Abdou NIANG
MIASHS
Université de Lille

Projet SAS

Thème : TITANIC

Notre travail porte sur trois tableaux de données train_url1, train_url2 et train_url3 portant sur les détails des voyageurs du Titanic que voici :

train_url1

Variable	Definition	Key
PassengerId	Identifiant du passager	Primary key
Name	Nom du passager	
survival	Survie	0 = No, 1 = Yes
Pclass (classe)	Classe du ticket	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sexe du pasager	

train_url2

Variable	Definition	Key
PassengerId	Identifiant du passager	Primary key
Age	Age en années	
Fare	Tarif du ticket	
embarked	Port d'Embarquation	C = Cherbourg, Q = Queenstown, S = Southampton

train_url3

Variable	Definition	Key
PassengerId	Identifiant du passager	Primary key
Name	Nom du passager	
survival	Survie	0 = No, 1 = Yes
Pclass (classe)	Classe du ticket	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sexe du pasager	
Age	Age en années	
Fare	Tarif du ticket	
embarked	Port d'Embarquation	C = Cherbourg, Q = Queenstown, S = Southampton

Étape 1 : Importation

Après importation et tri, nous avons rassemblé les trois tableaux en un seul (merge) nommé train_url sur lequel nous avons écrit toutes nos requêtes.

Voilà un petit aperçu du tableau train_url regroupant toutes les variables:

Obs.	PassengerId	Name	Survived	Pclass	Sex	Age	Fare	Embarked	age_num	Fare_num
1	1	Braund, Mr. Owen Harris	0	3	male	22	7.25	S	22.0	7.250
2	2	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	1	1	female	38	71.2833	C	38.0	71.280
3	3	Heikkinen, Miss. Laina	1	3	female	26	7.925	S	26.0	7.925
4	4	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	1	female	35	53.1	S	35.0	53.100
5	5	Allen, Mr. William Henry	0	3	male	35	8.05	S	35.0	8.050
6	6	Moran, Mr. James	0	3	male		8.4583	Q	.	8.458
7	7	McCarthy, Mr. Timothy J	0	1	male	54	51.8625	S	54.0	51.860
8	8	Palsson, Master. Gosta Leonard	0	3	male	2	21.075	S	2.0	21.070
9	9	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	1	3	female	27	11.1333	S	27.0	11.130
10	10	Nasser, Mrs. Nicholas (Adele Achem)	1	2	female	14	30.0708	C	14.0	30.070

ÉTAPE 2 : Analyse descriptive du fichier Train_url

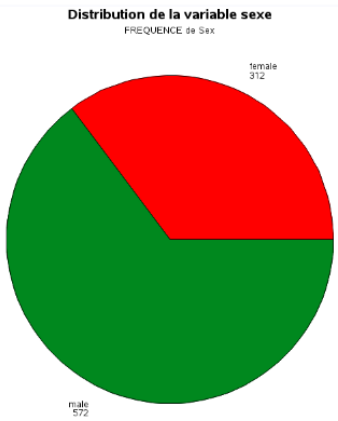
1. Nombre de survivants et non survivants

Survived				
Survived	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	543	61.43	543	61.43
1	341	38.57	884	100.00

2.Nombre de passagers hommes et femmes

Sex				
Sex	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
female	312	35.29	312	35.29
male	572	64.71	884	100.00

- On peut faire un diagramme secteur qui représente mieux les proportions **:**



3. Nombre de passagers par point d'embarquement

Embarked

Embarked	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
C	167	18.93	167	18.93
Q	76	8.62	243	27.55
S	639	72.45	882	100.00

Fréquence manquante = 2

4. Nombre de passagers hommes femmes par classe

C’est un tableau croisé avec les fréquences, les pourcentages et les effectifs marrginaux

La procédure FREQ

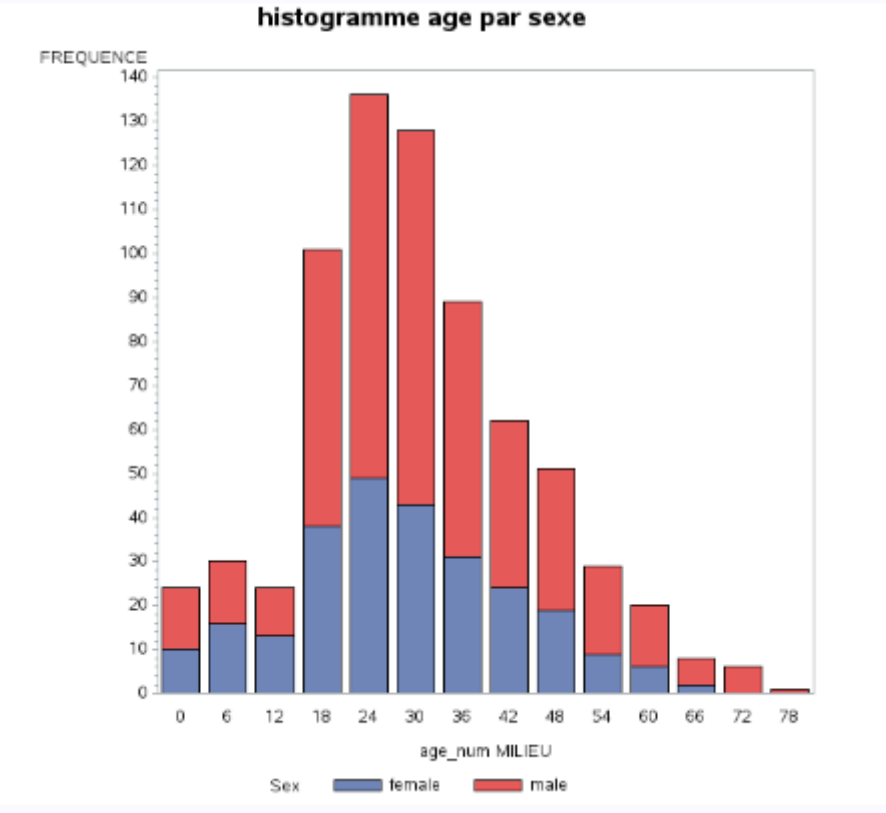
Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table de Sex par Pclass

Sex(Sex)	Pclass(Pclass) 1	2	3	Total
female	93 10.52 29.81 43.46	76 8.60 24.36 41.30	143 16.18 45.83 29.42	312 35.29
male	121 13.69 21.15 56.54	108 12.22 18.88 58.70	343 38.80 59.97 70.58	572 64.71
Total	214 24.21	184 20.81	486 54.98	884 100.00

ÉTAPE 3 : Analyse descriptive avancée

1. Faire un histogramme de la variable Age par sexe. Interpréter le graphique



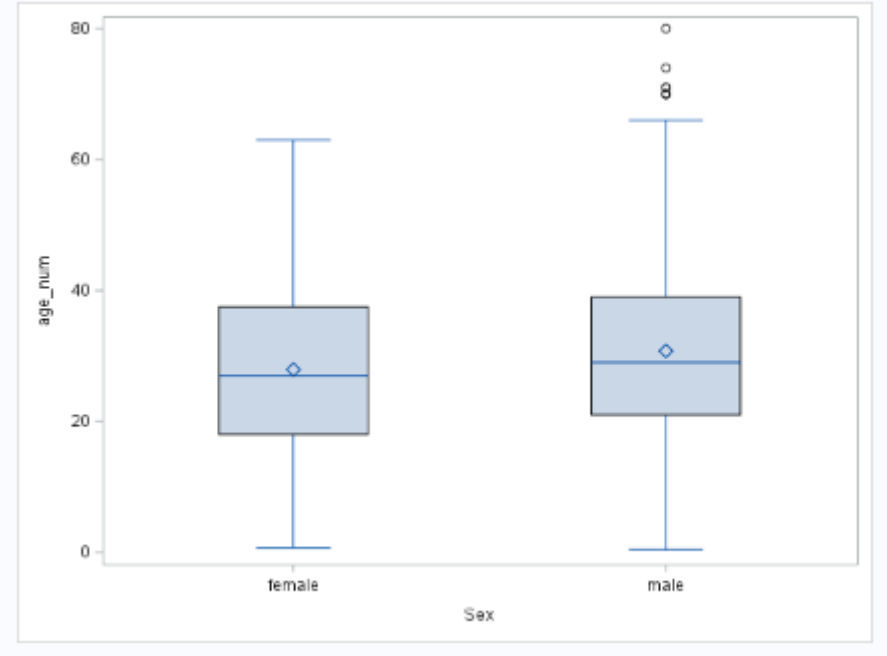
Interprétation de l'histogramme :

La proportion de passagers homme est plus importante que celle des femmes

On constate que la catégorie des personnes agées entre 18 et 40 ans est la plus représentée

La courbe de distribution des ages de passagers est identique à celle de la distribution d'une loi normale centrée réduite

2. Faire une boîte à moustaches de la variable Fare par classe. Interpréter le graphique



Interprétation :

D'abord on peut remarquer que le Boxplot de l'age des femmes présente des outliers (des valeurs aberrantes)

On pourrait pour avoir des données plus cohérents supprimer toutes les valeurs au delà de 65 ans qui semble être la valeur du min(max)*

Ensuite le point bleu dans chaque boite indique la moyenne qui semble être ce qu'on avait trouvé à savoir 27 ans pour les femmes et 30 pour les hommes

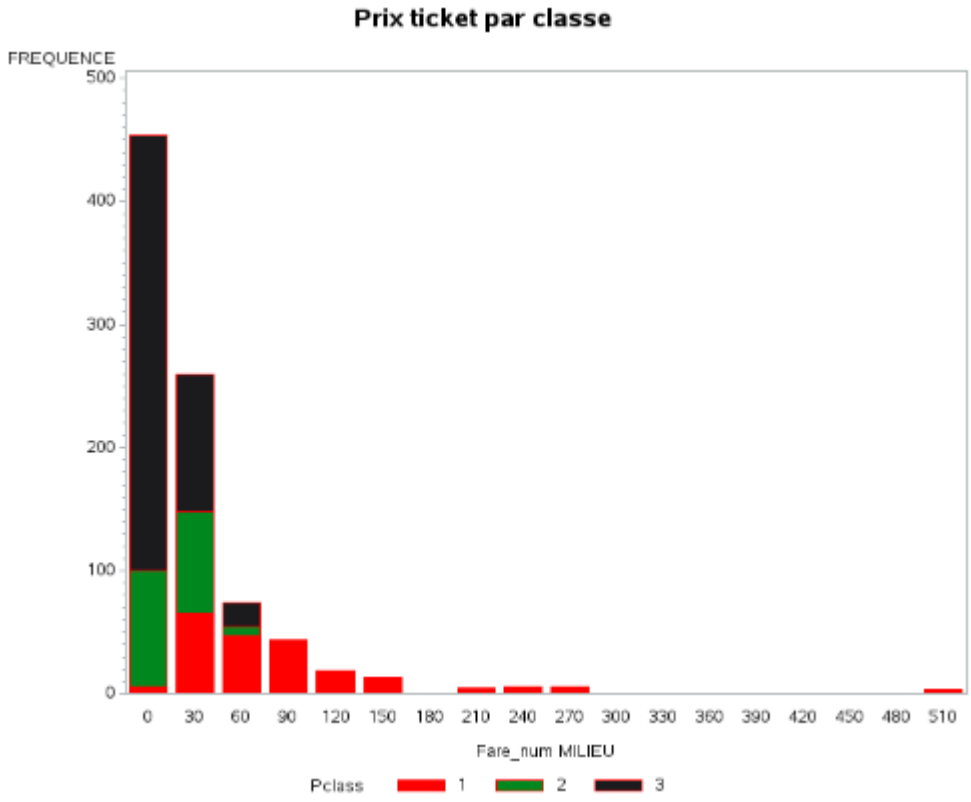
Et comme nous avons plus haut les ages les plus représentés sont entre 18 et 40 ans environ.

3. Pour la variable âge : Calculer la moyenne de la variables Fare par classe :

Remarque **:** Le ticket moyen en 3e est très faible par rapport aux autres en particulier en 1ere

Pclass	ticket_moyen
1	84.08229
2	20.66179
3	13.68589

4. histogramme de la variable Fare par classe. Interpréter le graphique



Interprétation :

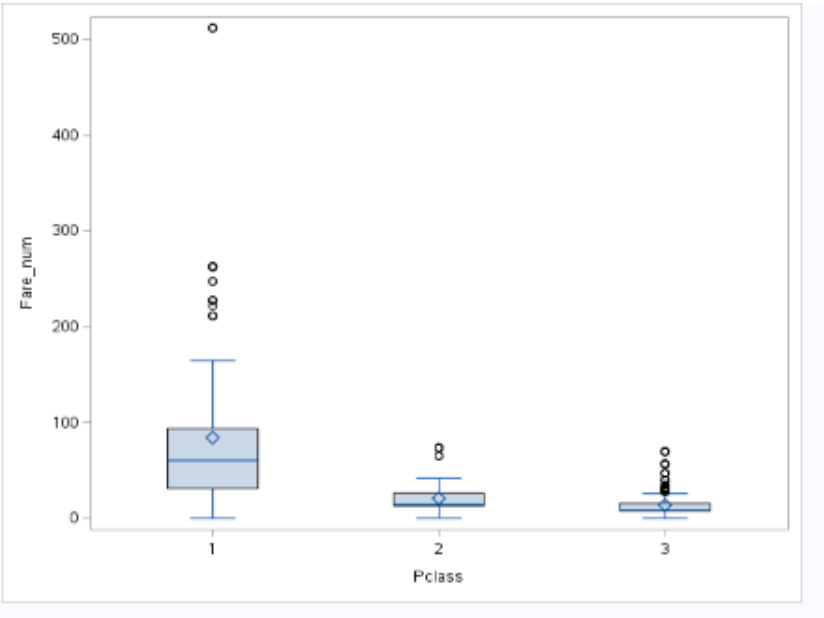
Comme on pouvait s'y attendre la proportion de passagers avec des tickets de 3e classe est la plus grande et leurs prix sont moins élevés.

Ils sont suivis des tickets de 2e et enfin ceux de 3e classé.

Cependant certaines valeurs semblent aberrantes en particulier le ticket de 510 environ.

Nous verrons avec un Boxplot si c'est le cas.

5. Faire une boîte à moustaches de la variable Fare par classe. Interpréter le graphique



Interprétation :

Nous remarquons encore des valeurs aberrantes en particulier pour la 1ere classe ou la moyenne est largement au dessus de la médiane.

Elle est presque proche du 3e quartile. Ce qui se comprend aisément parce que tiré par les grosses aberrantes.

C'est l'intérêt d'un Boxplot pour se rendre compte que certains outliers influencent énormément nos données et de facto font apparaître des incohérences.

C'est la même chose pour les deux autres boites restantes.

Conclusion : Donc le ticket moyen calculé plus haut doit être prix avec prudence !

Remarque **:** Dans une analyse de données de manière générale, on aurait sûrement supprimé les données aberrantes. Là, sans instruction dans l'énoncé, nous n'avons procédé à aucune suppression.

6. Le nombre d'hommes et de femmes par classe :

La procédure FREQ

Fréquence Pourcentage Pct de ligne Pct de col.	Table de Polasc par Sex		
	Sex(Sex)		
Polasc(Polasc)	female	male	Total
1	93 10.52 43.46 29.81	121 13.69 56.54 21.15	214 24.21
2	76 8.60 41.30 24.36	108 12.22 58.70 18.88	184 20.81
3	143 16.18 29.42 45.83	343 38.80 70.58 59.97	486 54.98
Total	312 35.29	572 64.71	884 100.00

7. Moyenne par classe et par sexe :

Polasc	Moyenne_Age
1	38.37185
2	29.87763
3	25.1475

Sex	Moyenne_Age
female	27.92692
male	30.77989

8. Les pourcentages de survivants et non survivants par sexe du fichier train_url

Fréquence Pourcentage Pct de ligne Pct de col.	Table de Survived par 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

Interprétation :

C'est un tableau de contingence/probabilité/croisé survivant/sex :

Les données sont fournies en nombres et pourcentages..., nous faisons le choix d'interpréter le **tableau des pourcentages :**

Le tableau train_url a 884 passagers dont 35.29% de femmes et 64.71% d'hommes
Il y a 8.94 % des femmes qui ont survécu et 26.36% ont péri
Il y a 52.49% des hommes qui ont survécu et 12.22% ont péri
Il y a 61.43% de survivants et 38.57% de non survivants.

A titre d'exemple le tableau des fréquences peut être lu comme suit :

Nous pouvons lire que les individus ayant survécu
et étant de sexe femme représente 14.51% (f11) des passagers (884)

9. Pourcentages de survivants par point d'embarquement

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table de Survived par Embarked

Survived(Survived)	Embarked(Embarked) 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

Fréquence manquante = 2

En tout 18.93% des passagers ont embarqué en C, 8.62% en Q et 72.45% en S.

8.50% des passagers qui ont embarqué au port C ont survécu et 10.43% ont péri,
5.22% au qui ont embarqué au port Q ont survécu contre 3.40% qui ont péri et enfin
47.85% au port S ont survécu contre 24.60% qui sont décédés (plus représentés en embarquement).

10. Calcul les pourcentages de survivants par sexe et par classe

La procédure FREQ

Table 1 de Sex par Pclass

Contrôle pour Survived=0

Fréquence Pourcentage Pct de ligne Pct de col.	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

Table 2 de Sex par Pclass

Contrôle pour Survived=1

Fréquence Pourcentage Pct de ligne Pct de col.	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

11.Variable pour différencier un adulte d'un enfant:

Aperçu :

vived	Pclass	Sex	Age	Fare	Embarked	age_num	Fare_num	age_Categorie
0	3	male	22	7.25	S	22	7.25	Adulte
1	1	female	38	71.2833	C	38	71.28	Adulte
1	3	female	26	7.925	S	26	7.925	Adulte
1	1	female	35	53.1	S	35	53.1	Adulte
0	3	male	35	8.05	S	35	8.05	Adulte
0	3	male		8.4583	Q	.	8.458	Enfant
0	1	male	54	51.8625	S	54	51.86	Adulte
0	3	male	2	21.075	S	2	21.07	Enfant

12. Le nombre de passagers adultes et le nombre d'enfants passagers

age_Categorie	Nbre_passager
Adulte	597
Enfant	287

13. Le nombre d'enfants et d'adultes qui ont survécu par classe ?

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table de age_Categorie par 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

14. Le nombre d'enfants et d'adultes qui ont survécu par classe et par point d'embarcation ?

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table 1 de Survived par Embarked				
Contrôle pour age_Categorie=Adulte				
	Embarked(Embarked)			
Survived(Survived)	C	Q	S	Total
0	45 7.56 12.20 42.86	15 2.52 4.07 75.00	309 51.93 83.74 65.74	369 62.02
1	60 10.08 26.55 57.14	5 0.84 2.21 25.00	161 27.06 71.24 34.26	226 37.98
Total	105 17.65	20 3.36	470 78.99	595 100.00
Fréquence manquante = 2				

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table 2 de Survived par Embarked				
Contrôle pour age_Categorie=Enfant				
	Embarked(Embarked)			
Survived(Survived)	C	Q	S	Total
0	30 10.45 17.24 48.39	31 10.80 17.82 55.36	113 39.37 64.94 66.86	174 60.63
1	32 11.15 28.32 51.61	25 8.71 22.12 44.64	56 19.51 49.56 33.14	113 39.37
Total	62 21.60	56 19.51	169 58.89	287 100.00