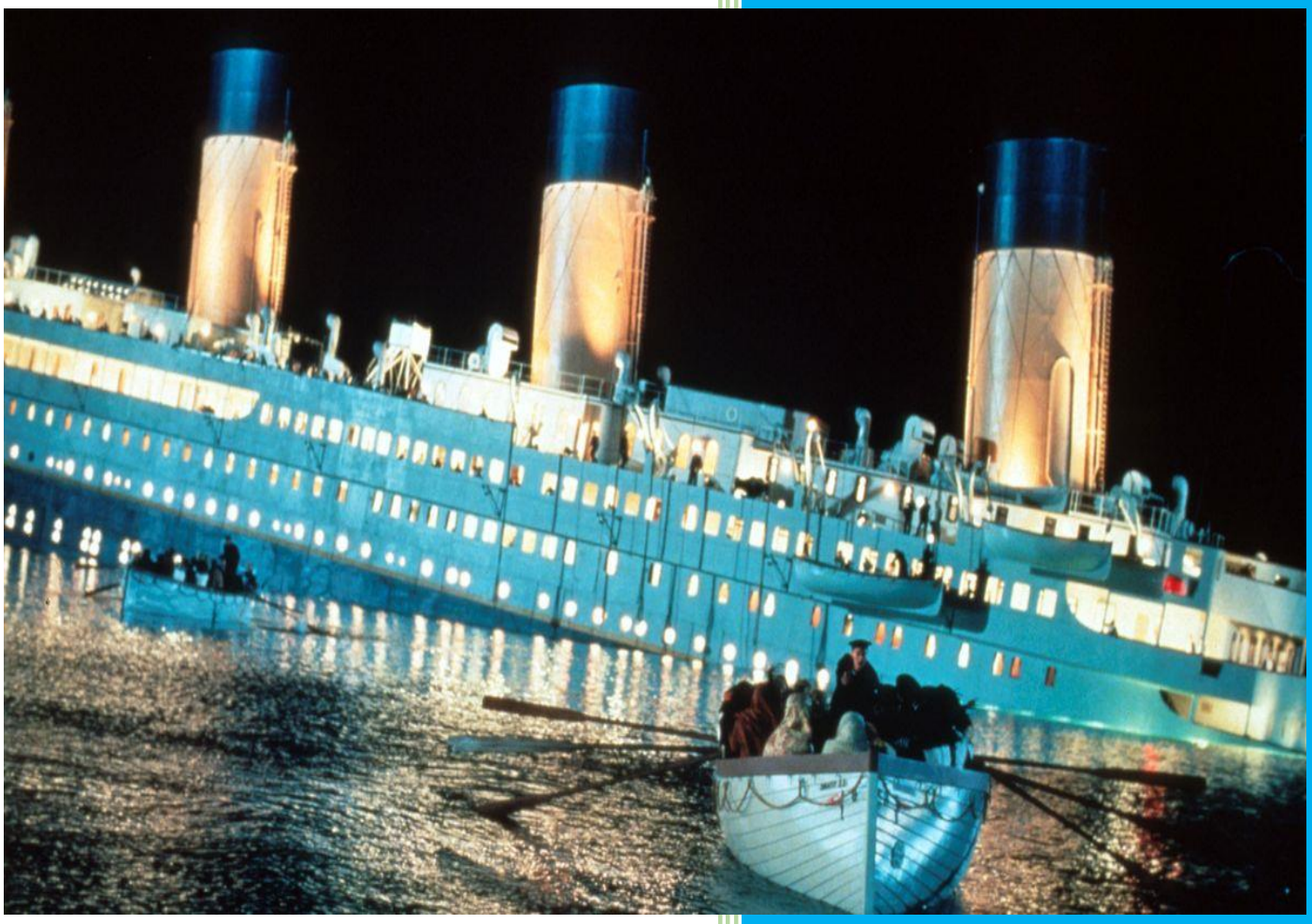
SAWADOGO
Amadou

Projet SAS sur les données Titanic



M1 – MIASHS - MQME

Contents

Introduction

Afin de comprendre les critères à remplir pour avoir plus de chance de survivre au naufrage survenu avec le bateau Titanic, cette étude est ainsi réalisée. Pour étudier cette base de données, nous utiliserons les outils de SAS. **Les codes seront placés en fichier joint en pdf**. Dans la suite de notre travail nous allons faire d'abord une analyse descriptive puis une analyse statistique pour chacune des variables afin de déterminer les tendances générales des différentes observations dans notre base de données. Enfin nous complèterons avec le tableau rempli des codes **SAS** sans la procédure **SQL**.

Partie 1

L'objectif de cette partie est tout d'abord de reconstituer le fichier Train qui est une partie du célèbre fichier Titanic. Ce fichier est divisé en trois sous-fichiers que je vous propose de reconstituer et ensuite d'effectuer une étude statistique du fichier en entier.

Etape 1 : Importation

Nous avons effectué une concaténation entre train_url1, train_url2 et train_url3 pour créer un fichier nommé train_url qui regroupe toutes les trois bases de données en une seule base de données que nous travaillerons là-dessus.

Etape 2 : Analyse descriptive du fichier Train_url

Rapport sur l'état d'embarquement et le nombre de survivants

Tout d'abord, nous avons décidé de faire un bilan après le naufrage. On remarque ainsi qu'il y a presque le double des personnes embarquées qui n'ont pas survécu à ce drame d'après l'illustration du tableau ci-dessous.

Survived				
Survived	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	543	61.43	543	61.43
1	341	38.57	884	100.00

Figure 1: Tableau de survie

Par la suite, nous allons faire le point à l'embarquement, par genre et par classe afin de bien visualiser nos données.

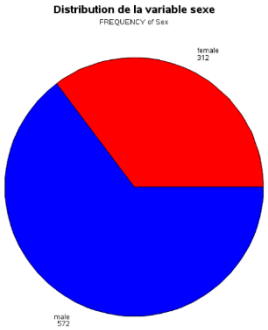


Figure 2: Camembert sur le genre

Embarked	Nombre_Passager
	2
Q	76
C	167
S	639

Figure 3: Tableau du point d'embarquement

Nous remarquons qu'il y a deux (02) plus d'homme qui ont embarqué par rapport aux femmes.

Concernant les quais d'embarquement, la majorité des personnes ont embarqué dans la quais **S** tandis qu'une poignée de personnes ont embarqué à partir de la quais **Q**.

Etape 3 : Analyse descriptive avancée

Dans cette partie, nous allons réaliser les tableaux croisés dynamiques afin de mieux étayer nos idées sur les chances de survie par rapport à l'âge, le genre, le point d'embarquement et la classe.

Pour la variable âge et genre

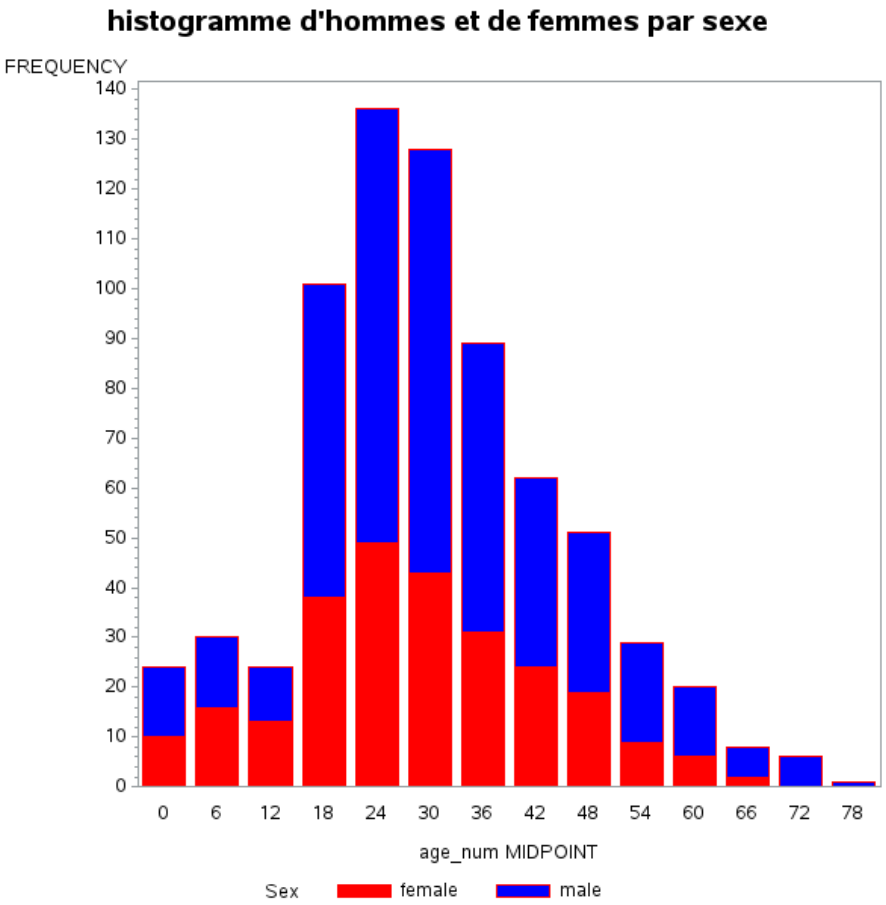


Figure 4: Répartition de l'âge par rapport au genre

Nous remarquons que les enfants âgés de 0 à 12 ans, il y a un peu plus de femmes par rapport aux hommes contrairement au reste ou il y a plus d'hommes.

Nous allons faire une boîte à moustache afin de mieux voir la répartition de nos données.

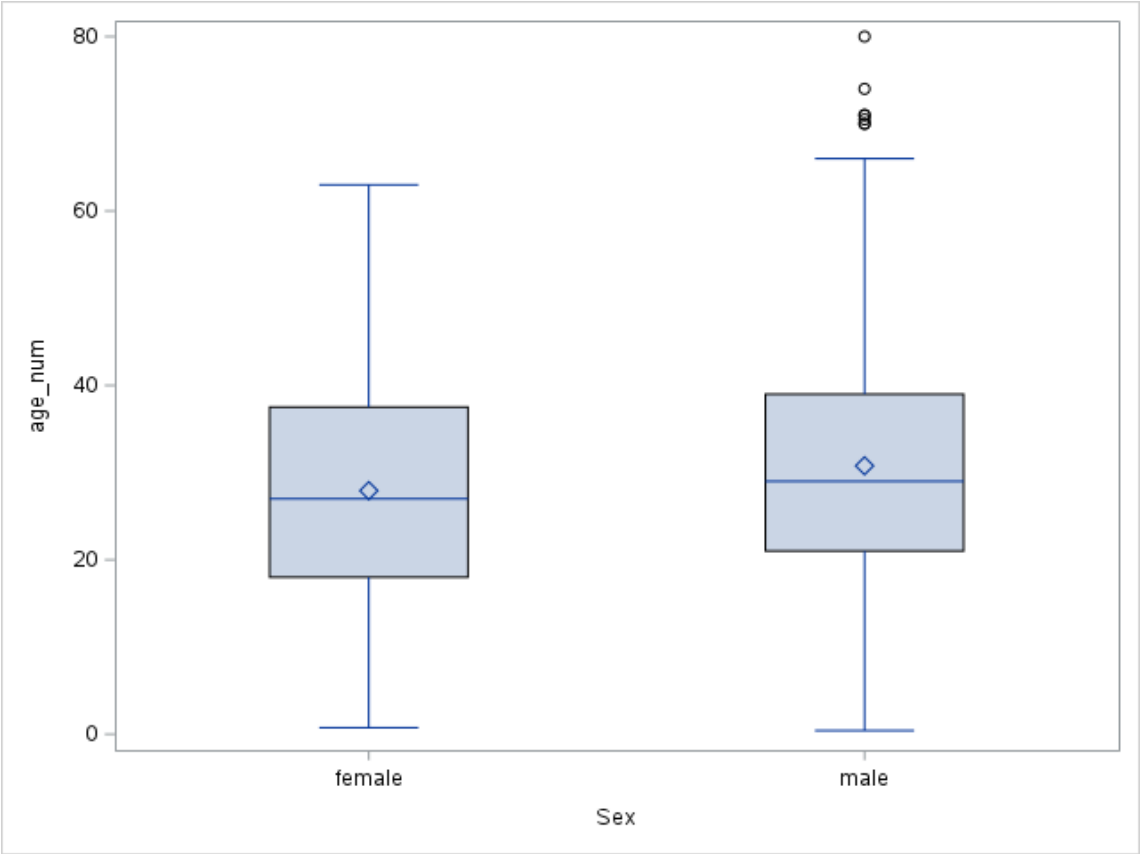


Figure 5: Box plot âge et genre

Nous remarquons que les données sont bien réparties, l'âge moyen est égale peu importe le genre et qu'il y a aussi des données aberrantes chez les hommes.

Pour la variable fare (prix du ticket) et la classe

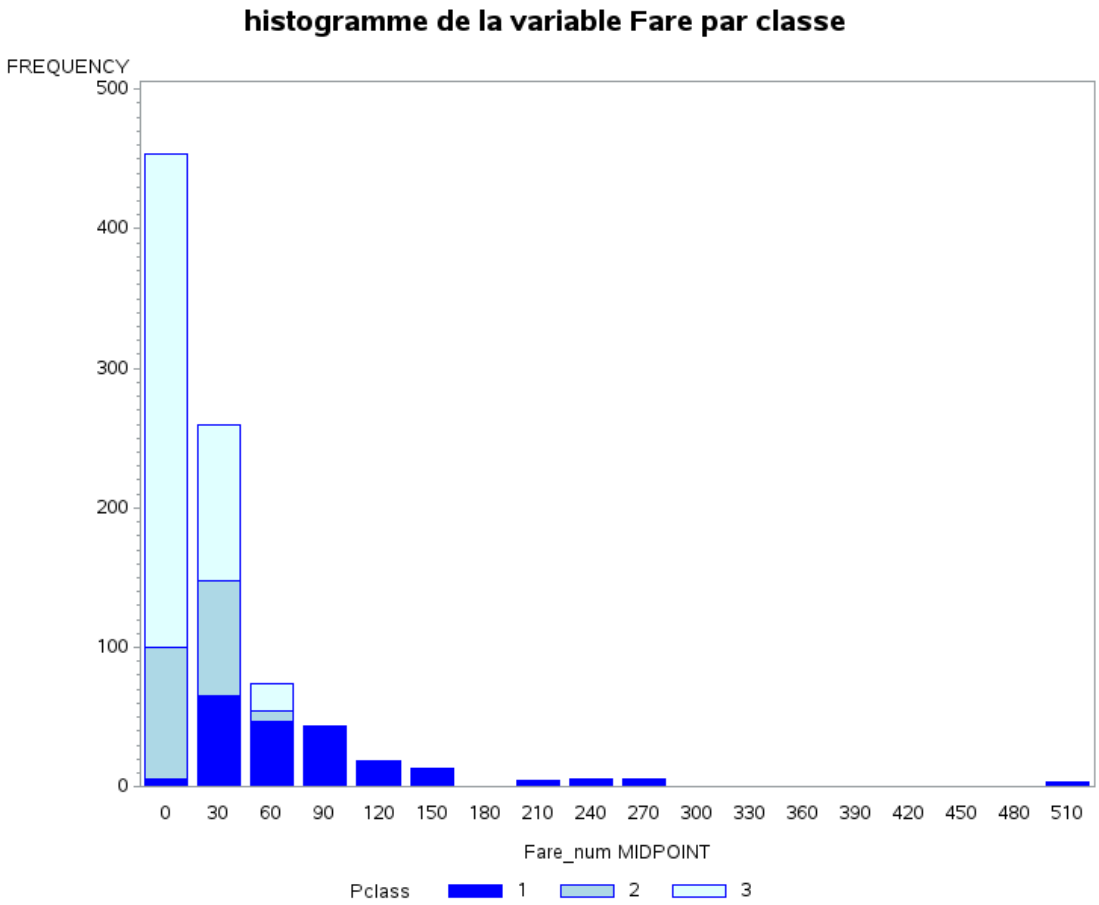


Figure 6: Histogramme de fare et classe

Nous remarquons ainsi que les tickets qui coutent moins de 30$ ont été achetés par la classe 3, pour ceux de 30$ plus par la classe 2 et les tickets les plus chers par la classe 1.

Pour voir la distribution en fonction des valeurs des tickets, nous allons faire une boîte à moustache.

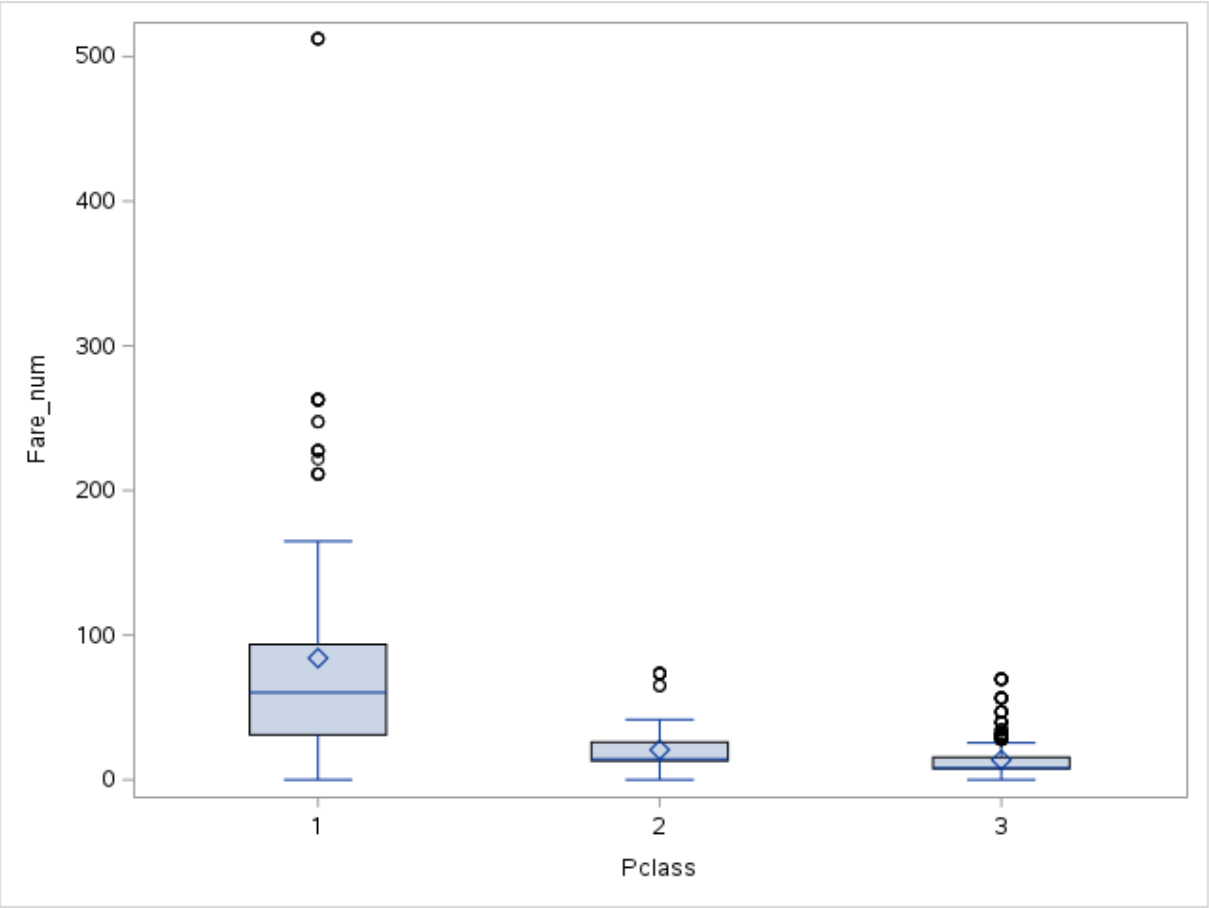


Figure 7: Box plot fare et classe

Nous remarquons que les données sont bien réparties, le prix moyen du ticket diffère énormément par rapport à chaque classe et qu'il y a aussi des données aberrantes dans chaque classe.

Les chances de survie par rapport aux autres variables

Nous allons tout d'abord **les chances de survie par rapport au genre.** Pour ce faire nous avons le tableau suivant :

Table of Survived by Sex			
	Sex(Sex)		
Survived(Survived)	female	male	Total
0	79 8.94 14.55 25.32	464 52.49 85.45 81.12	543 61.43
1	233 26.36 68.33 74.68	108 12.22 31.67 18.88	341 38.57
Total	312 35.29	572 64.71	884 100.00

Figure 8: Pourcentage de survivants par genre

Sur l'ensemble des personnes à bord, le dixième du nombre de femmes ont péri avec un quart de survivants contrairement aux hommes qui se retrouvent avec la moitié des décès et un dixième de survivants sur l'ensemble de la population.

En plus, seul 15% des femmes ont péri contre 85% chez les hommes.

Nous pouvons conclure que les femmes ont nettement plus de chances de survivre à ce naufrage par rapport aux hommes.

Nous allons voir maintenant **les chances de survie par rapport au quai d'embarquement.** Pour ce faire nous avons le tableau suivant :

Table of Survived by Embarked				
	Embarked(Embarked)			
Survived(Survived)	C	Q	S	Total
0	75 8.50 13.81 44.91	46 5.22 8.47 60.53	422 47.85 77.72 66.04	543 61.56
1	92 10.43 27.14 55.09	30 3.40 8.85 39.47	217 24.60 64.01 33.96	339 38.44
Total	167 18.93	76 8.62	639 72.45	882 100.00
Frequency Missing = 2				

Figure 8: Pourcentage de survie par rapport au quai d'embarquement

Sur l'ensemble des personnes à bord, seulement **5% et 8%** des personnes embarquées respectivement à **la quai Q et C** ont **péri** avec respectivement **3% et 10% de survivants** contrairement à **la quai S** qui se retrouvent avec **la moitié des décès** et **le quart de survivants** sur l'ensemble de la population.

En plus, seul **22% des quais C et Q** combinées ont péri contre **78% de la quai S**.

Nous pouvons conclure que les quais d'embarquement **C et Q donnent énormément plus de chances de ne pas mourir** à ce naufrage par rapport à la **quai S**.

Nous allons tout ensuite **les chances de survie par rapport au genre et la classe.** Pour ce faire nous avons le tableau suivant :

Table 1 of Sex by Pclass				
Controlling for Survived=0				
	Pclass(Pclass)			
Sex(Sex)	1	2	3	Total
female	2 0.37 2.53 2.53	6 1.10 7.59 6.19	71 13.08 89.87 19.35	79 14.55
male	77 14.18 16.59 97.47	91 16.76 19.61 93.81	296 54.51 63.79 80.65	464 85.45
Total	79 14.55	97 17.86	367 67.59	543 100.00

Figure 9: Tableau par genre et par classe pour les non survivants

Table 2 of Sex by Pclass				
Controlling for Survived=1				
	Pclass(Pclass)			
Sex(Sex)	1	2	3	Total
female	91 26.69 39.06 67.41	70 20.53 30.04 80.46	72 21.11 30.90 60.50	233 68.33
male	44 12.90 40.74 32.59	17 4.99 15.74 19.54	47 13.78 43.52 39.50	108 31.67
Total	135 39.59	87 25.51	119 34.90	341 100.00

Figure 10: Tableau par genre et par classe pour les survivants

La figure 9, nous reflète que :

Sur les 543 personnes à bord qui ont péri, seulement 2 femmes en première classe sont mortes ainsi que 6 personnes de la deuxième classe. Pour la troisième classe, il y a autant de mort coté femme qu'homme. On remarque 296 morts chez les hommes de la troisième classe soit plus de la moitié des morts.

La figure 10 reflète :

Sur les survivants, 68% sont de sexe féminin avec presque la même chance de survie peu importe la classe (environ 21%) contre 32% d'homme ayant survécu avec 13% pour les classes 1 et 3 (chacune) qui vaut le double de ceux de la classe 2.

Comme ce dicton le dit « Les femmes et les enfants d'abord », nous allons tout enfin **les chances de survie par rapport à la catégorie d'âge.** Pour ce faire nous avons le tableau suivant :

Table of age_Categorie by Survived			
	Survived(Survived)		
age_Categorie	0	1	Total
Adulte	369 41.74 61.81 67.96	228 25.79 38.19 66.86	597 67.53
Enfant	174 19.68 60.63 32.04	113 12.78 39.37 33.14	287 32.47
Total	543 61.43	341 38.57	884 100.00

Figure 11: Pourcentage de survivants et de non survivants par catégorie d'âge

Sur l'ensemble des personnes, 42% des adultes ont péri et le quart a survécu tandis que chez les enfants, 19% ont péri contre 13% qui ont survécu.

Chez les adultes uniquement, environ 38% des personnes adultes ont survécu.

Cependant, chez les enfants, nous avons le même pourcentage de survivants soit 39%.

Partie 2

Le but de cette partie est de réécrire les procédures en langage uniquement.

PROC SQL; SELECT * FROM CLIENTS WHERE cpcli <> 57500; QUIT;	DATA NOUVEAU_CLIENTS; SET CLIENTS; WHERE cpcli ne 57500; RUN;
PROC SQL; SELECT * FROM Clients WHERE villecli= "Saint-Avold" AND sexe= "H" AND age < 23; QUIT;	DATA NOUVEAU_CLIENTS; SET CLIENTS; WHERE villecli='Saint-Avold' AND sexe='H' AND age<23; RUN;

1. Restriction: where, upper, lower, between, is null

PROC SQL; SELECT * FROM Clients WHERE Age BETWEEN 18 AND 21; QUIT;	DATA NOUVEAU_CLIENTS; SET CLIENTS; WHERE Age >= 18 AND Age <= 21; RUN;
PROC SQL; SELECT * FROM Clients WHERE Age BETWEEN 18 AND 21; QUIT;	DATA NOUVEAU_CLIENTS; SET CLIENTS; WHERE Age >= 18 AND Age <= 21; RUN;

2. Restriction: where, in, like

PROC SQL; SELECT * FROM Clients WHERE sexe IN ('H','F'); QUIT;	DATA NOUVEAU_CLIENTS; SET CLIENTS; WHERE sexe IN ('H', 'F'); RUN;
PROC SQL; SELECT * FROM Clients WHERE Nomcli LIKE 'L%'; QUIT;	DATA NOUVEAU_CLIENTS; SET CLIENTS; WHERE substr(Nomcli, 1, 1) = 'L'; RUN;

3. Gestion des doublons : Distinct et renommage

PROC SQL; SELECT DISTINCT Villecli FROM Clients; QUIT;	PROC SORT DATA=Clients; BY Villecli; RUN; DATA NOUVEAU_CLIENTS; SET Clients; BY Villecli; IF FIRST.Villecli THEN OUTPUT; RUN;
PROC SQL; SELECT DISTINCT Villecli as ville, sexe FROM clients;	PROC SORT DATA=Clients; BY Villecli Sexe; RUN;

QUIT;	DATA NOUVEAU_CLIENTS (KEEP=Ville Sexe); SET Clients; BY Villecli Sexe; IF FIRST.Sexe THEN OUTPUT; RUN;

4. Les jointures : INNER JOIN - LEFT JOIN

PROC SQL; SELECT * FROM Clients cl INNER JOIN Locations Lo ON cl.codecli=lo.codecli INNER JOIN Films fi ON lo.codefilm =fi.codefilm WHERE cl.cpcli=57800; QUIT;	DATA CL; SET Clients; WHERE cpcli=57800; RUN; DATA LO; SET Locations; WHERE codecli IN (SELECT codecli FROM CL); RUN; DATA NOUVEAU_FILMS; MERGE LO(in=inlo) Films(in=infi); BY codefilm; IF inlo AND infi; RUN;
PROC SQL; SELECT * FROM Clients cl LEFT JOIN Locations Lo ON cl.codecli=lo.codecli INNER JOIN Films fi ON lo.codefilm =fi.codefilm WHERE fi.codefilm IN (1,5,7); QUIT;	DATA FILMS; SET Locations (WHERE=(codefilm IN (1,5,7))); RUN; DATA NOUVEAU_CLIENTS; MERGE Clients(in=incl) FILMS(in=inf); BY codecli; IF incl; RUN; DATA NOUVEAU_FILMS; MERGE FILMS(in=infi) Films(in=iff); BY codefilm; IF infi AND iff; RUN;

5. Requêtes et calculs arithmétiques

PROC SQL; SELECT * FROM produits WHERE (CodeCateg=1 AND PrixUnit > 50) OR (CodeCateg=3 AND PrixUnit < 90); QUIT;	DATA PRODUITS; SET produits; WHERE (CodeCateg=1 AND PrixUnit > 50) OR (CodeCateg=3 AND PrixUnit < 90); RUN;
PROC SQL; SELECT RefProd, PrixUnit *(UnitesStock-5) as soustraction FROM produits WHERE UnitesStock >= 5;	DATA PRODUITS; SET produits; IF UnitesStock >= 5 THEN soustraction = PrixUnit * (UnitesStock-5); DROP UnitesStock PrixUnit;

QUIT;	RENAME soustraction = PrixTotal; RUN;

6. Traitement conditionnel: CASE WHEN

PROC SQL; SELECT Refprod, Nomprod, PrixUnit, CASE WHEN PrixUnit<=50 THEN "Bas prix" WHEN 50 < PrixUnit <=90 THEN "Prix moyens" ELSE "Produits de luxe" END AS Gamme FROM Produits ; QUIT;	DATA PRODUITS; SET Produits; IF PrixUnit <= 50 THEN Gamme = "Bas prix"; ELSE IF PrixUnit <= 90 THEN Gamme = "Prix moyens"; ELSE Gamme = "Produits de luxe"; DROP PrixUnit; RUN;
PROC SQL; SELECT Refprod, UnitesStock, UnitesCom, NiveauReap, CASE WHEN UnitesCom > 40 THEN "Déjà commandé" WHEN UnitesCom < NiveauReap THEN "A Commander " WHEN UnitesCom = 0 THEN "N'est plus en stock " ELSE " Disponible " END AS Informations FROM Produits ; QUIT;	DATA PRODUITS; SET Produits; IF UnitesCom > 40 THEN Informations = "Déjà commandé"; ELSE IF UnitesCom < NiveauReap THEN Informations = "A Commander"; ELSE IF UnitesCom = 0 THEN Informations = "N'est plus en stock"; ELSE Informations = "Disponible"; DROP UnitesCom NiveauReap; RUN;

7. Agrégation

PROC SQL;
SELECT NoFour,
COUNT(*) AS Nombre_de_produits,
ROUND(AVG(PrixUnit),.001) AS PrixMoyen,
MIN(PrixUnit) AS PrixMinimum,
MAX(PrixUnit) AS PrixMaximum
FROM produits
GROUP BY NoFour
ORDER BY NoFour;
QUIT;
	data produits_summary;
	set produits;
	by NoFour;
	
	if first.NoFour then do;
	Nombre_de_produits = 1;
	PrixTotal = PrixUnit;
	PrixMinimum = PrixUnit;
	PrixMaximum = PrixUnit;
	end;
	else do;
	Nombre_de_produits + 1;
	PrixTotal + PrixUnit;
	if PrixUnit < PrixMinimum then PrixMinimum = PrixUnit;
	if PrixUnit > PrixMaximum then PrixMaximum = PrixUnit;
	end;
	
	if last.NoFour then do;
	PrixMoyen = PrixTotal / Nombre_de_produits;
	output;
	end;
	drop PrixUnit PrixTotal;
	run;

	proc sort data=produits_summary;
	by NoFour;
	run;
	
	proc print data=produits_summary;
	var NoFour Nombre_de_produits
	PrixMoyen PrixMinimum PrixMaximum;
	run;

PROC SQL; SELECT NoFour, COUNT(*) AS Nombre_de_produits, ROUND(AVG(PrixUnit)) AS PrixMoyen, MIN(PrixUnit) AS PrixMinimum, MAX(PrixUnit) AS PrixMaximum FROM produits GROUP BY NoFour HAVING Nombre_de_produits < 2 ORDER BY NoFour; QUIT;	PROC SORT DATA=produits; BY NoFour; RUN; PROC MEANS DATA=produits NOPRINT; BY NoFour; VAR PrixUnit; OUTPUT OUT=stats N(COUNT(*))=Nombre_de_produits MEAN= PrixMoyen MIN= PrixMinimum MAX= PrixMaximum; RUN; DATA stats; SET stats; IF Nombre_de_produits < 2; DROP _TYPE_ _FREQ_; RUN; PROC SORT DATA=stats; BY NoFour; RUN;

Conclusion

Dans le but de mener une étude dans de bonnes perspectives à travers une base de données, il est essentiel de connaître les variables ainsi que leurs différentes caractéristiques de tendance. A travers ce rapport, force est de constater que certains facteurs n'influent pas alors qu'à vue d'œil, sans données on dirait plus le contraire comme la chance de survive étant la même que tu sois enfant ou adulte. Il est donc nécessaire d'utiliser les statistiques pour avoir une meilleure appréciation qui sont plus réalistes et permettent une meilleure compréhension.