



Analyse de la population des chats et des chiens entre 2017 et 2020

Projet tutoré - Vincent Godard

Althéa FEUILLET - Karla KUDRIC - Laurent WU

G2M - 2023-2024



Introduction

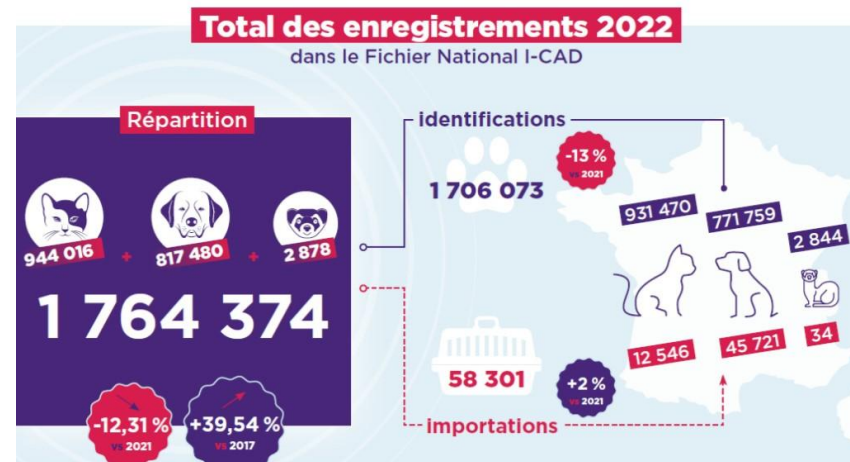
Évolution de la population des chats et des chiens par commune :

I-CAD :

I-CAD a pour mission de gérer le Fichier National d'Identification des Carnivores Domestiques en France

2022 :

- 1.764.374 animaux enregistrés par l'I-CAD
- 13% en plus par rapport à l'année précédente
- + d'importance envers l'identification des animaux



Problématique

Analyser la population des chats et des chiens entre
2017 et 2020

Données

- Données I-CAD
- Données saisies manuellement, donc non uniformes
- Données hors France
- Couche communes

Annee	Espece	Code_postal	Ville	Population
2018	CHIEN	68000	COLMAR	5331
2018	CHIEN	AUTRES	SUISSE	4479
2018	CHIEN	19300	EGLETONS	482
2018	CHIEN	29170	FOUESNANT	1428
2018	CHIEN	26120	CHABEUIL	1217
2018	CHIEN	32390	STE CHRISTIE	67
2018	CHIEN	30250	FONTANES	222
2018	CHIEN	45370	CLERY ST ANDRE	637
2018	CHIEN	59160	LOMME LILLE	4
2018	CHIEN	79000	BESSINES	215
2018	CHIEN	94300	VINCENNES	2076
2018	CHIEN	14500	MAISONCELLES LA JOURDAN	279
2018	CHIEN	54210	AZELOT	38
2018	CHIEN	78200	MANTES LA JOLIE	1479
2018	CHIEN	61800	ST PIERRE D ENTREMONT	73
2018	CHIEN	13360	LASCOURS	8
2018	CHIEN	56520	GUIDEL PLAGES	14
2018	CHIEN	60140	BAILLEVAL	264
2018	CHIEN	51500	RILLY LA MONTAGNE	90
2018	CHIEN	04300	FORCALQUIER	669
2018	CHIEN	61270	ST HILAIRE SUR RISLE	46
2018	CHIEN	27620	BOIS JEROME SAINT OUEN	12
2018	CHIEN	14600	ABLON	273

Tableau 1 :Exemple de données initiales (pour l'année 2018)

I. Data Cleaning



Permet le nettoyage des données à partir de la Base Adresse Nationale (BAN)

Utilisation :

- Pandas pour le traitement des données
- API de la BAN pour le géocodage

Optimisation :

- Programmation asynchrone (asyncio et httpx)
- Docker pour l'hébergement de l'API local
- Mise en cache des requêtes (cachetool)

Résultat :

- Traitement environ 10 heures pour une année
- Contient quelques erreurs

II. Data Correction

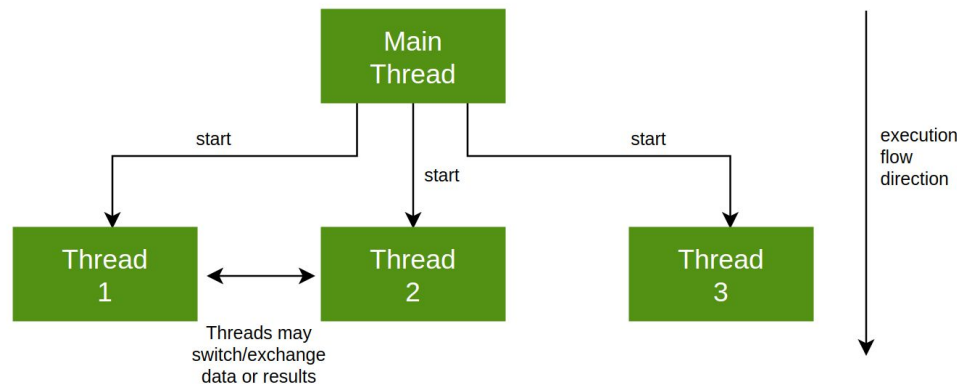
Permet la correction des données mal nettoyées de la BAN

Utilisation de :

- Pandas pour le traitement des tableaux
- Rapid fuzz pour la comparaison
- Joblib pour le multithreading

NLP = Natural Language Processing

Multithreading Programming



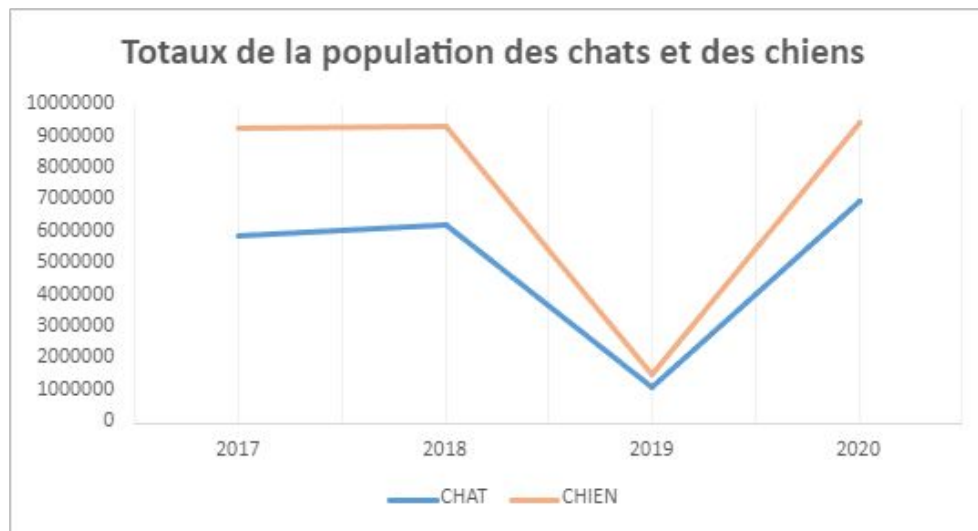
Science, B. O. C., & Science, B. O. C. (2023, 5 mai). The difference between Asynchronous and Multi-Threading | Baeldung on Computer Science. Baeldung on Computer Science.

<https://www.baeldung.com/cs/async-vs-multi-threading>

III. Data Geocoding

	CHAT		CHIEN	
	BAN/NLP	Manuel	BAN/NLP	Manuel
2017	97,82%	98,11%	97,67%	97,90%
2018	97,86%	98,12%	97,69%	97,87%
2019	98,07%	98,20%	97,61%	97,79%
2020	97,86%	98,11%	97,70%	97,87%

Tableau présentant la qualité du géocodage avant et après traitement manuel



- Le géocodage est correct avant et après le traitement manuel
- On rencontre un problème avec l'année 2019 qui a des totaux de population très faible

IV. Visualisation web

Création d'un tableau de bord interactif (dashboard) pour la visualisation des données

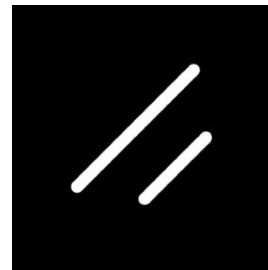
Outils utilisés :



TypeScript

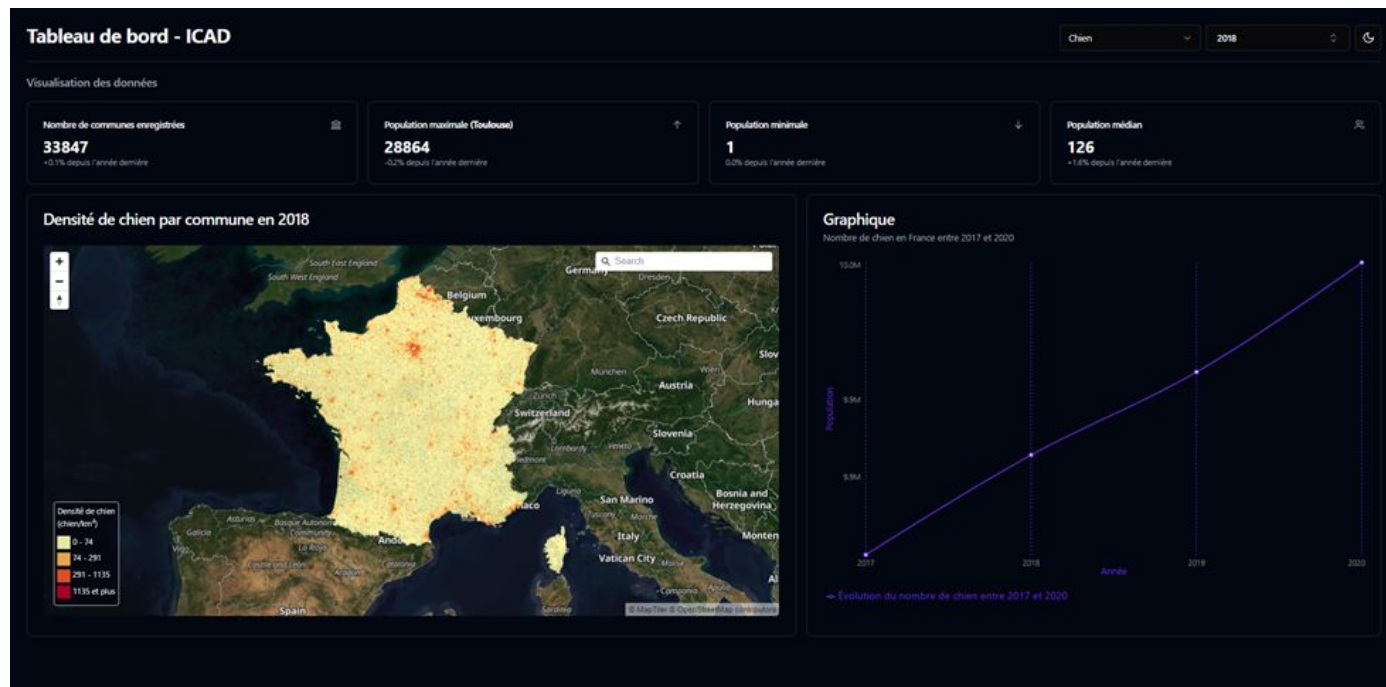


React



Shadcn UI

IV. Démonstration



Lien du site web : <https://wlaurent-hub.github.io/data-viz/>

V. Résultats

Résultats positifs

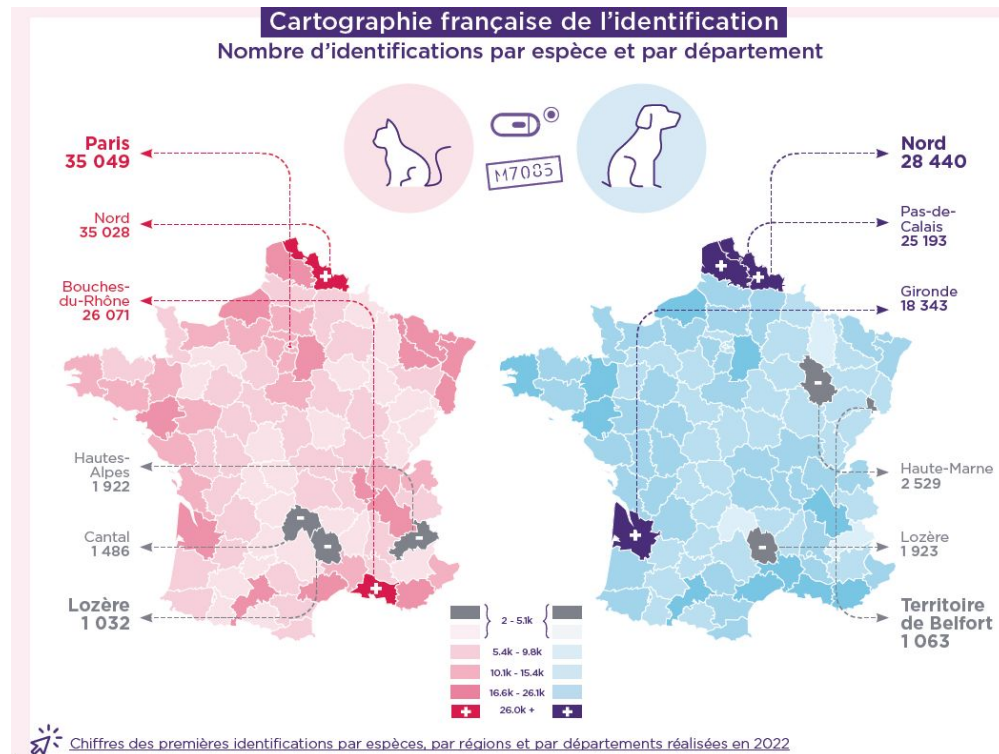
- Résultat précis grâce à la visualisation par commune
- On retrouve un schéma où le Nord de la France et le Sud-Est de la France sont très peuplés

Résultats négatifs :

- Les quartiers d'île de France n'ont pas de valeurs
- Les résultats dans d'autres communes ne sont pas identiques aux résultats de l'I-CAD

Carte de la population de chats et de chiens par département en 2020

(source : <https://www.i-cad.fr/articles/publications>)



VI. Critique

01.

Traitements longs

Datacleaning : 10h

Datacorrection : 2h

02.

Manque de données

De 2013 à 2016

03.

Analyses statistiques non convaincantes

ARIMA et GLM

RECOMMANDATIONS

Création d'un formulaire pour uniformiser les données d'entrées de l'I-CAD

◇ Conclusion

01.

Datacleaning

Problèmes de noms de communes et codes INSEE résolus

02.

Datacorrection

Proximité des noms
Seuil de 80%

03.

Géocodage

Association des communes à des coordonnées géographiques
Représentation spatiale

04.

Visualisation

Exploration dynamique des données
Filtrer les résultats par animal ou par année

- Concentration importante dans les grandes villes
- Manque de données -> 2019
- Proposition de formulaire en ligne pour la collecte des données

MERCI POUR VOTRE ATTENTION !

Sitographie

- Linear Regression Essentials in R - Articles - STHDA. (s. d.). STHDA - Accueil.
<http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/>
- Multiple Linear Regression in R - Articles - STHDA. (s. d.). STHDA - Accueil.
<http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/>
- Tutoriel : GLM sur données de comptage (régression de Poisson) avec R - DellaData. (s. d.). DellaData.
<https://delladata.fr/tutoriel-glm-sur-donnees-de-comptage-regression-de-poisson-avec-r/>

Bibliographie

- Andresen, M. A., Malleson, N., Steenbeek, W., Townsley, M., & Vandeviver, C. (2020). Minimum geocoding match rates : an international study of the impact of data and areal unit sizes. *International Journal of Geographical Information Science*, 34(7), 1306–1322. <https://doi.org/10.1080/13658816.2020.1725015>
- Ratcliffe, J. H. (2004). Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science*, 18(1), 61–72. <https://doi.org/10.1080/13658810310001596076>
- Soatiana RAHARIJAONA. (2023). La géomatique au service de l'étude des signalements de piqûres de tiques sur les chiens et les chats en France entre 2017 et 2019 [Mémoire de master 1 non publié]. Université Paris 8.