Analyse des jeux de données du Naufrage du Titanic

Analyse statistique des données selon les principices de la programmation lettrée

Thierno Ibrahima Sory Diallo, Mamadou Dian Diallo

March 13, 2018

Contents

1 Introduction

1.1 Context et Description des jeux de données

Le naufrage du RMS Titanic est l'un des naufrages les plus infâmes de l'histoire. Le 15 avril 1912, lors de son voyage inaugural, le Titanic a coulé après avoir heurté un Iceberg, tuant 1502 des 2224 passagers et membres d'équipage. Cette tragédie sensationnelle a choqué la communauté internationale et conduit à de meilleures règles de sécurité pour les navires.

L'une des raisons pour lesquelles le naufrage a entraîné une telle perte de vie était qu'il n'y avait pas assez de canots de sauvetage pour les passagers et l'équipage. Bien qu'il y ait eu de la chance à survivre au naufrage, certains groupes de personnes avaient plus de chances de survivre que d'autres, comme les femmes, les enfants et la classe supérieure.

Dans ce défi, nous allons compléter l'analyse des types de personnes susceptibles de survivre. En particulier, nous allons appliquer la programmation lettrée à travers l'outil RStudio pour derminer les chances de survie des différents groupes de personne.

Ajout des packages pour la construction de l'environnement de simulation

```
#Package
library(dplyr);
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2);
library(gridExtra);
```

```
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(scales);
```

1.2 Demarche d'acquisition des données

Dans le cadre de ce mini projet nous avons eu a nous posez un certains nombre de question sur les sujets d'études les plus pertinents pour nous et la question sur laquelle sera basé l'analyse des jeux de données qui seront recueillis. A la suite de cela, il y avait lieu d'aller sur un certains nombre de site dédiés, dont les liens sont indiqués dans la liste des références pour chercher les jeux données.

1.3 Question sur la quelle se base l'analyse

Quelles sont les chances de survie de tous les groupes de personnes lors du naufrage du Titanic en fonction de l'âge, du sexe et de la classe.

2 Methodes de travail

La premiere des choses a été d'analyser les données à fin de savoir comment mener notre étude. Cette premier phase d'analyse à porter essentiellement sur les variables clés(le sexe, la classe et l'âge) des passagers qui vont nous permettre de determiner les chance de survie de chaque groupe de personne lors du nauffrage. Après cette premier phase, il y'avait lieu de se poser la question de savoir est ce que les critères d'étude choisis sont pertinantes et a t on des données suffisamentes et complets nous permettant d'aboutir à un Résultat vrai.En suite on s'est penché sur les types de graphes qu'on pouvait utiliser en fonction des questions auquel on voulais des reponses et comment faire ces différentes representations.En fin on a choisis les différents types de répresentations à utiliser.

2.1 Choix de la façon dont les données doivent être représenter

On a choisis d'uiliser un certains nombre de répresentations à travers des fonctions proposées par ggplot2:

- Nuage de points avec la fonction : geom_point()
- Histogramme : geom_histogram()
- Polygone de fréquenc : geom_freqpoly()

2.2 Dictionnaire de données

- Name : Nom et prénom du passager.
- PClass : La catégorie de classe d'embarquement dans le Bateau : 1st: 1er Classe, 2nd: 2ème classe , 3rd: 3ème classe.
- Age : L'âge du passager.
- Sex : Le genre " Féminin ou Masculin "
- Survived : Survivants du Naufrage, si Survived=1 signifie que le passager a survécu, si Survived=0 il n'a pas survécu.
- SexCode : Codification du genre, si SexCode=1 alors le passager est de genre féminin et si SexCode=0 Masculin.

3 Analyse Statistique (Programmation lettrée)

3.1 Chargement des données

Le chargement des jeux de données est effectué à partir du fichier titanic.csv. Après chargement il crée un data frame "df" dans laquelle toutes les données seront stoquées selon les différents champs (Name, PClass, Age, Sex, Survived ,SexCode) qu'on pourra afficher à l'aide de la fonction "head()".

```
df <- read.csv("titanic.csv", head=TRUE )
head(df)
```

N	Name	PClass	Age	Sex	Survived	SexCode
1	Allen, Miss Elisabeth Walton	1st	29.00	female	1	1
2	Allison, Miss Helen Loraine	1st	2.00	female	0	1
3	Allison, Mr Hudson Joshua Creighton	1st	30.00	male	0	0
4	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25.00	female	0	1
5	Allison, Master Hudson Trevor	1st	0.92	male	1	0
6	Anderson, Mr Harry	1st	47.00	male	1	0

3.2 Traitement des données

Utilisation de la fonction "summary()" pour obtenir un résumé détaillé de la distribution de la variable "Age" du jeux de données.

```
summary(df$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.17   21.00   28.00   30.40   39.00   71.00     557
```

On constate que l'âge n'est pas connu pour près de la moitié des données de ce fait cela impact forcément notre étude s'il est fait en fonction de l'âge.

Création d'une nouvelle variable pour étudier l'impact des données "Age" non connues pour certains passagers.

```
df$MissingAge<- ifelse(is.na(df$Age),"Y","N")
```

Création d'une variable et initialisation de notre graphe "Nuage de points" dans lequel on pour ajouter de nouvelle fonction.

```
survie <- ggplot(df, aes(x=Sex, y=PClass, colour=Survived, shape=factor(Survived)) )+
theme_bw()
```

Ajout de la fonction "geom_point()" pour générer le graphe, la fonction "ggtitle()" pour lui donner un titre et print() pour sa visualisation.

```
survie <- survie + geom_point(size=6)
print(survie)
```

A travers ce graphe on constate que les passagers qui étais en première classe et qui étaient de sexe féminin avaient plus de chance de survie que ceux de la deuxième et troisième classe et encore moins de chance si on est de sexe masculin.

- Histogramme (PClass , Sex) Graphe pour la chance de survie des passagers en fonction de la classe et du sexe.

```
survie <- ggplot(df, aes(x=Sex, y=PClass, fill=factor(Survived)) ) +geom_histogram(stat='identity', p
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
survie <- survie +theme_bw()
print(survie)
```

- Histogramme (survived)

Figure 1: Nuage de points

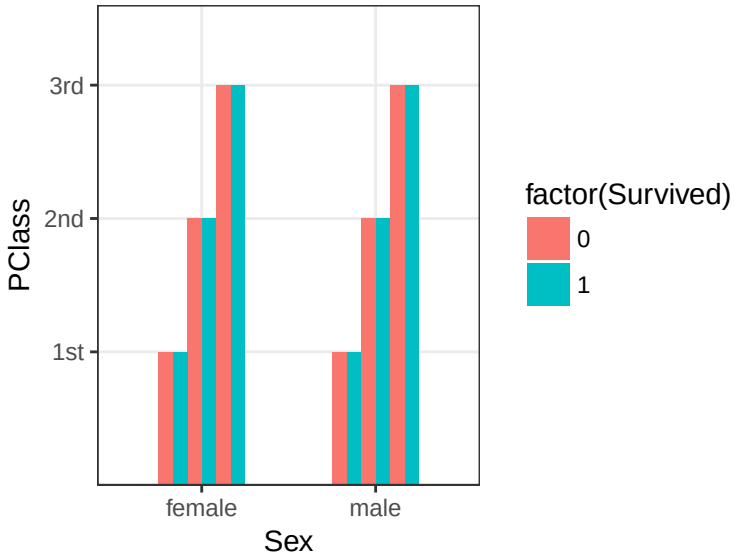

Figure 2: Histogramme

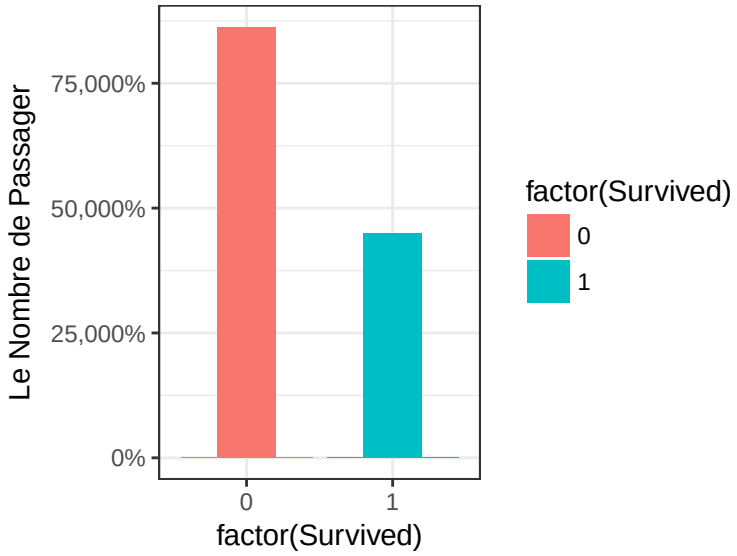

Figure 3: Histogramme

```
 # Calcul de nombre de survivant du titanic
  ggplot(df, aes(x=factor(Survived), fill=factor(Survived))) +
  theme_bw()+
  geom_bar(aes(y = (..count..)/sum(..count..))) +
  scale_y_continuous(labels = percent) +
  geom_bar(width=.4)+
  labs(y="Le Nombre de Passager")
```

- Calcul du pourtange de survie chez les hommes et chez le femme en fonction du sexe

```
# Pourcentage de survivant du Titanic
  prop.table(table(df$Sex,df$Survived))
```

```
##
##                  0         1
##   female 0.1172887 0.2345773
##   male   0.5399848 0.1081493
```

```
# Survivants : 447 personnes
## Homme :
## Femme :
# Morts : 866 personnes
## Homme :
## Femme :
```

On constate qu'il eu 34% de survivant en fonction du sexe dont 23% de femme et 11% d'homme et 66% qui n'ont pas survecu dont 12% de femme et 54% d'homme.

- Calcul du pourtange de survie chez les hommes et chez le femme en fonction de la categorie de classe qu'ils ont achete.

```
# Pourcentage de survivant du Titanic
  prop.table(table(df$PClass,df$Survived))
```

```
##
##                  0          1
##   1st 0.09900990 0.14699162
##   2nd 0.12185834 0.09063214
##   3rd 0.43640518 0.10510282
```

```
# Survivants : 447 personnes
## Homme :
```

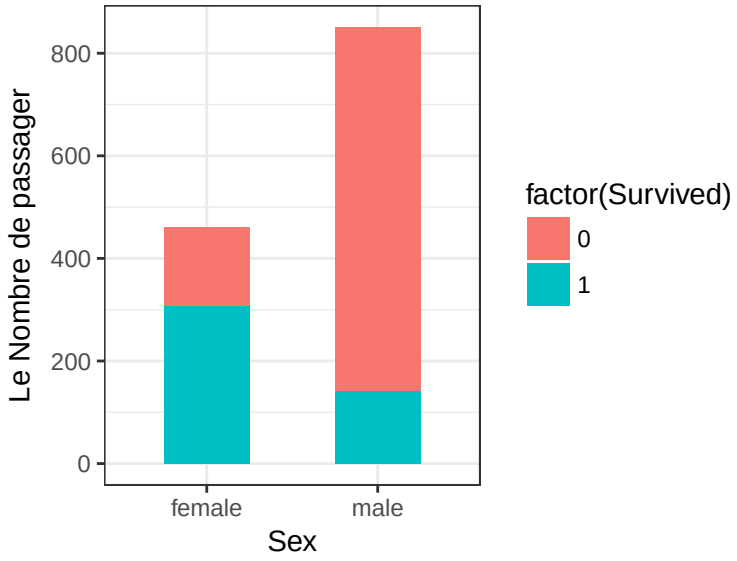

Figure 4: Histogramme

```
## Femme :
# Morts : 866 personnes
## Homme :
## Femme :
```

On constate qu'il eu 34% de survivant en fonction de la categorie de classe dont 15% pour la 1er classe, 9% pour la 2è classe, 10% pour la 3è classe et 66% qui n'ont pas survecu dont 10% pour la 1è classe, 12% pour la 2è classe et 44% pour la 3è classe.

- Calcul du pourtange de survie en fonction de la categorie de classe et du Sexe.

```
# Pourcentage de survivant du Titanic
  prop.table(table(df$Sex,df$PClass,df$Survived))
```

```
## , ,  = 0
##
##
##                  1st         2nd         3rd
##   female 0.006854532 0.009900990 0.100533130
##   male   0.092155369 0.111957350 0.335872049
##
## , ,  = 1
##
##
##                  1st         2nd         3rd
##   female 0.102056359 0.071591775 0.060929170
##   male   0.044935263 0.019040366 0.044173648
```

```
# Survivants : 447 personnes
# Morts : 866 personnes
# Femmes :302
# Hommes : 145
```

- Histogramme (Sex) Graphe pour la chance de survie des passagers en fonction de la Sexe.

```
  ggplot(df, aes(x = Sex, fill=factor(Survived))) +
  theme_bw()+
  geom_bar(width = .5)+
  labs(y="Le Nombre de passager")
```

- Histogramme (Pclass) Graphe pour la chance de survie des passagers en fonction de la classe.

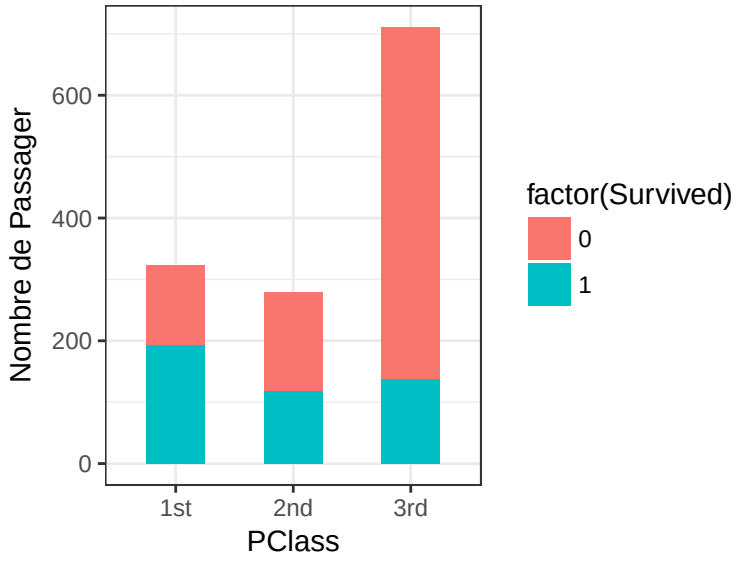

Figure 5: Histogramme

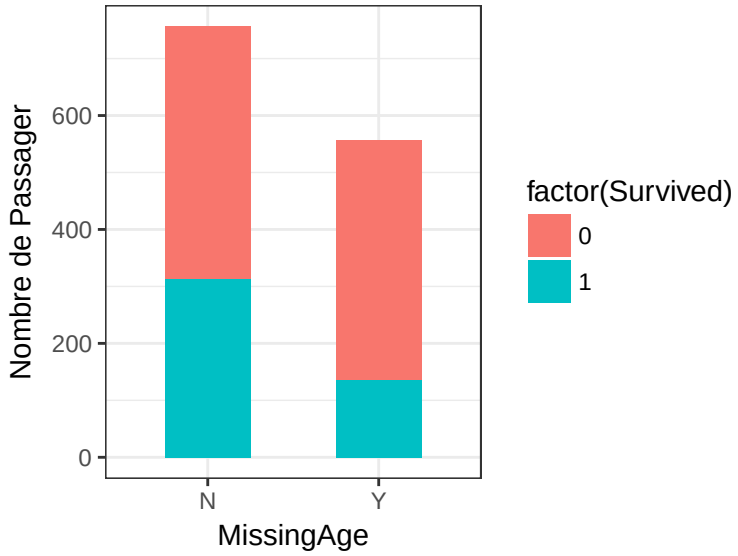

Figure 6: Histogramme

```
ggplot(df, aes(x=PClass, fill=factor(Survived))) +
theme_bw()+
geom_bar(width=.5)+
labs(y="Nombre de Passager")
```

- Histogramme (MissingAge) Graphe pour des passagers dont n'est pas connu

```
ggplot(df, aes(x=MissingAge, fill=factor(Survived))) +
theme_bw()+
geom_bar(width = .5)+
labs(y="Nombre de Passager")
```

Cette courbe montre qu'il y'a près de la moitié des passagers dont l'âge n'est pas connues.

- Histogramme(Age)

```
ggplot(df, aes(x=Age, fill=factor(Survived))) +
theme_bw()+
geom_histogram(position="dodge")+
labs(y="Nombre de Passager")
```

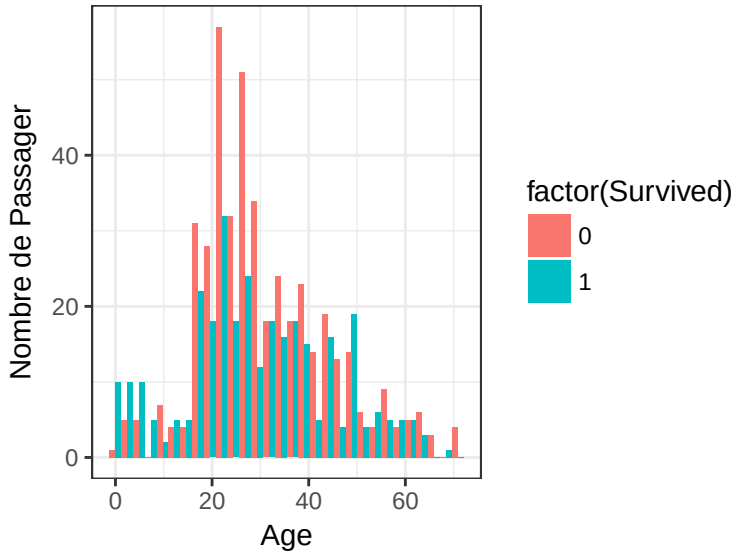

Figure 7: Histogramme

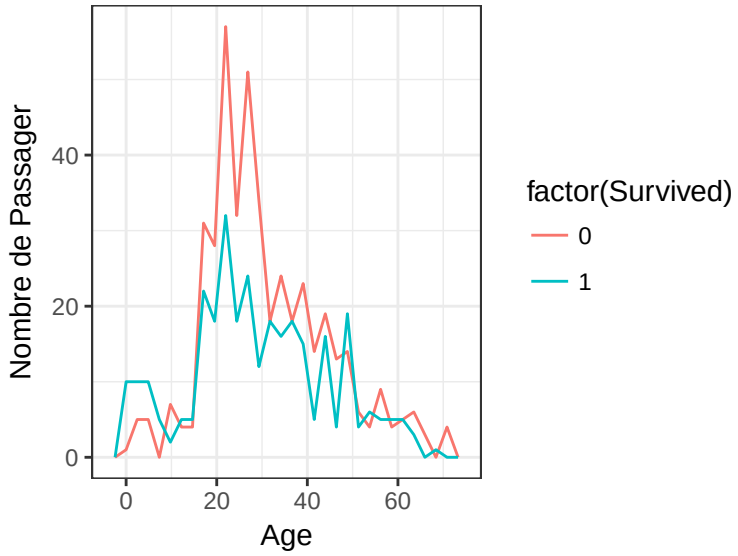

Figure 8: Histogramme

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 557 rows containing non-finite values (stat_bin).
```

Cette courbe montre que les personnes âgés de 18 ans de 40 ans avaient plus de chance de survie que ceux en dessus de 30 ans en ne tenant pas compte des passagers dont on ne connais pas l'âge.

- Polygone des fréquences

```
ggplot(df, aes(x=Age, colour=factor(Survived))) +
theme_bw()+
geom_freqpoly()+
labs(y="Nombre de Passager")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 557 rows containing non-finite values (stat_bin).
```

Cette courbe montre que les personnes âgés de 18 ans de 40 ans avaient plus de chance de survie que ceux en dessus de 30 ans en ne tenant pas compte des passagers dont on ne connais pas l'âge.

4 Conclusion

Pour déterminer la chance de survie des différents groupes de passagers, tenir compte du critère âge comme prévue au début va un peu fausser l'étude, vu que l'âge de plus de la moitié des passagers n'est pas connu. De ce fait on s'est plus pencher sur la classe et le sexe et ces deux critères montrent que les passagers de sexe féminin qui étaient en première classe avait plus de chance de survie que les autres.

5 Références

Documentation RStudio :

- https://www.rstudio.com/resources/cheatsheets/

Documentation de R Markdown :

- https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf
- https://rmarkdown.rstudio.com/authoring_basics.html
- http://archimede.mat.ulaval.ca/dokuwiki/lib/exe/fetch.php?media=r:communication:redactiondocumentsrmarkdown2017.pdf
- https://rmarkdown.rstudio.com/pdf_document_format.html#table_of_contents

Lien jeux de données :

- **FR Data.gouv :** https://www.data.gouv.fr/fr/
- **Insee :** https://www.insee.fr/fr/accueil
- **Kaggle :** https://www.kaggle.com/datasets
- **US Data.gov :** https://catalog.data.gov/dataset