Enseignate

[NGUYỄN Thị Minh Huyền]

## Fouille De Donnée

Données utilisées : Adult

**Jean Baptiste FLORIAL**

**Đào Thúy Hồng**

Année académique : 2016-2017



[**1.** **Préparer et importer le jeu de données :** 2](#_Toc470441144)

[**2.** **Statistiques sur des variables individuelles:** 6](#_Toc470441145)

[**2.1** **Statistiques élémentaires univariées discrètes :** 6](#_Toc470441146)

[**2.2** **Statistiques élémentaires univariées continues :** 9](#_Toc470441147)

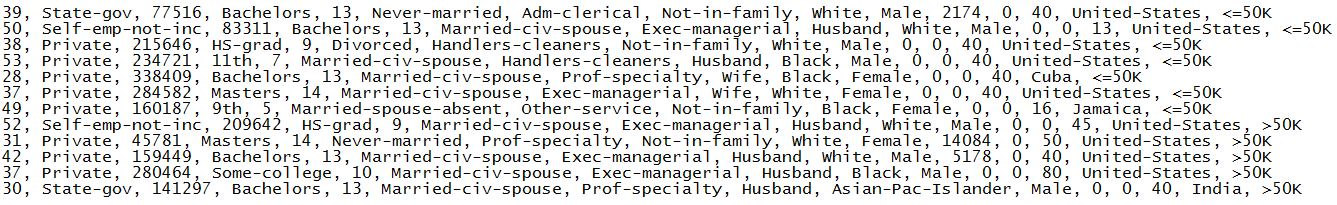
[**1.** **Des statistiques sur les liens entre des variables:** 11](#_Toc470441148)

[**1.1** **Test chi-deux :** 11](#_Toc470441149)

[**1.2** **One way ANOVA :** 13](#_Toc470441150)

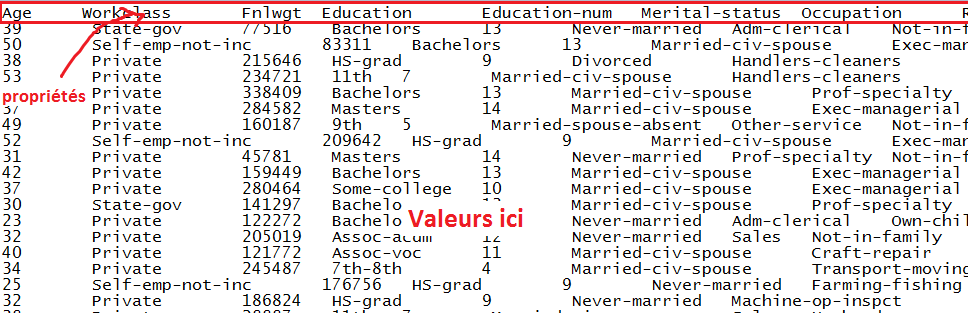
1. **Préparer et importer le jeu de données :**

Le jeu de donnée que nous avons téléchargé sur la ligne <http://archive.ics.uci.edu/ml/datasets/Adult> est en forme comme ça :



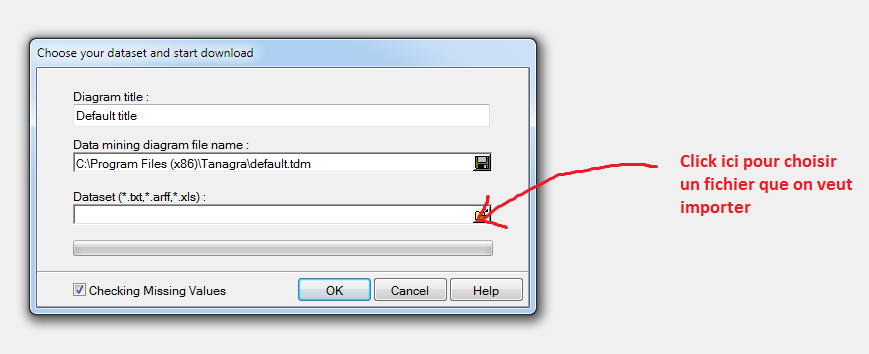
*Figure 1 : Forme du jeu de données.*

Chaque ligne est un échantillon. Dans chaque ligne, on a les valeurs des propriétés (on a déjà les décris dans le TP2) qui sont séparés par une virgule. Avant d’importer ce jeu de données, on doit convertir ce forme en colonne et ajouter le nom de chaque propriété. Nous utilisons le logiciel Open Excel pour faire ça.



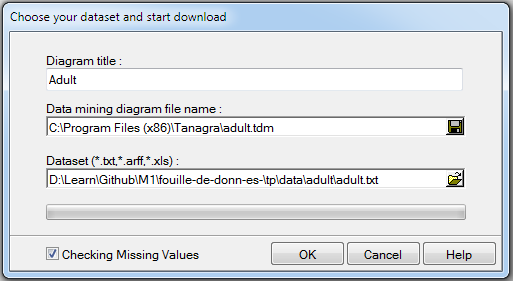
*Figure 2 : Jeu de données après de convertir.*

Pour importer ce fichier, on ouvre Tanagra et puis va **File/New** dans le menu.



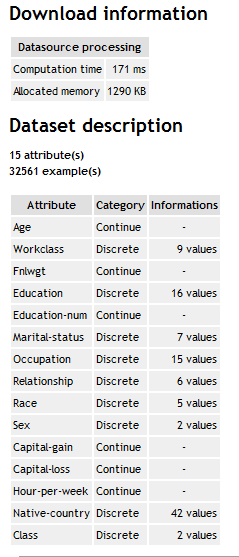
*Figure 3 : Interface pour importer le jeu de données.*

On entre le nom de diagramme **Nursery**, le nom et le chemin de fichier de fouille de données **C:\Program Files (x86)\Tanagra\nursery.tdm** et enfin choisit le chemin du jeu de données **D:\Learn\Github\M1\fouille-de-donn-es-\tp\data\nursery.txt**. Le résultat comme dans la capture suivante :



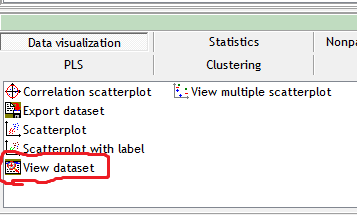
*Figure 4 : Capture du choix du jeu de données.*

On valide par **OK**, on a alors :



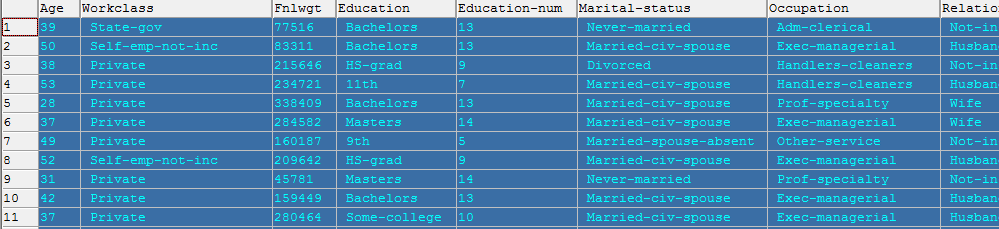
*Figure 5 : Description du jeu de données nursery.txt*

On peut voir en détail du jeu de données en glissant l’opérateur **View** **dataset** de catégorie **Data visualization** sur le **Dataset.**

****

*Figure 6 : Chosir View dataset.*

On fait un clic double sur l’opérateur  **View dataset 1** et obtient le résultat comme suivant :

****

*Figure 7 : Détail du jeu de données.*

1. **Statistiques sur des variables individuelles:**

Tout d’abord, on fait un coup d’oeil sur le statistique du jeu de données. Il y a 7 atributs discrets et une seule attribut continu.

* 1. **Statistiques élémentaires univariées discrètes :**

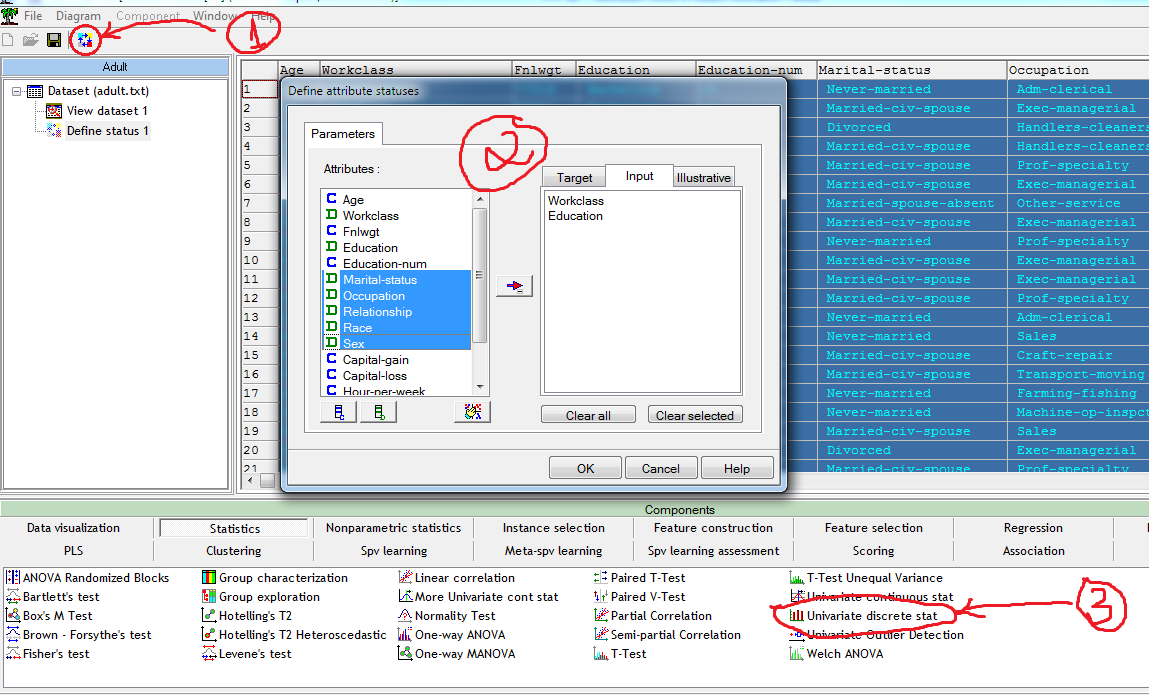
Pour voir le statistique des attributs discrets, on fait comme ça :

**Étape 1 :** Choisir **Define status** dans la barre de menu de logiciel. Ce qui fait apparaître une boîte de dialogue.

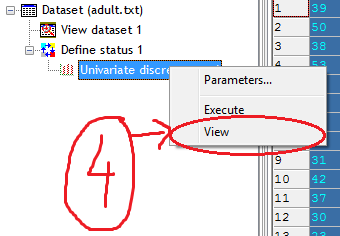
**Étape 2 :** Sellectionner les attributs discrets dans la boîte : *parents, has\_nurs, housing, finance, social, health* pour **Input** et valider par **OK.** Et puis, on fait un clic droit sur **Define status** et clicque sur **Execute**.

**Étape 3 :** Rajouter l’opérateur **Univariate discrete stat** de l’onglet **Statistics** dans **Define status**.

**Étape 4 :** Ensuite, on fait un clic droit sur **Univariate discrete stat** et clicque sur **View**.

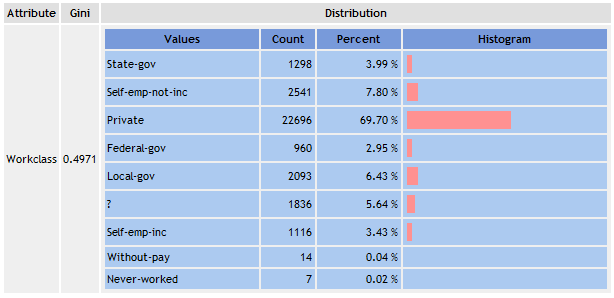


*Figure 8 : Trois premières étapes.*

**

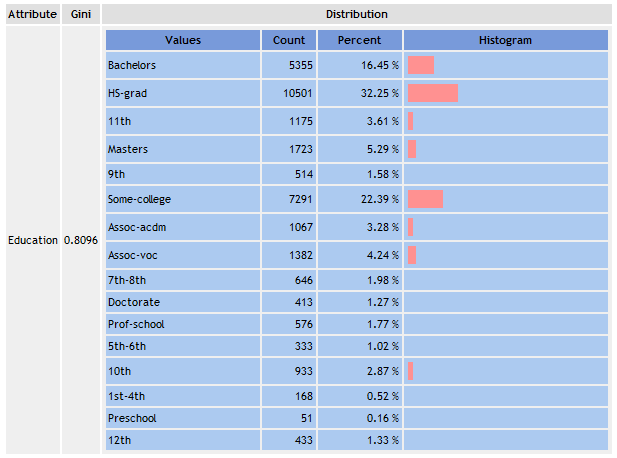
*Figure 9 : Capture de la étape 4.*

Après ces étapes, on obtient les statistiques élémentaires univariées suivantes :

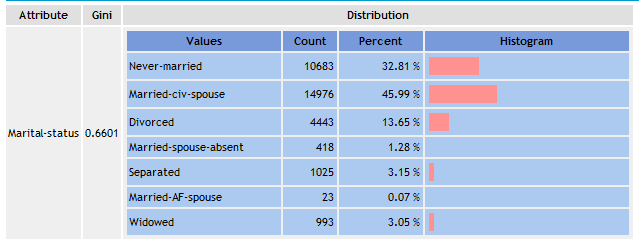


*Figure 10 : Statitisque de l’attribut* ***Workclass****.*

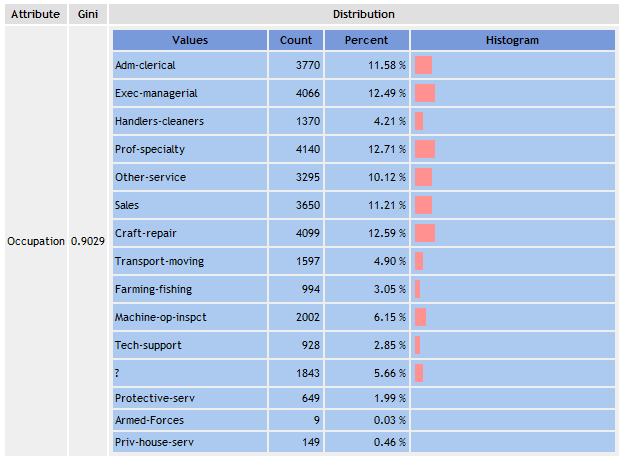
Selon la tableau au-dessus,



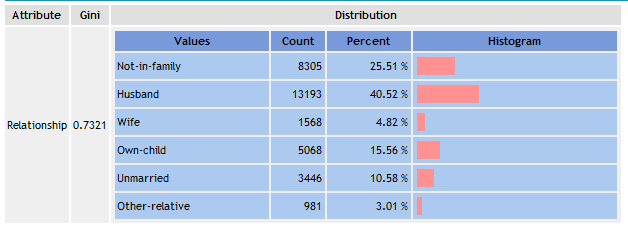
*Figure 11 : Statistique de l’attribut* ***Education****.*

******

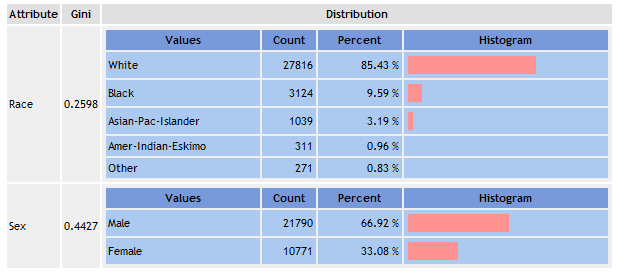
*Figure 12 : Statistique de l’attribut* ***Marital-status****.*

******

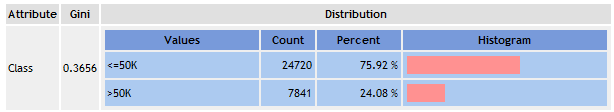
*Figure 13 : Statistique de l’attribut* ***Occupation****.*

******

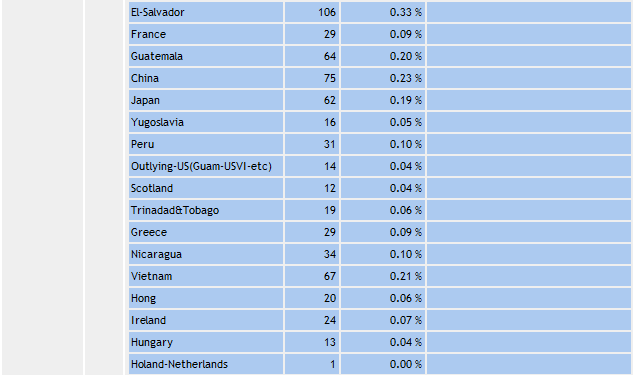
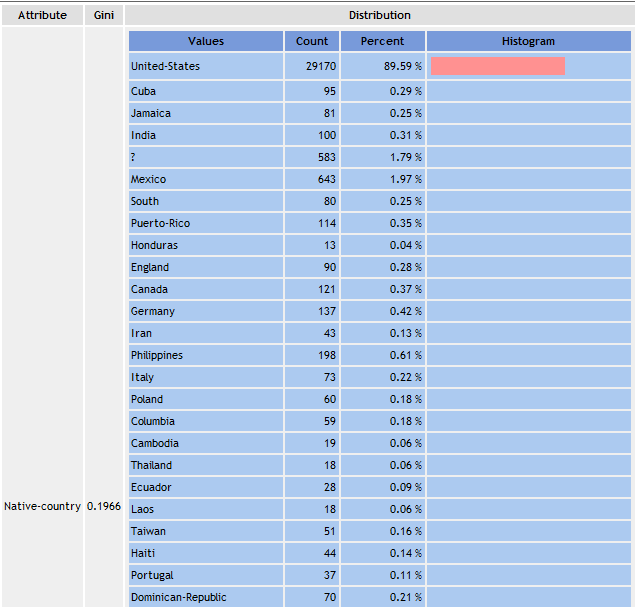
*Figure 14 : Statistique de l’attribut* ***Relationship****.*

******

*Figure 15 : Statistique des attributs* ***Race*** *et* ***Sex****.*

**

*Figure 16 : Statistique de l’attribut* ***Class****.*

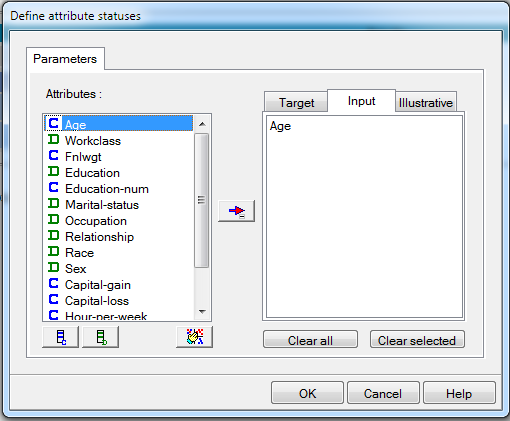
**

*Figure 17 : Satistique de l’attribut* ***Native-country****.*

Selon les statistiques des attributs, on voit que pour chaque attribut, la distribution de probabilité de chaque valeur est égal.

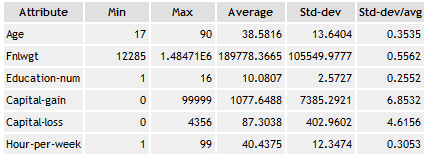
* 1. **Statistiques élémentaires univariées continues :**

On fait 4 étapes comme dans la partie 3.1, mais dans l’étape 2, au lieu de sélectionner les attributs discrets, on choisit l’atribut continu *children* – ça veut dire le nombre des enfants de chaque famille. Dans l’étape 3, on rajoute l’opérateur **Univariate continue stat** au lieu de rajouter l’opérateur **Univariate discrete stat**.



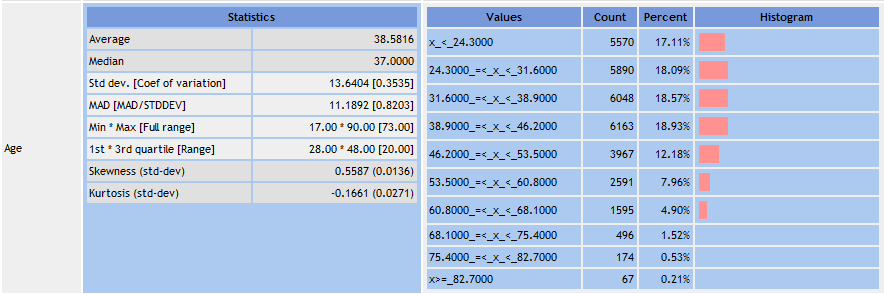
*Figure 18 : Choisir les attributs continus.*

Le résultat est obtenu comme suivant :

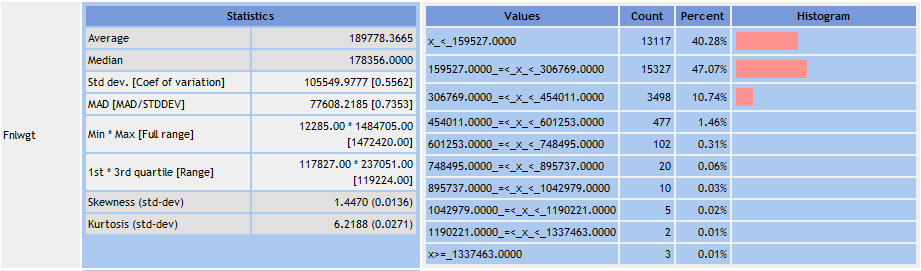


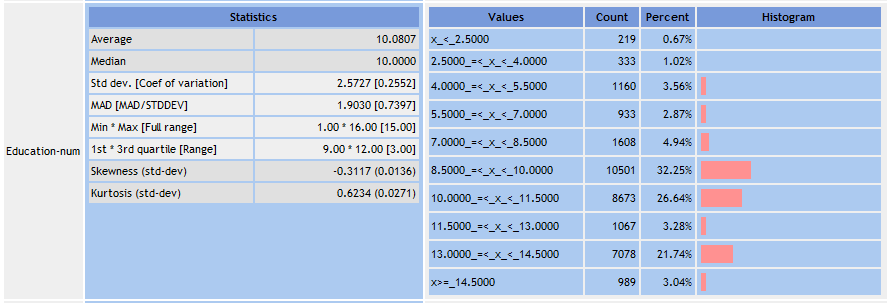
*Figure 19 : Statistique de les attributs continus.*

On peut voir en détail du statistique des atributsen utilisant l’opérateur **More univariate cont stat**. Ce qui nous donne le résultat au-dessous :

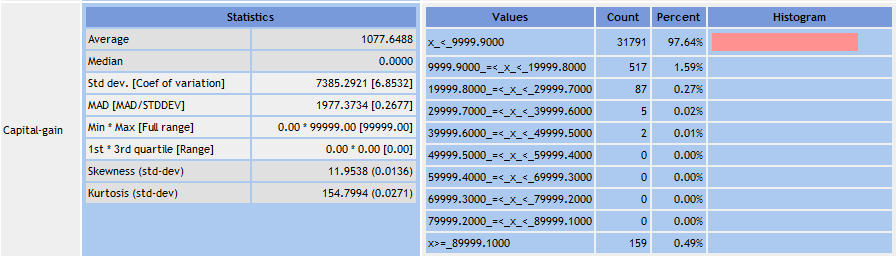


*Figure 20 : Détail du statistique de l’attribut* ***Age****.*

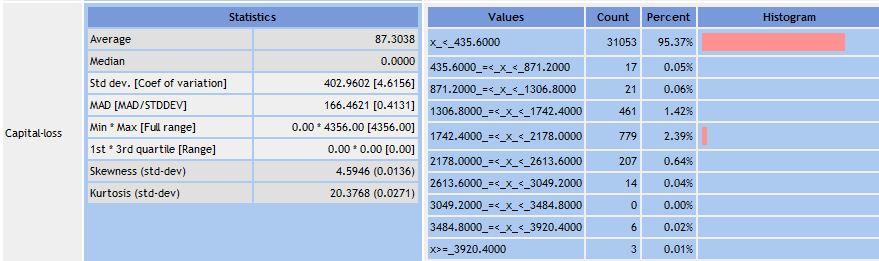
* Figure 21 : Détail du statistique de l’attribut* ***Fnlwgt****.*

**

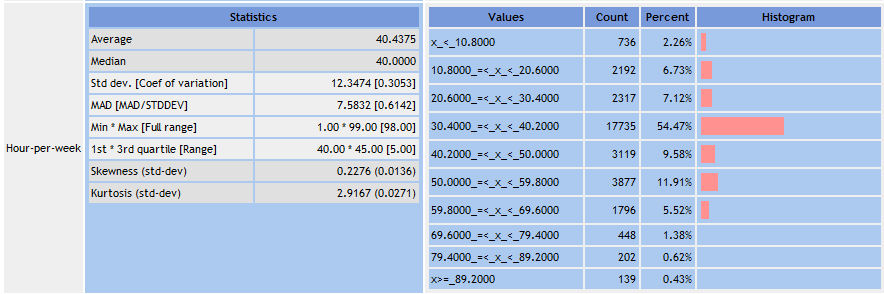
*Figure 22 : Détail du statistique de l’attribut* ***Education-num****.*

**

*Figure 23 : Détail du statistique de l’attribut* ***Capital-gain****.*

**

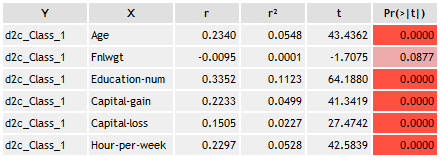
*Figure 24 : Détail du statistique de l’attribut* ***Capital-loss****.*

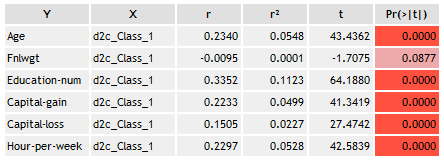
**

*Figure 25 : Détail du statistique de l’attribut* ***Hour-per-week****.*

En observant le statisque de l’attribut *children*, on trouve que la valeur minimum est 1, la valeur maximum est 4, la valeur moyenne est 2.5. Cet attribute a 4 valeur et le pourcent de chaque valeur est égale à 25%. L’écart-type est 1.1181, cette valeur n’est pas grande ça veut dire qu’ il n’existe pas de valeur qui sont loin la valeur moyenne.

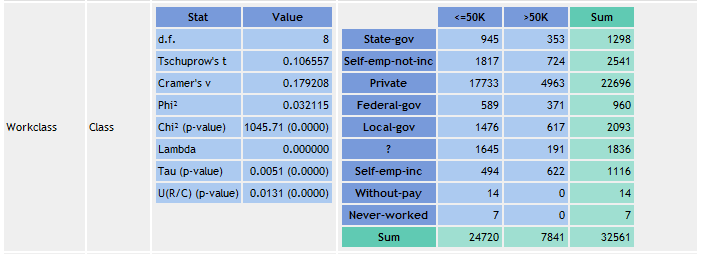
1. **Des statistiques sur les liens entre des variables:**
   1. **Corrélation linéaire :**

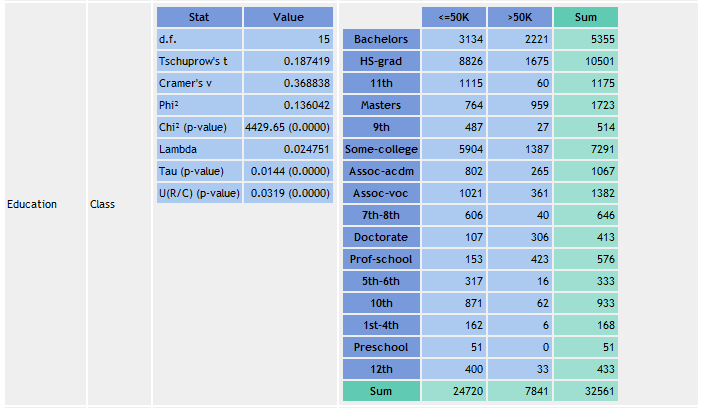
****

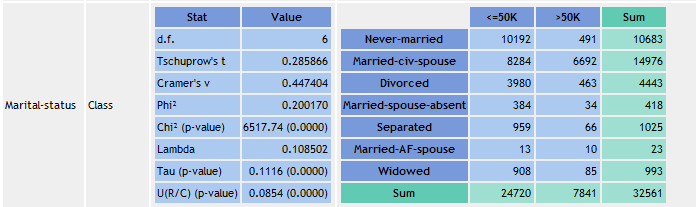
****

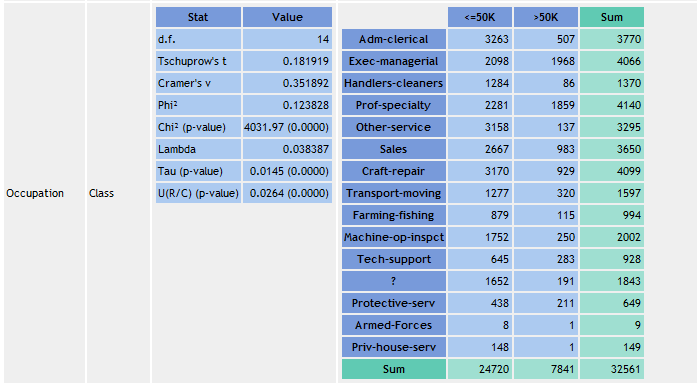
* 1. **Test chi-deux :**

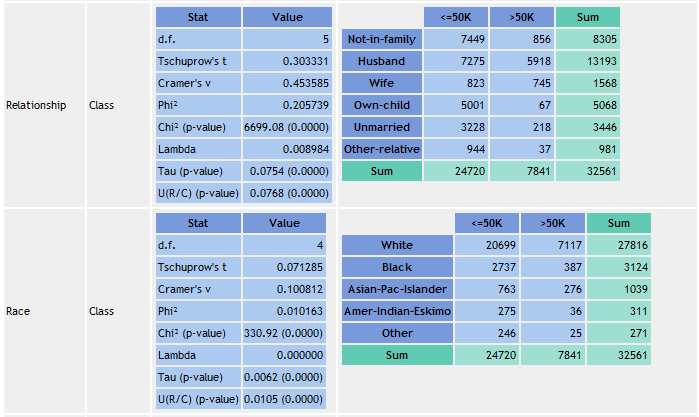
On a 9 attributs qualitatifs : ***Workclass, Education, Marital-status, Occupation, Relationship, Race, Sex, Native-country, Class***; on utilise l’opérateur **Contingence Chi-Square** dans Tanagra pour les verifier. On a le résultats au-dessous.

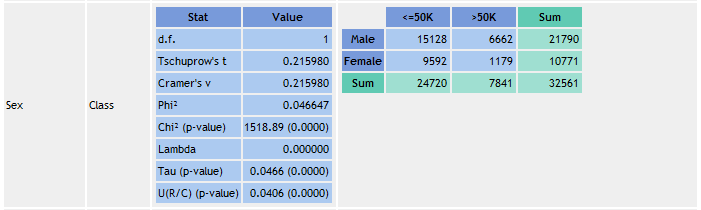


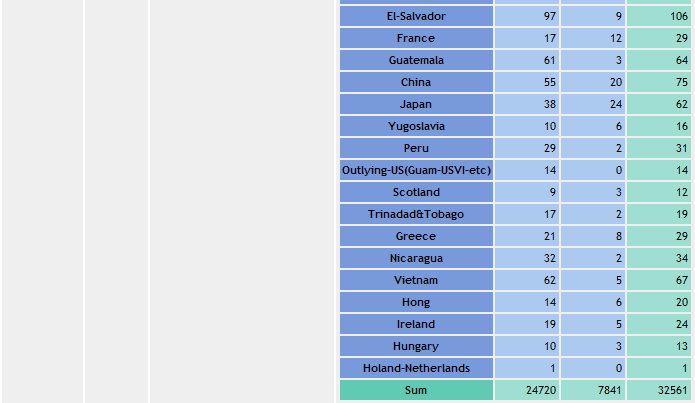
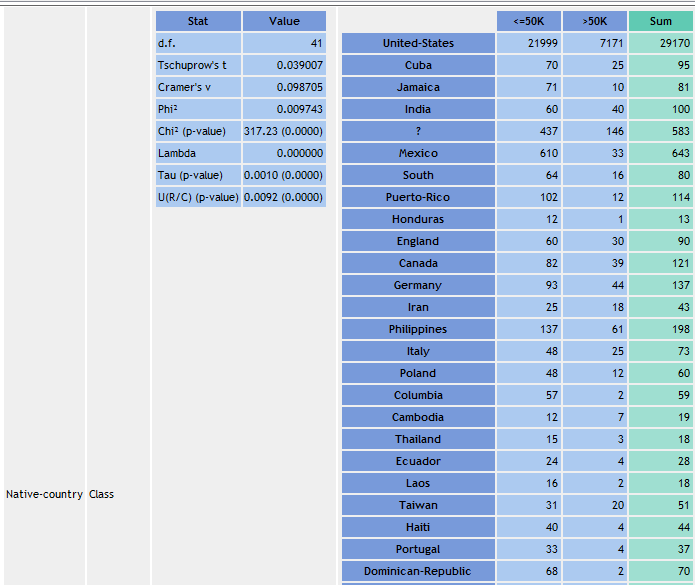






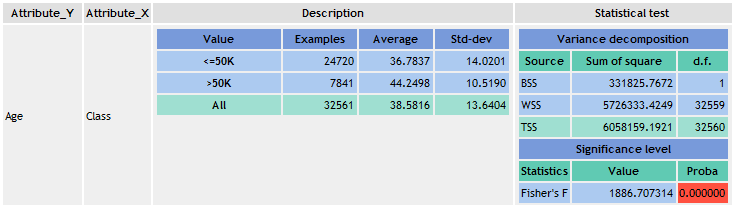




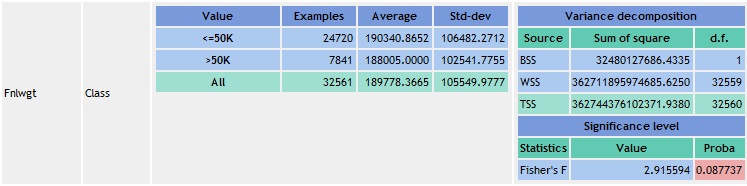


* 1. **One way ANOVA :**

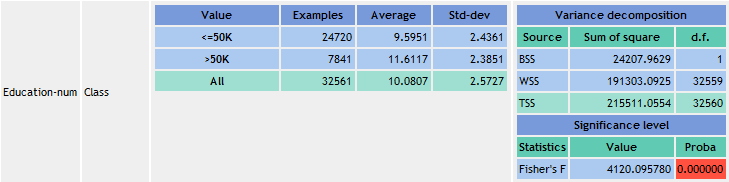
On vise à tester les différences significatives entre les moyennes en utilisant l’opérateur **One way ANOVA**. Cet opérateur demande l’attribut continu comme **Target** et l’attribut discret comme **Input.** Il y a un seul attribute continu dans ce jeu de données, alors on définit successivement *parents, has\_nurs, housing, finance, social, health* comme **Input**, *has\_nurs* comme **Target** pour verifier le lien entre eux.



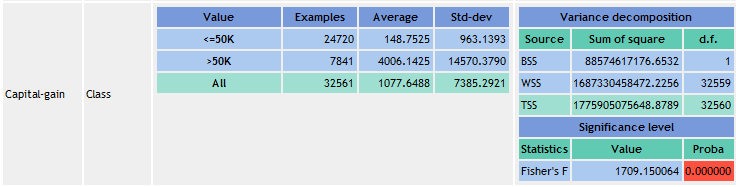
*Figure 19 : Lien entre* ***Age*** *et* ***Class****.*



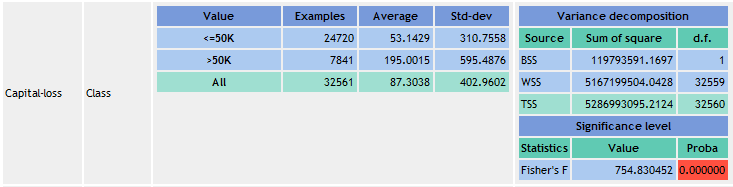
*Figure 20 : Lien entre “children” et has\_nurs”.*



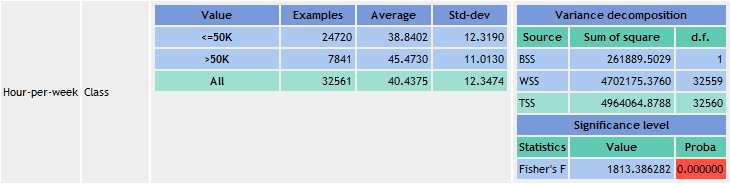
*Figure 20 : Lien entre* ***Education-num*** *et* ***Class****.*

**

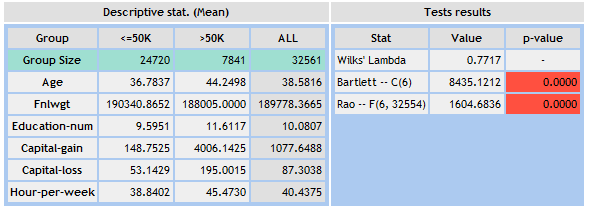
*Figure 21 : Lien entre* ***Capital-gain*** *et* ***Class****.*

**

*Figure 22 : Lien entre* ***Capital-loss*** *et* ***Class****.*

**

*Figure 23 : Lien entre* ***Hour-per-week*** *et* ***Class****.*



One way manova