Analyse dataset - Titanic

Lien d'analyse kaggle : https://www.kaggle.com/code/startupsci/titanic-data-science-solutions/notebook

Analyse du dataset Titanic

Exploratory Data Analysis

Analyse de forme

- **Identification de la target**: le target est la colonne « Survived »
  Ici on a (sur trainset):
  - 549 morts (0) : 61.6 %
  - 342 survivants (1) : 38.4 %

  L'équilibre des classes est acceptable mais toujours à surveiller.
- **Dimensions** :
  - trainset : 891 lignes , 12 colonnes
  - testset : 418 lignes, 11 colonnes
  - all dataset : 1309 lignes, 12 colonnes
- **Types :**
  - int64 : 5
  - object : 5 (catégoriel)
  - float64 : 2
- **Valeurs manquantes :**
  - Age : 177 valeurs manquantes, ≅ 20% missing value
    On pourrait remplir avec la valeur médiane ou créer un modèle pour prédire
  - Cabin : 687 valeus manquantes, ≅ 77% MV
    On pourrait supprimer ou encoder avec un modèle, prédire les valeurs
  - Embarked : 2 valeurs manquantes, ≅ 0.2% MV
    On pourrait remplir avec mode

  Les valeurs manquantes pour chaque colonnes semblent n'avoir aucune relation/ correlation.

- **Duplicats** : aucune ligne dupliquée

Analyse de fond

Signification des variables

Nom de la colonne	Type	Description
PassengerId	Numérique (int)	Identifiant unique du passager. Ne contient pas d'information utile pour la prédiction.
Survived	Binaire (0 ou 1)	Variable cible. 0 = mort, 1 = survécu.
Pclass	Catégorielle (1, 2, 3)	Classe du billet : 1 = première, 2 = deuxième, 3 = troisième. Proxy du statut socio-économique.
Name	Texte	Nom complet du passager. Peut contenir des titres (Mr, Mrs, Miss...) utiles pour extraire des informations.
Sex	Catégorielle (male, female)	Sexe du passager. Très corrélé avec la survie (femmes ont survécu en plus grand nombre).
Age	Numérique (float)	Âge du passager. Contient des valeurs manquantes. Important pour évaluer la vulnérabilité.
SibSp	Numérique (int)	Nombre de frères/sœurs ou conjoints à bord. Utilisé pour déterminer si la personne voyageait en famille.
Parch	Numérique (int)	Nombre de parents ou enfants à bord. Combine avec SibSp pour inférer la taille de la famille.
Ticket	Texte	Numéro du billet. Souvent non standardisé, peu utile sans gros nettoyage.
Fare	Numérique (float)	Prix du billet. Peut être un indicateur du statut économique. Corrélé à Pclass.
Cabin	Texte	Numéro de cabine. Beaucoup de valeurs manquantes. Peut contenir l'information sur la localisation sur le bateau.

Nom de la colonne	Type	Description
Embarked	Catégorielle (C, Q, S)	Port d'embarquement : C = Cherbourg, Q = Queenstown, S = Southampton. Peut refléter des différences socio-économiques ou géographiques.

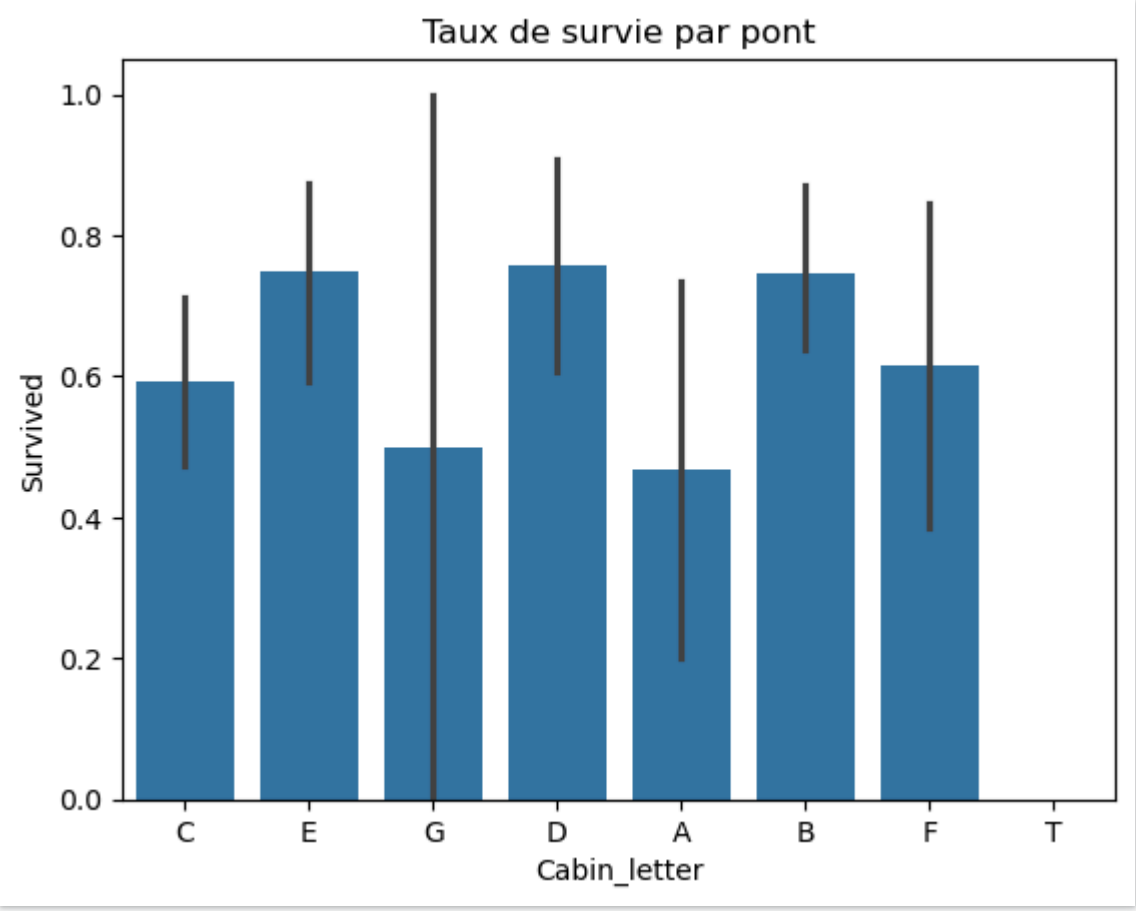
Analyse des variables par type

- **Variable numérique :**
  - **Age**: la pluspart des personnes à bord se situaient entre 20 et 38 ans; avec une médiane de 28 ans
  - **Fare** : il a y 15 lignes où Fare = 0, cela pourrait indiquer des membres d'équipages, cette colonne semble contenir beaucoup d'outliers aussi (à revoir)
- **Variable discrète :**
  - **SibSp et Parch** : on constate que beaucoup de persones ont voyagé seul. On consate que 50% (≅ 354 passagers) de ceux qui n'ont pas voyager seul ont survécu et à peu près 30% (≅ 537 passagers) ont survécu.Cela suggère que la présence de proche à bord a été favorable à la survie peut être grâce à l'entraide lors de l'évacuation, la priorisation lors de l'embarquement ou effets psychologique.
- **Variable catégorielle :**
  - **Sex** : On compte 65% de male et 35% de female dans le trainset. On constate que 74 % de femmes ont survécu et seulement 19% des hommes. Cela est probablement du à la règle de priorité « les femmes et les enfants » d'abord pour l'accés au canots.
  - **Pclass** : On compte 55 % des passagers ont voyagé en $3^{ème}$ classe, 24 % en $2^{ème}$ et 20% en $1^{ère}$ classe. On constate aussi que: 63% de ceux qui ont voyagé en première classe en survécu, 47% en deuxième classe et 24% en troisième classe. Cela s'explique peut être par la priorisation des riches lors de l'évacuation.
  - **Cabin** : Voici un tableau récapitulatif qui catégorise les passagers selon leurs cabine (≅ selon le pont proche de leur cabine).

Cabin	Nombre de passagers
A	15
B	47

Cabin	Nombre de passagers
C	59
D	33
E	32
F	13
G	4
T	1

N'oublions pas qu'il y a 77% de NaN pour cette colonne.



Par ailleurs, après création d'une nouvelle variable "Has_cabin", on constate que 30% de ceux qui n'ont pas eu de cabine ont survecu et 67% pour ceux qui en ont. Il est à revoir plus tard la relation avec les autres variables.

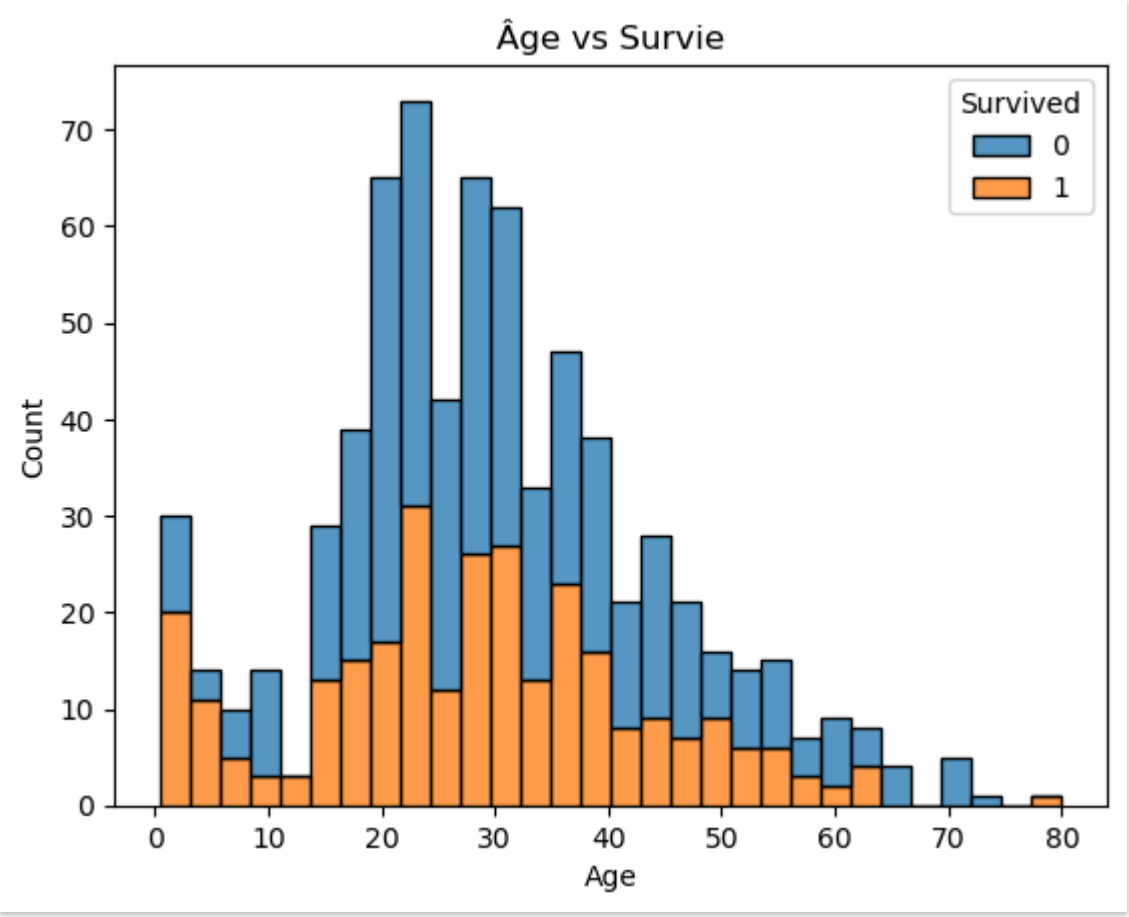
- **Embarked :**
  - S (port Southampton) : représente 72% des passagers
  - C (port Cherbourg) : 19% des passagers
  - Q (port Queenstown) : 9% des passagers

    La relation entre Embarked et Survived est à revoir mais :

- C : 55% de survivants
- Q : 39%
- S : 34%

Analyse de corrélation features vs target

- **Age vs Survived :**
  - Enfant : 58% de suvivant
  - Ado : 43%
  - Adulte : 39%
  - Moyen age : 40%
  - Vieux : 23%



On peut dire que les enfants et les jeunes survivents plus que les adultes et les vieux.

- **Fare vs Survived :**
  - (-0.001, 7.91]$: 20% de chance de survie
  - (7.91, 14.454]$: 30%
  - (14.454, 31.0]$: 45%
  - (31.0, 512.329]$: 58%

On constate donc que les personnes de la classe 1 ont plus de chance de survie par rapport à la $3^{ème}$ classe.

Corrélations inter-features

- Voici les variables avec un taux de **corrélation élevé** :
  - 'Family_size' & 'SibSp' : 0.89
  - 'Family_size' & 'Parch' : 0.78
  - 'Is_alone' & 'Family_size' : 0.69
  - 'Has_cabin' & 'Pclass' : 0.73

Nous allons considéré une corrélation comme suffisament élevée à partir de $|r| \geqslant 0.7$

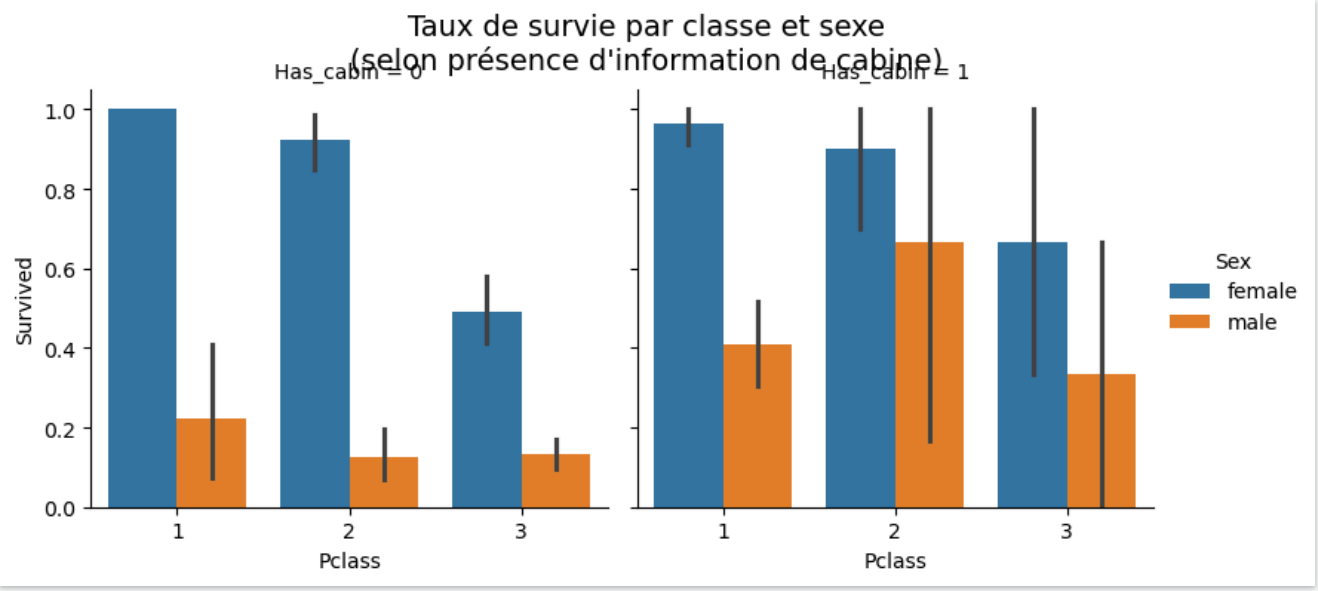
- 'Family_size' est fortement corrélée à 'SibSp' (r = 0.89) et à 'Parch' (r = 0.78). Cela suggère que 'Family_size' est redondante, car elle dérive probablement de ces deux variables. On pourrait envisager d'en supprimer une pour éviter la multicolinéarité.
- 'Is_alone' est inversement lié à 'Family_size' (r = 0.69), ce qui est logique : plus une personne a une grande famille à bord, moins elle est seule. À considérer lors de la sélection de features.
- 'Has_cabin' et 'Pclass' (r = 0.73) pourraient être liées à la richesse. Cette corrélation mérite d'être prise en compte si on utilise ces deux variables dans un modèle.

Détection d'outliers

- **Age** : Grâce à la méthode quantile, on a découvert 11 outliers pour la colonne 'Age' , tous dans le groupe d'age de vieux. Mais aucun ne semble avoir une valeur suspecte donc ce sont peut être des valeurs réels.
- **Fare :** La variable `Fare` présente une distribution très asymétrique, avec plusieurs valeurs extrêmement élevées (jusqu'à 512). L'analyse par la méthode des quartiles (IQR) confirme la présence d'outliers, principalement situés dans la partie haute de la distribution. Ces valeurs correspondent souvent à des passagers de 1re classe ayant payé des billets premium. Elles ne sont pas erronées, mais peuvent biaiser certains modèles. Selon le type de modélisation, ces valeurs pourront être conservées, transformées (via une échelle logarithmique), ou isolées dans une variable catégorique indiquant un "haut tarif

Relation entre valeurs manquantes et la target

Cabin Nan vs target



On constate que:

- **femmes** : elles ont a peu près les mêmes chances de survie avec ou sans cabin
- **hommmes** : les hommes mal enregistrés ont beaucoup moins de chance de survivre par rapport à ceux qui ont été bien enregistrés.