* **Membre de groupe:**

1. **Madani Mohamed El-Habib – G4**
2. **Herrouel Nor El-Yacine – G4**
3. **Benkahla Mohamed – G2**
4. **Hagani Wael – G4**

**Projet FDD**

**Dataset Adult**

1. **Introduction**

Ce projet vise à trouver le meilleur modèle pour prédire si le revenu d'un individu est supérieur à 50 000 Sur la base des données fournies par le référentiel d'apprentissage automatique de l'UCI appelé "Adult Dataset" par Étudiez différentes techniques de reconnaissance de formes. Les variables de prédiction incluent des informations telles que Âge, éducation, profession, sexe, pays d'origine et autres informations pertinentes sur l'individu. L'approche impliquait le prétraitement des données où les données manquantes, la représentation des caractéristiques catégorielles, la normalisation et le déséquilibre des données ont été traités. En outre, différentes techniques de reconnaissance de formes à la fois.

Des techniques de classification statistique et sans distribution ont été mises en œuvre.

De plus, le choix de divers thèmes et techniques de réduction dimensionnelle comme les principaux L'analyse des composants, l'annulation de caractéristiques récursives, etc. sont utilisées pour extraire les caractéristiques les plus représentatives

Pour améliorer les performances et aussi pour réduire la complexité de calcul.

Étant donné que l'ensemble de données était déséquilibré, des mesures de performances telles que la zone sous le récepteur La courbe caractéristique et le score F1 de la métrique groupée ont été utilisés pour sélectionner le meilleur modèle.

1. **Prétraitement**

Afin de rendre les données brutes données adaptées à la classification, un prétraitement a été effectué. La les fonctionnalités appartenaient à l'une des catégories suivantes :

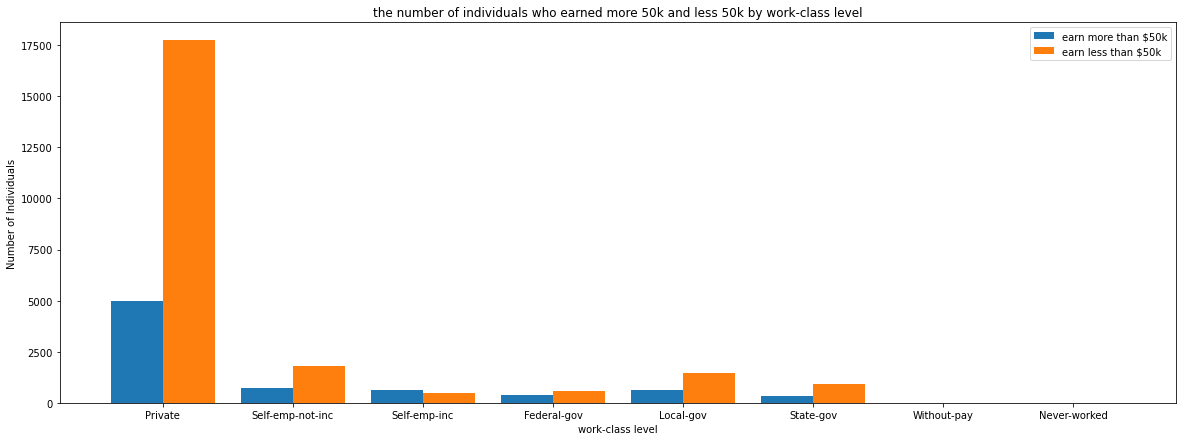
* **Caractéristiques numériques (6 au total) :** Les informations fournies par ces fonctionnalités sont au format numérique. Les caractéristiques qui appartiennent à cette catégorie sont l'âge, le fnlwgt, le numéro d'éducation, le gain en capital, la perte en capital et heures par semaine.
* **Caractéristiques catégorielles (8 au total) :** Les informations fournies par ces fonctionnalités sont catégorielles format. Les caractéristiques qui appartiennent à cette catégorie sont la classe de travail, l'éducation, l'état matrimonial, profession, relation, race, sexe et pays d'origine.

| **Attribut** | **Description** | **Type** |
| --- | --- | --- |
| age | représentant l'âge de et l'adulte. | Nombres |
| workclass | représentant le type de travail de l'adulte. | Catégorielle (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked) |
| fnlwgt |  | Nombres |
| education | représentant le niveau d'éducation qu'un adulte avait | Catégorielle (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool) |
| eductation-num | représentant le numéro du niveau d'éducation qu'un adulte avait | Nombres |
| marital | représentant les options distinctes qui décrivent la relation d'une personne avec un autre significatif | Catégorielle (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse) |
| occupation | représentant l'activité/le travail dans lequel s'exerce un adulte | Catégorielle (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces) |
| relationship | représentant les relations qu'un adulte pourrait avoir avec une personne | Catégorielle (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried) |
| race | représentant la race d'un adulte | Catégorielle (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black) |
| sex | représentant le sexe d'un adulte | Catégorielle (Male,Female) |
| capital-gain | représentant le concept économique défini comme le bénéfice réalisé sur la vente d'un bien qui a pris de la valeur sur la durée de détention pour un adulte | Nombres |
| capital-loss | représentant la différence entre un prix de vente inférieur et un prix d'achat ou un prix de revient supérieur d'une immobilisation admissible | Nombres |
| hours-per-week | représentant les heures qu'une personne travaille dans une semaine | Nombres |
| country | représentant le pays d'origine d'un adulte | Catégorielle (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands) |
| income | représentant le revenu d'un adulte | Catégorielle (<=50k , >50k) |

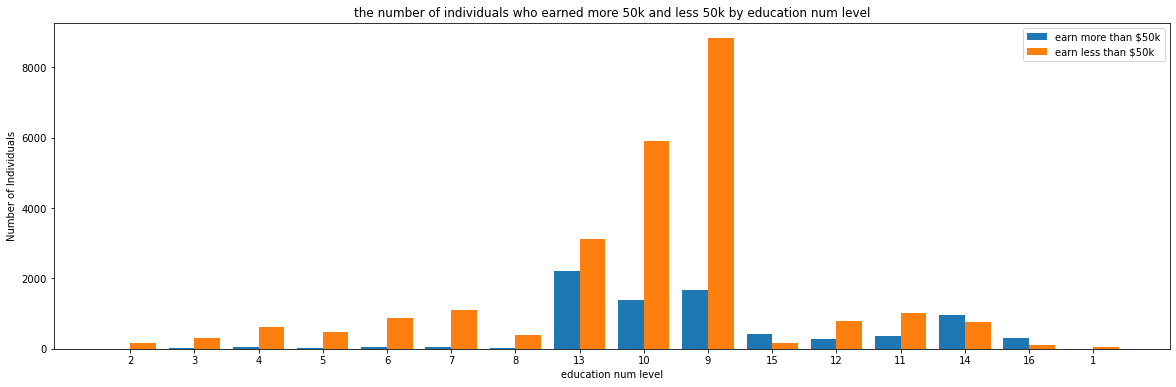
1. **Gestion des valeurs manquantes :**

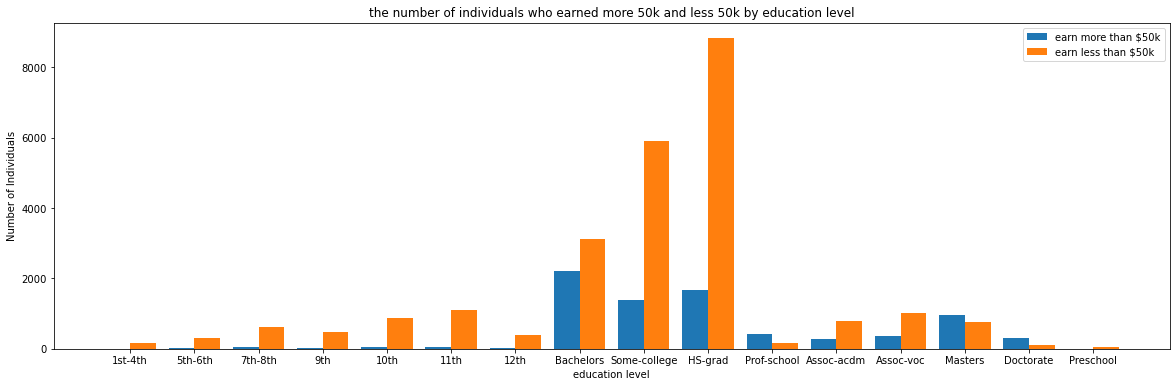
Certaines des caractéristiques (seulement catégoriques dans ce cas) pour certains points de données contiennent des valeurs inconnues. Il faut remplacer ces valeurs inconnues par la valeur qui convient le mieux correspond au contexte. La description des différentes techniques utilisées pour traiter les données manquantes est la en dessous de:

* Supprimer les lignes contenant des valeurs manquantes : C'est en quelque sorte la méthode la plus simple et la meilleure pour traiter les données manquantes car l'ensemble de la classification dépend de l'ensemble de données réel, c'est-à-dire sans aucun hypothèses faites sur les valeurs manquantes et ainsi les résultats seront une représentation fidèle des données recueillies des données d'entrée. Cependant, l'adoption de cette méthode peut également entraîner une perte d'informations dans de nombreux cas. Dans notre cas, il y avait un total de 4262 valeurs manquantes et il nous restait encore une quantité considérable de données (30161 sur 32560) après suppression des valeurs manquantes.
* Imputation en mode indépendant de la classe : Étant donné que toutes les valeurs manquantes dans notre ensemble de données appartiennent à la caractéristiques catégorielles, le mode est le meilleur substitut statistique. D'autres mesures comme la moyenne ou d'autres moyens de l'interpolation des données peut ne produire des valeurs correspondant à aucune des catégories. Dès lors, toute la valeur manquante dans une caractéristique particulière a été remplacée par la catégorie la plus fréquente de cette fonctionnalité. Cela a du sens intuitivement car la valeur manquante le long d'une caractéristique particulière est plus susceptible de contenir la catégorie qui se produit le plus. Par exemple, nous remplaçons les valeurs manquantes dans la fonction "WorkClass" par "Private" comme on peut le voir sur le tracé de l'histogramme pour la classe ouvrière dessous. Cependant, il y a un certain degré d'incertitude qui y est attaché (en raison des hypothèses formulées) et cela pourrait avoir un impact positif ou négatif sur le résultat final du classement. Cette méthode a été utilisée pour les trois caractéristiques avec des données manquantes : classe de travail, profession et pays d'origine.

1. **Traitement des caractéristiques catégorielles :** En analysant les données brutes fournies, les observations suivantes ont été fabriquées.

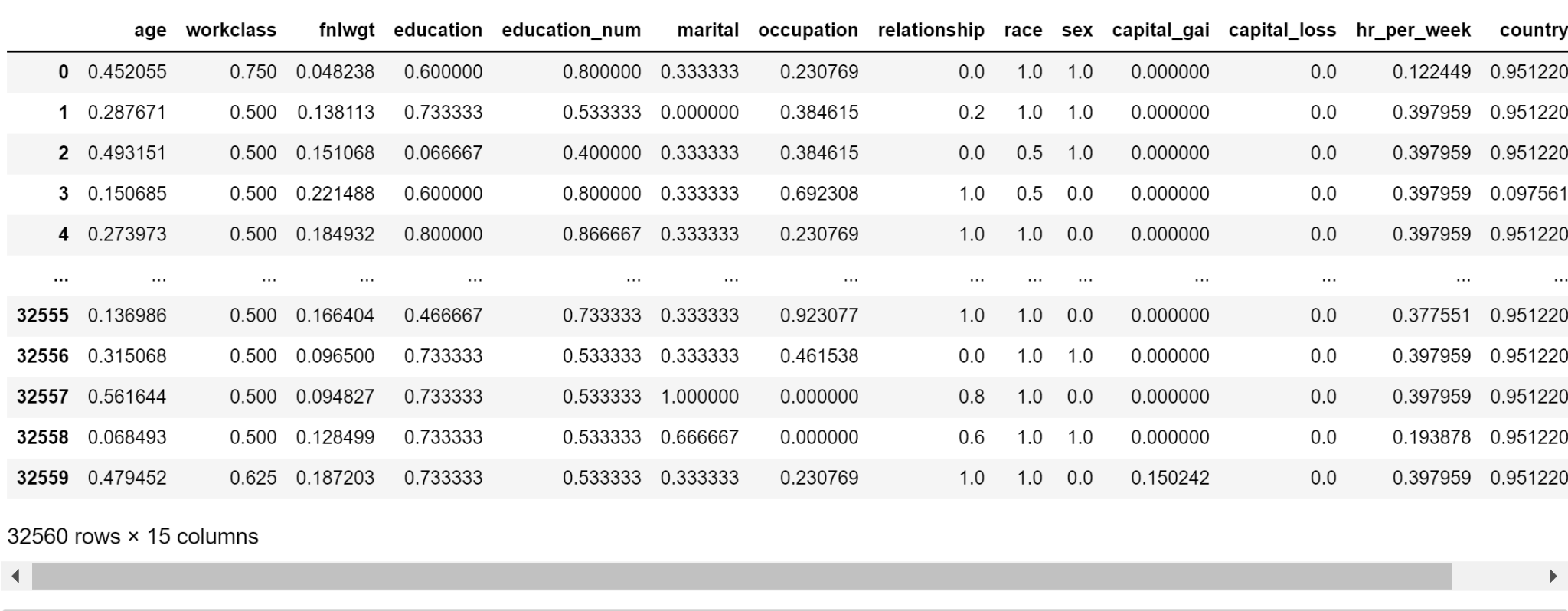
* On observe également que les mêmes informations concernant l'éducation sont fournies dans deux caractéristiques - l'une en catégorie (éducation) et l'autre en numérique (numéro d'éducation). Depuis dans le tâche de prédiction du revenu, selon notre logique, le niveau d'éducation joue un rôle essentiel.

****



* D'après les graphiques ci-dessus, il est entendu que le chiffre de l'éducation et de l'éducation fournit en fait les mêmes informations. Étant donné que le numéro d'éducation est une caractéristique numérique et a une manière logique d'organiser et leur classement tout nombre plus élevé indique une qualification scolaire supérieure, nous gardons cela et ignorons l'équivalent catégorique de l'éducation. Les autres caractéristiques catégorielles sont restées les mêmes. En outre, toutes les caractéristiques catégorielles ont été encodées en valeurs numériques à l'aide d'un seul chiffrement à chaud. Parmi les trois caractéristiques, il y a valeurs manquantes, l'occupation et la classe ouvrière semblent avoir le sens de l'ordre par rapport à la variable cible (revenu). Cependant, le codage de nommage n'a pas été utilisé en raison d'un manque de

Informations sur la catégorie à évaluer plus haut que l'autre. Par conséquent, l'encodage d'un hot Il a été utilisé pour traiter toutes les classes au même niveau.

1. **Normalisation :** les caractéristiques numériques des données obtenues lors des étapes précédentes n'ont pas besoin d'être la même échelle. Par conséquent, toutes les caractéristiques numériques sont mises à l'échelle à une moyenne nulle et un écart type avant un traitement ultérieur. La fonction Standard Scalar() offerte par la bibliothèque scikit learn a été utilisée pour y parvenir.

## **Fractionner les données :** Maintenant, lorsque toutes les variables catégorielles sont transformées et que toutes les caractéristiques numériques sont normalisées, nous devons diviser nos données en ensembles d'apprentissage et de test. Nous partageons 80 % pour la formation et 20 % pour les tests.

# **Exploration :**

* Le nombre de personnes qui gagnent plus de 50 000 $ et leur pourcentage : Obtenez les adultes qui gagnent plus de 50 000 $ en utilisant l' attribut de revenu et précisez-le à > 50k et obtenez la longueur du tableau pour obtenir le nombre

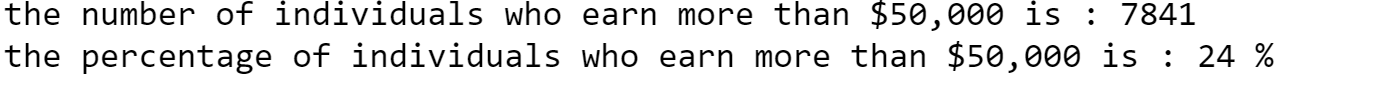
**objet = jeu de données[condition]**

en divisant le nombre d'adultes qui gagnent plus de 50 000 avec le nombre

total d'adultes dans l'ensemble de données, après cela, nous le multiplions

par 100 pour obtenir le pourcentage

**objet = (>50k-adultes/total-adultes)\*100**

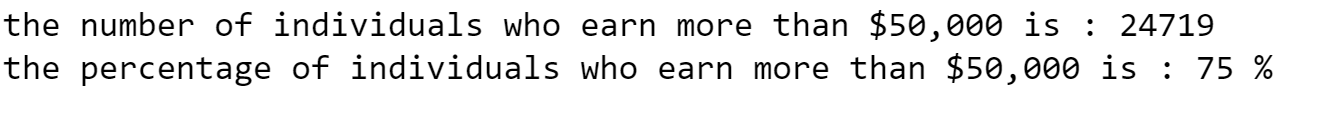
****

* Le nombre de personnes qui gagnent moins de 50 000 $ et leur pourcentage : Obtenez les adultes qui gagnent moins de 50 000 $ en utilisant l' attribut de revenu et précisez-le à <= 50k et obtenez la longueur du tableau pour obtenir le nombre

**objet = jeu de données[condition]**

en divisant le nombre d'adultes qui gagnent moins de 50k avec le nombre total d'adultes dans l'ensemble de données, après cela nous le multiplions par 100 pour obtenir le pourcentage

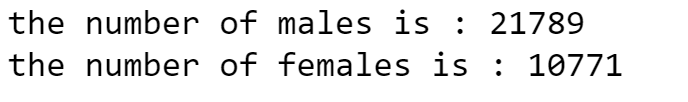
**objet = (<=50k-adultes/total-adultes)\*100**

****

* Le nombre de mâles et de femelles :

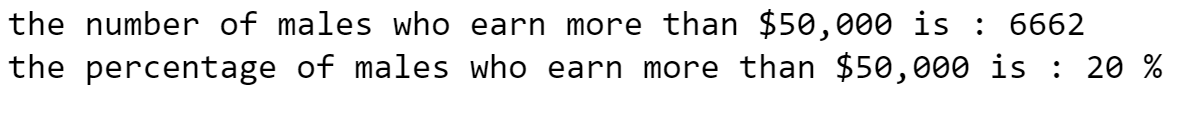
affichez le nombre d'hommes en utilisant l'attribut sex et précisez qu'il s'agit d'un homme et obtenez la longueur du tableau pour obtenir le nombre

affichez le nombre d'hommes en utilisant l'attribut sex et précisez qu'il s'agit d'une femme et obtenez la longueur du tableau pour obtenir le nombre

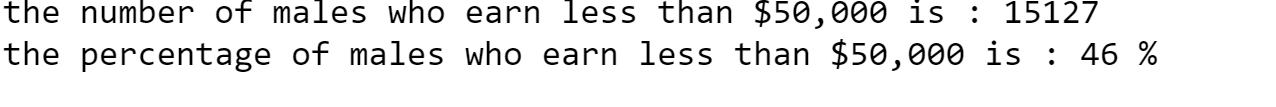


* Le nombre d'hommes qui gagnent plus de 50 000 $ et leur pourcentage : afficher le nombre d'hommes qui gagnent plus de 50 000 $ en utilisant l'attribut sexe et en précisant qu'il s'agit d'un homme et en utilisant le revenu

attribuez-le et précisez-le pour qu'il soit> 50k et obtenez la longueur du tableau pour obtenir le nombre

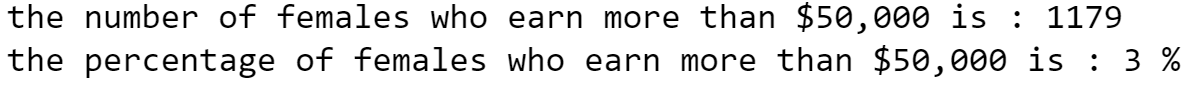


* Le nombre d'hommes qui gagnent moins de 50 000 $ et leur pourcentage : afficher le nombre d'hommes qui gagnent moins de 50 000 $ en utilisant l'attribut sexe et en précisant qu'il s'agit d'un homme et en utilisant l'attribut revenu et en précisant qu'il est <= 50 000 et en obtenant la longueur du tableau pour obtenir le nombre

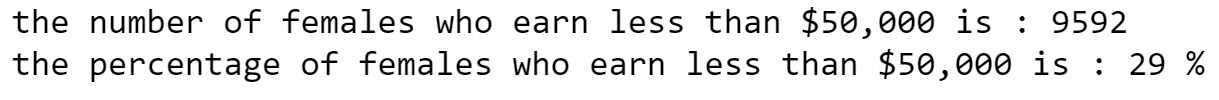


* Le nombre de femmes qui gagnent plus de 50 000 $ et leur pourcentage : afficher le nombre de femmes qui gagnent plus de 50 000 $ en utilisant l'attribut sexe et en précisant qu'il s'agit d'une femme et en utilisant le revenu

attribuez-le et précisez-le pour qu'il soit> 50k et obtenez la longueur du tableau pour obtenir le nombre

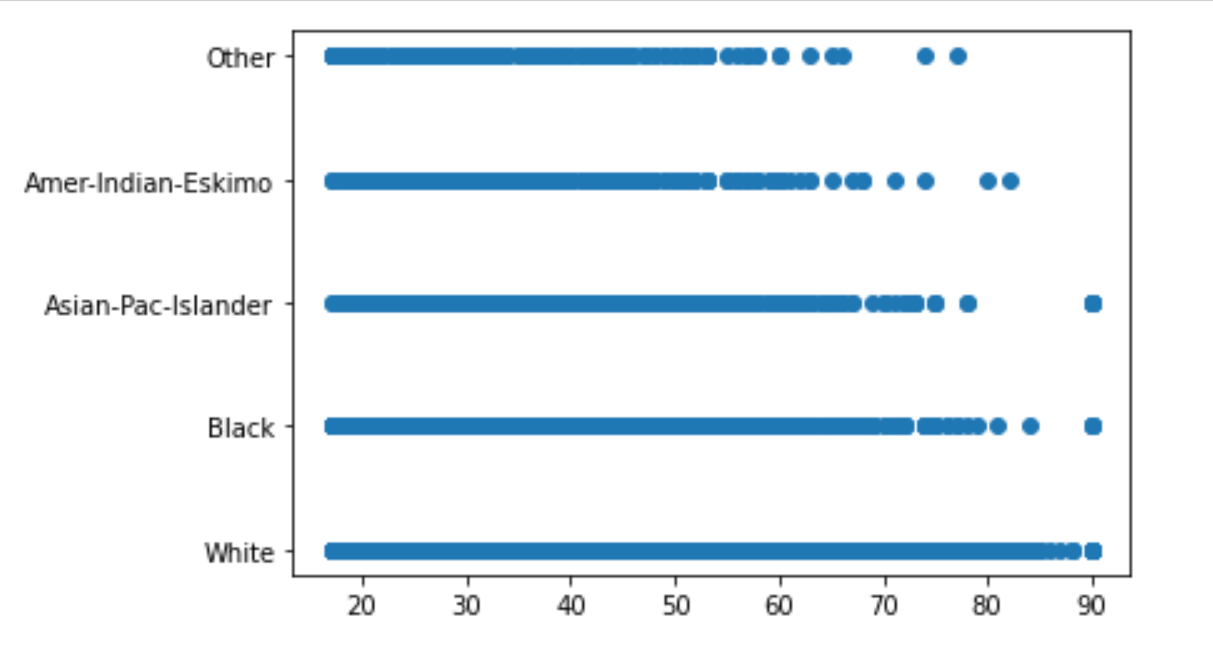


* Le nombre de femmes qui gagnent moins de 50 000 $ et leur pourcentage : afficher le nombre de femmes qui gagnent moins de 50 000 $ en utilisant l'attribut sexe et en précisant qu'il s'agit d'une femme et en utilisant l'attribut revenu et en précisant qu'il est <= 50 000 et en obtenant la longueur du tableau pour obtenir le nombre



* nuage de points montrant la répartition par âge des enregistrements :

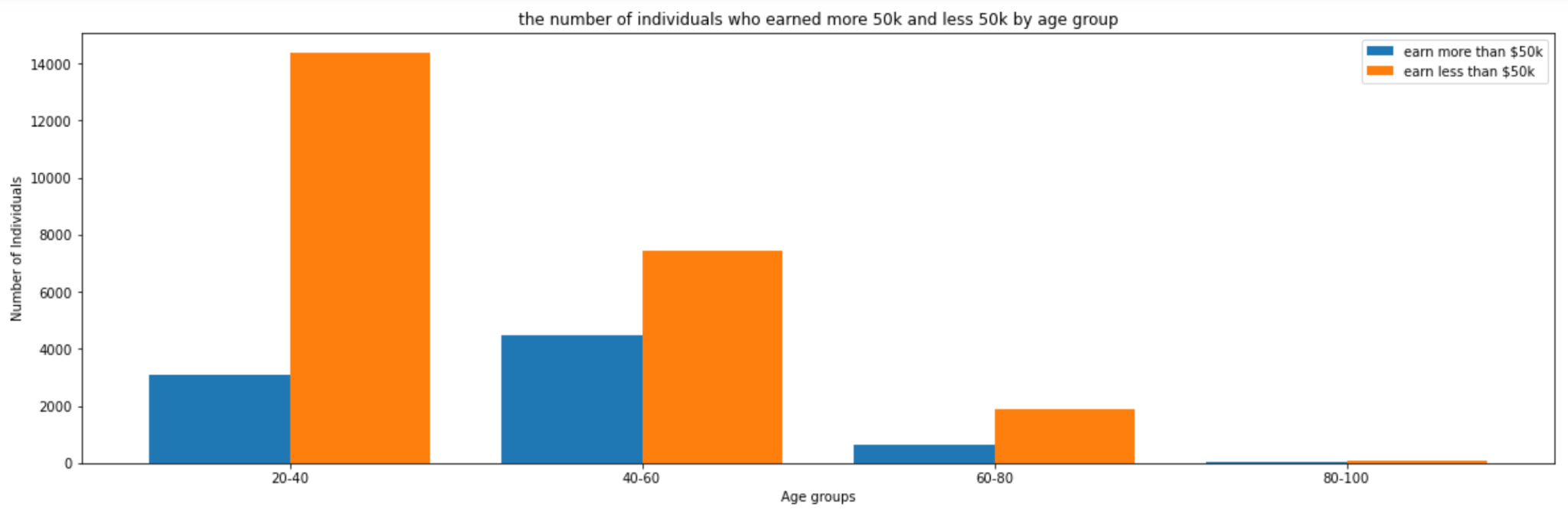
afficher le nuage de points de (âge et race) en utilisant la fonction de dispersion fournie par la bibliothèque matplotlib



* graphique de barres empilées montrant le nombre d'individus qui ont gagné plus/moins 50 000 $ par groupe d'âge :

afficher un graphique de barres implantées à l'aide de la fonction de barre calculée par la bibliothèque matplotlib, le nombre d'adultes qui gagnent plus de 50 000 $ en utilisant l'attribut de revenu et en le précisant à > 50 000 et en utilisant l'attribut d'âge pour obtenir les groupes d'âge et obtenir la longueur du tableau pour obtenir le numéro

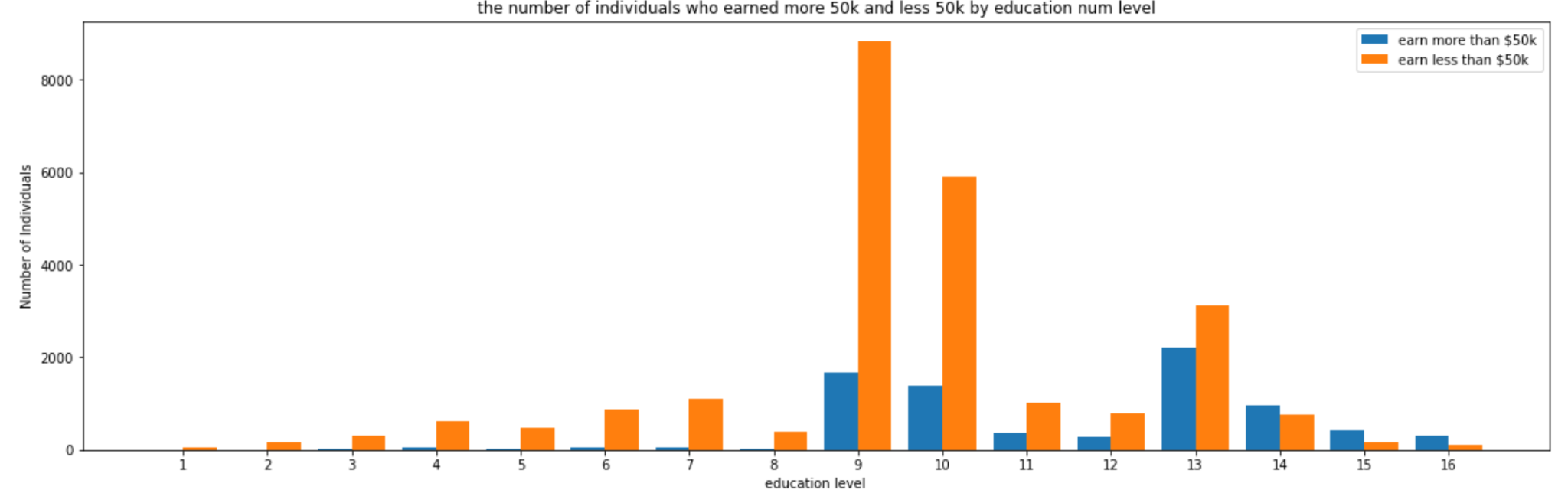
afficher un graphique de barres implantées à l'aide de la fonction de barre calculée par bibliothèque matplotlib, le nombre d'adultes qui gagnent plus de 50 000 $ en utilisant l'attribut de revenu et en le précisant <=50k et en utilisant l'attribut âge pour obtenir les groupes d'âge et en obtenant la longueur du tableau pour obtenir le nombre



* graphique de barres empilées montrant le nombre d'individus qui ont gagné plus/moins 50 000 $ par niveau d'éducation :

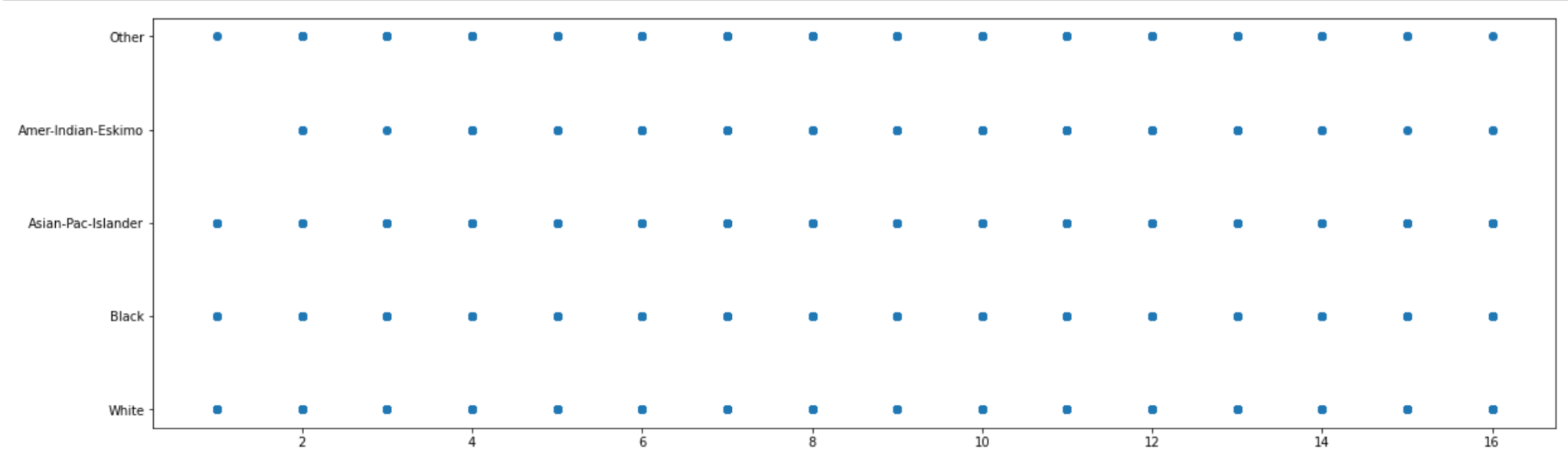
afficher un graphique de barres jalonnées à l'aide de la fonction de barre calculée par matplot library, le nombre d'adultes qui gagnent plus de 50 000 $ en utilisant l'attribut de revenu et en le précisant à > 50 000 et en utilisant l'attribut d'éducation pour préciser les groupes d'éducation et obtenir la longueur du tableau pour obtenir le numéro

afficher un graphique de barres empilées à l'aide de la fonction bar fournie par la bibliothèque matplotlib, le nombre d'adultes qui gagnent moins de 50 000 $ en utilisant l'attribut de revenu et le préciser <=50k et en utilisant l'attribut education pour préciser les groupes d'éducation et obtenir la longueur du tableau pour obtenir le nombre



* nuage de points montrant la distribution des niveaux d'éducation des enregistrements :

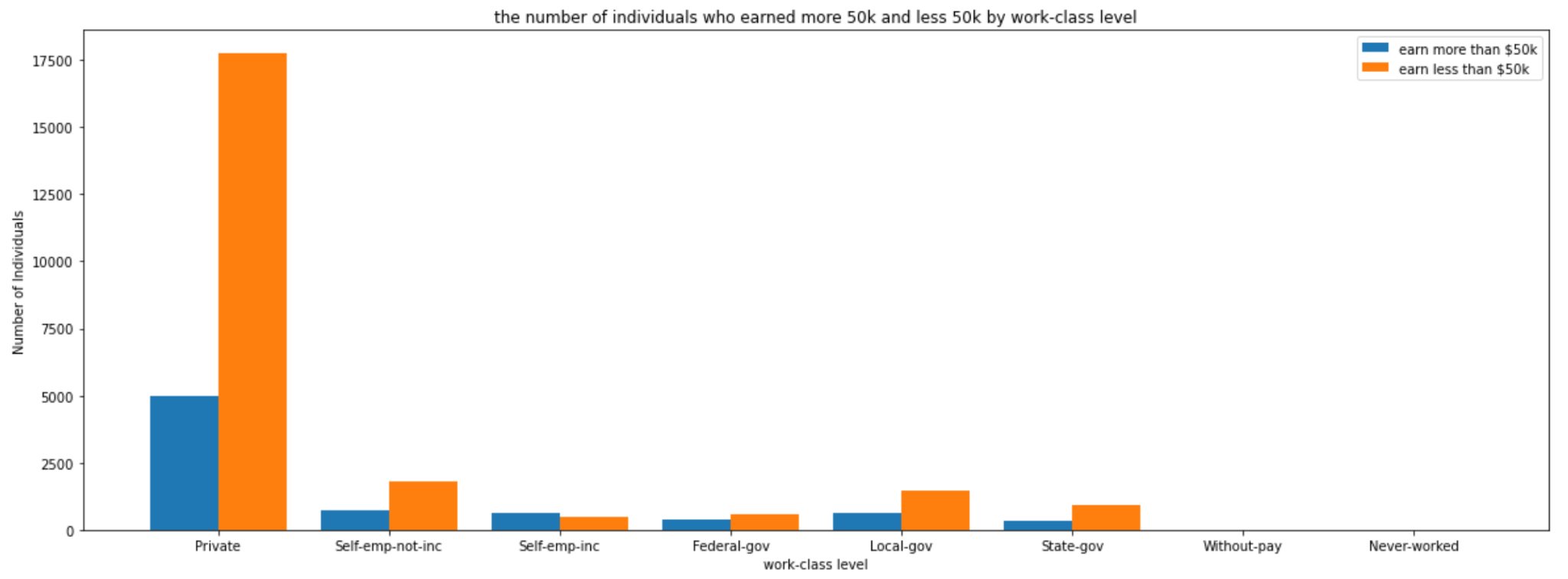
afficher le nuage de points de (éducation et race) en utilisant la fonction de dispersion fournie par la bibliothèque matplotlib



* graphique de barres empilées montrant le nombre de personnes qui ont gagné plus/moins 50 000 $ par catégorie de travail :

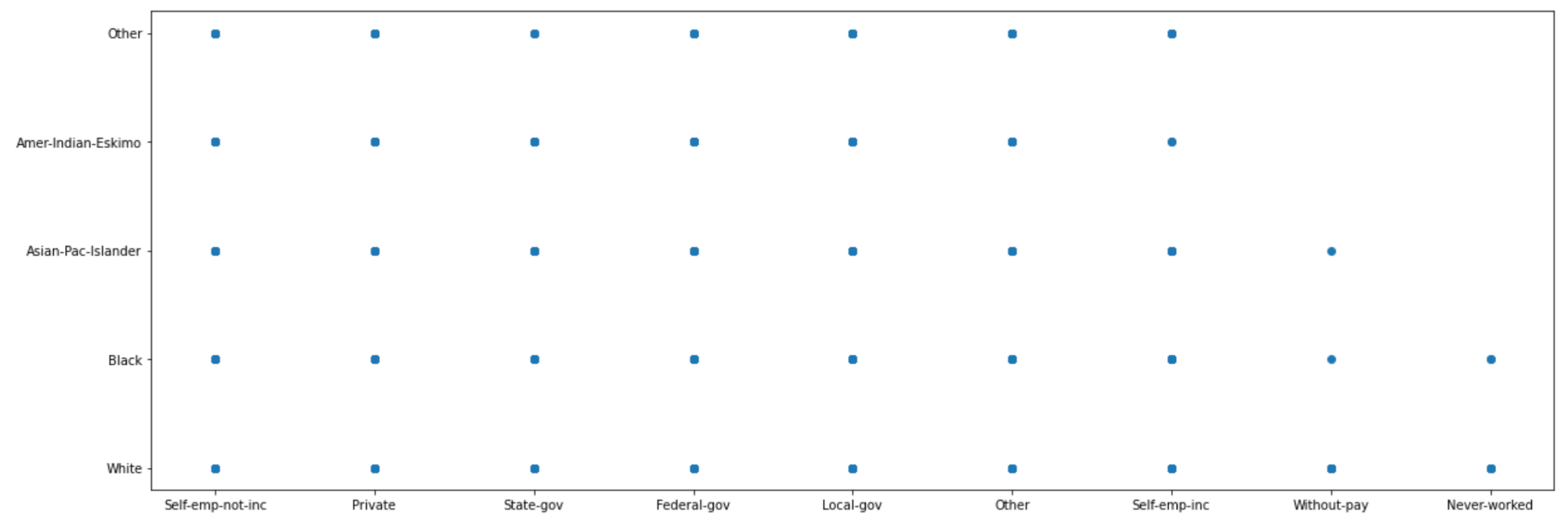
afficher un graphique de barres implantées à l'aide de la fonction de barre calculée par la bibliothèque matplotlib, le nombre d'adultes qui gagnent plus de 50 000 $ en utilisant l'attribut income et le préciser à > 50 000 et en utilisant l'attribut workclass pour préciser les groupes de classe de travail et obtenir la longueur du tableau pour obtenir le nombre

afficher un graphique de barres implantées à l'aide de la fonction de barre calculée par la bibliothèque matplotlib, le nombre d'adultes qui gagnent moins de 50 000 $ en utilisant l'attribut income et le préciser à <=50k et en utilisant l'attribut workclass pour préciser les groupes de classe de travail et obtenir la longueur du tableau pour obtenir le nombre



* nuage de points montrant la distribution des enregistrements par classe de travail:

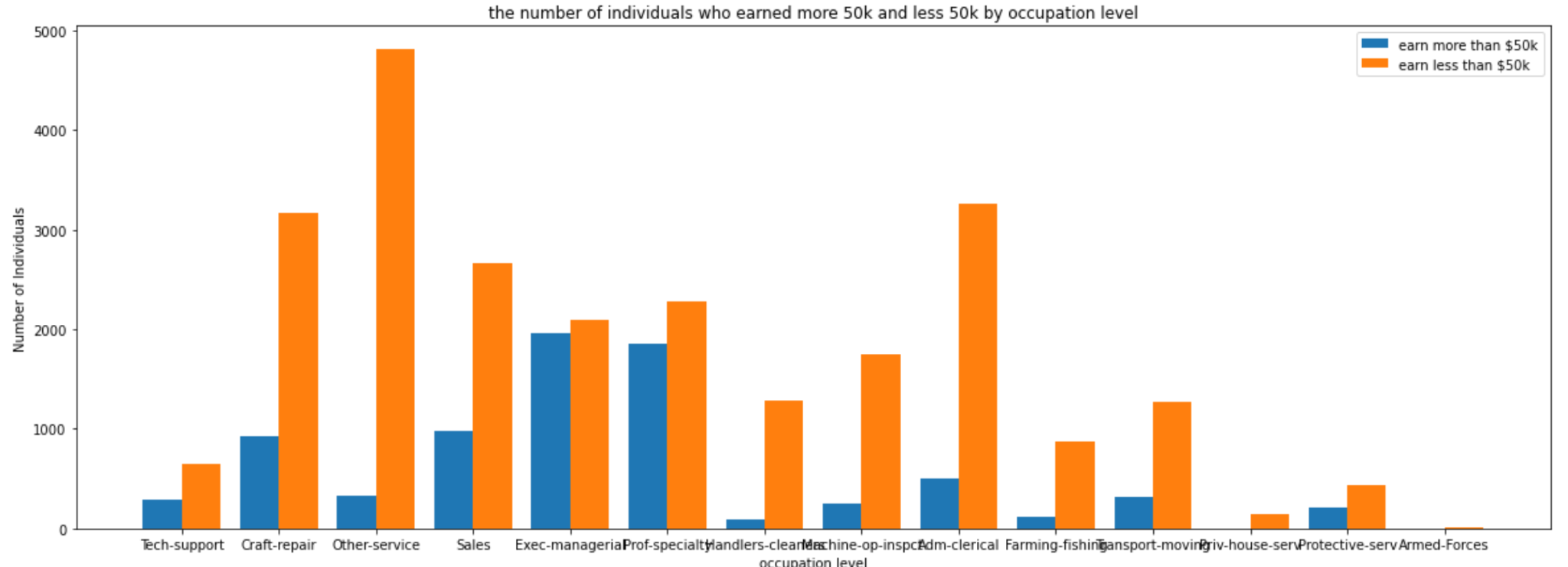
afficher le nuage de points de (classe ouvrière et race) en utilisant la fonction de dispersion fournie par la bibliothèque matplotlib



* graphique de barres empilées montrant le nombre d'individus qui gagnent plus/moins 50 000 $ par profession :

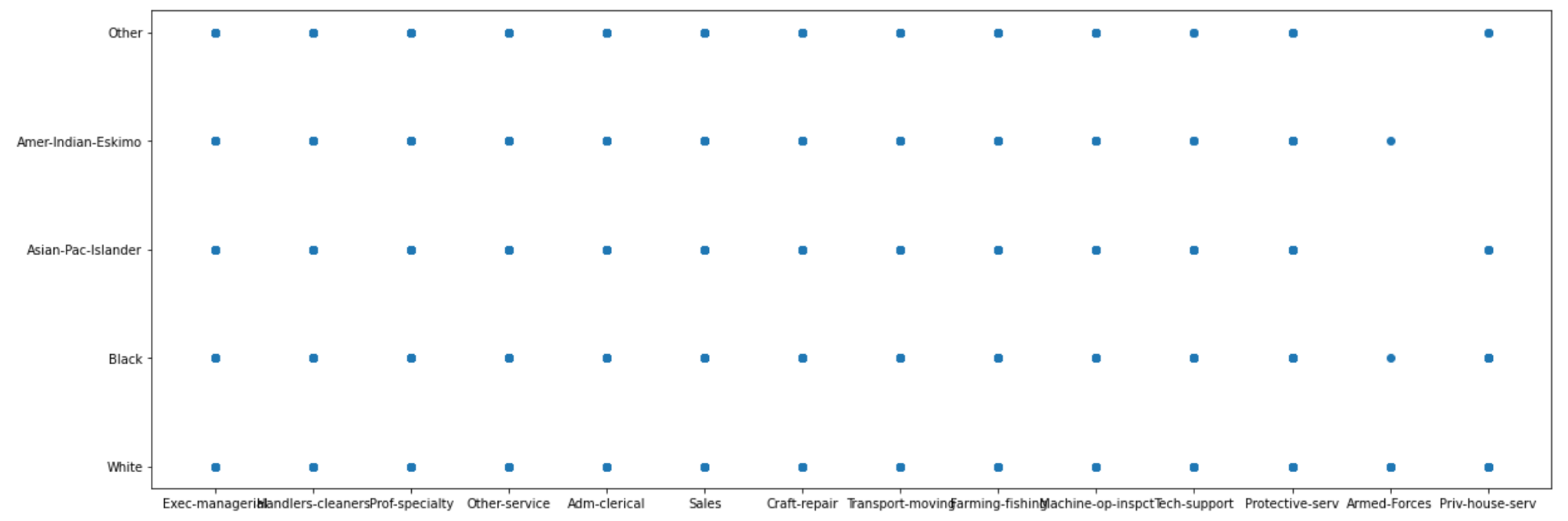
afficher un graphique de barres implantées à l'aide de la fonction de barre calculée par bibliothèque matplotlib, le nombre d'adultes qui gagnent plus de 50 000 $ en utilisant l'attribut de revenu et le précisent à > 50 000 et en utilisant l'attribut occupation pour préciser les groupes d'occupation et obtenir la longueur du tableau pour obtenir le nombre

afficher un graphique de barres implantées à l'aide de la fonction de barre calculée par bibliothèque matplotlib, le nombre d'adultes qui gagnent moins de 50 000 $ en utilisant l'attribut de revenu et le précisent à <=50k et en utilisant l'attribut occupation pour préciser les groupes d'occupation et obtenir la longueur du tableau pour obtenir le nombre



* nuage de points montrant la répartition des professions des enregistrements :

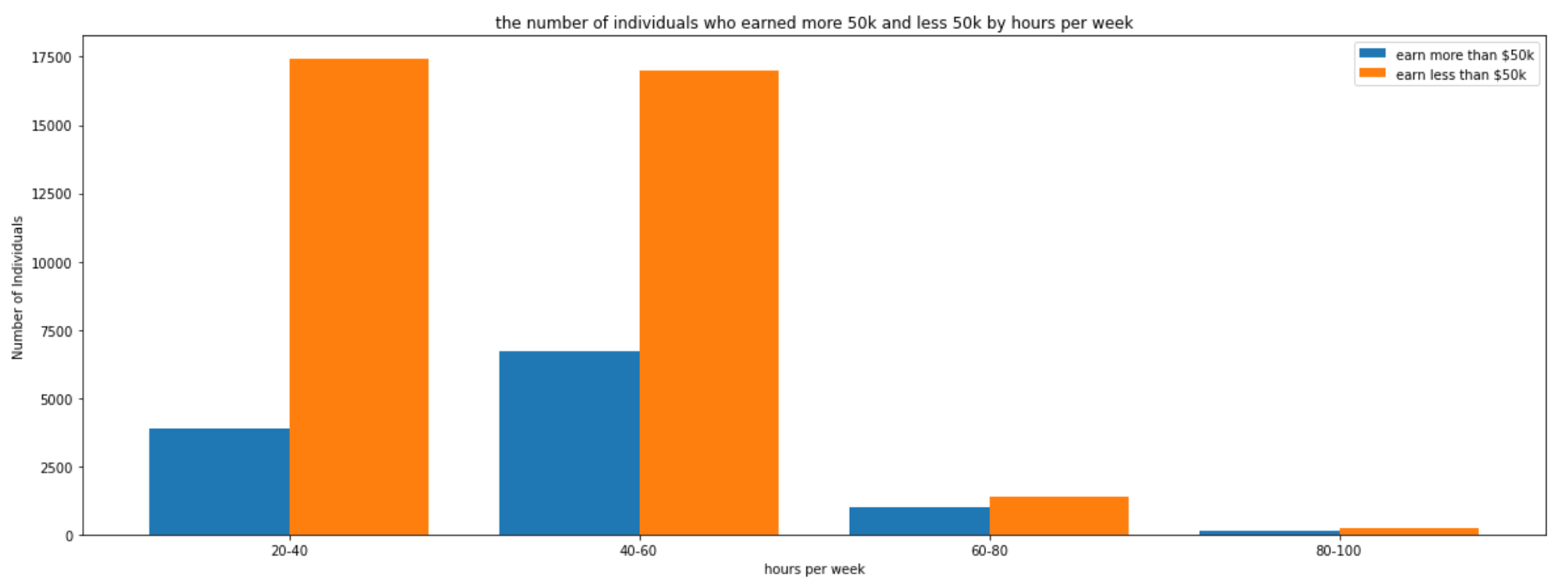
afficher le nuage de points de (profession vs race) en utilisant la fonction de dispersion fournie par la bibliothèque matplotlib



* graphique de barres empilées montrant la relation entre le nombre d'heures travaillées et le revenu :

afficher un graphique de barres jalonnées à l'aide de la fonction de barre calculée par la bibliothèque matplotlib, le nombre d'adultes qui gagnent plus de 50 000 $ en utilisant l'attribut de revenu et en précisant qu'il est > 50 000 et en utilisant l'attribut hrs per week pour obtenir l'heure par semaine des groupes et obtenir le longueur du tableau pour obtenir le nombre

afficher un graphique de barres jalonnées à l'aide de la fonction de barre calculée par la bibliothèque matplotlib, le nombre d'adultes qui gagnent moins de 50 000 $ en utilisant l'attribut de revenu et en précisant qu'il est <=50k et en utilisant l'attribut hrs per week pour obtenir l'heure par semaine des groupes et obtenir le longueur du tableau pour obtenir le nombre

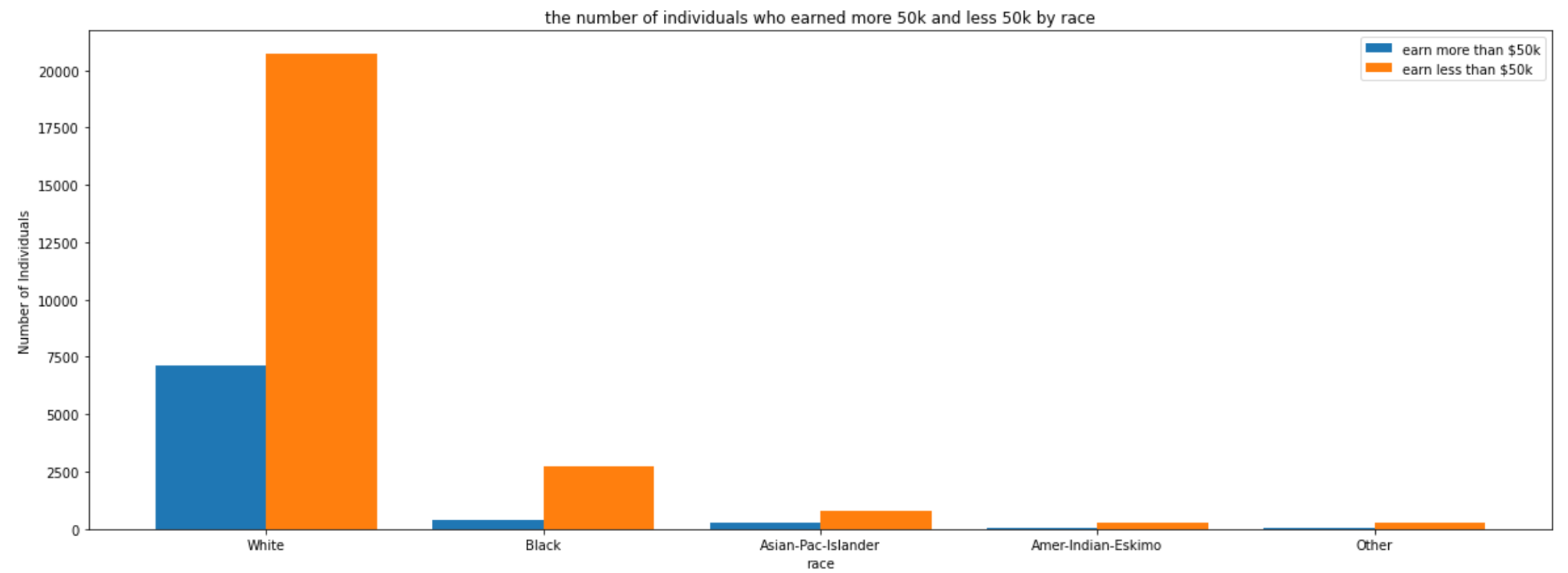


* graphique de barres empilées montrant la relation entre la course et le revenu :

afficher un graphique de barres jalonnées à l'aide de la fonction de barre calculée par la bibliothèque matplotlib, le nombre d'adultes qui gagnent plus de 50 000 $ en utilisant l'attribut de revenu et en le précisant > 50 000 et en utilisant l'attribut de race pour obtenir les groupes de race et obtenir la longueur du tableau pour obtenir le nombre

afficher un graphique de barres implantées à l'aide de la fonction de barre calculée par la bibliothèque matplotlib, le nombre d'adultes qui gagnent moins de 50 000 $ en utilisant l'attribut de revenu et le préciser à <= 50k

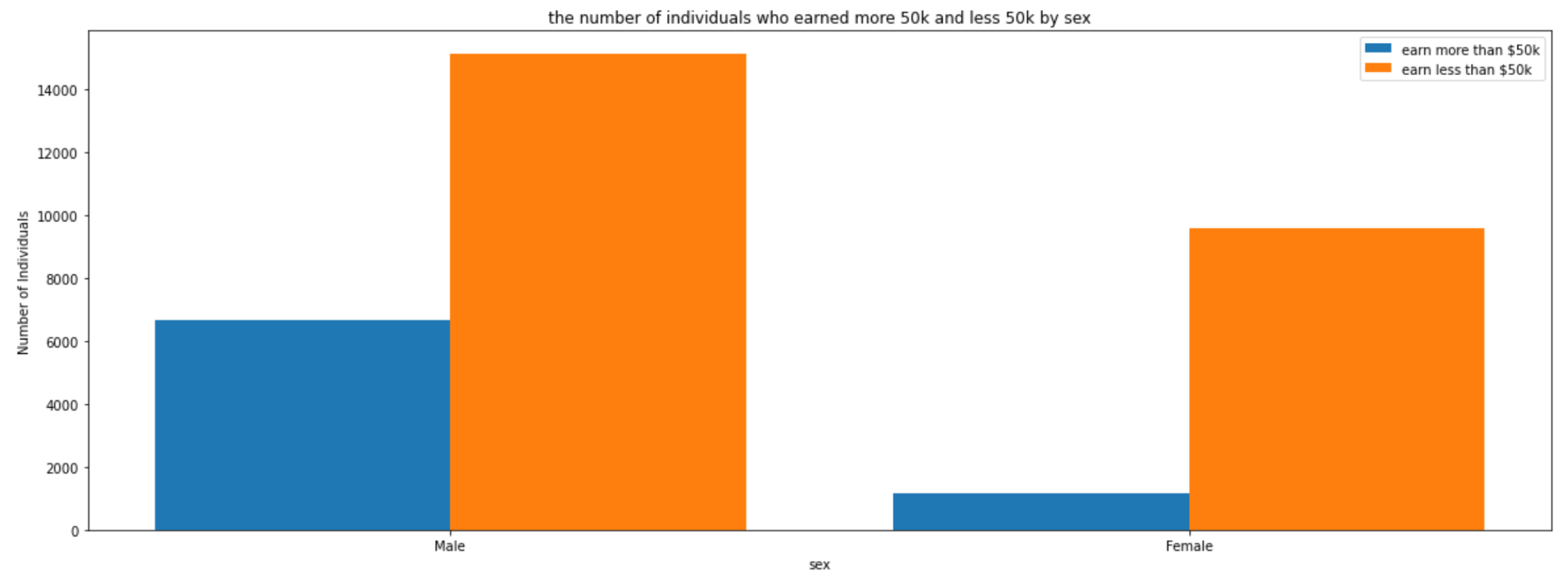
et en utilisant l'attribut race pour obtenir les groupes de race et en obtenant la longueur du tableau pour obtenir le nombre



* graphique de barres empilées montrant la relation entre le sexe et le revenu :

afficher un graphique de barres implantées à l'aide fonction de barre calculée par la bibliothèque matplotlib, le nombre d'adultes qui gagnent plus de 50 000 $ en utilisant l'attribut de revenu et en le précisant > 50 000 et en utilisant l'attribut de sexe pour obtenir le sexe groupes et obtenir la longueur du tableau pour obtenir le nombre

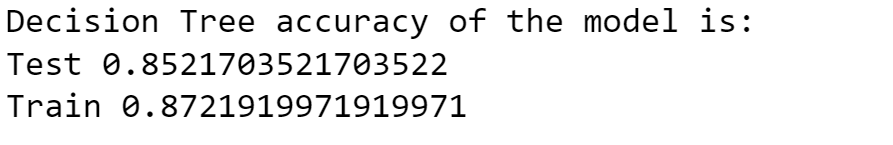
afficher un graphique de barres implantées à l'aide fonction de barre calculée par la bibliothèque matplotlib, le nombre d'adultes qui gagnent moins de 50 000 $ en utilisant l'attribut de revenu et en le précisant <= 50k et en utilisant l'attribut de sexe pour obtenir le sexe groupes et obtenir la longueur du tableau pour obtenir le nombre



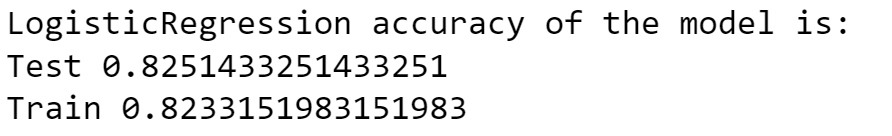
# **Prédiction :**

Certains des algorithmes de classification disponibles dans scikit-learn :

* **Arbres de décision:**

****

* **Régression logistique:**



* **Observations:**
* Arbres de décision : 85,21%
* Logistic Regression : 82,51%

La précision du modèle optimisé sur les données de test sur le classificateur de Arbres de décision est de 85,21 % .

* **La matrice de confusion:**

****

# **Conclusion :**

Nous avons fait une analyse de classification d'apprentissage automatique de bout en bout et nous avons appris et obtenu plusieurs informations sur les modèles de classification et les clés pour en développer un avec de bonnes performances.