Rapport des données d'analyse

1. Étude exploratoire

1.1. Les personnes qui travaillent dans le privé sont au nombre de : **22696.**

1.2. Les pays d'origine des femmes dans ce jeu de données sont : United-States, Trinadad&Tobago, Philippines, Canada, Holand-Netherlands, Honduras, South, Hungary, Mexico, Taiwan, Cambodia, Columbia, Guatemala, Japan, Cuba, Puerto-Rico, Dominican-Republic, England, Germany, Iran, France, Scotland, Portugal, Nicaragua, El-Salvador, Italy, Ecuador, Peru, India, Jamaica, Laos, Poland, China, Outlying-US(Guam- USVI-etc), Haiti, Hong, Vietnam, Ireland, Yugoslavia, Greece, Thailand et Inconnu (?).

1.3. Le pourcentage d'hommes ayant un niveau éducatif HS-grad est de :
32.63%

1.4. Nous identifions les données manquantes comme les données représentées par le symbole **' ?'** Les attributs manquant de valeurs sont : **classe.travail, occupation, pays.natal.**

Le nombre de valeurs manquantes par attribut est :

Attribut	Nombre de valeur manquante
ocupation	**1843**
classe.travail	**1836**
pays.natal	**583**

L'attribut avec le nombre maximal de valeurs manquantes est : **occupation**.
L'attribut avec le nombre minimal de valeurs manquantes (mais > 0) est : **pays.natal**.

1.5. Les valeurs la plus élevée et la moins élevée de l'attribut nombre d'années d'éducation sont respectivement **16** et **1**.

2. Statistiques et probabilités

2.1. Les trois valeurs les plus fréquentes du nombre d'années d'éducation sont : **9**, **10** et **13**.
Correspondent-elles à un intervalle complet ? : **NON**

2.2. La moyenne du nombre d'années d'éducation est de : **11.4305**. La médiane est de : **12.00**.

2.3. L'âge moyenne de notre échantillon étant de **43.567**, voici la liste des pays ayant les moyennes d'âges les plus élevées pour leurs ressortissants sont : Ecuador (90.00), Hungary (69.00), Yugoslavia (66.00), Scotland (62.00), Jamaica (60.00), Columbia (57.00), Poland (50.75), Greece (50.00), China (49.55), Trinadad&Tobago (49.50), France(49.00), Honduras (47.00), El-Salvador (46.00), Italy (45.60), Guatemala (45.50), Germany (45.07), Vietnam (44.83), India (44.53), Cuba (44.00), Inconnu « ?» (43.97) et Canada (43.88)

2.4. Le pourcentage de femmes ayant une maîtrise avec un salaire >50K est de **67.61 %.** Chez les hommes, ce pourcentage est de **90.90 %.**

2.5. Le pourcentage de personnes ne vivant pas en famille avec un diplôme universitaire est de **44.88%.**

2.6. La race avec la proportion la plus élevée de salaires ≤50K est : **Autre**. Pour les salaires >50K, la race correspondante est : **Asiatique**.

2.7. **Oui il y a une différence significative dans la répartition des salaires entre les hommes mariés et les hommes seuls.** Sens : **Les hommes mariés ont tendance à avoir plus des salaires >50K**

2.8. La variance des âges dans le jeu de données complet est de **186.061**, tandis que pour l'échantillon, elle est de **149.912**.

2.9. L'attribut le plus corrélé au salaire entre les années d'éducation et l'âge est : années d'éducation. Nous trouvons : Corrélation âge – salaire : **0.210 ;** Corrélation nombre.education – salaire : **0.345.**

2.10. Après calcul nous trouvons : Moyenne d'âge des femmes avec un salaire <=50K : **38.46** et Moyenne d'âge des femmes avec un salaire >50K : **43.16**. Différence : 43.16−38.46 = **4.7 ans.**
La comparaison des âges moyens montre que **Oui, l'âge semble être un facteur d'influence.** Cela peut être traduit par des arguments comme l'expérience professionnelle.

2.11. Après calcul nous trouvons : Écart-type des heures par semaine (jeu complet): **12.3474** et Écart-type des heures par semaine (échantillon): **11.7295** La différence entre les écarts types des heures travaillées par semaine pour le

jeu complet et l'échantillon est de **0.6179.**

3. Tests d'hypothèses

3.1. En testant l'hypothèse H0 que la moyenne des heures travaillées par semaine sur le jeu complet n'est pas plus élevée que 40 heures/sem, nous avons obtenu les résultats suivants :

- z-score : **20.51**
- p-value : **0.0000**

Par conséquent, **on rejette l'hypothèse nulle H0 : la moyenne des heures travaillées est significativement supérieure à 40**

3.2. En testant l'hypothèse H0 que les femmes travaillent, en moyenne, plus d'heures que 40 heures/sem, nous avons obtenu les résultats suivants :

- z-score : **-2.16**
- p-value : **0.9847**

Ainsi, **On ne peut pas rejeter l'hypothèse nulle H0 : il n'y a pas de preuve suffisante que les femmes travaillent plus de 40 heures par semaine.**

3.3. En testant l'hypothèse H0 que le salaire moyen des femmes est égal à celui de la population (μ femmes = μ population), nous avons obtenu les résultats suivants :

- z-score : **-11.93**
- p-value : **0.0000**

D'où, **l'on rejette l'hypothèse nulle H0 : il existe un écart significatif entre le salaire moyen de la population et celui des femmes.**