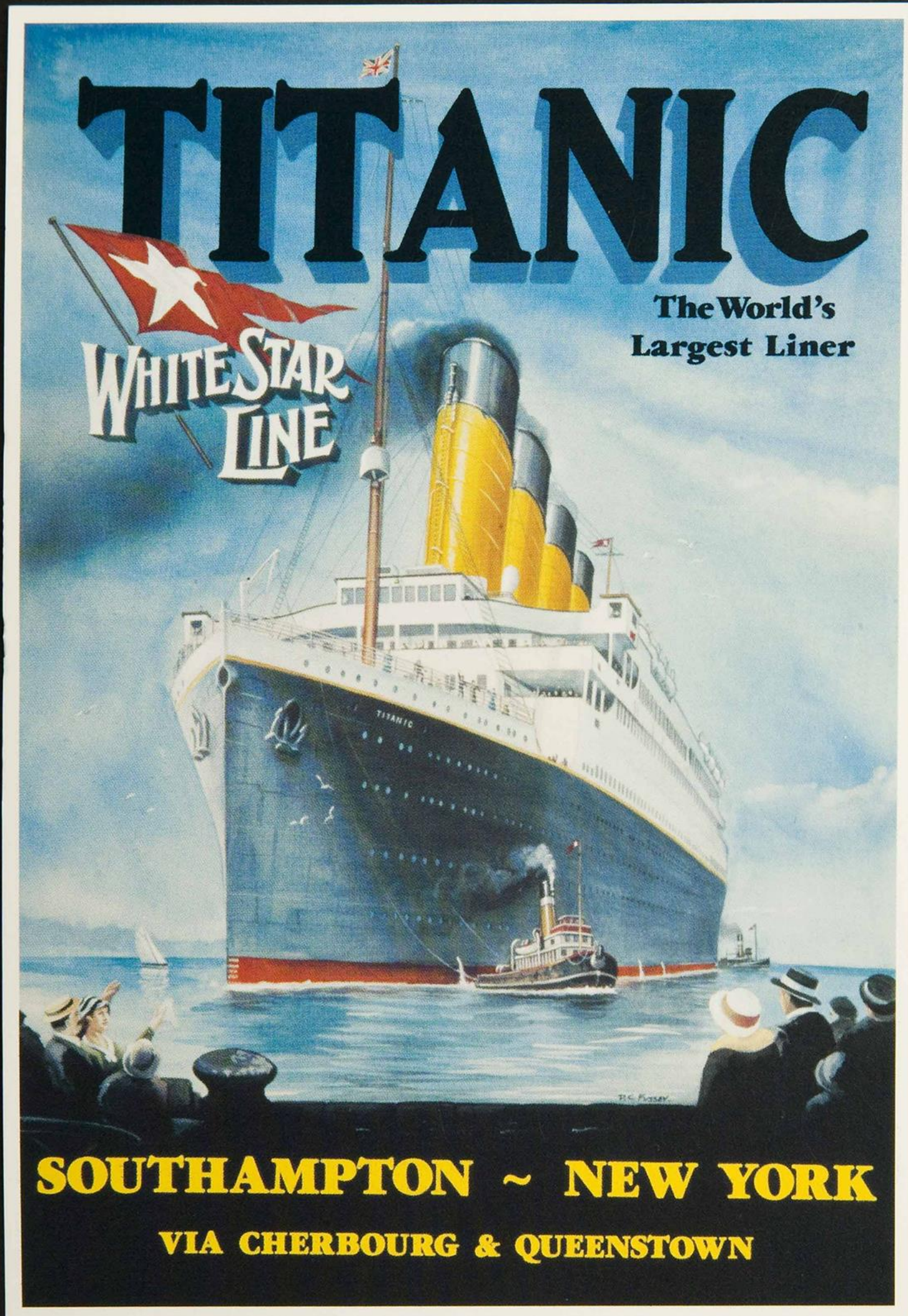

TITANIC
The World's
Largest Liner
WHITE STAR
LINE
TITANIC
SOUTHAMPTON ~ NEW YORK
VIA CHERBOURG & QUEENSTOWN

Rapport Projet Titanic

Damien / Jason

1. Exploration, correction et nettoyage du dataset

1.1 Exploration du dataset

Pour l'exploration du dataset nous n'avons rien fait de spécial de plus que d'habitude si ce n'est les commandes de base d'exploration, describe, head, dtypes.

	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

Voilà l'entête du dataset et l'explications pour chaque colonne (features) sachant que notre target lors de ce projet est "Survived" :

PassengerID : L'ID (unique) du passager, l'index.

Survived : Indique la survie ou non du passager, valeur binaire 0 (non survécu) ou 1 (survécu) type INT

Pclass : Classe du ticket du passager, 1ère (1), 2ème (2), 3ème (3) type INT

Name : Nom du passager, type STRING.

Sex : Sexe du passager, homme ou femme type STRING

Age : Âge du passager type INT

SibSp : Nombre de frères/soeurs ou conjoints du passager à bord

Parch : Nombre de parents/enfants du passager à bord

Ticket : Numéro du ticket

Fare : Prix du billet du passager

Cabin : Numéro de la cabine du passager

Embarked : Port d'embarcation du passager : C = Cherbourg, Q = Queenstown, S = Southampton

1.1 Correction et nettoyage

Nous avons décidé de supprimer les colonnes 'Cabin','Ticket','Name','PassengerId' sachant qu'elles ne nous sont pas utiles pour notre utilise et ne seront de toute façon pas pertinente pour une régression.

Pour les valeurs manquantes, il y en avait 177 dans Age et 2 dans Embarked.

Pour l'âge nous avons divisé le nombre de valeurs manquantes par 3 (59), et répartis équitablement les valeurs entre la moyenne, la moyenne + l'écart type, et moyenne-écart-type.

Et pour Embarked, sachant qu'il n'y a que 2 valeurs manquantes et que la valeur la plus présente dans cette colonne est 'S', les valeurs ont alors été remplacés par S.

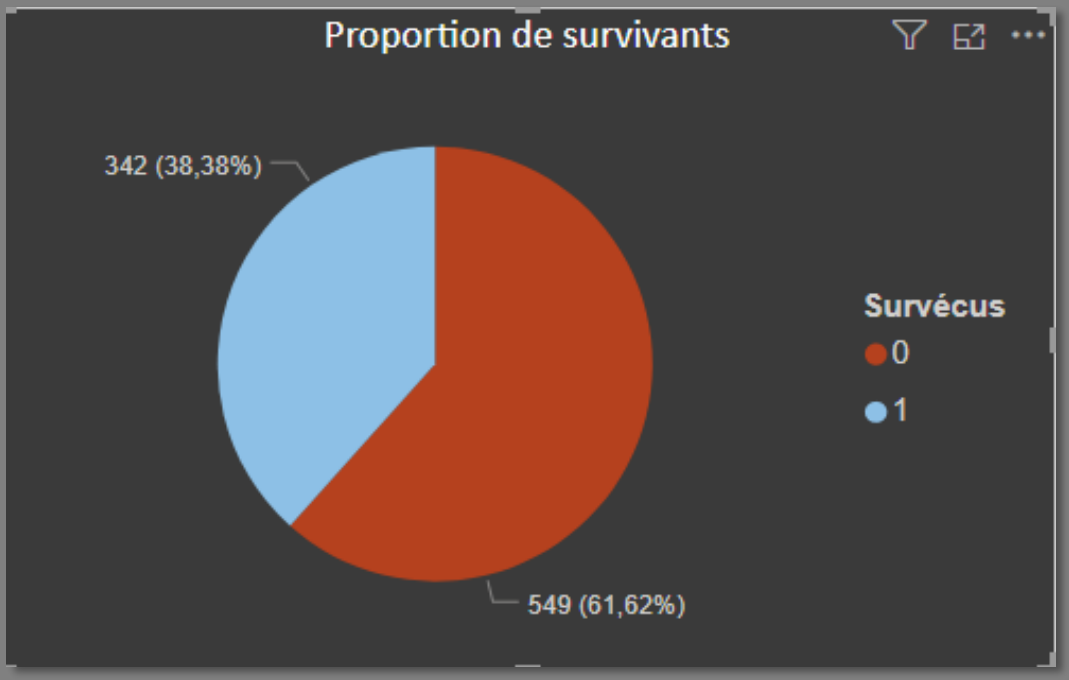
Pour finir d'explorer un peu plus le dataset, nous avons fait une matrice de corrélation

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	0.012658
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000

Malheureusement cette matrice de corrélation ne nous indique pas grand-chose de particulier si l'on regarde notre target Survived, il n'y a quasiment rien de corrélé si ce n'est "Pclass" qui lui est légèrement corrélé négativement.

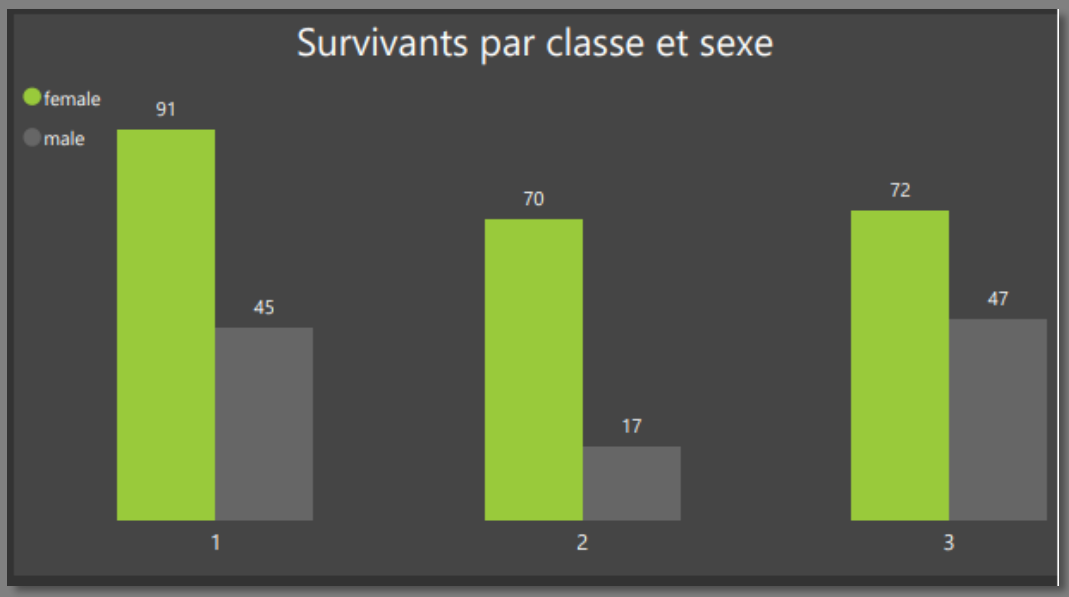
2. Exploration visuel des données

Proportion du nombre de survivants



Ici nous avons le nombre et la proportion de survivants. 0 étant affecté à "Non survécu" et 1 "Survécu. Il y a donc 61% de personnes qui ont péries sur le naufrage parmi les 891 total.

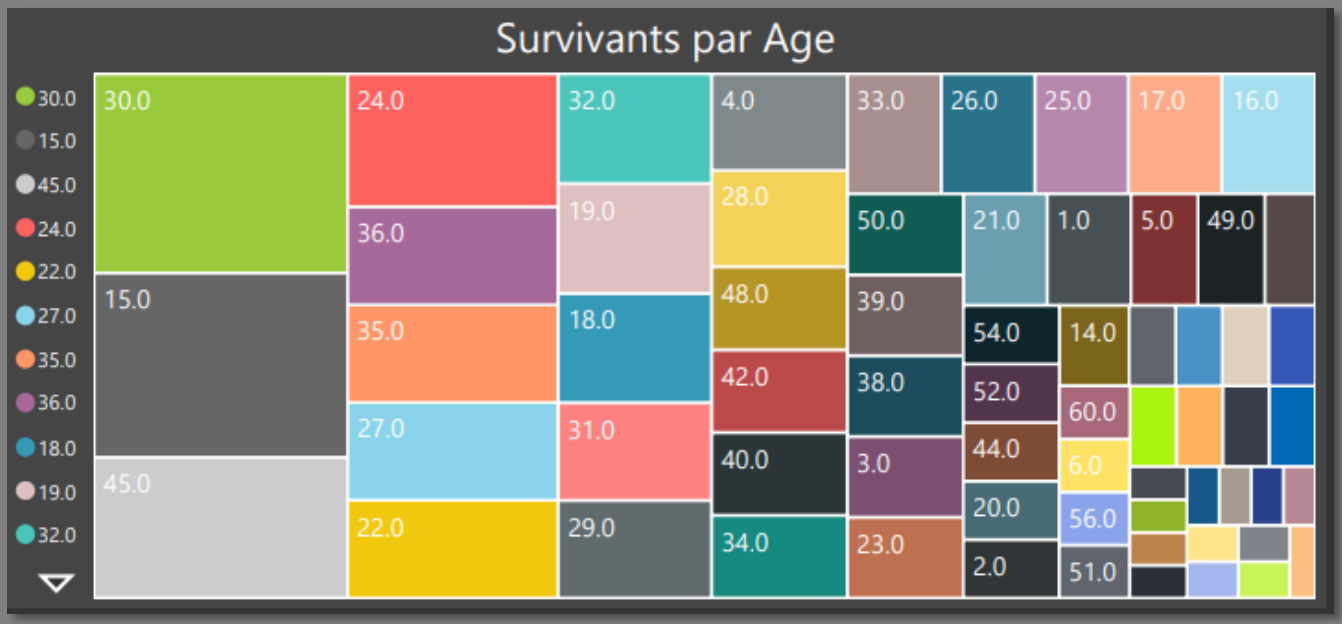
Nombre de survivants par classe et par sexe



Un diagramme un peu plus représentatif et qui note bien une légère tendance des classes mais surtout des sexes. On remarque qu'il y a toujours bien plus de femmes survivantes que d'homme peu importe la classe du billet.

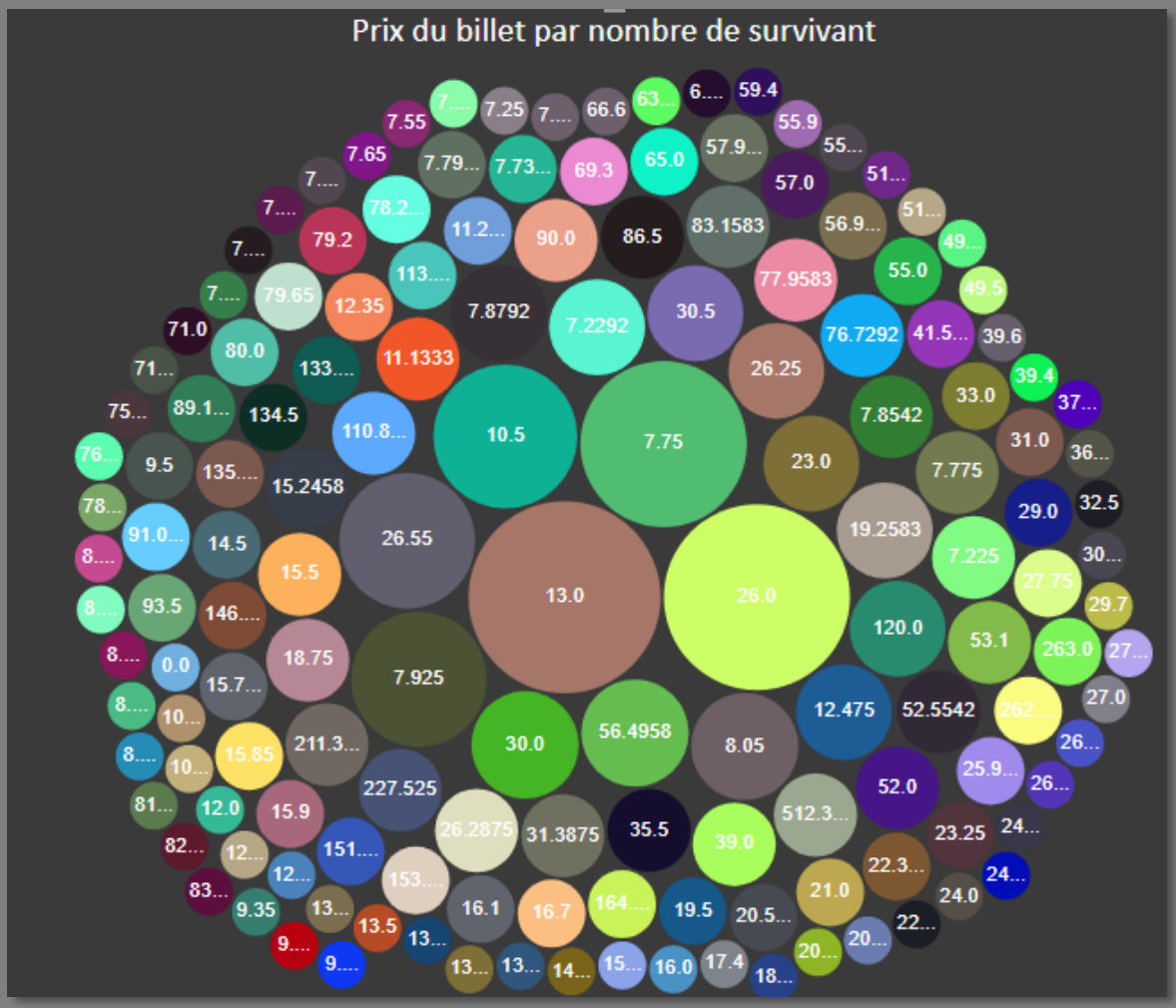
Au niveau des classes, on ne remarque aucune réelle différence entre les 2 et 3ème classe, alors que les premières classes, pourtant bien moins nombreuses que les 2 et 3ème dans les données, ont plus de survivants. Il y aurait une endance de chance de survie supérieur en étant une femme en 1ère classe.

Treemap des survivants par âge



Treemap des survivants par âge par simple but de visualistion générale de la tendance de l'âge des survivants, on peut rapidement observer que la majorité des survivants sont des adultes.

Nombre de survivant par prix du billet



Nous voulions rapidement observer si le prix du pouvait avoir une tendance sur la survie de la personne, on peut observer qu'au final pas tant que ça, là ou il y a plus de survivants sont les billets qui coutent le moins cher.

Attention cependant, ce graphique n'est pas exhaustif et soumis à un souci de proportionnalité. En effet, certes on observe qu'il y a plus de survivants avec des

billets moins chers, mais il y avait également bien + de personne avec des billets peu chers, et moins avec des billets chers. Donc on ne sait pas vraiment s'il y avait "autant" de chance de survie peu importe le prix du billet avec cette seule observation.

3. Tests d'Hypothèse

3.1 Les femmes et les enfants d'abord

3.1.1 Les enfants ont été privilégiés lors du naufrage

Question : Les enfants ont-ils été privilégiés lors du naufrage.

Hypothèse H0 : les enfants n'ont pas été privilégiés.

Hypothèse H1 : les enfants ont été privilégiés.

Test utilisé : comparaison de moyennes de 2 échantillons (Test de Student).

Conditions : les données suivent une loi Normale, le nombre d'échantillons > 30

Etapes :

- Sélection des données à comparer :
    - Ages des enfants.
    - Ages de l'ensemble des passagers.
- Nettoyage des données.
- Application du Test de Student

Conclusion :

La p-value est égale à 3.907198e-56, ce qui est largement inférieur à 0.05. On peut donc rejeter l'hypothèse H0 et conclure, que les moyennes des enfants ayant été sauvés et de toutes les personnes sauvées ne sont pas identiques.

3.1.2 Les femmes ont été privilégiées lors du naufrage

Question : Les femmes ont-elles été privilégiées lors du naufrage.

Hypothèse H0 : les femmes n'ont pas été privilégiées.

Hypothèse H1 : les femmes ont été privilégiées.

Test utilisé : Test de comparaison à une proportion (Test de Pearson).

Etapes :

- Sélection des données à comparer :
  - Nombre de femme sauvés.
  - Total des passagers.
- Nettoyage des données.
- Application du Test de Pearson.

Conclusion :

La p-value est égale à 1.005984e-11, ce qui est largement inférieur à 0.05. On peut donc rejeter l'hypothèse H0 et conclure, que les femmes ont été privilégiées lors du naufrage.

3.2 Influence du prix du billet

3.2.1 Le prix du billet a une influence sur la survie d'un passager

Question : Le prix du billet a une influence sur la survie d'un passager.

Hypothèse H0 : Le prix du billet a une influence.

Hypothèse H1 : Le prix du billet a une influence.

Test utilisé : comparaison de moyennes de 2 échantillons (Test de Student).

Conditions : les données suivent une loi Normale, le nombre d'échantillons > 30

Etapes :

- Sélection des données à comparer :

- Prix du billet des survivants.
- Prix du billet de l'ensemble des passagers.

- Nettoyage des données.
- Application du Test de Student

Conclusion :

La p-value est égale à 3.632247e-08, ce qui est largement inférieur à 0.05. On peut donc rejeter l'hypothèse H0 et on peut conclure que, le prix du billet a une influence sur la survie d'un passager.

4.Régression linéaire

Premièrement un changement dans le dataset a été effectué pour passer du qualitatif au quantitatif sur certaines colonnes pour les adapter à la régression linéaire

Un remplacement a été effecuté dans la feature "Sex", en remplacant les "male" par 0 et les "female" par 1. Pareil pour la feature "Embark", 0 pour S, 1 pour C et 2 pour Q.

Pour ces régressions linéaires, j'ai à chaque fois pris 70% du dataset comme training set, et les 30% restants comme test set.

4.1 Régression linéaire à une seule variable explicative

Avec l'âge comme seule feature :

```
R² = : 0.010903361918819243
RMSE =  0.48486914365225775
```

On obtient un R^2 de 0.01 et une erreur quadratique moyenne de 0.48 (le R^2 est censé tendre vers 1 pour une meilleure qualité de modèle, et le RMSE vers 0 pour une erreur minimisée) ce qui est très mauvais et très loin d'être un bon modèle. L'âge seule ne convient donc pas vraiment.

Avec le sexe comme seule feature :

```
R² = : 0.19723096014166897
RMSE =  0.4087163303408202
```

Avec le sexe on obtient un résultat quand même plus convaincant qu'avec seulement l'âge, mais ça reste néanmoins médiocre pour un modèle.

Avec la classe comme seule feature :

```
R² = : 0.08537822829131647
RMSE =  0.45791374277467006
```

Avec le prix du billet comme seule feature :

```
R² = : 0.0210893377936906
RMSE =  0.4704628540897182
```

Et la classe et le prix du billet font quasiment pareil qu'avec seulement l'âge, c'est à dire un très mauvais score.

4.1 Régression linéaire multivariée

Ici nous avons pris toutes les features dans la partie d'avant et les avons toutes réunies pour faire une régression linéaire multivariée.

```
R² = : 0.3376719572365149
RMSE =  0.38252608540492467
```

Le score est légèrement mieux en mettant toutes les features ensemble, mais on observe bien que ce n'est tout de même pas assez et qu'il faudrait changer de solution de choix de modèle.

On peut voir que tous les score, que ça soit avec une seule feature ou toutes les features en même reste très bien quoiqu'il arrive en régression linéaire. Il n'arrive pas à prédire de façon efficace et fait moins bien que le hasard. Ceci est sûrement dû au fait que la régression linéaire n'est pas du tout adaptée à ce type de problème qui est au final un problème de classification.

On cherche à savoir si tel ou tel passager, selon plusieurs critères déterminant (ou pas) à bien plus de chance de survivre (1) ou de ne pas survivre (0), c'est donc un problème de régression logistique "simple" binaire avec deux résultats possibles.

Le problème vient du fait que, la régression linéaire qui est censé prédire une variable quantitative grâce à d'autres variables quantitatives n'y arrive car une conversion a été faite, les données ont été transformés et ça n'est pas adapté.

5.Régression Logistique

L’objectif est de trouver un modèle qui nous permettra de prédire si un passager est mort ou vivant en se basant sur les données du dataset.

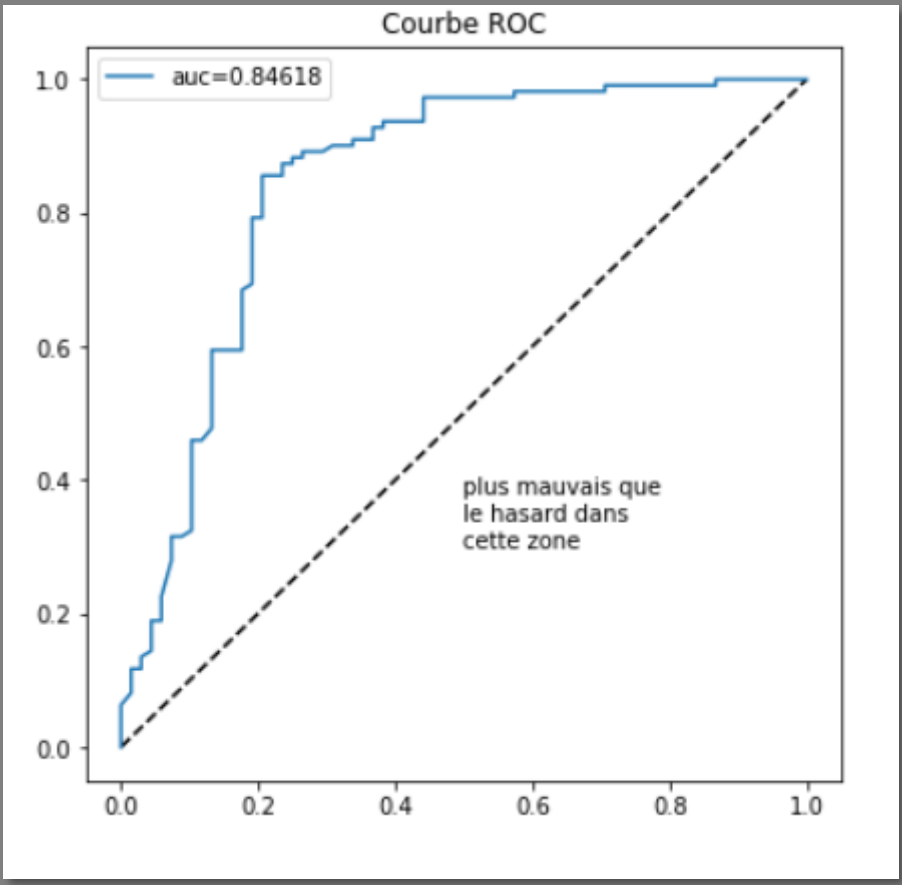
Etapes :

- Importation des données : read.csv
- Analyse des données :
    - On constate qu’il y a beaucoup de valeurs manquantes dans la colonne Age et Cabin.
    - 5 colonnes sur les 12 sont qualitatives.
- Choix des colonnes :
    - Choix de la Target y et des data X :
        - Y : l’objectif de l’exercice est de prédire, si un passager est mort ou vivant. Nous avons sélectionné la Colonne Survived.
        - X : Le reste des features.
    - Suppression des faetures de X non pertinentes pour notre analyse :
        - Cabin : Trop de valeurs manquantes et valeurs uniques.
        - PassengerId : Valeurs uniques.
        - Name : Valeurs uniques.
        - Ticket : Valeurs uniques.
- Nettoyage des données :
    - Colonne Age : En se basant sur la moyenne et l’écart type des valeurs de la colonne (15,30,45). Nous avons remplacé les valeurs manquantes, par tranches.
    - Colonnes Sex et Embarked : Transformation des valeurs qualitatives en quantitatives, via LabelEncoder de Sklearn.
    - Utilisation de RFECV de Sklearn pour terminer la sélection.
- Entrainement du modèle : LogisticRegression (algorithme imposé).
- Evaluation du modèle :
    - Score : 82% résultat correct, qui pourrait être optimisé.

- Matrice de confusion :

	Predis Vrai	Predis Faux
Actuels Vrai	99	12
Actuels Faux	19	49

- Justesse (Accuracy score) : 0.84



- Application du modèle au données test fournit et exports du résultat en CSV.