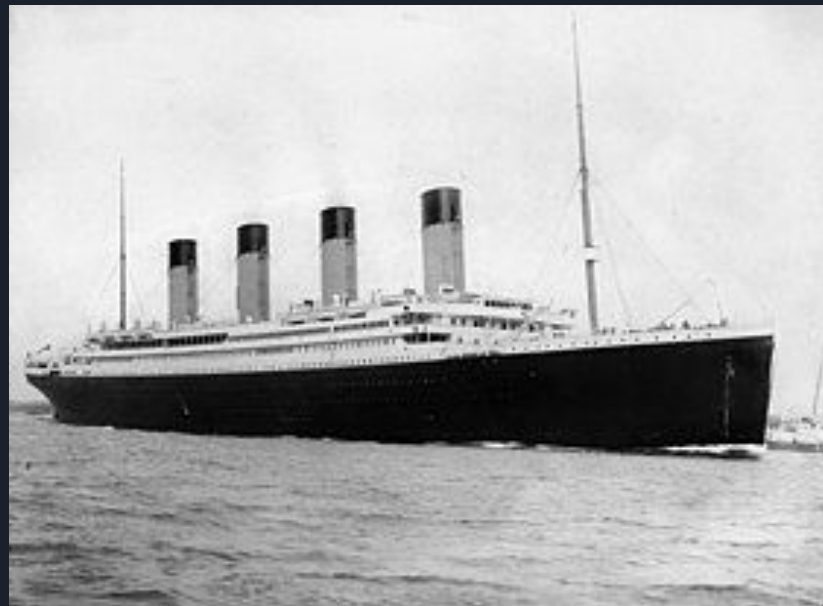


Le Titanic

Karima HARET
Thibaut FIOKA



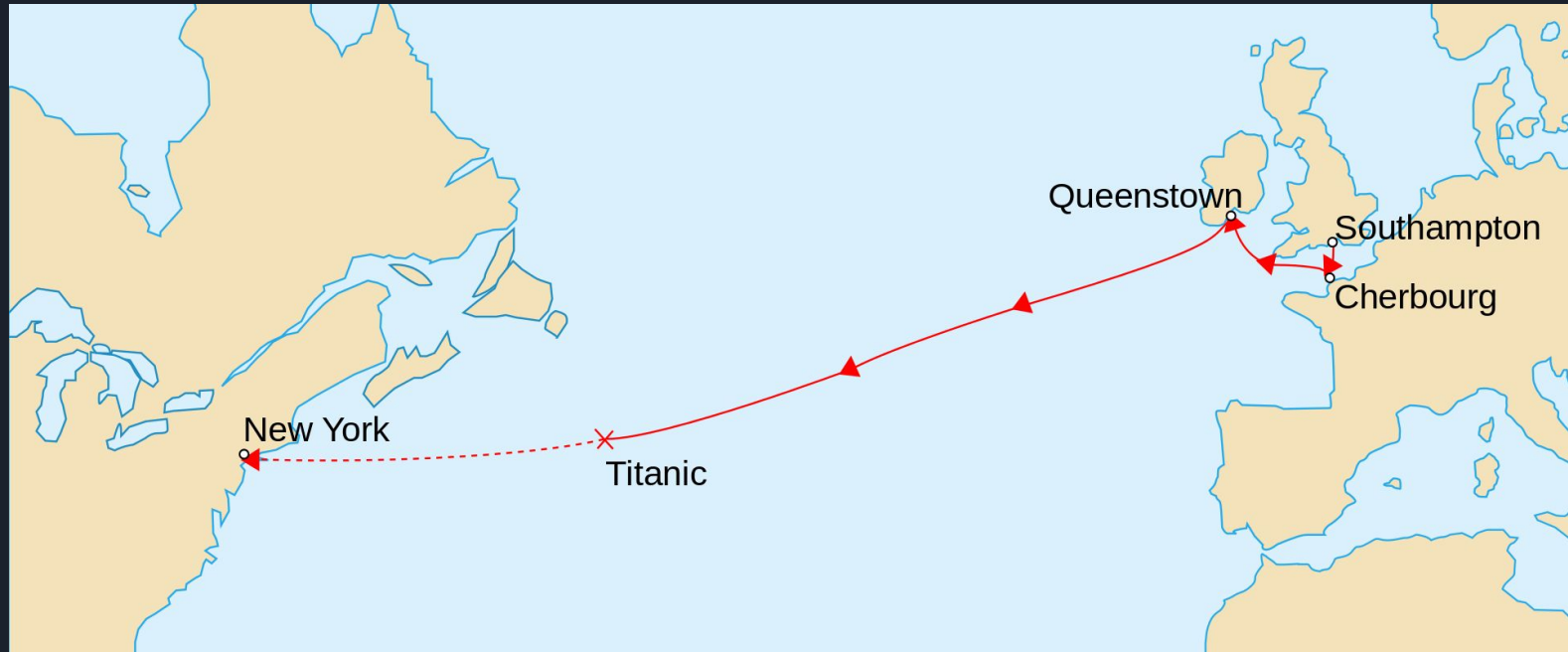


Sommaire

1. Introduction
2. Exploitation des données
3. Tests statistiques
4. Régression linéaire
5. Régression logistique
6. Conclusion

Introduction

Histoire



Introduction

Présentation des données

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
12	13	0	3	Saunderscock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.0500	NaN	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NaN	S

Introduction

Nettoyage de la base

- Suppression de la colonne “Cabin”
- Remplacement des valeurs manquantes de la colonne “Age” par la moyenne
- suppression des 2 lignes dans lesquelles l’information “Embarked” était manquante

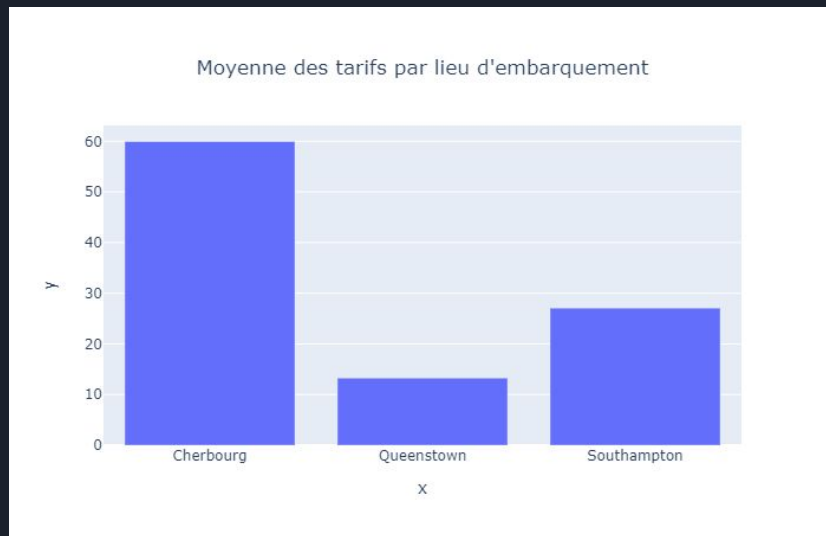
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
12	13	0	3	Saunderscock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.0500	NaN	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NaN	S

On passe donc de 891 à 889 lignes.

Exploitation des données

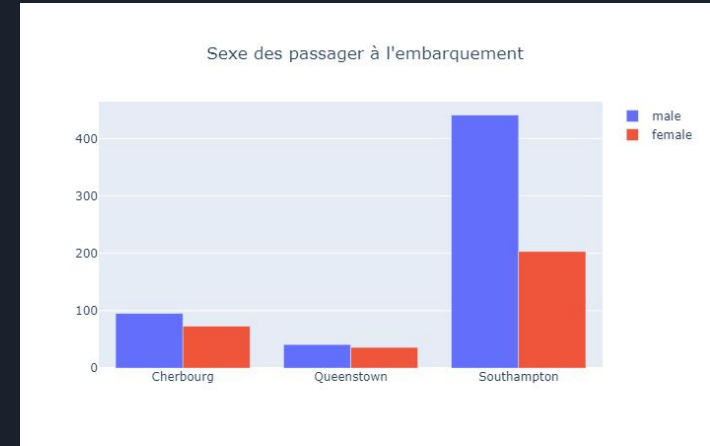
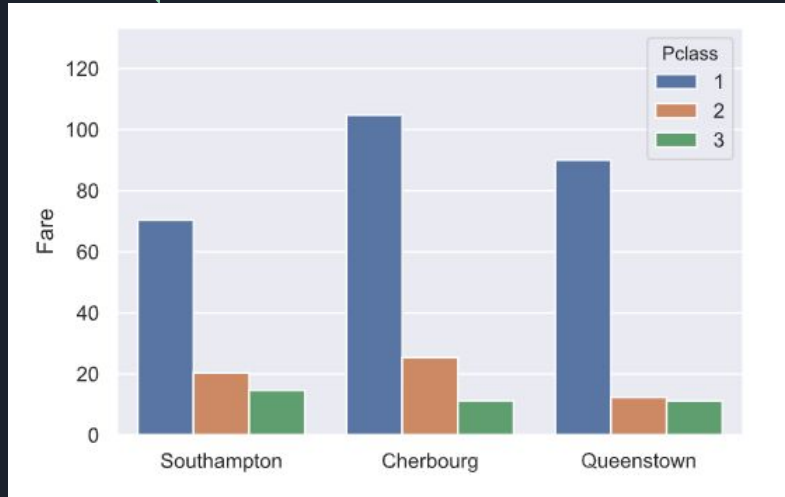
Embarquement

A l'embarquement des 889 passagers que nous avons sélectionnés, 644 ont embarqués à Southampton (Angleterre), 168 à Cherbourg (France) et 77 à Queenstown (Irlande).



Exploitation des données

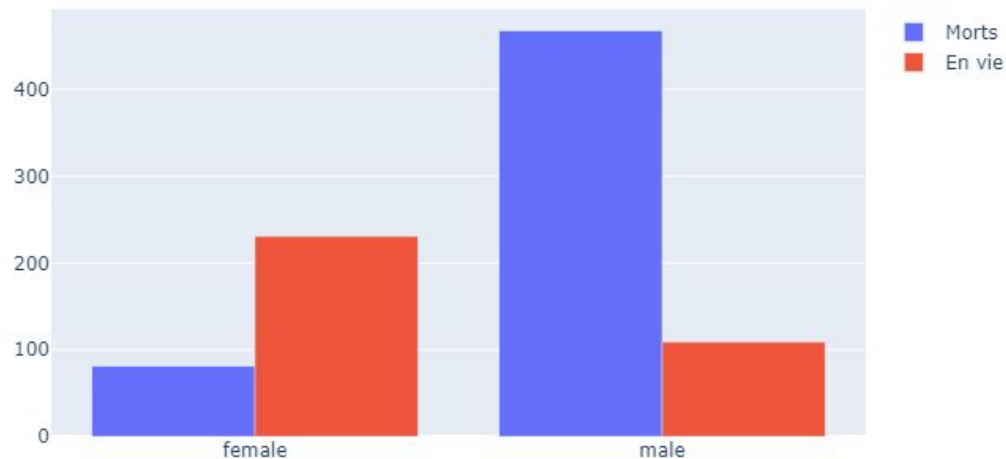
Embarquement



Exploitation des données

Naufrage : qui a survécu ?

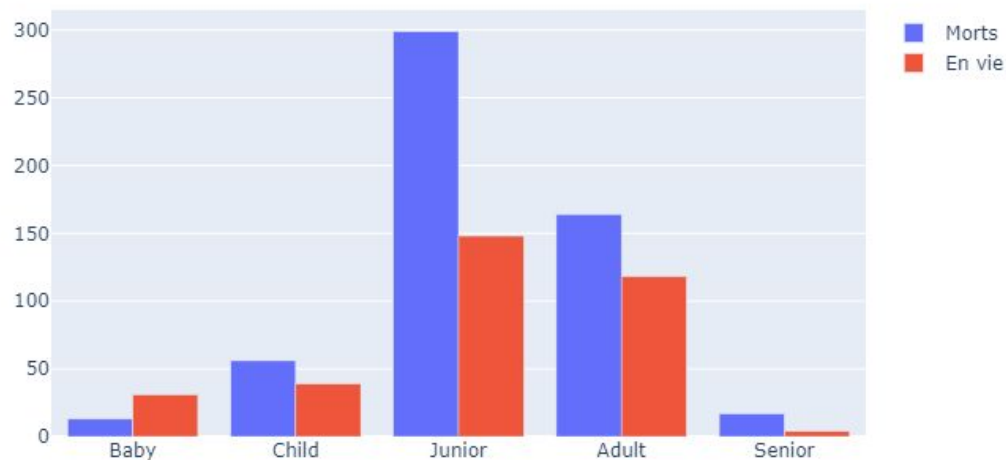
Etat des passager en fonction du sexe



Exploitation des données

Naufrage : qui a survécu ?

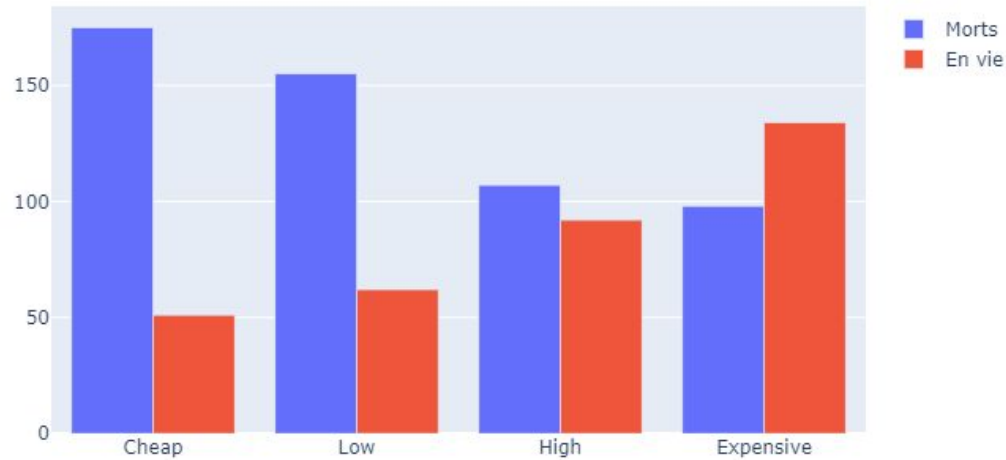
Etat des passager en fonction de leur âge



Exploitation des données

Naufrage : qui a survécu ?

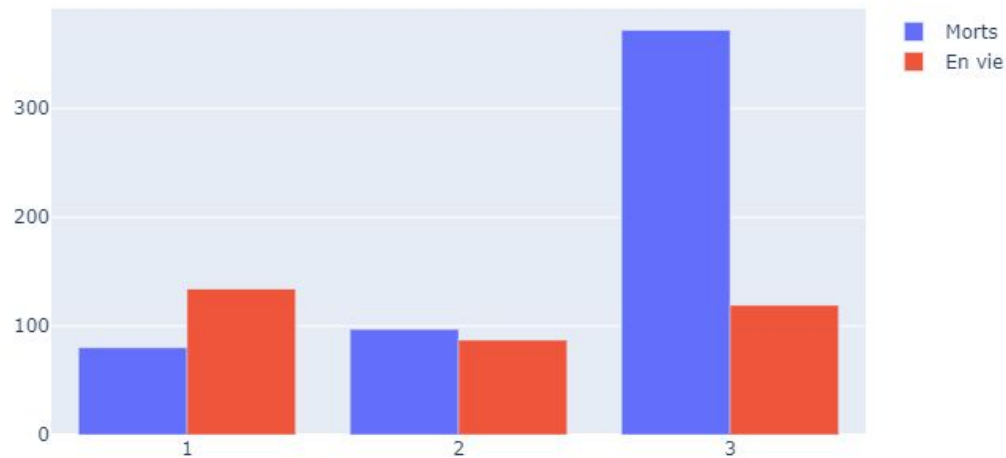
Etat des passager en fonction du prix du ticket



Exploitation des données

Naufrage : qui a survécu ?

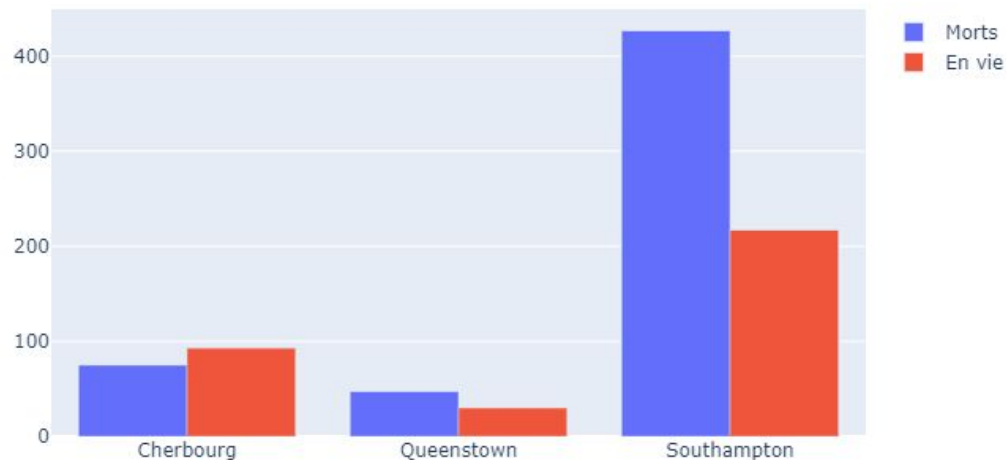
Etat des passager en fonction de la classe de ticket



Exploitation des données

Naufrage : qui a survécu ?

Etat des passager en fonction du point d'embarquement





Tests Statistiques

Hypothèses

D'après les graphes précédents, nous pouvons dégager plusieurs hypothèses :

- 1 : Les femmes ont été privilégiées durant le sauvetage
- 2 : Les enfants (passagers étant âgés de moins de 18 ans) ont été privilégiés durant le sauvetage
- 3 : Les passagers ayant payé un ticket plus cher ont été privilégiés durant le sauvetage
- 4 : Les passagers ayant une meilleure classe ont été privilégiés durant le sauvetage
- 5 : Les passagers ayant embarqué à Cherbourg ont été privilégiés durant le sauvetage



Tests statistiques

Hypothèses

Vérification de l'hypothèse 1 :

Avec le test du khi-2, nous obtenons une p-value bien inférieure à alpha, ce qui indique que l'hypothèse 1 est validée.

```
pvalue = 3.7799096665574906e-58 alpha = 0.05  
Les deux variables ne sont pas indépendantes.
```

Vérification de l'hypothèse 2 :

Avec le test du khi-2, nous obtenons une p-value bien inférieure à alpha, ce qui indique que l'hypothèse 2 est validée.

```
pvalue = 0.00034261369274988666 alpha = 0.05  
Les deux variables ne sont pas indépendantes.
```

Vérification de l'hypothèse 3 :

Avec le test du khi-2, nous obtenons une p-value bien inférieure à alpha, ce qui indique que l'hypothèse 3 est validée.

```
pvalue = 1.553812876621213e-11 alpha = 0.05  
Les deux variables ne sont pas indépendantes.
```



Tests Statistiques

Conclusion

Les femmes, les enfants, et les passagers ayant payé cher leur billet ont été privilégiés durant le sauvetage.




Régression Linéaire

Régression multiple sur plusieurs variables

Dans cette partie, nous cherchons à savoir si la survie peut s'exprimer en fonction de l'âge, la classe, le prix du billet et le sexe. On suppose qu'il existe une relation linéaire du type :

$$survived = \alpha_0 + \alpha_1 Age + \alpha_2 classe + \alpha_3 Fare + \alpha_4 Sex$$

Pour cela, nous décomposons notre jeu de données en deux parties Train et Test. La partie Train nous servira à faire l'apprentissage de notre modèle et Test pour tester sa performance en s'appuyant sur l'erreur des moindres carrés.



Régression Linéaire

Régression multiple sur plusieurs variables

Les résultats obtenus sont résumés dans le tableau ci-dessus :

α_0	α_1	α_2	α_3	α_4	R^2	erreur
0.8375	-0.0056	-0.1924	0.00004	0.4612	0.46	0.46

Ce qui est de loin insuffisant car le coefficient de corrélation R^2 est éloigné de 1. Nous allons donc faire des régressions linéaires en ne prenant en compte que certaines variables.



Régression Linéaire

Régression Linéaire une à une

Voici les R^2 obtenus pour chaque variable, prise individuellement :

Variable	Modèle	R^2
Age	$Survived = 0.5 - 0.004 \cdot Age$	-0.029
Classe	$Survived = 0.509 - 0.004 \cdot Classe$	0.097
Fare	$Survived = 0.509 - 0.004 \cdot Fare$	0.015
Sexe	$Survived = 0.509 - 0.004 \cdot Sexe$	0.39

Nous remarquons que la variable "Sexe" a le plus gros coefficient R^2 .



Régression Linéaire

Régression Linéaire une à une

Nous allons faire une régression avec des combinaison de la variable « Sexe » et les autres variables. Les résultats se résument ci-dessous :

Variable	Modèle	R ²
Sexe + Classe	$Survived = 0.59 - 0.48.Sexe - 0.16.Classe$	0.45
Sexe+ Classe+ Fare	$Survived = 0.58 - 0.48.Sexe - 0.16.Classe + 0.0002.Fare$	0.46

Nous remarquons que le R² ne change pas même si on ajoute d'autres variables après la variable « Sexe » et reste loin de 1.



Régression Linéaire

Conclusion

La régression linéaire n'est pas envisageable sur notre jeu de données.



Régression Logistique

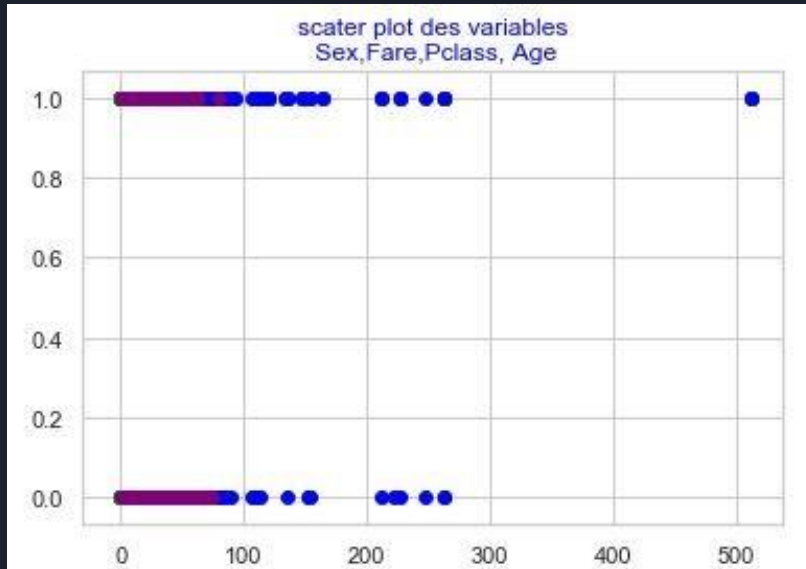
Il s'agit d'expliquer la « survie » en fonction des autres variables comme précédemment mais en utilisant la fonction « logit » suivante :

$$F(x) = \frac{e^x}{1+e^x} \text{ où } x = \alpha_0 + \alpha_1.Age + \alpha_2.classe + \alpha_3.Fare + \alpha_4.Sex$$

$$\text{Logit}(x) = \alpha_0 + \alpha_1.Age + \alpha_2.classe + \alpha_3.Fare + \alpha_4.Sex$$

Régression Logistique

Nous avons tracé le nuage de points de la variable survie et fonction des 4 autres, le résultat est dans le graphe ci-dessus :



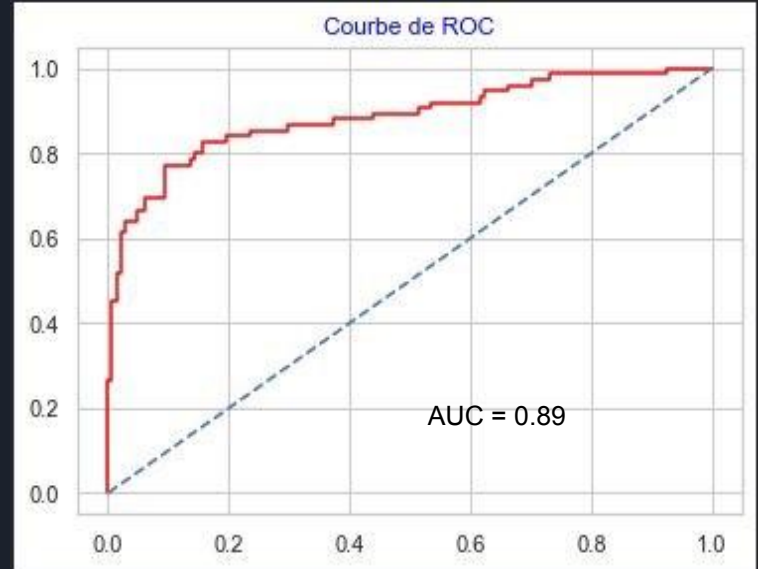
Régression Logistique

Paramètres estimés :

$$\alpha = \begin{pmatrix} 2.1445 \\ -1.100 \\ 2.272 \\ -0.003 \\ 0.0008 \end{pmatrix}$$

Notre modèle de régression logistique s'écrit :

$$\text{Logit}(x) = 2.145 + -1.100\text{Age} + 2.272\text{classe} + -0.003\text{Fare} + 0.0008\text{Sex}$$





Conclusion

Ce projet nous aura permis de lire scientifiquement des données historiques liées à un évènement tragique. Nous sommes donc maintenant en mesure de dire que l'adage dont il a été question dans ce dossier a été respecté.

De même, nous avons démontré que la survie d'un passager était très liée à sa richesse. Cette dernière conclusion nous prouve que des privilèges étaient encore valables sur ce paquebot titanesque.

Ainsi nous avons pu comparer la performance de deux algorithmes de Machine Learning qui sont la régression logistique et linéaire. L'algorithme de régression logistique est le plus adapté dans notre cas d'étude.