```
** Abdou NIANG **
** Master MIASHS **

/* CREATION DE LIBRAIRIE */;

libname TEST "/home/u63207901/EXAMEN_BASSENE";


** ETAPE 1 : IMPORTATION **

* 1. Ecrivez un script qui lit les trois fichiers de train_url1, train_url2 et train_url3*;


proc import datafile= "/home/u63207901/EXAMEN_BASSENE/train_url1.xlsx"
out = TEST.train_url1
dbms = xlsx replace;
getnames=yes;
run;

proc import datafile= "/home/u63207901/EXAMEN_BASSENE/train_url2.xlsx"
out = TEST.train_url2
dbms = xlsx replace;
getnames=yes;
run;


proc import datafile= "/home/u63207901/EXAMEN_BASSENE/train_url3.xlsx"
out = TEST.train_url3
dbms = xlsx replace;
getnames=yes;
run;

*2 Concaténation *

*-Script qui vous permet d'effectuer une concaténation entre train_url1 et train_url2*
                    *pour créer un fichier nommé train_url1_2*

*Tri des tableaux*;

proc sort
data=TEST.train_url1;
    by passengerId;
run;

proc sort
data=TEST.train_url2;
    by passengerId;
run;

*Concaténation en train_url 1_2*;

data TEST.train_url1_2;
merge TEST.Train_url1 TEST.Train_url2;
by passengerId;
run;

*Impression*;

proc print data=TEST.train_url1_2;

*Concaténation en train_url*;

data TEST.train_url;
set TEST.train_url1_2 TEST.Train_url3;
run;

proc print data=TEST.train_url;

*Impression*;

proc print data=TEST.train_url;


** ETAPE 2 : Analyse descriptive du fichier Train_url **;

** 1. Nombre de survivants et non survivants **;

Proc Freq data=TEST.train_url;
Tables survived ;
run;
```

```
** 2. Nombre de passagers hommes et femmes **;

Proc Freq data=TEST.train_url;
Tables sex;
run;

** On peut visualiser les proportions hommes/femmes grace à un diagramme secteur **;

PROC GCHART DATA = TEST.train_url;
PIE sex;
title "Distribution de la variable sexe";
goptions colors=(red green);
RUN;
QUIT;


** 3. Nombre de passagers par point d'embarquement **;

Proc Freq data=TEST.train_url;
Tables Embarked;
run;


** 4. Nombre de passagers hommes femmes par classe **;

PROC FREQ DATA = TEST.train_url;
TABLES Sex * Pclass;
RUN;


*** ETAPE 3 : Analyse descriptive avancée ***

**1. Pour la variable âge **

    ** 1.1 Calculer la moyenne de la variable âge par sexe **;

data   TEST.train_url;
  set  TEST.train_url;
    age_num = input(age ,5.);
run;

PROC SQL;
    Select sex, AVG(age_num ) as Moyenne_Age
        from TEST.train_url
        group by sex;
QUIT;

    ** 1.2 Faire un histogramme de la variable Age par sexe. Interpréter le graphique **;

PROC GCHART DATA=TEST.train_url;
    VBAR age_num/ subgroup= sex;
    title "histogramme age par sexe";
    RUN;
    goptions colors=(red green);
QUIT;

** Interprétation **

** 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 réd


    ** 1.3 Faire une boîte à moustaches de la variable Fare par classe. Interpréter le graphique **;

PROC SGPLOT  DATA = TEST.train_url;
   VBOX age_num / category = sex;
RUN;

** /bold Interprétation **

    ** D'abord on peut remarquer que le boxplot de l'age des femmes présente des outliers ( des valeurs abérrantes)**
    ** On pourrait pour avoir des données plus cohérents supprimer toutes les valeurs au delà de 65 ans qui semble etre l
    **Ensuite le point bleu dans chaque boite indique la moyenne qui semble etre ce qu'on avait trouvé à savoir 27 ans po
    ** Et comme nous avons plus haut les ages les plus représentés sont entre 18 et 40 ans environ **



** 2. Pour la variable âge **;

    ** 2.1 Calculer la moyenne de la variables Fare par classe **;
```

```
data   TEST.train_url;
  set  TEST.train_url;
    Fare_num = input(Fare ,5.);
run;
```

```
PROC SQL;
    Select Pclass, AVG(Fare_num) as ticket_moyen
        from TEST.train_url
        group by Pclass;
QUIT;
```

```
    ** Remarque : Le ticket moyen en 3e est évident tres faible par rapport aux autres **


    ** 2.2 histogramme de la variable Fare par classe. Interpréter le graphique **;
```

```
proc gchart data=TEST.train_url;
vbar Fare_num/ subgroup=Pclass;
title "Prix ticket par classe";
goptions colors=(red green blb)
run;
```

```
    ** Interprétation :**

    ** Comme on pouvait s'y attendre la proportion de passagers avec des tickets de 3e classe est la plus grande et leurs
    ** 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 **



    ** 2.3 Faire une boîte à moustaches de la variable Fare par classe. Interpréter le graphique **;
```

```
PROC SGPLOT  DATA = TEST.train_url;
   VBOX Fare_num / category = Pclass;
RUN;
```

```
    ** Nous remarquons encore des valeurs aberrantes en particulier pour la 1ere classe ou la moyenne est largement au de
    ** Elle est presque proche du 3e quartile. Ce qui se comprend aisément parce que tiré par les grosses aberrantes **
    ** C'est l'interet d'un boxplot pour se rendre compte que certains outliers influencent énormément nos données et de
    ** C'est la meme chose pour les deux autres boites restantes **
    Conclusion :
    ** Donc le ticket moyen calculé plus haut doit être prix avec prudence ! **



** 3. Determiner **

    ** 3.1 Le nombre d'hommes et de femmes par classe **;
```

```
PROC Freq  data =TEST.train_url;
    Tables Pclass *sex;
run;            /* C'est un tableau de contingence */



    ** 3.2 La moyenne de la variable Age par classe et par sexe **;
```

```
Proc SQL;
Select Pclass, avg(age_num) as Moyenne_Age
    from TEST.train_url
    group by Pclass;
quit;
```

```
Proc Sql;
Select Sex, avg(age_num) as  Moyenne_Age
    from TEST.train_url
        group by Sex;
quit;



** 4. Calculer **

** 4.1 Les pourcentages de survivants et non survivants par sexe du fichier train_url **;
```

```
Proc freq Data = TEST.train_url ;
    Tables Survived * Sex;
Run;
```

```
** Interpréter les résultats **

** C'est un tableau de contigence survivant/sex :

** Les données sont fournies en nombres et pourentages..., nous faisons le choix d'interpréter le tableau des pourcentage
** 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 etre lu comme suit :
Nous pouvons lire que les individus ayant survécu
et étant de sexe femme représente 14.51% (f11) des passagers (884);


** 4.2 Calcul des pourcentages de survivants par point d'embarquement **;

Proc freq data =TEST.train_url;
    Tables Survived * Embarked ;
run;


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

    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 ).



** 4.3 Calcul les pourcentages de survivants par sexe et par classe **;


Proc freq data =TEST.train_url;
    Tables Survived * Sex * Pclass;
run;



** 5. Variable pour différencier un adulte d'un enfant: **;

data TEST.train_url4;
set TEST.train_url;
if age<18 then age_Categorie="Enfant";
else age_Categorie="Adulte";
run;

** 6. Déterminer **
** 6.1 Le nombre de passagers adultes et le nombre d'enfants passagers **;

Proc sql;
    Select age_Categorie, count(age_Categorie) as Nbre_passager
        from TEST.train_url4
        group by age_Categorie;
quit;


** 6.2 Le nombre d’enfants et d’adultes qui ont survécu par classe ?**;

Proc sql;
Select  age_Categorie, count(Survived) as Nbre_passager_1
        from TEST.train_url4
        where Survived=1
        group by age_Categorie;
quit;


Proc sql;
    Select  age_Categorie, count(Survived) as Nbre_passager_0
        from TEST.train_url4
        where Survived=0
        group by age_Categorie;
quit;


Proc freq data =TEST.train_url4;
    Tables age_Categorie * Survived  ;
run;


** 6.3 Le nombre d’enfants et d’adultes qui ont survécu par classe et par point d’embarcation ? **;
```

```
Proc freq data = TEST.train_url4;
    Tables age_Categorie * Survived* Embarked ;
run;
```