|  |
| --- |
| **RAPPORT**  **TP FOUILLE DE DONNÉES** |

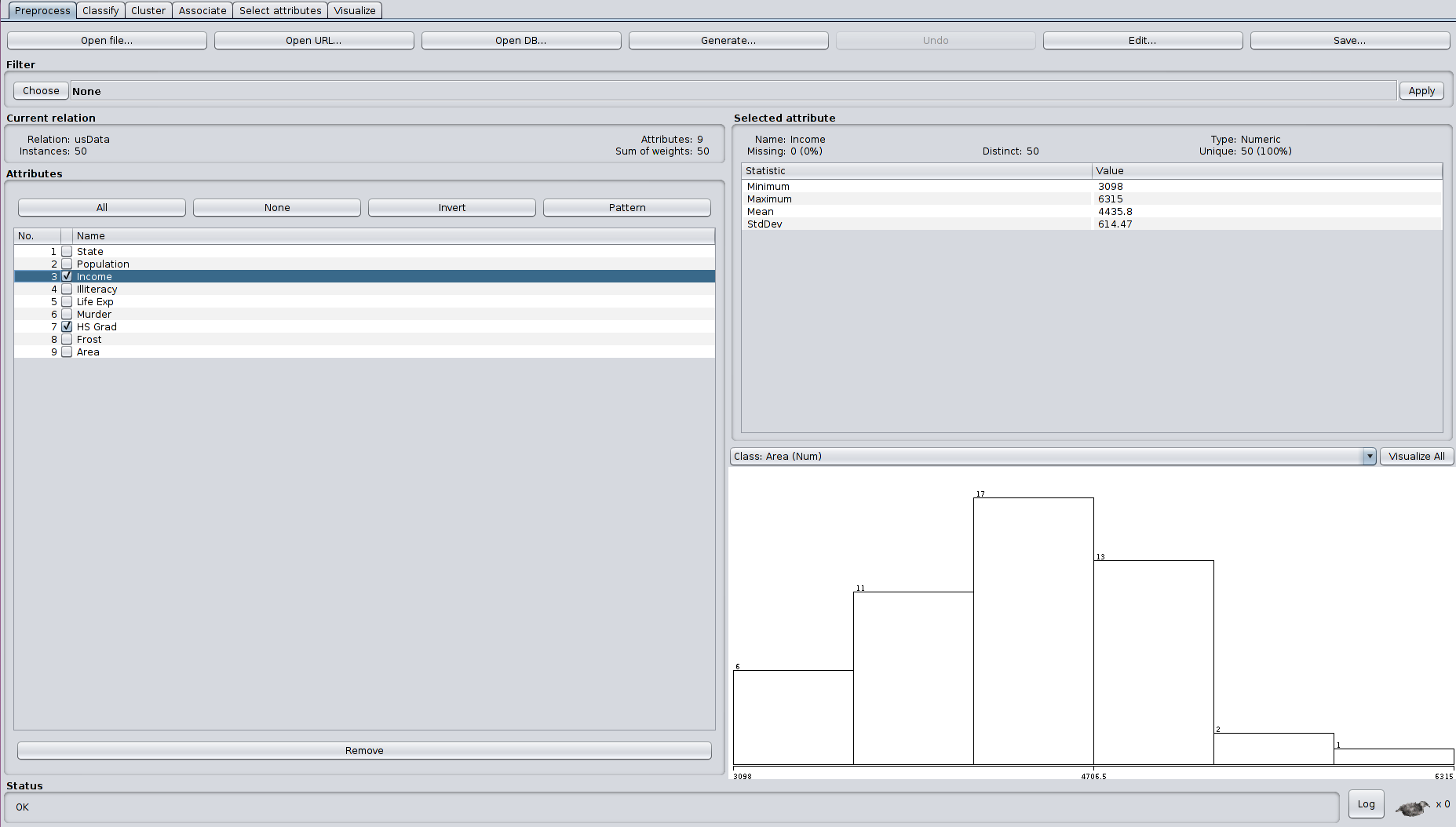
|  |
| --- |
| **Youssouf IBRAHIM ELMI**  **Mohamed Amine TOUCHENE**  **Mohamadou LO** |
|  |



# EXERCICE 1

**Question 1**

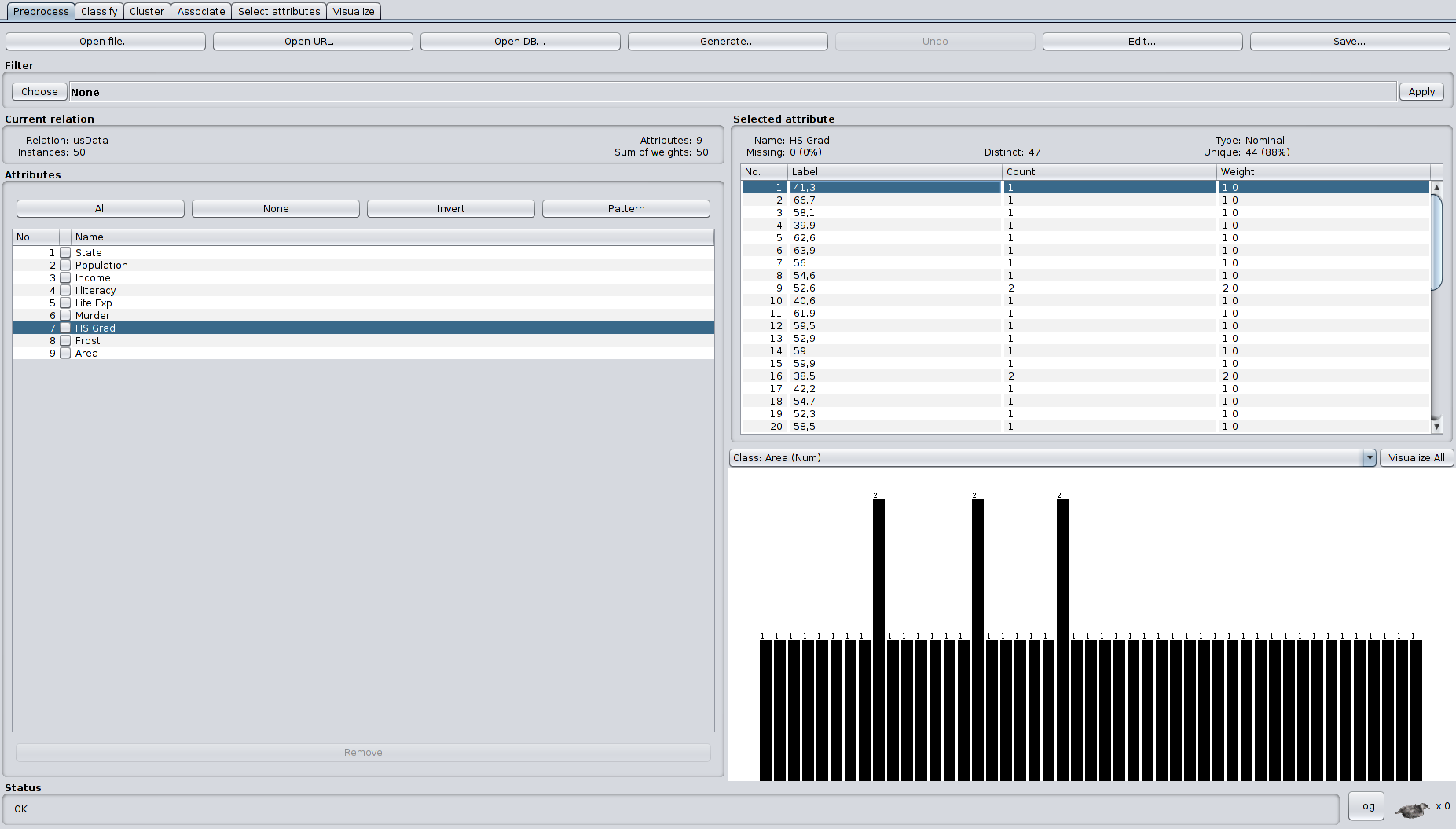
Dans un premier temps nous devrons sélectionner les deux attributs sur les quels on va faire l'analyse avec les différents algorithmes demander. Voici le statistique des deux attributs qui sont : **Income** et **HS Grad**.



La population " **Income** " minimum de 3098 et un maximum de 6315.

* **6** Etats sur 50 sont dans l'intervalle de [3098 ;3635].
* **11** Etats sur 50 sont dans l'intervalle de [3635 ;4171].
* **17** Etats sur 50 sont dans l'intervalle de [4171 ;4707].
* **13** Etats sur 50 sont dans l'intervalle de [4707 ;5243].
* **2** Etats sur 50 sont dans l'intervalle de [524 ;5779].
* **1** Etat sur 50 est a une seuil de PIB supérieur à 5779.

La moyenne de PIB par habitants est de **4435.8**.



Le pourcentage de " HS Grad " est d'un minimum de 37.8% et d'un maximum de 67.3% .

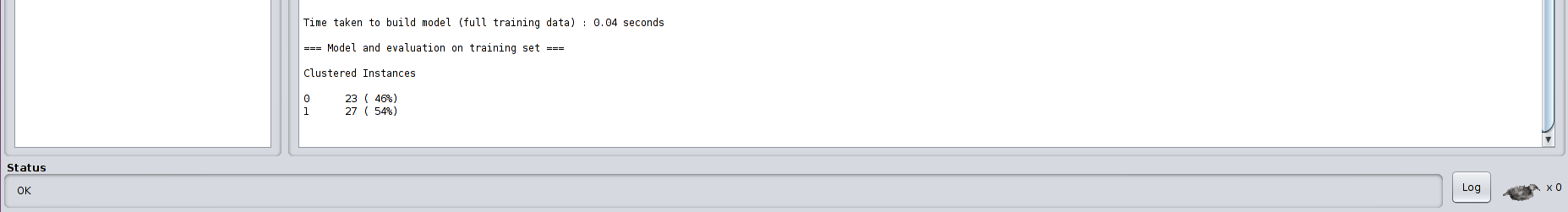
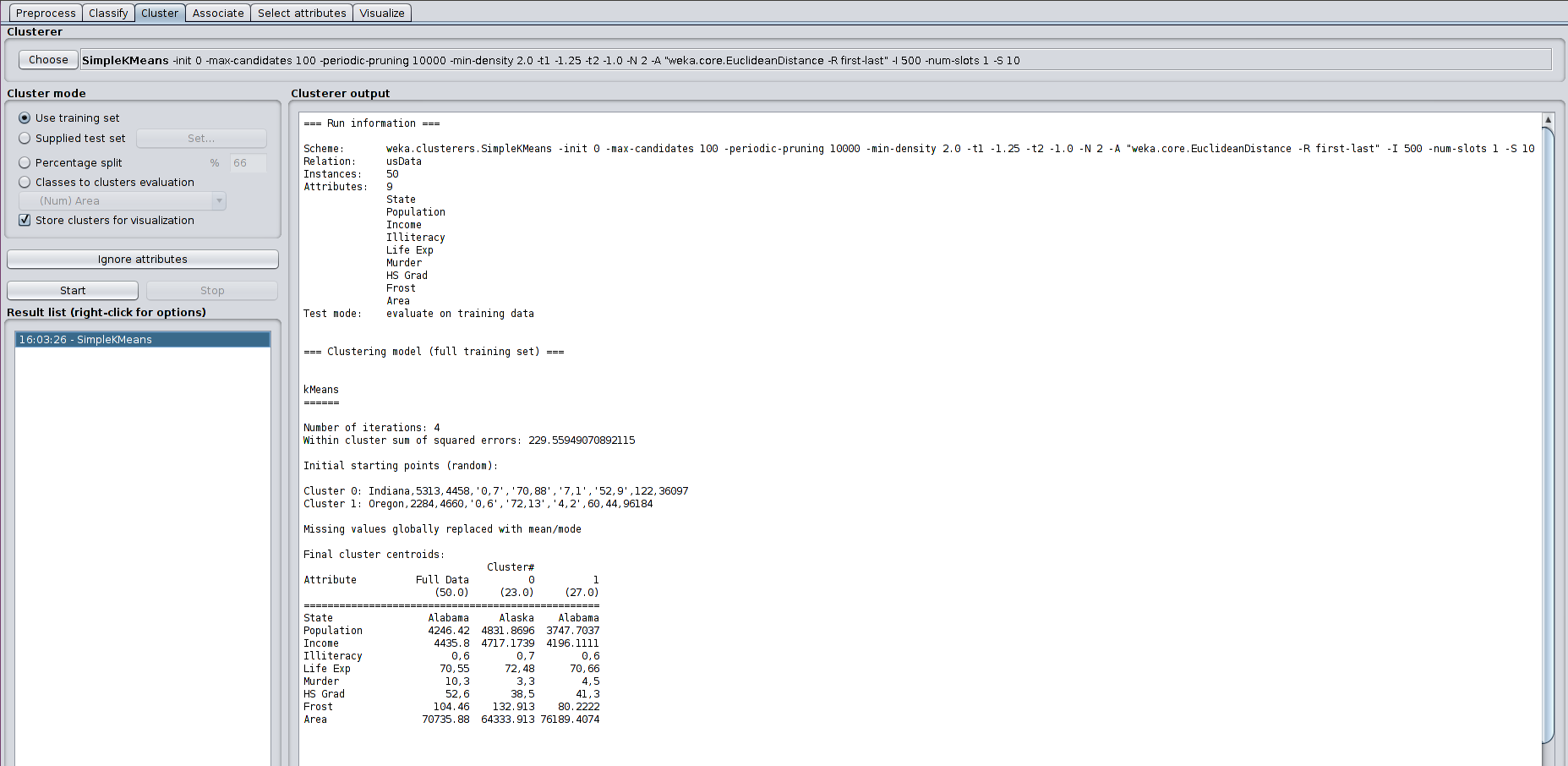
* On a 44 Etats respectives sur 47.
* On a 3 Etats respectives (52,6 ; 38,5 ; 52,7 ; 57,6 ) sur 47.

**Les résultats des algorithmes de clustering SimpleKMeans, EM et Canopy.**

# SimpleKMeans

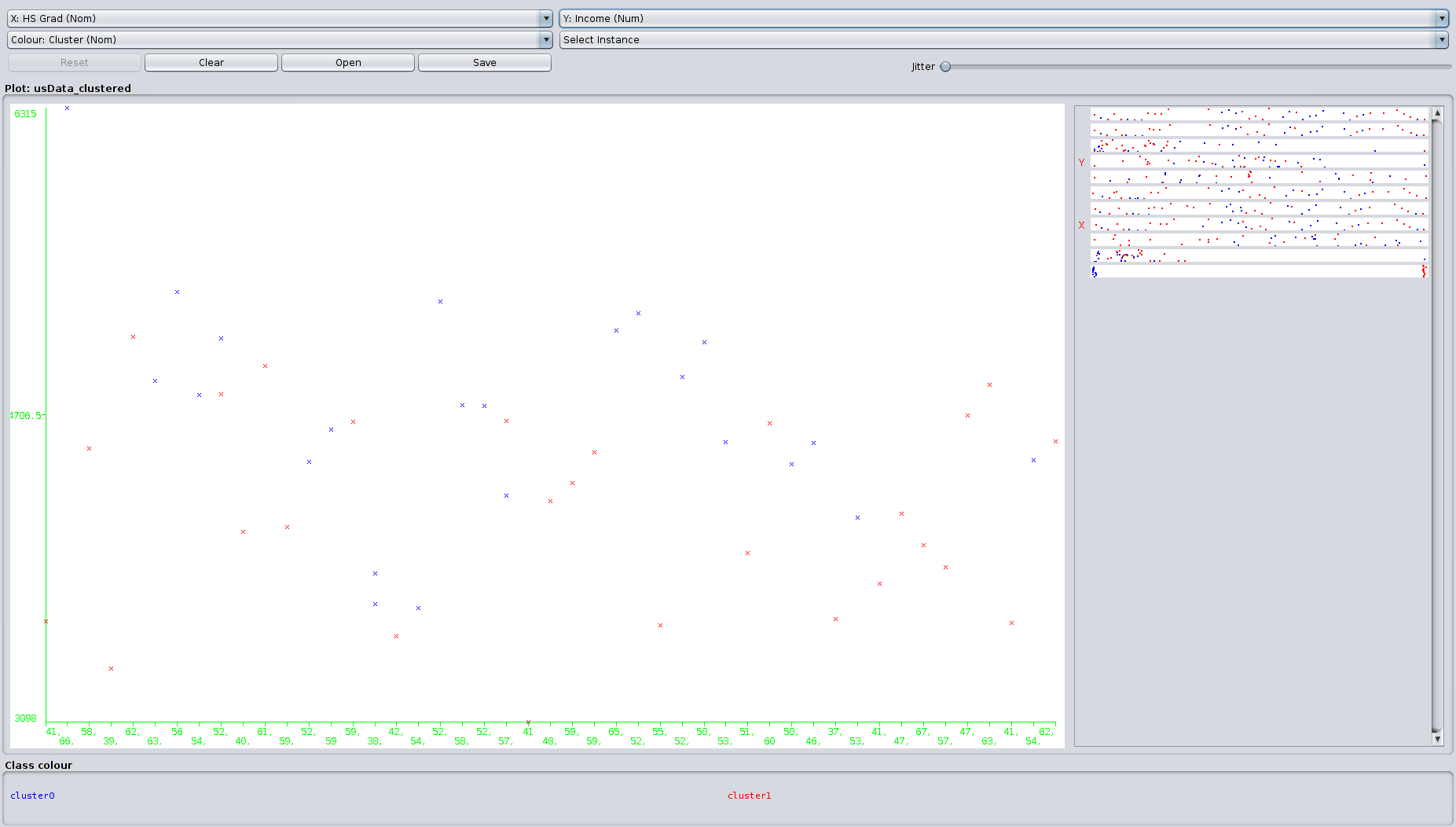
On a lancer l'algorithme de cluster SimpleKMeans, le résultat est la suivante :

Pour le premier cluster, **23** états sur 50 sont classer, soit **46%** des états. Pour le second cluster, **27** états sur 50, soit **54%** des états. Avec un taux d'erreur de 61% on a la répartition suivante :



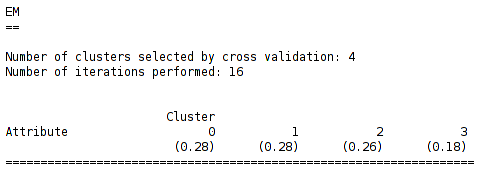
Quand on visualise l'affectation du cluster dans un graphe et qu'on choisie les deux axes ( X;Y)

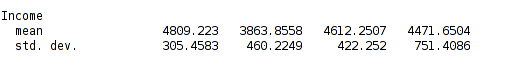
respectivement le "HS Grad" et "Income", on remarque que le premier cluster ( c’est-à-dire 10/23 états sont à un PIB par habitants faible ( inférieure à la moyenne). Ce dernier analyse montre que majoritairement, plus le PIB est petit et plus le taux des diplômé du supérieur est faible.

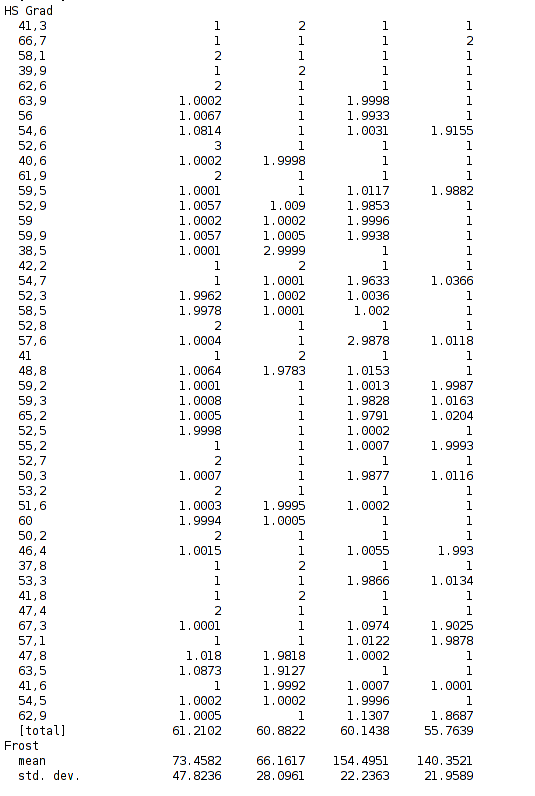


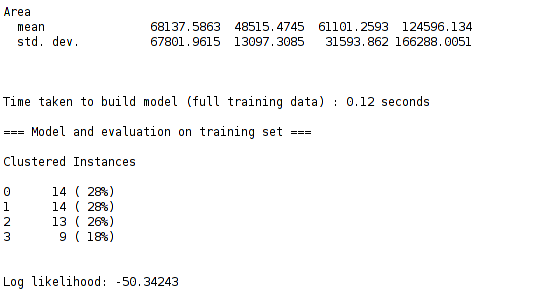
# EM

Dans cet algorithme le résultat est légèrement inverser à savoir que dans un premier cluster on a 14 états sur 50 sont classer dans un premier cluster ( soit 28% des états). Le deuxième cluster, on a 14 états sur 50 sont classer en cluster ( soit 28% des états). Le troisième cluster, on a 13 états sur 50 sont classer dans en ( soit 26% des états). Le quatrième, on a 09 états sur 50 sont classer en cluster ( soit 18% des états).



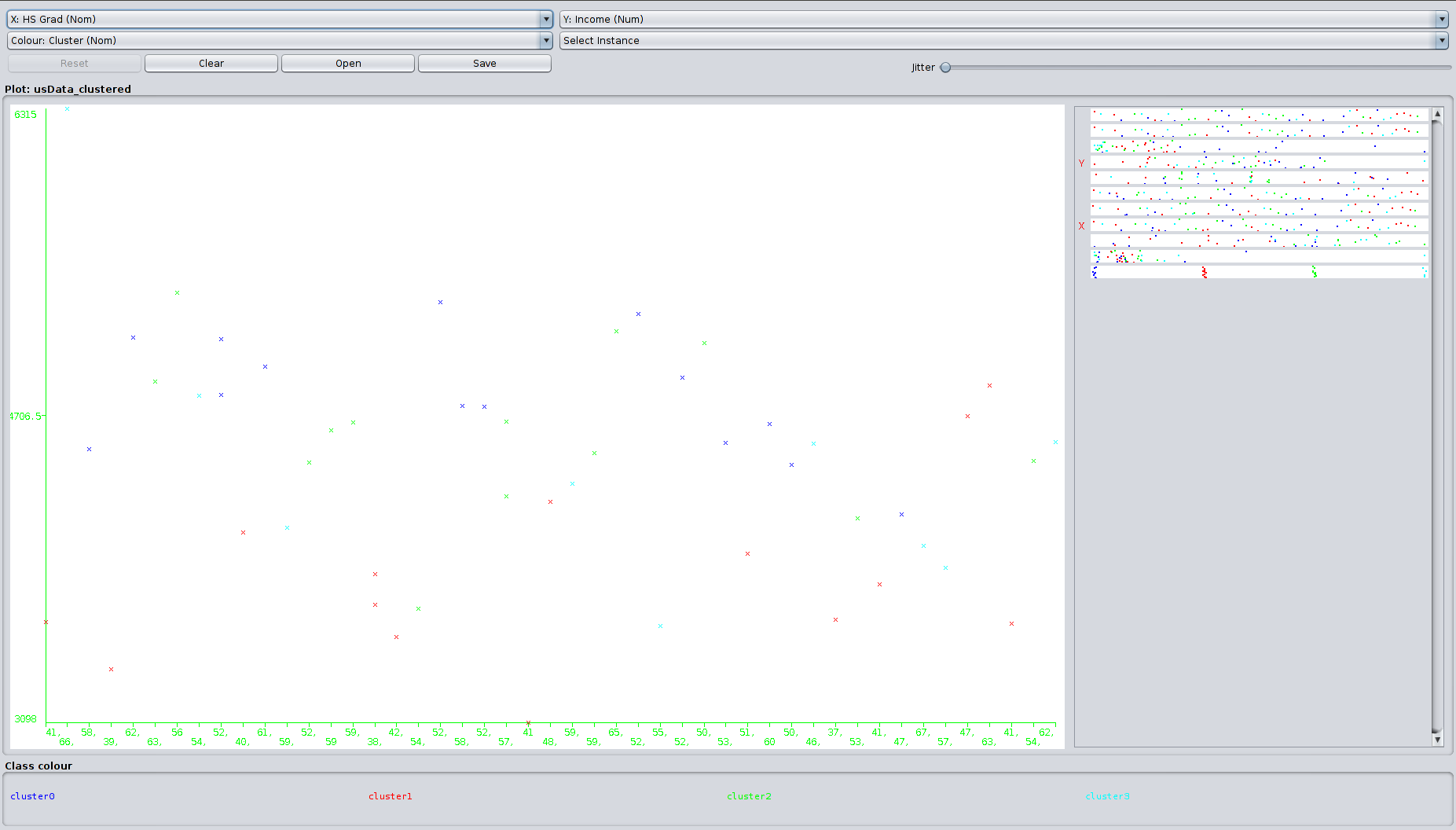






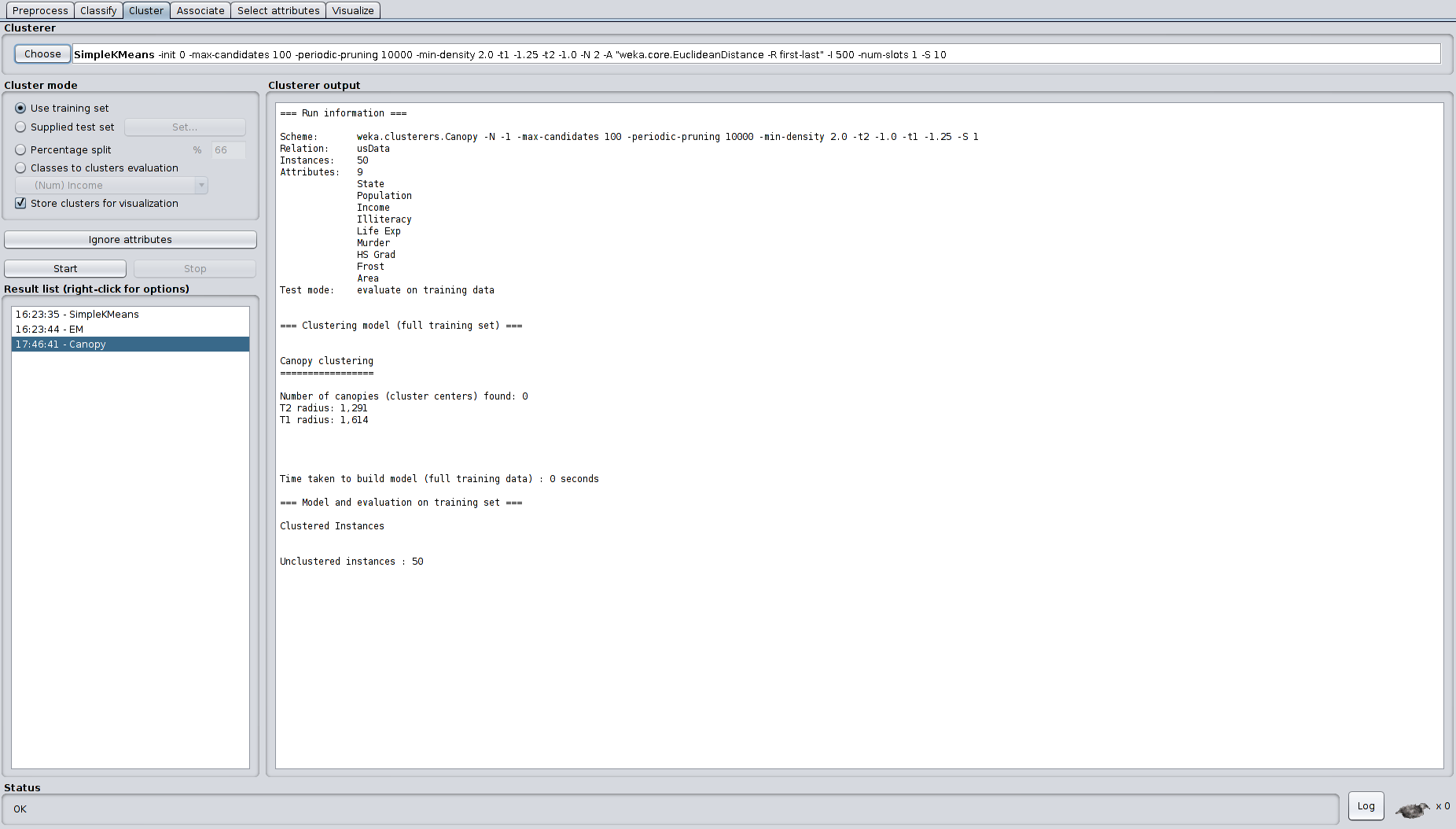
La précédente répartition montre qu'avec l'algorithme EM le dernier cluster regroupe une bonne majorité des états qui ont une " Income " faible avec un " HS Grad " inférieur à la moyenne. Le cluster 0,1 et 2 (

c’est-à-dire le trois premiers cluster) regroupe des états repartie sur un PIB plus ou moins dans la moyenne et un taux des diplômé du supérieur largement bien et même supérieurs à la moyenne. La visualisation de ce dernier résultat dans un graphe avec comme abscisse les " HS Grad " et comme ordonnées les " Income ".



# CANOPY

Pour l’algorithme canopy, on obtient aucun cluster, c’est à dire qu’on a eu 0 états de clusters sur 50.

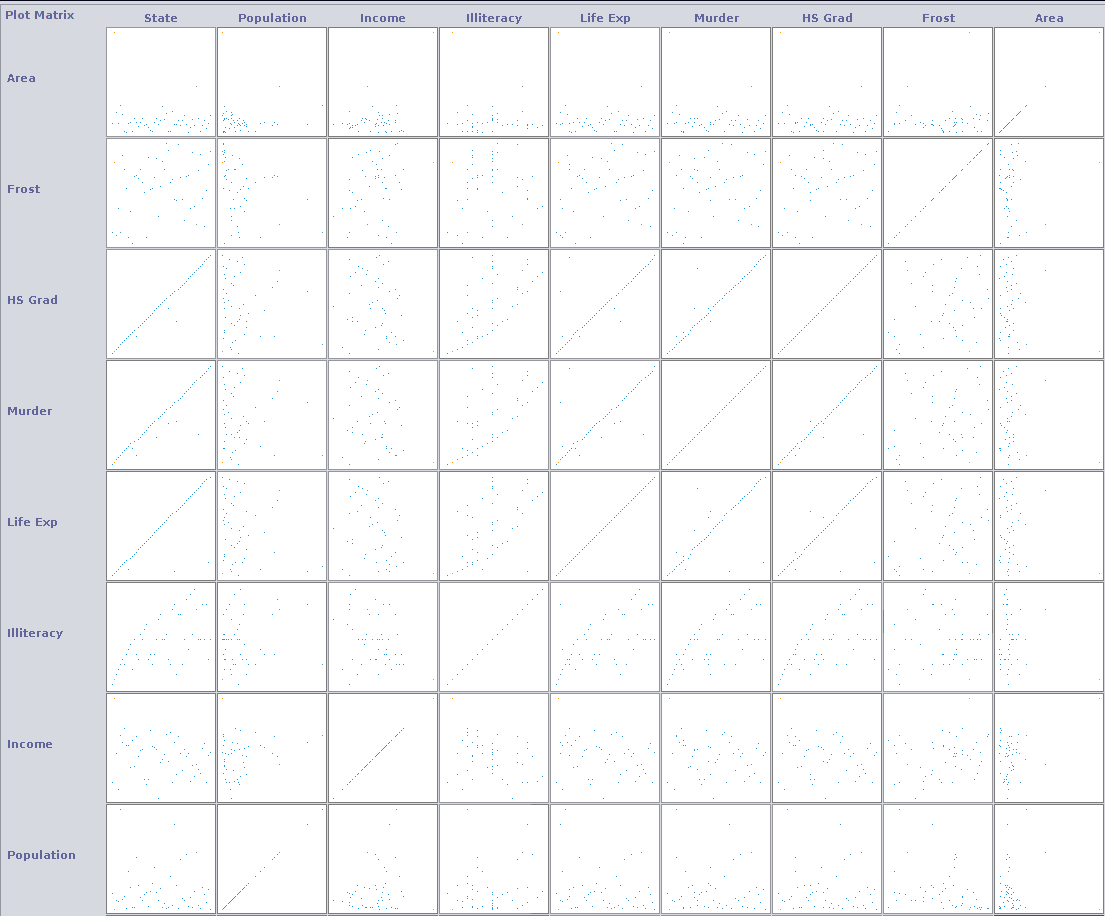


Comme on a eu aucun cluster, la répartition avec l’algorithme Canopy affiche aucun graphe puisqu’on a eu « 50 états unclustered »

# Question 2 :

Dans un premier temps, on réalise une visualisation des ensembles des attributs chacune en fonction des autres.

Nous aurons une visualisation qui est :



Prenons l’exemple le « Murder » en fonction « Life Exp ». Nous remarquons que l’espérance de vie est plus ou moins bonne mais avec le taux de meurtre, nous observons que l’espérance de vie diminue en se concentrant sur un taux inférieure à la moyenne d’après le graphe ci-dessous.

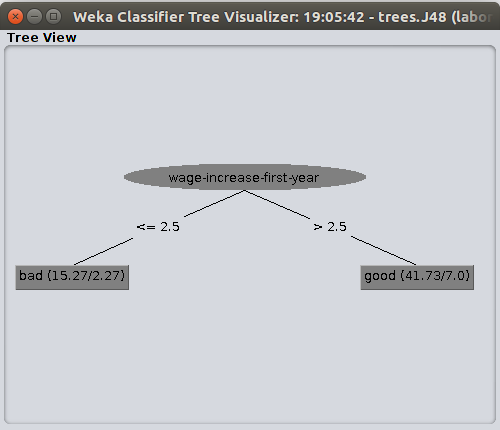
# 

# EXERCICE 2 :

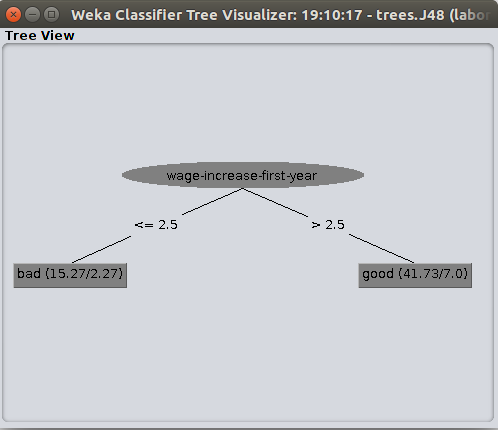
Afin de faire la classification et pouvoir tracer l'arbre de décision, on va devoir attribuer plusieurs valeurs à l'indice de confiance **(confidenceFactor)**.

Initialement on lui attribue la valeur : **0.10**.

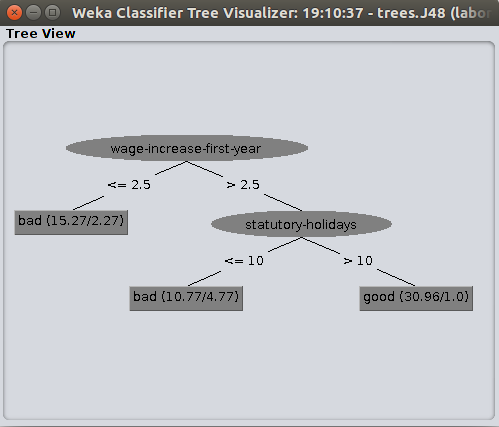
l'arbre obtenu est le suivant :



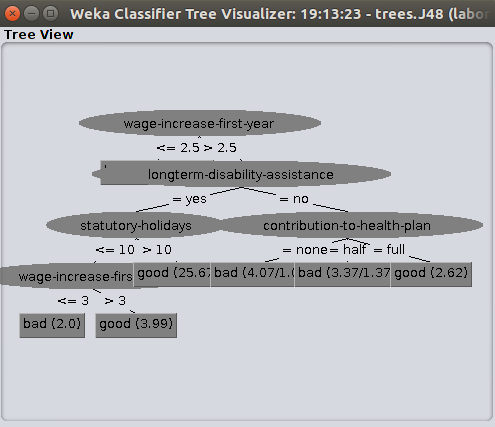
Il est de meme pour un indice de confiance à **0.20** ou l’arbre n’est toujours pas bien exploitable



Avec une valeur augmenter à **0.25**, on obtient un arbre ayant plus de nœuds que le précédent, le resultat est illustré dans la figure suivante :



En vue d’avoir un arbre avec plus d’information, la valeur **0.50** est attribuée pour le confidenceFactor. La figure suivante montre que l’arbre est bien exploitable avec un taux d’erreur moins considérable par rapport aux autres cas.



On assigne respectivement la valeur **0 .80** et **1.0** pour le facteur de confidence, on remarque que l’arbre généré n’est pas exploitable (voir les deux figures suivantes). De ce fait la valeur idéal pour le facteur est égale à **0.5**:

