a csv file that contains the tweet, tweet id, timestamp, text, rating numerator and denominator, dog name, etc.

Image Predictions File

- Image prediction file, based breed of dog is in each tweet, I downloaded the image prediction file programmatically from Udacity's servers using the requests library.

Additional Data via the Twitter API

- Additional data collection including "Retweet count "and "Favorite count", using python's tweepy library.

Assessing Data

After gathering the data and storing them in **DataFrames**, assessing the data for quality and tidiness.

Data were assessed based on quality and tidiness.

- **Low quality** is data has content issues such as missing, inaccurate.
  So I doing I removing unnecessary columns, converting data types, and removing outliers.
- **Untidy** is data has structural issues.
  So I doing gathering dog stages from multiple columns into one, creating a "prediction" column (dog, not dog, maybe dog), and combining the three datasets into one.

Cleaning Data

It is where we will fix the quality and tidiness issues that we identified in the assess step.

this section consists of the cleaning portion of the data wrangling

- Define
- Code
- Test

Some steps for cleaning data

- Drop missing value
- Drop unnecessary columns
- Replace empty value with a space
- Create new column "dog_stage" and merge 4 column
- Replace space with NaNs
- Change the data type from timestamp to datetime
- Create new columns (year, month, day, time)
- Create new column (WeekDay)
- Change datatype for (p1_dog, p2_dog, p3_dog) to integer

- 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

The data has been cleaned of impurities and is now ready for analysis and created initial visuals using Matplotlib in Python.