I. Introduction

```
data_df.shape
```

```
(1969, 24)
```

	tweet_id	rating_numerator	rating_denominator	img_num	p1_conf	p2_conf	p3_conf
count	1.969000e+03	1969.000000	1969.0	1969.000000	1969.000000	1969.000000	1.969000e+03
mean	7.371736e+17	10.554611	10.0	1.202133	0.593177	0.134028	6.050621e-02
std	6.798196e+16	2.191284	0.0	0.559267	0.272044	0.099926	5.092227e-02
min	6.660209e+17	0.000000	10.0	1.000000	0.044333	0.000010	2.160900e-07
25%	6.758981e+17	10.000000	10.0	1.000000	0.362835	0.054322	1.648340e-02
50%	7.095570e+17	11.000000	10.0	1.000000	0.588230	0.117566	4.981050e-02
75%	7.931355e+17	12.000000	10.0	1.000000	0.841987	0.194207	9.150480e-02
max	8.924206e+17	15.000000	10.0	4.000000	0.999984	0.488014	2.734190e-01

```
data_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1969 entries, 0 to 1968
Data columns (total 24 columns):
tweet_id              1969 non-null int64
timestamp             1969 non-null datetime64[ns, UTC]
source                1969 non-null object
text                  1969 non-null object
expanded_urls         1969 non-null object
rating_numerator      1969 non-null float64
rating_denominator    1969 non-null int64
name                  1326 non-null object
dog_stage             1969 non-null object
jpg_url               1969 non-null object
img_num               1969 non-null int64
p1                    1969 non-null object
p1_conf               1969 non-null float64
p1_dog                1969 non-null bool
p2                    1969 non-null object
p2_conf               1969 non-null float64
p2_dog                1969 non-null bool
p3                    1969 non-null object
p3_conf               1969 non-null float64
p3_dog                1969 non-null bool
retweet_count         1969 non-null object
favorite_count        1969 non-null object
text_clear            1969 non-null object
senti_polarity        1969 non-null float64
dtypes: bool(3), datetime64[ns, UTC](1), float64(5), int64(3), object(12)
memory usage: 344.2+ KB
```

There are 1969 rows and 24 columns in the dataset.

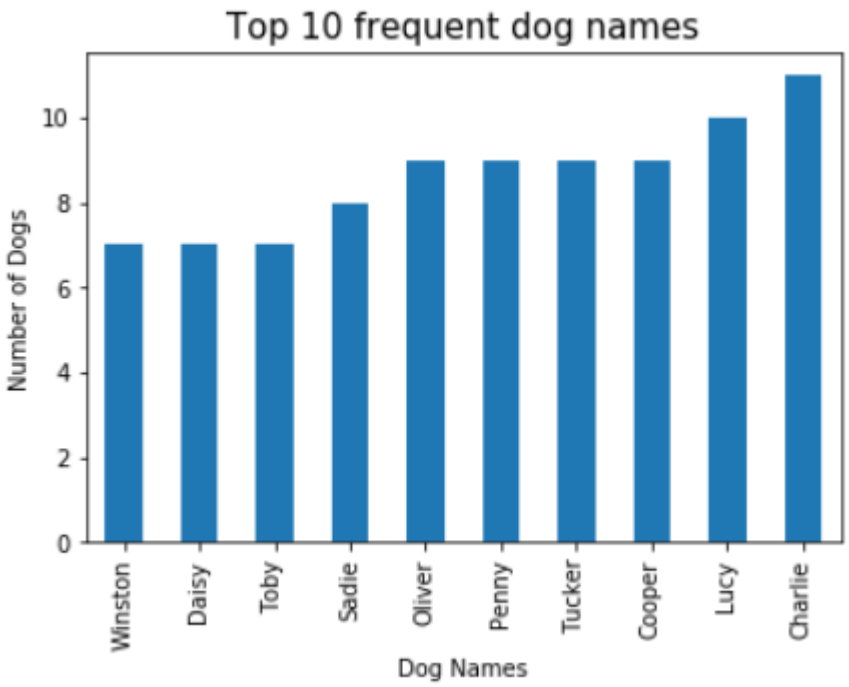
The student would like to investigate into the dataset regarding following questions:

- Which are top frequent names?
- How are average rating different among dog stages?
- Which dog types are accurately predicted most?
- What do people say in the review?

II. Analysis

1. Top frequent dog names

```
data_df['name'].value_counts()[0:10].sort_values(ascending=True).plot(kind = 'bar')
plt.ylabel('Number of Dogs')
plt.title('Top 10 frequent dog names', size=15)
plt.xlabel('Dog Names')
plt.plot();
```



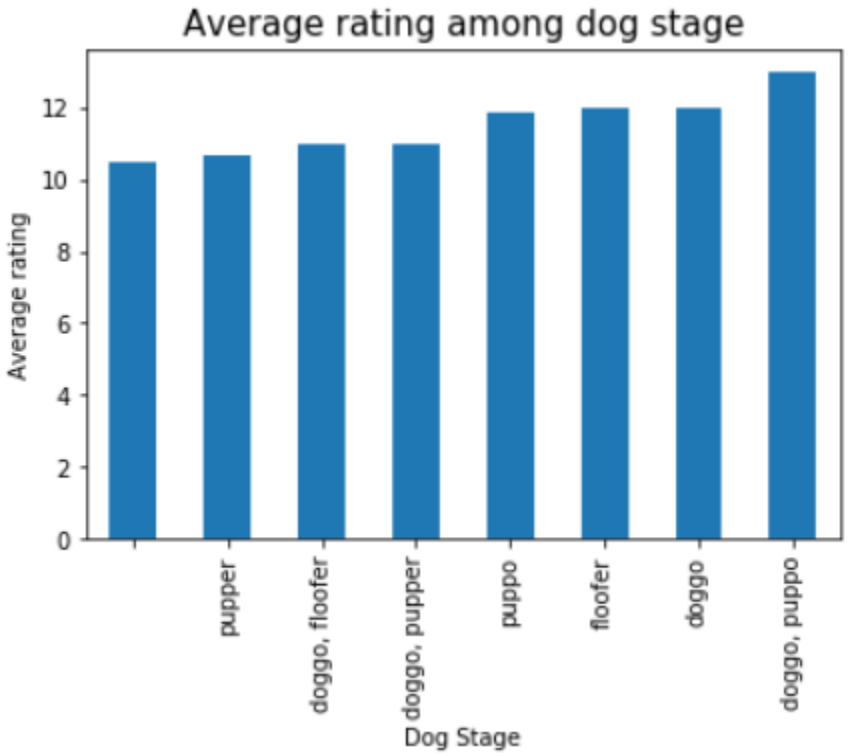
This is the top ten frequent dog names including: Charile, Lucy, Cooper, Tucker, Penny, Oliver, Sadie, Toby, Daisy and Winston. Among them, Charlie and Lucy are the two most popular names.

2. Average Rating

```
data_df.dog_stage.value_counts()
```

```
                1672
pupper           197
doggo             63
puppo             19
doggo, pupper      9
floofer            7
doggo, floofer     1
doggo, puppo       1
Name: dog_stage, dtype: int64
```

```
# Create a variable and store the average rating numerator for each dog stage
avg_rating = data_df.groupby('dog_stage').rating_numerator.mean().sort_values(ascending=True).plot(kind = 'bar')
plt.ylabel('Average rating')
plt.title('Average rating among dog stage', size=15)
plt.xlabel('Dog Stage')
plt.plot();
```



Regarding average rating among dog stage, multiple stage (doggo, puppo) has the highest average rating (more than 12/10). Doggo and Floofer ranks second and third respectively.

3. Dog Prediction

The student would like to investigate into dog prediction according to the first prediction. Step includes:

- Identify top 10 most frequently predicted dog types
- Among them, identify % correction prediction regarding prediction confidence > 0.5 and Prediction is True

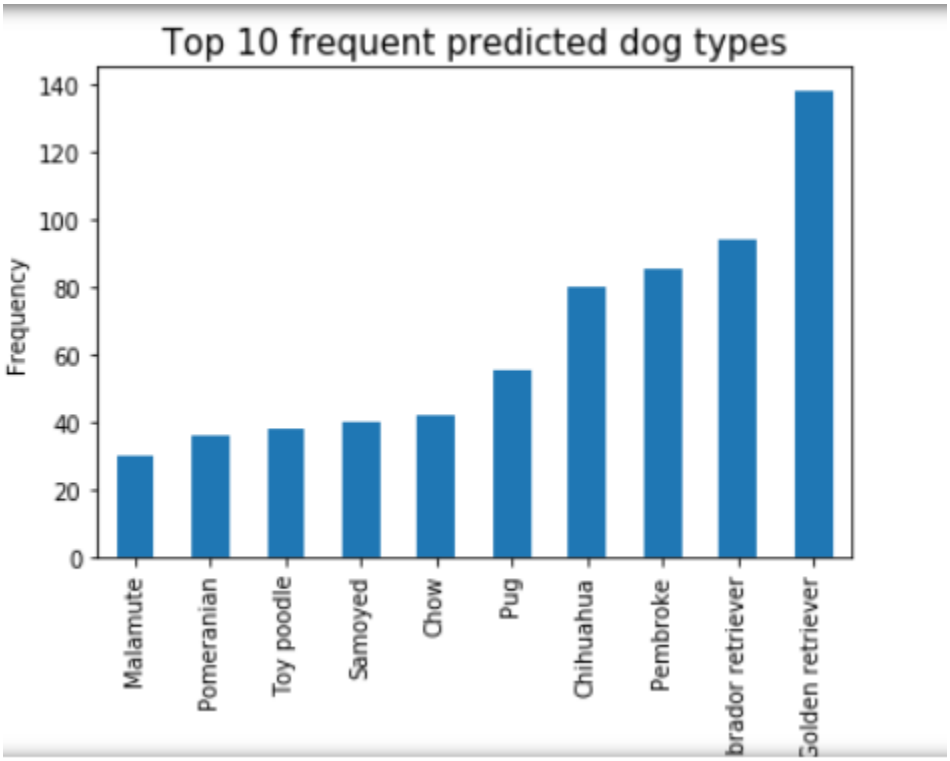
```
#Create first prediction subset with relevant attributes
first_pre = data_df[['tweet_id', 'p1', 'p1_conf', 'p1_dog']]
first_pre.head()
```

	tweet_id	p1	p1_conf	p1_dog
0	892420643555336193	Orange	0.097049	False
1	892177421306343426	Chihuahua	0.323581	True
2	891815181378084864	Chihuahua	0.716012	True
3	891689557279858688	Paper towel	0.170278	False
4	891327558926688256	Basset	0.555712	True

```
first_pre.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1969 entries, 0 to 1968
Data columns (total 4 columns):
tweet_id    1969 non-null int64
p1          1969 non-null object
p1_conf     1969 non-null float64
p1_dog      1969 non-null bool
dtypes: bool(1), float64(1), int64(1), object(1)
memory usage: 63.5+ KB
```

```
first_pre.p1.value_counts()[0:10].sort_values(ascending=True).plot(kind = 'bar')
plt.ylabel('Frequency')
plt.title('Top 10 frequent predicted dog types', size=15)
plt.xlabel('Dog')
plt.plot();
```



Golden retriever, Laborador retriever and Pembroke are the top 3 most frequently predicted. Among top 10, Malmute is the least frequently predicted.

```
# Store the prediction names of 10 most frequent predictions in a seperate variable
pred_name = first_pre.p1.value_counts().head(10).index.values
pred_name
```

```
array(['Golden retriever', 'Labrador retriever', 'Pembroke', 'Chihuahua',
       'Pug', 'Chow', 'Samoyed', 'Toy poodle', 'Pomeranian', 'Malamute'],
      dtype=object)
```

```
# Store the instances when algorithm's first prediction has been successful in a list
true_counts = []
for item in pred_name:
    x = first_pre[(first_pre.p1 == item) & (first_pre.p1_conf > 0.5) & (first_pre.p1_dog == True)]['p1_conf'].count()
    true_counts.append(x)
true_counts
```

```
[114, 64, 67, 48, 44, 26, 29, 23, 28, 18]
```

```
# Store the value counts of 10 most frequent predictions in a seperate variable
total_predictions = first_pre.p1.value_counts().head(10).values
total_predictions
```

```
array([138,  94,  85,  80,  55,  42,  40,  38,  36,  30], dtype=int64)
```

```
# Calculate the prediction efficiency in a seperate column
eff_result['% correct prediction'] = (eff_result['prediction_correct'] / eff_result['prediction_total'])*100
eff_result
```

	prediction_name	prediction_total	prediction_correct	% correct prediction
0	Golden retriever	138	114	82.608696
1	Labrador retriever	94	64	68.085106
2	Pembroke	85	67	78.823529
3	Chihuahua	80	48	60.000000
4	Pug	55	44	80.000000
5	Chow	42	26	61.904762
6	Samoyed	40	29	72.500000
7	Toy poodle	38	23	60.526316
8	Pomeranian	36	28	77.777778
9	Malamute	30	18	60.000000

Among top 10 frequently predicted dog types, Golden retriever are the most frequently and correctly predicted. Pug this the name that ranks second regarding percent of correct prediction with 80% though there are only 55 times of predictions. Although Labrador retriever has high frequency of prediction, the percentage of accuracy is just around 68%.

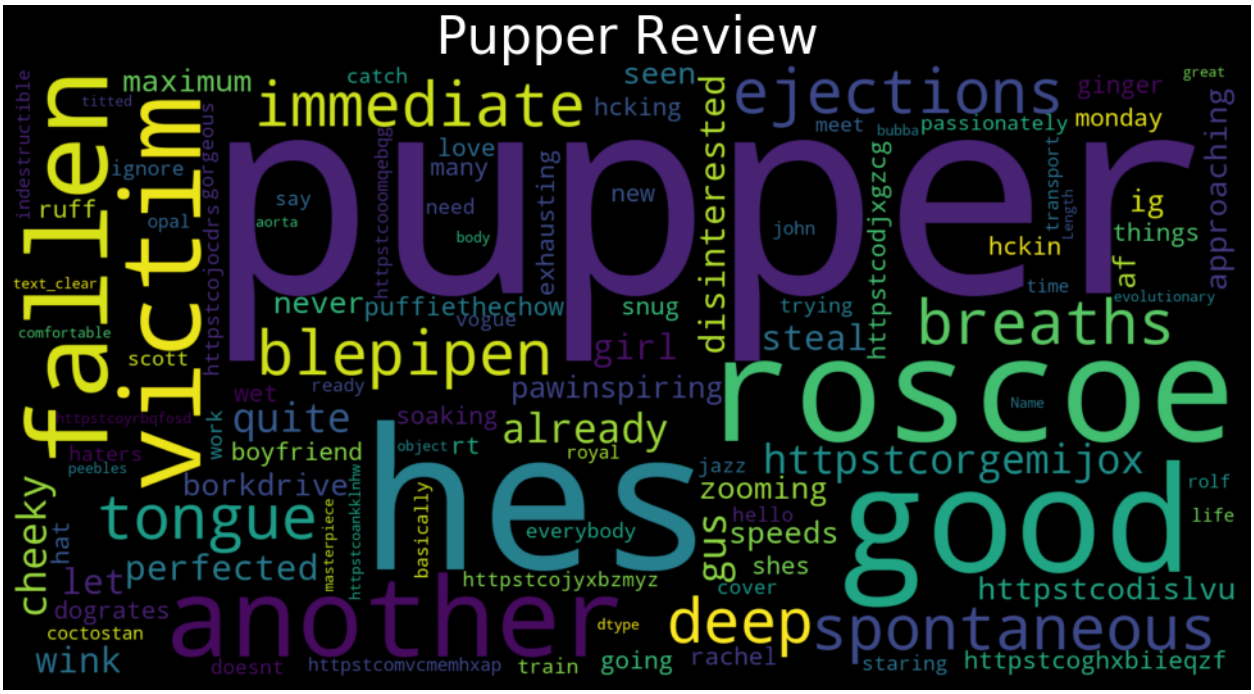
4. Text analysis

The student applied the NLP to identify top keywords are mentioned most in the review and which insights can be inferenced from. Reviews in general and those regarding pupper (the most popular dog stage in the dataset) are analyzed.

Dog Review in general

Top meaningful keyword include: look, big, fan, mystical, Phineas, appears.

Pupper dog



Top meaningful keyword include: pupper, fallen victim, blepipen, spontaneous, tongue, breaths, ejections, immediate, cheeky.

It seems that people always have a special feeling for dog and also take special care of them according to keyword analysis.