☼SUNBURST ☼

S2.06 Analyse de données, reporting et datavisualisation

RAPPORT FINAL

Présentation de l'ensemble de la SAE et des tâches réalisées

14/06/2024

Groupe : Galton

Thevasenan Satujan
Avenel Kyllian
Diallo Thierno
Tep Mathieu

Résumé

Dans ce rapport, nous expliquons nos différentes tâches réalisées lors de cette SAE. Lors du jalon 1, nous avons développé nos compétences dans la compétence Traiter où nous avons exploré la base de données. Le jalon 2 consistait à développer nos compétences dans la compétence Traiter et Analyser en analysant la base de données pour réaliser une étude de marché. Et pour le jalon 3, nous avons reformulé notre étude pour expliquer nos résultats non pas à un enseignant mais à un client.

Summary

In this report, we explain our various tasks completed during this project. During phase 1, we developed our skills in the "Processing" (Traiter) competency where we explored the database. Phase 2 involved developing our skills in both the "Processing" (Traiter) and "Analyzing" (Analyser) competencies by analyzing the database to conduct a market study. And for phase 3, we reformulated our study to explain our results not to a teacher but to a client.

I. Compte rendu étape 1 (traiter)

Dans cette étape, nous avons commencé par décrire la base données issus de l'INSEE et le nettoyage éventuel à effectuer.
Nous avons observé que ce document présente plusieurs thèmes socio-économiques allant de la démographie à l'emploi, en passant par les entreprises, les revenus et le logement. Chaque thème est abordé à travers une série de variables décrivant différents aspects comme la population par âge et genre des communes de France.

Pour faire une description complète de cette base de données, nous avons réalisé plusieurs tâches et sous-tâches.

Tâche 1 : Compréhension de la base de données

Dans un premier temps, nous avons cherché à comprendre la base de données. Pour cela, nous avons réalisé une lecture sur Excel pour une meilleure compréhension. Nous avons compris que le fichier présente deux bases de données de l'INSEE, encodées en UTF-8 sous format CSV, qui décrivent les données sur plusieurs indicateurs. Nous avons aussi observé que cette base de données est parue le 27/02/2024.

Il est séparé en deux fichiers, un fichier "**dossier-complet**" contenant plusieurs informations dans lequel chaque ligne représente une commune et un fichier "**meta_dossier_comple**t" décrivant toutes les variables de notre fichier "dossier-complet" ainsi que toutes leurs modalités et leurs types.

Le fichier "**dossier-complet**" possède 35001 lignes et 1883 colonnes, où chaque colonnes représentent une variable.

Pour avoir une meilleure compréhension de la base de données. Nous avons ouvert le fichier meta_dossier_complet et on a étudié les entêtes des colonnes, on a eu une meilleure compréhension du fichier.
C'est à ce moment qu'on a compris qu'il y avait plusieurs thèmes sur notre fichier et que les thèmes étaient une colonne importante dans notre base de données.

Pour avoir une meilleure compréhension de la base de données, Thierno et Kyllian ont déterminé le sens des variables existant dans notre base de données. Pendant ce temps Mathieu et Satujan ont créé un google doc ou on pourra mettre l'explication des variables présents dans chaque thème de notre base de données.

Tâche 2 : Compréhension des variables de la base de données

Suite à la compréhension de notre base de données, nous avons commencé par avoir une certaine compréhension des variables de notre base de données. Satujan et Mathieu ont

analysé les variables du premier thème tandis que Kyllian et Thierno ont analysé le deuxième thème. Nous avons aussi fait des recherches externes pour comprendre qu'est ce que nos données représentais

Tâche 3 : Description des variables de la base de données

Tout d'abord, nous avons identifié les variables pour chaque thème, la base de données de l'Insee comporte 12 thèmes, nous sommes quatre dans l'équipe, nous nous sommes répartis de la façon suivante :

Thème de la base de donnée	Décrit par
Évolution et structure de la population	Mathieu et Thierno
Caractéristique de l'emploi au sens du recensement	Satujan et Kyllian
Couples - Familles - Ménages	Mathieu
Évolution et structure de l'entreprise	Thierno
Diplôme et Formation	Kyllian
Caractéristiques des établissements	Satujan
Évolution des revenus et pauvreté des ménages	Thierno
Naissances et décès domiciliés	Thierno
Population active, emploi et chômage au sens du recensement	Satujan
Salaire et revenus d'activité	Kyllian
Tourisme	Kyllian
Logement	Mathieu

Tâche 4 : Rédaction du rapport et nettoyage du fichier

Après avoir fait une description des variables, Satujan et Mathieu ont réalisé un google et ont rédigé le rapport du Jalon 1.
Pendant ce temps Thierno et Kyllian, ont réalisé une activité cible et des variables pertinentes pour notre jalon 2.
Satujan et Mathieu ont rédigé la description des variables de façon plus détaillée pendant que Thierno avec l'aide de Kyllian a rédigé un script python pour extraire le nom des villes du docs meta et commencer le nettoyage du fichier.
Satujan et Mathieu, encore en duo, ont rédigé l'explication du nettoyage du fichier.
Ensuite, collectivement nous avons relu le rapport et apporter les corrections nécessaires.
Après avoir rendu ce rapport, Thierno a réalisé des tâches supplémentaires pour rendre notre gestion de groupe plus propre avec notamment l'automatisation de la feuille de temps de groupe. Ensuite, vu qu'on avait quelques incompréhensions sur le code, Thierno nous a expliqué le code.
Nous avons pris 89h en groupe pour terminer le jalon 1.

II. Compte rendu étape 2 (traiter et analyser)

A. Définir la problématique

a. Problématique

Quel est le lieu optimal pour installer mon entreprise d'aide aux personnes âgées ?

b. Contexte

Un jeune entrepreneur souhaite développer son entreprise d'aide aux personnes âgées par l'accompagnement et le soin de cette population.
L'aide aux personnes âgées désigne un ensemble de mesures et de services mis en place pour accompagner les seniors dans leur vie quotidienne et maintenir leur autonomie. Elle s'adresse aux personnes âgées en perte d'autonomie, qu'elle soit physique, cognitive ou sociale.
Ainsi le jeune entrepreneur se demande quel est le lieu optimal pour installer son entreprise d'aide aux personnes âgées ?

Par optimal il faut comprendre une ou plusieurs communes ayant :

- des revenus stables
- des revenus assez haut
- des revenus faiblement dispersés
- une proportion de personnes âgées plutôt élevé
- une proportion de personnes âgées vivant seules
- une proportion de personnes ayant un diplôme dans le domaine de la santé (CAP, jusqu'à BAC+2)

c. Objectifs

Les objectifs de l'aide aux personnes âgées sont multiples :

- Permettre aux personnes âgées de vivre à leur domicile le plus longtemps possible en leur apportant l'aide nécessaire pour accomplir les actes essentiels de la vie quotidienne (lever, coucher, toilette, repas, habillage...)
- Prévenir la dépendance en favorisant le maintien des capacités physiques et cognitives des personnes âgées
- Soutenir les aidants familiaux en leur apportant du répit et des conseils
- Améliorer la qualité de vie des personnes âgées en leur permettant de rester autonomes et actives

d. Services proposés

L'aide aux personnes âgées se présentera sous différentes formes au sein de son entreprise :

- Aide à domicile : intervention d'un aide-ménagère ou d'un auxiliaire de vie pour l'aide à la toilette, au ménage, aux courses, à la préparation des repas...
- Livraison de repas : livraison de repas à domicile équilibrés et adaptés aux besoins des personnes âgées
- Aide à la téléassistance : installation d'un système d'alerte qui permet aux personnes âgées de contacter un service d'assistance en cas de problème

e. Facteurs clés de localisation et de marché

Afin de mener à terme son projet, cet entrepreneur désir savoir dans quel endroit doit-il s'installer pour développer son activité.
Cette question peut être décomposée en plusieurs questions plus simples, à savoir :

- Quelle est la population cible ?
- Quelle est la zone avec le plus de personnes âgées ?
- Dans quel endroit il y a le moins de concurrents ?
- Quelle est la meilleure zone pour faciliter le recrutement de personnel ?
- Quel lieu facilite le déplacement des personnes âgées et à mobilité réduite (pour les activités) ?
- Quel est le niveau de revenu de vie des populations (famille) ?

f. Nos contraintes

Pour nous le meilleur endroit doit respecter les critères suivants :

- avoir un minimum de 1000 habitants
- avoir le maximum de clients possible
- le maximum de mains d'oeuvres possible
- le minimum de concurrents possible
- Pouvoir se déplacer facilement (transport, vélo,...) [voir classeur1.csv]
- Avoir une pension de retraite supérieure à 0 car nous voulons des communes ayant au minimum des pensions pour les personnes âgées.

g. Spécificité de notre problématique

- Proposer des activités aux personnes dans la semaine permettant de les faire sortir, faire des exercices physiques ou loisirs

- Installation d'un "système d'alarme" avec une télécommande par exemple qui va dire si le patient a un problème et lui attribuer une personne pour le soigner ou l'aider.
- Assistance de la personne mais pas 24h/24 exemple 3h le matin, 2h l'après-midi et 2 le soir (permet de baisser les coûts)

h. Hypothèses

1. Plus il y a de personnes âgées dans une ville, plus il y aura de chances de trouver des personnes âgées vivant seules dans cette même ville.
2. Plus le nombre de ménage avec personnes âgées vivant seules est élevé, moins il y aura des inégalités de richesses, car ils sont en majorité inactifs
3. Plus le revenu médian est haut, plus la pension des retraites sera élevé
4. Lorsque l'indice de dépendance calculé est élevé, la pension des retraités suit cette augmentation en étant aussi élevé
5. Le meilleur lieu pour implanter une entreprise d'aide aux personnes âgées sont les villes reculées car les retraités ont tendance à vouloir séjourner pour profiter dans des lieux calme et éloignés des métropoles.

B. Description de la population cible

Pour répondre à notre problématique, nous avons défini des populations ciblées.
Avec tout d'abord :

a. Les Clients potentiels

Qui ? : Les personnes âgées, personnes âgées seules
Où ? : Grâce à plusieurs indicateurs, on doit déterminer la bonne ville, à domicile
Âge : au delà de 60 ans
Quelles sont leurs habitudes, leurs loisirs : casanier, jeux de société, activités physiques, ...
Quels sont leurs problèmes : Personnes seules, malades (dépression, santé,...) handicapées,....)

b. Les profils de nos futures employées

Sans emploi (chômeurs actif)
Dynamique (personne d'âge jeune à partir de 20 ans)
Personne ayant fait des études dans le domaine de la santé (BEP, CAP, BAC +2)

Profil des individus : en bonne santé, dévoué et sens de l'accompagnement

c. Contraintes démographique

Chaque ville doit avoir au moins 1000 habitants.

d. Contraintes géographique

Des départements, régions ou communes qui facilitent le transport (transport en commun) et d'autres critères comme une ville avec beaucoup de potentiel clients

e. Contraintes revenues

Famille avec un revenu assez élevé pour pouvoir se payer nos services.
Chaque ville doit avoir une pension de retraite supérieure à 0%. Nous avons décidé de choisir une pension de retraite supérieure à 0% pour avoir suffisamment de villes dans nos analyses, de plus une pension inférieure n'est pas pertinente dans notre étude.

C. Description des variables

Pour réaliser notre étude de marché, nous avons créé les variables suivantes :

1. Proportion des personnes de 60 ans et plus dans les communes.

Le code de cette variable est **proportion_60p,** elle représente la proportion des personnes de 60 ans ou plus dans chaque commune.
Cette variable est quantitative continue. Elle a été créée à partir des variables population globale en 2020 puis la même variable pour les individus mais avec des tranches d'âge différentes de 60 à 74 ans, 75 à 89 ans et 90 ans et plus.
Nous avons utilisé une formule pour créer cette variable statistique : (P20_POP6074 + P20_POP7589 + P20_POP90P) / P20_POP
Cette formule calcule le nombre total de la population ayant un âge de 60 ans ou plus, puis fait le rapport (divise par P20_POP) avec la variable "population globale en 2020" pour obtenir la proportion des individus de 60 ans ou plus.

2. Proportion des individus de 75 ans et plus dans les communes

Le code de cette variable est **proportion_75p,** elle représente la proportion des personnes de 75 ans et plus dans chaque commune. C'est une variable quantitative continue. Elle a été créée à partir des variables population globale en 2020, puis la même variable pour les individus mais avec des tranches d'âges différentes de 75 à 89 ans et 90 ans ou plus.
La formule pour créer cette variable statistique est la suivante : ((P20_POP7589 + P20_POP90P) / P20_POP)
Cette formule fait la somme des individus ayant une tranche d'âge de 75 à 89 ans et 90 ans ou plus puis fait le rapport avec la variable "population globale en 2020" pour obtenir la proportion des individus de 75 ans ou plus.
La variable a pour modalité potentielle les valeurs de 0 à 1.

Nous souhaitons utiliser cette variable pour déterminer si les personnes âgées installées dans une ville s'y établissent à long terme. En comparant les villes ayant des proportions élevées de personnes de plus de 60 ans à celles ayant des proportions élevées de personnes de plus de 75 ans, nous pourrons identifier si les personnes âgées ont tendance à s'installer durablement dans ces villes.

3. Proportion des individus de 55 ans et plus vivant seuls

Nous avons nommé cette variable **proportion_55p_s.**
Cette variable représente la proportion des personnes de 55 ans ou plus vivant seules dans chaque commune. C'est une variable quantitative continue. Elle a été créée à partir des variables population globale en 2020 puis la même variable pour les individus qui vivent seules mais avec des tranches d'âges différentes de 55 à 64 ans, 65 à 79 ans et 80 ans ou plus.
La formule pour créer cette variable statistique est la suivante : (P20_POP5564_PSEUL + P20_POP6579_PSEUL + P20_POP80P_PSEUL) / P20_POP
Cette formule fait la somme des individus ayant une tranche d'âge de 55 ans ou plus vivant seuls, ensuite nous faisons la rapport avec la variable de la population globale en 2020 pour obtenir la proportion des individus vivant seuls

4. Proportion des individus de 65 ans et plus vivant seuls

Nous avons nommé cette variable **proportion_65p_s.**

Cette variable représente la proportion des personnes de 75 ans ou plus vivant seules dans chaque commune. C'est une variable quantitative continue. Elle a été créé à partir des variables suivante : P20_POP, P20_POP6579_PSEUL, P20_POP80P_PSEUL
La formule pour créer cette variable statistique est la suivante :
(P20_POP6579_PSEUL + P20_POP80P_PSEUL) / P20_POP.

5. Nombre moyen de personnes de plus de 60 ans dans un ménage.

Nous avons nommé cette variable **nb_moy_60p.**
Cette variable représente le nombre moyen de personnes de plus de 60 ans dans un ménage dans chaque commune. C'est une variable quantitative continue. Elle a été créé à partir des variables suivante : P20_MEN, P20_POP6074, P20_POP7589, P20_POP90P
La formule pour créer cette variable statistique est la suivante :
((P20_POP6074 + P20_POP7589 + P20_POP90P) / P20_MEN)
Cette formule permet de faire la somme des personnes ayant 60 ans ou plus. Puis fait le rapport avec le nombre de ménages pour l'année 2020 pour enfin obtenir le nombre moyen de personnes de plus de 60 ans dans un ménage dans chaque commune.

6. Variables utilisées

Nous avons estimé que les variables présentes dans le fichier initial n'avaient pas besoin de modification. Nous les utilisons pour déterminer le niveau de vie des populations dans chaque commune.

- PPEN21 : variable quantitative discrète représentant la part des pensions de retraite dans chaque ville est le pourcentage que représentent les pensions, retraites et rente. La part des pensions, retraites et rentes dans le total des revenus fiscaux de la zone géographique observée.
- MED21 : variable quantitative discrète représentant le salaire médian dans chaque ville
- RD21 : variable quantitative discrète représentant le rapport interdécile des salaires dans chaque ville

7. Indice de dépendance

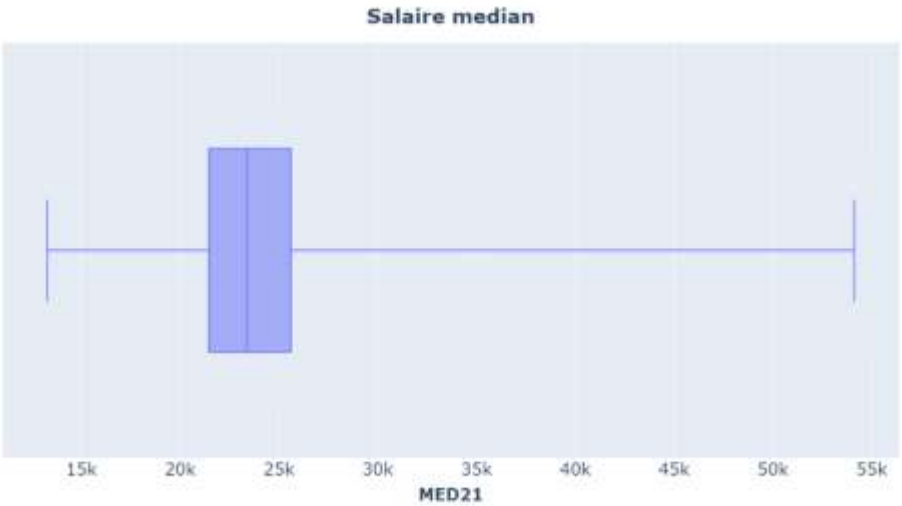
Cette variable permet de mesurer le niveau de dépendance d'une commune, permettant ainsi d'indiquer un besoin élevé ou non de services d'aide à domicile pour les personnes âgées. Elle nous permet de déterminer la proportion de personnes âgées de plus de 60 ans par rapport à celle des jeunes de plus de 15 ans (15 à 59). C'est une variable quantitative continue. Elle est obtenue à partir des variables

suivantes : P20_POP6074, P20_POP7589, P20_POP90P, P20_POP1529, P20_POP3044, P20_POP4559.
Formule : (P20_POP6074 + P20_POP7589 + P20_POP90P) / (P20_POP1529 + P20_POP3044 + P20_POP4559).

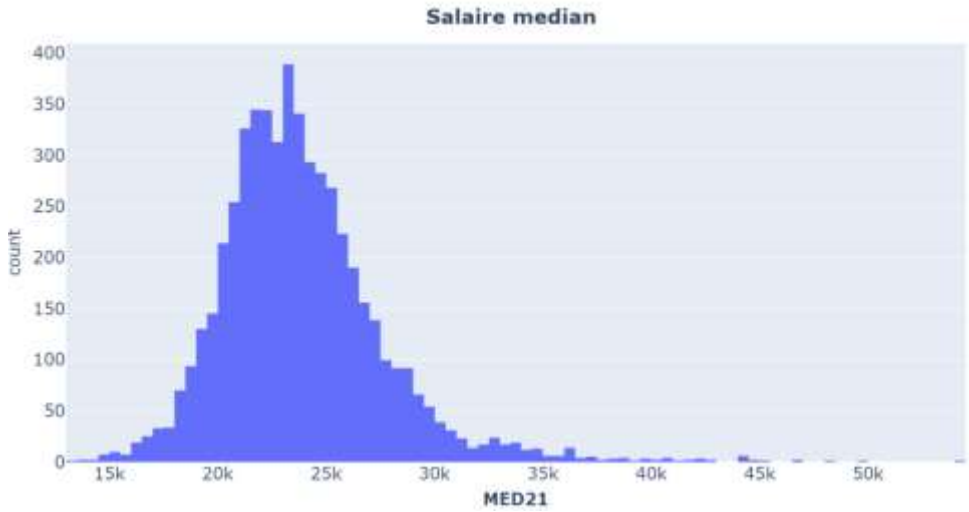
Plus cette proportion est grande, plus il est probable qu'il y ait un fort besoin pour aider les personnes âgées. Grâce à cette variable, si l'on constate qu'il y a de nombreuses personnes âgées et peu de jeunes, alors cette commune aura besoin de services d'aide à domicile.

D. Analyse descriptive simple.

1. Revenu médian des villes : MED21

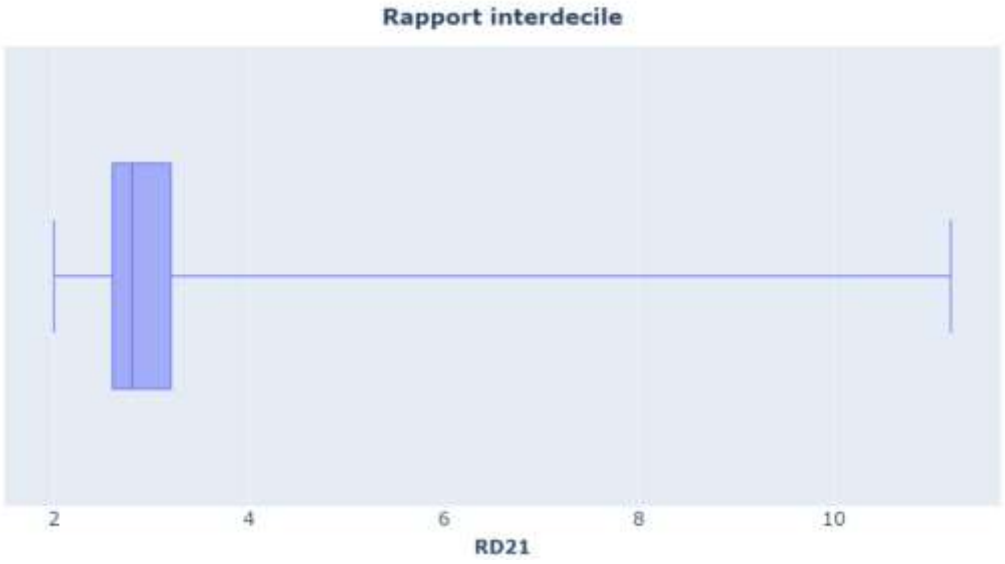


Cette boîte à moustache présente la distribution du salaire médian par commune. Le salaire médian minimum est de 13 283 € et le salaire médian maximum est de 54 120 €, ce qui montre un très grand écart entre les salaires. Cependant, le premier quartile est de 21 450 €, le troisième quartile est de 25 600 €, et la médiane est de 23 370 €, ce qui indique que 50 % des communes ont un salaire médian très proche de cette valeur, ce qui témoigne un revenu assez similaire entre les villes, hormis celle qui sont très éloignées. Le salaire étant un indicateur utilisé dans le calcul du niveau de vie, nous pouvons supposer que le niveau de vie semble assez similaire dans ces communes.



De plus, l'histogramme des salaires médian a une allure de cloche (forme de la loi normale) et est très concentré autour de la moyenne (qui est très proche de la médiane). Cela indique que les villes avec un salaire médian plus élevé et plus faible que les quartiles ne sont pas nombreuses et que la majorité des villes ont un salaire médian très proche de Q1 et Q3. On conclut donc que le salaire médian est relativement le même dans les villes et tourne autour des 23 000€.

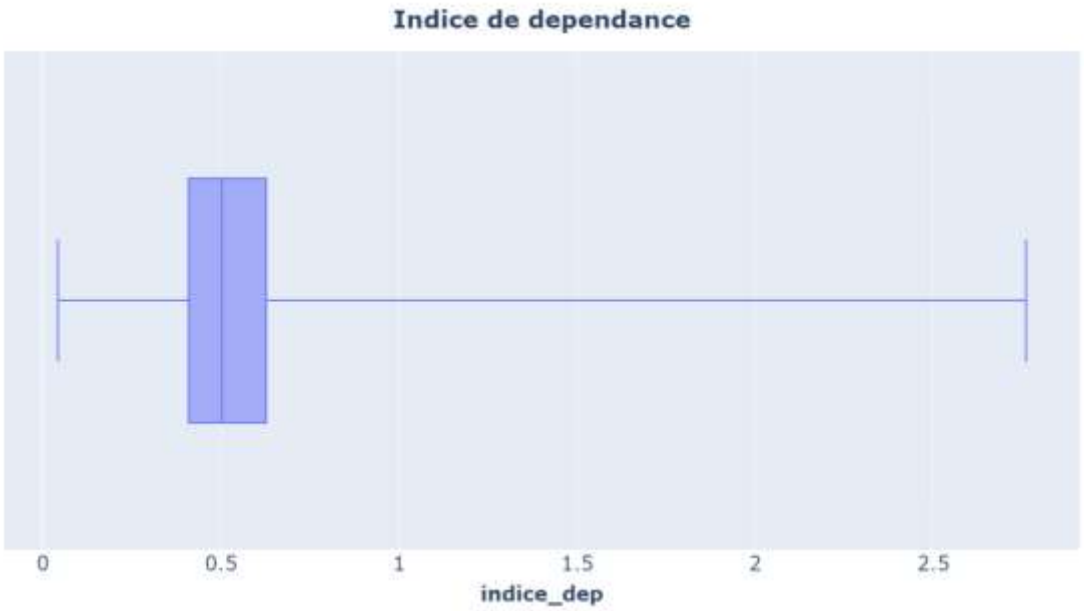
2. Rapport interdécile : RD21



Le rapport interdécile est un indicateur statistique utilisé pour mesurer les inégalités de revenus au sein d'une population. Dans notre cas, on a un rapport interdécile (rd) minimale de 2, ce qui signifie, dans toute les communes, le revenu des 10 % les plus riches de la population est au moins deux fois plus élevé que le revenu des 10 % les plus pauvres, ce qui est significative et montre une certaine inégalité de revenu dans les communes. Nous constatons que le troisième quartile est de 3,18 donc les 75% des communes ont un rd plus petit que 3. Toutefois, il y

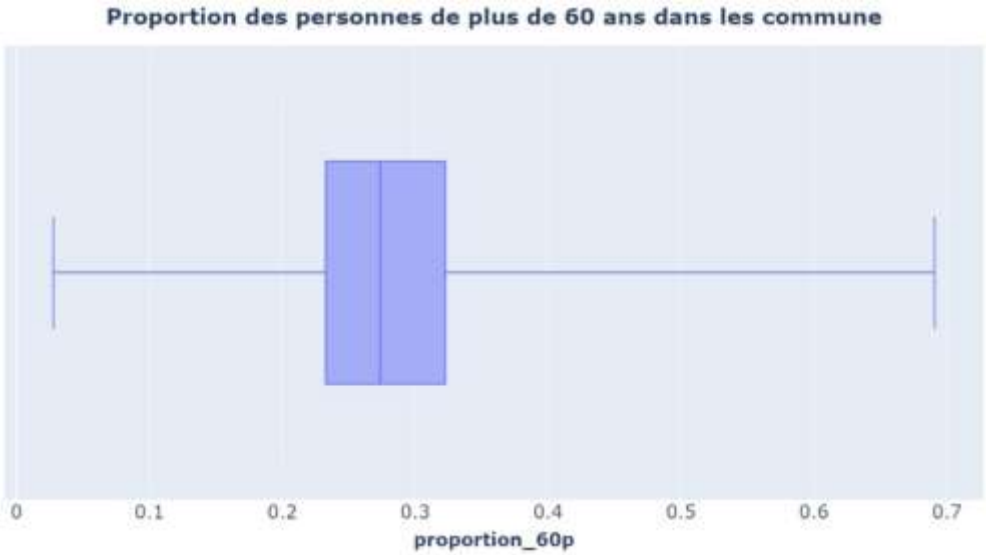
a des communes avec un rd de plus de 11 (max), qui témoigne d'une très grande inégalité de revenu dans les populations.

3. Indice de dépendance



Cette variable indique le niveau de besoin de services d'aide à aux personnes âgées à domicile dans les communes.
Elle va nous permettre de savoir quelle est la proportion de personnes âgées par rapport au jeune, s' il y a de nombreuses personnes âgées, et pas assez de jeunes, alors cette commune aura besoin d'entreprise de la sorte. Pour la variable indice_dep, la boite à moustache se place entre les valeurs 0 à 2.76 représentant ainsi le minimum et le maximum. De plus, nous remarquons que 75% des valeurs sont en dessous de 0.6.
Cela signifie que dans 75% des communes, il y a beaucoup plus de jeunes que de personnes âgées. On remarque que dans le reste (25%) l'indice de dépendance est grand, ce qui indique qu'il manque de jeunes pour s'occuper des personnes âgées.

4. Proportion des plus de 60 ans



La variable proportion_60p représente la proportion des personnes de notre population possédant 60 ans ou plus dans les communes.
La boîte à moustache se situe entre la valeur minimale et maximale de nos données soit entre 0 et 0.7 Tout d'abord, nous pouvons apercevoir que notre boite se trouve approximativement au centre de notre axe des abscisses soit environ 0.3 avec le premier quartile se trouvons à 0.23, le troisième quartile environ à 0.32 et une médiane de cette boite à environ 0.27.

Pour l'histogramme de cette variable, nous avons donc une importante significativité des occurrences dans la même tranche de valeurs que nous avons donné via les quartiles, nous apercevons une importante croissance des occurrences au niveau du premier quartile montant jusqu'à la valeur extrême soit environ 182, puis cette dernière se voit décroître à partir du troisième quartile pour ainsi avoir vers la valeur max des abscisses une très faible occurrences. Ainsi, cela vient montrer tout de même une importante présence des personnes ayant 60 ans et vivant dans les communes.

5. Proportion des 75 ans et plus

Pour la variable proportion_75p, elle représente la proportion des personnes qui ont 75 ans ou plus et vivent dans une commune.
Pour la boite à moustache, nous avons la boite se situant à environ 0.1. D'après nos valeurs et la représentation, nous avons le premier quartile, la médiane ainsi que le troisième quartile sachant que chacune de ses valeurs sont assez proches telle que, le premier quartile est de 0.075 et le troisième quartile est de 0.12.

D'après l'histogramme, nous avons une concentration des valeurs autour de
0.1. C'est donc aussi le cas pour l'histogramme qui d'après la représentation,
nous apercevons cette même concentration des occurrences. De plus, nous
voyons une tendance croissante jusqu'à la valeur de la médiane, puis suivant
la médiane nous avons une tendance décroissante dans quasiment l'ensemble des occurrences de la représentation. Ainsi, d'après, les

occurrences totales des personnes ayant 75 ans et vivant dans une commune, nous pouvons dire qu'il y a tout de même une forte concentration
des personnes de 75 ans ou plus vivant dans les villes de nos données.

6. Proportion des personnes de 55 et plus vivant seuls.

La variable proportion_55p_s représente la proportion des individus de notre population ayant 55 ans et plus qui vivent seules dans les communes.
La boîte à moustache montre la valeur minimale et maximale de nos données soit entre 0 et environ 0.3.
Nous pouvons observer que le premier quartile se trouve vers 0.07, le troisième quartile est d'environ 0.12. La médiane de cette boite à moustache est d'environ 0.085.

Pour l'histogramme de cette variable nous avons donc une importante significativité des occurrences dans la même tranche de valeurs que nous avons données via les quartiles, nous apercevons une importante croissance des occurrences lorsqu'on se rapproche du premier quartile

montant jusqu'à la valeur extrême soit environ 395 puis cette dernière se voit décroître un peu avant le troisième quartile pour ainsi avoir vers la valeur max des abscisses une très faible occurrences allant jusqu'à 0 lorsque x se rapproche de 0.3. Ainsi, cela vient montrer tout de même une importante présence des personnes ayant plus de 55 ans et vivant dans les communes.

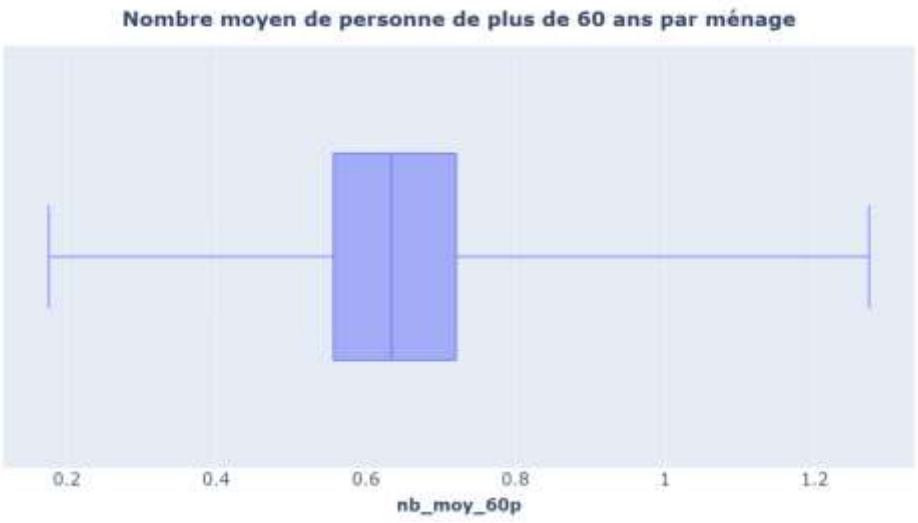
7. Proportion des personnes de 65 ans vivant seuls

La variable proportion_65p_s représente la proportion des individus de notre population ayant 65 ans et plus qui vivent seules dans les communes. La boîte à moustache montre la valeur minimale et maximale de nos données soit entre 0 et environ 0.23. Nous pouvons observer que le premier quartile se trouve vers 0.04, le troisième quartile est d'environ 0.075. La médiane de cette boite à moustache est d'environ 0.058.

Pour l'histogramme de cette variable, nous avons donc une importante significativité des occurrences dans la même tranche de valeurs que nous avons données via les quartiles, nous apercevons une importante

croissance des occurrences lorsqu'on se rapproche du premier quartile montant jusqu'à la valeur extrême soit environ 250 puis cette dernière se voit décroître un peu avant le troisième quartile pour ainsi avoir vers la valeur max des abscisses une très faible occurrences allant jusqu'à 0 lorsque la proportion des personnes se rapproche de 0.15. Ainsi, cela vient montrer tout de même une importante présence des personnes vivant seules ayant 65 ans et plus dans les communes.

8. Nombre moyen de personnes de plus de 60 ans



La variable nb_moy_60p permet de donner le nombre moyen de ménages de notre population ayant 60 ans ou plus vivant seules dans les communes.
Nous pouvons observer que le premier quartile se trouve vers 0.5, le troisième quartile est d'environ 0.7. La médiane de cette boite à moustache est d'environ 0.65. On observe qu'il y a 50% des valeurs entre 0.6 et 0.7 montrant qu'en moyenne on a une proportion de personnes âgées par ménage assez faible cependant, il y a certaines communes où cette proportion est forte.

9. Part des pensions de retraite par ville

Pour la variable PPEN21, c'est une variable qui montre la répartition des
pensions de retraite des personnes âgées selon les communes de nos données. Tout d'abord, nous apercevons une boite à moustache qui montre
qu'il y a un premier et troisième quartile ainsi qu'une médiane situées dans un intervalle de valeurs d'environ 20 à 40, pour la boite a moustache
complète située entre environ 10 à 70. Ainsi, ces valeurs vont alors nous
permettre de visualiser les mêmes données mais avec l'histogramme qui
devait être plus ou moins dans les mêmes tranches de valeurs.

A travers l'histogramme, nous retrouvons bien cette répartition. La représentation de l'histogramme montre une forte présence des communes qui possèdent une part de pension de retraite de 20 à 40 et a une forme de cloche. De plus, cette répartition montre que sur un

ensemble de plus de 300 communes qui possèdent une part de pension de retraite de 30 pour l'année 2021. C'est-à-dire que pour ces 300 communes, les 30% de revenus générés par les communes sont distribués aux retraités. Toutefois, il y a aussi certaines communes qui possèdent une part beaucoup plus élevée montrant donc que les revenus générés par certaines communes, sont davantage distribués pour les parts de pensions de retraite.

10. Synthèse des variables univariées

En conclusion, les données révèlent une relative homogénéité des niveaux de vie dans la plupart des villes, avec une majorité des salaires médians proches de 23 370€ et un RD moyen de 2.9.
Certaines communes affichent des disparités de richesse marquées. L'indice de dépendance et les proportions d'habitants âgés montrent une population majoritairement équilibrée entre jeunes et personnes âgées, bien que les besoins en services pour personnes âgées soient évidents dans certaines communes.
La présence significative de personnes âgées vivant seules suggère une demande potentielle pour des services de soutien.
Les pensions de retraite et le revenu médian soulignent l'importance des revenus des personnes âgées, c'est-à- dire qu'il ont assez d'argent pour pouvoir se payer des services d'aide à domicile. Globalement, ces analyses indiquent une relative stabilité économique avec des zones d'inégalité et des besoins spécifiques pour les services aux personnes âgées.

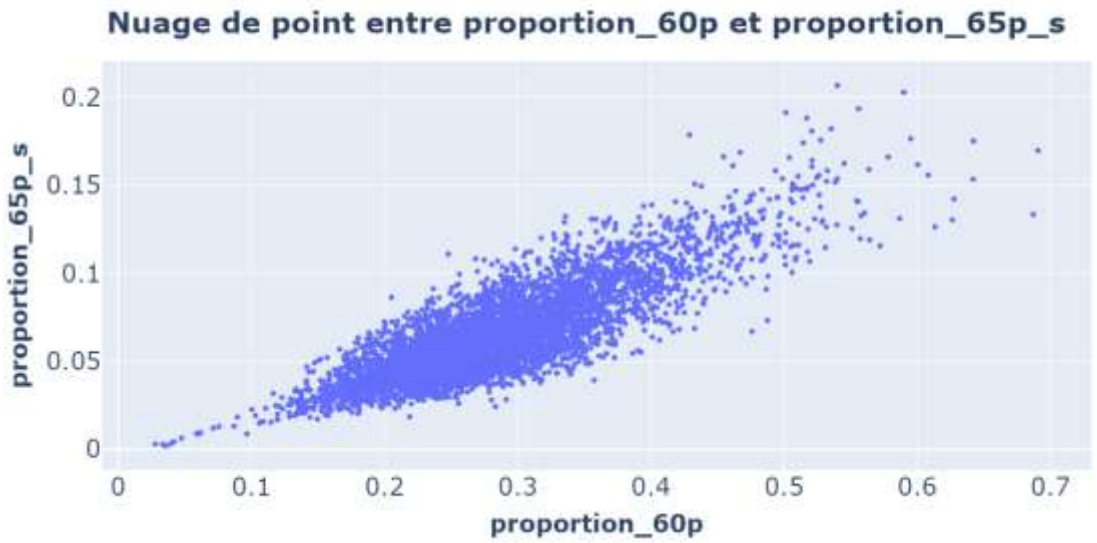
E. Analyse multivariée

Rappel des hypothèses présentées (en I-h) :

- Plus il y a de personnes âgées dans une ville, plus il y aura de chances de trouver des personnes âgées vivant seules dans cette même ville.
- Plus le nombre de ménage avec personnes âgées vivant seules est élevé, moins il y aura d'inégalités de richesses, car les personnes âgées sont en majorité inactifs
- Plus le revenu médian est haut, plus la part des pensions des retraites sera élevé
- Lorsque l'indice de dépendance calculé est élevé, la pension des retraites suit cette augmentation en étant aussi élevé
- Le meilleur endroit pour implanter une entreprise d'aide aux personnes âgées se trouve dans les villes reculées, car les personnes âgées préfèrent généralement résider dans des lieux calmes et éloignés des grandes villes.

Relation entre la proportion des personnes de plus de 60 ans et celle des personnes de plus de 65 ans vivent seules

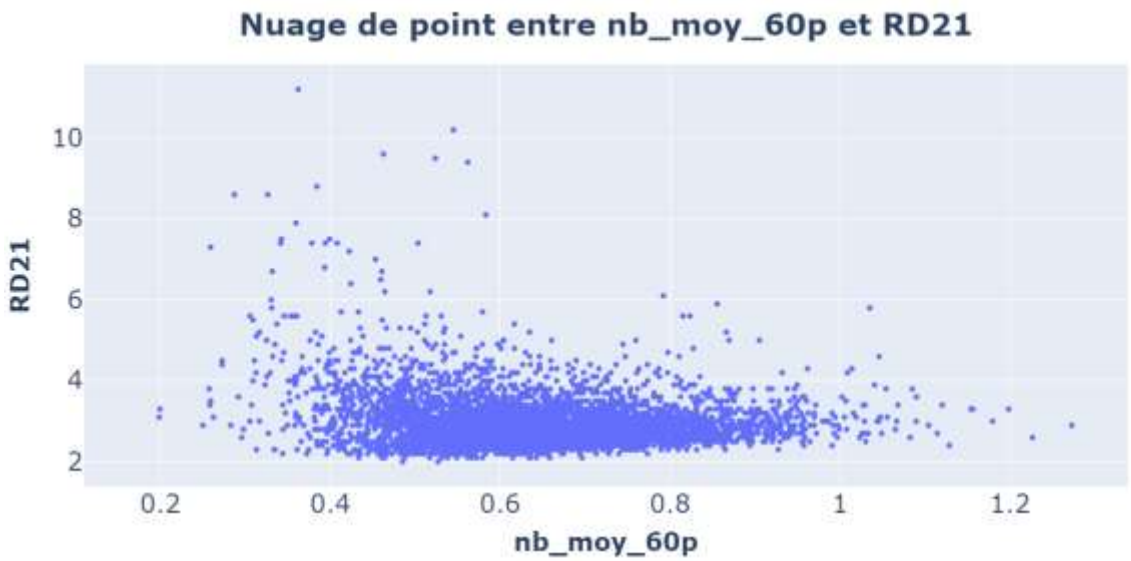
Pour répondre à notre première hypothèse “Plus il y a de personnes âgées dans une ville, plus il y aura de chances de trouver des personnes âgées vivant seules dans cette même ville“. Nous avons étudié la corrélation entre la variable des personnes âgées de plus de 60 ans et des personnes âgées de plus de 60 ans vivant seules.



Dans cette visualisation, on peut constater qu’il y a une certaine relation entre les variables cela se voit à l’aide du nuage de points formant une forme de droite montrant un lien linéaire entre elles. Cela se montre par le nuage de point, qui montre une évolution en fonction de la proportion des personnes ayant 60 ans ou plus. En effet, plus la proportion des personnes ayant 60 ans augmente, plus la proportion des personnes ayant 60 ou plus vivant seule augmente, cela semble donc corrélée linéairement. A l’aide des calculs que nous avons réalisés pour les variables, nous avons trouvé un coefficient de corrélation qui est de 84%. Comme il est proche de 100, il semble donc que les variables aient une corrélation linéaire entre nos variables, confirmant ainsi notre interprétation graphique.

Relation entre le rapport interdécile et le nombre moyen de personnes âgées par ménage

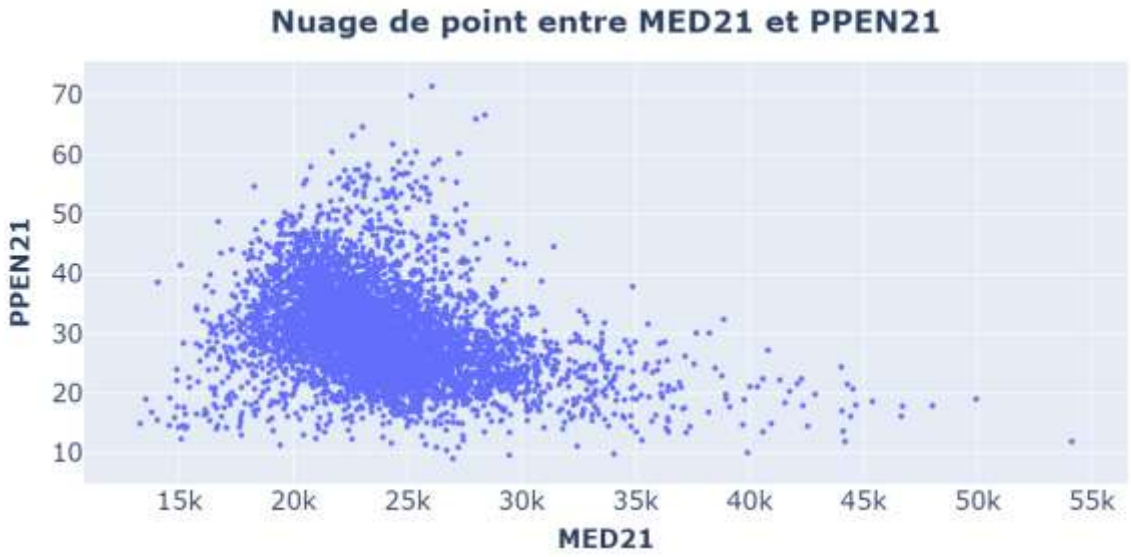
Pour répondre à l’hypothèse “Plus le nombre moyen de personnes âgées dans les ménages est élevé, moins il y aura d'inégalités de richesses, car les personnes âgées sont en majorité inactifs”. Nous avons étudié la corrélation entre les deux variables.



Dans ce nuage de point, on observe que les données **RD21** et le **nb_moy_60p** sont regroupées entre un rapport interdécile de 2 à 4 et un nombre moyen de 0.4 à 1, ainsi les deux variables semblent ne pas être corrélées linéairement, ce qui infirment notre hypothèse. Pour vérifier notre interprétation, nous avons réalisé des calculs de corrélation entre les deux variables. Nous avons obtenu un coefficient de corrélation de -0,18, il semble donc qu'il n'y ait pas de corrélation linéaire entre nos deux variables. Cela vient donc réfuter notre hypothèse.

Relation entre la variable du revenu médian et de la pension de retraite

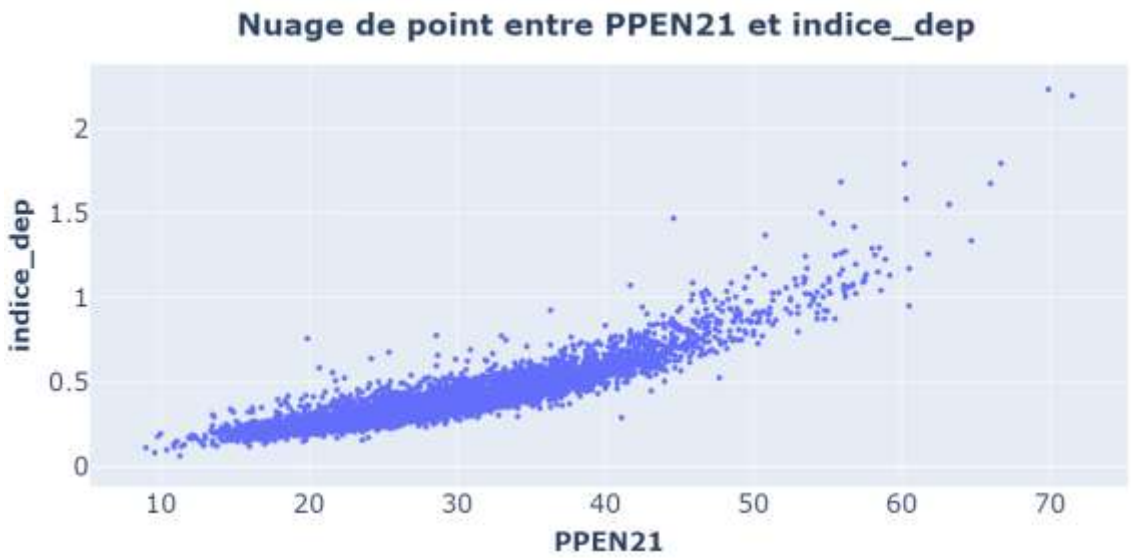
Nous voulons savoir si le revenu médian des personnes influence la part des pensions de retraite. À travers cette hypothèse, nous avons choisi de voir s' il y a une dépendance entre ces variables afin d'en déduire des réponses sur le territoire où nous voulons nous installer.



Ce graphique nous présente la relation entre la variable **MED21** et **PPEN21,** cette visualisation nous permet de voir le revenu médian des personnes en fonction des parts de pensions de retraites, nous observons que les points sont regroupés dans un intervalle de 17 000 à environ 30 000 pour une pension de retraite comprise entre 20 et 45%.
Avec ce graphique, on voit plutôt que, lorsque le salaire médian est compris entre 17 000 et 30 000, la part des pensions de retraite est haute et quand nous observons que le salaire médian est très haut, la part des pensions de retraite devient de plus en plus faible. Ainsi en nous fiant uniquement au graphique, on peut supposer qu'il n'y a pas de relation linéaire entre ces deux variables.
Afin de vérifier notre analyse, nous avons calculé le coefficient de corrélation, nous avons obtenu une corrélation de -0,32. Le coefficient étant proche de 0, il ne semble pas avoir de relation linéaire entre ces deux variables.
Ce qui infirme notre hypothèse qui était la suivante : "Plus le revenu médian est haut, plus la part des pensions des retraites sera élevée" étant donné la relation négative entre ces deux variables, on ne peut pas interpréter les variables entre elles.
Cependant, dans notre analyse nous avons remarqué qu'il y a une forte concentration de points, on pourra en déduire que lorsque le salaire médian est compris entre 17 000 et 30 000, la part des pensions de retraite entre 20 et 45% est très élevée.

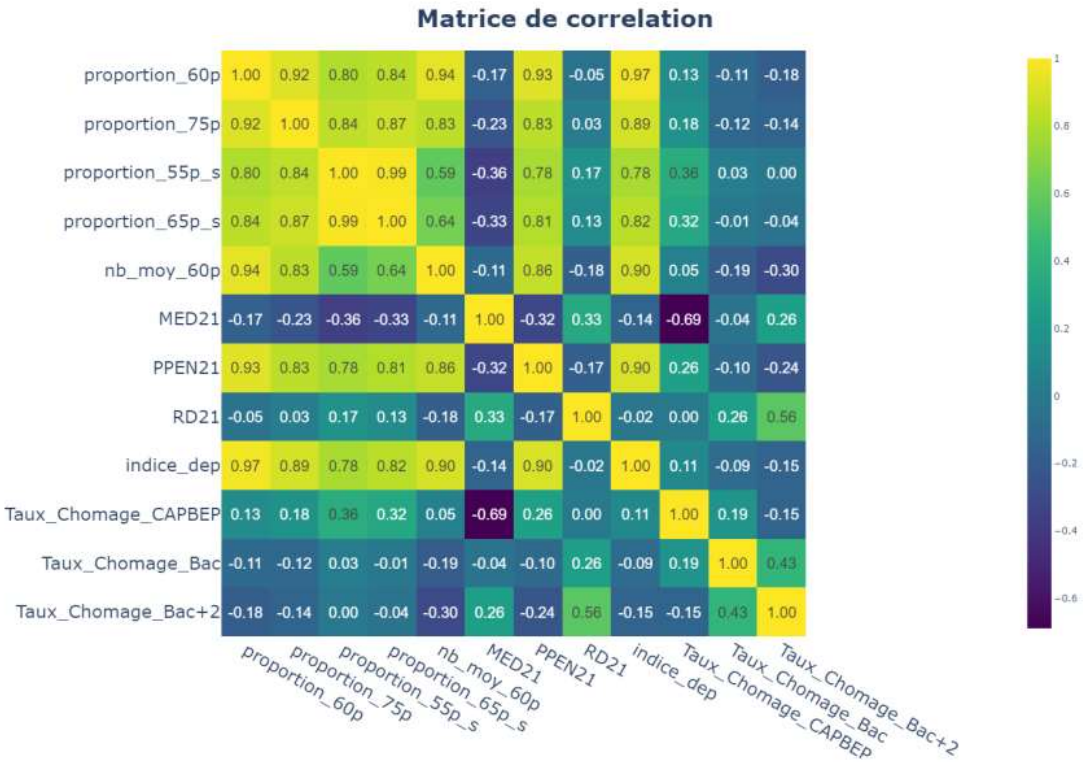
Relation entre la variable indice de dépendance et pension de retraite

Afin de savoir si lorsque l'indice de dépendance est élevé, la pension de retraite suit cette évolution. Nous avons réalisé le graphique suivant.



Dans ce nuage de points, on observe que les points forment presque une droite indiquant qu'il existe une relation linéaire entre la variable **PPEN21** représentant la part des pensions de retraite et la variable **indice_dep** représentant la variable indice dépendance.
Lorsque la part des pensions de retraite augmente, l'indice de dépendance évolue aussi.
Nous obtenons un coefficient de correlation 90%, il semble donc avoir une relation linéaire entre ces variables.
Ce nuage de points affirme notre hypothèse "Lorsque l'indice de dépendance calculé est élevé, la pension des retraités suit cette augmentation en étant aussi élevée ".

Matrice de corrélation des variables



Cette visualisation montre la matrice de corrélation entre nos variables statistiques, à travers cette matrice nous avons la corrélation entre toutes nos variables créées. C'est-à-dire que pour toutes nos variables créées dans l'étape précédente, nous avons l'ensemble des combinaisons des corrélations entre les différentes variables. C'est donc le coefficient de corrélation de Pearson qui indique si les variables mises en jeu sont liées. C'est-à-dire que si le coefficient est proche de 1 ou -1, il semble y avoir un lien entre les variables. Un coefficient de -1, indique une corrélation inverse entre les variables. Enfin, une corrélation proche de 0, indique donc qu'il n'y a aucun lien entre les variables. Nous nous sommes basés sur ce modèle pour étudier les liens entre nos variables permettant d'affirmer ou réfuter nos hypothèses.

$$r = \frac{cov(X,Y)}{\sigma_X . \sigma_Y}$$

$$avec$$

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n}$$

$$et$$

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}}$$

F.Modèle

Dans l’optique de répondre à la problématique posée, nous avons créé plusieurs variables statistique tel que la proportion de personnes de plus de 65 ans ou le nombre moyen de personnes âgées par ménage.

Dans un premier temps, nous avons effectué des analyses univariées pour examiner chacune de ces variables de manière individuelle. Cette étape nous a permis de comprendre la distribution et les caractéristiques principales de chaque variable. Via cette analyse, nous avons pu sélectionner les variables les plus pertinentes et fiables pour notre problématique.

Ensuite, nous avons procédé à des analyses bivariées afin d'explorer les relations et les interactions entre les différentes variables. Cette analyse nous a aidé à identifier les corrélations significatives et les éventuelles dépendances entre les variables. Sur la base de ces analyses, nous avons sélectionné certaines variables clés qui se sont révélées être les plus pertinentes pour notre étude et rejeter certaines variables car elle se comportait de la même manière que certaines variables qui avaient déjà été sélectionnées.

Nous leur avons ensuite attribué des poids spécifiques en fonction de leur importance relative basée sur leur influence sur les résultats globaux et les objectifs de notre problématique. Ces poids permettent de mieux refléter l'impact de chaque variable dans les modèles que nous mettrons en place.

Ainsi, les variables retenues et les poids attribués sont :

Nom variable	*Poids*	*Raison*
proportion_65p_s Personne de plus de 65 ans ou plus vivant seul	: 0.27	Nous voulons des villes avec des personnes âgées vivant seules.
proportion_60p Personne de plus de 60 ans ou plus	: 0.21	Nous voulons des villes avec des personnes âgées
RD21 Rapport interdécile :	: 0.16	Nous voulons des communes avec un niveau de richesse très similaire.
MED21 Médiane de niveau de vie	: 0.16	Nous voulons des villes avec des revenus assez élevés, en la mettant en relation(sélectionner les communes avec le revenu médian le plus élevé) avec le rapport interdécile(sélectionner les communes avec le RD le plus faible), on obtient les villes ou les populations ont un niveau de vie élevé, tout en étant similaire.
PPEN21 Part des pensions de retraite	: 0.13	Pour que nos service soit accessible à la majorité des personnes
Taux_Chomage_Bac+2 Part des chômeurs qui ont au moins un bac+2	: 0.07	Cela pourrait être de potentiel employé, encore plus si ils ont un diplôme dans l'aide aux personnes âgées

Puis, nous avons normalisé les variables pour qu'elles se trouvent toutes dans **[0 ; 1]** en appliquant la formule suivante :

$$X_{new} = \frac{X - min_{col}}{max_{col} - min_{col}}$$

Puis nous avons créé une colonne score avec la formule suivante où n est le nombre de variables n poids est le poids attribué à chaque variables après avoir été normalisée :

$$score = \sum_{i=1}^{n} poid_i.variable_i$$

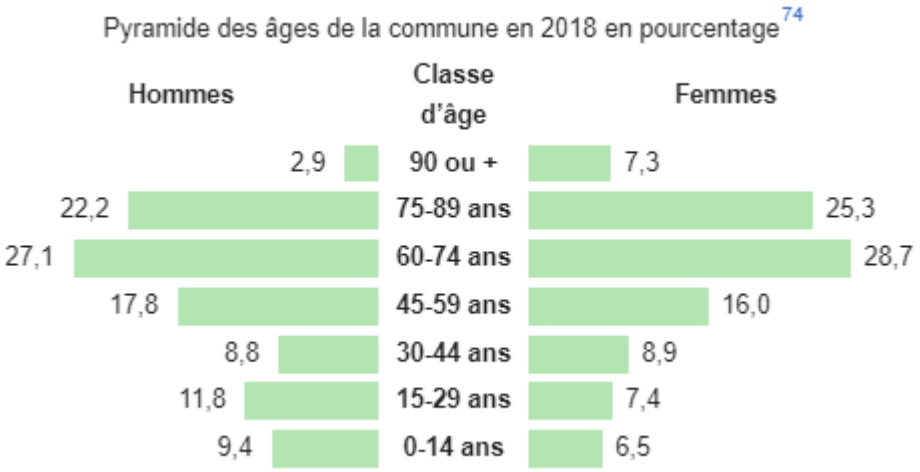
Après cela nous obtenons un classement des villes.

Les 5 communes avec le score le plus élevé sont :

Code commune	***Nom de la commune***	***Score***
33009	Arcachon	0.66
56005	Arzon	0.65
62826	Le Touquet-Paris-Plage	0.64
85114	Jard-sur-Mer	0.62
14220	Deauville	0.60

Vérification du résultat.

- Arcachon est une ville située dans le département de la Gironde, dans la région de la Nouvelle-Aquitaine au Sud-Ouest de la France
- On constate que la ville d'Arcachon sort du lot notamment par son score qui est le plus élevé sur l'ensemble des communes de France mise en jeu.



- Les habitants d'Arcachon (Gironde) gagnent en moyenne 2 611 € nets par mois, soit 31 332 € nets par an
- Emploi

Données 2020	Arcachon	Moyenne des villes
Actifs en emploi	2 607	90,0 %
Chômeurs	507	10,0 %
Inactifs	1 437	13,6 %

G. Tâches Étape 2

- Pour cette étape, nous avons commencé par la recherche de problématique, cette partie nous l'avons fait tous ensemble car chacun de nous avait décrit une partie des variables dans l'étape 1. Mais également, afin de confronter nos idées et de choisir la meilleure. Nous avons fait pareil pour la recherche d'hypothèses. A la fin de cette partie, nous avons choisi la problématique et les hypothèses qui semblaient les plus pertinentes.
- Après avoir choisi la problématique, nous nous sommes tous penchés sur la description du lieu idéale (population, employeur, transport). Etant donné que nous avions une grande base de données, nous avons profité de cette partie pour fixer des contraintes que les villes devaient respecter.
- Ensuite, Thierno et Kyllian ont commencé à rechercher des variables qui serait utile et comment les combiner afin de répondre à la problématique, pendant ce temps, Satujan et Mathieu ont commencé la rédaction et description de la problématique et des hypothèses
- Puis, Thierno à rédiger le script pour la création de variable, de sélection des données en fonction des contraintes, d'analyse (calcul statistique et graphique) univariée et bivariée. Pendant ce temps, les autres membres du groupe s'occupaient du rapport.
- Puis nous sommes passées à l'analyse et l'interprétation des résultats obtenus dans la partie précédente. Chacun, ses occupés d'analyse (3 graphique(et résultat) univariée et de 1 bivariée, puis à fait valider son travail par les autres membres du groupe. Puis ensemble, nous avons décidé quel sont les variables que nous allons conservé pour notre modèle et quel importance nous les attribuons
- Après cela, Thierno réalise le script python pour le modèle.
- Enfin, nous nous somme repartie les différente partie du rapport qui rester a rédiger(2 à 2)

Tâches	Réalisé par
Chercher les problématiques	Thierno, Kyllian, Satujan, Mathieu
Chercher les hypothèses	Thierno, Kyllian, Satujan, Mathieu
Définir les variables utiles	Thierno, Kyllian
Réflexion sur la viabilité de la problématique	Thierno, Satujan, Mathieu, Kyllian
Création de variable statistique en python	Thierno
Analyse univariée	Thierno
Analyse bivariée	Thierno
Début création de model pour choisir la commune	Thierno
Rédaction rapport	Kyllian, Satujan, Mathieu
Correction rapport	Thierno, Kyllian, Satujan, Mathieu
Modification des variables dans le rapport	Thierno, Kyllian, Satujan, Mathieu
Réalisation des corrélations	Thierno
Réflexion sur les hypothèses	Thierno, Kyllian, Satujan, Mathieu
Étude des corrélations	Thierno, Kyllian, Satujan, Mathieu
Rédaction des corrélations	Thierno, Kyllian, Satujan, Mathieu
Réalisation du modèle pour déterminer les villes optimales	Thierno
Rédaction des corrélations entre les variables	Satujan, Mathieu
Rédaction conclusion	Thierno, Kyllian
Vérification de l'entièreté du rapport du Jalon 2	Thierno, Kyllian, Satujan, Mathieu

Nous avons réalisé 126h en groupe pour la réalisation de cette étape.

III. Compte rendu étape 3 (valoriser)

Lien vidéo : https://youtu.be/3BGD3hc3c-s

Tâche 1 : Réalisation du diaporama

L'objectif de l'étape 3 était de réaliser une vidéo présentant nos résultats à un client. Pour ce faire, nous avons créé une présentation Google Slides. Kyllian a choisi le thème. Après interprétation des consignes, nous avons décidé quelles parties de notre rapport de l'étape 2 conserver. Ensuite, nous avons structuré en plusieurs parties le contenu que nous allions présenter. En groupe, nous avons créé un sommaire pour notre diaporama, structuré en plusieurs parties, afin de raconter une histoire cohérente dans notre présentation.

Nous avons collaboré pour élaborer le diaporama et préparer notre présentation vidéo. Chacun a réalisé une partie du diaporama.

Partie du diaporama	Réalisé par
Description du sommaire	Thierno, Kyllian, Satujan, Mathieu
Qui sommes-nous ?	Thierno, Kyllian, Satujan, Mathieu
Présentation du client	Thierno, Kyllian, Satujan, Mathieu
Interrogation sur l'activité	Mathieu
Population et indicateur utilisé	Satujan
Analyse des indicateurs	Kyllian et Satujan
Analyse relationnelle entre les indicateurs	Kyllian
Modèle mis en place	Thierno
Conseil pour le lieu idéal	Thierno et Mathieu

Thierno et Mathieu ont amélioré le visuel du diaporama pendant que Kyllian et Satujan ont commencé la rédaction du rapport final. Kyllian a rédigé les tâches réalisées lors de l'étape 2, tandis que Satujan s'est occupé des étapes 1 et 3.

Avant de commencer la vidéo, Thierno a amélioré le diaporama en ajoutant des animations à chaque diapositive. Ensuite, nous avons fait une dernière correction collective du support visuel pour qu'il soit sans erreurs et clair.

Tâche 2 : Réalisation de la vidéo

Pour la réalisation de la vidéo, nous nous sommes répartis les sections du sommaire du diaporama et avons enregistré une voix off. Pour cette tâche, nous nous sommes rendus dans une salle où nous avons utilisé le micro de notre monteur, Mathieu. Plutôt que de simplement lire notre rapport, nous avons décidé de rendre la présentation plus engageante et claire pour notre client, en expliquant la création de l'entreprise dans le lieu idéal. Nous avons effectué de nombreuses prises pour chaque section, travaillant ensemble pour que chacun puisse donner son avis sur les performances des autres.

Partie	Présenté par
Introduction et Description du sommaire	Kyllian
Qui sommes-nous ?	Satujan
Présentation du client	Mathieu
Interrogation sur l'activité	Thierno
Population et indicateur utilisé	Kyllian
Analyse des indicateurs	Thierno, Mathieu, Satujan et Kyllian
Analyse relationnelle entre les indicateurs	Kyllian, Satujan, Mathieu et Thierno
Modèle mis en place	Thierno
Conseil pour le lieu idéal	Mathieu

Ensuite, Mathieu a réalisé le montage de la vidéo dans lequel il coupait les parties inutiles, les blancs et a rajouté une musique de fond pour que la vidéo soit plus plaisante à regarder. Pour réaliser la vidéo cela nous a pris 6 heures pour la réaliser avec le montage inclus.

La réalisation de la vidéo et du rapport finale nous a pris en tout 87 heures

Conclusion Générale

Pour conclure sur la SAE-206, elle a pu nous permettre d'avoir un aperçu sur la façon dont peuvent se dérouler des projets en équipe mais aussi des projets que nous pouvons réaliser pour la suite de nos études ou carrière professionnelle. Ce projet nous a permis de mettre en avant nos connaissances acquises au cours de l'année que nous avons pu mettre à profit pour cette étude de marché. Notre sujet avait aussi pour but de nous renseigner davantage sur un secteur d'activité, sur des situations et problèmes qui existent de nos jours, de davantage s'instruire. En effet, que ce soit sur notre sujet d'étude ou bien sur les connaissances, l'ensemble de ce projet nous a réellement permis d'améliorer nos compétences. Enfin, il y a un équipier qui nous a abandonnés en cours de route (P.S : on était pas surpris), cependant l'ambiance du groupe a aussi permis d'avoir de sérieuses conversations et de sérieuses idées permettant ainsi d'accomplir notre étude.

Bibliographie

Insee : Base du dossier complet | Insee

Lien vidéo : https://youtu.be/3BGD3hc3c-s