



Master science et technique

Filière : système d'information décisionnel et imagerie

MINI PROJET DATA MINING

Thème

Extraction de la connaissance à partir d’une dataset en utilisons Python

Présenté Par :

-BOUIGADERN ABDELAZIZ

Encadré Par :

-MOHAMED SABIRI

Année Universitaire 2023-2024

**I-Introduction :**

**-Description de dataset :**

La base de données Titanic décrit le statut de survie des passagers individuels sur le Titanic.

-Variable descriptions

* Pclass : Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
* Survival: Survival (0 = No; 1 = Yes)
* name :Name
* sex :Sex
* age :Age
* sibsp :Number of Siblings/Spouses Aboard
* parch :Number of Parents/Children Aboard
* ticket :Ticket Number
* fare :Passenger Fare (British pound)
* cabin :Cabin
* embarked :Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
* boat :Lifeboat
* body Body :Identification Number
* home.dest : Home/Destination

-Notes spéciales

* Pclass : est un indicateur du statut socio-économique (SES)

1er ~ Supérieur; 2ème ~ Moyen; 3ème ~ Basse

* Age : est en années; Fractionnel si l'âge est inférieur à un (1)

Si l'âge est estimé, il est sous la forme xx.5

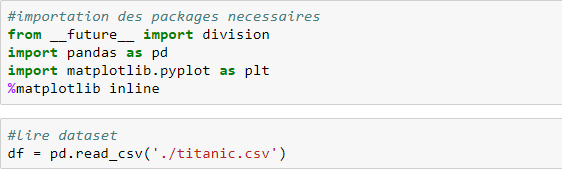
* Fare : Le prix est en livres britanniques d'avant 1970 ()

Facteurs de conversion: 1 = 12s = 240d et 1s = 20d

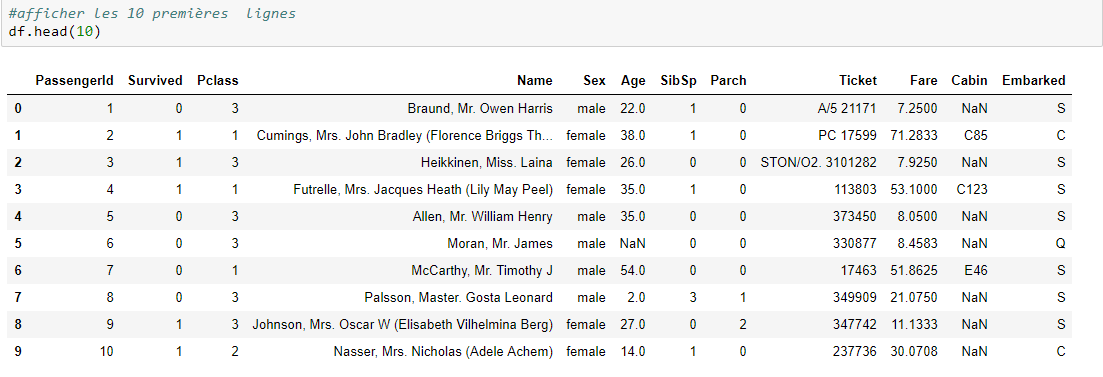
En ce qui concerne les variables de la relation familiale (c'est-à-dire sibsp et parch), certaines relations :

* **Sibling**: Frère, sœur, du passager à bord du Titanic
* **Spouse**: Mari Ou femme du passager à bord du Titanic
* **Parent**: Mère ou père du passager à bord du Titanic
* **Child**: Fils, Fille, Stepson, ou belle-fille du passager à bord du Titanic

**Lecture du fichier :**



**APERÇU DU FICHIER :**

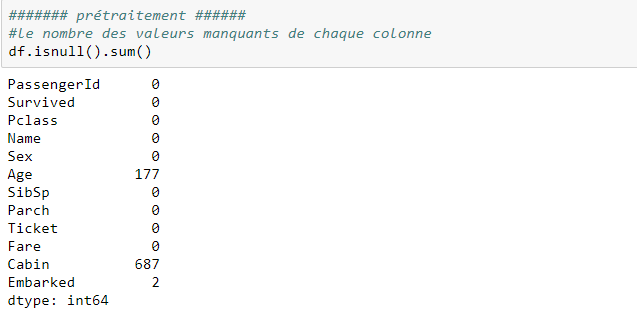


**II- PRE-TRAITEMENT :**

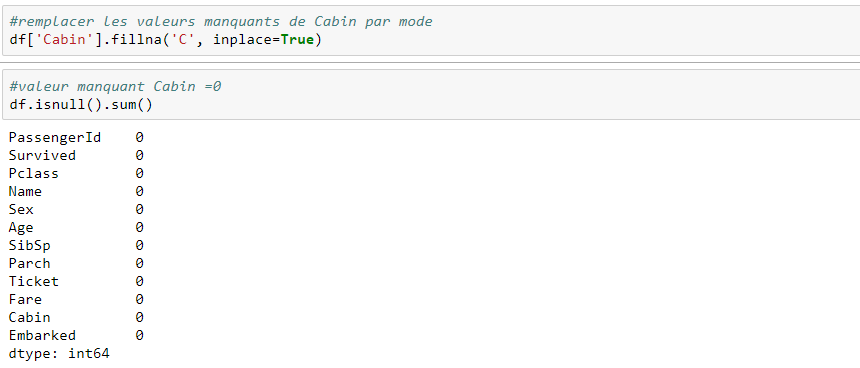
Dans cette partie consiste à faire quelque modification pour donner une bonne étude statistique :

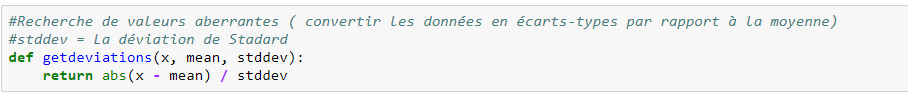
* **Manipuler les valeurs manquantes :**

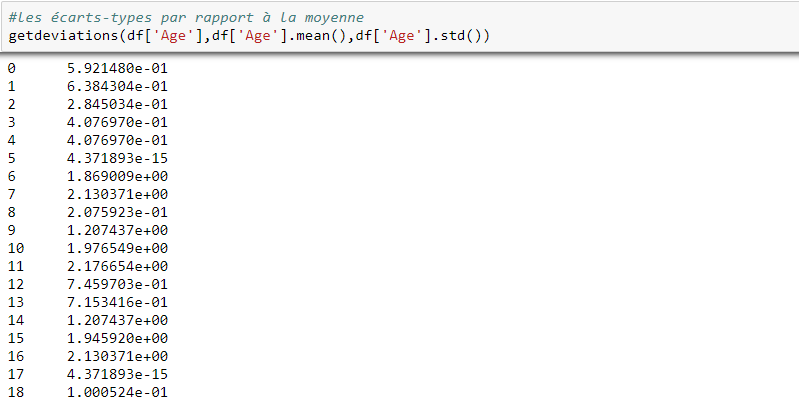
En premier on va compter les valeurs manquantes dans chaque colonne par script suivant :



On remarque que les variables (Age, Embarked, Cabin) contient des valeurs manquants, et pour nettoyer notre dataset :

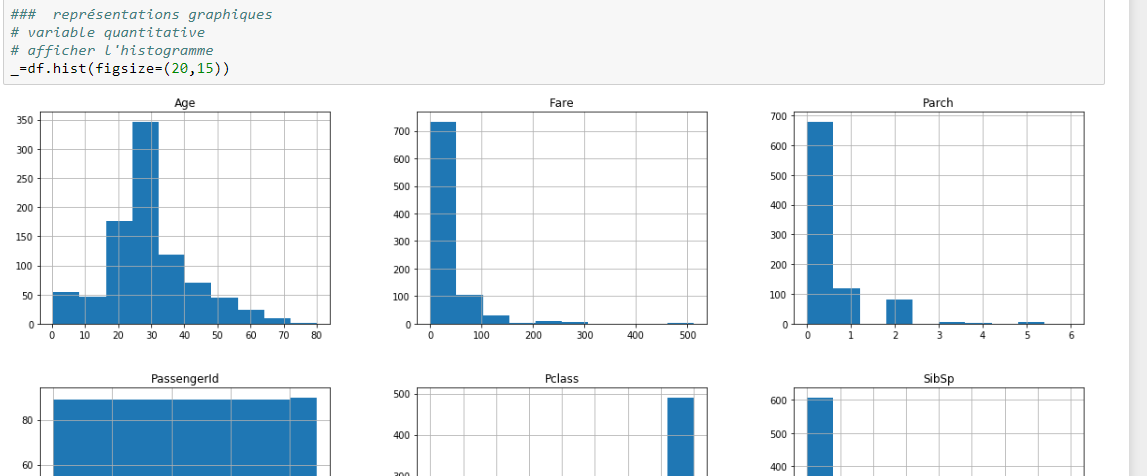
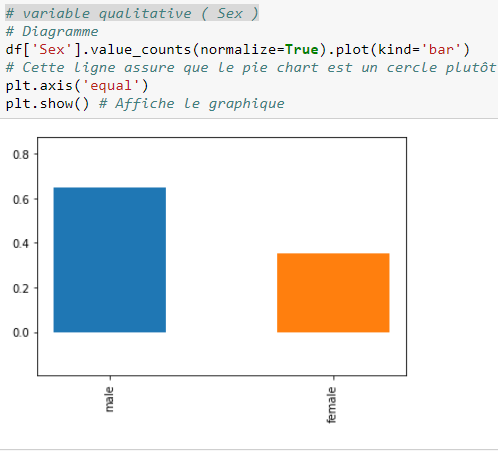
* on va tout d’abord calculé la moyenne de la variable ‘ âge’, en suite on va remplacer ses valeurs manquantes par la moyenne calculé.
* Et pour vérifier on a lancé un petit script où on remarque que le nombre des valeurs manquants de la variable Age est égale à 0 parés notre nettoyage.
* Et pour la variable  ‘Embarked’ On va grouper selon la variable en suite on va calculer la fréquence de chaque valeurs de cet variable et on a remplacé nos valeurs manquantes par valeurs la plus fréquente qui est la valeur ‘S’.
* Et on suite on a effectué a lancé un script pour vérifier le bon déroulement de notre nettoyage.
* Même principe on traite les valeurs manquantes de variable Cabin par mode.
* **Recherche des valeurs aberrantes :**

Le but de cette partie est convertir les données en écarts-types par rapport à la moyenne pour connaitre les valeurs aberrantes ****

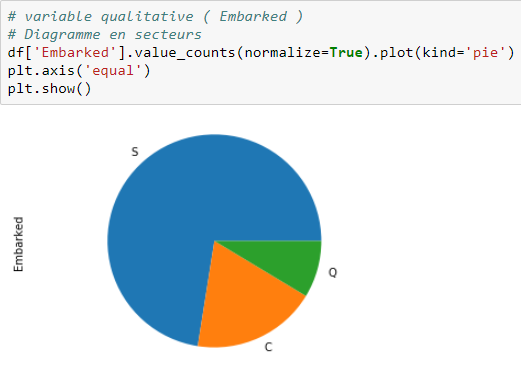


On remarque que tous les observations sont proches entre eux donc on ne possède pas des valeurs aberrants.

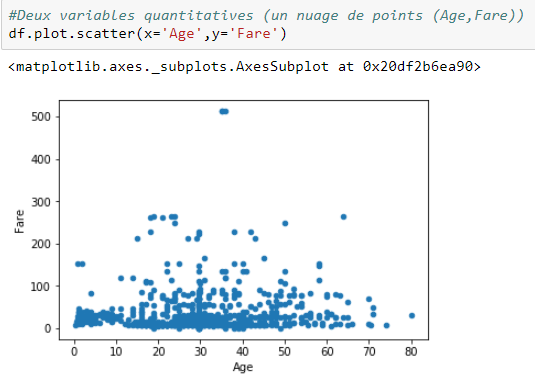
**III-Analyse du Donnés :**

* **Représentations graphiques :**
* **Pour variable quantitative (histogramme)**
* **Pour variable qualitative**

On constate que le Sexe masculin est plus fréquent que féminin.

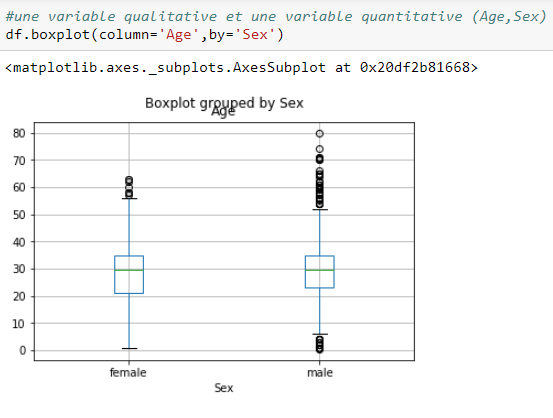


Dans ce diagramme on voit que pour la variable Embarked on trouve la valeur S après la valeur C et enfin la valeur Q

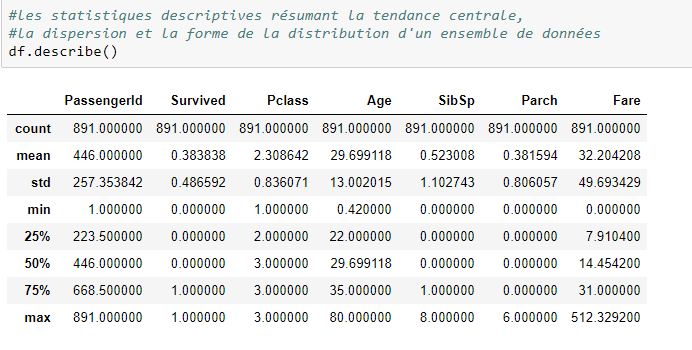
* **Deux variables quantitatives :**

Le nuage de points indique le degré de corrélation entre deux ou plusieurs variables liées. Chaque unité représente un point dans le nuage.

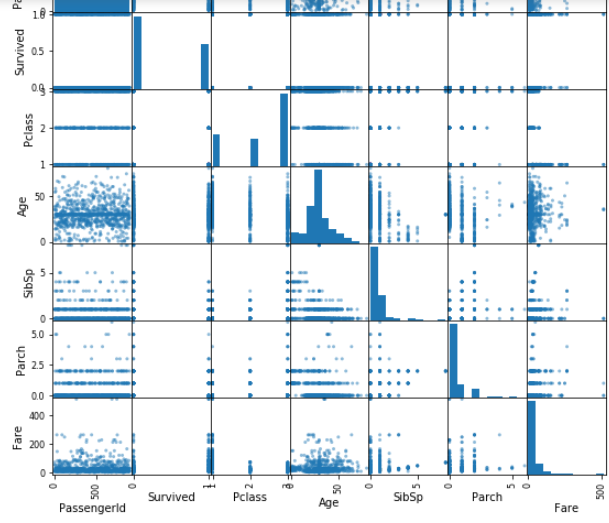
L'abscisse (Age) augmente de gauche à droite tandis que l'ordonnée (Fare) augmente de bas en haut. À mesure que les valeurs augmentent, les points se déplacent de la base inférieure gauche vers la droite inférieure.

* **une variable qualitative et une variable quantitative (boîte à moustache)**

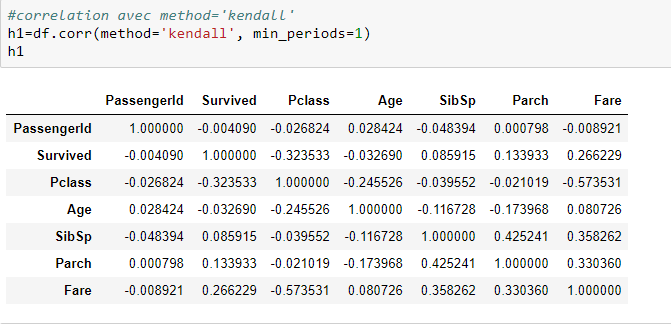
Dans cette feuille de travail, Age correspond à la variable de graphique et Sex correspond à la variable de catégorie pour le regroupement. Le graphique présente la loi de distribution des Age pour chaque type de Sexe.

****On remarque que les valeurs éloignées qui se situent entre 1,5 et 3 longueurs de boite à partir de la bordure inférieure moins que les valeurs éloignées de la bordure supérieure de la boite.

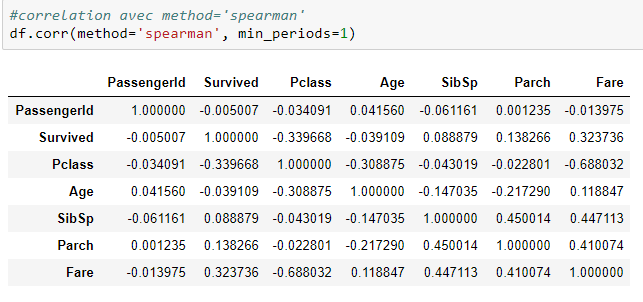
Comme vous pouvez le voir, nous obtenons pas mal d'informations ici avec juste une invocation rapide de .describe (). Nous obtenons le nombre, qui est combien de lignes nous avons pour chaque colonne. Nous obtenons alors la moyenne, ou la moyenne, de toutes les données dans cette colonne. STD est l'écart type pour chaque colonne. Min est la valeur minimale de cette ligne. 25% correspond à la marque du 25ème centile, et ainsi de suite à 75%. Enfin, nous obtenons max, qui est la valeur la plus élevée pour cette colonne.

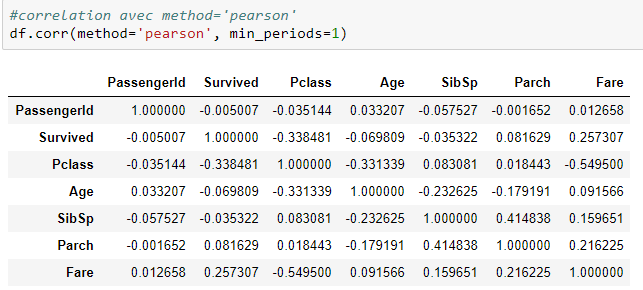
****

* **Corrélation entre deux variables quantitatives**

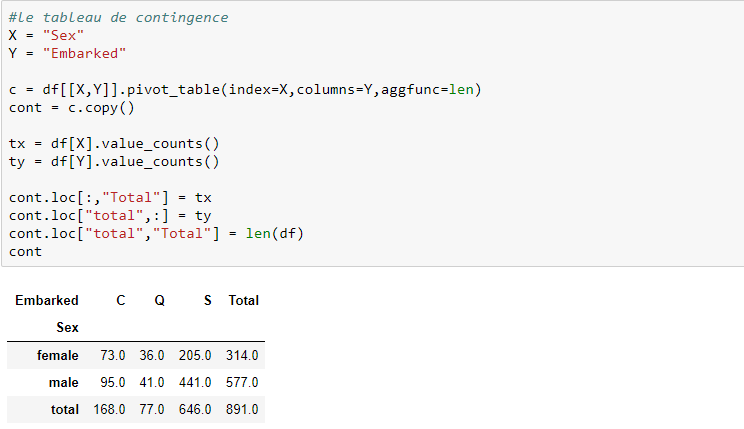
****

Ici, nous obtenons la corrélation de chaque colonne par rapport à l'autre. Comme vous pouvez le voir, nous obtenons un tableau de comparaison. Évidemment, tous les variables ne sont pas tous très étroitement corrélés.

* **Corrélation avec méthode de Spearman**
* **Corrélation avec méthode de pearson**

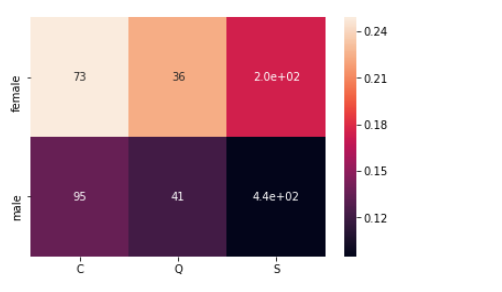
****

* **Corrélation entre deux variables qualitatives (tableau de contingence)**

****

 La fréquence des observations présentant à la fois la caractéristique i pour la variable Sexe, et la caractéristique j pour la variable Embarked.

Apres on ajoute une mesure statistique. (Cette mesure est calculable pour chacune des cases du tableau de contingence.) Appliquer sur cette mesure un seuil au-delà duquel on dira que les 2 variables sont corrélées. (Une contribution au non indépendance)

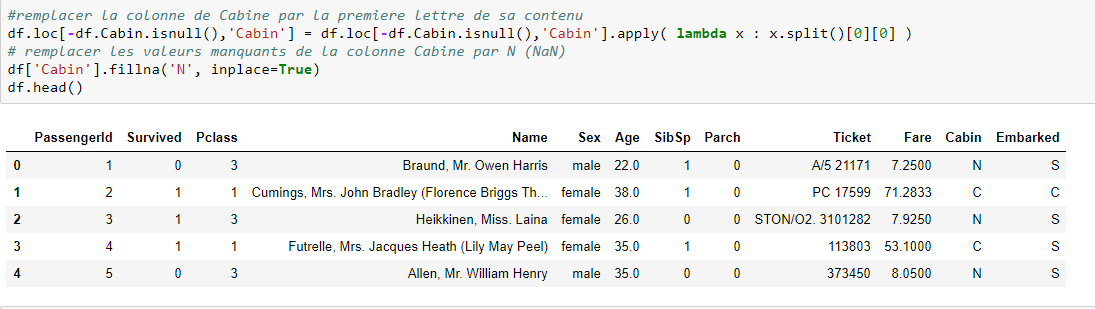
****

On constate d’après le seuil que Sexe féminin et Embarked\_C sont corrélés.

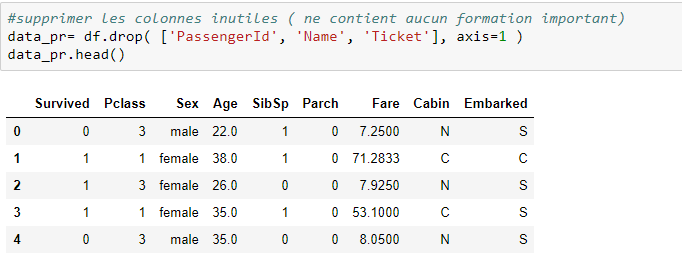
**III-Extraction de connaissances par « Règles d’association »**

Pour faire cette Extraction il faut faire quelque modification pour bien Etablir cette connaissance :

* Remplacer la colonne de Cabine par la première lettre de son contenu :

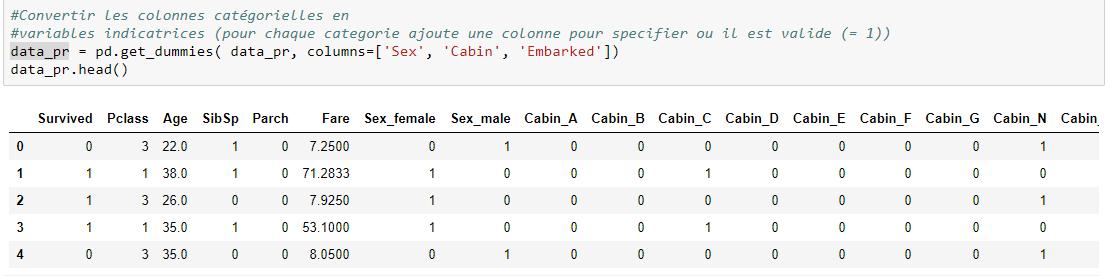
****

Apres on supprime les colonnes inutiles qui ne contient aucun information important comme PassengerId, Name , Ticket avec les script suivant :

****

* Convertir les colonnes catégorielles en variables indicatrices (0 ou 1) :

Pour chaque catégorie ajoute une colonne et donner la valeur 1 pour spécifier où il est valide. On fait cette transformation pour les colonnes (Sex, Cabin ,Embarked). Avec ce script :

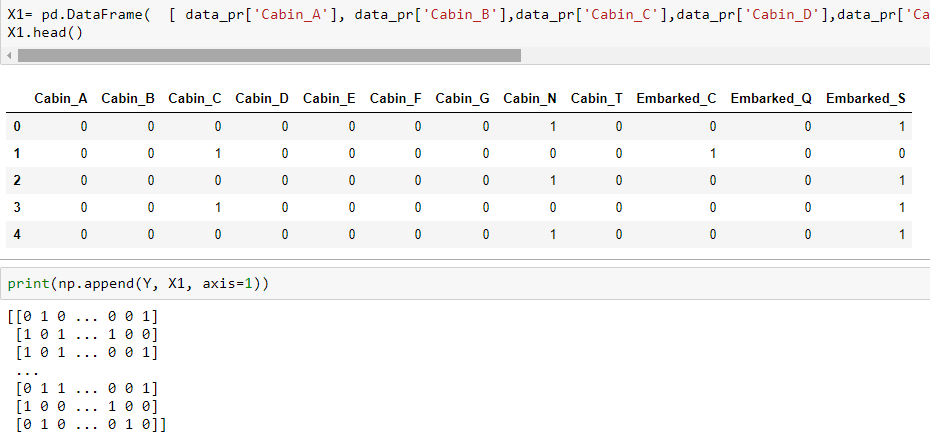
****

* On traite la variable Survived

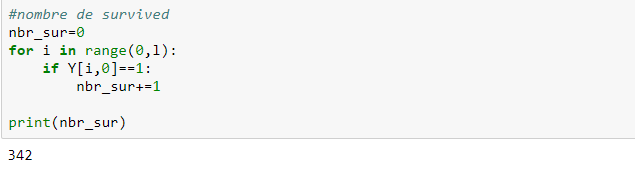
****On devise cette colonne en deux colonnes une colonne contient Survived et d’autre contient not Survived.

Donc après cette transformation on a la table suivant :

Y=['Survived','not\_Survived','Sex\_female','Sex\_male'] et ajoute d’autre colonne pour faire la règle [‘Cabine\_A’ …… , ‘Cabine\_T’,’Embarked\_C’, … ,’Embarked\_S’]

****

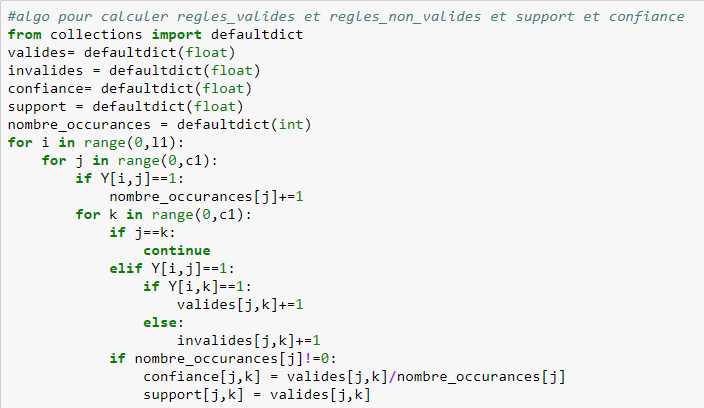
-on peut calculer le nombre de Survived :

****

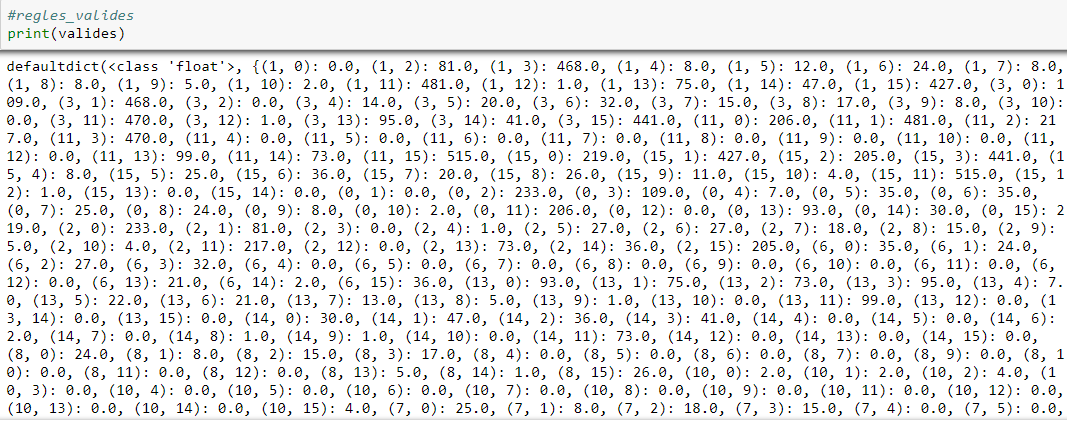
* On utilise maintenant l’algorithme a priori :

L’algorithme Apriori est un algorithme d’exploration de données , dans le domaine de l’apprentissage des règles d’association. Il sert à reconnaître des propriétés qui reviennent fréquemment dans un ensemble de données et d’en déduire une catégorisation. Chaque règle d’association a deux mesures : le « support » et la « confiance ». Le support d’un ensemble d’items est définit comme la fréquence d’apparition simultanée des items figurant dans l’ensemble des données.

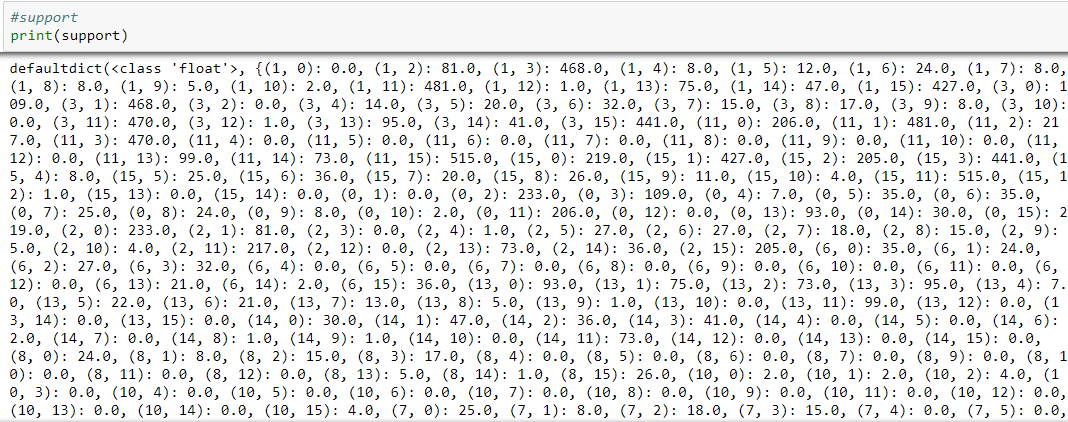
On calcule aussi regles\_valides ,regles\_non\_valides, support et confiance de chaque deux variables pour déterminer la règle la plus fréquente.

****

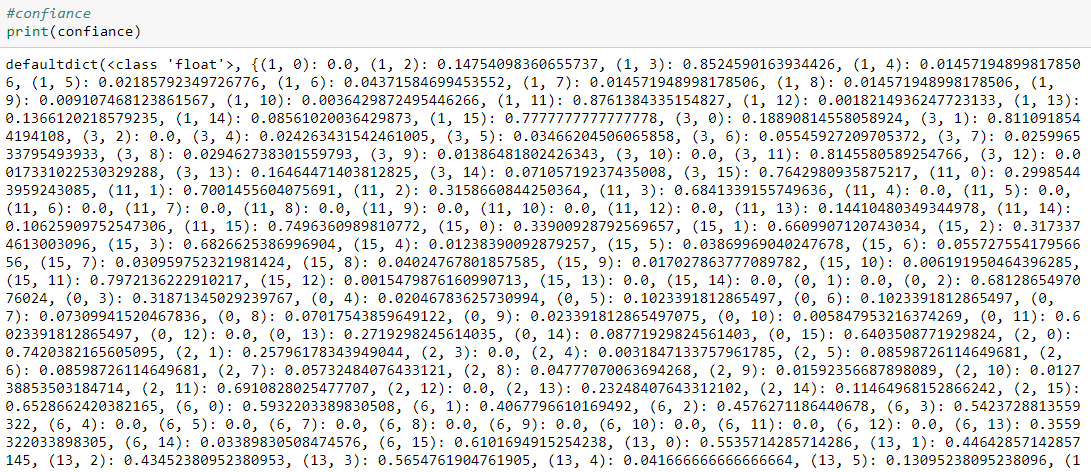
**Donc On peut afficher les regles\_valides entre chaque deux variable :**

****

**Et le support :**

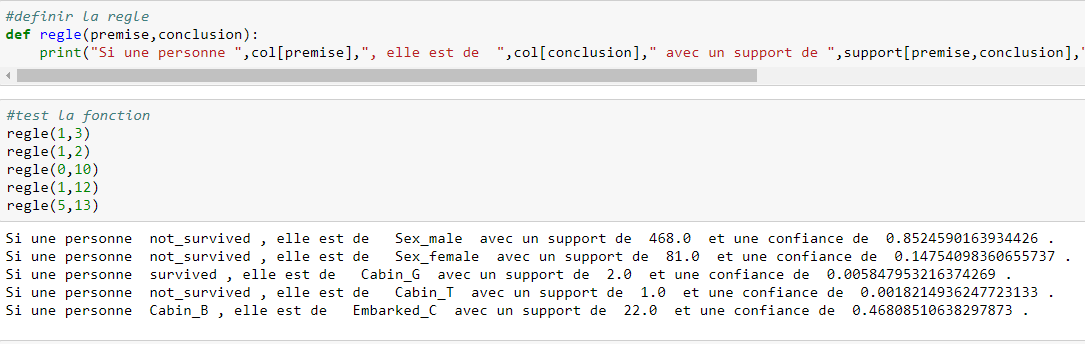
****

**La confiance :**

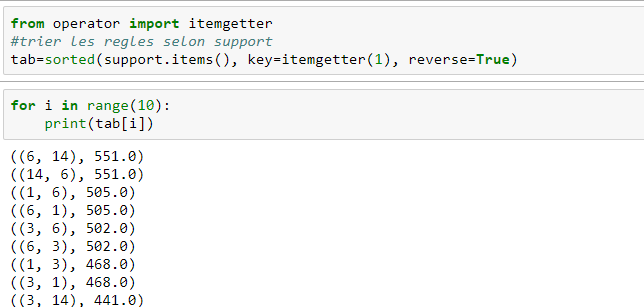
****

On définit une fonction qui affiche pour chaque prémisse et conclusion en entrée, le Support et la confiance que nous venons de calculer.

Et faire quelque test :

****

Généraliser le programme précèdent pour afficher les 10 premières règles en les triant par support avec ce script :

****

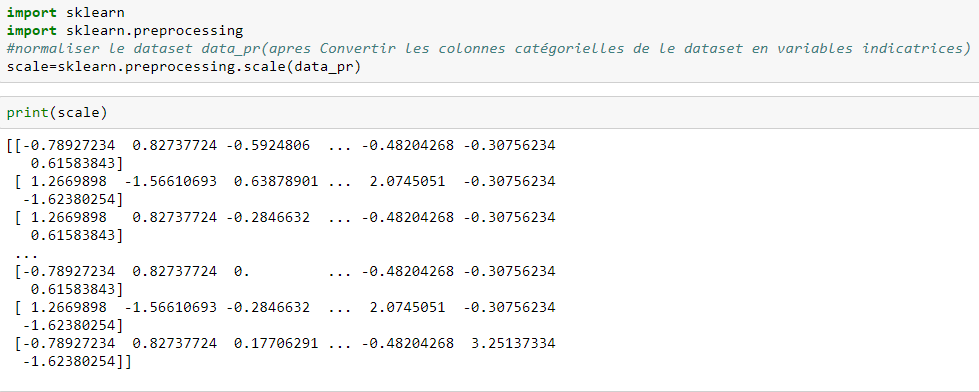
On remarque que l’indice 6 et 11 et plus fréquente il correspond à Cabin\_C et Embarked\_S donc la règle est suivant :

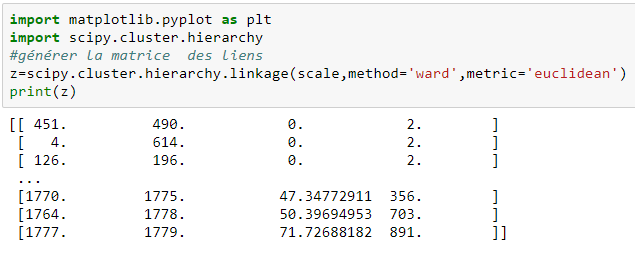
*« Si une personne Cabin\_C, elle est d’Embarked\_S avec un support de 551.0 et une confiance de 0.7386058981233244. »*

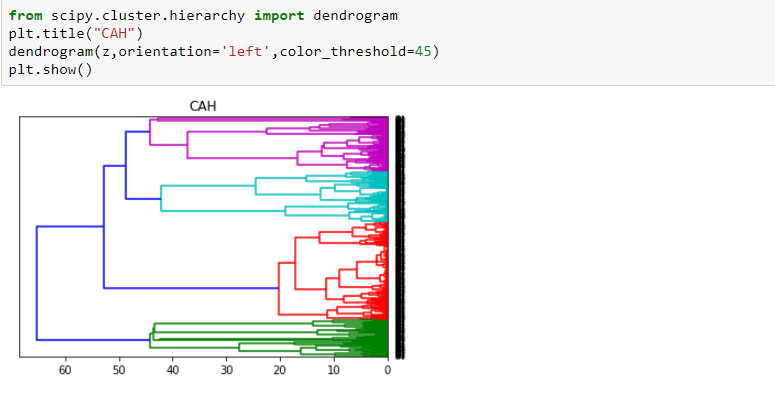
**IV-Extraction de connaissances par « Classification Ascendante Hiérarchique ou d’autre type de classification»**

Une première approche consiste à utiliser la Classification Ascendante Hiérarchique (CAH). Le graphe qui en résulte permet de se faire une idée visuelle des différents regroupements et d’intuiter le nombre de classes. Les variables étant toutes numériques, nous utiliserons la distance euclidienne comme mesure de dissimilarité et la distance de “Ward” comme mesure de dissimilarité inter-classe (cette distance vise à maximiser l’inertie inter-classe)

-importer les packages nécessaires avant de normaliser le dataset data\_pr (cette data qui construit après Convertir les colonnes catégorielles de le dataset en variables indicatrices)

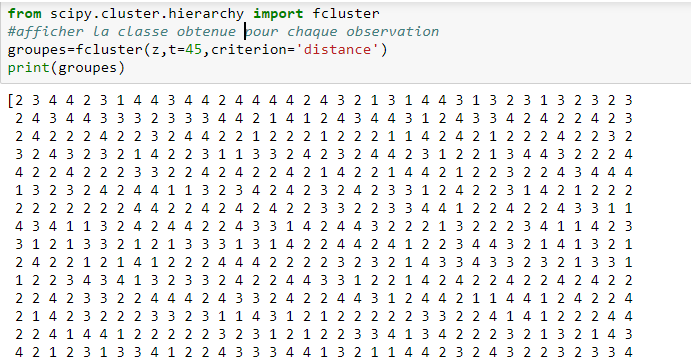
****

Apres générer la matrice des liens :

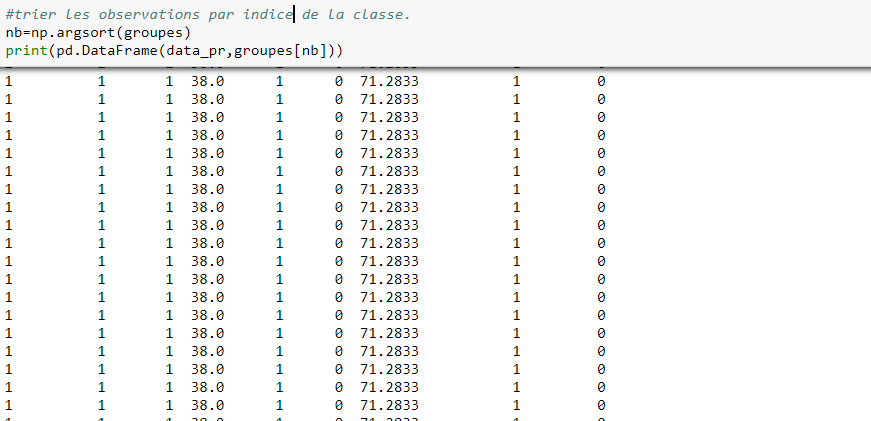
****

4 groupes constitués

Nous allons alors afficher la classe obtenue pour chaque observation

****

Trier les observations par indice de la classe.

****

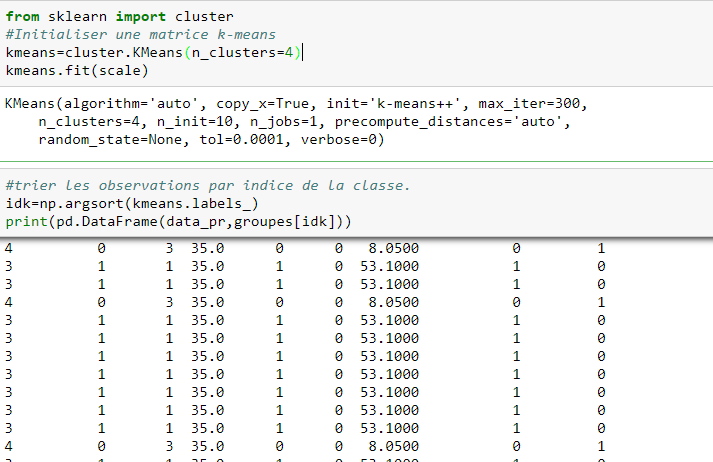
En coupant à une distance d’environ 45, nous pouvons former quatre groupes. Néanmoins on voit bien que pour beaucoup de groupes le regroupement se fait à une distance élevée ce qui inciterait à former plus de groupes (pour avoir plus d’homogénéité). Si nous avions coupé à 40, nous aurions alors 11 groupes (et la encore il faudrait couper plus bas).

Néanmoins la coupure aux alentours des 45 est intéressante. En effet, même si ces groupes n’apparaissent pas parfaitement homogènes sur la figure, si je regarde les ascensions à l’intérieur de ces groupes, je retrouve bien des similarités de difficulté (évaluées selon ma propre expérience) entre les éléments qui le composent.

Nous pouvons à présent essayer d’utiliser une méthode de classification qui n’est pas basée sur des arbres, par exemple les K-Means

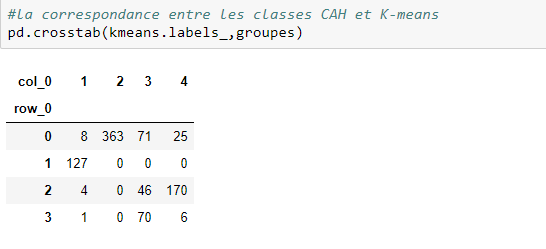
**K-means**

Initialiser une matrice k-means et trier les observations par indice de la classe :

****

On donne le paramètre n\_clusters=Nombre de classe trouve dans CAH (4).

Et enfin Afficher la correspondance entre les classes CAH et K-means



On remarque qu’il n y a aucun groupe correspond à k-means et a CAH en même temps, Dans cet exemple chaque méthode possède ses propres groupes.