DATA WRANGLING REPORT

Project Introduction

There are many definition of data wrangling, I chose to define it in the most simple way, " data wrangling is the process of transforming raw data into easy and analyse to use way.

The objective of this task was to have a hands on practice on data wrangling with data from twitter information provided by the udacity learning portal. To be specific , the topic of concern was the tweeter reviews on dogs,(dog_rates, and weratedogs). this account provides rating of user dogs at a number ranging from 1 too 10. This project has been structured in the chronological order to handle different task that includes below:

Gathering Data.

Assessing Data.

Data Cleaning.

Data Gathering / Data collection /Fetching.

This task used data gathered from different sources which includes. I made an observation of three data sources.

a) **Comma separated value-csv file format.** (Twitter-Archive-Enhanced.csv). This data was provide in the udacity learning portal with the following . this file contained the data variables and respective values of the dogs from the tweets made.link.https://video.udacity-data.com/topher/2018/November/5bf60c1e_twitter-archive-enhanced-2/twitter-archive-enhanced-2.csv.

b) **Tab separated values-tsv (Image-Predictions.tsv).** This file as well was provide on the learning portal. This file contained the details of the dogs image from the tweet for predictions. The details include the name link of the image and the image file extension and media used to upload the image. The link to the data is as follows : https://video.udacity-data.com/topher/2018/November/5bf60c69_image-predictions-3/image-predictions-3.tsv

c) **JavaScriptOjectNotation-json (tweet_json.txt)-** it was also provide on the learning portal. This source contained information similar to contents in above files which were in a dictionary format. The data needed to be pre-processed before it could be ready to be used for analysis.

d) **Tweeter Application inter-phase api -** this is by calling the inter-phase and fetching the data from the internet. The api requires one to have a developer account in order to access the data. The parameters passed include the. The consumer key, access token, consumer secret as well as the access secret.\

Assessing Data.

a) **Visual assessment.** i opened the csv data with Microsoft excel where I was able to make assessment from. It was to my observation that the data had issues that compromised both data quality and data tidiness.

i) Data tidiness.

Structure- with regards to structure, the data set loaded was not in the correct structure. Data structure is considered as a tidy factor. I observed that the there were variable which were not in proper structure, for instance the column of entities, it was poorly structure hence making the data untidy and not ready for analysis. Both raw's and columns were stored in the same variable.

ii) Data semantics - as for this attribute. The dint obey this rule by the fact that it dint not have atomic attributes in each columns. I could observe the column of user instance having the details of user in a dictionary form; user id, user string etc. This qualified the data set to be untidy.

b) Programmatic assessment.

This is the use of programmatic tools to examine the data applying appropriate calculations to create insights and find patterns. I used pandas to check for a summary statistics on the data and as well making visuals. As part of checking for data tidiness, I noticed that the data formats were not well formatted and several columns had incorrect data types in them.for instance the timestamp was an

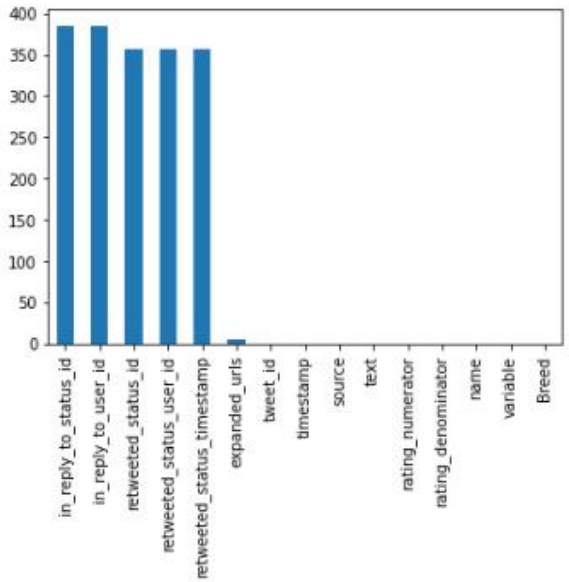
object type when it should have a date format variable.n among this were some of the findings I observed.

In [7]:
```
# getting the data types and information
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   tweet_id                    2356 non-null   int64  
 1   in_reply_to_status_id       78 non-null     float64
 2   in_reply_to_user_id         78 non-null     float64
 3   timestamp                   2356 non-null   object 
 4   source                      2356 non-null   object 
 5   text                        2356 non-null   object 
 6   retweeted_status_id         181 non-null    float64
 7   retweeted_status_user_id    181 non-null    float64
 8   retweeted_status_timestamp  181 non-null    object 
 9   expanded_urls               2297 non-null   object 
 10  rating_numerator            2356 non-null   int64  
 11  rating_denominator          2356 non-null   int64  
 12  name                        2356 non-null   object 
 13  doggo                       2356 non-null   object 
 14  floofer                     2356 non-null   object 
 15  pupper                      2356 non-null   object 
 16  puppo                       2356 non-null   object 
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

In [35]:
```
data.isna().sum().sort_values(ascending=False).plot(kind='bar')
```

Out[35]: <AxesSubplot:>



Data Quality Issues.

1. Data inconsistency. -at glance of my visual assessment, I noticed there were too many columns with none values.
2. Data incompleteness - there were several missing data also still with visual assessment. Further, programmatic data assessment revealed that several columns had a huge a percentage of missing data. For instance in reply tweet status variable had most missing data. The observation is as shown below.
3. Poor data organization . the data was poorly organized and searching through the data was very difficult and a challenging task to archive.
4. Duplicated data. Made an examination of duplicated raw, to my findings I realized that there were several tweet ids that were duplicated, this in itself bridged the quality requirement.
5. Inaccurate data - there were variables with inaccurate data values in them, the were multiple similar data items being stored on different columns.
6. Poorly data definition - the were several data variables with poorly defined names. Rating nominator and rating denominator should have probably been named as max rate and min rating.
7. Data mislabeling- there data that were wrongly labeled. This was making the data understanding a difficult as well as paralyse the analysis process.

Data cleaning

This involved the process of tidying and preparing the data ready to be applied analytic. The above challenges were handled to harmonize the data. A successful assessment suggested creation of function to assist in achieving this goal. The data variables were cleaned with respect to the nature of the problem the columns had. I made a copy of the original data set after which I proceeded to clean the data. I removed the columns that had a huge percentage of missing values. The other values with missing values were identified for formats and nature of columns. The integer types, a fill of none and null values was made with impute strategy of median. The choice of imputation was motivated by the fact that the median values would the most occurring.

The column of timestamp was converted from the object types it originally had to a data type variables. Duplicated tweet ids were removed. This paved way for a smooth analysis. I combined the column of floofer, doggo, pupper,and popper were melted and a single column of name breed was created. The columns values with Nan values

and the none values were removed there after. The visual assessment was not be enough to uncover this thus programmatic assessment come in handy.

I must admit this challenged me, I was however able to achieve the goals for the project from help from previous classes and the learning materials, I also received assistance from googles where I was stack over which I would always get an idea that worked. After cleaning the data. I went a head with my next phase.

Conclusions

data analysis insights is as important as the quality and tidiness of data. Data wrangling helps to achieve the above. For that therefore, a better approach to ensure that data is well structured deserves great attention of emphasis. Understanding the data takes the dimension of both programmatic as well as that one of visual assessment, depending on the nature of data, you might be required to use one or even both. Its important to therefore clean and wrangle the data before beginning the analysis of it.