My Wrangle Report

PRESENTED BY: ONOJAKE MERCY

Data Analyst Nano degree student, Udacity

Introduction of the Report

This report is basically to explain the processes that occured in the wrangling of the data provided by werateDogs.

This project is a combination of three data sets which generally shows us the ratings of dogs on the @werate_dogs twitter profile.

My Wrangling Efforts

STEP 1: GATHERING DATA

I was able to gather this data through the Udacity Classroom Platform. The required files were:
- 1. Twitter archive file
- 2. The tweet image predictions

This was also done by importing necessary packages needed.

STEP 2: ACCESSING THE DATA

Using the jupyter notebook, I was able to access and assess the data in the data frames.

STEP 3: CLEANING THE DATA

This section is basically to:

- Determine the datatype of each column: This is done to know the exact type of data used in each column wether it is an integer, float etc.

- Check for missing and duplicate values: This is done so as to identify any value of that is empty or is duplicated. Using values like this could lead to wrong analysis leading to misrepresentation of the data.

- Remove missing values: After identifying the missing values, it is necessary that the missing values are removed so as to make the data clean and the analysis free from errors.

STEP 4: STORING THE DATA

This involves storing the merged data frame in a new and single data frame after which further analysis would be carried out.

STEP 5: ANALYZING AND VISUALIZING THE DATA

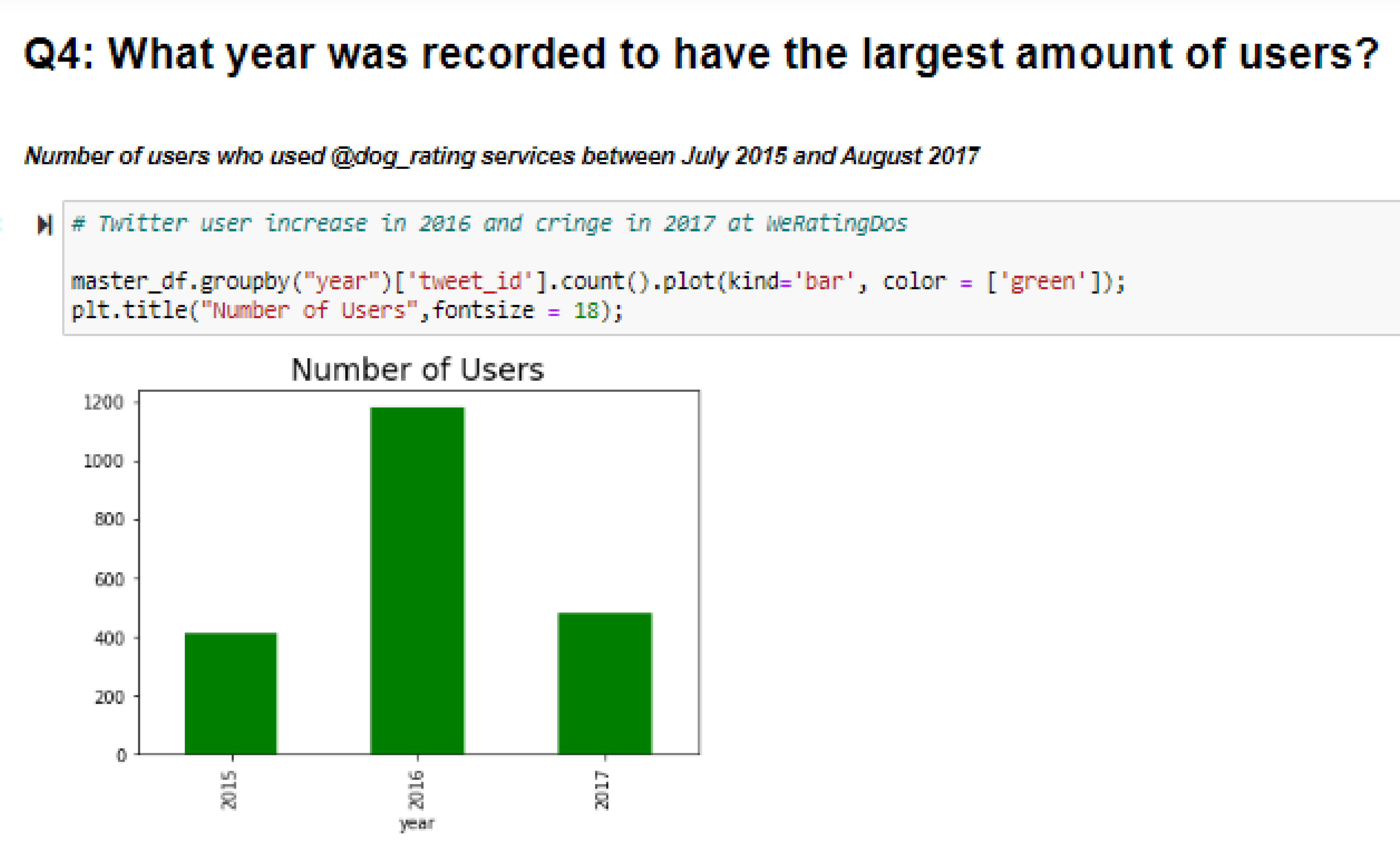
Now, there is a consolidation of the data sets, the data can be visualized through the use of bar charts, pie charts and many other visualization tools.

STEP 6: REPORTING THE DATA

At the end of the data analysis and visualizations, a report on the analysis carried out should be given and this will be contained in the "Conclusions" and "Limitations" heading.

Visualising the Data

Here's an example;



Q4: What year was recorded to have the largest amount of users?

Number of users who used @dog_rating services between July 2015 and August 2017

```
# Twitter user increase in 2016 and cringe in 2017 at WeRatingDos

master_df.groupby("year")['tweet_id'].count().plot(kind='bar', color = ['green']);
plt.title("Number of Users",fontsize = 18);
```

Reporting the Data

CONCLUSION

At the end of the analysis, the following can be deduced;
1. pupper is the one with the most population amongst the dog stages.
2. No image was lost.
3. The year 2016 had the most users.
4. Golden_retriever is the most popular dog breed.

LIMITATION

Being that the information (data) provided were supplied by users, the degree of accuracy cannot be determined or ascertained.
In essence, this data is subject to bias.

Tidiness Issues

1- I created a dog stage column to put all type of dogs classification.

2- I changed the'timestamp' from string to date format day, month , year columns.

3- I combined three different data frames into one master data set.

Quality Issues

1- The'rating_denominator' columns standard should be set at 10.

2- The dog rating should be calculated.

3- Change 'None' to empty cell in doggo,floofer,pupper,puppo to add in dogs stage, after that delete doggo,floofer,pupper,puppo columns.

4- Delete unnecessary columns that will not be used in analysis

5- Delete the 'timestamp' column.

6- Replace 'None' with NaN to indicate the missing values.

7- The dog names format should be consistent. I made the first letter capital for all the names.

8- The column 'name' has an error value 'a' and was removed or corrected.

9- Drop duplicates values from "jpg_url" column.

10- I converted the time stamp column or "create_date" column into real date having their individual columns. i.e. "Day" has its own column, "time", "month_name" etc all have their columns.

Thank you!