Udacity Data Analyst Nanodegree

P3: Wrangle and Analyze Data

Author: Challa Sri Venkata Divya Madhuri

Date: Feb 4, 2018

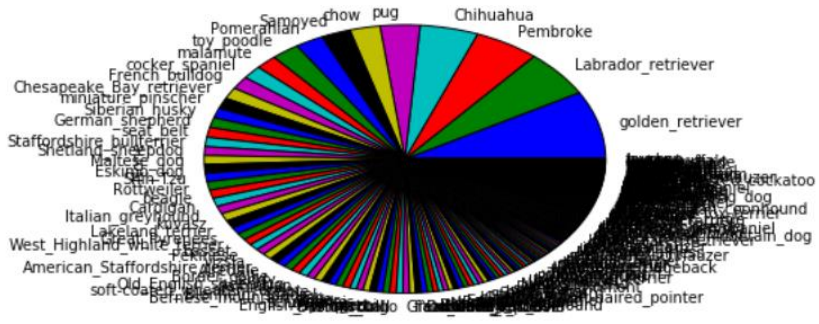
Analysis and visualization

1. Ploting piechart to find the difference in 'Possible_breed_1' column between clean dataset (having only dog entries) and unclean dataset

To find the difference in distribution of the column ‘Possible_breed_1’ before and after cleaning, separate pie-charts are plotted for the ‘Possible_breed_1’ variable for both clean dataframe and the old dataframe before cleaning.

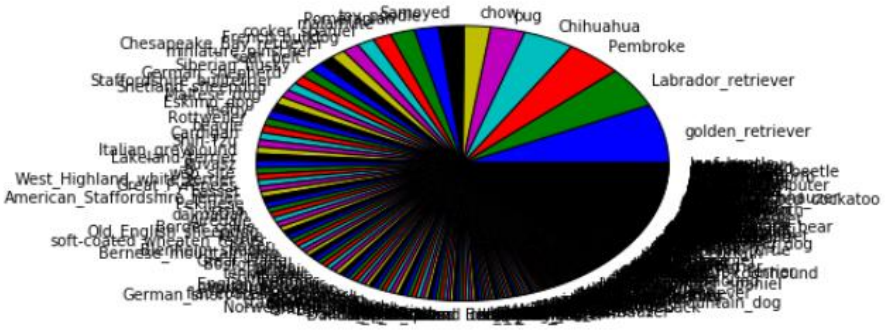
In [42]:
```
import matplotlib.pyplot as plt
%matplotlib inline
# plot all possible dog breeds for df2_clean which has only dogs
df2_clean.Possible_breed_1.value_counts().plot(kind ='pie')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0xf3cc390>
```



In [46]:
```
# plot all possible dog breeds for df_unclean which has things other than dogs
df_unclean.Possible_breed_1.value_counts().plot(kind ='pie')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x112a90b8>
```



The pie-charts clearly showed that the proportion of dog breeds is higher in the case of cleaned dataframe (df2_clean) than that in unclean dataframe (df_unclean). This is because after removing the non dog items like paper-towel, dishwasher , refrigerator, teddy, bathtub etc., the distribution of each dog breed can now be seen more clearly. But the presence of such non dog entries in the uncleaned dataframe made its pie-chart more congested and to lack clarity. From the pie-chart of clean dataframe , we can clearly spot that the 5 most popular dog breeds are golden_retriever, Labrador_retriever, Pembroke, Chihuahua and pug.

2. Analyze 'Overall_rating' column to find number of entries with Overall_rating >=1

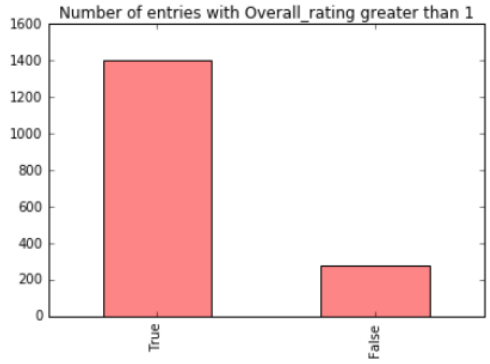
A new column ‘rating_morethan_one’ is created to contain rows for which Overall_rating >=1 and then a barchart is plotted on its value_counts, as shown :-

In [47]:
```
# plotting number of entries having overall_rating >=1
df2_clean['rating_morethan_one'] = df2_clean.Overall_rating >=1
df2_clean.rating_morethan_one.value_counts()
```

Out[47]:
```
True     1406
False     280
Name: rating_morethan_one, dtype: int64
```

In [48]:
```
df2_clean.rating_morethan_one.value_counts().plot(kind ='bar',
                              title ='Number of entries with Overall_rating greater than 1',
                              color = '#fe8585')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x12849b00>
```



3. Insights from the Analysis

- As seen from the two piecharts plotted above, the clean dataset i.e df2_clean has higher proportion of dog breeds like golden_retriever, Labrador_retriever, Chihuahua, Pembroke etc , but there is comparatively lesser proportion of dogs in df_unclean (the unclean dataset).
- Only 280 entries of dogs have Overall_ratings less than 1 , and the rest 1406 entries have Overall_rating greater than 1.
- The 5 most popular dog breeds from the above dataset are golden_retriever, Labrador_retriever, Pembroke, Chihuahua and pug.

Storing Data

The final cleaned dataframe df2_clean is stored as CSV file named twitter_archive_master.csv

In [37]:
```
df2_clean.to_csv('twitter_archive_master.csv')
```