Daulard Mathieu

Projet 8 : L’obésité dans le monde

(Projet personnel)

Sommaire

Introduction

Définition de l’obésité :

Qu’est-ce que l’obésité ? Pour l’adulte, l’OMS définit le surpoids et l’obésité comme suit :

il y a surpoids quand l’IMC est égal ou supérieur à 25

il y a obésité quand l’IMC est égal ou supérieur à 30.

L’IMC est la mesure la plus utile du surpoids et de l’obésité dans une population car, chez l’adulte, l’échelle est la même quels que soient le sexe ou l’âge du sujet. Il donne toutefois une indication approximative car il ne correspond pas forcément au même degré d’adiposité d’un individu à l’autre.
Pour les enfants, il faut tenir compte de l’âge pour définir le surpoids et l’obésité.

Calcul de l’IMC : Il correspond au poids en kilogrammes divisé par le carré de la taille exprimée en mètres (kg/m2).

L’IMC s’applique aux deux sexes et à toutes les tranches d’âge adultes. Il doit toutefois être considéré comme une indication approximative car il ne correspond pas nécessairement au même pourcentage de masse graisseuse selon les individus. L’IMC n’est pas encore utilisable dans le cas des enfants.

Evolution de l’obésité dans le monde :

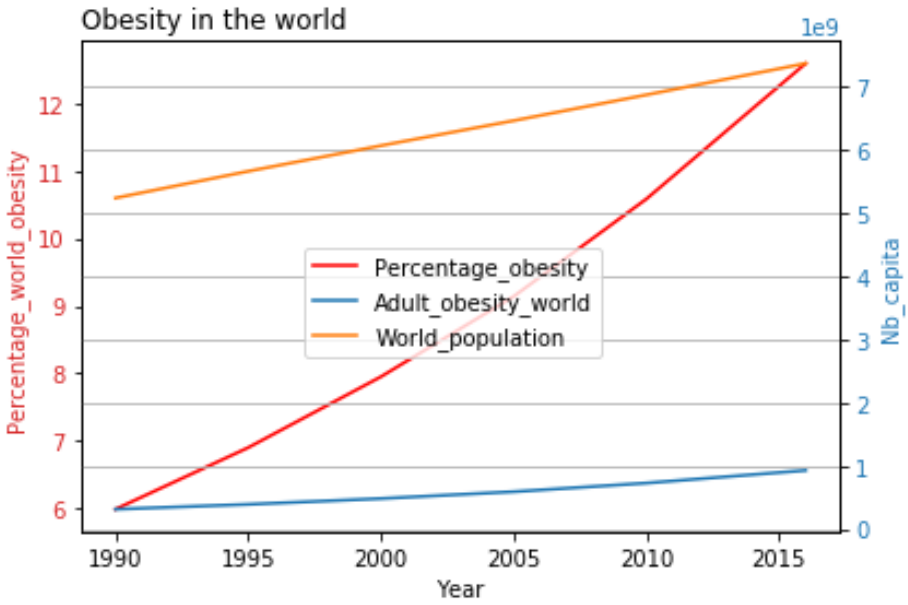


Figure 1: Evolution de l'obésité dans le monde

L’obésité dans le monde a doublé en 25 ans passant de 6 % en 1990 à quasiment 13 % en 2016. Ce phénomène est en croissance forte à l’échelle mondiale et pour l’instant nous semblons incapable de le juguler.

Problématique : Quels sont les facteurs sur lesquels le monde doit intervenir pour réduire l’obésité de sa population ?

Données :

L’obésité est un phénomène qui semble multi-factoriel, il a fallut donc choisir plusieurs approches.

Tout d’abord l’alimentation, comme le recommande le site français Mangerbouger.fr pour équilibrer sa santé, il faut réduire la consommation d’alcool, les aliments gras, sucrés et salés.

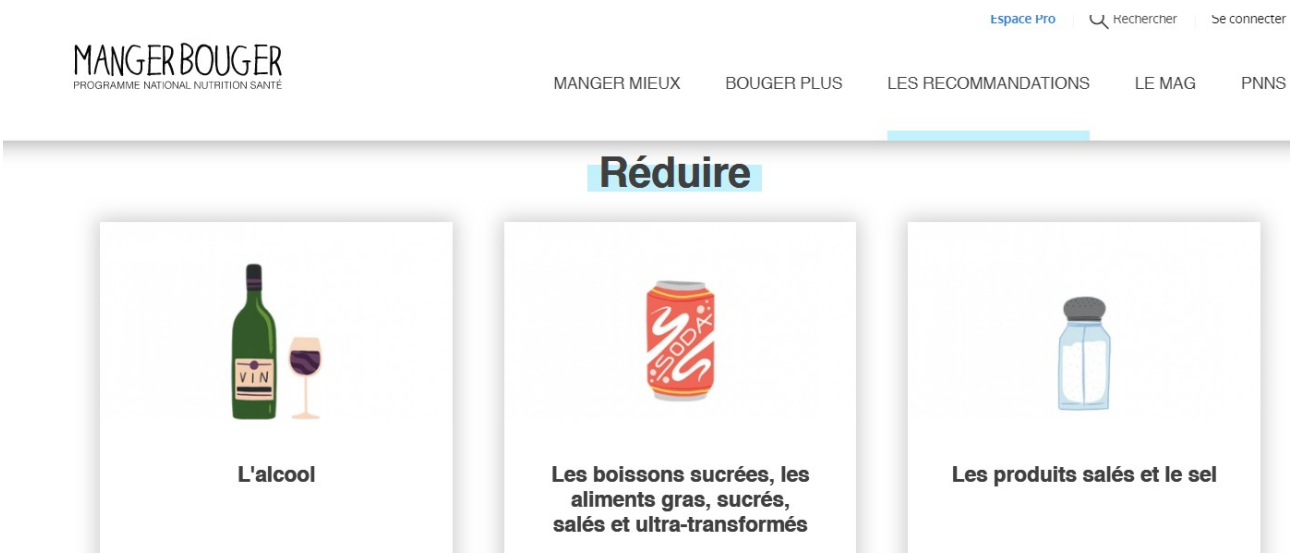


Figure 2: Extrait site mangerbouger.fr

Afin de pouvoir étudier la consommation de ces produits dans le monde, l’étude des bases de données de la FAO semble la plus adaptée.

Figure 3: Sigle FAO

L’Organisation des Nations unies pour l’alimentation et l’agriculture (connue sous les sigles ONUAA ou, plus couramment, FAO soit en anglais *Food and Agriculture Organization of the United Nations*) est une organisation spécialisée du système des Nations unies, créée en 1945 à Québec.

Deux tables ont été créées pour étudier l’alimentation grasse la première est nommées Supply dans laquelle on récupère la consommation d’aliments par pays et par habitant en kcal, en matière grasse et protéine. On pourra donc étudier la part de graisse dans l’alimentation de chaque pays.

La deuxième table dispo alim étudie la part de la viande dans l’alimentation du pays. Sur le même site on parle aussi de réduire l’apport en viande dans notre alimentation, en étudiant la part de la viande dans l’alimentation, on va pouvoir étudier ce phénomène supposé néfaste.

Une table a aussi été créée pour étudier l’impacte de la consommation d’alcool, la table alcool qui donne la consommation d’alcool par pays et par habitant. Ainsi qu’une table undernourrish qui étudie le taux de population sous-alimenté du pays.

Afin d’étudier d’autres facteurs comme le manque de sport, ainsi que l’urbanisation, l’étude de la base de données de la banque mondiale semble aussi pertinente :

Figure 3: Sigle World Bank

Le Groupe de la Banque mondiale est l’une des principales sources de financement et de savoir pour les pays en développement. Il se compose de cinq institutions engagées en faveur de la réduction de la pauvreté, d’un plus grand partage de la prospérité et de la promotion d’un développement durable.

C’est cette banque de données qui a permis de récupérer les taux d’obésité des adultes par pays avec la table adult_obesity_main, ainsi que la table child_obesity_main pour les taux d’obésité des enfants.

La concentration des population semblait un facteur intéressant à étudier par rapport aux changements d’habitudes qu’elle impose aux population. C’est pour cela que l’on a créé la table pop_urban qui étudie le taux d’urbanisation par pays.

Nous avons aussi créé une table poverty, pour étudier l’impact de la pauvreté du pays sur son taux d’obésité. Ainsi que deux tables adult_lsport et child_lsport pour étudier le manque de sport chez les populations adultes, ainsi que chez les enfants. Et une dernière table pour étudier l’impact de la prise des psychotropes sur l’obésité avec la table %Use_of_mental_disorder_substance.

Même si les informations provenant de wikipedia peuvent manquer de précision, en faisant attention on peut trouver des informations intéresssante :

Figure 2: Sigle Wikipédia

Wikipédia est une encyclopédie alimentée sur internet par des volontaires. Universelle, multilingue et fonctionnant sur le principe du wiki, chacun peut y collaborer immédiatement. Wikipédia a pour objectif d'offrir un contenu libre, objectif et vérifiable que chacun peut modifier et améliorer, sans nécessité de s'enregistrer. Tous les articles de Wikipédia sont un *travail en progression* qui peut être modifié et amélioré par tout le monde.

Ici on a pu trouver le nombre d’habitant par restaurant Mac Donald par pays pour connaître l’importance de la présence de fast food dans un pays, Mac Donald étant le plus représentatif. On a donc créé la table Macdo pour étudier la densité des fast food par pays.

Nettoyage :

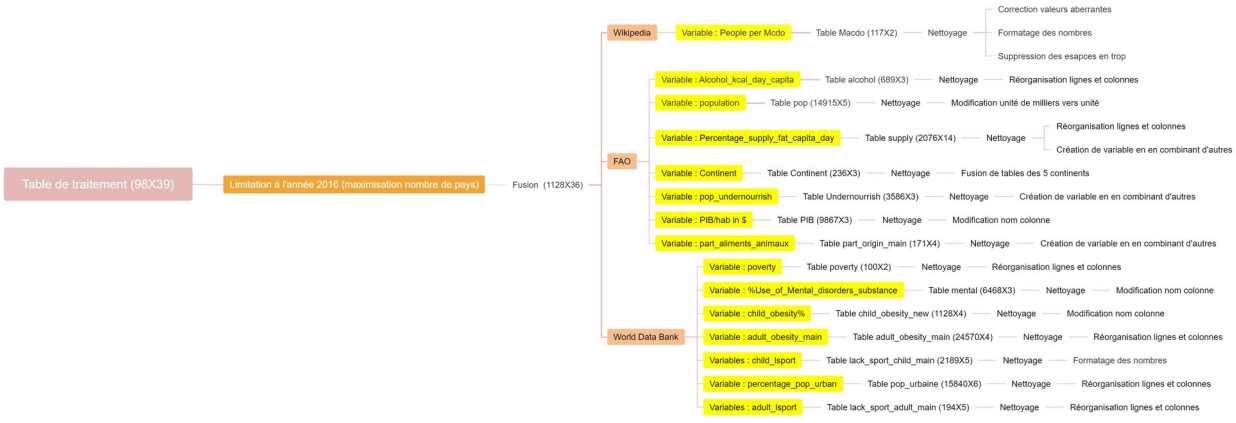


Figure 3: Carte mentale nettoyage données

La carte mentale ci-dessus représentent les processus de nettoyage des différentes tables pour obtenir la table finale. En raison des origines multiples des tables, les noms des pays ne correspondaient pas toujours. Pour pallier à ce problème un algorithme de fuzzy matching a été testé. Cependant les résultats n'étaient pas assez satisfaisants.
Par conséquent une table de correspondance a été créée manuellement sur Excel pour permettre la gestion des multiples dénominations de certains pays.

I. Analyse exploratoire

1. Les approches possibles :

Tout d'abord pour différencier les populations au niveau obésité nous pourrions être tenté d'analyser les pays par continent, pour cela nous allons afficher via des box plot la répartition de l'obésité par continent :

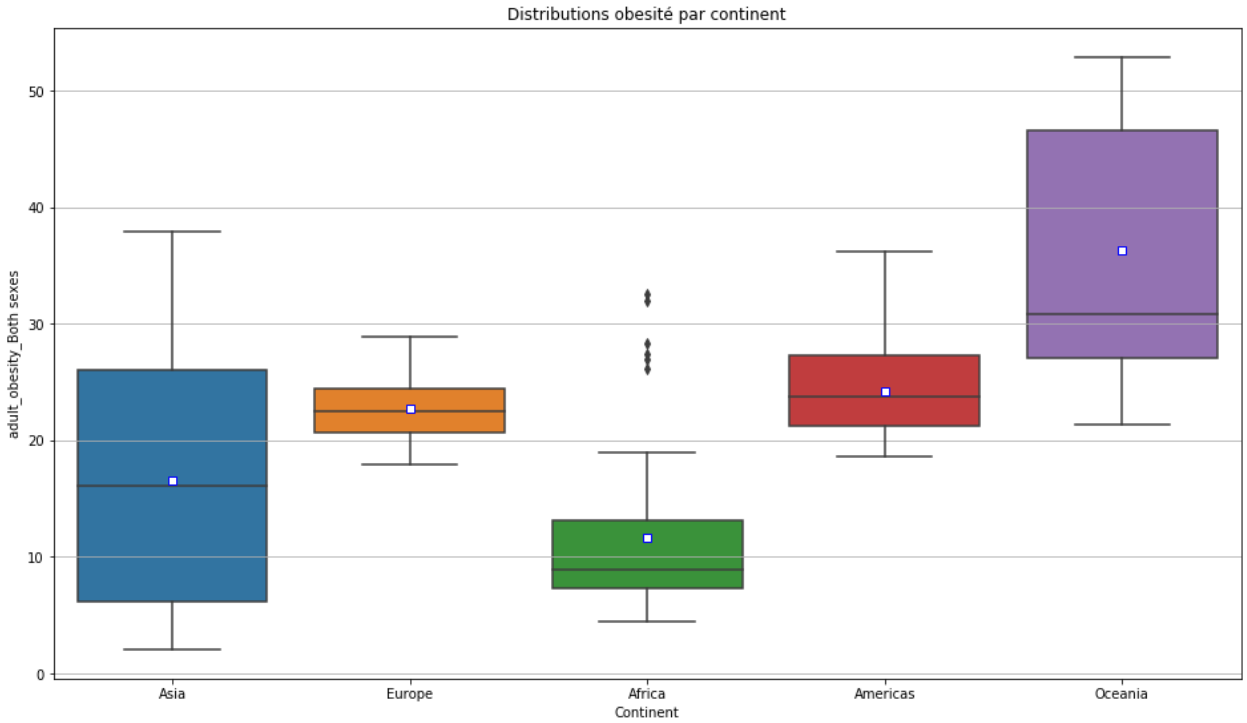


Figure 4: Boîte à moustache de la répartition de l'obésité par continent

En analysant le box plot, on remarque plusieurs points :

- On a deux continents l'Asie et l'Océanie avec des variances fortes par rapport à la moyenne, même si l'absence d'outliers traduit une bonne homogénéité de ces ensembles
- On a deux continents l'Amérique et l'Europe qui semblent au même niveau
- Visuellement deux groupes de continent se distinguent : Asie et Afrique avec une obésité plus faible par rapport à Europe, Amérique et Océanie
- Le continent Africain et le seul avec des outliers traduisant une inégalité forte entre les pays d'Afrique face à l'obésité

Nous pouvons déjà partiellement conclure que le classement des pays par continent pour expliquer l'obésité est imparfaite, nous n'arrivons pas par exemple à différencier l'Amérique de l'Europe. L'Afrique semble être divisée en deux groupes face à l'obésité. Concernant l'Asie et l'Océanie malgré des répartitions homogènes, la plage d'obésité que couvre ces deux continents est trop importante pour les classer efficacement.

Nous allons néanmoins analyser les outliers du continent africain pour avoir des premiers éléments pour savoir ce qui différencie les pays à forte obésité face à ceux avec une faible obésité, analyser dans un seul continent permet de diminuer les effets géographiques qui pourrait rendre difficile la différenciation.

Pour l’étude de ces outliers, nous allons utiliser le test de student pour définir les variables qui différencient vraiment les outliers du reste des pays africains :

```
child_obesity% vient d'une distribution normale avec pvalue = 0.12658319992898928
Les deux groupes ont une variance équivalente pour child_obesity% avec une pvalue = 0.0672540708444689
Test de student pvalue = 0.04212360481659284
percentage_pop_urban vient d'une distribution normale avec pvalue = 0.6846434452789939
Les deux groupes ont une variance équivalente pour percentage_pop_urban avec une pvalue = 0.8892113151184726
Test de student pvalue = 0.01887023106148944
```

Figure 6: Test de Student

On remarque deux variables avec une pvalue inférieure à 0.05 indiquant que la moyenne des deux groupes sur ces variables sont bien différentes. Par conséquent elles permettent de différencier ces deux groupes.

Outliers :

Country	child_obesity%	percentage_pop_urban	pop_undernourrish	adult_lsport_bothsex	child_lsport_bothsex	PIB/hab in $	Percentage_supply_fat_capita_day
Algeria	28.587	71.459	3.21	33.599998	83.8	3941.141870	0.215689
Egypt	23.543	42.732	4.76	31.000000	87.5	2824.310351	0.155952
Morocco	31.619	61.360	3.70	26.200001	87.3	2928.608681	0.182467
Tunisia	26.352	68.346	0.00	30.400000	81.5	3665.802374	0.251427

Figure 6: Liste des pays outliers du continent africain

Non outliers :

Country	child_obesity%	percentage_pop_urban	pop_undernourrish	adult_lsport_bothsex	child_lsport_bothsex	PIB/hab in $	Percentage_supply_fat_capita_day
Benin	20.538	46.229	7.36	15.900000	75.9	789.075269	0.177630
Botswana	27.537	67.933	18.52	21.700001	87.5	6953.144156	0.262215
Ghana	8.943	54.749	7.37	21.799999	87.5	1950.234376	0.125908
Kenya	13.460	26.105	23.44	15.400000	86.8	1427.704067	0.184755
Mauritania	10.968	51.962	12.01	41.299999	87.2	1526.646410	0.203507
Mauritius	30.348	40.910	7.92	29.799999	82.2	9691.909714	0.281906
Mozambique	19.212	34.926	33.06	5.600000	87.1	414.055163	0.156185
Namibia	11.525	47.961	12.72	33.400002	87.4	4551.399560	0.191009
Senegal	7.682	46.296	10.00	23.100000	88.5	1235.452194	0.253453
Tanzania	11.302	32.333	24.51	6.500000	82.1	918.375034	0.198045
Uganda	13.914	22.624	0.00	5.500000	85.7	612.791914	0.186881
Zambia	35.937	42.438	0.00	22.100000	89.3	1292.999808	0.204582
Zimbabwe	11.215	32.296	0.00	26.799999	86.6	1272.335448	0.236788

Figure 6: Liste des pays non outliers du continent africain

Les outliers ayant un taux d’obésité plus fort ont en moyenne une obésité infantile plus forte ainsi qu’un pourcentage de population urbaine plus important. Ceci nous donne deux premières pistes pour l’étude des facteurs d’obésité dans le monde.

2. Analyse des corrélations :

Afin de définir les facteurs influençant l'obésité mondiale, nous allons créer une matrice de corrélation. Plus le facteur de corrélation s'approche de 1 plus les deux variables sont corrélées, par contre plus il s'approche de -1 plus les variables sont anti-corrélées.
Dans notre cas nous allons rechercher les variables avec des corrélations les plus proches de 1 ou -1 par rapport aux variables des taux d'obésité.

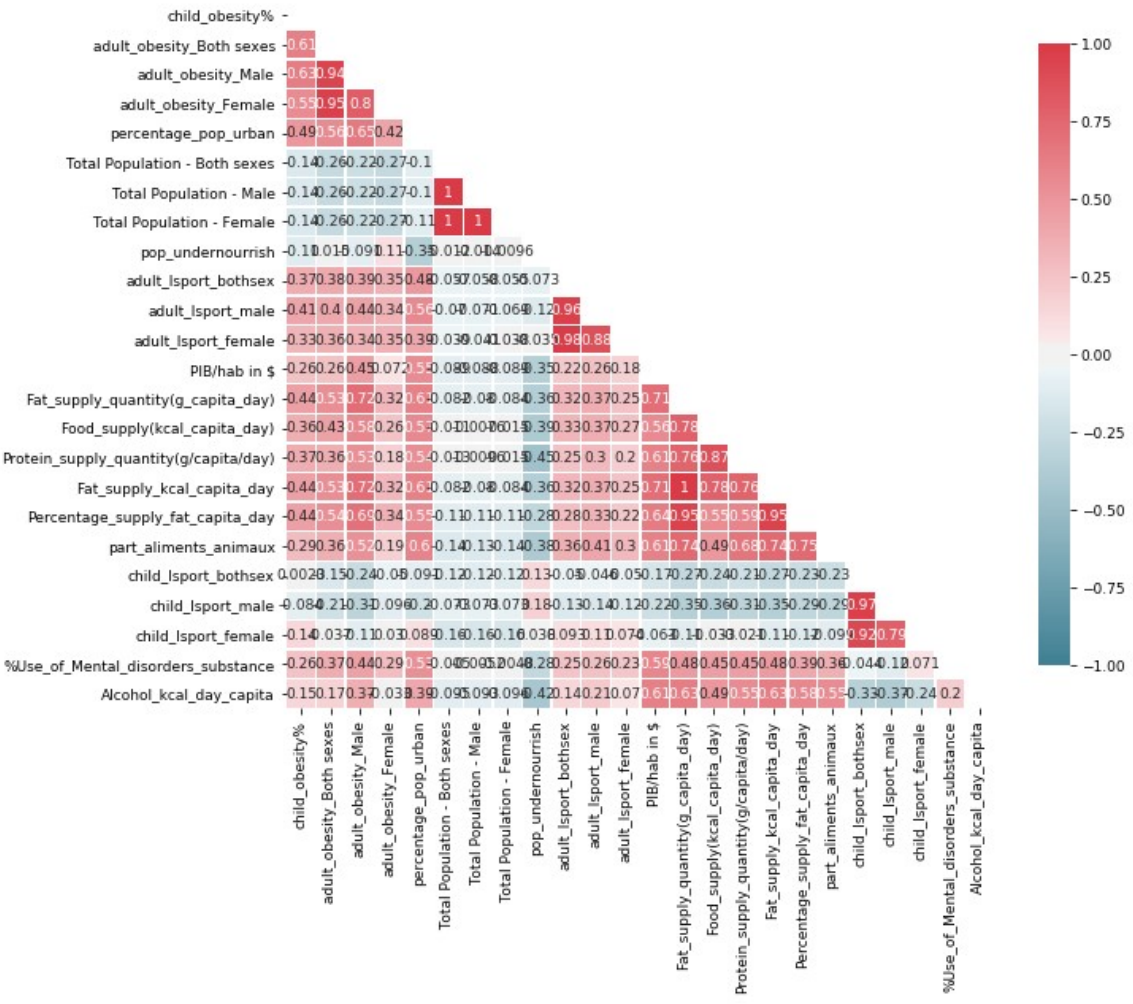


Figure 5: Matrice des corrélations

Ci-dessous les variables qui présentent une corrélation notable :

Child_obesity % : Une des premières variables que nous avons trouvé grâce aux outliers en Afrique. Nous avons ici une corrélation de 0.6

percentage_pop_urban : La seconde variables que l'on avait remarqué avec les outliers, concernant l'obésité des hommes on est à 0.65, pour les femmes la corrélation est plus faible à 0.42

Percentage_supply_fat_capita_day : Bonne corrélation (0.69) avec l'obésité masculine. Cependant cette variable est fortement corrélée à deux autres variables Fat_supply_quantity

et Fat_supply_kcal_capita_day (0.95). Nous ne garderons qu'une seule variable pour éviter un phénomène de colinéarité.
Part_aliments_animaux : Corrélée avec l'obésité masculine, mais très faiblement avec l'obésité féminine
adult_lsport_male : Le manque de sport chez l'homme est moyennement corrélée (0.44) avec l'obésité chez l'homme
%Use_of_Mental_disorders_substance : L'utilisation de psychotropes est moyennement corrélée (0.44) avec l'obésité chez l'homme

Une des premières observations importantes suite à cette matrice des corrélations est que beaucoup de variables présentent des corrélations assez élevées avec l'obésité chez l'homme. En revanche ces valeurs de corrélations ne se retrouvent pas du côté des femmes.
Afin de valider ou non notre capacité de prédiction de l'obésité chez l'homme ainsi que chez la femme nous allons tester des modèles de régression linéaire.

Avant cela, la réalisation d'une ACP a été réalisée. Cependant aucune information complémentaire a la matrice des corrélations n'a été trouvée. Il a tout de même été remarqué que la variable obésité de la femme avait une mauvaise qualité de représentation comparée à celle de l'homme. Ceci coïncide avec notre idée que nous serions peu à même de trouver les facteurs d'obésité de la femme avec nos variables actuelles.

3. Régression obésité homme :

Pour les hommes notre modèle de régression explique 82 % de la variance, ce qui est significatif.

OLS Regression Results

Dep. Variable:	adult_obesity_Male	R-squared:	0.822
Model:	OLS	Adj. R-squared:	0.790
Method:	Least Squares	F-statistic:	25.27
Date:	Thu, 15 Oct 2020	Prob (F-statistic):	1.48e-24
Time:	22:06:58	Log-Likelihood:	-276.31
No. Observations:	98	AIC:	584.6
Df Residuals:	82	BIC:	626.0
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	17.6194	0.448	39.323	0.000	16.728	18.511
child_obesity%	1.8890	0.583	3.242	0.002	0.730	3.048
percentage_pop_urban	3.4797	0.738	4.714	0.000	2.011	4.948
Total Population - Male	-0.9050	0.481	-1.880	0.064	-1.862	0.052
pop_undernourrish	0.6328	0.562	1.125	0.264	-0.486	1.751
adult_lsport_male	0.4104	0.585	0.702	0.485	-0.753	1.573
child_lsport_male	-1.8372	0.515	-3.565	0.001	-2.862	-0.812
PIB/hab in $	-1.2587	0.765	-1.644	0.104	-2.781	0.264
Percentage_supply_fat_capita_day	1.7398	0.783	2.221	0.029	0.182	3.298
%Use_of_Mental_disorders_substance	1.1734	0.648	1.811	0.074	-0.115	2.462
Alcohol_kcal_day_capita	-0.4345	0.830	-0.523	0.602	-2.086	1.217
People per Mcdo	0.3950	0.482	0.819	0.415	-0.565	1.355
Africa	-1.5987	0.481	-3.321	0.001	-2.556	-0.641
Americas	0.0586	0.392	0.150	0.882	-0.721	0.838
Asia	-1.4365	0.471	-3.049	0.003	-2.374	-0.499
Oceania	0.9315	0.584	1.596	0.114	-0.229	2.092
Europe	2.9650	0.522	5.681	0.000	1.927	4.003

Omnibus:	17.635	Durbin-Watson:	1.804
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24.133
Skew:	0.859	Prob(JB):	5.75e-06
Kurtosis:	4.720	Cond. No.	7.47e+15

Figure 6: Régression linéaire obésité masculine

Ci-dessous les variables les plus importantes dans l'ordre décroissant :

1. Percentage_pop_urban
2. Europe
3. Child_obesity %
4. child_lsport_male
5. Percentage_supply_fat_capita_day

Ces variables sont celles qui avaient une corrélation élevée dans la matrice des corrélations.

4. Régression obésité femme :

OLS Regression Results

Dep. Variable:	adult_obesity_Female	R-squared:	0.638
Model:	OLS	Adj. R-squared:	0.572
Method:	Least Squares	F-statistic:	9.652
Date:	Fri, 16 Oct 2020	Prob (F-statistic):	1.30e-12
Time:	16:39:19	Log-Likelihood:	-319.72
No. Observations:	98	AIC:	671.4
Df Residuals:	82	BIC:	712.8
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	23.2010	0.698	33.250	0.000	21.813	24.589
child_obesity%	2.5143	0.907	2.771	0.007	0.709	4.319
percentage_pop_urban	3.5712	1.150	3.107	0.003	1.284	5.858
Total Population - Male	-1.6732	0.749	-2.233	0.028	-3.164	-0.182
pop_undernourrish	0.5541	0.876	0.633	0.529	-1.188	2.296
adult_lsport_male	0.6176	0.910	0.678	0.499	-1.193	2.429
child_lsport_male	-1.8690	0.803	-2.329	0.022	-3.465	-0.273
PIB/hab in $	-2.8587	1.192	-2.398	0.019	-5.230	-0.487
Percentage_supply_fat_capita_day	1.6820	1.220	1.379	0.172	-0.744	4.108
%Use_of_Mental_disorders_substance	1.1146	1.009	1.105	0.272	-0.892	3.121
Alcohol_kcal_day_capita	-1.5435	1.293	-1.194	0.236	-4.116	1.029
People per Mcdo	0.9818	0.751	1.307	0.195	-0.513	2.476
Africa	0.3653	0.750	0.487	0.627	-1.126	1.857
Americas	1.0224	0.610	1.676	0.098	-0.191	2.236
Asia	-2.0209	0.734	-2.754	0.007	-3.481	-0.561
Oceania	-1.0197	0.909	-1.122	0.265	-2.827	0.788
Europe	3.1570	0.813	3.884	0.000	1.540	4.774

Omnibus:	14.163	Durbin-Watson:	2.012
Prob(Omnibus):	0.001	Jarque-Bera (JB):	15.629
Skew:	0.854	Prob(JB):	0.000404
Kurtosis:	3.952	Cond. No.	7.47e+15

Figure 7: Régression linéaire obésité féminine

Pour les femmes notre modèle de régression explique 64 % de la variance, ce qui est peu significatif.

Ci-dessous les variables les plus importantes dans l'ordre décroissant :

1. Percentage_pop_urban
2. Europe
3. PIB/hab in $
4. child_obesity %
5. Asie

On retrouve une partie des variables qui expliquent l'obésité des hommes. Cependant le pourcentage de variance étant plus faible, cela prouve que qu'il nous manque une information pour mieux prédire l'obésité des femmes.

Suite à ces résultats, nous concentrerons notre analyse sur l'obésité chez l'homme.

II. Analyses de l'obésité chez les hommes

1. Le random forest :

Nous allons reprendre notre régression pour l'analyse de l'obésité masculine. Actuellement 82 % de la variance est trouvée par le modèle, ce qui est un bon score. L'objectif de cette partie est de chercher à améliorer ce score avec potentiellement un modèle plus élaboré.

Une des pistes que nous allons étudier repose sur la méthode dite du « bagging » et des méthodes parallèles.
Un des exemples de ces algorithmes, dits ensemblistes, est le random forest (forêt aléatoire) qui est l'assemblage de plusieurs arbres de décision.

Dans notre cas le point de départ est une racine mère qui contient l'ensemble de nos individus. Les itération suivantes consistent à scinder en deux groupes sur lesquels on va chercher à maximiser l'homogénéité de chaque groupe en minimisant la variance à l'intérieur du groupe.

Cependant, ce type de modèle a une grande tendance au sur-apprentissage. Pour pallier à ce problème on associe plusieurs arbres pour former une forêt, dans laquelle chaque arbre est constitué d'un sous-ensemble aléatoire des individus et des variables. L'objectif étant de réduire les effets de colinéarité.

Test d'un premier random forest :

Entrée [125]:
```
# Instantiate model with 100 decision trees
rf = RandomForestRegressor(n_estimators = 100, random_state = 20)
# Train the model on training data
resulto = rf.fit(features, y)
```

Entrée [126]:
```
# Use the forest's predict method on the test data
predictions = resulto.predict(features)
# Calculate the absolute errors
errors = abs(predictions - y)
# Print out the mean absolute error (mae)
print('Mean Absolute Error:', round(np.mean(errors), 2), 'degrees.')
```

```
Mean Absolute Error: 1.62 degrees.
```

Entrée [127]:
```
# Calculate mean absolute percentage error (MAPE)
mape = 100 * (errors / y)
# Calculate and display accuracy
accuracy = 100 - np.mean(mape)
print('Accuracy:', round(accuracy, 2), '%.')
```

```
Accuracy: 83.31 %.
```

Figure 8: Code Random Forest

On a un MAE de 1,62 % et une précision de 83 %, ce qui correspond à une amélioration de 1 % par rapport au modèle de régression simple ce qui est beaucoup sur des niveaux de précision élevés.

Il faut cependant vérifier qu'il n'y a pas de sur-apprentissage du modèle. Cela se traduirait par le fait que notre modèle serait juste adapté à nos données d'entraînement et donc peu capable de prédire sur de nouvelles données.

Pour ça nous allons diviser notre jeu de données en deux parties, une partie test et l'autre d'entraînement :

Régression linéaire simple

```
R² du model d'entraînement :  0.8232543842576523
R² du model de test :  0.7523582003388379
```

Figure 9: Précision régression linéaire simple

Malgré une perte de précision sur nos données de test, cette baisse de précision reste acceptable (en partie due à la faible volumétrie de pays), ce qui indique que notre modèle n'est pas en sur-apprentissage

2. Optimisation avec GridSearch

Random Forest

```
R² du model d'entraînement :  78.45 %.
R² du model de test :  67.52 %
```

Figure 10: Précision Random Forest

Une nouvelle fois nous remarquons une perte de précision entre le jeu d'entraînement et le jeu de test. On remarquera la précision plus faible que la régression linéaire simple.

Le random forest est un modèle complexe qui demande beaucoup d'ajustement au niveau des hyperparamètres afin d'être performant. Nous allons utiliser la fonction GridSearch qui va tout simplement permettre de tester plusieurs combinaisons d'hyperparamètres.

```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done  25 tasks      | elapsed:   11.4s
[Parallel(n_jobs=-1)]: Done 146 tasks      | elapsed:   44.1s
[Parallel(n_jobs=-1)]: Done 349 tasks      | elapsed:  1.7min
[Parallel(n_jobs=-1)]: Done 500 out of 500 | elapsed:  2.4min finished
RandomizedSearchCV(cv=5, estimator=RandomForestRegressor(), n_iter=100,
                   n_jobs=-1,
                   param_distributions={'bootstrap': [True, False],
                                        'max_depth': [10, 20, 30, 40, 50, 60,
                                                      70, 80, 90, 100, 110,
                                                      None],
                                        'max_features': ['auto', 'sqrt'],
                                        'min_samples_leaf': [1, 2, 4],
                                        'min_samples_split': [2, 5, 10],
                                        'n_estimators': [10, 231, 452, 673, 894,
                                                         1115, 1336, 1557, 1778,
                                                         2000]},
                   random_state=42, verbose=2)
```

Figure 12: Itération GridSearch

Voici le résultat des hyper-paramètres optimisés :

```
]: rf_random.best_params_

{'n_estimators': 673,
 'min_samples_split': 5,
 'min_samples_leaf': 1,
 'max_features': 'sqrt',
 'max_depth': 90,
 'bootstrap': False}
```

Figure 11: Paramètres optimisés de Random Forest

Nous avons entraîné un nouveau modèle avec ces hyper-paramètres et nous obtenons une précision de 75,31 % sur le jeu de test. La précision est donc grandement améliorée.

Regardons désormais le poids des variables :

```
Variable: Percentage_supply_fat_capita_day Importance: 0.22
Variable: percentage_pop_urban Importance: 0.16
Variable: child_obesity%       Importance: 0.13
Variable: PIB/hab in $         Importance: 0.09
Variable: Total Population - Male Importance: 0.07
Variable: %Use_of_Mental_disorders_substance Importance: 0.07
Variable: pop_undernourrish    Importance: 0.05
Variable: child_lsport_male    Importance: 0.05
Variable: Asia                 Importance: 0.05
Variable: adult_lsport_male    Importance: 0.03
Variable: Alcohol_kcal_day_capita Importance: 0.03
Variable: Europe               Importance: 0.03
Variable: Africa               Importance: 0.02
Variable: Americas             Importance: 0.01
Variable: Oceania              Importance: 0.01
```

Figure 12: Poids des variables du Random Forest

Ci-dessous les variables les plus importantes dans l'ordre décroissant pour le modèle de régression linéaire :

1. Percentage_pop_urban
2. Europe
3. Child_obesity %
4. child_lsport_male
5. Percentage_supply_fat_capita_day

Pour le random_forest :

1. Percentage_supply_fat
2. Percentage_pop_urban
3. Child_obesity %
4. PIB/hab in $
5. Total population - male

Concernant le random forest, on remarque la disparition de la notion de continent, ainsi que le manque de sport chez les enfant de sexe masculin. A noter l'apparition de la notion de PIB/hab in $, ainsi que l'importance de la taille de la population masculine dans le pays.

Cependant nous pouvons envisager un nouveau modèle : le XGBoost. Cet algorithme s'inspire du random forest mais intègre la méthode de descente du gradient afin d'améliorer l'erreur de l'itération précédente. En revanche, il faut faire attention à la profondeur de correction afin d'éviter un sur-apprentissage.

3. Tentative d'amélioration plus poussée avec le XGBOOST

XGBOOST :

```
5]: param = {
        'eta': 0.2,
        'max_depth': 3,
        'objective': 'reg:squarederror'}

    steps = 30  # The number of training iterations
    df = X.drop("const", axis=1)
    df = df.rename(columns = lambda x: x.replace(' ', '_'))
    X_train, X_test, y_train, y_test = train_test_split(df, y, test_size=0.33, random_state = 42)
    D_train = xgb.DMatrix(data=X_train, label=y_train)
    D_test = xgb.DMatrix(data=X_test, label=y_test)
    model = xgb.train(param, D_train, steps)
```

Figure 14: Code XGBOOST

Précision sur le jeu d'entraînement :

```
Accuracy = 91.08%.
```

Figure 14: Précision entraînement XGBOOST

Précision sur le jeu de test :

```
Accuracy = 76.34%.
```

Figure 14: Précision test XGBOOST

On remarque une précision sur le jeu d'entraînement très élevée (91,08%). C'est nettement plus qu'avec le modèle précédent (+13 points). Cependant, la précision sur le jeu de test reste similaire au modèle précédent (76,34%), cela laisse penser à un léger sur-apprentissage.

Ci-dessous les variables les plus importantes pour ce modèle :

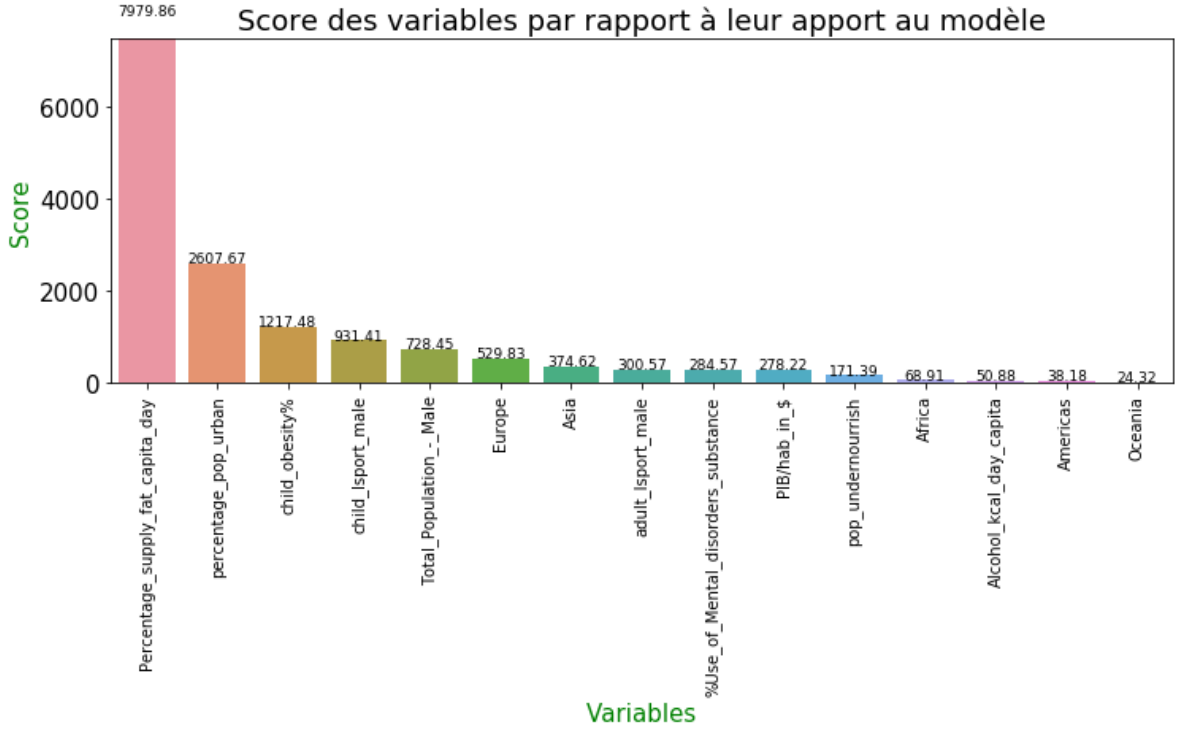


Figure 14: Apport des variables au XGBOOST

Dans l'ordre décroissant :

1. Percentage_supply_fat
2. Percentage_pop_urban
3. Child_obesity %
4. Child_lsport_male
5. Total population – male

Par rapport au random forest on perd le PIB/hab, mais on récupère le manque de sport chez les enfants garçons.

On remarque à travers ces 3 modèles, 3 variables qui restent importantes peu importe l'algorithme :

- Percentage_supply_fat
- Percentage_pop_urban
- Child_obesity%

Dans la dernière partie nous allons étudier ces trois variables, leur influence et leur lien par rapport aux politiques de lutte contre l'obésité.

III. L’influence de nos variables

1. Classification :

a. Détermination du nombre de clusters

Via les 3 variables trouvées précédemment, nous allons tenter de recréer une carte de l’obésité dans le monde. Effectivement lors de la première partie nous avons pu observer que l’obésité n’était pas homogène entre les pays au sein d’un même continent. Nous allons donc aborder un problème de clusterisation non-supervisée. Pour ce faire nous allons utiliser l’algorithme du Kmeans.

Afin de déterminer le nombre optimal de cluster nous étudions via l’indicateur de silhouette, ainsi que la méthode du coude :

Coude

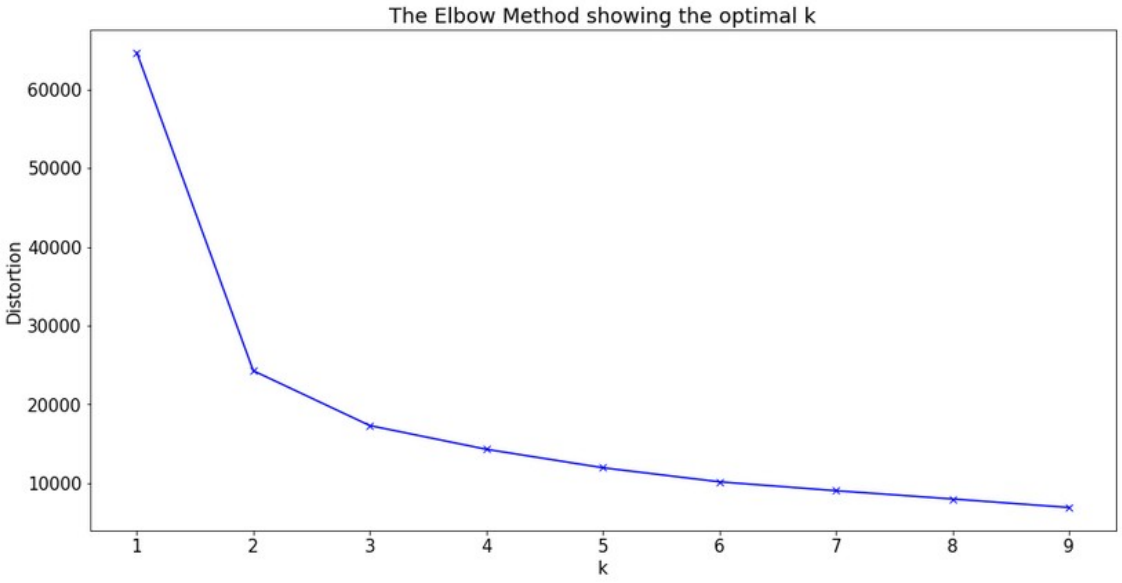


Figure 13: Méthode du Coude

Silhouette

```
Automatically created module for IPython interactive environment
For n_clusters = 2 The average silhouette_score is : 0.5315020114819231
For n_clusters = 3 The average silhouette_score is : 0.3869828339956205
```

Figure 14: Scores silhouette

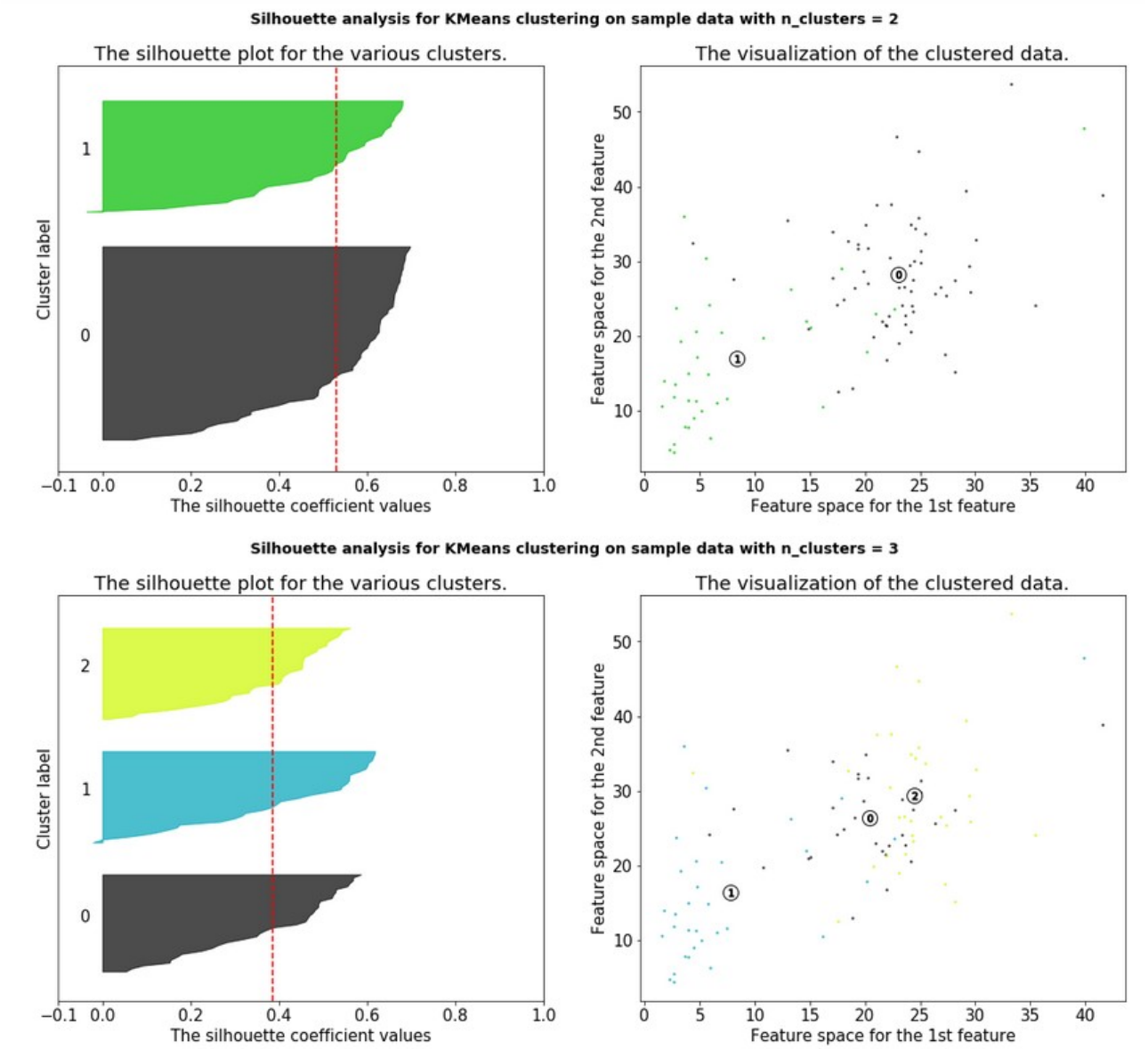


Figure 15: Représentation graphique Silhouette

Le résultat de ces deux indicateurs nous indique d'utiliser 2 clusters, cependant, d'un point de vue opérationnel, avoir seulement 2 clusters est trop faible nous décidons donc de faire un Kmeans avec 3 clusters.

b. Analyse des clusters

Après avoir appliquer l'algorithme pour obtenir 3 clusters, nous pouvons analyser chaque cluster sur chacune des variables :

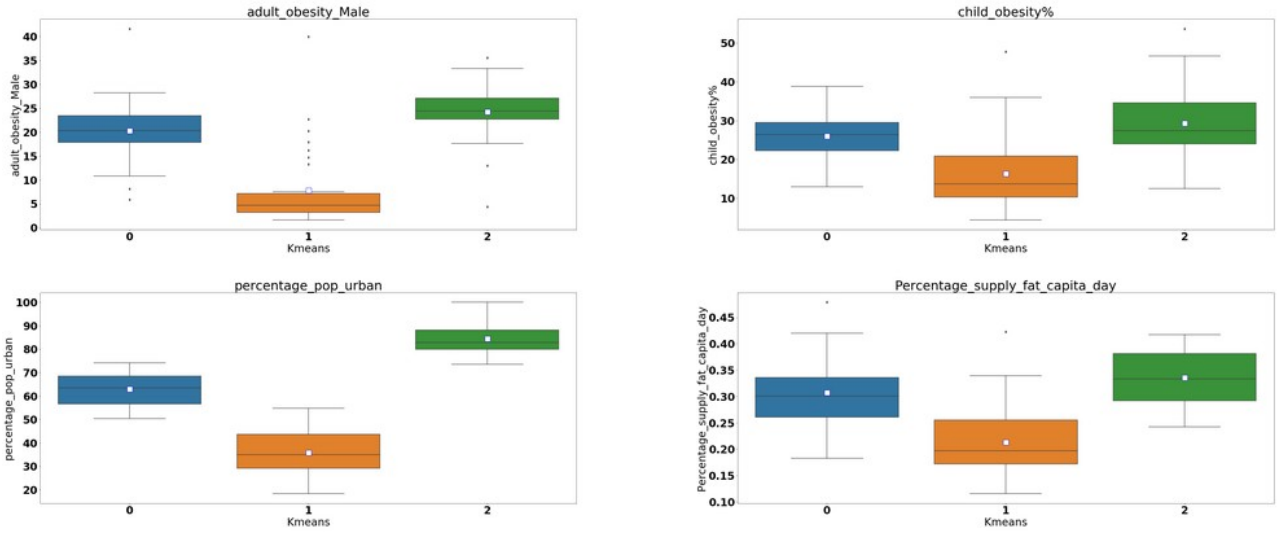


Figure 16: Représentation graphiques des clusters pour chaque variables

Les distributions et les moyennes des variables sur chacun des groupes sont bien différentes. Par ailleurs les groupes semblent homogènes. On notera tout de même que la variable obésité dans le cluster 1 a plusieurs outliers. Le pourcentage d'urbanisation est la variable qui différencie le mieux ces 3 clusters avec des moyennes bien différenciées et des groupes homogènes et compactes.

Voici la nouvelle carte dessiner par notre méthode de classification :

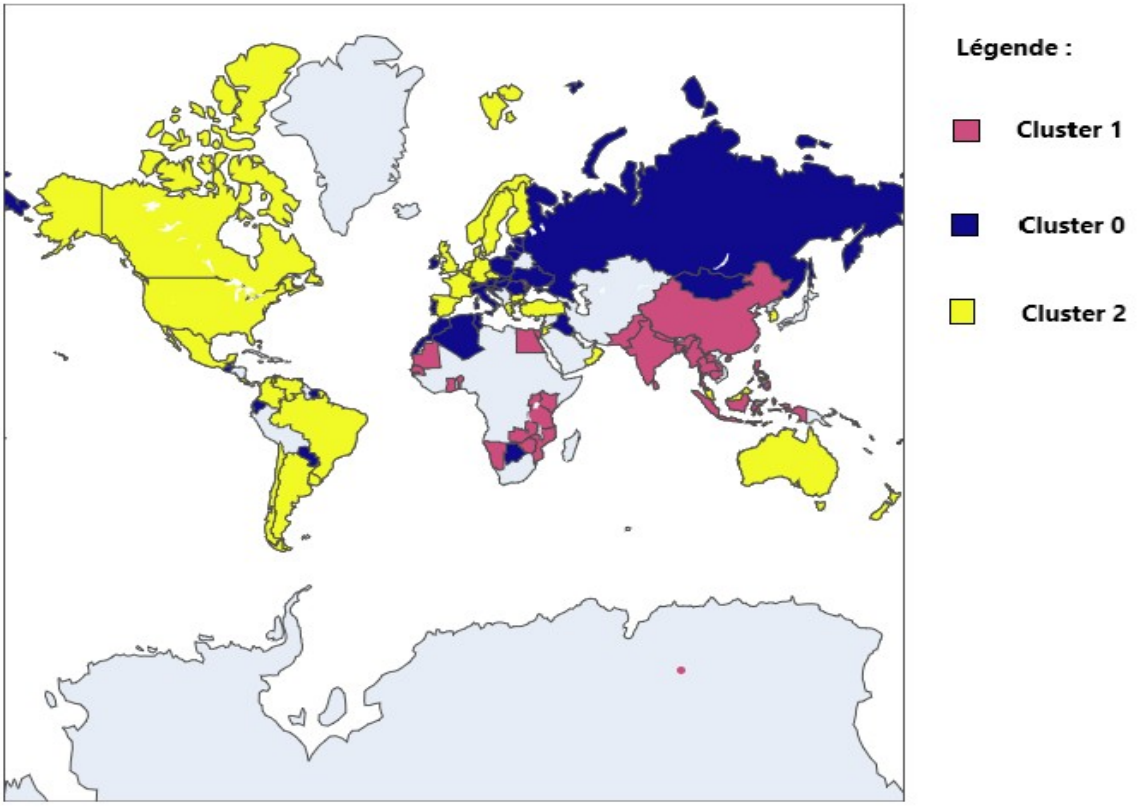


Figure 17: Représentation des clusters sur carte du monde

Le cluster 1 représentant les pays ayant l'obésité les plus faibles rassemble une grande partie de l'Asie et de l'Afrique. Cette observation correspond à notre analyse de départ où ces deux continents concentraient les pays avec les plus faibles taux d'obésité. Le cluster 2, qui comprend l'obésité la plus forte, concentre la majorité de l'Amérique, la moitié de l'Europe et une bonne partie de l'Océanie.

Cette nouvelle carte caractérise mieux les pays par rapport à l'obésité masculine. Nous allons désormais analyser les 3 autres variables qui constituent le Kmeans.

2. Le taux de graisse :

a. Lien fast food, aliments gras

Cette variable quantifie la part de graisse dans l'alimentation des habitants du pays. Il peut donc être intéressant de faire une analogie entre cette variable et l'implentation des fast food dans le monde. Nous avons trouvé sur internet un jeu de données indiquant le nombre d'habitant par Mac Donald implantés dans le pays :

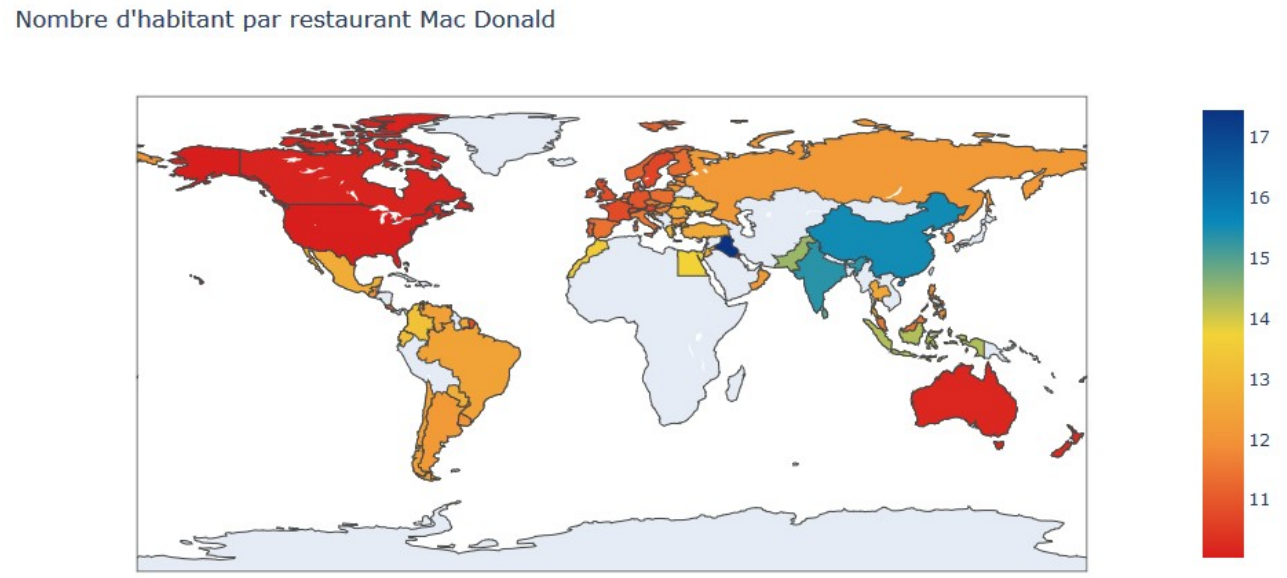


Figure 18: Densité habitants par restaurants Mac Donald

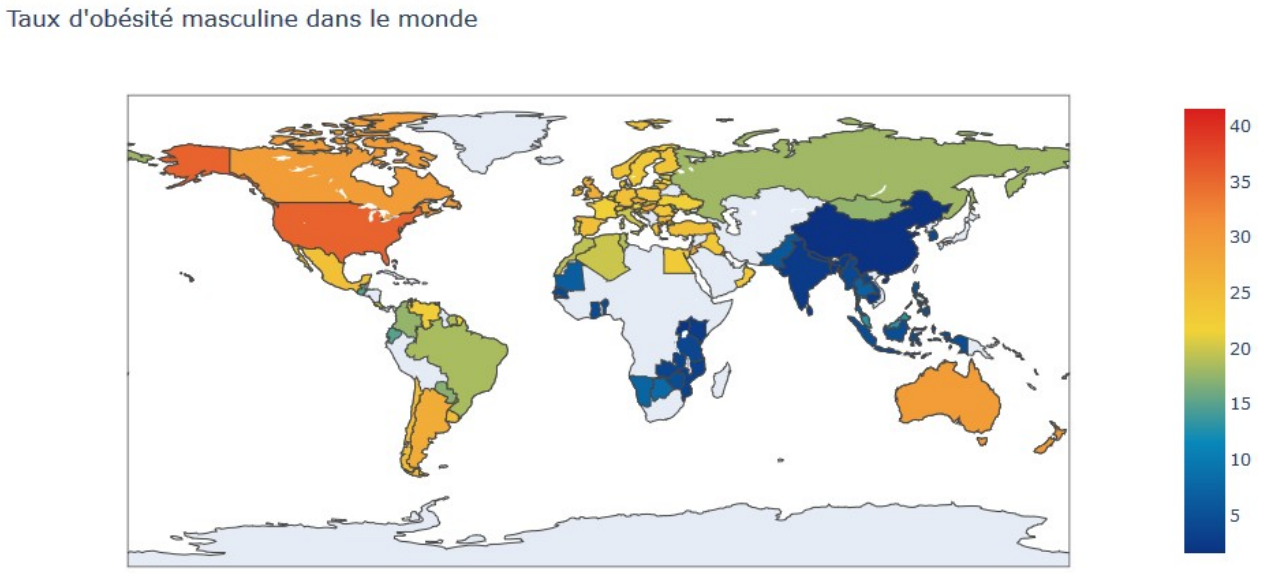


Figure 19: Taux d'obésité adulte masculine représenté sur une carte du monde

Logiquement, nous pouvons observer une correspondance entre les densités de Mac Donald et les taux d'obésité masculine : plus la densité de Mac Donald est importante plus le taux d'obésité est important (l'inverse est également vrai).

b. Les aliments gras une cible de l'obésité des instances étatiques :

Prenons l'exemple du CDC (Centers for Disease Control and Prevention) aux Etats-Unis. 4 causes de l'obésité ont été identifiées chez les américains : le comportement, l'environnement social, la génétique et facteurs combinés (drogues et maladies chroniques)

Figure 21: Extrait site du CDC américain concernant les causes de l'obésité adulte

D'après le CDC, l'alimentation grasse fait partie du critère « comportement ».

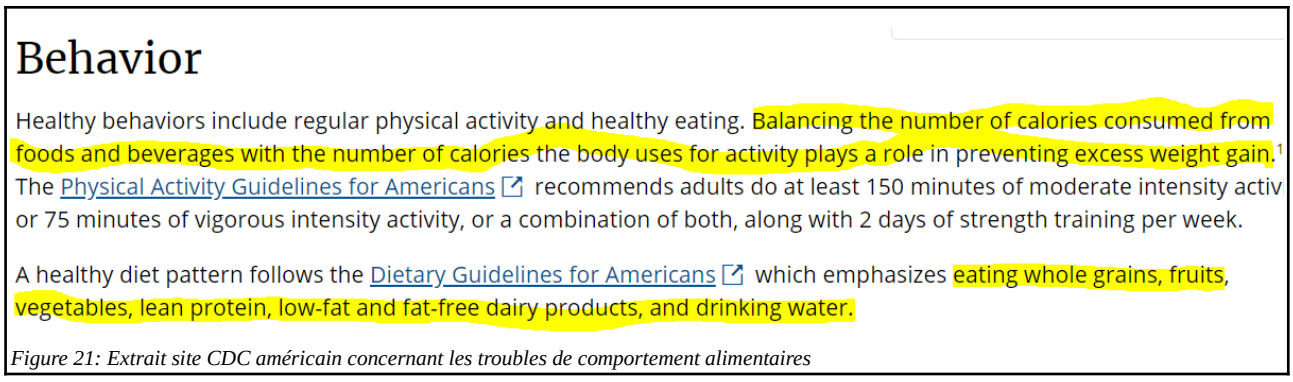


Figure 21: Extrait site CDC américain concernant les troubles de comportement alimentaires

Comme précisé dans l'article ci-dessus, l'important pour éviter une prise de poids est d'équilibrer les calories consommées par rapport à nos activités quotidienne. Cependant le besoin calorique dépend d'une personne à une autre, notamment en fonction de la constitution et de l'intensité de ses activités.
Par conséquent, dans le cadre de cette analyse, il était préférable de prendre la variable du taux de graisse dans l'alimentation car cette variable est rationalisée à l'individu.
Par ailleurs, la recommandation du CDC coïncide avec les résultats de cette étude : manger des aliments avec un faible taux de graisse est lié à un taux d'obésité faible.

Par conséquent, nous avons bien identifié un facteur de l'obésité.

3. Taux d’urbanisation :

a. La croissance de l’urbanisation :

Le taux d’urbanisation représente le pourcentage de la population concentrée dans des agglomérations urbaines en opposition à la population rurale.

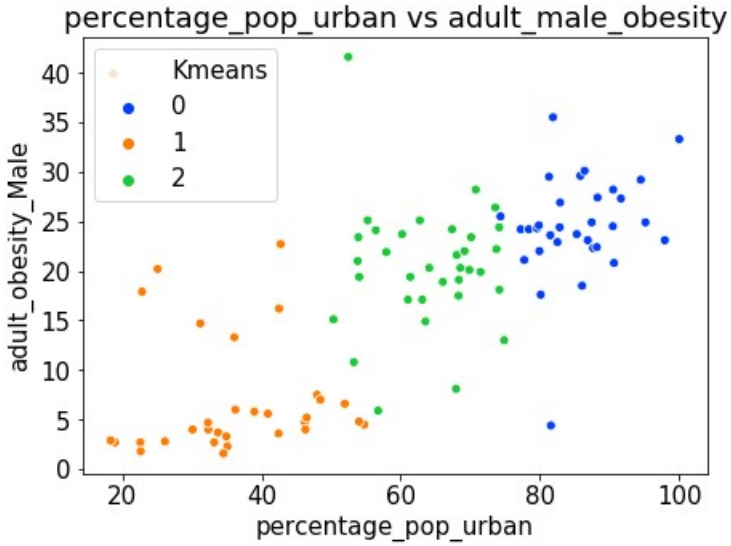


Figure 20: Graphique opposant l'obésité adulte masculine au taux d'urbanisation

D’après le graphique ci-dessus, plus le taux d’urbanisation est élevé et plus le taux d’obésité masculine est élevé. Cela se confirme avec les clusters du Kmeans : le cluster 0 représente les pays avec un fort taux d’obésité et un fort taux d’urbanisation à l’inverse du cluster 1.

b. L'urbanisation, facteur de progrès mais aussi d'obésité :

Un rapport de l'OMS de 2017 fait le lien entre urbanisation et obésité. Ce rapport met notamment en évidence le rôle crucial des villes pour mettre en place des aménagements pour faciliter l'accès aux activités physiques à la population urbaine, car on estime que le taux d'urbanisation mondiale en 2030 atteindra 80 %.

Figure 21: Extrait page de garde rapport OMS de 2017 concernant l'activité physique en ville

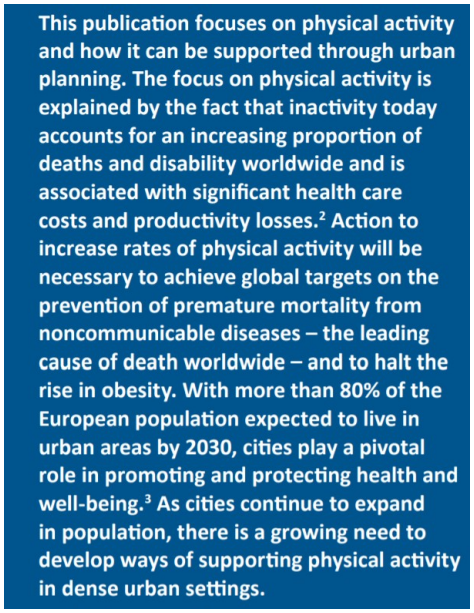

This publication focuses on physical activity and how it can be supported through urban planning. The focus on physical activity is explained by the fact that inactivity today accounts for an increasing proportion of deaths and disability worldwide and is associated with significant health care costs and productivity losses.[2] Action to increase rates of physical activity will be necessary to achieve global targets on the prevention of premature mortality from noncommunicable diseases – the leading cause of death worldwide – and to halt the rise in obesity. With more than 80% of the European population expected to live in urban areas by 2030, cities play a pivotal role in promoting and protecting health and well-being.[3] As cities continue to expand in population, there is a growing need to develop ways of supporting physical activity in dense urban settings.

Figure 22: Extrait introduction rapport OMS 2017 - le sport en ville

Le but de ce rapport est justement de montrer avec l'exemple de Copenhague comment la mise en place dans les villes de politiques qui promeuvent l'activité physique permet sur le moyen terme de réduire l'obésité en zone urbaine.

Concentrer la population dans des zones où la pratique du sport est moins possible à cause des différentes cohabitations urbaines, doit obliger les villes à refaire une place aux zones dédiées à l'activité physique.

Nous avons à nouveau bien identifié un facteur de l'obésité.

4. L’obésité infantile :

a. Impact de l’obésité infantile sur l’obésité adulte masculine

Ici l’obésité infantile serait un facteur explicatif de l’obésité masculine adulte :

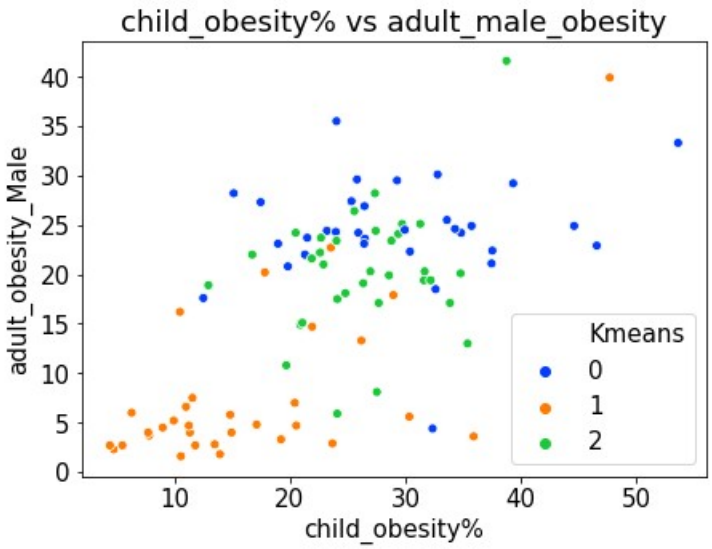
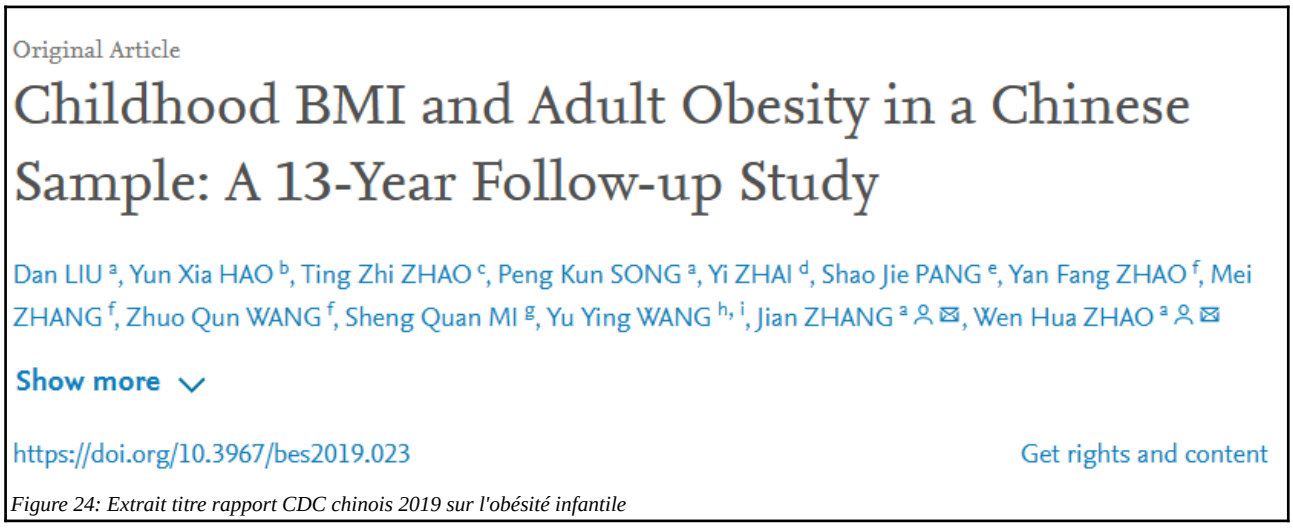


Figure 24: Représentation graphique taux obésité adulte masculine opposée à l'obésité infantile

On a effectivement une corrélation entre le taux d’obésité des adultes et le taux d’obésité chez les enfants. D’après ce graphique on peut identifier pourquoi les méthodes silhouette et coude suggéraient un Kmeans à deux clusters : cluster 1 versus le cluster 0 et 2.

b. Un effort de prévention à faire du côté des enfants

Le CDC chinois a notamment fait des études en 2019 pour étudier des enfants obèses et non obèses afin de vérifier si durant leur vie adulte ils étaient prédisposés ou non à devenir obèse.


Original Article

Childhood BMI and Adult Obesity in a Chinese Sample: A 13-Year Follow-up Study

Dan LIU [a], Yun Xia HAO [b], Ting Zhi ZHAO [c], Peng Kun SONG [a], Yi ZHAI [d], Shao Jie PANG [e], Yan Fang ZHAO [f], Mei ZHANG [f], Zhuo Qun WANG [f], Sheng Quan MI [g], Yu Ying WANG [h, i], Jian ZHANG [a], Wen Hua ZHAO [a]

Show more

https://doi.org/10.3967/bes2019.023 Get rights and content

Figure 24: Extrait titre rapport CDC chinois 2019 sur l'obésité infantile

Results

The percentage of non-obese children who grew up to be non-obese adults was 62.6%, and that of obese children who grew up to be obese adults was 80.0%. There was a significant association between childhood body mass index (BMI) and adulthood BMI with a β regression coefficient of 3.76 [95% confidence interval (*CI*): 1.36-6.16], and between childhood obesity and adulthood obesity with an odds ratio of 5.76 (95% *CI*: 1.37-24.34). There was no statistical difference between parental obesity at baseline and children's adulthood obesity, after adjustment of confounders. Male participants and those aged 10.0-13.0 years had a higher risk of adulthood obesity with odds ratios of 2.50 (95% *CI*: 1.12-5.26) and 3.62 (95% *CI*: 1.17-11.24), respectively.

Conclusion

Childhood obesity is an important predictor of adulthood obesity.

Figure 24: Extrait résultats rapport CDC chinois 2019 obésité infantile

Cette article démontre que 62,6 % des enfants non-obèses sont devenus des adultes non-obèses et que 80 % des enfants obèses sont restés obèses une fois adulte.
Comme écrit dans l'article et comme prouvé lors de notre analyse, l'obésité infantile est un prédicteur important de l'obésité masculine adulte.

Conclusion

L’obésité est un phénomène qui n’a pas les mêmes origines d’un sexe à l’autre. De plus, l’obésité masculine présente des facteurs facilement identifiables, tandis que chez la femme les causes semblent multiples et plus profondes. Lors de cette analyse nous nous sommes donc concentrés sur l’explication de l’obésité masculine.

L’urbanisation croissante des pays est un enjeu important pour la santé des populations, notamment par rapport à l’aménagement urbain et la place que cet aménagement laisse pour les activités physiques.
L’accès à une nourriture saine, notamment moins riches en graisse, est aussi un facteur déterminant pour lutter contre l’obésité masculine adulte. Par ailleurs, la prévention est nécessaire dès l’enfance car une obésité infantile créée de fortes prédispositions à une obésité adulte.

Tous les facteurs identifiés dans les modèles de régression font échos aux politiques d’aujourd’hui qui sont tournées vers la prévention de l’obésité chez l’enfant, l’équilibre alimentaire et la promotion du sport.
Cependant les modèles n’expliquent pas toute la variance. Il y a à fortiori des facteurs manquants dans cette analyse comme par exemple le poids de la génétique ou encore les maladies chroniques.

Index des figures