Base de données : Enquêtes Emploi réalisées par l'Insee en 2015 et 2016
Champ de l'étude : Salariés en France de 1962 à 2016.

0. Exploration et description de la base

Appel des librairies
library(dplyr)
library(ggplot2)
library(tidyverse)
library(oaxaca)
library(tableone)
library(forcats)

Stockage des dataset dans deux objets
db_2015 <- enq_emploi_salarie_2015
db_2016 <- enq_emploi_salarie_2016

Consolidation des deux dataset
db <- rbind(db_2015, db_2016)

Synthèse des caractériques des variables des dataset
Caractéristiques des tables
str(db_2015)
str(db_2016)
str(db)

- Etat des lieux

Les bases originales ont été traitées càd que les valeurs aberrantes et manquantes ont été omises.
La base consolidée contient 48 489 observations représentant des salariés français de 2015 et de 2016 et 18 variables (12 qualitatives et 6 quantitatives)

- Variables qualitatives :

> "lnais" indique Lieu de naissance de l'enquête contient 2 modalités (1= français | 2= étranger).
> "nfrred" spécifie la nationalité de l'enquêté contient 3 modalités (1= français de naissance | 2= français par naturalisation | 3= étranger).
> "naiper" indique le lieu de naissance du père contient 2 modalités (1= français | 2= étranger).
> "nfrp" spécifie la nationalité du père contient (1= français de naissance | 2= français par naturalisation | 3= étranger).
> "naimer" indique le lieu de naissance de la mère contient 2 modalités (1= français | 2= Etranger).
> "nfrm" spécifie la nationalité de la mère contient 3 modalités (1= français de naissance | 2= français par naturalisation | 3= étranger).
> "sexe" spécifie le sexe de l'individu contient 2 modalités (1= popnonimmigreeme | 2= popimmigreeme)
> "ddipl" indique le diplôme le plus élevé obtenu contient 7 modalités ordinales (1= diplôme>Bac+2 | 3= diplôme:Bac+2 | 4= diplôme:Bac ou brevet | 5= diplôme:CAP,BEP | 6= diplôme:brevet des collèges | 7= diplôme:si aucun diplôme ou CEP)
> "couple" indique le statut marital contient 2 modalités ordinales (1= Couple | 2= Seul)

> "cstotr" indique la catégorie socio-professionnelle contient 4 modalités ordinales (1= Cadre | 2= Profession intermédiaire | 3= Employé | 4= Ouvrier)
> "horaic" indique la nature des horaires contient 3 modalités ordinales (1= horaires hebdomadaires popnonimmigreeogènes | 2= horaires alternés | 3= horaires variables)
> "nuitc" indique Travail de nuit contient 3 modalités ordinales (1= travail de nuit plus de la moitié du temps | 2= travail de nuit moins de la nuit | 3= pas de travail de nuit)

- Variables quantitatives :

> "salaire" indique le salaire mensuel net retiré de la profession principale, corrigé des valeurs aberrantes et manquantes.
> "heures" spécifie le nombre d'heures travaillées en moyenne par mois dans l'emploi principal (heures supplémentaires comprises), corrigé des valeurs aberrantes et manquantes.
> "date" indique l'année de début dans l'entreprise.
> "age" renseigne l'âge en années.
> "annee_enq" renseigne l'année d'enquête.

Question1 : Déterminer la part d'immigrés dans la population active française

> Selon la définition de l'Insee un immigré est une personne née étrangère à l'étranger et résidant en France. Les personnes nées françaises à l'étranger et vivant en France ne sont donc pas comptabilisées.

> Six variables mobilisées relatives à l'origine "immigrée" ou "non française" de l'enquêté: "lnais", "nfrred", "naimer", "nfrred", "naiper", "nfrp"

- Création d'une nouvelle variable catégorielle relative à l'origine des enquêtés, elle prend la modalité "1" quand l'enquêté est immigré sinon "0" quand il est né en France.
  db$poporigine <- ifelse((db$lnais=="2" & db$nfrred!="1") | (db$lnais=="1" & db$nfrred=="1") & ((db$naimer=="2" & db$nfrm!="1") | (db$naiper=="2" & db$nfrp!="1")),1,0)

- Restitution d'un tableau de contingence (tri croisé) indiquant la part relative à la population immigrée dans la population.
  table(db$poporigine)

- Restitution d'un tableau de contingence (tri croisé) indiquant la part relative à la population immigrée dans la population active.
  prop.table(table(db$poporigine))
  print(prop.table(table(db$poporigine)))

> Interprétation : 16,49% des enquêtés est d'origine immigrée.

Question #2 : Analyser l'écart de salaire moyen entre la population immigrée et la population née en France

- Statistiques descriptives

- Calcul de la moyenne de salaire entre les deux populations.
  summarise(group_by(db,poporigine),mean(salaire))

> Interprétation :

- **La population immigrée perçoit un salaire mensuel moyen de 1 644€.**
- **La population née en France perçoit un salaire mensuel moyen de 1 793€.**
- **La population immigrée perçoit un salaire mensuel moyen 9,07% plus faible (soit 149€ de moins) par rapport à la population née en France.**

- Création de deux tables distinctes échantillonnées : la population immigrée (modalité "1") et de la population née en France (modalité "0").
  db_popimmigree <- subset(db,db$poporigine==1)
  db_popfrancais <- subset(db,db$poporigine==0)

- Création d'une variable dichotomique relative au lieu de naissance de l'individu, modalité "1" si il est né à l'étranger sinon "0" si il est né en France.
  db$lieunais <- ifelse(db$lnais=="2",1,0)

- Restitution d'un tableau de contingence (tri croisé) indiquant la part relative de la population immigrée dans la population
  prop.table(table(db$lieunais))
  print(prop.table(table(db$lieunais)))

> Interprétation : 90% des individus sont nés en France et 10% des individus sont nés à l'étranger.

- Test d'hypothèses afin de déterminer si l'écart salarial entre les deux populations est significativement différent de zéro.

- Test d'égalité des variances (Test de Fisher) salariales des deux sous-échantillons.
  var.test(db_popimmigree$salaire,db_popfrancais$salaire)

  - p-value = 0,000229 (< 0,05) par conséquent on accepte l'hypothèse alternative (H1), les variances ne sont pas égales.

- Test d'égalité des moyennes (t-test) salariales des deux sous-échantillons (à variances non égales)
  t.test(db_popimmigree$salaire,db_popfrancais$salaire,paired=F,var.equal=F,alternative="two.sided",conf.level=0.95)

  - p-value = 2.2e-16* (< 0,05) par conséquent on accepte l'hypothèse alternative (H1) et on rejette l'hypothèse nulle selon laquelle les salaires des deux groupes sont égaux.

> Interprétation : La population immigrée et la population née en France perçoivent des salaires significativement différents de zéro au seuil de 5%.

- Comparaison de la distribution des salaires entre la population immigrée et de la population née en France.
  summary(db_popimmigree$salaire)
  summary(db_popfrancais$salaire)

- Visualisation graphique via des boîtes à moustaches.
  db$poporigine_graph <- factor(db$poporigine,c("1","0"),labels=c("Immigrée","née en France"))
  ggplot(db)+aes(y=salaire,x=poporigine_graph)+geom_boxplot()+xlab("")+ylab("Salaire mensuel")+ggtitle("Salaire mensuel selon l'origine")

> Interprétation :
- Salaires plus dispersés pour la population immigrée que la population née en France.
- Il y a très peu de valeurs atypiques car la base a été nettoyée, néanmoins on constate trois enquêtés nés en France perçoivent des salaires > à 7 000€.

<u>Question #3 : Comparaison des différentes caractéristiques entre la population immigrée et la population née en France</u>

- Création de la variable « ancienneté »

- Conversion de la variable "années" en nombre d'années (population)
  db$datantn <- as.numeric(as.character(db$datant))
- Création de la variable 'ancienneté' (population)
  db$anc <- db$annee_enq - db$datantn
- Conversion de la variable "années" en nombre d'années (échantillon immigré)
  db_popimmigree$datantn <- as.numeric(as.character(db_popimmigree$datant))
- Création de la variable 'ancienneté' (échantillon immigré)
  db_popimmigree$anc <- db_popimmigree$annee_enq - db_popimmigree$datantn
- Conversion de la variable "années" en nombre d'années (échantillon natif français)
  db_popfrancais$datantn <- as.numeric(as.character(db_popfrancais$datant))
- Création de la variable 'ancienneté' (échantillon français)
  db_popfrancais$anc <- db_popfrancais$annee_enq - db_popfrancais$datantn

- Analyse descriptive

- Moyenne des variables catégorielles par échantillon.
  prop.table(table(db_popimmigree$sexe))
  prop.table(table(db_popfrancais$sexe))
  prop.table(table(db_popimmigree$couple))
  prop.table(table(db_popfrancais$couple))
  prop.table(table(db_popimmigree$ddipl))
  prop.table(table(db_popfrancais$ddipl))
  prop.table(table(db_popimmigree$cstotr))
  prop.table(table(db_popfrancais$cstotr))
  prop.table(table(db_popimmigree$horaic))
  prop.table(table(db_popfrancais$horaic))
  prop.table(table(db_popimmigree$nuitc))
  prop.table(table(db_popfrancais$nuitc))

- Moyenne des variables quantitatives
  summarise(group_by(db,poporigine),mean(age),mean(heures),mean(anc))

> Interprétation :

	Pop immigrée	Pop née en France	Différences
Salaire	1 644 €	1 793€	- 149€
âge	43	42	1
sexe: homme	50,50%	49%	2%
sexe: femme	49%	51%	-2%
ancienneté	44	41	3
Pas de diplôme	22%	11%	11%
diplôme Brevet	5%	5%	0%
diplôme BEP	22%	28%	-6%
diplôme BAC	19%	20%	-1%
diplôme BAC 2	12%	16%	-4%
diplôme BAC sup	19%	19%	0%
Statut marital : couple oui	69%	69%	0%
Statut marital : couple non	31%	31%	0%
CSP 3	13%	15%	-2%
CSP 4	22%	29%	-7%
CSP 5	35%	32%	3%
CSP 6	29%	23%	6%
Nombres d'heures	34,9	36,3	- 1
Nature des horaires 1	77%	75%	2%
Nature des horaires 2	7%	7%	0%
Nature des horaires 3	15%	18%	-3%
Travail de nuit 1	4%	3%	0%
Travail de nuit 2	6%	6%	-1%
Travail de nuit 3	90%	91%	-1%

- Les attributs "sexe", "ancienneté", "couple", "nature des horaires" et "travail de nuit" sont distribués de manière relativement homogène.
- La population née en France est mieux préparée pour la vie active avec un niveau d'études supérieures plus élevée.
- Prépondérance de la population née en France dans les postes hiérarchiquement plus élevés.
- La population immigrée reste professionnellement plus longtemps active (4 années supplémentaires).
- La population immigrée travail davantage de nuit.

- Test d'hypothèses afin de déterminer quels sont les différences plus significatives expliquant l'écart salarial entre les deux populations

- Test d'indépendance du khi-deux (χ^2) pour les variables 'sexe', diplome', 'CSP', 'horaire', 'travail de nuit' des deux populations
  table_result_chitest <- CreateTableOne(data = db, vars=c("ddipl","sexe","cstotr","nuitc","couple"), strata="poporigine", test = TRUE, testExact = chisq.test)
  print(table_result_chitest)

- Test de student pour les variables 'heures', 'salaire', 'age', 'couple', 'horaires' des deux populations
  table_result_ttest <- CreateTableOne(data = db,
  vars=c("anc","whor","salaire","age","couple","horaic"), strata="poporigine", test = TRUE, testExact = oneway.test)
  print(table_result_ttest)

> Interprétation : On constate des différences significatives entre les deux populations concernant les variables suivantes : 'âge', 'ancienneté', 'diplôme', 'nombre d'heures', 'catégorie socio-professionnelle', 'horaires semblables d'une semaine à l'autre', 'horaires variables d'une semaine à l'autre'.

- Modélisation des variables significativement différentes une une afin de déterminer si elles ont un effet sur le salaire

- Modélisation afin d'estimer de l'effet de l'âge sur le salaire
  mm_sal_age <- lm(salaire~poporigine+age,data=db)
  summary(mm_sal_age)
  > Interprétation :
  - **L'ordonnée à l'origine (constante) "Alpha α": les individus nés en France perçoivent un salaire moyen d'entrée de 1 276€.**
  - **Coefficient directeur "Beta β" (): à âge équivalent, les immigrés perçoivent un salaire mensuel moyen inférieur de 142,7€ .**
  - **À population équivalente, chaque date d'anniversaire des employés engendre l'augmentation du salaire annuel de 12,2€.**

- Modélisation afin d'estimer de l'effet de l'ancienneté sur le salaire
  mm_sal_anc <- lm(salaire~poporigine+anc,data=db)
  summary(mm_sal_anc)
  > Interprétation :
  - **L'ordonnée à l'origine (constante) "Alpha α": les individus nés en France perçoivent un salaire moyen d'entrée de 1 500€.**
  - **Coefficient directeur "Beta β" (): à ancienneté équivalente, les immigrés perçoivent un salaire mensuel moyen inférieur de 75,9€.**
  - **À population équivalente, l'augmentation d'une année d'ancienneté engendre une augmentation du salaire mensuel moyen de 23€.**

- Modélisation afin d'estimer de l'effet du nombre d'heures travaillés sur le salaire

  db$whor <- db$salaire/db$heures
  mm_sal_heure=lm(whor~poporigine+heures,data=db)
  summary(mm_sal_heure)

  > Interprétation :
  - **L'ordonnée à l'origine (constante) "Alpha α": les individus nés en France perçoivent un salaire horaire moyen d'entrée de 35€.**
  - **Coefficient directeur "Beta β" (): à nombre d'heures travaillés égales, le salaire horaire moyen mensuel des immigrés est inférieur de 1,97€ à celui des natifs.**
  - **À population équivalente, chaque mois entraine l'augmentation du salaire horaire moyen de 0,36€.**

- Modélisation afin d'estimer de l'effet de la "catégorie socio-professionnelle" sur le salaire

  mm_sal_cstotr <- lm(salaire~poporigine+cstotr,data=db)
  summary(mm_sal_cstotr)

  > Interprétation :
  - **L'ordonnée à l'origine (constante) "Alpha α": les individus nés en France perçoivent un salaire moyen d'entrée de 2850€.**
  - **Coefficient directeur "Beta β" ():**
  - **À catégorie socio-professionnelle équivalente, les immigrés perçoivent un salaire mensuel moyen inférieur de 89€.**
  - **À population équivalente :**
  - **La différence salariale entre les professions intermédiaires et les cadres s'élève à 909€.**
  - **La différence salariale entre les employés et les cadres s'élève à 1501€.**
  - **La différence salariale entre les ouvriers et les cadres s'élève à 1325€.**

- Modélisation afin d'estimer de l'effet de la "diplôme" sur le salaire

mm_sal_ddipl <- lm(salaire~poporigine+ddipl,data=db)
summary(mm_sal_ddipl)

> Interprétation :
- **L'ordonnée à l'origine (constante) "Alpha α": les individus nés en France perçoivent un salaire moyen d'entrée de 2370€.**
- **Coefficient directeur "Beta β" ():**
- **À catégorie socio-professionnelle équivalente, les immigrés perçoivent un salaire mensuel moyen inférieur de 105€.**
- **À population équivalente :**
- **La différence salariale entre les individus ayant un diplôme bac+2 et les individus ayant un diplôme supérieur à bac+2 s'élève à 348€.**
- **La différence salariale entre les individus ayant un diplôme brevet ou bac et les individus ayant un diplôme supérieur à bac+2 s'élève à 717€.**
- **La différence salariale entre les individus ayant un diplôme CAP ou BEP et les individus ayant un diplôme supérieur à bac+2 s'élève à 765€.**
- **La différence salariale entre les individus ayant un diplôme brevet des collèges et les individus ayant un diplôme supérieur à bac+2 s'élève à 863€.**

- **La différence salariale entre les individus n'aucuns diplômes et les individus ayant un diplôme supérieur à bac+2 s'élève à 1029€.**

- Modélisation afin d'estimer de l'effet des "horaires semblables d'une semaine à l'autre " sur le salaire
  mm_sal_horaic=lm(salaire~poporigine+horaic,data=db)
  summary(mm_sal_horaic)
  > Interprétation :
  - **L'ordonnée à l'origine (constante) "Alpha α": les individus nés en France perçoivent un salaire moyen d'entrée de 1 793€.**
  - **Coefficient directeur "Beta β" ():**
  - **À horaires semblables d'une semaine à l'autre équivalentes, les immigrés perçoivent un salaire mensuel moyen inférieur de 148,7€.**

Question #4 : analyser l'écart de salaire moyen entre la population immigrée et la population née en France (toutes choses égales par ailleurs)

- Retraitement des variables
- Recodage
  db$ddipl <- fct_recode(db$ddipl, "7" = "6")
  db$nuitc <- fct_recode(db$nuitc, "1" = "2")
  db$horaic_un <- ifelse(db$horaic=="1",1,0)
  db$horaic_deux <- ifelse(db$horaic=="2",1,0)
  db$horaic_trois <- ifelse(db$horaic=="3",1,0)

- Transformation logarithmique de la variable à expliquer salaire afin d'induire la normalité et stabiliser la variance des résidus
  db$lsalaire <- log(db$salaire)

- Estimation de l'effet simultané de toutes les variables mobilisées "toutes choses égales par ailleurs" sur le salaire
  mm_sal_all <-
  lm(lsalaire~poporigine+age+sexe+couple+anc+ddipl+cstotr+heures+horaic_un+horaic_deux+horaic_trois+nuitc,data=db)
  summary(mm_sal_all)

> Interprétation :
"Toutes choses égales par ailleurs", à population équivalente :
- **Les individus immigrés perçoivent un salaire mensuel moyen inférieur de 0,82% à celui des individus nés en France, une différence de -8,3 pt contre le résultat de la question n°2.**
- **Les femmes perçoivent un salaire mensuel moyen inférieur de 8,65% à celui des hommes.**
- **L'augmentation d'une année d'ancienneté engendre une augmentation de 0,94% du salaire mensuel moyen.**

Question #5 : Décomposition des différences observées entre les deux populations concernant l'écart de salaire

- Retraitement des variables

- Transformation logarithmique de la variable à expliquer salaire afin d'induire la normalité et stabiliser la variance des résidus (pour les deux echantillons)
  db_popimmigree$lsalaire <- log(db_popimmigree$salaire)
  db_popfrancais$lsalaire <- log(db_popfrancais$salaire)

- Décomposition agrégée afin d'estimer une équation de salaire pour mesurer l'effet de chaque caractéristique sur le salaire au sein de chaque population, « toutes choses égales par ailleurs »
- Population immigrée
  model_popim <-
  lm(lsalaire~age+sexe+couple+anc+ddipl+cstotr+heures+horaic+nuitc,data=db_popimmigree)
  beta_popim <- coef(model_popim)
  summary(model_popim)
- Population née en France
  model_popfr <-
  lm(lsalaire~age+sexe+couple+anc+ddipl+cstotr+heures+horaic+nuitc,data=db_popfrancais)
  beta_popfr <- coef(model_popfr)
  summary(model_popfr)

- Calcul des parties expliquée et inexpliquée de l'écart de salaire entre la population immigrée et née en France
  var_expli_popim <-
  model.matrix(~age+sexe+couple+anc+ddipl+cstotr+heures+horaic+nuitc,data=db_popimmigree)
  var_expli_popim_moy <- apply(var_expli_popim,2,mean)
  var_expli_popfr <-
  model.matrix(~age+sexe+couple+anc+ddipl+cstotr+heures+horaic+nuitc,data=db_popfrancais)
  var_expli_popfr_moy <- apply(var_expli_popfr,2,mean)
  *exp1 <- (var_expli_popfr_moy-var_expli_popim_moy)*beta_popfr*
  sum(exp1)
  *inexp1 <- (beta_popfr-beta_popim)*var_expli_popim_moy*
  sum(inexp1)
  *exp2 <- (var_expli_popfr_moy-var_expli_popim_moy)*beta_popim*
  sum(exp2)
  *inexp2 <- (beta_popfr-beta_popim)*var_expli_popfr_moy*
  sum(inexp2)

> Interprétation :

a. Comparaison des déterminants du salaire entre la population immigrée et la population née en France

	Population française	Population immigrée
Constante	6,2184***	6,0497***
age	0,0032***	0,0022***
sexe2	-0,089***	-0,0785***
couple	0,0586***	0,0618***
anc	0,0095***	0,009***
ddipl3	-0,0178**	-0,0001

ddipl4	-0,1207***	-0,0835***
ddipl5	-0,165***	-0,0864***
ddipl7	-0,2753***	-0,1735***
cstotr4	-0,1357***	-0,1649***
cstotr5	-0,3004***	-0,3396***
cstotr6	-0,2734***	-0,2934***
heures	0,0343***	0,0393***
horaic2	0,1175***	0,0864***
horaic3	-0,0189***	-0,0373***
nuitc3	-0,0391***	-0,0164***

b. et c. Calcul des parties expliquée et inexpliquée de l'écart de salaire entre les hommes et les femmes

	Modèle 1	Modèle 2
Différence	0,1018	0,1018
Partie expliquée	0,1104 (108%)	0,1096 (108%)
Partie inexpliquée	-0,0086 (-8%)	-0,0078 (-7%)

Si la population d'origine immigrée et la population native française avaient exactement les mêmes caractéristiques individuelles et d'emploi considérées, la population d'origine immigrée devraient percevoir en moyenne des salaires inférieures à la population native française : -11% selon le modèle 1 et -10,9% selon le modèle 2.

Les individus d'origine immigrée perçoivent en moyenne des salaires supérieurs de 0,86% selon le modèle 1 et de 0,76% selon le modèle 2 à ceux des individus en France.

<u>Question #6 : Déterminer la contribution de chaque variable explicative à chacune des 2 parties</u>

- Décomposition détaillée de l'écart de salaire entre la population immigrée et celle née en France
  decompo_detaillee <-
  oaxaca(formula=lsalaire~age+sexe+couple+anc+ddipl+cstotr+heures+horaic+nuitc | poporigine, data=db)
  decompo_detaillee$twofold$variables[[2]]

> Interprétation :

Principaux éléments explicatifs de l'écart de salaire moyen observé entre la population née en France et la population immigrée (pondération population française)

	Partie expliquée	Ecart avec la différence
age	0,00162	1,59
sexe2	- 0,00110	- 1,08
couple	- 0,00014	- 0,14
anc	0,03017	29,63
ddipl3	- 0,00076	- 0,75

ddipl4	- 0,00214	- 2,10
ddipl5	- 0,00962	- 9,44
ddipl7	0,03082	30,26
cstotr5	0,00859	8,43
cstotr6	0,01635	16,05
heures	0,04649	45,65
horaic2	- 0,00007	- 0,07
horaic3	- 0,00046	- 0,45
nuitc3	0,00006	0,06

Contribution positive entre les individus immigrées et les individus nés en France s'explique par l'ancienneté, 29% de l'écart salarial s'explique par le fait que la population née en France a davantage d'ancienneté d'emploi que la population immigrée.

Contribution positive entre les individus immigrées et les individus nés en France s'explique par le nombre d'heures, 45% de l'écart salarial s'explique par le fait que la population née en France a un nombre d'heures travaillées plus élevé que la population immigrée.

Contribution négative entre les individus immigrées et les individus nés en France s'explique par les diplômes de niveau 'CAP' et 'BEP', 9% de l'écart salarial ne s'explique pas par le fait que la population immigrée est davantage sous diplômée que la population née en France.