act_report

July 25, 2021

1 Act Report

[1]:
```
import sqlite3
from sqlite3 import Error
import pandas as pd

def create_connection(db_file):
    """ create a database connection to the SQLite database
        specified by db_file
    :param db_file: database file
    :return: Connection object or None
    """
    conn = None
    try:
        conn = sqlite3.connect(db_file)
        return conn
    except Error as e:
        print(e)

    return conn
```

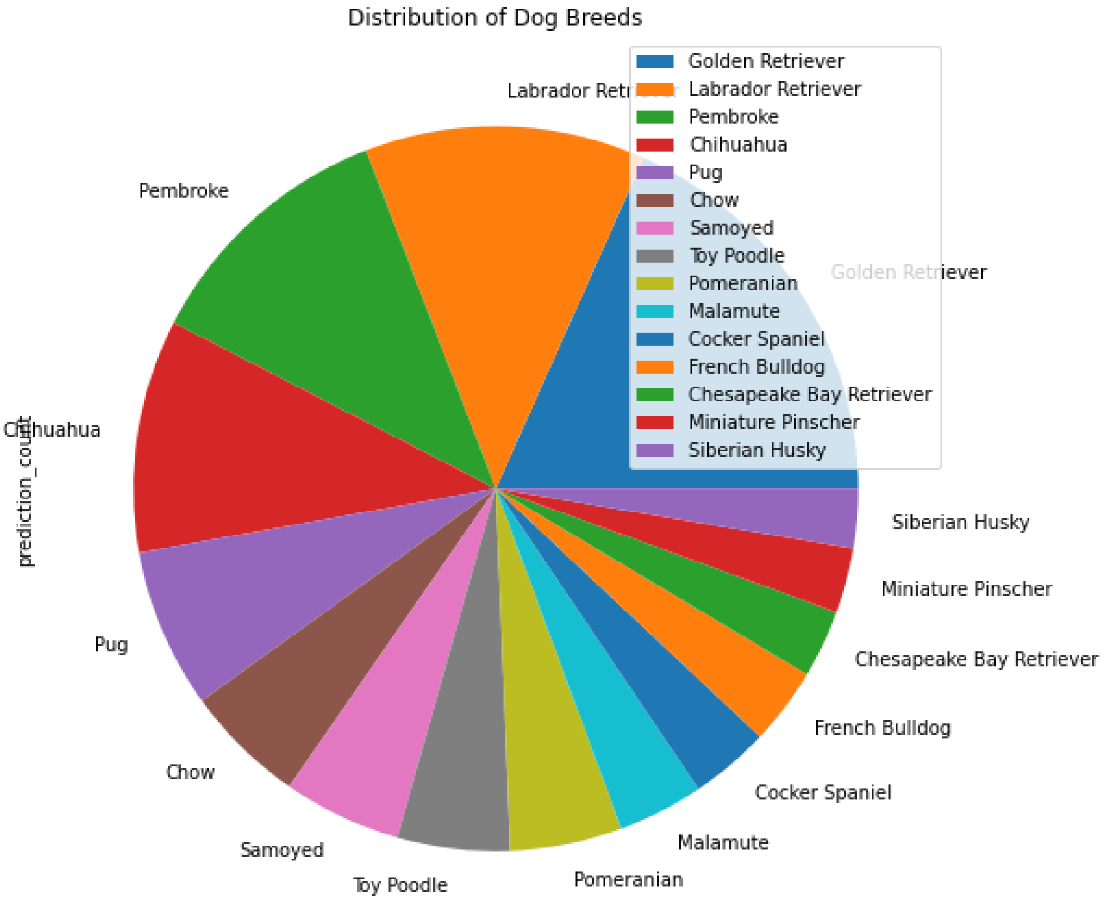
[2]:
```
database = "../data/master.db"
conn = create_connection(database)
```

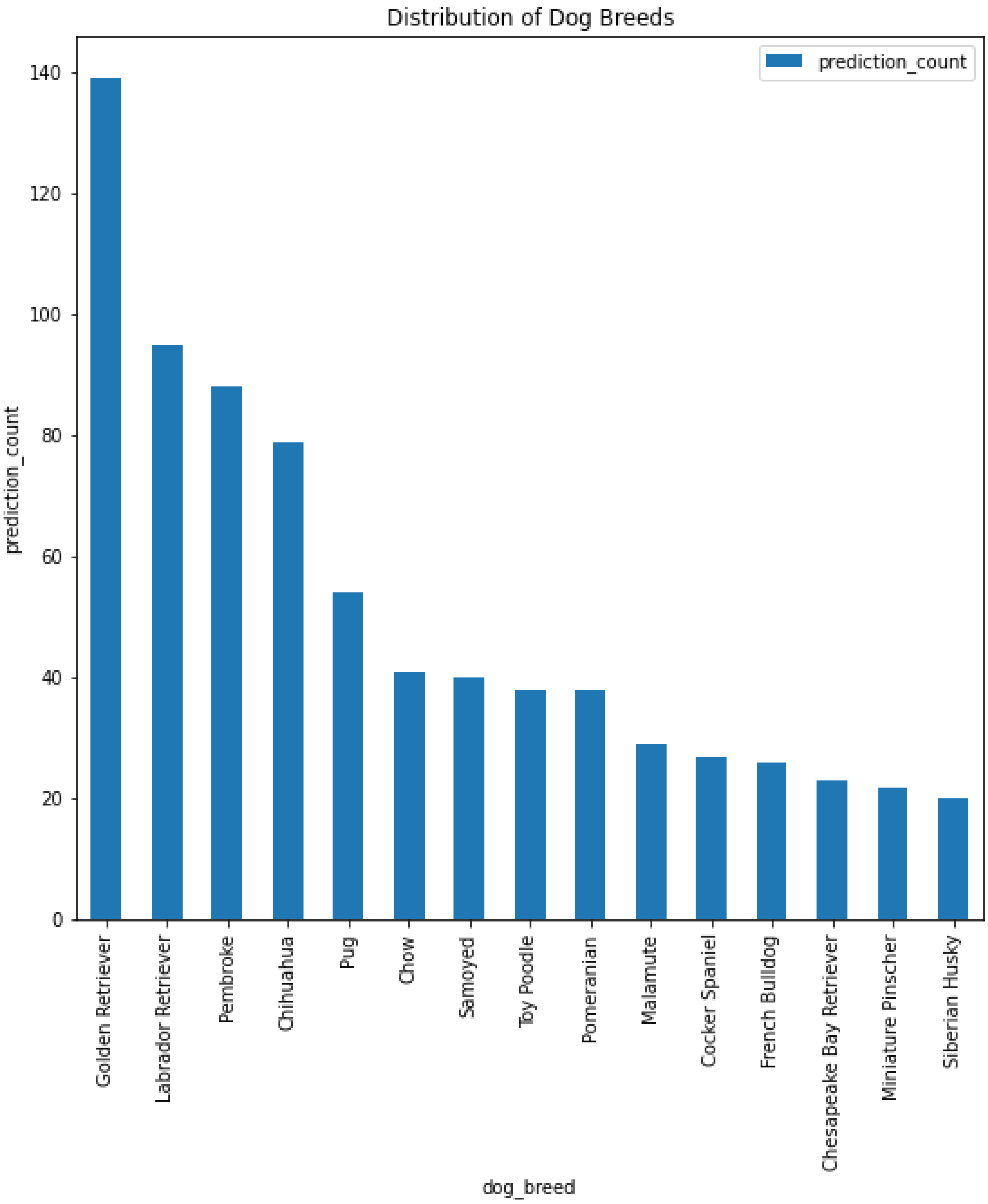
1.1 Distribution of Dog Breeds

[3]:
```
sqlite_select_query = """SELECT prediction AS dog_breed, COUNT(prediction) AS␣
 ↪prediction_count
                         from df_twt_archive_master WHERE breed_predicted ==␣
 ↪True
                         GROUP BY prediction ORDER BY prediction_count DESC␣
 ↪LIMIT 15"""

df = pd.read_sql_query(sqlite_select_query, conn)
df = df.set_index('dog_breed')
df.plot(kind = 'pie', y='prediction_count', title='Distribution of Dog␣
 ↪Breeds',figsize=(9,9))
```

```
df.plot(kind = 'bar', y='prediction_count', title='Distribution of Dog Breeds',␣
 ↪ylabel='prediction_count',figsize=(9,9))
df
```

```
[3]:                              prediction_count
     dog_breed
     Golden Retriever                          139
     Labrador Retriever                         95
     Pembroke                                   88
     Chihuahua                                  79
     Pug                                        54
     Chow                                       41
     Samoyed                                    40
     Toy Poodle                                 38
     Pomeranian                                 38
     Malamute                                   29
     Cocker Spaniel                             27
     French Bulldog                             26
     Chesapeake Bay Retriever                   23
     Miniature Pinscher                         22
     Siberian Husky                             20
```


Distribution of Dog Breeds
prediction_count
Golden Retriever
Labrador Retriever
Pembroke
Chihuahua
Pug
Chow
Samoyed
Toy Poodle
Pomeranian
Malamute
Cocker Spaniel
French Bulldog
Chesapeake Bay Retriever
Miniature Pinscher
Siberian Husky



There is a peak in number for golden retrievers. This number progressively goes down across the top 15 breeds.

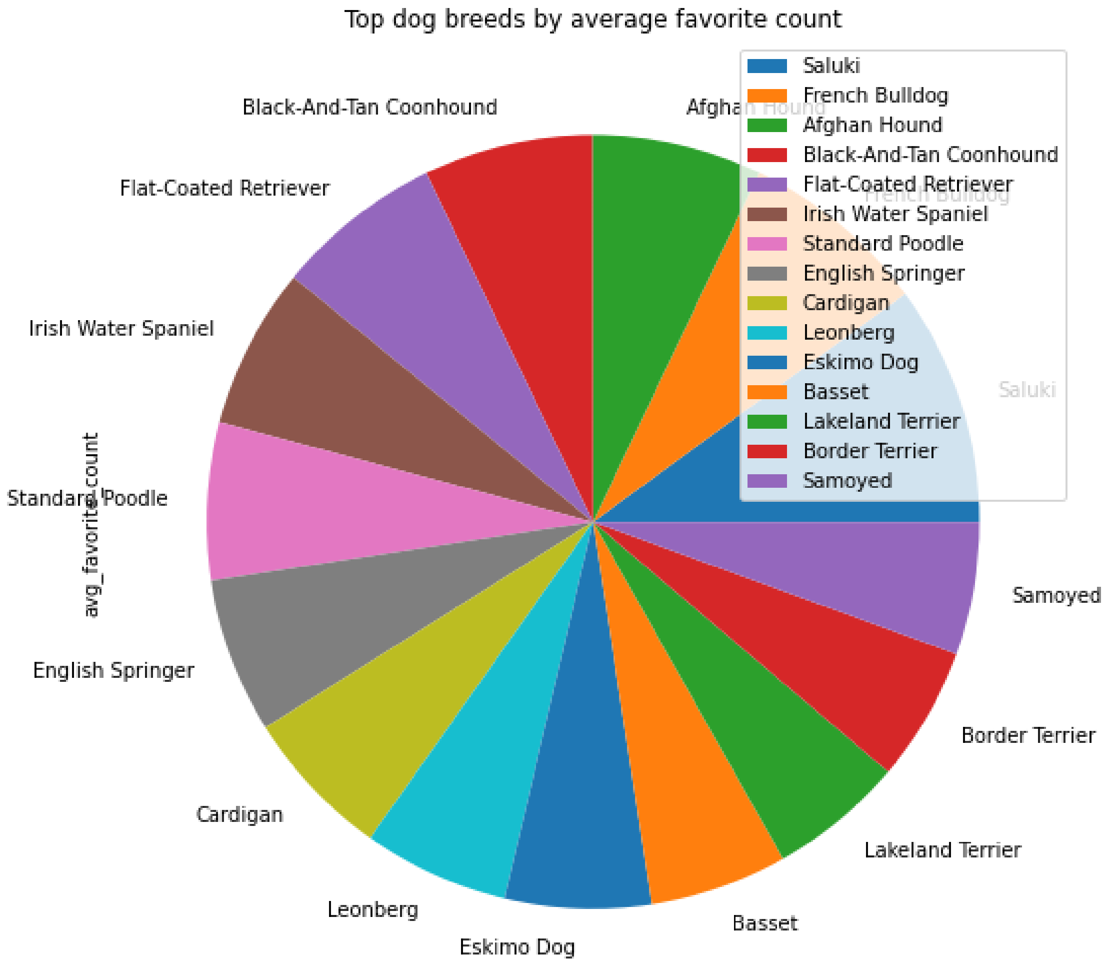
1.2 Top dog breeds by average favourite count

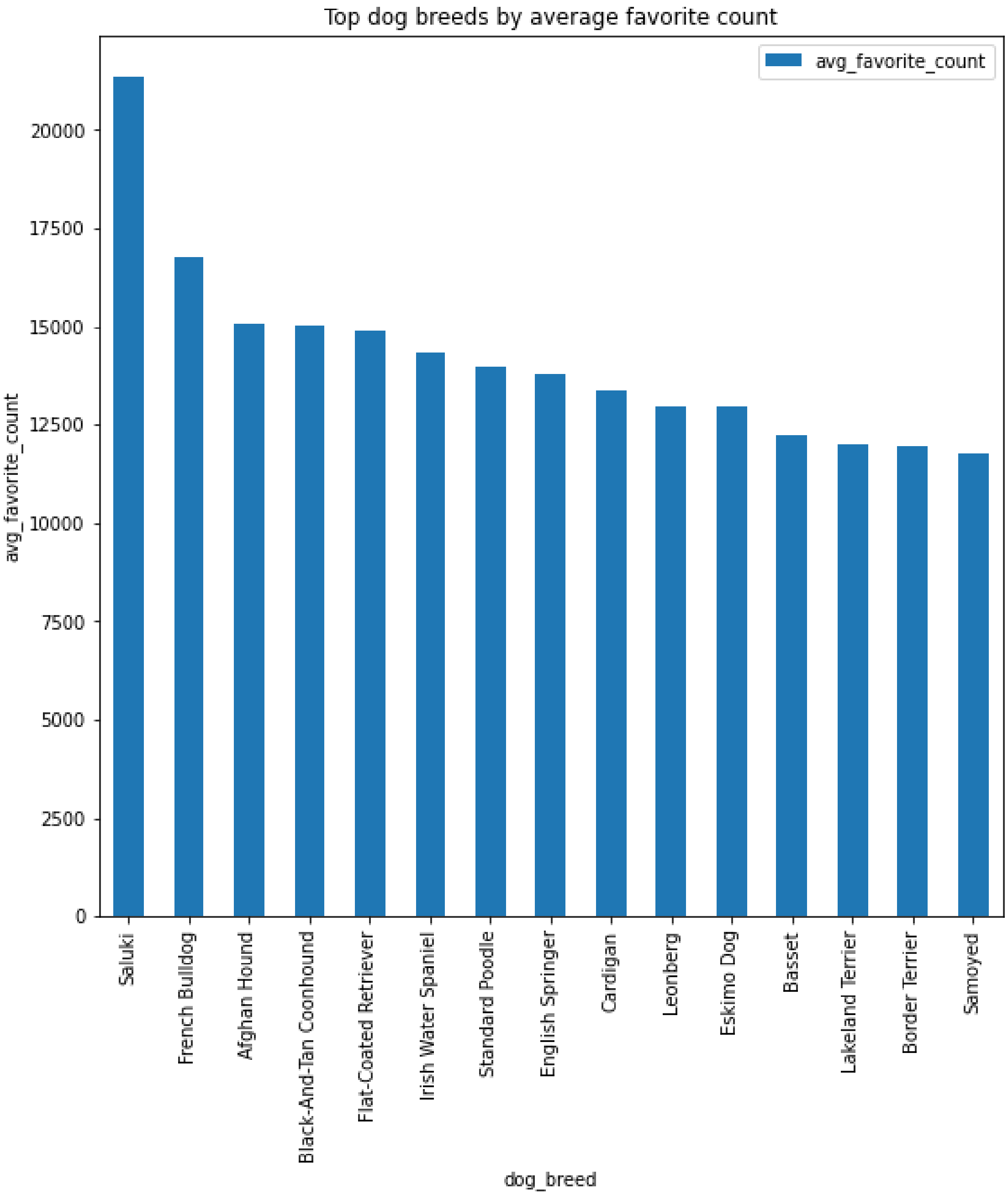
[4]:

```
sqlite_select_query = """SELECT prediction AS dog_breed, avg(favorite_count) AS␣
 ↪avg_favorite_count
                         from df_twt_archive_master WHERE breed_predicted ==␣
 ↪True
                         GROUP BY prediction ORDER BY avg_favorite_count DESC␣
 ↪LIMIT 15"""

df = pd.read_sql_query(sqlite_select_query, conn)
df = df.set_index('dog_breed')
df.plot(kind = 'pie', y='avg_favorite_count', title='Top dog breeds by average␣
 ↪favorite count', figsize=(9,9))
df.plot(kind = 'bar', y='avg_favorite_count', title='Top dog breeds by average␣
 ↪favorite count', ylabel='avg_favorite_count',figsize=(9,9))
df
```

[4]:

```
                         avg_favorite_count
dog_breed
Saluki                         21329.000000
French Bulldog                 16756.360000
Afghan Hound                   15084.666667
Black-And-Tan Coonhound        15019.000000
Flat-Coated Retriever          14882.500000
Irish Water Spaniel            14325.000000
Standard Poodle                13974.285714
English Springer               13795.222222
Cardigan                       13388.235294
Leonberg                       12974.333333
Eskimo Dog                     12951.277778
Basset                         12223.307692
Lakeland Terrier               11998.625000
Border Terrier                 11961.285714
Samoyed                        11776.461538
```


Top dog breeds by average favorite count
Saluki
French Bulldog
Afghan Hound
Black-And-Tan Coonhound
Flat-Coated Retriever
Irish Water Spaniel
Standard Poodle
English Springer
Cardigan
Leonberg
Eskimo Dog
Basset
Lakeland Terrier
Border Terrier
Samoyed
avg_favorite_count
Black-And-Tan Coonhound
Afghan Hound
Flat-Coated Retriever
French Bulldog
Irish Water Spaniel
Saluki
Standard Poodle
Samoyed
English Springer
Border Terrier
Cardigan
Lakeland Terrier
Leonberg
Basset
Eskimo Dog

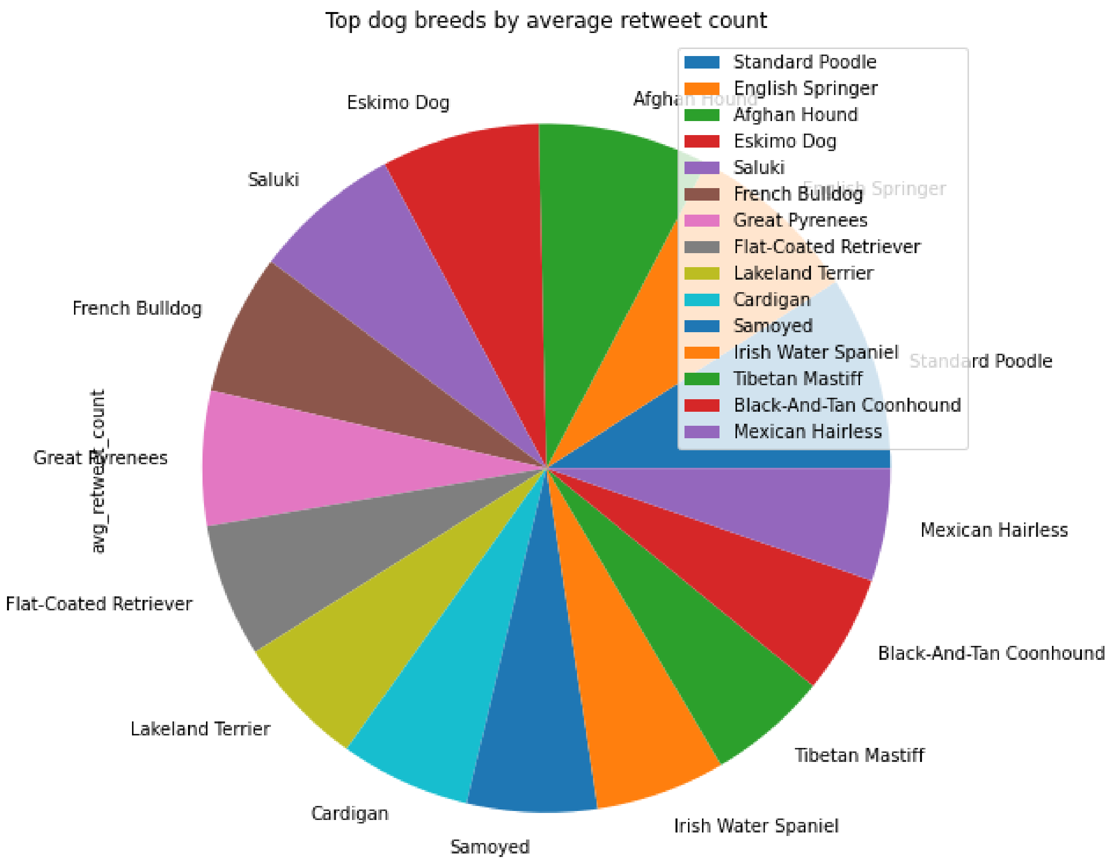


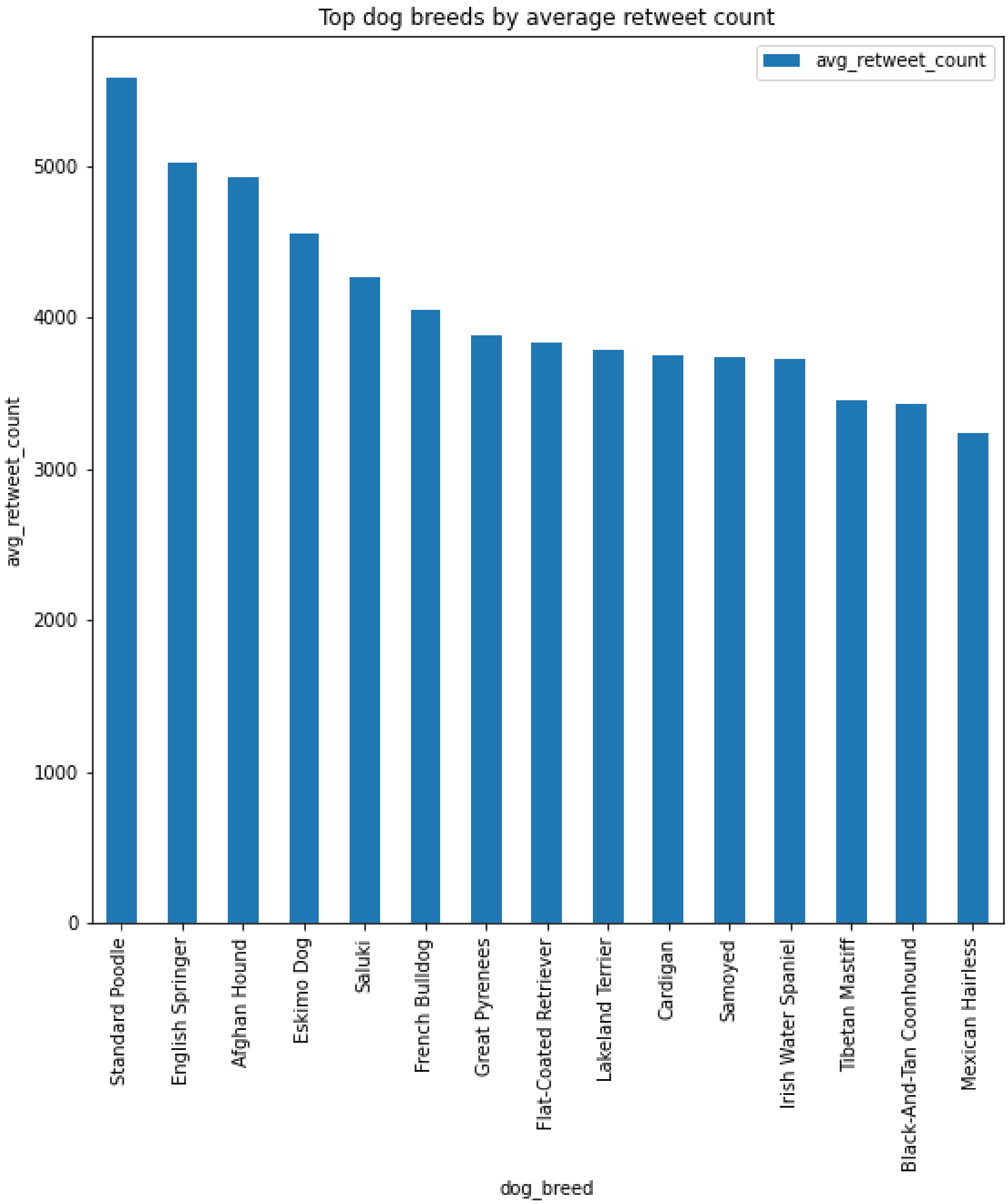
There is a peak for the Saluki breed. The distribution looks even across most of the top 15 breeds from African Hound onwards. None of the most common dog breeds are in the top 15 for average favorite count.

1.3 Top dog breeds by average retweet count

```
[5]: sqlite_select_query = """SELECT prediction AS dog_breed, avg(retweet_count) AS␣
      ↪avg_retweet_count
                              from df_twt_archive_master WHERE breed_predicted ==␣
      ↪True
                              GROUP BY prediction ORDER BY avg_retweet_count DESC␣
      ↪LIMIT 15"""

     df = pd.read_sql_query(sqlite_select_query, conn)
     df = df.set_index('dog_breed')
     df.plot(kind = 'pie', y='avg_retweet_count', title='Top dog breeds by average␣
      ↪retweet count', figsize=(9,9))
     df.plot(kind = 'bar', y='avg_retweet_count', title='Top dog breeds by average␣
      ↪retweet count', ylabel='avg_retweet_count',figsize=(9,9))
     df
```

```
[5]:                          avg_retweet_count
     dog_breed
     Standard Poodle                5577.428571
     English Springer               5015.555556
     Afghan Hound                   4928.333333
     Eskimo Dog                     4550.333333
     Saluki                         4263.250000
     French Bulldog                 4047.200000
     Great Pyrenees                 3877.214286
     Flat-Coated Retriever          3827.875000
     Lakeland Terrier               3790.000000
     Cardigan                       3752.941176
     Samoyed                        3738.564103
     Irish Water Spaniel            3730.333333
     Tibetan Mastiff                3452.750000
     Black-And-Tan Coonhound        3425.500000
     Mexican Hairless               3234.750000
```


Top dog breeds by average retweet count
Standard Poodle
English Springer
Afghan Hound
Eskimo Dog
Saluki
French Bulldog
Great Pyrenees
Flat-Coated Retriever
Lakeland Terrier
Cardigan
Samoyed
Irish Water Spaniel
Tibetan Mastiff
Black-And-Tan Coonhound
Mexican Hairless
avg_retweet_count
Eskimo Dog
Afghan Hound
English Springer
Saluki
French Bulldog
Standard Poodle
Great Pyrenees
Mexican Hairless
Flat-Coated Retriever
Black-And-Tan Coonhound
Lakeland Terrier
Tibetan Mastiff
Cardigan
Irish Water Spaniel
Samoyed



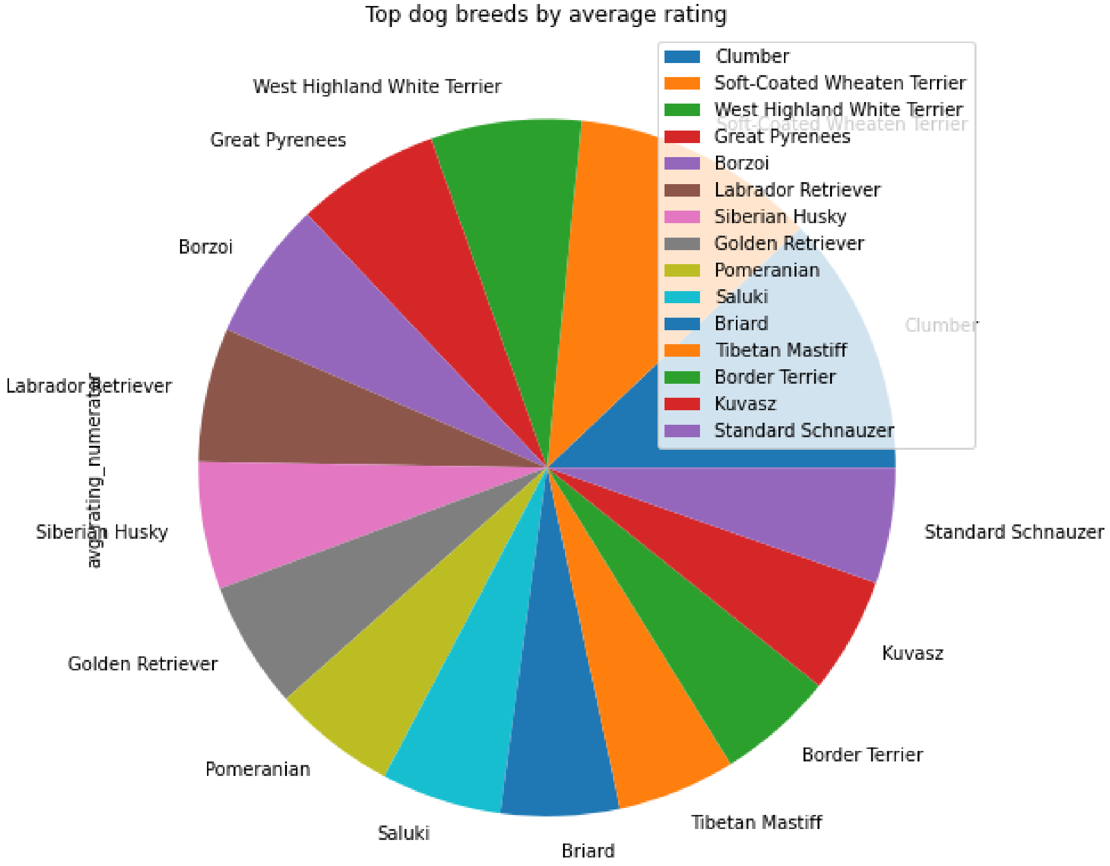
Great Pyrenees is the only breed that appears in the top 15 average retweet count but not in the top 15 average favorite count.

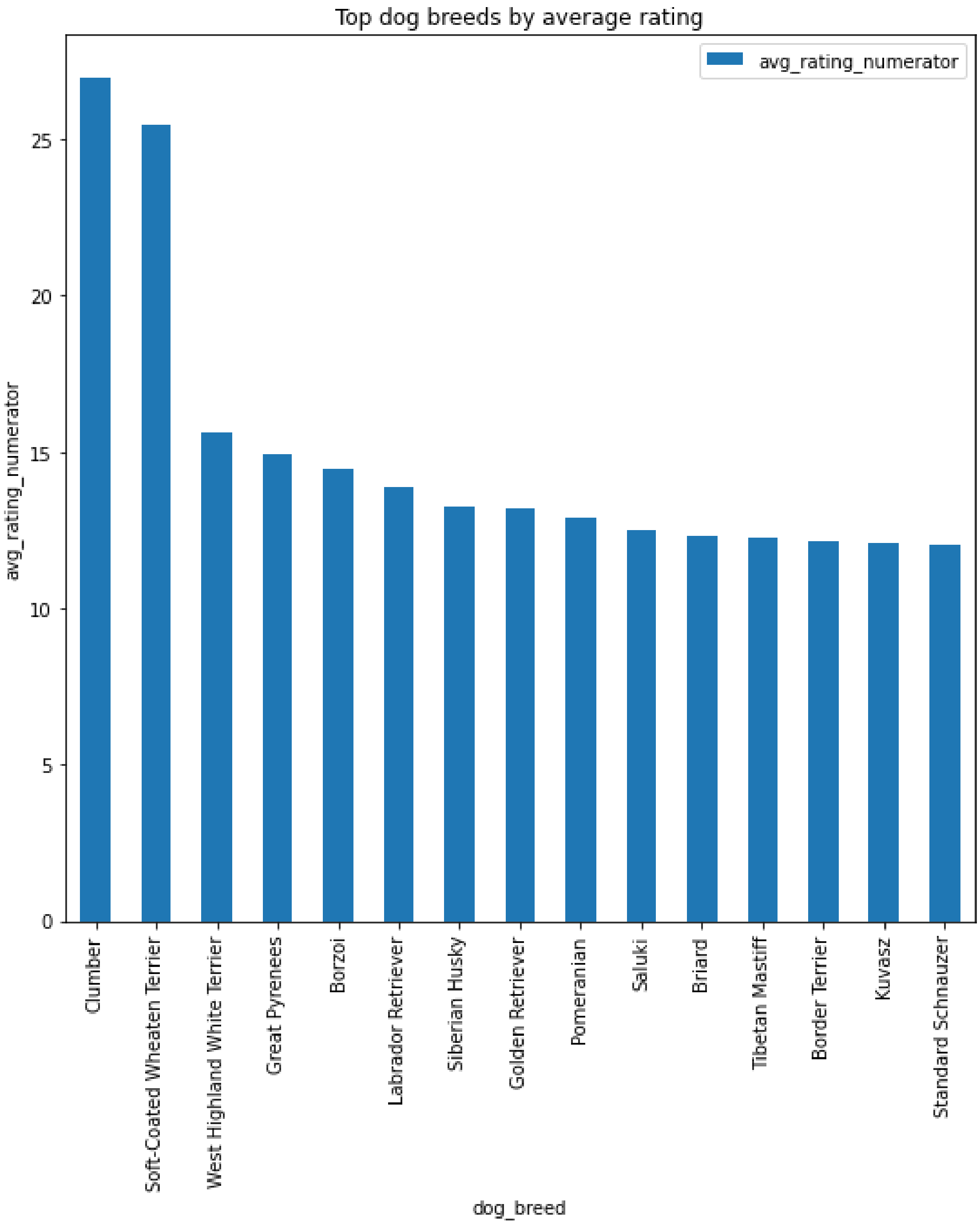
1.4 Top dog breeds by average rating

[6]:
```
sqlite_select_query = """SELECT prediction AS dog_breed, avg(rating_numerator)␣
↪AS avg_rating_numerator, avg(favorite_count) AS avg_favorite_count,␣
↪avg(retweet_count) AS avg_retweet_count
                         from df_twt_archive_master WHERE breed_predicted ==␣
↪True
                         GROUP BY prediction ORDER BY avg_rating_numerator DESC␣
↪LIMIT 15"""

df = pd.read_sql_query(sqlite_select_query, conn)
df = df.set_index('dog_breed')
df.plot(kind = 'pie', y='avg_rating_numerator', title='Top dog breeds by␣
↪average rating', figsize=(9,9))
df.plot(kind = 'bar', y='avg_rating_numerator', title='Top dog breeds by␣
↪average rating', ylabel='avg_rating_numerator',figsize=(9,9))
df
```

[6]:
```
                             avg_rating_numerator  avg_favorite_count  \
dog_breed
Clumber                                 27.000000         6364.000000
Soft-Coated Wheaten Terrier             25.454545         1968.727273
West Highland White Terrier             15.642857         5735.142857
Great Pyrenees                          14.928571        11196.428571
Borzoi                                  14.444444         5431.777778
Labrador Retriever                      13.905263        10158.054348
Siberian Husky                          13.250000         6132.400000
Golden Retriever                        13.208633        10892.618705
Pomeranian                              12.868421         7142.289474
Saluki                                  12.500000        21329.000000
Briard                                  12.333333         8277.666667
Tibetan Mastiff                         12.250000        10566.500000
Border Terrier                          12.142857        11961.285714
Kuvasz                                  12.062500         5079.875000
Standard Schnauzer                      12.000000         1742.000000

                             avg_retweet_count
dog_breed
Clumber                            1536.000000
Soft-Coated Wheaten Terrier         670.818182
West Highland White Terrier        1293.857143
Great Pyrenees                     3877.214286
Borzoi                             1611.111111
Labrador Retriever                 3147.380435
Siberian Husky                     1416.300000
Golden Retriever                   3034.705036
Pomeranian                         2366.894737
```

```
Saluki                          4263.250000
Briard                          2448.333333
Tibetan Mastiff                 3452.750000
Border Terrier                  2820.571429
Kuvasz                          1446.125000
Standard Schnauzer               752.000000
```





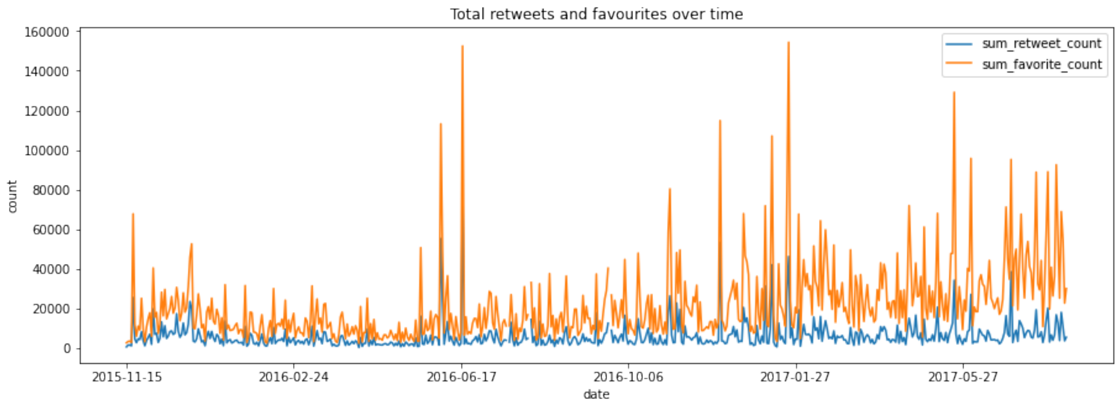
1.5 Total retweets and favourites over time

```
[7]: import pandas as pd
     sqlite_select_query = """SELECT date, SUM(retweet_count) AS sum_retweet_count,␣
      ↪SUM(favorite_count) AS sum_favorite_count
```

```
                         from df_twt_archive_master WHERE breed_predicted ==␣
 ↪True
                         GROUP BY date ORDER BY date ASC"""

df = pd.read_sql_query(sqlite_select_query, conn)
df = df.set_index('date')
df.plot(kind = 'line', y=['sum_retweet_count', 'sum_favorite_count'],␣
 ↪title='Total retweets and favourites over time',␣
 ↪ylabel='count',figsize=(15,5))
df
```

[7]:

```
            sum_retweet_count  sum_favorite_count
date
2015-11-15              518.0              2549.0
2015-11-16             1470.0              3273.0
2015-11-17             1448.0              3430.0
2015-11-18             1049.0              2765.0
2015-11-19            25450.0             67730.0
…                         …                   …
2017-07-27             3724.0             25124.0
2017-07-28            17982.0             68880.0
2017-07-29            10664.0             54451.0
2017-07-31             3576.0             22599.0
2017-08-01             5416.0             30024.0

[563 rows x 2 columns]
```



There are noticeable peaks in favorite count in 2016-06-17, 2017-01-27 and 2017-05-27. Favorite count rises across the years, whereas retweet count raises more gradually with fewer peaks.