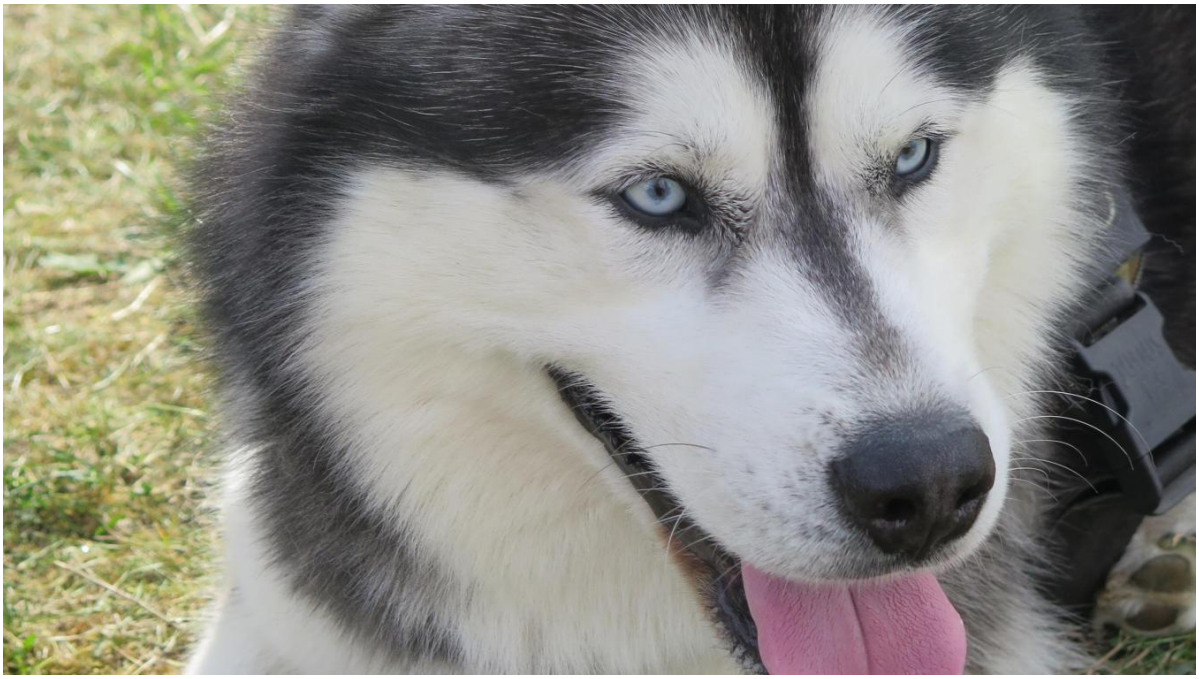
Projet 2 :

Analyse et Visualisation de l'ensemble des données de l'archive de tweets « *WeRateDogs* »



RAPPORT ACT

CONTEXTE DU PROJET

Pour le compte du deuxième projet de notre formation d'analyste de données, il nous est proposé de mener l'analyse et la visualisation d'un ensemble des données provenant de l'archive de tweets ***WeRateDogs.*** WeRateDogs est un compte Twitter qui évalue les chiens des gens avec un commentaire humoristique sur le chien. Cet ensemble de données est réparti dans trois (03) datasets différentes à savoir :

- L'archive twitter améliorée de WeRateDogs «***twitter-archive-enhanced.csv*** »;
- Les données supplémentaires regroupés dans le fichier « ***tweet-json.txt*** » ;
- Le fichier de prédiction d'image « ***image-predictions.tsv*** »

Ce rapport présente les étapes de l'analyse et de la visualisation de quelques variables de l'ensemble de données propre issus des données originelles traitées.

ANALYSE DES DONNEES

L'analyse de cet ensemble des données a été minimisée à trois observations.

Observation 1 : Quel stade de chien a reçu le plus de favoris ?

stage	favorite_count
doggo	80
floofer	8
pupper	218
puppo	24

Nous pouvons observer que les images qui sont au stade de « pupper » sont celles qui ont récolté le plus de favori.

Observation 2 : combien d'image de chien ont été marqué comme favoris ?

```
False    2346
True        8
Name: favorited, dtype: int64
```

Nous observons que seul huit (08) chiens ont été marqué comme favoris contre 2346. On peut déduire que la quasi-totalité des chiens n'ont pas reçu de favoris.

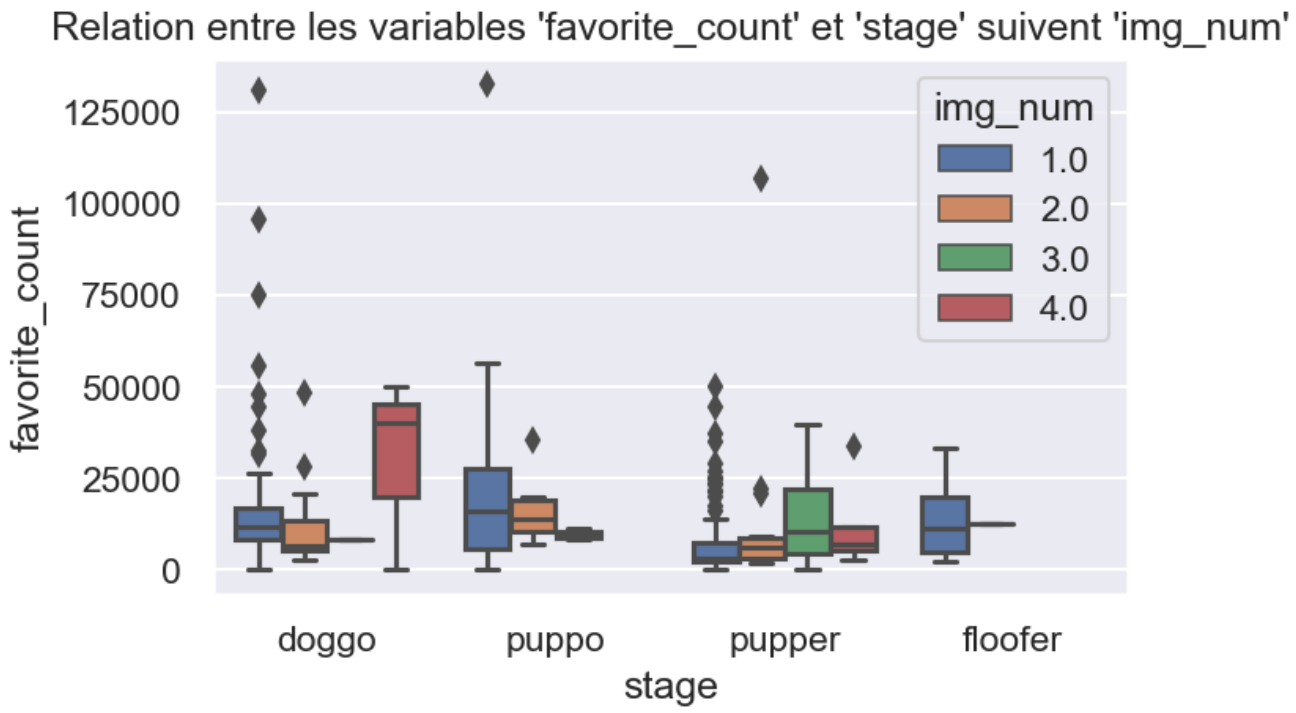
Observation 3 : quel numéro d'image a reçu le plus grand nombre de favoris ?

img_num	favorite_count
1.0	1779
2.0	197
3.0	66
4.0	31

Comme on pouvait s'y attendre, les premières images des retweets ont reçu le plus grand nombre de retweet. Cela confirme que se sont bien les images numéro 1 de chaque retweet qui ont une prédiction sure.

VISUALISATION DES DONNEES

Pour cette analyse, nous avons produit une seule visualisation car la visualisation n'était pas l'objectif central du projet.
Le graphe présenté ci-dessous met en évidence la relation de corrélation entre ne nombre de favori, les stades de chien et leur numéro.



Nous pouvons souligner quelques observations à partir de cette visualisation :

1. Les troisièmes images des tweets ne sont classées qu'au stade « pupper » ;
2. Il n'existe pas de corrélation entre les premières images et les stades de chien ;
3. Les quatrièmes images ont reçu un plus grand nombre de favori au stade « doggo » avec un nombre de favori médian au tour de 40000.
4. Aucune des images de deuxième rang n'a atteint le stade « flooter ».