ACP\_University

Moussa

2024-08-18

##### Import of dataset

library(readxl)  
universite <- read\_excel("~/GitHub/SORADATA/ACP-R/Dataset/universite.xlsx")

## New names:  
## • `` -> `...1`

str(universite)

## tibble [10 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ ...1 : chr [1:10] "Droit, sciences politiques" "Sciences économiques, gestion" "Administration économique et sociale" "Lettres, sciences du langage, arts" ...  
## $ Licence-F : num [1:10] 69373 38387 18574 48691 62736 ...  
## $ Licence-H : num [1:10] 37317 37157 12388 17850 21291 ...  
## $ Master-F : num [1:10] 42371 29466 4183 17672 13186 ...  
## $ Master-H : num [1:10] 21693 26929 2884 5853 3874 ...  
## $ Doctorat-F: num [1:10] 4029 1983 0 4531 1839 ...  
## $ Doctorat-H: num [1:10] 4342 2552 0 2401 907 ...  
## $ Total-F : num [1:10] 115773 69836 22757 70894 77761 ...  
## $ Total-H : num [1:10] 63352 66638 15272 26104 26072 ...  
## $ Licence : num [1:10] 106690 75544 30962 66541 84027 ...  
## $ Master : num [1:10] 64064 56395 7067 23525 17060 ...  
## $ Doctorat : num [1:10] 8371 4535 0 6932 2746 ...  
## $ Total : num [1:10] 179125 136474 38029 96998 103833 ...

#### Les statistiques descriptives

summary(universite)

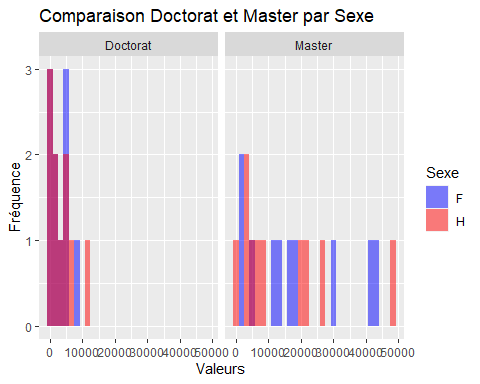
## ...1 Licence-F Licence-H Master-F   
## Length:10 Min. : 1779 Min. : 726 Min. : 1963   
## Class :character 1st Qu.:19570 1st Qu.:15566 1st Qu.: 5910   
## Mode :character Median :31353 Median :19571 Median :15132   
## Mean :38901 Mean :25490 Mean :18238   
## 3rd Qu.:59225 3rd Qu.:37277 3rd Qu.:26518   
## Max. :94346 Max. :54861 Max. :43016   
## Master-H Doctorat-F Doctorat-H Total-F   
## Min. : 811 Min. : 0.0 Min. : 0.0 Min. : 4148   
## 1st Qu.: 3948 1st Qu.: 600.8 1st Qu.: 472.8 1st Qu.: 27330   
## Median : 7155 Median :3006.0 Median : 2476.5 Median : 56940   
## Mean :14341 Mean :3041.8 Mean : 3424.0 Mean : 60181   
## 3rd Qu.:21382 3rd Qu.:4500.0 3rd Qu.: 5009.5 3rd Qu.: 76044   
## Max. :48293 Max. :7787.0 Max. :11491.0 Max. :145149   
## Total-H Licence Master Doctorat   
## Min. : 1552 Min. : 2505 Min. : 3167 Min. : 0   
## 1st Qu.: 22833 1st Qu.: 33052 1st Qu.: 9565 1st Qu.: 1074   
## Median : 27399 Median : 71043 Median :21536 Median : 5734   
## Mean : 43255 Mean : 64391 Mean :32579 Mean : 6466   
## 3rd Qu.: 65817 3rd Qu.: 82375 3rd Qu.:61696 3rd Qu.:10248   
## Max. :114645 Max. :135396 Max. :65371 Max. :15898   
## Total   
## Min. : 5700   
## 1st Qu.: 45957   
## Median :100416   
## Mean :103436   
## 3rd Qu.:153135   
## Max. :213618

aggregate(Master~Doctorat, data = universite, FUN = mean)

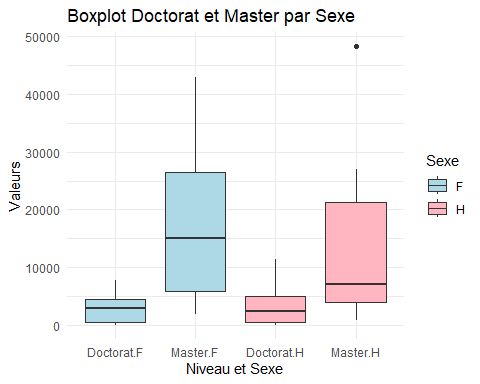
## Doctorat Master  
## 1 0 7067  
## 2 28 3167  
## 3 516 6135  
## 4 2746 17060  
## 5 4535 56395  
## 6 6932 23525  
## 7 8371 64064  
## 8 10873 19547  
## 9 14759 63463  
## 10 15898 65371

library(ggplot2)  
  
# Préparer les données pour ggplot  
universite\_long <- tidyr::pivot\_longer(universite, cols = c(`Doctorat-F`, `Doctorat-H`, `Master-F`, `Master-H`),   
 names\_to = c("Niveau", "Sexe"), names\_sep = "-", values\_to = "Valeurs")  
  
# Créer l'histogramme avec ggplot pour Doctorat et Master  
ggplot(universite\_long, aes(x = Valeurs, fill = Sexe)) +   
 geom\_histogram(alpha = 0.5, position = "identity") +  
 facet\_wrap(~Niveau) +  
 labs(title = "Comparaison Doctorat et Master par Sexe", x = "Valeurs", y = "Fréquence") +  
 scale\_fill\_manual(values = c("blue", "red", "green", "orange"))

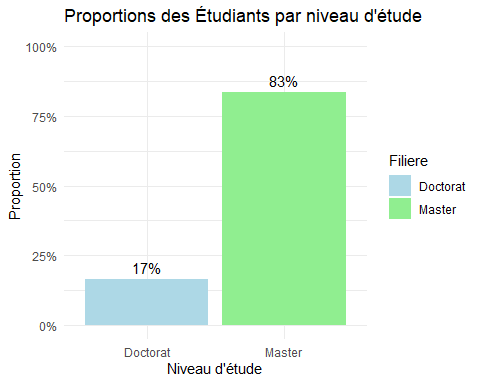
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



library(ggplot2)  
  
# Préparer les données au format long  
universite\_long <- tidyr::pivot\_longer(universite, cols = c(`Doctorat-F`, `Doctorat-H`, `Master-F`, `Master-H`),   
 names\_to = c("Niveau", "Sexe"), names\_sep = "-", values\_to = "Valeurs")  
  
# Créer le boxplot avec ggplot  
ggplot(universite\_long, aes(x = interaction(Niveau, Sexe), y = Valeurs, fill = Sexe)) +  
 geom\_boxplot() +  
 labs(title = "Boxplot Doctorat et Master par Sexe", x = "Niveau et Sexe", y = "Valeurs") +  
 scale\_fill\_manual(values = c("lightblue", "lightpink")) +  
 theme\_minimal()



# Calculer le total pour chaque filière  
total\_doctorat <- sum(universite$`Doctorat-F`) + sum(universite$`Doctorat-H`)  
total\_master <- sum(universite$`Master-F`) + sum(universite$`Master-H`)  
  
# Calculer le total général  
total\_global <- total\_doctorat + total\_master  
  
# Calculer les proportions  
proportion\_doctorat <- total\_doctorat / total\_global  
proportion\_master <- total\_master / total\_global  
  
# Créer un vecteur des proportions  
proportions <- c(proportion\_doctorat, proportion\_master)  
  
# Créer un vecteur des noms des filières  
filieres <- c("Doctorat", "Master")  
library(ggplot2)  
  
# Préparer les données dans un data frame  
proportions\_df <- data.frame(  
 Filiere = filieres,  
 Proportion = proportions  
)  
  
# Créer le barplot avec ggplot2  
ggplot(proportions\_df, aes(x = Filiere, y = Proportion, fill = Filiere)) +  
 geom\_bar(stat = "identity") +  
 scale\_y\_continuous(labels = scales::percent, limits = c(0, 1)) +  
 labs(title = "Proportions des Étudiants par niveau d'étude", x = "Niveau d'étude", y = "Proportion") +  
 scale\_fill\_manual(values = c("lightblue", "lightgreen")) +  
 theme\_minimal() +  
 geom\_text(aes(label = scales::percent(Proportion)), vjust = -0.5)



#### ACP via FactomineR

library(FactoMineR)  
library("factoextra")

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

# Sélectionner uniquement les colonnes numériques  
numerical\_vars <- sapply(universite, is.numeric)  
  
# Créer un sous-data frame avec uniquement les colonnes numériques  
universite\_numeric <- universite[, numerical\_vars]

resultat\_ACP<-PCA(universite\_numeric, graph = FALSE)  
print(resultat\_ACP)

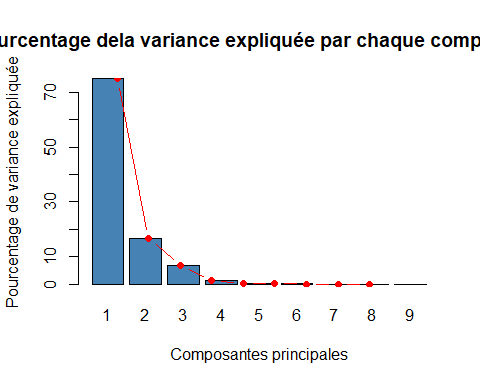
## \*\*Results for the Principal Component Analysis (PCA)\*\*  
## The analysis was performed on 10 individuals, described by 12 variables  
## \*The results are available in the following objects:  
##   
## name description   
## 1 "$eig" "eigenvalues"   
## 2 "$var" "results for the variables"   
## 3 "$var$coord" "coord. for the variables"   
## 4 "$var$cor" "correlations variables - dimensions"  
## 5 "$var$cos2" "cos2 for the variables"   
## 6 "$var$contrib" "contributions of the variables"   
## 7 "$ind" "results for the individuals"   
## 8 "$ind$coord" "coord. for the individuals"   
## 9 "$ind$cos2" "cos2 for the individuals"   
## 10 "$ind$contrib" "contributions of the individuals"   
## 11 "$call" "summary statistics"   
## 12 "$call$centre" "mean of the variables"   
## 13 "$call$ecart.type" "standard error of the variables"   
## 14 "$call$row.w" "weights for the individuals"   
## 15 "$call$col.w" "weights for the variables"

### Le choix de l’axe ou de dimension

valeurspropes<-resultat\_ACP$eig  
valeurspropes

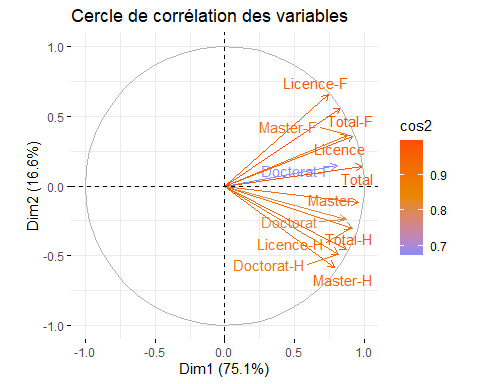
## eigenvalue percentage of variance cumulative percentage of variance  
## comp 1 9.008252e+00 7.506876e+01 75.06876  
## comp 2 1.988651e+00 1.657209e+01 91.64086  
## comp 3 8.032154e-01 6.693461e+00 98.33432  
## comp 4 1.680335e-01 1.400279e+00 99.73460  
## comp 5 2.304292e-02 1.920243e-01 99.92662  
## comp 6 8.805451e-03 7.337875e-02 100.00000  
## comp 7 1.985892e-31 1.654910e-30 100.00000  
## comp 8 1.112184e-32 9.268203e-32 100.00000  
## comp 9 2.901728e-33 2.418106e-32 100.00000

barplot(valeurspropes[,2],names.arg = 1:nrow(valeurspropes),  
 main = "Pourcentage dela variance expliquée par chaque composante",  
 xlab = "Composantes principales",  
 ylab = "Pourcentage de variance expliquée",  
 col = "steelblue")  
lines(x=1:nrow(valeurspropes),valeurspropes[,2],  
 type = "b",pch=19,col="red")



### Le cercle de corrélation ou il exite un effet de taille

fviz\_pca\_var(resultat\_ACP,  
 col.var = "cos2", # Coloration par le cosinus carré  
 gradient.cols = c("#888AFB","#E78800","#FC4E07"), # Définir les couleurs du gradient  
 repel = TRUE, # Éviter le chevauchement des étiquettes  
 title = "Cercle de corrélation des variables") # Titre du graphique



fviz\_pca\_ind(resultat\_ACP,   
 col.ind = "cos2", # Coloration en fonction du cos2  
 gradient.cols = c("blue", "white", "red"), # Palette de couleurs  
 repel = TRUE) + # Éviter le chevauchement des étiquettes  
 scale\_color\_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0.50) +  
 theme\_minimal() +  
 labs(title = "Visualisation des individus selon Cos2",  
 color = "Cos2")

## Scale for colour is already present.  
## Adding another scale for colour, which will replace the existing scale.

