

Cours INF8100

Fait par

Bouaoune Youghourta BOUY28059509

Présenté à

**Mrabah Nairouz**

Date de remise : 22 octobr 2022

5.2 Exploration des données :

1)Quel est le nombre de notes données par tous les utilisateurs ?

réponce 1:

100836

2)Quel est le nombre d’utilisateurs ?

réponce 2:

610

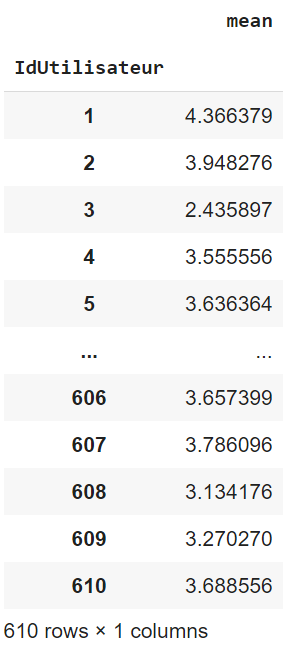
2)Quel est le nombre de films ?

réponce 2.2:

9742

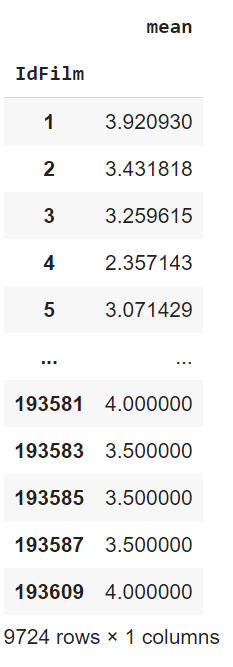
3)Quel est le nombre moyen de notes par utilisateur ?

réponce 3:



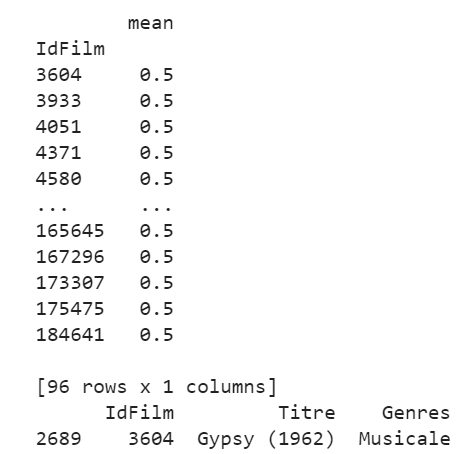
4)Quel est le nombre moyen de notes par film ?

réponce 4:



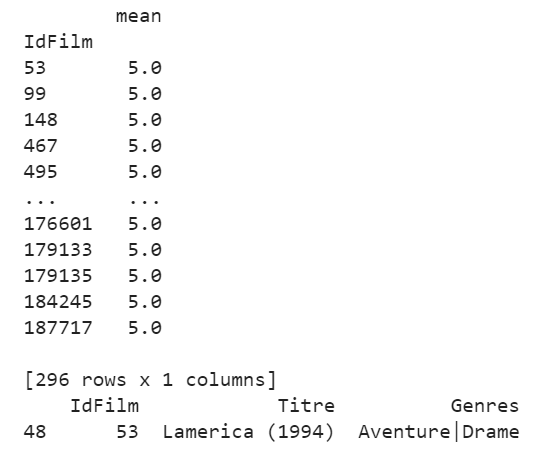
5)Quel film a la note moyenne la plus basse?

réponce 5:



6) Quel film a la note moyenne la plus élevée?

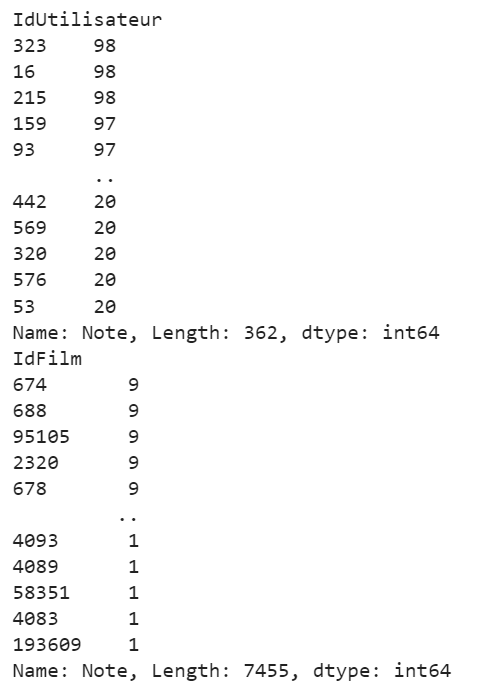
réponce 6:



5.3 Transformation des données :

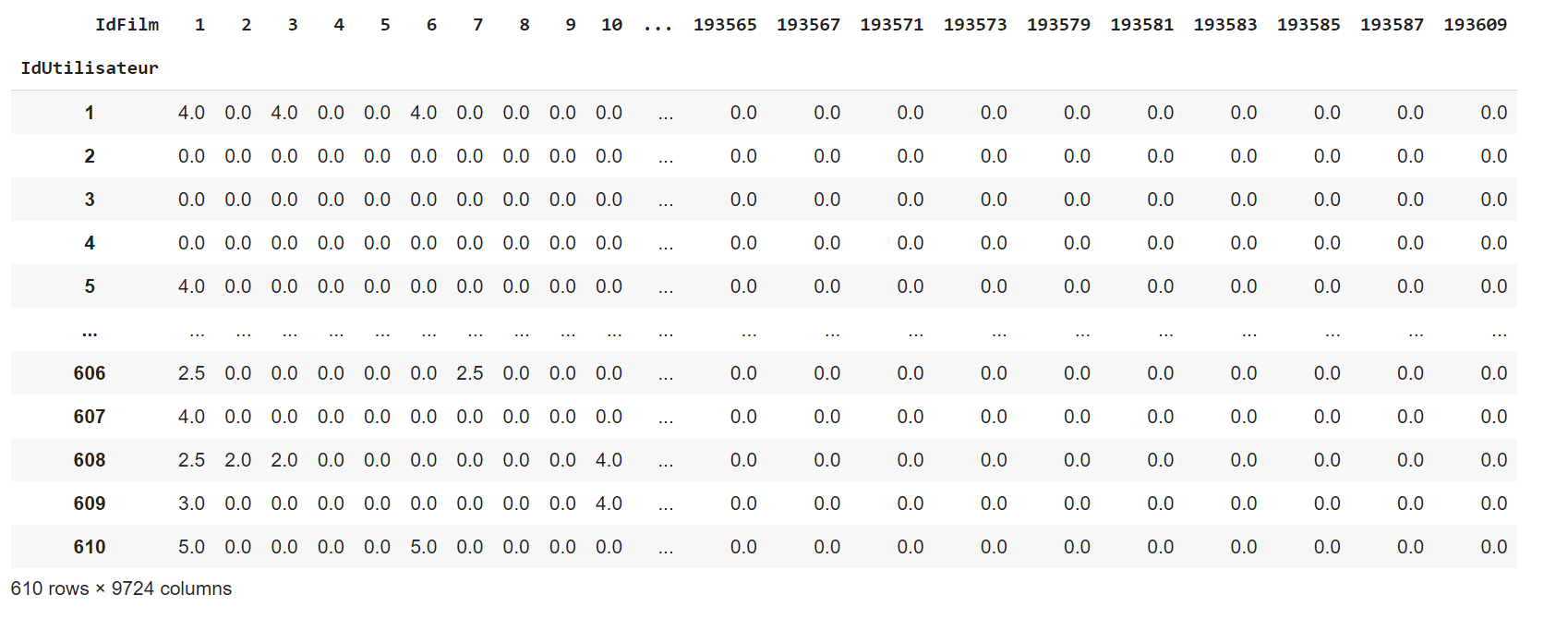
1)Filtrez les utilisateurs qui ont votés moins de 100 films et les films qui sont votées par moins de 10 utilisateurs? Remarque: vous pouvez utiliser pandas.DataFrame.groupby(). Dans la suite, on utilisera uniquement la matrice filtrée:

réponce 1:



2)Construisez la matrice utilisateur-film? Remarque: vous pouvez utiliser la méthode pandas.DataFrame.pivot():

réponce 2:



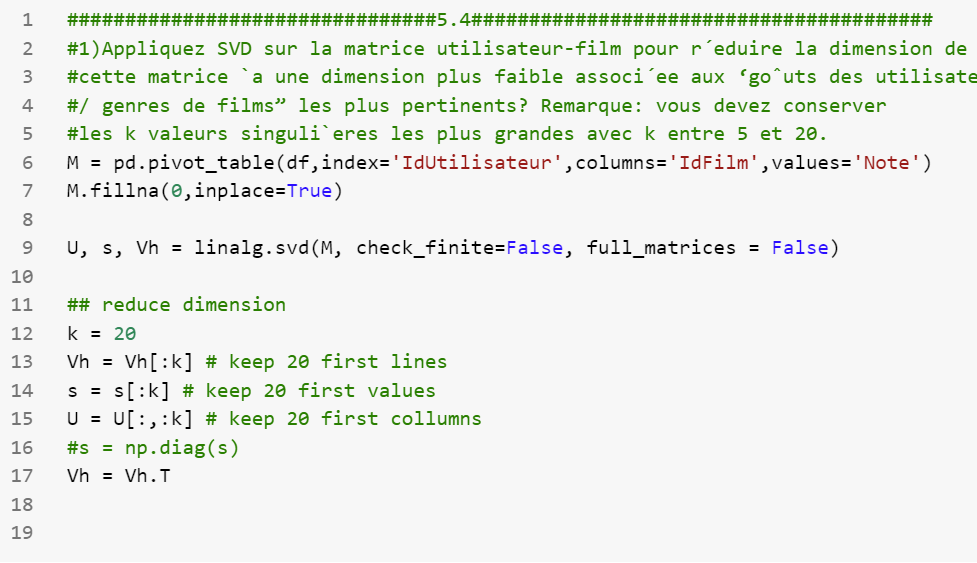
3)Calculez le taux de parcimonie de la matrice utilisateur-film en utilisant la formul s = A/B

réponce 3:

0.016999683055613623

5.2 Filtrage collaboratif :

1)Appliquez SVD sur la matrice utilisateur-film pour réduire la dimension de cette matrice `a une dimension plus faible associée aux 'goûts des utilisateurs genres de films” les plus pertinents? Remarque: vous devez conserver les k valeurs singuli`eres les plus grandes avec k entre 5 et 20.



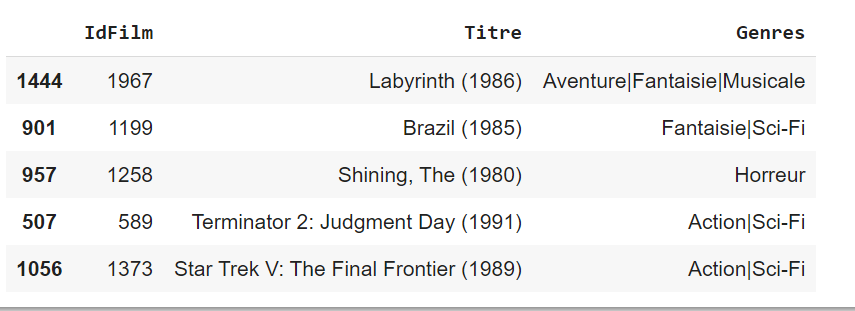
2)Cherchez les 5 films les plus similaires au premier film ? Remarque: vous devez utiliser la fonction scipy.spatial.distance.cosine sur les lignes de la matrice V (après la réduction de la dimension).

réponce 2:



3)Affichez les films recommandés pour le premier utilisateur? Remarque:vous devez afficher l’identifiant de chaque film recommandé, son titre et son genre. Il ne faut pas inclure les films que l’utilisateur a déjà regardés.

réponce 3:



6 Partie 2:

6.2 Exploration des données:

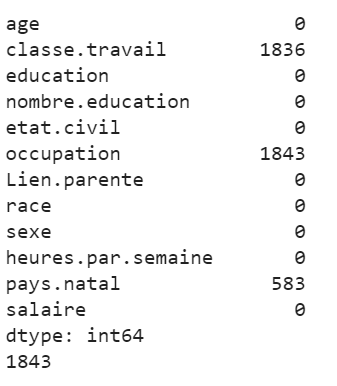
1) Quel est la taille du jeu de donn´ees ?

réponce 1:

390732

2)Combien de valeurs manquantes y a-t-il pour chaque colonne ? Quelle est la colonne (feature) qui a le plus de valeurs manquantes ?

réponce 2:



la colonne (feature) qui a le plus de valeurs manquantes :

occupation 1843

3)Combien d’hommes y a-t-il dans ce jeu de données ?

réponce 3:

Le nombre d’homme dans le jeu de données :21790

4)Quel est l'âge le plus élevé et le plus bas de ce jeu de données ?

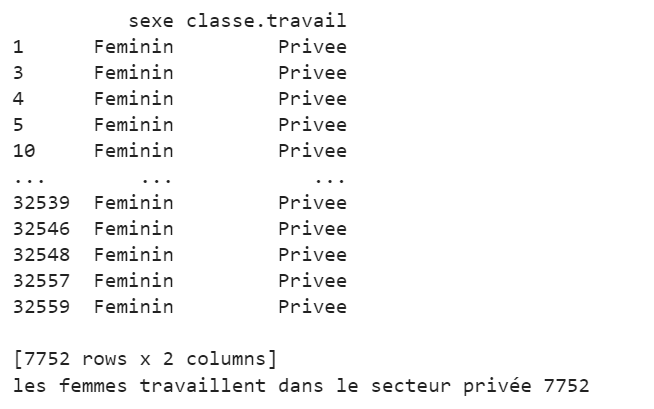
réponce 4:

age max 90

age min 17

5)Combien de femmes travaillent dans le secteur privé?

réponce 5:



6)Quelle est la liste des différents pays représentés dans la colonne représentant le pays natal/d’origine de chaque personne de notre jeu de donn´ees?

réponce 6:

La liste des pays:

0 United-States

9 ?

26 Mexico

27 Greece

47 Vietnam

63 China

82 Taiwan

90 India

95 Philippines

156 Trinadad&Tobago

157 Canada

166 South

211 Holand-Netherlands

216 Puerto-Rico

268 Poland

284 Iran

287 England

412 Germany

447 Italy

487 Japan

508 Hong

535 Honduras

726 Cuba

781 Ireland

921 Cambodia

938 Peru

954 Nicaragua

1004 Dominican-Republic

1044 Haiti

1117 El-Salvador

1180 Hungary

1258 Columbia

1268 Guatemala

1765 Jamaica

2299 Ecuador

2368 France

2985 Yugoslavia

3079 Scotland

3267 Portugal

3804 Laos

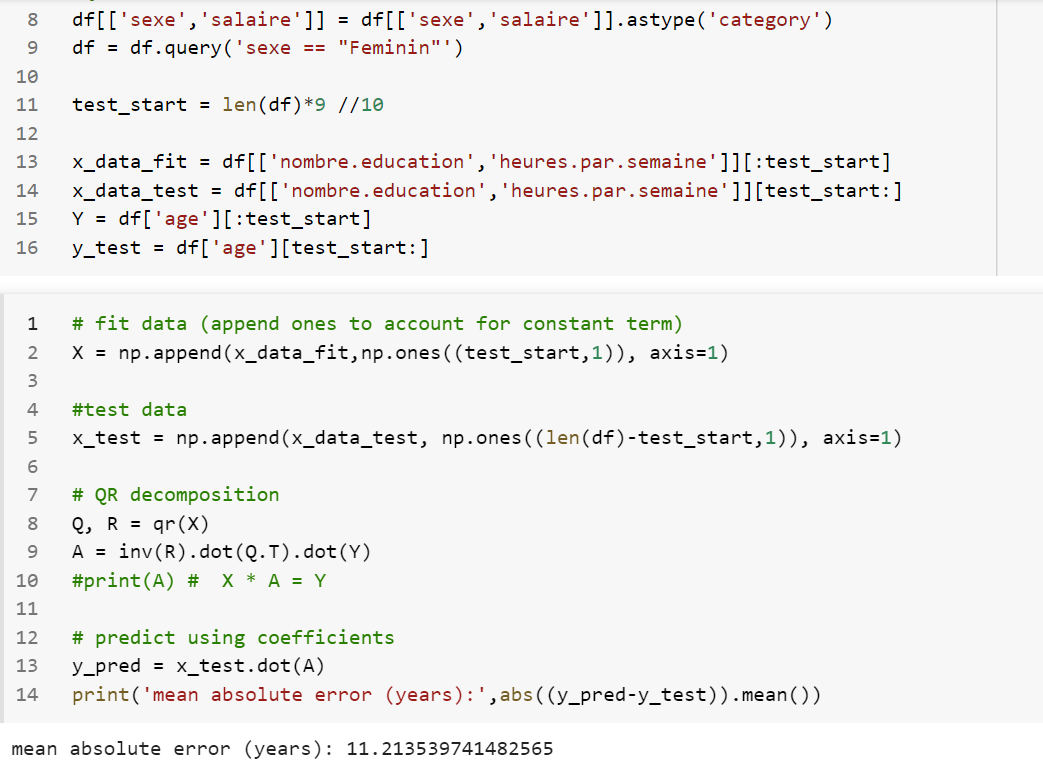
4460 Thailand

5581 Outlying-US(Guam-USVI-etc)

6.3 Algebre lineaire:

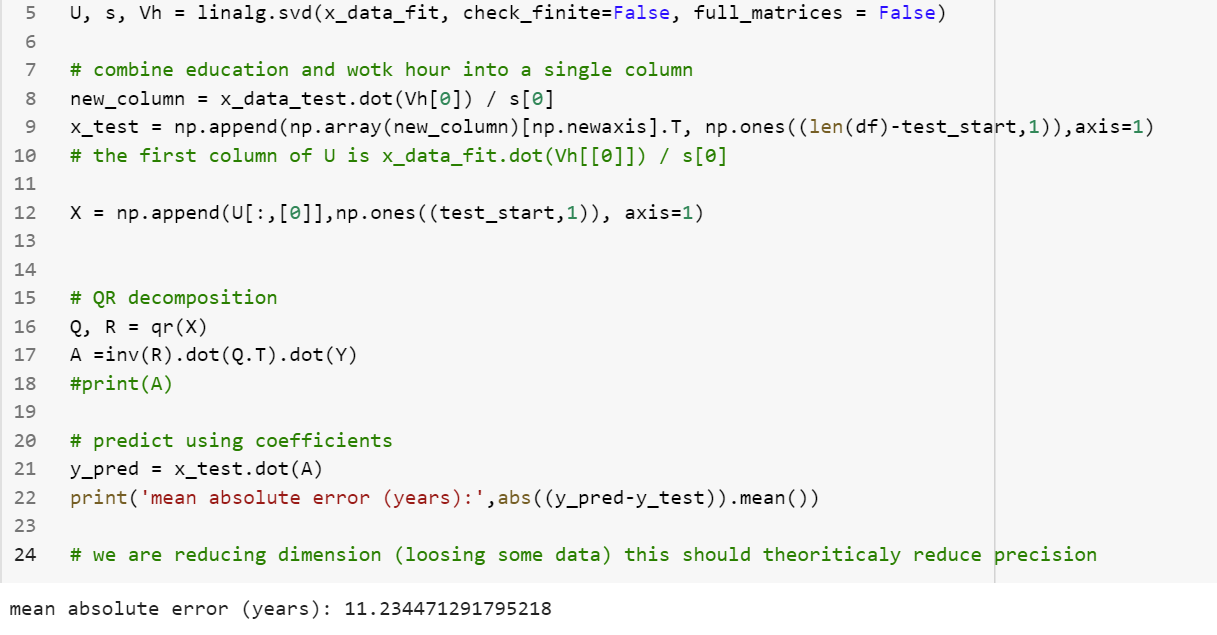
1. Peut-on pr´edire l’ˆage d’une personne (femme) en fonction de son niveau d’´education et son nombre d’heures travaill´ees par semaine ? Vous devez passer par la technique de résolution d' équations linéaires par la factorisation QR. Vous devez tester votre mod`ele sur les données de tests

réponce 1:

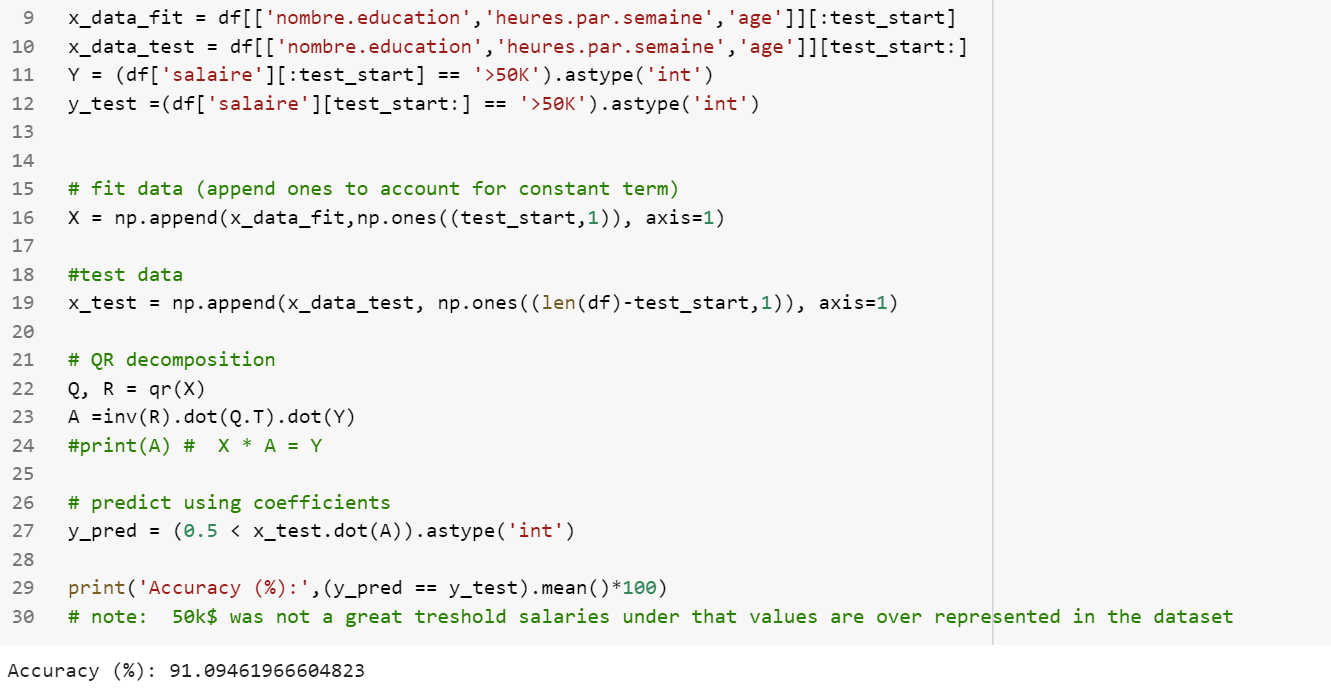


2)Si le niveau d'éducation et le nombre d’heures travaillées étaient réduits à 1 seule dimension qui capture les informations importantes dès 2 (SVD) ? Est-ce que la prédiction serait améliorée ou pas ? Vous devez tester votre modèle sur les données de tests.

réponce 2:



3)Peut-on pr´edire si le salaire d’une femme sera ≤ 50K en fonction de son niveau d'éducation, son ˆage et son nombre d’heures travaillées par semaine Vous devez passer par la technique de r´esolution d’´equations lin´eaires.Remarque : puisque le Y `a pr´edire est binaire alors ce n’est normalement pas un problèeme de r´egression linèaire. Vous devez donc adapter votre solution en conséquence. Par exemple, si la prédiction est inférieure à 0,5 alors le salaire est ≤ 50k sinon c’est > 50k. Vous devez tester votre mod`ele sur les données de tests



6.4 Analyses statistiques et probabilistes:

1)Quelle est la moyenne des âges ? Quelle est la moyenne du nombre d’heures travaillées par semaine ? et les valeurs médianes

réponce 1:

Moyenne d'age= 38.58164675532078

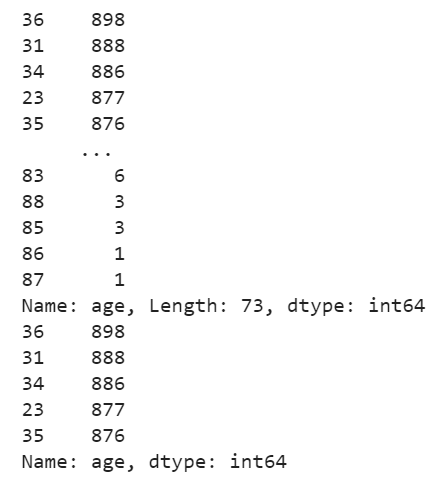
Median d’ age= 37.0

Moyenne d’ heuer= 40.437455852092995

median d’ heur= 40.0

2)Quels sont les âges (prendre les 5 premiers) les plus représentatifs de ce jeu de donn´ees ? Représentent-ils une tranche d'âge en particulier (ex:30-35 ans) ?

réponce 2:



les ˆages (prendre les 5 premiers) les plus représentatifs de ce

jeu de données [23-36]

3) Comment est la variance (à quel point les données sont éparpillées/dispersées)des heures travaillées par semaine ? Et les âges ?

réponce 3:

la variance des heures.par.semaine = 152.45899505045415

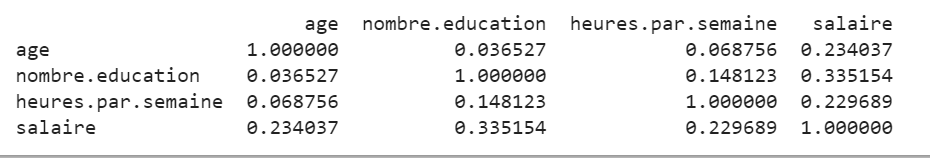
Ecat\_type\_heur = 12.347428681731843

la variance des âges = 186.06140024880162

ecat\_type\_age = 13.640432553581343

4)Est-ce que le niveau d’éducation, comparé à l’âge influe le plus sur la valeur du salaire ? En d’autres termes, est-ce que le niveau d’éducation a plus de poids sur le salaire que l’âge ??

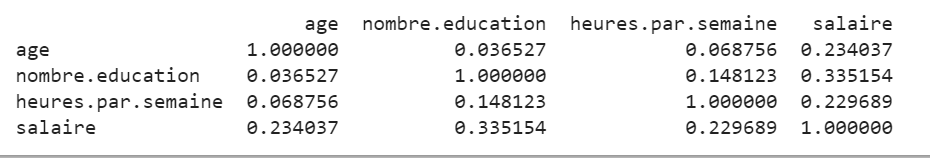
Réponse 4:



Oui parce que la corrélation entre le niveau d’éducation et le salaire (0.335154) est supérieure à la corrélation entre l’âge et le salaire (0.234037)

5)Est ce que le niveau d'éducation, comparée au nombre d’heures travaillées par semaine influe le plus sur la valeur du salaire ? En d’autres termes est que le niveau d'éducation a plus de poids sur le salaire que le nombre d’heures travaillées par semaine

Réponse 5:



Oui parce que la corrélation entre le niveau d’éducation et le salaire (0.335154) est supérieure à la corrélation entre l nombre d’heures travaillées par semaine et le salaire (0.229689)

6)A quel point l'âge est-il un facteur qui influe le salaire? Pour répondre à cette question vous devez vérifier que l'âge moyen de ceux qui gagnent moins de 50k est plus bas que ceux qui gagnent plus de 50k.

Réponse 6:

le nombre d age moyenne pour les salaries <=50k est : 36.78373786407767

le nombre d age moyenne pour les salaries >50k est : 44.24984058155847

le nombre d'âge moyenne pour les salariés <=50k < le nombre d age moyenne pour les salariés >50k == True

7)`A quel point le nombre d’heures travaillées par semaine est un facteur qui influe le salaire? Pour répondre à cette question vous devez vérifier que le nombre moyen d'heures travaillées par semaine pour ceux qui gagnent moins de 50k est plus bas que ceux qui gagnent plus de 50k.

Réponse 7:

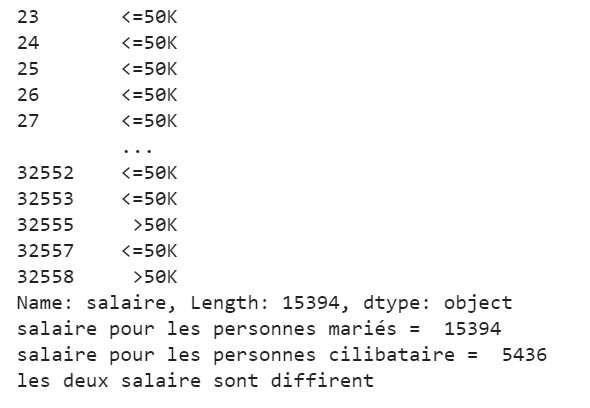
le heures.par.semaine moyenne pour les salariés <=50k est : 38.840210355987054

le heures.par.semaine moyenne pour les salariés >50k est : 45.473026399693914

le heures.par.semaine moyenne pour les salariés <=50k < le heures.par.semaine moyenne pour les salariés >50k == True

8) Est ce qu’il y a une différence de salaire entre les personnes mariées et les personnes célibataires ? Quel groupe gagne le mieux leurs vies ?

Réponse 8:



Les personnes marié gange le mieux leurs vies

9)Quelle est la proportion des hommes qui ont un doctorat et gagnent plus de 50k ? et chez les femmes ?

Réponse 9:

la taille de la population des hommes qui ont un doctorat et gagnent plus de 50k = 256

la taille totale de df : 21790

le pourcentage des hommes ont NV doctorat et un salaire de >50K : % 1.174850849013309

la taille de la population des femmes qui ont un doctorat et gagnent plus de 50k = 50

la taille totale de df : 10771

le pourcentage des femme ont NV doctorat et un salaire de >50K : % 0.46420945130442853

10)Quelle est la proportion des femmes mariées qui n’ont pas fait de masters ni de doctorat ? et chez les hommes ?

Réponse 10:

la proportion des femmes mariées qui n’ont pas fait de masters ni de doctorat % 15.940952557794077

la proportion des hommes mariées qui n’ont pas fait de masters ni de doctorat % 56.74162459843965

11) Les ressortissants de quel pays ont les salaires les plus élevés uniquement dans notre échantillon?

Réponse 11:

pays ont les salaires les plus ´elev´es uniquement dans notre échantillon:

pays.natal

United-States 7171

? 146

Philippines 61

Germany 44

India 40

Canada 39

Mexico 33

England 30

Italy 25

Cuba 25

Japan 24

China 20

Taiwan 20

Iran 18

South 16

Poland 12

France 12

Puerto-Rico 12

Jamaica 10

El-Salvador 9

Greece 8

Cambodia 7

Yugoslavia 6

Hong 6

Ireland 5

Vietnam 5

Ecuador 4

Haiti 4

Portugal 4

Scotland 3

Thailand 3

Hungary 3

Guatemala 3

Laos 2

Dominican-Republic 2

Peru 2

Trinadad&Tobago 2

Columbia 2

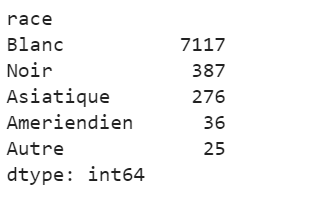
Nicaragua 2

Honduras 1

12):Quelle race a les salaires les plus élevés uniquement dans notre ́échantillon?

Réponse 12:

les salaires les plus élevés uniquement dans notre échantillon:



6.5 Tests d'hypothese :

1)Est-ce qu’on peut dire que la moyenne d’heures travaillées par semaine est autour de 40h/semaine ? Est-ce que l’affirmer serait significatif ?

Réponse 1:

p-values 1.6483286793366825e-10

rejeter l' hypothèse nulle

Alore le test est significatif

2)Est ce que les hommes ont en moyenne un salaire supérieur à 50k comparé aux femmes ? Est-ce significatif ?

Réponse 2:

homme std = 0.46071870187301667

femme\_std = 0.31221622020176376

p-value 0.0

rejeter l' hypothèse nulle

Alore le test est significatif

3)Est ce qu’il y a un lien quelconque entre l'âge et le nombre d’heures travaillées par semaine ? Est-ce significatif ?

Réponse 3:

la moyenne de la df\_heur1 = 40.933173637982925

la moyenne de la df\_heur2 = 39.94170761670762

p-value 4.24686791127587e-13

rejeter l' hypothèse nulle

Alore le test est significatif