## Démarche proposée

L’idée du test est très simple : nous avons des prix, nous avons des caractéristiques. Pouvez vous proposer quelques résultats précis qui permettent d’expliquer les prix constatés ? Pour donner quelques idées, possible d’analyser par dimension géographique, taille, statut,…L’idée, dans l’ordre, et selon le temps dont vous disposez (il est entendu qu’une durée de référence pour faire un test de recrutement est de l’ordre d’une heure) :

* Poser par écrit les questions auxquelles vous souhaiteriez pouvoir répondre. De façon plus approfondie : décrire la démarche souhaitée, ses attendus et ses limites. Sachant que vous n’aurez pas clairement pas le temps de tout faire, cela permet de voir comment vous aborderiez le sujet)
* Pour quelques éléments choisis, proposer et mettre en œuvre une démarche d’analyse en R permettant d’obtenir quelques premiers résultats.
* Signaler quelles seraient vos priorités avec plus de temps

En retour, l’idée est donc de nous envoyer un word (ou équivalent) ainsi qu’un .R au moins avec quelques lignes de code. Le lien vers un repo GitHub fonctionne très bien aussi bien sûr.

# Proposition

Sujet : **l’évolution des prix des chambres d’EHPAD avant le début de la crise sanitaire après le pic de crise en 2022**

Sous-questions : Est-ce que les prix ont flambé avec les protocoles de sécurité renforcés ? Est-ce que les établissements mobilisent désormais plus de main-œuvre ? Est-ce que des disparités existent par département ?

1. **Qu’il y a-t-il effectivement dans les données ? Quelles clés d’appariement sont disponibles ? Quelles variables sont utilisables sur l’ensemble de la base ?**

Les données fournies portent sur le prix des EHPAD de 2018 jusqu’à début 2024. On peut donc traiter le sujet avant-après pic COVID. Les données brutes de prix des EHPAD doivent être jointes avec la clé Finess pour obtenir davantage d’information. Les données récentes Finess sont très désagréables à utiliser mais 2019 et 2022 peuvent s’ouvrir avec un code standard pour les .csv. On remarque qu’avec l’appariement le SIRET des établissements est disponible, **cela nous permet de récupérer la tranche d’effectifs de l’établissement** via un appariement à SIRENE (disponible sur data.gouv.fr) ou via un appel API SIRENE v3. De même, nous récupérons le département. => Nous pouvons donc, en principe, traiter l’ensemble des questions.

1. **Peut-on aller dans le détail des prestations EHPAD liées au COVID ? Quels pourcentage de NA sur les variables clés ?**

Les données de détail des prestations ne permettent pas vraiment de creuser les potentiels prestations supplémentaires que le COVID a institué, même s’il y a une variable à champ texte libre qui pourrait contenir des informations, mais sûrement au mieux médiocre pour notre étude.

On surveille le % de NA soit par des tableaux, soit par au cours du travail via le package *tidylog.* On perd une 20ène d’établissements qui n’ont pas de finess correspondant entre les deux bases de données, sur +6000 établissements c’est acceptable. La base SIRENE est exhaustive et il n’y a pas de perte lors de l’appariement, mais les variables de tranches d’effectifs ne sont pas systématiquement remplies (important % de « 00 » ou de « »).

Les données prix nécessitent de léger recodages.

1. **Quels résultats ?**

* **Est-ce que les prix ont flambé ?**
  + Non a priori, +11% au max sur 3 ans dans le 92 il y a pire, moyenne autour de 7-6-5%.
* **Disparités départementales ?**
  + Oui et assez fortes, parfois du simple au double, notamment en île de France
* **Augmentation des effectifs ?**
  + Rien de visible à partir des données SIRENE

1. **Quelles limites ?**

Données SIRENE insuffisamment précises et parcellaires (énormément de NA sur la variables tranche d’effectifs). Il doit y avoir un moyen plus efficace pour récupérer l’information en termes d’effectifs (SIASP 😉).

1. **Priorités ?**

Expertiser les données manquantes, trouver des sources à appareiller, améliorer les connaissances sur le système médico-social en France.