ANALYSIS REPORT

A report for the analysis step of master_df

CREATED BY
Habiba Hisham Hatata

Analysis steps

In the analysis step, I started by asking questions related to the dataset or questions that I expect to find their answers in the dataset columns.

These are the questions asked in the analysis step

- What factors affect the favourite counts?
- Most and least common dog names.
- The most common dog age stage.

Q1: What factors affect the favourite counts?

To find an answer to this question, we must test the strength of correlation between the favourite counts columns and the columns we suspect there is a relation to them.

The first column was the language column. I thought that maybe a specific language can get high favourite counts, which will indicate the increased interest of a certain people or nation in a type of dog or dogs generally.

Here, I created a pie chart to show the relationship between the language column and the favourite counts column.

```
# create a pie chart to show the correlation between language and favorite_count
master_df.groupby('language')['favorite_count'].mean().plot(kind = 'pie', autopct = '%1.1f%%')
```

Despite the small number of tweets

In other languages, compared to

English tweets. The Indonesian tweets

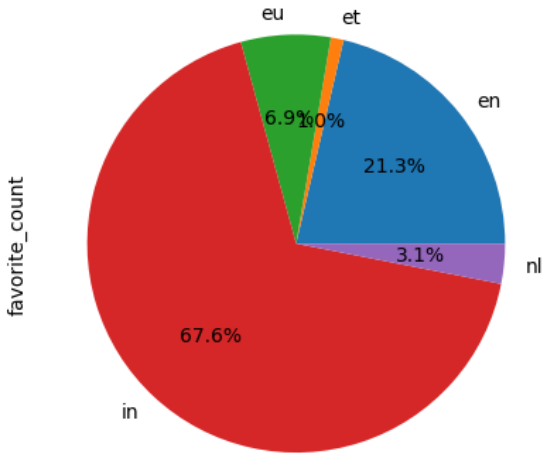
Show a high favourite count

We can concentrate on this

The Indonesian people are interested

In having dogs or they only follow the

WeRateDogs Twitter page

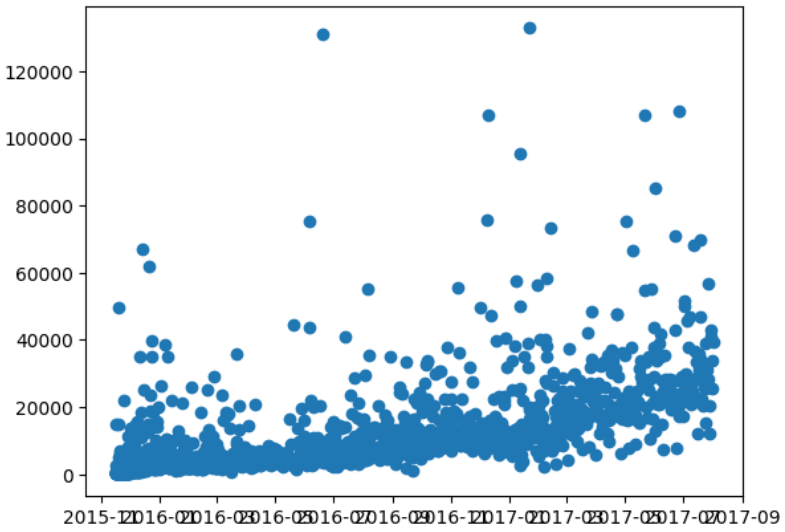


The second column was the timestamp column, I wanted to see the favourite counts change over the years.

```
plt.scatter(master_df['timestamp'], master_df['favorite_count'])
#no strong correlation between the timestamp and the favourite counts
```

From this scatter plot, we can see that

There is no strong correlation between the timestamp and the favourite counts. Which shows that the tweets are randomly distributed over the years, we can also Concentrate that the tweets are mostly getting low fourite counts in the range between 0-20000 And some outliers get more than 100000 counts!

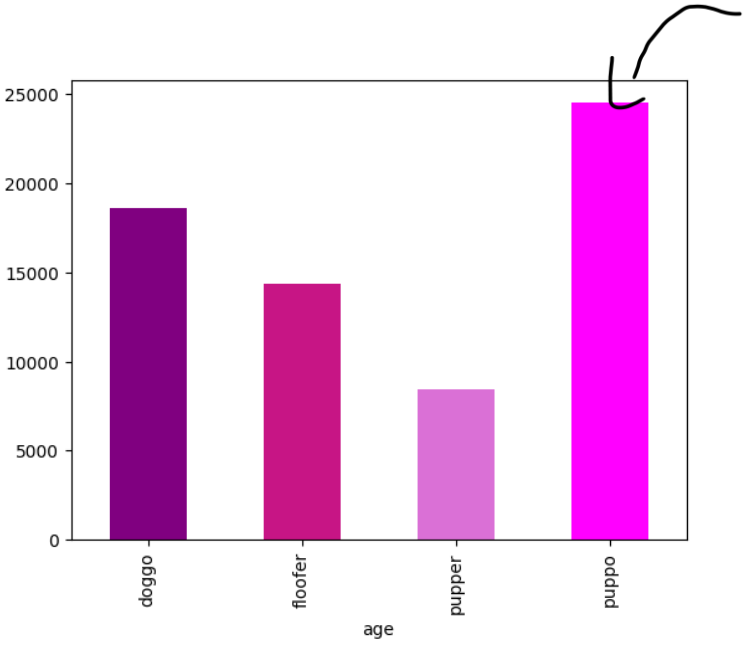


The third column was the age column. I was curious to know which age stage is getting higher favourite counts.

```
new_df = master_df[master_df['age'] != '__']
colours =['purple','mediumvioletred','orchid','fuchsia']
new_df.groupby('age')['favorite_count'].mean().plot(kind = 'bar' , color = colours)
#tweets with puppos are more popular and get higher favourite counts
```

From this bar chart, we can see that the Puppo Age stage is getting higher favourite counts than Other stages which show a high popularity

The least stage with favourite counts is the Pupper Which shows a low popularity



A1 :

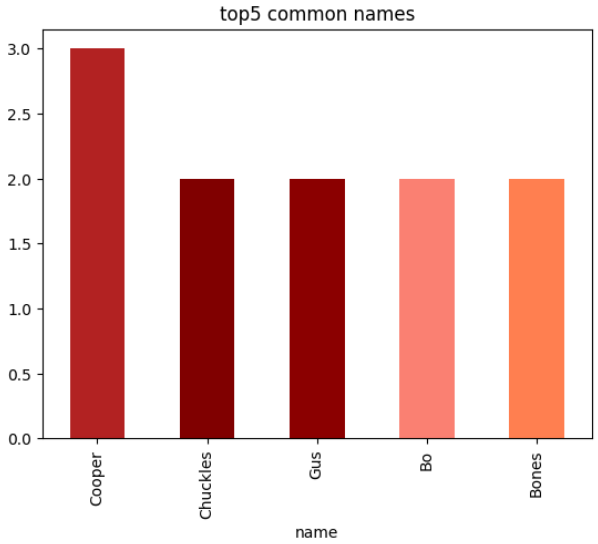
→ From these charts, we can find that the main factor affecting the favourite counts is the age stage, as the age stage with favourite counts is puppo, and the lowest is pupper. Then the second factor is the language, as the Indonesian tweets get a higher number of favourites, however, only 7 tweets are written using different languages. The least factor is the timestamp, which shows a low correlation in the scatter plot

Q2: Most and least common dog names.

I was curious to know the most common name that people usually use to call their dogs. I was thinking of something like Oliver or Bayerd.

```
[83] new_df = master_df[master_df['name'] != 'unknown']
     colours = ['firebrick', 'maroon' ,'darkred' ,'salmon' , 'coral']
     new_df['name'].value_counts().head().plot(kind = 'bar' , color = colours , title = 'top5 common names')
```

After looking at the chart, I found that the name Cooper was the most common. I didn't expect that at all.



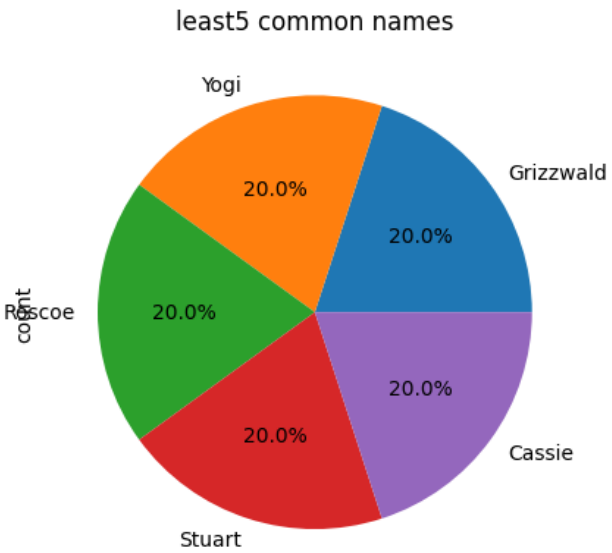
What about the least common names?

```
[84] new_df['name'].value_counts().tail().plot(kind = 'pie', title = 'least5 common names' , autopct = '%1.1f%%')
```

These are the five least common names amongst dogs

Which means that people show less interest in getting

These names of their dogs

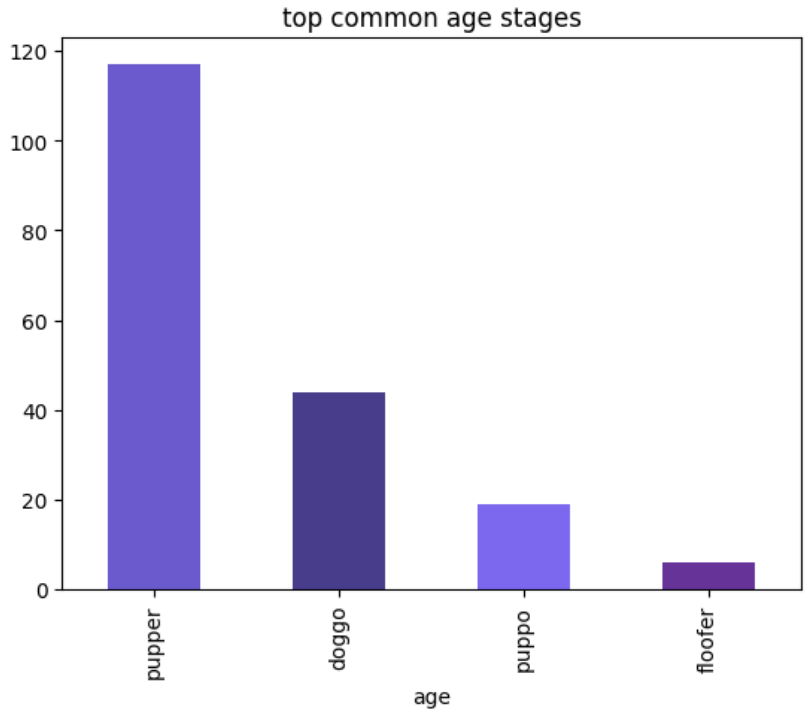


A2:

→ As shown in the bar, Cooper is a very common name in dogs, as many people prefer to give their dog that name. On the other hand, some names showed a lower percentage in the pie chart, like Cassie and Grizzwald

Q3: The most common dog age stage

```
new_df = master_df[master_df['age'] != '__']
colours = ['slateblue','darkslateblue','mediumslateblue','rebeccapurple']
new_df['age'].value_counts().plot(kind = 'bar' , color = colours , title = 'top common age stages')
#the most common stage is the 'pupper' which means that people prefer to get puppers more than any other age stage
```



A3:

→ From the bar chart above we can see that the most common stage for people to buy is the pupper and the least one is the floofer which means that people show more interest to get a pupper more than other age stages Which was the opposite of what was expected because in the bar chart that shows the correlation between the dog age stages and the favourite counts the puppers were the least in favourite counts

Report conclusion

After this long analysis process, we can find that the factors that affect the favourite counts are the language and the dog stage. The most common name is Cooper, and the least common names, which all have 20% favourite counts, are like Cassie and Grizzward. The most common dog age stage is the pupper, and the least common one is the floofer