Les passagers du Titanic

Le jeu de données **Titanic** est l'un des ensembles de données les plus emblématiques en science des données. Il provient de la base publique mise à disposition par **Kaggle**, et contient des informations sur les passagers du RMS Titanic, célèbre paquebot ayant fait naufrage lors de son voyage inaugural en avril 1912.

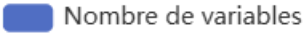
Ce jeu de données est composé de 1309 observations et de 12 variables dont :

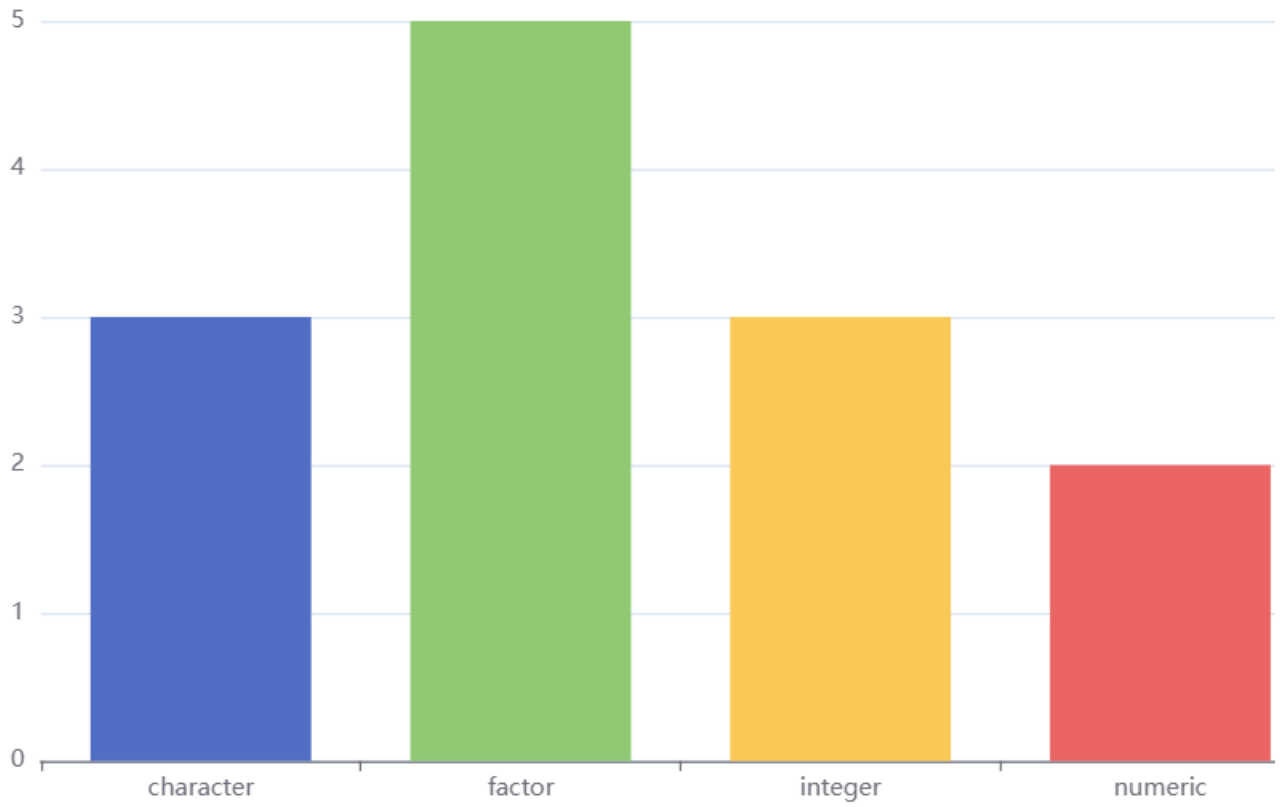
L'identifiant du passager -> `PassengerId`

La variable **`Survived`** qui prend comme valeur 0 si la personne a survécu 1 sinon

Le sexe : Homme ou Femme

Variables	Explication
`PassengerId`	Identifiant unique pour chaque passager (numéro de ligne)
`Survived`	**Statut de survie** : `1 = a survécu`, `0 = décédé`
`Pclass`	**Classe du billet** (catégorie socio-éco) : `1 = 1ère`, `2 = 2ème`, `3 = 3ème`
`Name`	Nom complet du passager (inclut le titre : Mr., Mrs., etc.)
`SEX`	**Sexe** du passager
`Age`	**Âge** du passagers
`Sibsp`	Nombre de frère et sœur et/ou de conjoint a bord
`parch`	Nombre de **parents et/ou enfants** a bord
`Ticket`	Numéro de **billet**
`Fare`	**Tarif payé pour le billet en £**
`Cabin`	Numéro de **cabine**
`Embarked`	Port d'**embarquement** • `C = Cherbourg` • `Q = Queenstown` • `S = Southampton`





1 EDA

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	
4	1	1	Futrelle, Mrs. Jacques	female	35	1	0	113803	53.1000	C123

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
			Heath (Lily May Peel)							
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583	
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.0750	
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	

Présentation du jeu de données

{Table-1} Table des 10 1er valeurs du jeu de données

Any nan values in the data ?

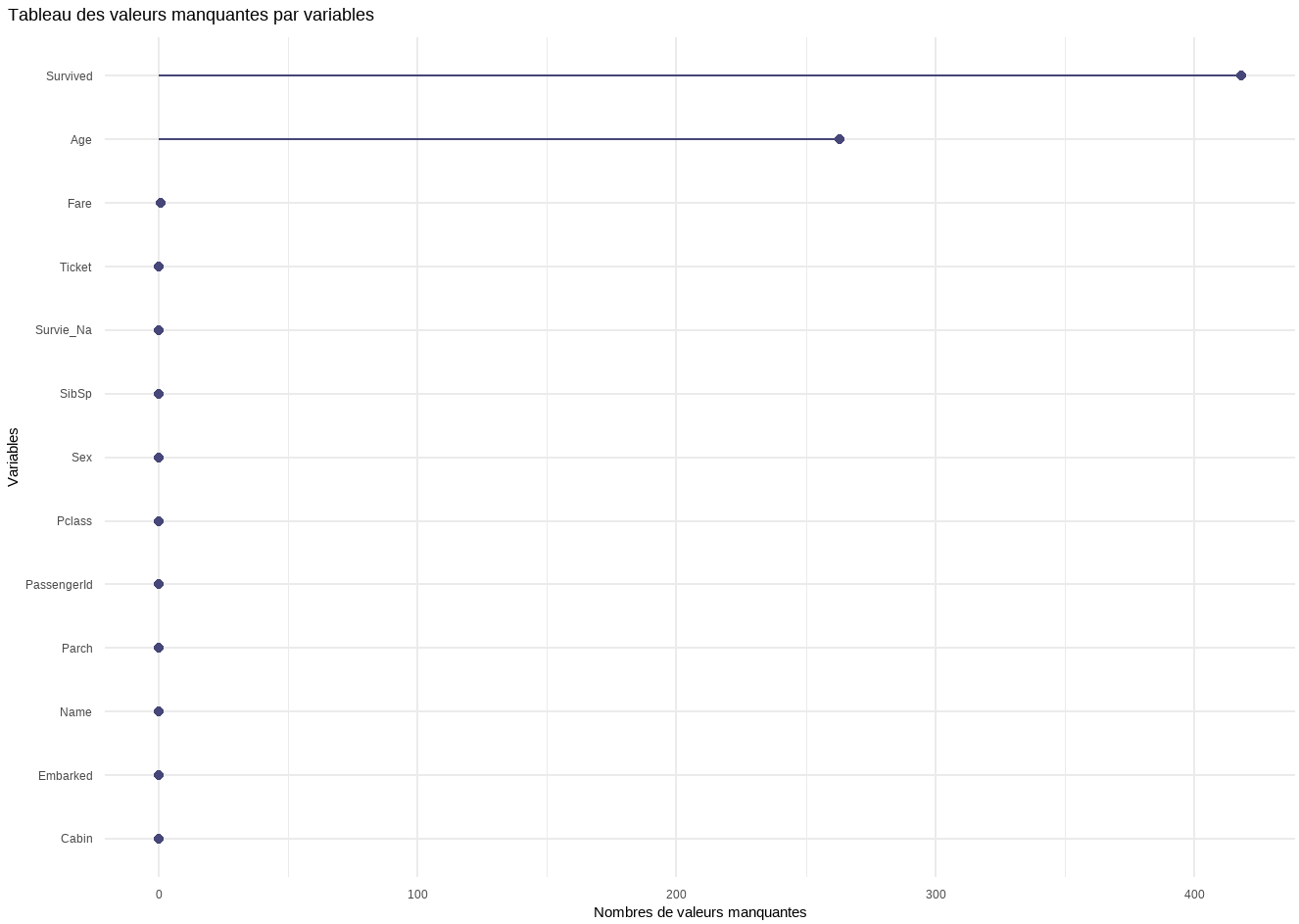


figure :2 : Tableau récapitulatif des valeurs manquantes par variables

La variable `Survivant ,Age et Fare` possèdent des valeurs manquantes respectivement (418,263 et 1) le fait d'avoir des données manquantes peut être problématique dans le sens ou si on ne gère pas ces données manquantes on risquerait de perdre des informations car ces valeurs manquantes touche 52,1% des données

```
'data.frame':   1309 obs. of  13 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
 $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs
Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr  "" "C85" "" "C123" ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
 $ Survie_Na  : Factor w/ 3 levels "0","1","Informations inconnues": 1 2 2 2 1 1 1 1 2 2 ...
```

Analyse de la variable cible

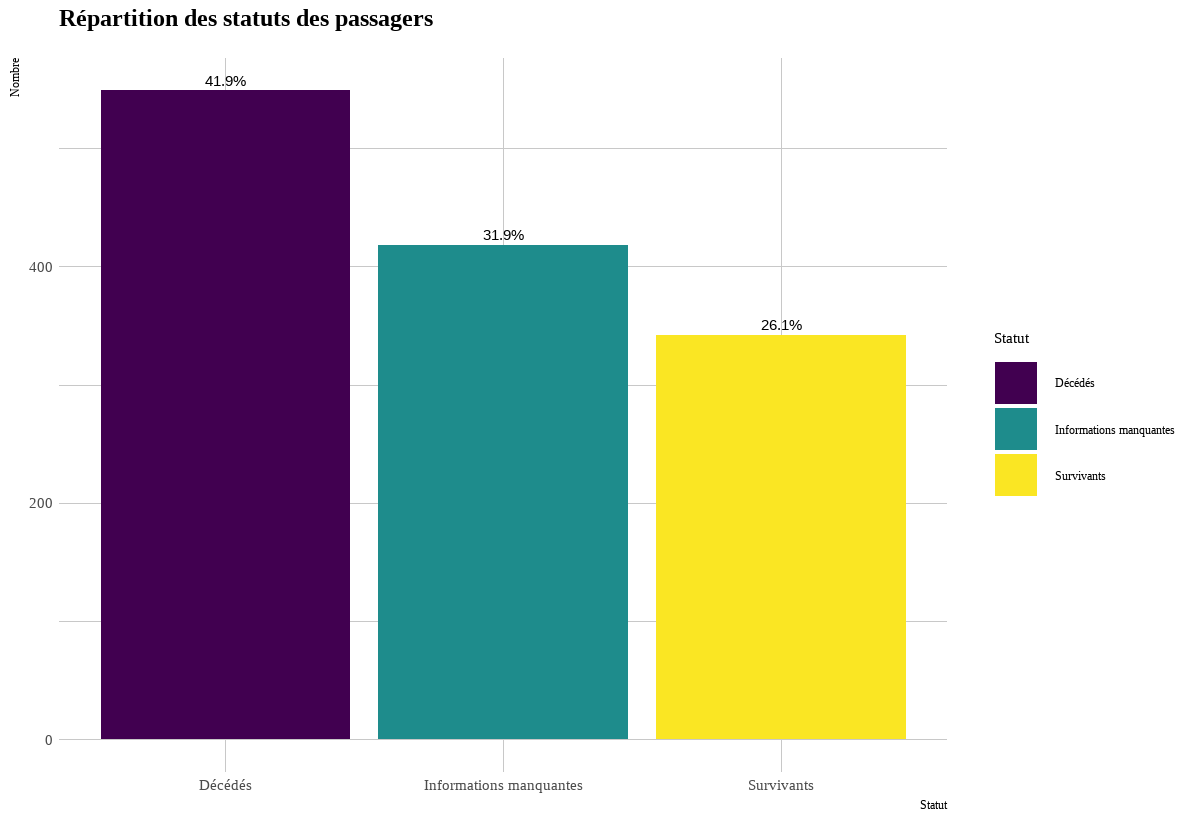
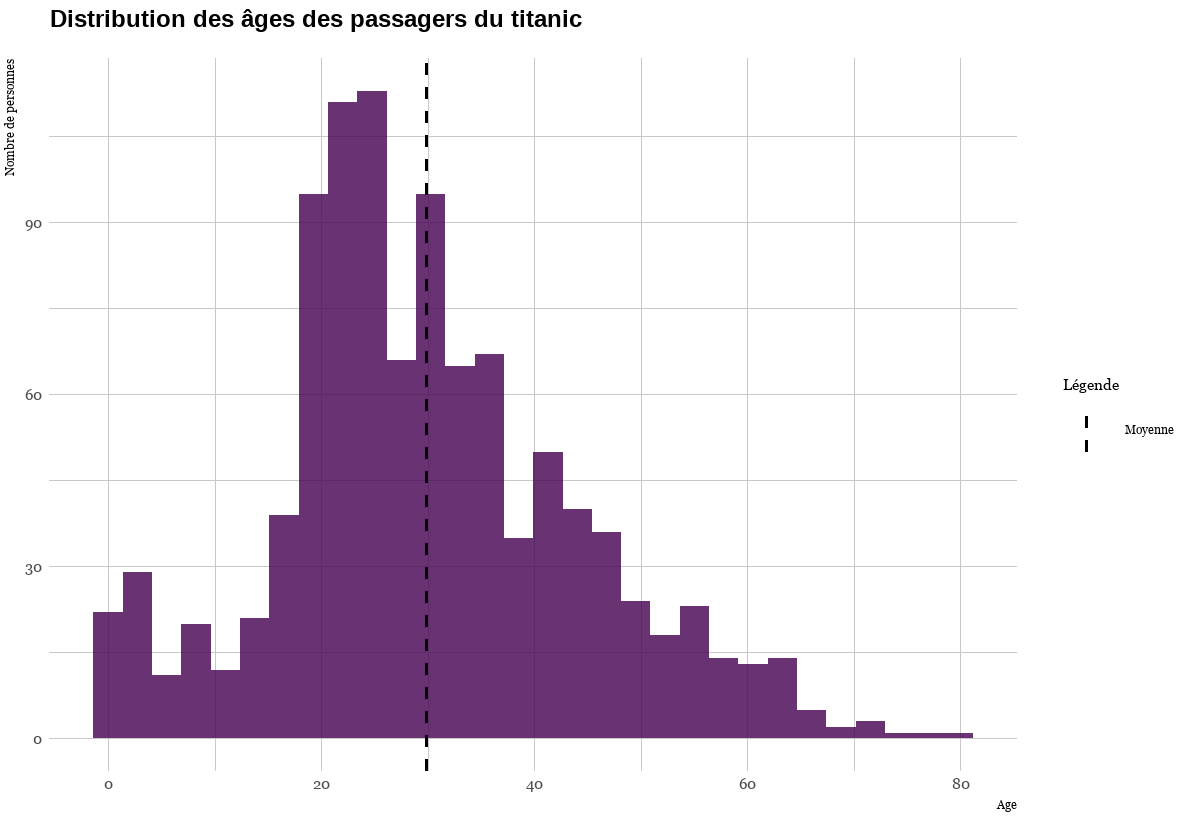


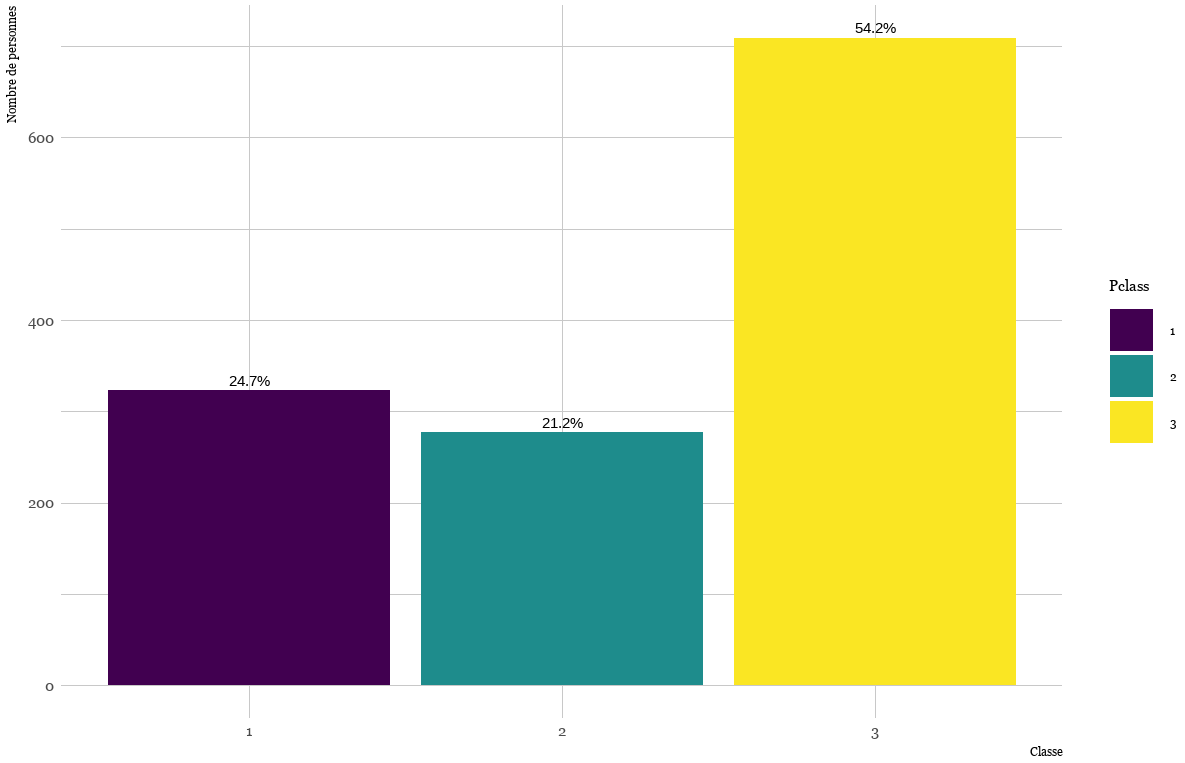
Figure 2 : Répartition des statuts des passagers du tit

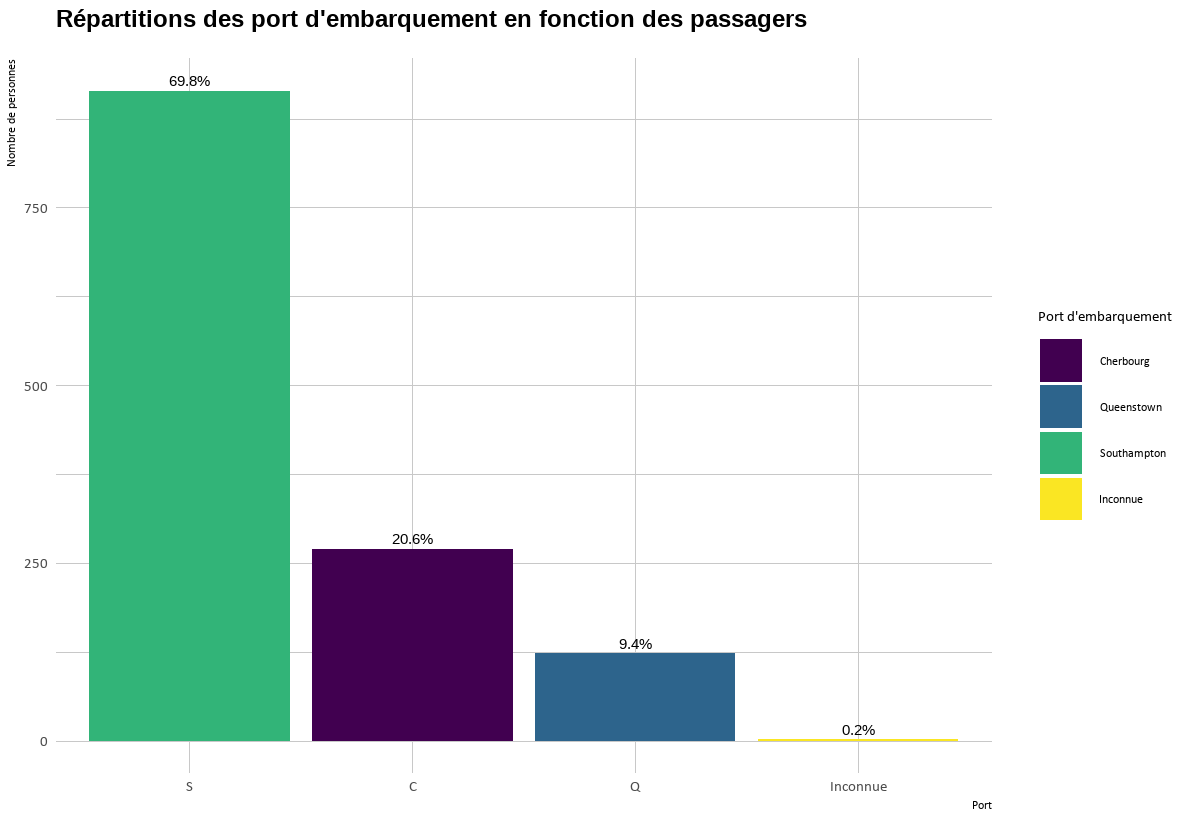
Lors de la décompte on observe que 41.9 % des passagers sont décédés ,26.1% sont des survivants et pour le reste ont ne dispose pas d'informations peut-être qu'elles sont portés disparus



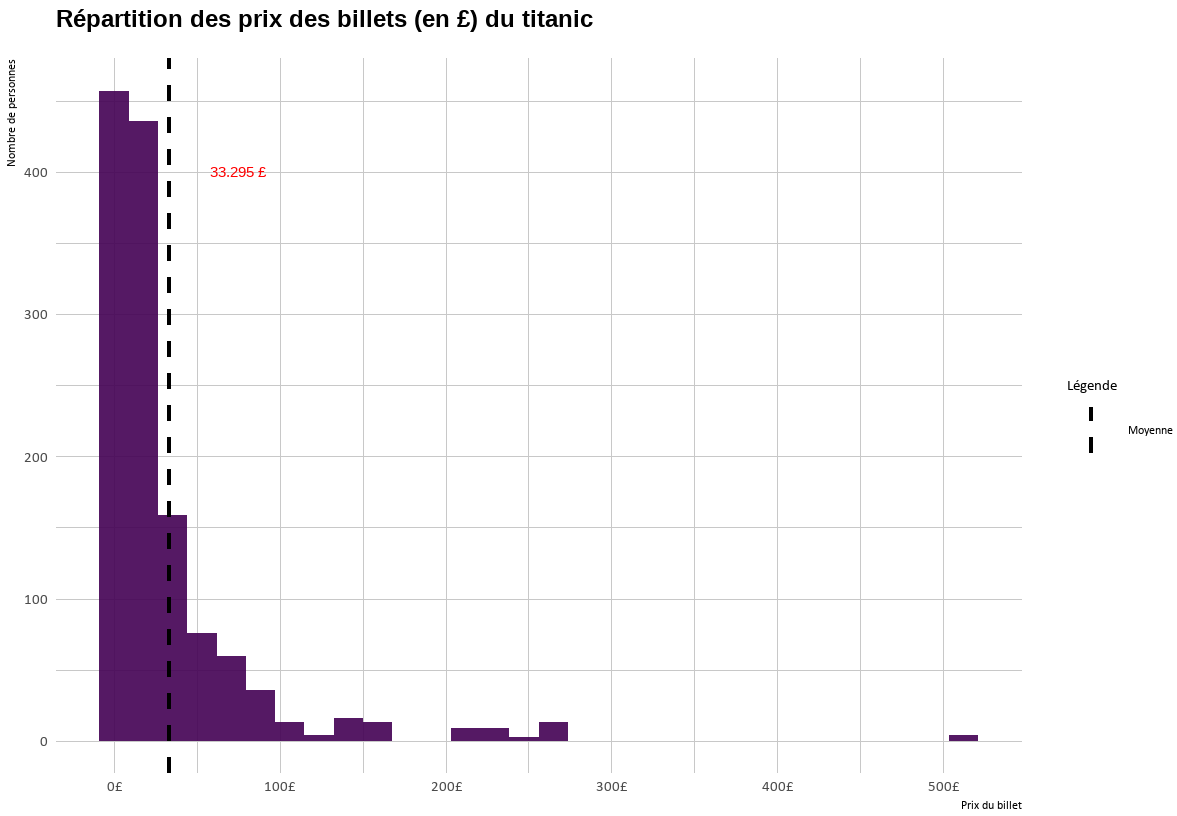
Distribution des ages parmis les passagers du titanic

Répartition des class en fonction des statut des passagers





Près de 70 % des passagers du **Titanic** ont embarqué à **Southampton**, contre 20,6 % à **Cherbourg** et 9,4 % à **Queenstown**.



Nombre de parents et/ou enfants a bord	Effectif	Proportion
0	1002	76.5 %
1	170	13 %
2	113	8.6 %
3	8	0.6 %
4	6	0.5 %
5	6	0.5 %
6	2	0.2 %
9	2	0.2 %

Parmi les passagers du `Titanic` 76.5 % voyage sans leurs enfants et ou leurs parents contre 13% qui voyage avec un seul parent ou enfant

Nombre de frère ou sœur et/ou de conjoint a bord	Effectif	Proportion

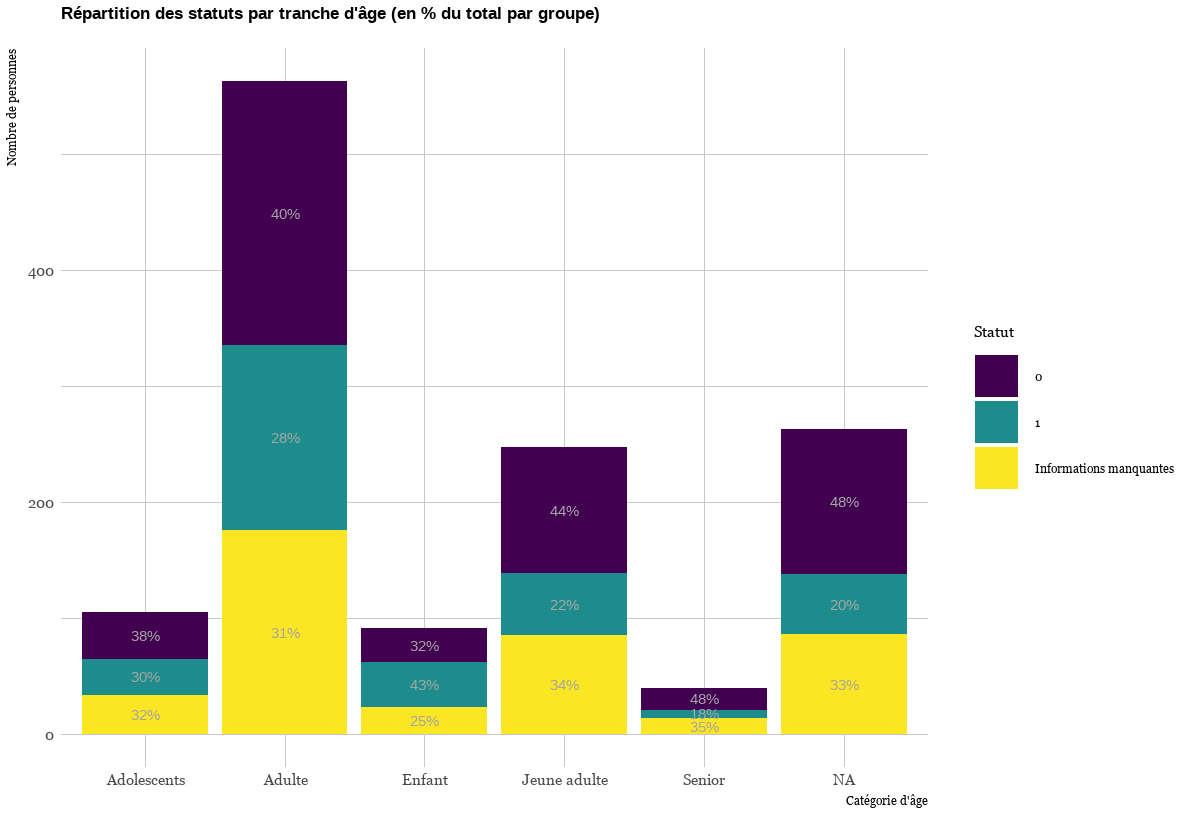
0	891	68.1 %
1	319	24.4 %
+2	57	4.4 %
2	42	3.2 %

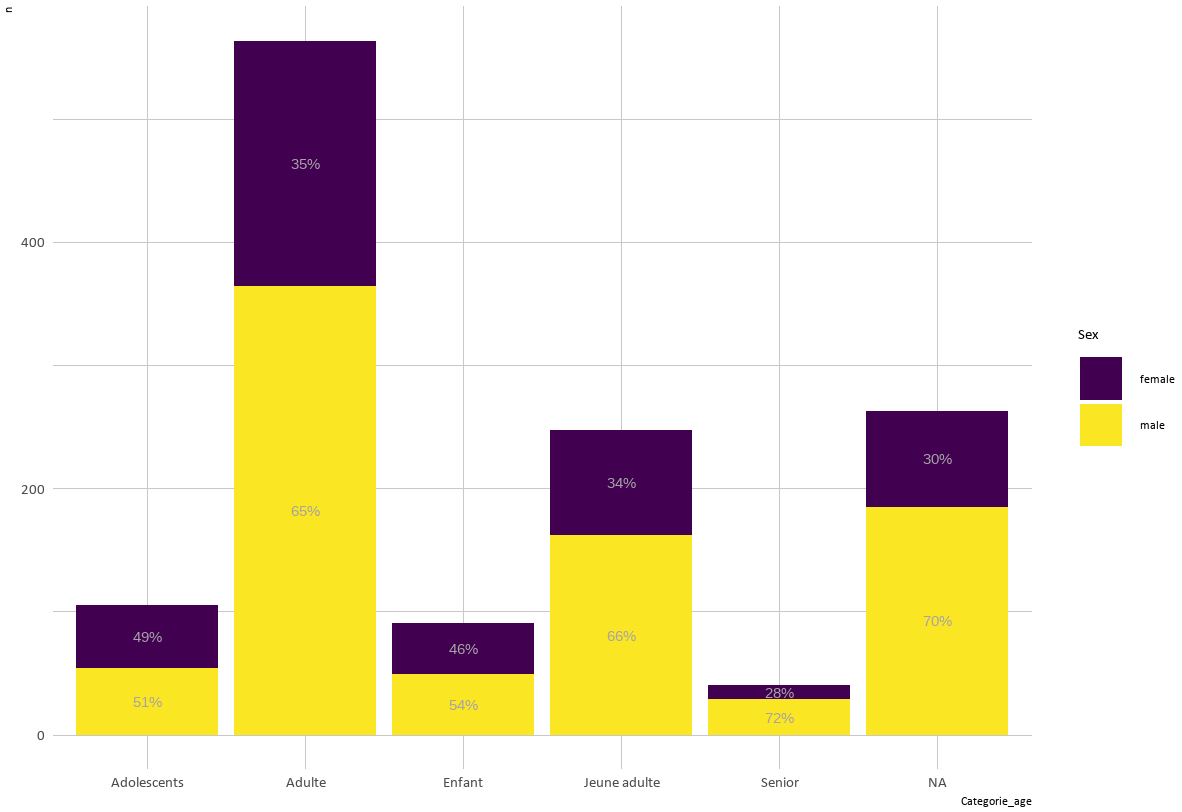
68,1 % des passagers voyageaient sans conjoint ni frère ou sœur, contre 24,4 % accompagnés d’un proche familial (frère, sœur ou conjoint)

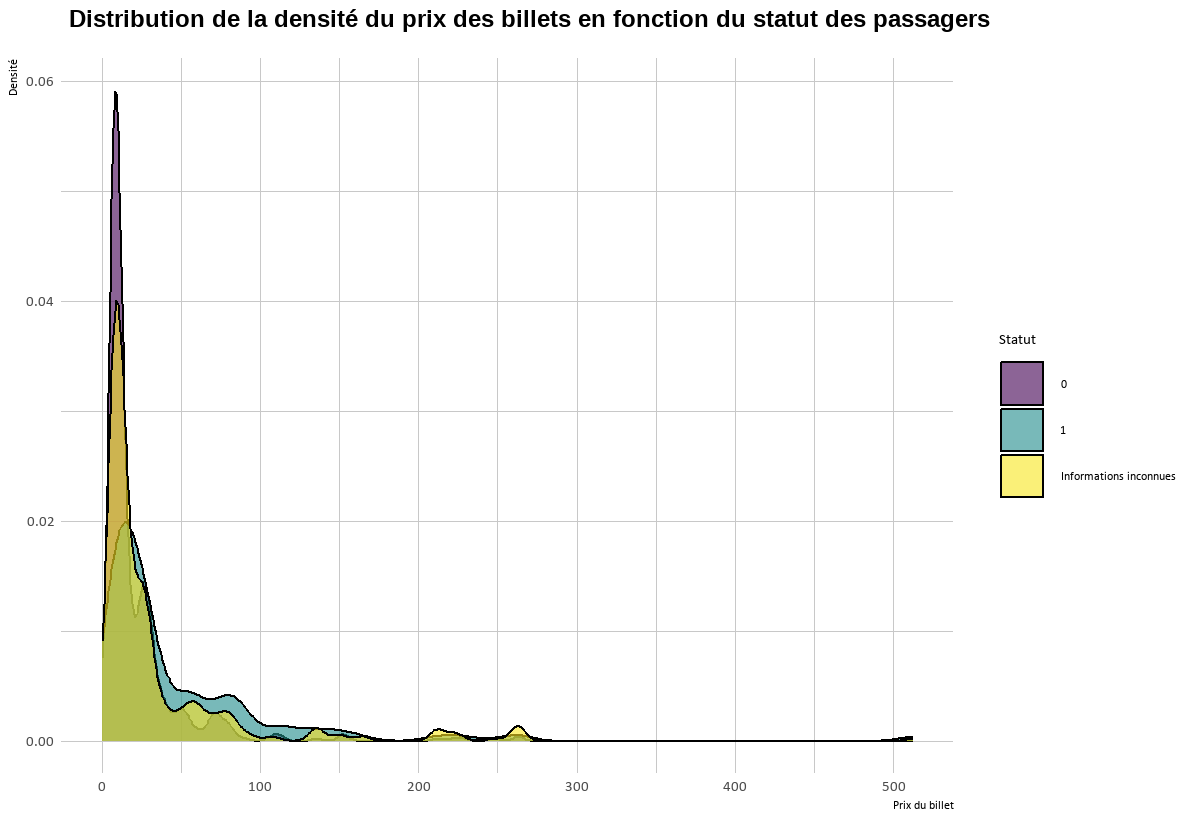
1.1 Analyse bivariée

```
Adolescents       Adulte       Enfant Jeune adulte       Senior         <NA>
        105          563           91          247           40          263
```

Tranche d'age	Nombre de personnes
Adulte	563
Inconnu	263
Jeune adulte	247
Adolescents	105
Enfant	91
Senior	40


Répartition des statuts par tranche d'âge (en % du total par groupe)
Nombre de personnes
400
200
0
40%
28%
31%
38%
30%
32%
32%
43%
25%
44%
22%
34%
48%
18%
35%
48%
20%
33%
Adolescents
Adulte
Enfant
Jeune adulte
Senior
NA
Catégorie d'âge
Statut
0
1
Informations manquantes


n
400
200
0
35%
65%
49%
51%
46%
54%
34%
66%
28%
72%
30%
70%
Adolescents
Adulte
Enfant
Jeune adulte
Senior
NA
Categorie_age
Sex
female
male



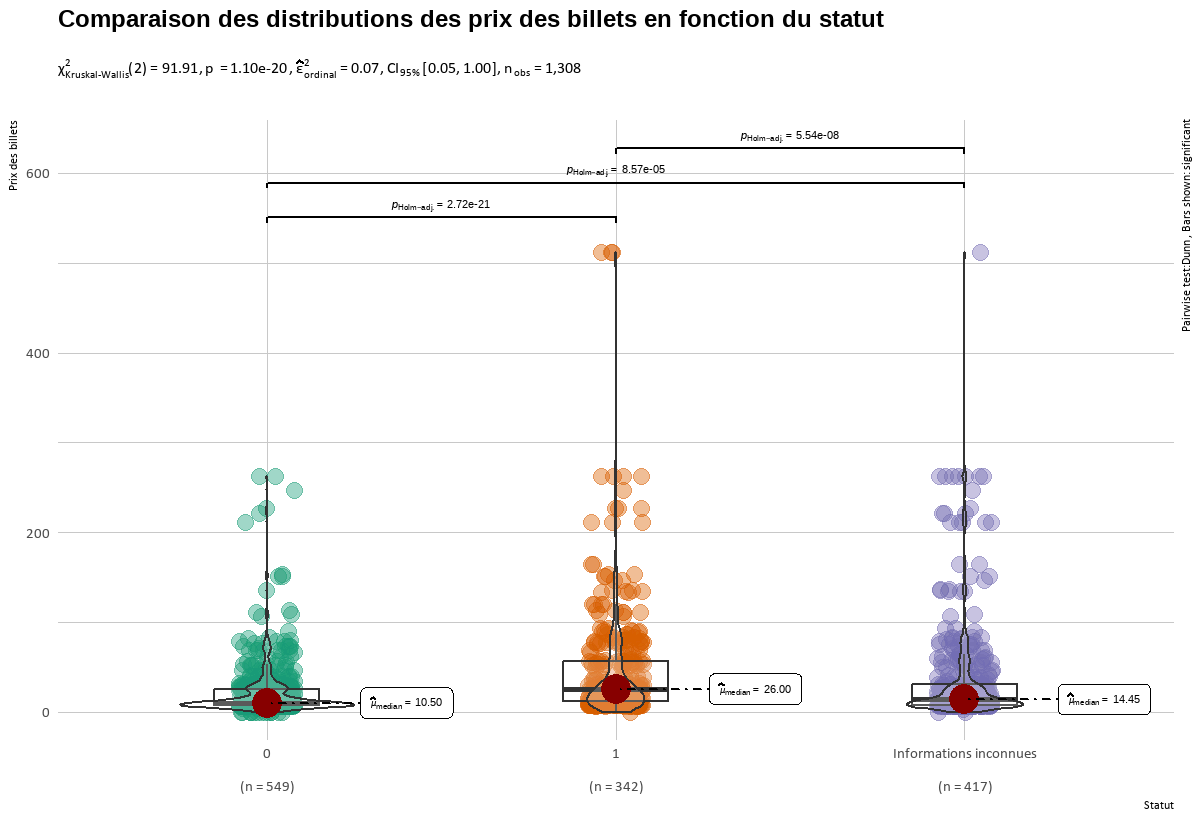
- Les distribution sont assez identique elles suivent toute la même tendance :
  - La plupart des billets ont étaient achetés autour de 25£ car c'est le prix auquel la densité est maximal
  - Un deuxieme pic (moins importante que la 1er) est observé aux alentour de 75£
  - Ont observe également des valeurs extrême des billets qui ont étaient vendus a + 200 £

1.2 Test statistique

Dans cette partie nous allons faire des test de significativité afin d'apprendre d'avantage sur les différences (potentiellement existantes) entre les différentes caractéristiques de la base donnée avec la variable cible (`Survivived)`

Existe-t-il des différences prix en fonction du statut ?

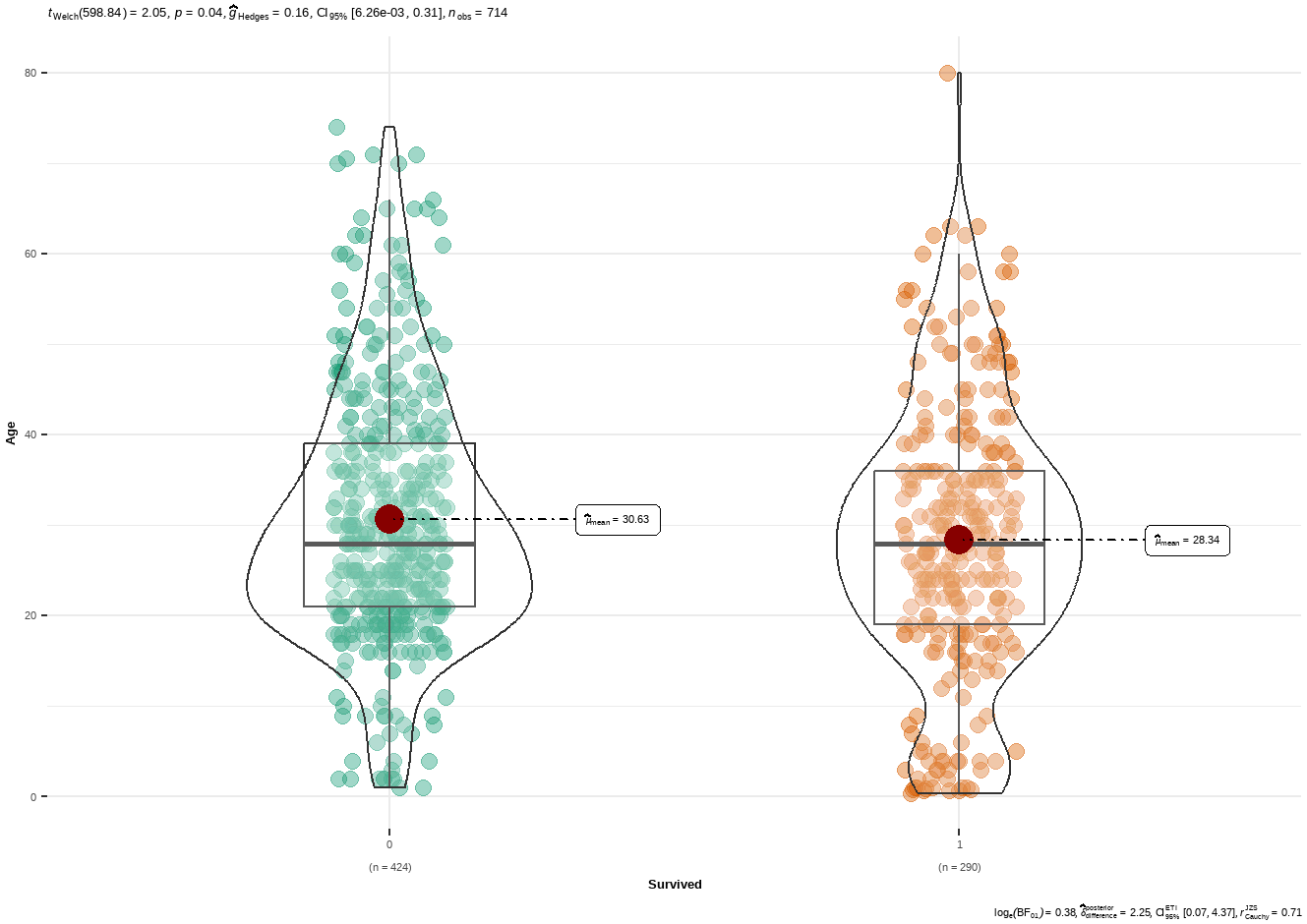
> Nous utilisons le test de `Kruskal-Wallis` pour évaluer s'il existe **des différences significatives dans la distribution des prix des billets (`Fare`) selon le statut de survie des passagers**. Le test permet de répondre à la question : **"Le prix du billet varie-t-il en fonction du fait qu'un passager ait survécu ou non ?"**



- Le test de Kruskal-Wallis indique que les prix des billets diffèrent significativement selon le statut des passagers ($p < 0.001$)
- Les comparaisons par paires à l'aide du test de `Dunn` , avec correction de `Holm` , confirment des différences significatives entre chacun des groupes
- Cependant vu que l'effet $\hat{\epsilon}^2_{ordinal} = 0.07$ cela suggère que les différences existe bel et bien mais elles sont **modérées**.Ces différences sont statistiquement significatives, même après correction des p-values (méthode de `holm` **)** pour comparaisons multiples

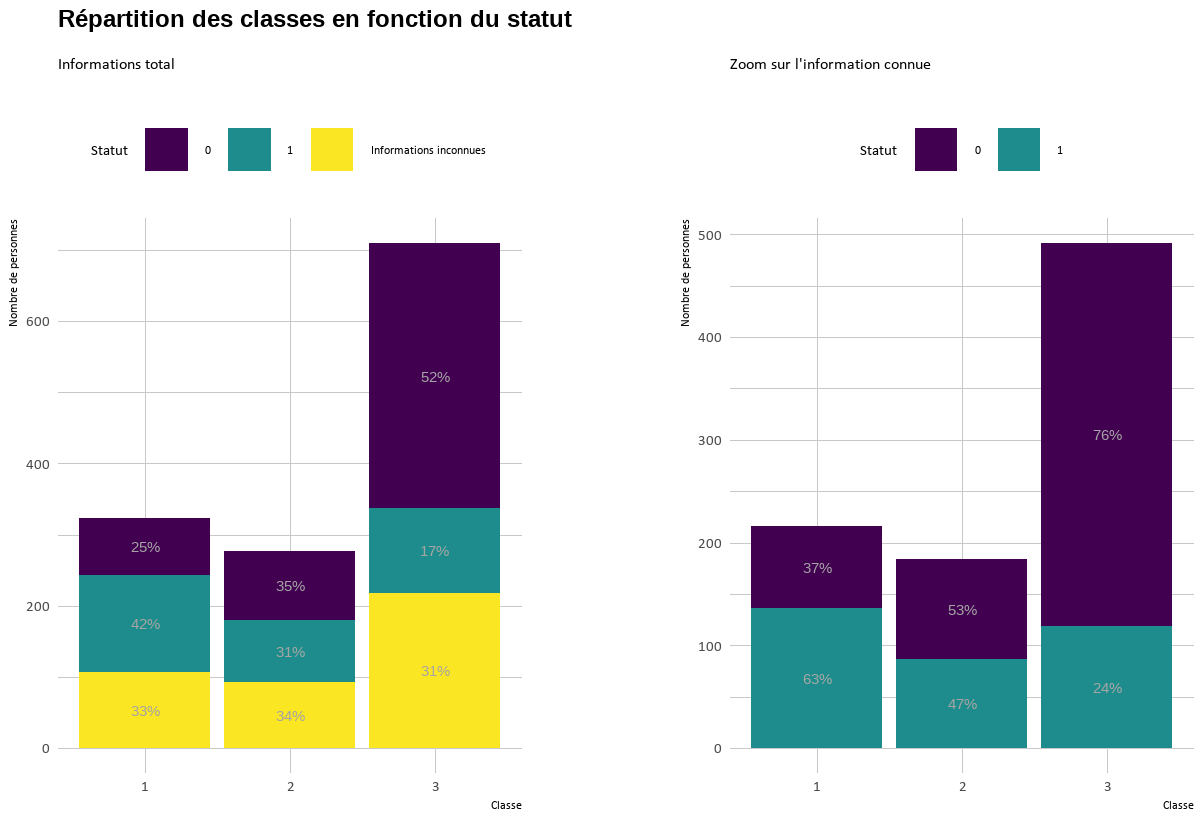
Analysons maintenant si la distribution de l'âge en fonctions du statut est significativement différente ou pas , cela nous permettrait de savoir si l'age des individus ont pût influencer ou pas le statut des individus.

Pour cela nous allons faire un test de **Welch**



Les survivants avaient un **age moyen de 28.34 ans contre 30.63 ans** pour les non-survivants. Bien que la différence d'âge moyen soit faible, le test de **Welch** nous dit que la différence est significative (probabilité critique < 5% mais > 1%) . Cela suggère que **l'âge pourrait jouer un rôle modéré dans la probabilité de survie**. Toutefois, la **taille de l'effet**, mesurée par l'indice de `Hedges` $\hat{g}_{Hedges} = 0.16$ est considérée comme **faible** selon la règle de Cohen

Nous allons maintenant analyser la relation entre la classe des passagers (`Pclass`) et leur probabilité de survie.



D'après les données disponibles, on observe que les passagers de **1re classe** présentaient un **taux de survie nettement plus élevé** que ceux des classes inférieures.
En effet, **42 % des passagers de 1re classe ont survécu**, contre **31 % en 2e classe** et seulement **17 % en 3e classe**.

À l'inverse, la **classe tertiaire (3e classe)** enregistre un **taux de mortalité particulièrement élevé** : **52 %**, contre **35 % en 2e classe** et seulement **25 % en 1re classe**.

Ces chiffres suggèrent une possible **inégalité d'accès aux moyens de sauvetage** selon la classe sociale.
Cela soulève une **hypothèse importante** :

> *Le fait d'appartenir à une classe supérieure augmente-t-il réellement les chances de survie, ou cette différence n'est-elle qu'un artefact dû à la **répartition déséquilibrée des passagers** entre les classes ?*

Cette question justifie la mise en œuvre d'un **test statistique d'indépendance**, afin de déterminer si la variable `Pclass` est **statistiquement associée au statut de survie**, au-delà de simples différences d'effectifs.

Cette analyse vise à déterminer si la **classe sociale des passagers**, représentée par la variable `Pclass`, a eu un **impact significatif sur leur probabilité de survie**. Comprendre cette relation permet d'**évaluer l'influence des inégalités sociales** dans les chances de survie lors du naufrage du Titanic.

```
[1] "Le test de khi deux nous renvoie :"
```

```
    Pearson's Chi-squared test

data:  table(df$Survived, df$Pclass)
X-squared = 102.89, df = 2, p-value < 2.2e-16

[1] "Les tests de significativités (post-hoc) par paires \n avec la méthode de fisher et un
ajustement des probabilité critique avec la méthode de holm:"
```

Comparison	p.Gtest	p.adj.Gtest	p.Chisq	p.adj.Chisq
1 : 2	0.00163	0.00163	0.00165	0.00165
1 : 3	0	0	5.21e-23	1.56e-22
2 : 3	1.42e-08	2.84e-08	7.04e-09	1.41e-08

Afin d'identifier plus précisément les différences entre les classes de passagers, nous avons effectué des **comparaisons par paires** entre les modalités de la variable `Pclass`, à l'aide du **test post-hoc du Chi²** (`pairwiseNominalIndependence`), avec une **correction des p-values selon la méthode de Holm**.

Les résultats montrent que :

- Il existe une **différence significative entre la 1re et la 2e classe** (p-value ≈ 0.001), indiquant une variation claire des taux de survie entre ces deux groupes.
- La **différence entre la 1re et la 3e classe est particulièrement marquée**, ce qui confirme l'hypothèse d'une **inégalité d'accès aux moyens de sauvetage**. Les passagers de 1re classe étaient probablement **mieux situés**, avec des cabines plus proches des **ponts supérieurs** ou des **canots de sauvetage**, facilitant leur évacuation.
- Un **écart significatif** est également observé entre la **2e et la 3e classe**, dans le même sens.

1.2.1 En conclusion :

Ces résultats confirment que **la probabilité de survie varie significativement selon la classe du passager**. En particulier, les passagers de 1re classe ont un **taux de survie nettement supérieur** à ceux des 2e et 3e classes.

Cela renforce l'hypothèse d'une **inégalité structurelle liée à la classe sociale**, influant sur les **conditions d'évacuation** et l'**accès aux dispositifs de sauvetage**, lors du naufrage du Titanic.

1.3 Gestion des Valeurs manquantes et feature engineering

1.3.0.1 Gestion des valeurs manquantes

La gestion des valeurs manquantes est une étape **essentielle en analyse de données**, car elle permet d'**affiner les résultats** en estimant des valeurs absentes de manière cohérente. Cela nous aide à nous rapprocher des **valeurs réelles** que l'on aurait pu observer dans la population ou l'échantillon étudié.

Cependant, toutes les techniques d'imputation **ne garantissent pas la fiabilité** des résultats : chaque situation ou type de données nécessite **une méthode d'imputation adaptée**.

Dans cette partie, nous allons tester **différentes stratégies d'imputation** afin de mieux gérer les données manquantes et d'**enrichir notre base de données**, dans le but d'améliorer les performances des modèles de machine learning

	Pourcentage des données manquantes
Cabin	0.7746
Survived	0.3193
Age	0.2009
Embarked	0.0015
Fare	0.0008
PassengerId	0
Pclass	0
Name	0
Sex	0
SibSp	0

1–10 of 12 rows Previous **1** 2 Next

Ont pourrais être tenté de supprimé la variable `Cabin` car elle offre pas assez d'informations 77,46% des données sont manquantes .

Avant de rentré en *`détails`* sur les technique adapter a cette variable ont doit comprendre comment elle est construite ?

La variable **`Cabin`** indique la cabine attribuée à chaque passager, généralement composée d'une lettre représentant le pont du navire (**`Deck`**) suivie d'un numéro.

Bien que cette variable contienne de nombreuses valeurs manquantes, l'information extraite du pont peut révéler des indices importants sur le statut social ou la position des passagers, et ainsi influencer leur probabilité de survie. (**`hypothèse`**)