1) L’entreprise souhaite un modèle permettant de déterminer le revenu potentiel d’une personne à partir seulement de son pays d’origine et du revenu de ses parents .

**Trouver un modèle efficace pour faire une telle prédiction ?**

**Insight et recommandation**:

On peut prédire de manière efficace le revenu d’une personne, en utilisant le logarithme du gdppp , la classe de revenue des parents et l’indice de gini.

2) J’ai importé les données de la wid depuis openclassrooms.

Les données importées portent sur la période 2004-2011 mais la très grande majorité est entre 2006 et 2010 (+ de 98% des données).

Le gdpppp n’a pas changé de façon significatif entre 2006 et 2010, après de petits ajustement fait en fonction de la croissance mondiale(augmentation pour les années 2006 et 2009), J’ai considéré que j’étais suffisamment proche des données de 2008 pour dire que mon étude porté uniquement sur 2008.

3)J’ai souhaité ajouter la population par pays à mon tableau principale. J’ai donc téléchargé les données de la fao sur la population en 2008. Malheureusement les nom des pays de ma table sont en iso3 alors que la fao a simplement le nom du pays en entier en francais. J’ai du chargé une table de correspondence iso3, nom francais afin de réaliser ma jointure.

4) Certain pays ne figure malheureusement pas sur la liste je les ai donc ajouté

D’autres ne sont pas écris exactement de la même façon, je les ai donc modifié (beacoup de problème de majuscule)

5) Je calcul l’indice de gini de chaque pays à l’aide de la fonction gini et j’ai ajouté une colonne gini à ma table principale.

6) Puis après plusieurs jointures j’obtient ma table principale qui rassemble toutes les données de mon études à ce stade.

7) Les années utilisées par la WID vont de 2004 à 2011, avec une forte proportion de 2007 et 2008. Mon étude portera seulement sur l’année **2008**

116 pays sont présents dans l'étude de la WID

Mon étude portera sur **111** d’entre eux (Environ 50% des pays recensés par la FAO)

Environ **6,2 milliards** de personnes sont couvertes par l'étude, soit environ **88%** de la population mondiale recensée par la FAO en 2008.

Les quantiles utilisés par la WID sont des **percentiles**, pour l'ensemble des pays

L'utilisation des percentiles est une **bonne méthode**, car cela permet de considérablement réduire la taille de l'échantillon (versus un individu par personne), tout en préservant suffisamment d'information.

Le dollar PPP est une unité qui permet de comparer le pouvoir d'achat entre deux pays sans distorsion due aux taux de change.

8) La distribution des revenus varie selon les pays (échelle logaritmique)

J’ai choisis 10 pays pour montrer la différence entre trois groupe de pays occidentaux, afriquain et l’amérique du sud.

La bolivie, le brésil, la colombie, l’Allemagne, la France, le japon la suede le ghana l’ouganda et les états unis

On peut ici identifier 3 groupes de courbes les pays développés qui sont très haut est assez stable les pays d’amérique du sud qui partent de très bas et qui finissent très haut et les pays d’afrique qui partent de très bas mais qui ne monte pas beaucoup.

Les états unis sont à part leur courbe est beaucoup moins stable que celle des autres pays développés, c’est le pays développé qui part le plus bas mais c’est aussi celui qui finit le plus haut.

9) Courbe de Lorenz

Ici la courbe de lorenz permet d’identifier si un pays est plus ou moins inégalitaire. La droite y = x représente une égalité de revenu parfaite entre les individus,  plus la courbe d’un pays s’éloigne de cette droite plus il est inégalitaire.

Les groupes sont les même sauf que les états unis ont rejoint le groupe des pays d’Afrique. On peut aussi voir que cette fois se sont les pays d’amérique du sud qui se retrouvent les plus bas derrière les pays d’afrique.

la suède dépasse assez facilement les autre pays développé.

On peut à partir de la courbe de lorenz calculer l’indice de Gini. C’est l’aire entre la courbe étudié et la droite parfaite divisé par l’aire sous la droite parfaite. Plus l’indice de gini est grand plus le pays est inégalitaire .

10) Les indices de Gini des différents pays semble se rapprocher, les pays qui avait un indice de gini très élevé descendent alors que les autres ont tendances a monter (sauf la France)

11) Les pays de l’europe de l’est domine très largemement le top5 c’est pourtant très loin d’être la région la plus riche du globe.

Le worst 5 se compose de 3 pays d’amérique du sud et de l’afrique du sud là encore se ne sont pas les pays les plus pauvre de la planète loin de là, l’afrique du sud est même l’un des plus riche pays d’afrique. (la république centrafricaine en revanche est un des pays les plus pauvre du monde..)

La France est 39em du classement sur 111 dans le milieu haut

12) Au premier abord on pourrait croire que l’indice de Gini est directement corrélé avec le gdpppp du pays mais quand on regarde plus précisément on voit que ce n’est pas aussi simple.

Les Etats unis au niveau des pays africains, des pays d’Amérique du sud dans le worst5 alors que les pays africains sont bien plus Pauvre, trois pays d’europe de l’est leader dans le top5 alors que leurs gdpppp sont plutôt moyen etc…

13) Je récupère les données d’une étude sur le coeficient d’élasticité, elle me donne les coefficient d’élacticité par région

14) Je décide donc de classer mes pays par région, pour faire cela j’ai téléchargé sur le site de la FAO une table de pays pour chaque région et j’ai ensuite réalisé des jointure puis j’ai ajouté une colonne région dans ma table principale

15) On réalise 1000fois la variable ln(Yparent) qui suit la loi normal centré réduite.

On réalise ensuite 1000 fois la variable epsilone qui suit la loi normal centré réduite c’est l’erreur.

On prend alpha = 0

A l’aide de la formule qui donne le revenu des enfant en fonction de ln(Yparent), je calcul les 1000 revenus Ychild associés. Je calcul ensuite les mille revenus Yparent associés.

Je partage maintenant mes 1000 revenus enfant en centile qui représenteront leurs classes de revenu pareil pour mes revenus parents

Je crée ensuite une table avec tout mes revenu enfant leur classe associer et la classe du parent associé.

Je regroupe par classe enfant, classe parent et je compte le nombre de fois qu’apparait chaque combinaison. Je détermine ensuite pour chaque classe enfant les proba conditionnelles des classes parents.

16) C’est la représentation de la distribution conditionnelle pour 10 quantiles et un coef d’elasticité de 0,1

On peut voir que quand le coef d’élasticité est faible la classe parent a très peu d’influence sur la classe enfant

17) C’est la représentation de la distribution conditionnelle pour 10 quantiles et un coef d’elasticité de 0,9

On peut voir que quand le coef d’élasticité est élevé la classe parent a beaucoup d’influence sur la classe enfant

18) Un coefficient d’élasticité élevé nous permet donc de mieux prévoir le revenu de l’enfant,

quand on connait le revenu des parents. Il sera donc sans doute plus facile de faire nos

prédiction sur des pays ayant un coefficient d’élasticité élevé.

19)J’ai réalisé un test ANOVA pour prédire le revenu des clients en fonction de leur pays d’origine.

La P-Value est infiniment faible, le pays est donc un facteur d'influence statistiquement fiable.

Le modèle permet d'expliquer 48.6% de la somme des carrés. Ceci est un bon début, mais ce n'est pas suffisant pour pouvoir déterminer de manière fiable le revenu d'un individu lambda.

L’analyse graphique de notre modèle n’est pas satisfaisante non plus en effet la variance des résidus varie beaucoup (beaucoup de résidus sont très éloigne de la ligne 0, hétéroscédacité) et les résidus ne sont pas normalement distribué en effet dans le deuxieme graphique la courbe s’éloigne très fortement de la droite théorique de répartition des résidus surtout sur la fin)

Le modèle n’est en plus validé par aucun tests

20)je réalise une régression linéaire pour expliqué le revenu enfant à partir de l’indice de gini et du gdpppp

La P-Value est aussi infiniment faible

Le modèle permet d'expliquer 44% de la somme des carrés. C’est moins que le modèle précédent.

Ce n’est donc pas satisfaisant.

21) Le modèle n’est lui non plus pas validé par les tests. Et n’est pas non plus satisfaisant au niveau graphique en effet les résidu ne respecte ni l’hypothèse d’homoscadicité ni l’hypothèse d’homogénéité.

On peut aussi voir de nombreux outlier. De nombreuses valeur dépasse très largement le seuil levier.

22)Cette je vais prendre le log du revenu enfant et le log du gdpppp

P-value très faible => facteur d’influence fiable

Mais Le modèle permet d'expliquer 66% de la variance. C’est largement mieux

23) Les graphique sont bien plus satisfaisant

Les point sont très regroupé autour de la ligne zéro dans le premier

Les points suivent à peu près la droite dans le qqplot et l’histogramme des résidu est en forme de cloche.

La fonction logarithme a permis de linéariser les données

Tous les test sont encore négatifs, on peut supposer que c’est à cause des outliers toujours présents en grand nombre dans ce modèle.

24) on ajoute cette fois le revenu des parents des individus.

La P-value est encore très faible.

Le modèle permet d'expliquer 71% de la variance. C’est le meilleur taux obtenu depuis le début

25) On est dans le même cas que précédemment les graphique sont satisfaisant mais les test reviennent négatifs sans doute à cause des nombreux outliers.

Ce modèle est le plus satisfaisant c’est celui qui permet les meilleurs prédictions c’est celui que nous allons utilisé pour prédire les revenus des enfants

26) En incluant la **classe de revenu des parents, l’analyse des résidus est**

sensiblement la même que sur le modèle précédent (ils suivent une loi

normale et sont globalement de même variance) mais on gagne **5 points** sur le coefficient de détermination pour atteindre **0.7129.**

La variance totale est expliquée à **71%** par le pays de naissance et le revenu des parents et à **19%** par les autres facteurs non considérés dans le modèle.

Un indice de **gini plus élevé semble défavoriser plus de personnes qu’il n’en favorise.** Ceci est mis en évidence par le coefficient négatif devant l’indice au sein du modèle.

Attention néanmoins l’indice de gini était positif dans le modèle sans log. Et même si l’indice de gini est négatif il est surtout très faible.

27) Ce projet m’a permis de m’améliorer au niveau de la récupération et du traitement des données. J’en ai aussi appris beaucoup sur les hypothèses de validités des modèles de régression linéaires.  
  
J’ai trouvé la mission 3 un peu confuse.  
Il manque un petit texte pour expliquer quelle est le but des 12 opérations.