PROJET AVEC R

****

Analyse des données du naufrage du Titanic (1912).

$\Longrightarrow$ **REALISER PAR :**

▷**Prènom et Nom : ZAKARIA ELHARCHAOUI**

Description des données

- **1) La Premiére question** :

  Utilisation du data frame train contient un échantillon de passagers du Titanic

  load(”C:/Users/admin/Documents/titanic_train.Rdata”)

  View(train)

- **2) La Deuxiéme question** :

  – 2-1) le nombre d’observations:

  nrow(train) ou bien:

  nobservation=dim(train)[1]

  En tapant : nobservation

  RESULTAT: 594

  – 2-2) le nombre de variables:

  ncol(train) ou bien :

  nvariable=dim(train)[2]

  En tapant : nvariable

  RESULTAT: 12

  – 2-3) le nom des variables:

  names(train)

  RESULTAT:

  ”PassengerId” ”Survived” ”Pclass” ”Name” ”Sex” ”Age” ”SibSp” ”Parch” ”Ticket” ”Fare” ”Cabin” ”Embarked”

  – 2-4) Types des variables quantitatives ou qualitatives:

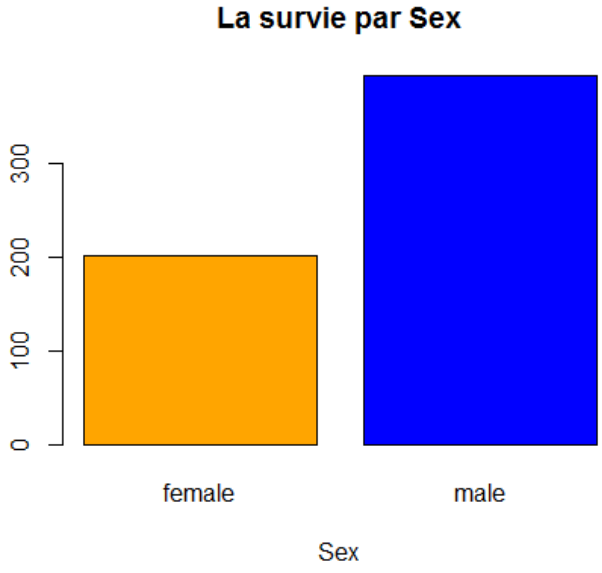
Survived	**Pclass**	**Name**
Qualitative	Qualitative	Qualitative
Age	**Sibsp**	**Parch**
Quantitative	Qualitative	Qualitative
Fare	**Cabin**	**Embarked**
Quantitative	Qualitative	Qualitative
Sex	**Ticket**	**PassengerId**
Qualitative	Qualitative	Qualitative

Table types des variables.

– 2-5) Le nombre de valeurs manquantes
En Utilisant la commande:
table(is.na(train))
On obtient le total de valeurs manquantes
RESULTAT:FALSE : 6543 TRUE: 585
Donc 585 valeurs manquantes
En utilisant la commande
summary(train)
on retrouve le nombre de valeurs manquantes pour chaque variable.
Ce qui nous permet de relever les deux variables
table (is.na(train$Cabin))
RESULTAT:FALSE : 131 TRUE: 463
Cabin avec 463 valeurs manquantes
table (is.na(train$Age))
RESULTAT:FALSE : 473 TRUE: 121
La variable Age contient 121 valeurs manquantes.

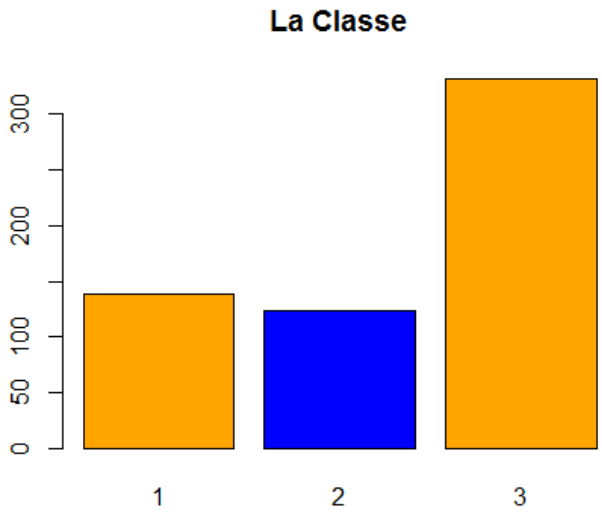
- **3) La Troisiéme question** :

– 3-1)décrire la variable Sex
On exécutons la commande suivante:
summary(train$Sex)
on retrouve le nombre des hommes et des femmes :
female 201 male 393
barplot(table(train$Sex),main="La survive par Sex",xlab="Sex",col=c("orange","blue"))



avec la commande :
prop.table(table(train$Sex))
female 0.3383838 male 0.6616162
Nous remarquons en analysant ces résultats que les hommes
représentent 66.1% et les femmes représentes 33.8% du total des passagers.

– 3-2)décrire la variable Pclasse
On exécute les commandes suivantes:
summary(train$Pclass)
On retrouve la fréquence de chaque classe par le
résultat suivant :
classe 1 :0.2340067 classe 2 2 :0.2087542 classe3 3 :0.5572391
Nous remarquons en analysant ces résultats que la
classe 3 représente 55.7% du total des passagers,
suivie de la Classe 1 23.4%
pour afficher l'histogramme relatif à la distribution de la
variable survie.On utilise la commande suivante:
barplot(table(train$Pclass),main="La Classe",col=c("orange","blue"))



– 3-3)décrire la variable Age
En exécutant la commande suivantes:
summary(train$Age)
On retrouve les caractéristiques suivantes :

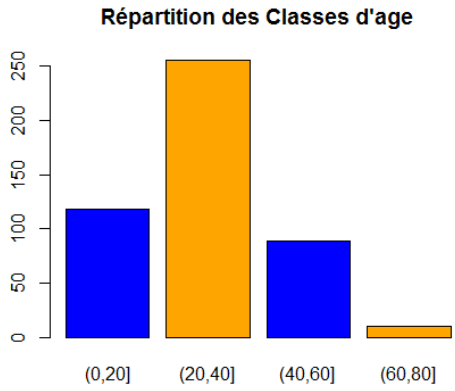
```
> summary(train$cAge)
 (0,20] (20,40] (40,60] (60,80]    NA's
    118     255      89      11     121
```

La moyenne des ages est de 29.58 ans. Les ages sont répartis
entre 0.75(minimum) et 71 ans(maximum).
On calcule l'ecart-type qui mésure la dispersion autour de la moyenne :
sd(table(train$Age),na.rm = T)
l'ecart-type est de : 5.330236.
barplot(table(train$Age))
0.2037037
sum(is.na(train$Age))/nrow(train)
les valeurs manquantes représentent 20.3% des modalités de la variable :"Age".

- **4) La Quatriéme question** :
  - 4-1) La nouvelle variable cAge qui catégorises Age à l'aide de la fonction cut()

  Pour Construire une nouvelle variable cAge qui catégorise
  l'Age selon 4 Fractions,
  on utilise la commande suivante:
  train$cAge=cut(train$Age,breaks=(0:4)*20)
  summary(train$cAge)
  (0,20] (20,40] (40,60] (60,80] NA's
  118 255 89 11 121
  en proportion:
  (0,20] (20,40] (40,60] (60,80]
  24 53 18 2

  Nous remarquons que les passagers Agés entre 20 et 40 ans
  représentent plus que la moitié des passagers avec une
  proportion de 53%, suivis des de la tranche[0,20[avec 24%. cette nouvelle variable est additionnée à notre base de
  données "train". Nous pouvons la visualiser on utilisant la
  commande suivante:
  barplot(table(train$cAge),main="Répartition des Classes d'age",col=c("blue","orange"))



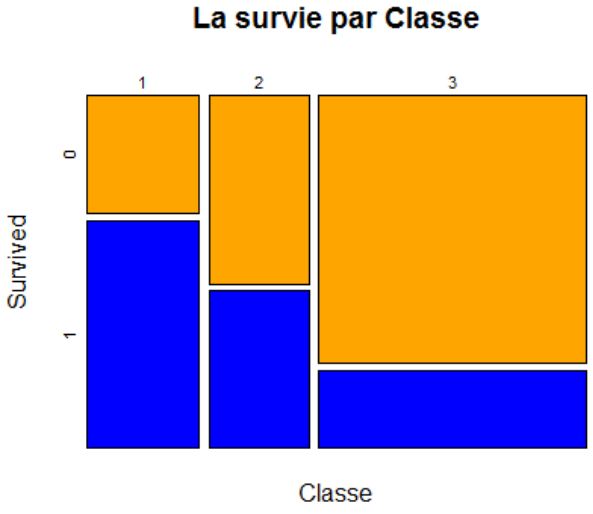
Liens entres les variables

- **5) La Cinquième Question** :
  - 5.1)Lien entre P et S:

    table(train$Pclass,train$Survived)

    plot(table(train$Pclass,train$Survived))

    Le resultat affiché :



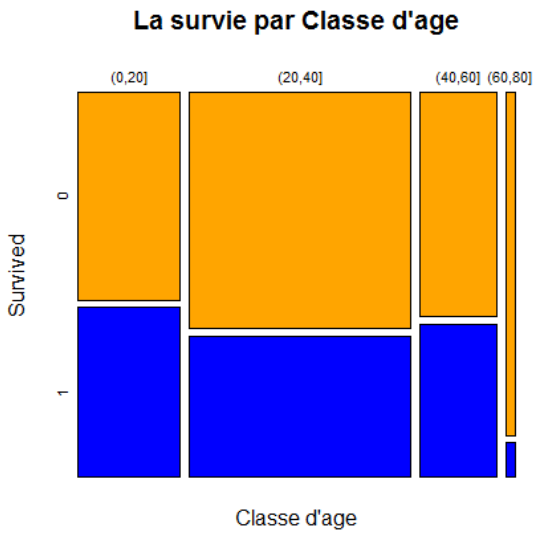
Interprétation : Ce sont les passagers de la classe 1 qui ont le plus survécu (91 personnes) .Par contre 257 personnes de la classe 3 n'ont pas survécu.Nous supçonnons que le fait d'etre de la classe 3 diminue ta changer de survivre.

– 5.2)Lien entre A et S

table(train$Age, train$Survived)

plot(table(train$Age, train$Survived))

Le resultat affiché :



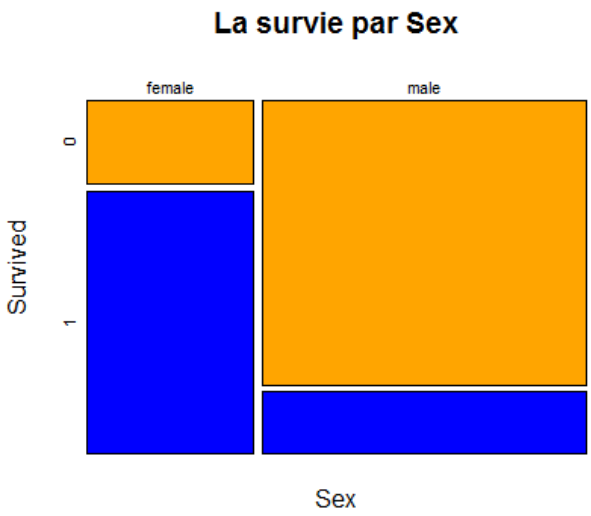
	0	1
(0,20]	65	53
(20,40]	160	95
(40,60]	53	36
(60,80]	10	1

Interprétation :les passagers adultes entre (20,40] ont connu le nombre de non survecu le plus élevé 160 personnes contre 95 survecu.

– 5.3)Lien entre Sx et S

table(train$Sex, train$Survived)

plot(table(train$Sex,train$Survived),main="La survie par Sex",xlab="Sex", ylab="Survived",col=c("orange","blue"))



Interprétation: les passagers de sexe féminin ont survécu avec un nombre de 152 contre seulement 69 hommes survivants.

- 6) **sixiéme question:**

  Ce jeu de données représente le nombre de passagers du Titanic, répartis par Classe, Sex et par Age. La première hypothèse : Nous supposons que la probabilité de survie d'un passager au bord de titanic dépend significativement des variables : sex , Classe et l'Age. Autrement dit , ces 3 variables arrivent à expliquer la probabilité de servivre. Cette hypothèse est basée sur les constats suivants:

  Les femmes de la classe 1 ont les meilleures chances de survie, suivi par les femmes de la classe 2 m et ensuite les femmes de la classe 3. Enfin, on prévoit que l'âge sera négativement liée à la probabilité de survie l'age entre (60,80] chances de survie étaient trop faibles. Nous chercherons donc la modélisation de la probabilité de survie en fonction des probabilité (fréquence en l'occurence) des 3 variables précitées.

Prédiction de la survie

- **7)La septième question:**
  - 7-1) Problabilité de survie sachant sex

  prop.table(table(train$Sx, train$S),1)

  LA RESULTAT ‘

```
                0         1
female 0.2437811 0.7562189
male   0.8244275 0.1755725
```

  $\mathbb{P}$(S=1|Sx=female) = 75%

  $\mathbb{P}$(S=1|S=male) = 17%

  - 7-2) Probabilité de survie sachant la classe

  prop.table(table(train$P, train$S),1)

  LA RESULTAT ‘

```
          0         1
1 0.3453237 0.6546763
2 0.5483871 0.4516129
3 0.7764350 0.2235650
```

  $\mathbb{P}$(S=1|P=1) = 65%

  $\mathbb{P}$(S=1|P=2) = 45%

  $\mathbb{P}$(S=1|P=3) = 22%

  - 7-3)Probabilité de survie sachant Classe Age

  prop.table(table(train$cA, train$S),1)

  LA RESULTAT ‘

```
                0         1
(0,20]  0.55084746 0.44915254
(20,40] 0.62745098 0.37254902
(40,60] 0.59550562 0.40449438
(60,80] 0.90909091 0.09090909
```

$\mathbb{P}$(S=1|cAge=(0,20]) = 45%

$\mathbb{P}$(S=1|cAge=(20,40]) = 37%

$\mathbb{P}$(S=1|cAge=(40,60]) = 40%

$\mathbb{P}$(S=1|cAge=(60,80]) = 0.9%

- **8) La huitieme question :**

  – **8-1) Creation de la variable S_P la probabilité de la survie sachant Pclass**

  S_P=prop.table(table(train$Pclass,train$Survived),margin=2)

  -noms aux lignes et aux colonnes pour faciliter l'accès

  colnames(S_P)=c('Unsurvived','survived') rownames(S_P)=c('Classe 1','Classe 2','Classe 3')

```
          Unsurvived  survived
Classe 1  0.1286863 0.4117647
Classe 2  0.1823056 0.2533937
Classe 3  0.6890080 0.3348416
```

```
> S_P['Classe 3','survived']
[1] 0.3348416
```

  – **8-2) Creation de la variable S_Sx la probabilité Survie sachant Sex**

  S_Sx=prop.table(table(train$Sex,train$Survived),margin=2)

  colnames(S_Sx)=c('Unsurvived','survived')

```
> S_Sx
          Unsurvived  survived
female  0.1313673 0.6877828
male    0.8686327 0.3122172
```

  –**8-3) Creation de la variable S_Ca la probabilité Survie sachant cAge**

  S_Ca=prop.table(table(train$cAge,train$Survived),margin=2)

  colnames(S_Ca)=c('Unsurvived','survived')

```
          Unsurvived    survived
(0,20]  0.225694444 0.286486486
(20,40] 0.555555556 0.513513514
(40,60] 0.184027778 0.194594595
(60,80] 0.034722222 0.005405405
```

– 8-4) Construire la table S

S=prop.table(table(train$Survived))

names(S)=c('Unsurvived','survived')

```
> S
Unsurvived   survived
 0.6279461  0.3720539
```

- 9) La Neuviéme question :

9-1)prob_prediction(Sx, P, cAge)

```
prob_prediction=function(Sx,P,cAge){

classage=ceiling(cAge/20)
return ((S_Sx[Sx,"survived"]*S_P[P,"survived"]*S_Age[classage,"survived"]*probsurv[2])
        /((S_Sx[Sx,"survived"]*S_P[P,"survived"]*probsurv[2]*S_Age[classage,"survived"] )+
            (S_Sx[Sx,"Unsurvived"]*S_P[P,"Unsurvived"]*probsurv[1]*S_Age[classage,"Unsurvived"])))}

```

```
> prob_prediction("female",1,23)
        1
0.9017163
```

Nous testons notre formule de probabilité afin de vérifier si l'hypothèse suposée est valide.

$\mathbb{P}$(S=1|Sx=female,P=1,cAge=23) = 0.901.

Effectivement les femmes de classe 1 ont une probabilité de 90% de survivre.

```
> prob_prediction("male",3,55)
          1
0.09864242
```

$\mathbb{P}$(S=1|Sx=male,P=3,cAge=55) = 0.09.

les hommes de classe 3 ont une probabilité de 0.9% de survivre.