WRANGLE REPORT

Table des matières

INTRODUCTION

L'ensemble des données à étudier dans le cadre de ce projet est l'archive de tweets de l'utilisateur de Twitter @dog_rates, également connu sous le nom de **WeRateDogs**. Le but de ce document est de faire des analyses visuelles sur le fichier twitter_archive_master.csv issue du nettoyage de données détaillé sur le fichier Wrangle_report. Ainsi nous allons dans un premier temps nous poser les bonnes questions pour comprendre ce jeu de données(I) puis faire de la visualisation des données et l'analyse de celles-ci(II) .

I- Business Understanding :

Les questions business auxquelles nous allons essayer de répondre sont les suivantes :

- Quelle est la répartition des retweets?
- Quelle est la répartition des notes ?
- Quelle est la répartition des favoris(Likes) ?
- Quelle est la répartition des chiens par catégories ?
- Quelle est la relation entre le nombre de retweet et les notes ?
- Quelle est la relation entre le nombre de like et les notes?
- Quelle est la relation entre le nombre de tweet et le nombre de favoris ?

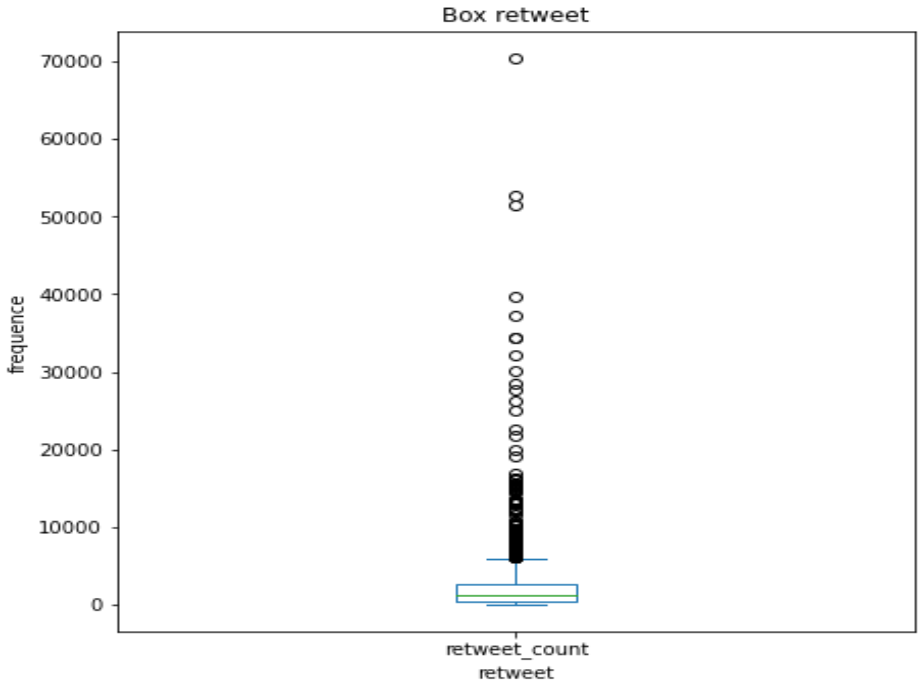
II- Visualisation et analyse :

Pour plus de clarté l'analyse se fera en fonction du fait que la visualisation fait intervenir plusieurs variables (univariée) ou plusieurs variables (bi variées)

1- Analyse Univariée:

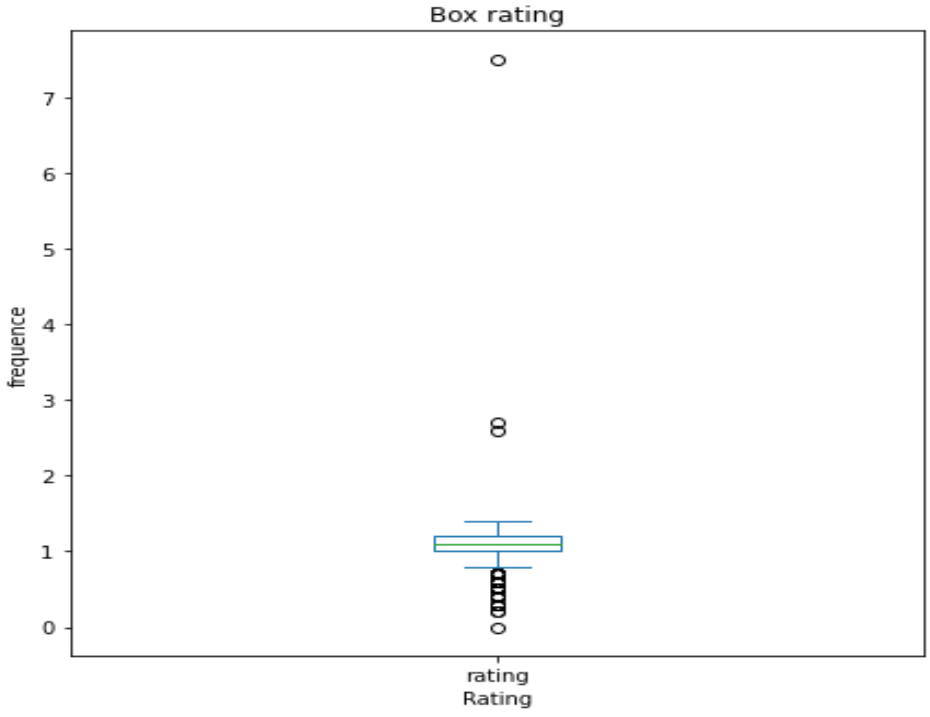
L'analyse univariée nous permet de mieux comprendre la répartition de nos échantillons de données à étudier. Nous utiliserons les box pour les variables quantitatives et l'histogramme pour les variables qualitatives

a- Quel est la répartition des retweets?



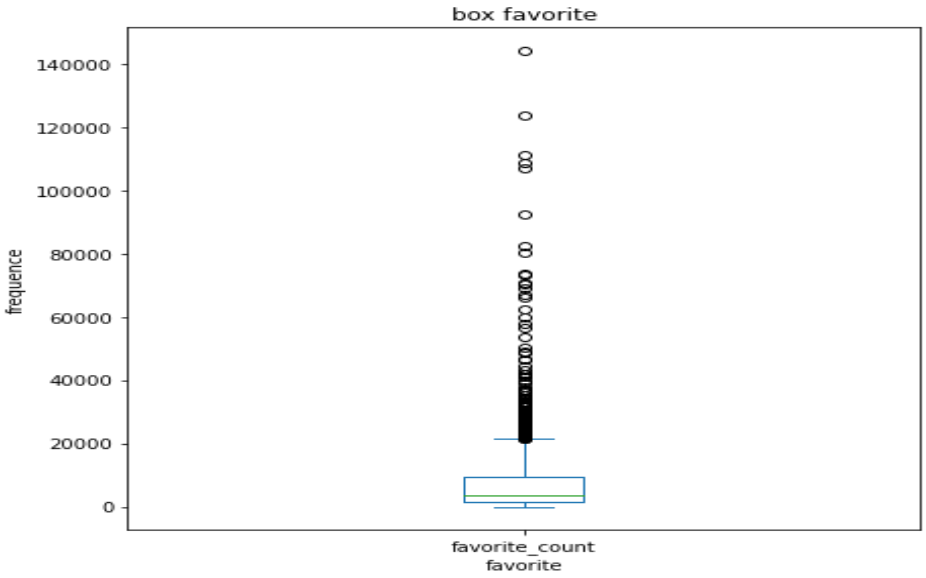
La moitié des internautes retweetent en moyenne 1300 fois les images des chiens de races. On remarque cependant que certaines valeurs sont extrêmes (valeurs abérrantes) mais ils ne le sont qu'en apparence car une étude de ces valeurs aberrantes montrent qu'il s'agit des chiens de race et peuvent être plusieurs sur la photo.

b- <u>Quelle est la répartition des notes?</u>



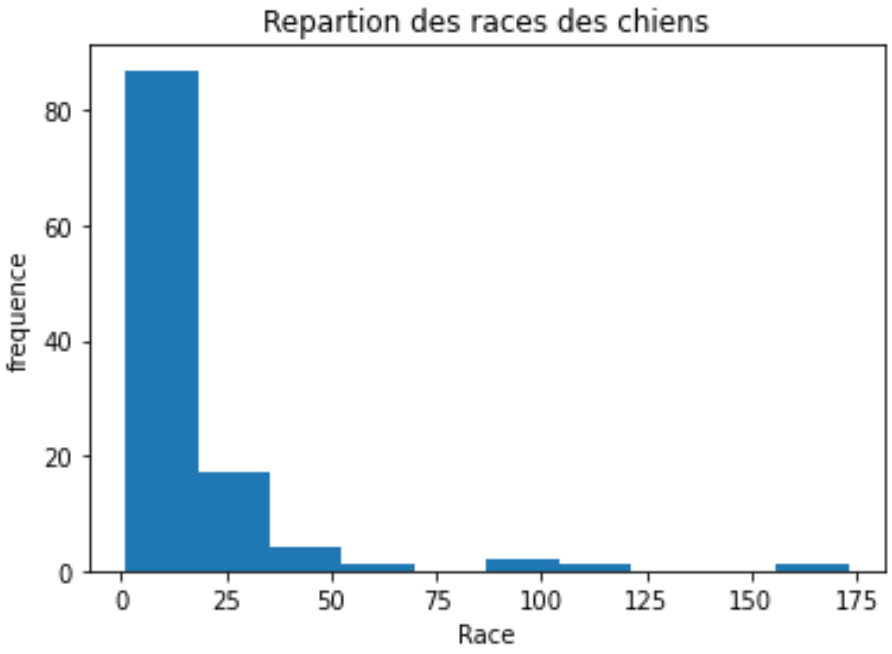
On remarque ici que 25% des notes sont comprises entre 0,8 et 1,2,75% des notes sont comprises entre 1,4 et 1,8 on remarque qu'il y'a des notes abérrantes de 2,8 et 3 et en dessous de 0,8.

c- Quelle est la répartition des favoris ?



En moyenne nous remarquons que les internautes like les photos des chiens trois fois. Pour certains chiens ces notes peuvent aller jusqu'à plus de 140000 fois.Le nombre de like compte des valeurs aberantes(140000 fois).

d- Quels sont les différentes catégories de race de chiens :

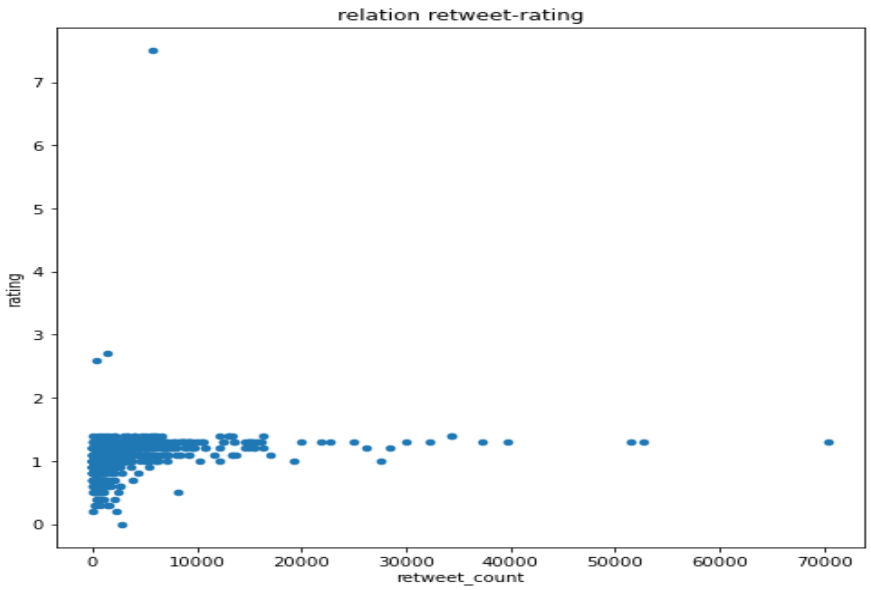


Il ressort de cette observation que les races de chiens les plus réprésentées golden retriever,Labrador retriever,Pembroke,Chihuahua,pug sont-ils alors pour autant les mieux notés?

2- **Analyse Bi variée :**

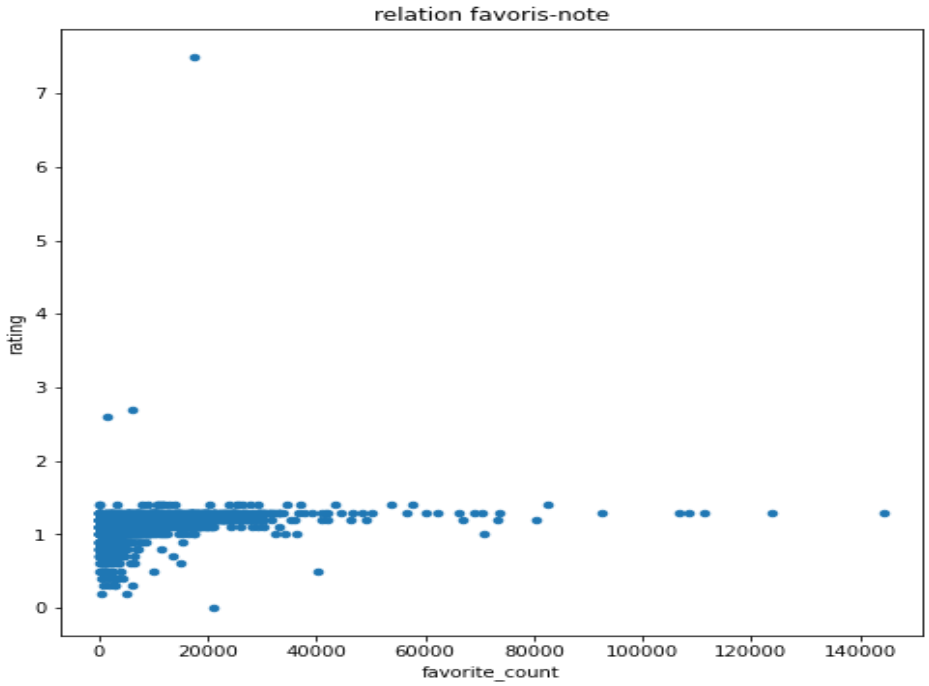
Il s'agit pour nous d'analyser les relations entre deux variables.

a-Relation entre le nombre de retweets et les notes



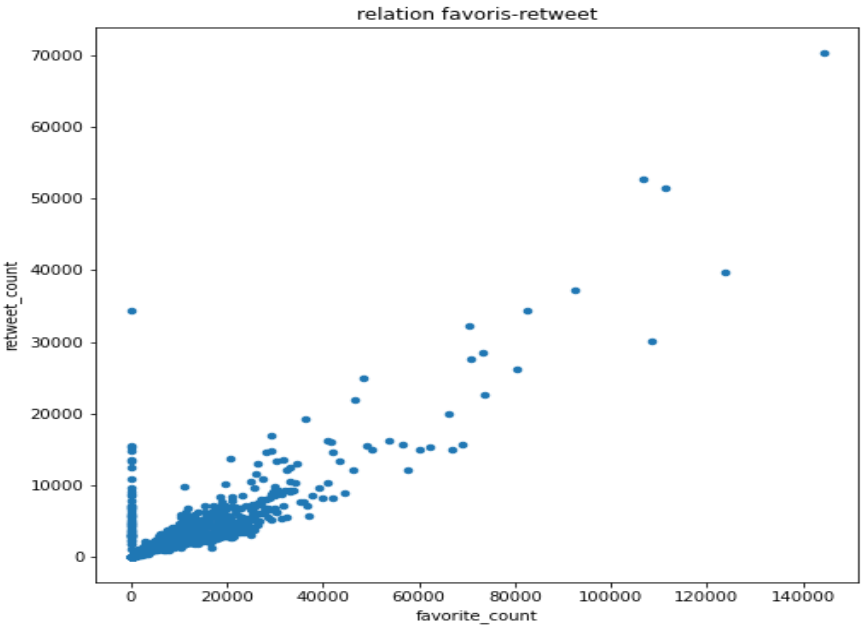
Les notes de chiens croissent avec les retweet. Lorsque le nombre de retweet augmente les notes ont tenancs à aller vers la hausse. Celle prouve que le nombre de retweet influence la note obtenue par image. En revanche à plus de 6000 cette note a tendance à devenir constante et se situe à 1.3.

b-Relation entre le nombre de likes(favoris) et les notes :



On remarque qu'il n'existe pas une corrélation linéaire entre le nombre de like et de retweet c'est pas forcément parce qu'ils aiment une photo qu'il le retweet.Par contre on remarque que plus les likes sont élevés plus les internautes les notes bien et cela devient constant à partir de 1500 likes environ.

c- Relation entre le nombre de like et le nombre de retweet :



Le nombre de likes influencent grandement le nombre de retweet.Plus une image est aimée plus il a des chances de le retweeter.

CONCLUSION

Nous avons en définitive essayé de comprendre les répartitions des échantillons qui ont été soumises à notre étude. Nous avons en deuxième analyse étudié la relation entre deux de ces variables et parvenus à la conclusion que les notes qui se présente comme notre variable principale est dépendante non seulement des retweet mais aussi du notre de ligne. Cette remarque nous permet de connaitre comment organiser une campagne de communication sur les réseaux sociaux.