A. Description des données

Dans ce travail, nous utilisons un jeu de données décrivant une population de personnes (32 561 au total) en mettant l'accent sur les informations pertinentes à leurs revenus annuels. Au lieu des valeurs exactes de ce revenu, une séparation en deux classes est utilisée (≤ 50 000 , > 50 000). Les entrées sont décrites par douze attributs représentant différents traits des personnes (valeurs traduites en français lorsque justifié) :

- Age : Entier positif
- Travail (nature du poste occupé) : Nominal = {*Privee, Travailleur-Auto-inc, Travailleur-Auto-non-inc, Gouv_Local, Gouv-Provincial, etc.*}
- Éducation (le plus haut niveau éducatif acquis) : Nominal = {*Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, Assoc-acdm, Assoc-voc, Prof-school, HS-grad, Some-college, Bachelors, Masters, Doctorate*}
- Nombre (années) d'éducation : Entier positif
- État civil (statut matrimonial) : Nominal = {*Jamais_marié, Marié_civil, Veuf, Divorcé, Séparé, etc.*}
- Occupation (secteur de l'emploi) : Nominal = {*Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, etc.*}
- Lien de parenté avec la personne du même ménage : Nominal ={*Pas_dans_famille, Célibataire, Mari, etc.*}
- Race: Nominal ={*Blanc, Noir, Asiatique, Amérindien, Autre*}
- Sexe : Nominal
- Heures (travaillées) par semaine : Entier positif
- Pays d'origine : Nominal
- Revenu: Ordinal = {≤ 50 000 , > 50 000}

B. Travail à faire

Utiliser *Python* pour établir les réponses aux questions de cet énoncé. Il n'est pas interdit de faire les calculs à l'aide d'un autre outil ou à la main, toutefois les réponses qui ne sont pas appuyées par un code *Python* fonctionnant ne seront pas prises en compte.

1. Étude exploratoire

1. Combien de personnes travaillent-elles dans le privé ?
2. Quels sont les pays d'origine des femmes dans ce jeux de données ?
3. Quel est le pourcentage d'hommes ayant un niveau éducatif *HS-grad* ?
4. Quelles sont les **attributs** pour lesquels **manquent** de valeurs ? Combien de valeurs manquent par attribut ?Quels attributs ont le **nombre** de valeurs **manquantes maximal** ? Et **minimal** (mais > 0) ?
5. Quelles sont les valeurs la **plus élevée** et la **moins élevée** de l'attribut *nombre d'années d'éducation* ?

2. Statistiques et probabilités

Pour la suite, nous considérons le jeux de données complet ainsi qu'un échantillon dont la taille est 10% de la taille complète (3256 lignes). Pour des raisons de déterminisme, on prendra les premiers 10% des enregistrements (lignes 1 à 3256).

1. Extraire les 3 valeurs du nombre d'années d'éducation les plus fréquentes. Correspondent-elles à un intervalle complet (ex. 8, 9 et 10) ?
2. **Moyenne** des nombre d'*années d'éducation* ? Et **médiane** ?
3. Quels pays ont les **moyennes** des *âges* de leurs ressortissants les **plus élevée**s ?
4. Quelle % des femmes ayant une maîtrise ont un *salaire* >50K ? Quelle est ce % chez les hommes ?
5. Quelle % des personnes **ne vivant pas en famille** ont un diplôme universitaire (baccalauréat, maîtrise ou doctorat) ?
6. Quelle *race* a la proportion la plus élevée de *salaires* ≤50K ? Qu'en est il pour la valeur >50K ?

7. Existe-t-il une **différence** dans les *salaires* des hommes **mariées** et les hommes **ne vivant pas en famille** ? Dans quel sens va-t-elle ?
8. Quelle est la **variance** des *âges* dans notre **jeu** de données **complet** ? Quelle est sa valeur pour **l'échantillon** ?
9. Parmi les *années d'éducation* et l'*âge*, quel attribut a plus d'influence sur la valeur du salaire ? Dit autrement, lequel des deux est plus **corrélé** avec le salaire ? Pour y répondre, il faudrait coder l'attribut ordinal salaire par des nombres, par ex., ≤50K par 1 et >50K par 2.
10. Est-ce que *l'âge* est un facteur d'influence important pour le *salaire* des femmes ? Répondez en comparant les *ages* moyens dans les groupes de femmes à salaires ≤50K et >50K, respectivement.
11. Quelle est la différence entre les **écarts types** des *heures par semaine* pour le jeux complet et pour l'échantillon ?

3. Tests d'hypothèses

Dans cette partie, nous allons considérer le jeu de données complet comme étant notre « population » et notre attention sera portée sur l'échantillon des 1ers 3256 lignes qui nous servira comme base pour la validation de diverses hypothèses. On assume aussi que l'attribut *nombre d'heures par semaine* suit une loi normale.

1. On émet l'hypothèse que la **moyenne** des heures par semaine sur le jeu de données complet n'est pas plus élevée que 40 hrs/sem. En vous basant sur **l'échantillon**, et en utilisant un **écart type** de la population égal à 12,35 , testez cette hypothèse à l'aide d'un **z-test**. Prenez un niveau de signification de 5%.
    - *N.B.* Pour des raisons pédagogiques, on ignore la valeur effective de la moyenne pour le jeu complet : ce qu'on veut tester est si la valeur observée sur l'échantillon est suffisamment extrême pour permettre de rejeter l'hypothèse H_0 que vous aurez formulée.
2. On émet une autre hypothèse : les femmes travaillent, en moyenne, plus d'heures que la **moyenne** hypothétique de 40 hrs/sem. En vous basant uniquement sur la partie pertinente de **l'échantillon**, testez cette hypothèse à l'aide d'un **z-test**. Prenez un niveau de signification de 10%.
3. Finalement, existe-t-il un écart significatif (à 5%) entre le salaire moyen **de la population** et celui des femmes ? Pour y répondre, utilisez le codage suggéré par la question 9. Avec ce codage, l'écart type de la population est arrondi à 0,43.