MK

DATA PROJECTS

sas

Titanic Data Analysis

SAS - Statistical Analysis Software

SOMMAIRE

Importation des données

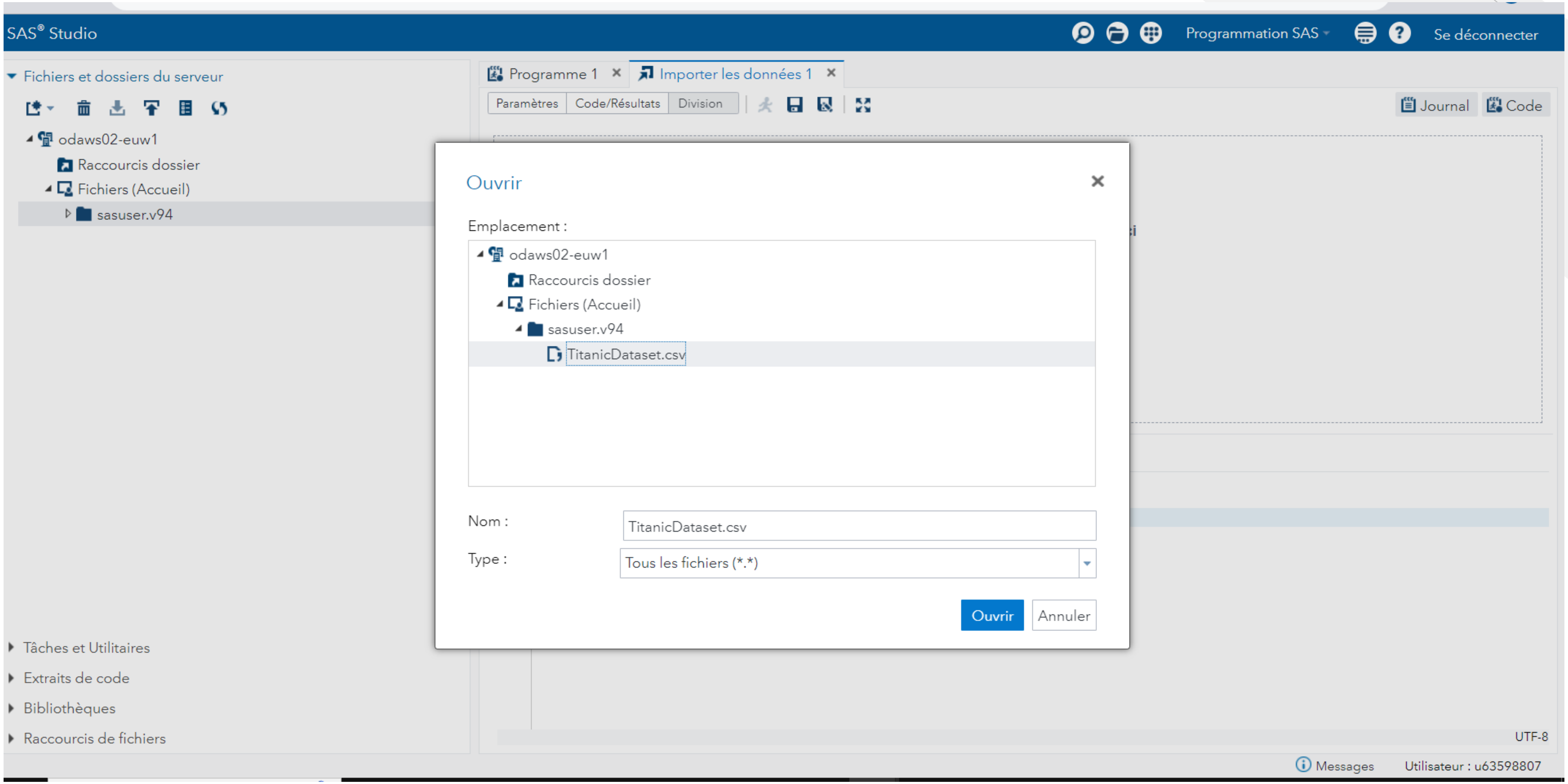
Étude des données

Data Processing

Régression Logistique

Importation des données

Importation des données (1/3)



Ouvrir

Emplacement :

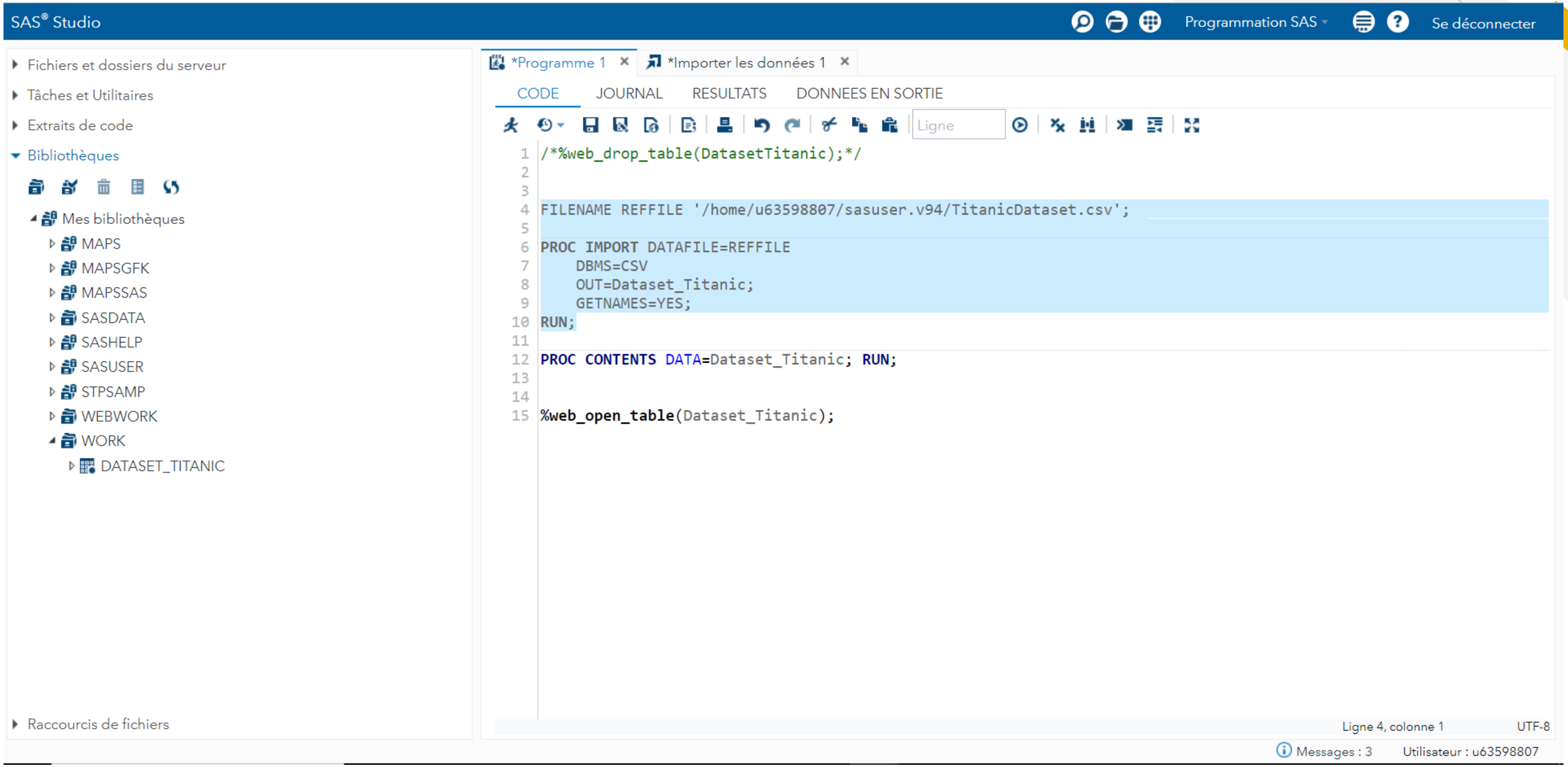
- odaws02-euw1
  - Raccourcis dossier
  - Fichiers (Accueil)
    - sasuser.v94
      - TitanicDataset.csv

Nom : TitanicDataset.csv

Type : Tous les fichiers (*.*)

Ouvrir | Annuler

Importation des données (2/3)



```
/*%web_drop_table(DatasetTitanic);*/


FILENAME REFFILE '/home/u63598807/sasuser.v94/TitanicDataset.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=Dataset_Titanic;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=Dataset_Titanic; RUN;


%web_open_table(Dataset_Titanic);
```

Importation des données (3/3)

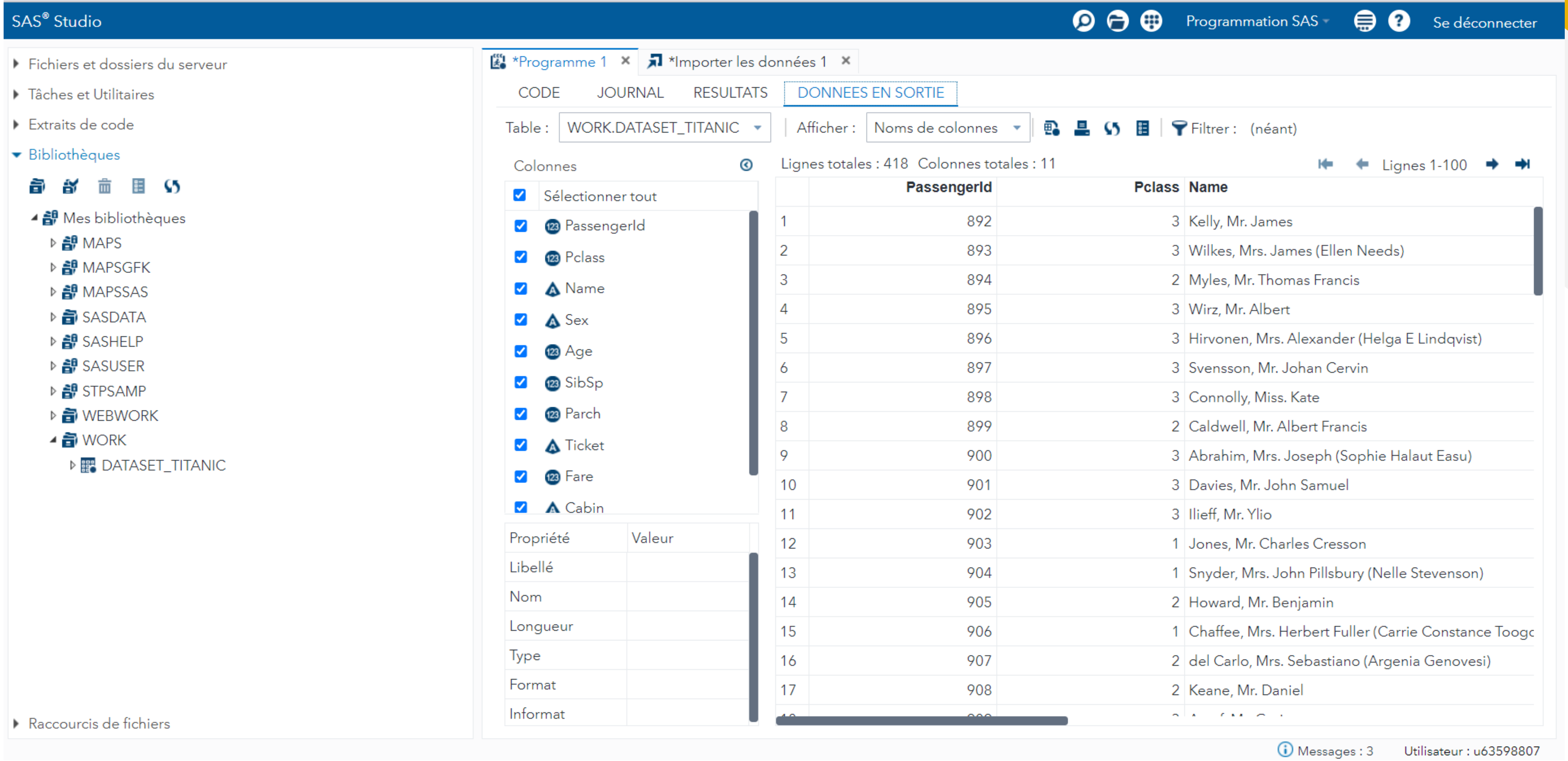


Table : WORK.DATASET_TITANIC

Afficher : Noms de colonnes

Filtrer : (néant)

Colonnes

- ☑ Sélectionner tout
- ☑ PassengerId
- ☑ Pclass
- ☑ Name
- ☑ Sex
- ☑ Age
- ☑ SibSp
- ☑ Parch
- ☑ Ticket
- ☑ Fare
- ☑ Cabin

Propriété	Valeur
Libellé	
Nom	
Longueur	
Type	
Format	
Informat	

Lignes totales : 418 Colonnes totales : 11

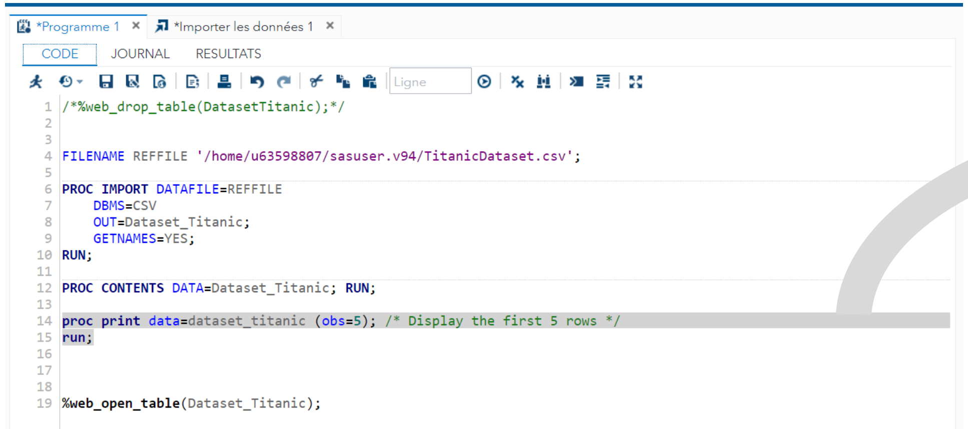
	PassengerId	Pclass	Name
1	892	3	Kelly, Mr. James
2	893	3	Wilkes, Mrs. James (Ellen Needs)
3	894	2	Myles, Mr. Thomas Francis
4	895	3	Wirz, Mr. Albert
5	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)
6	897	3	Svensson, Mr. Johan Cervin
7	898	3	Connolly, Miss. Kate
8	899	2	Caldwell, Mr. Albert Francis
9	900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)
10	901	3	Davies, Mr. John Samuel
11	902	3	Ilieff, Mr. Ylio
12	903	1	Jones, Mr. Charles Cresson
13	904	1	Snyder, Mrs. John Pillsbury (Nelle Stevenson)
14	905	2	Howard, Mr. Benjamin
15	906	1	Chaffee, Mrs. Herbert Fuller (Carrie Constance Toogc
16	907	2	del Carlo, Mrs. Sebastiano (Argenia Genovesi)
17	908	2	Keane, Mr. Daniel

Utilisateur : u63598807

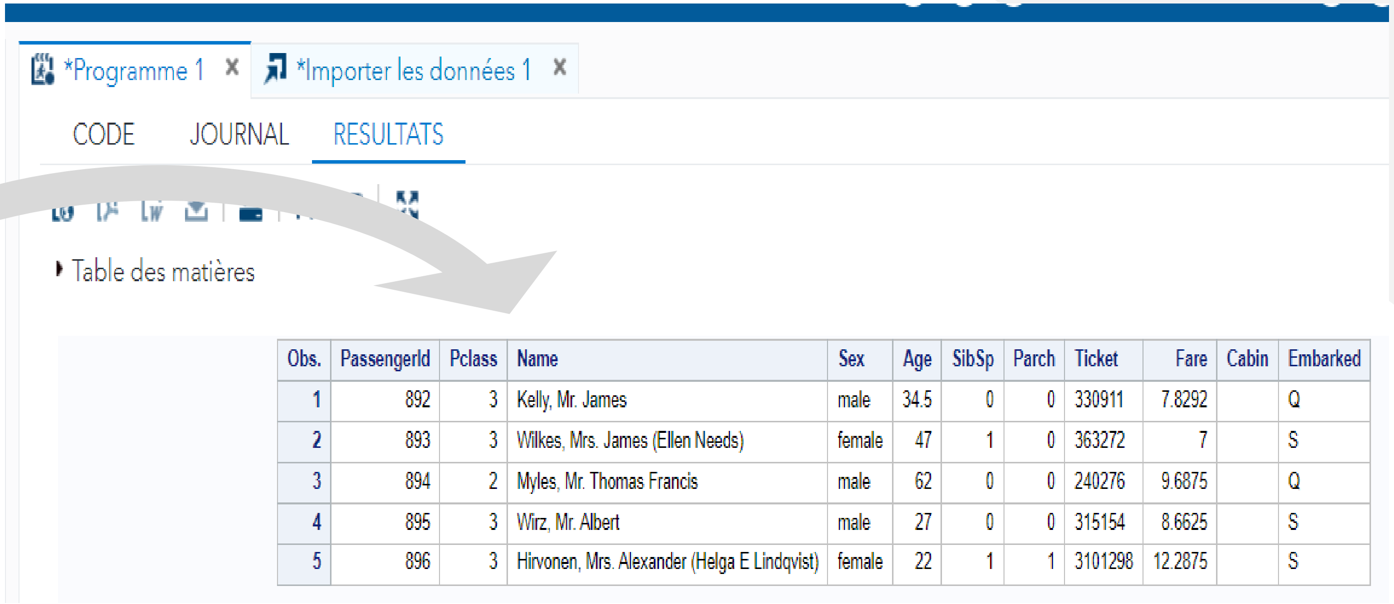
ales of main products in 2013
N FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
26% BN&T
74% IT&C
Share of market activity
Changes in the activity of the active and passive market is uncertain. Established positive trends in various market segments.
Projected sales of main pro
25%
75%
300 250 200 150 100 50 0
45 26 78 151 30 220 257
Passive market share

Étude des données

Étude des données (1/3)

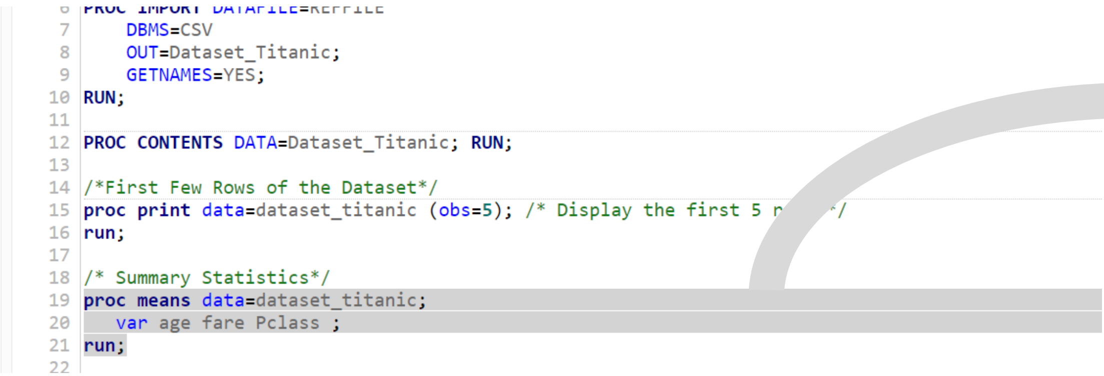


```
/*%web_drop_table(DatasetTitanic);*/


FILENAME REFFILE '/home/u63598807/sasuser.v94/TitanicDataset.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=Dataset_Titanic;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=Dataset_Titanic; RUN;

proc print data=dataset_titanic (obs=5); /* Display the first 5 rows */
run;



%web_open_table(Dataset_Titanic);
```

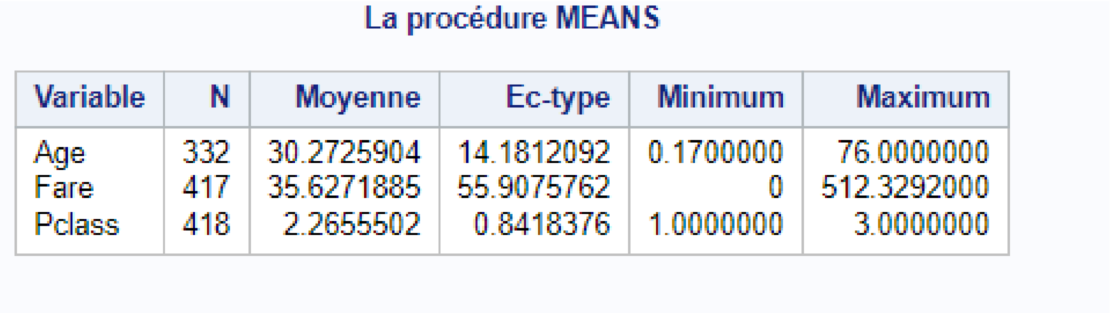


Obs.	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292		Q
2	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47	1	0	363272	7		S
3	894	2	Myles, Mr. Thomas Francis	male	62	0	0	240276	9.6875		Q
4	895	3	Wirz, Mr. Albert	male	27	0	0	315154	8.6625		S
5	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22	1	1	3101298	12.2875		S

Étude des données – Statistiques (2/3)



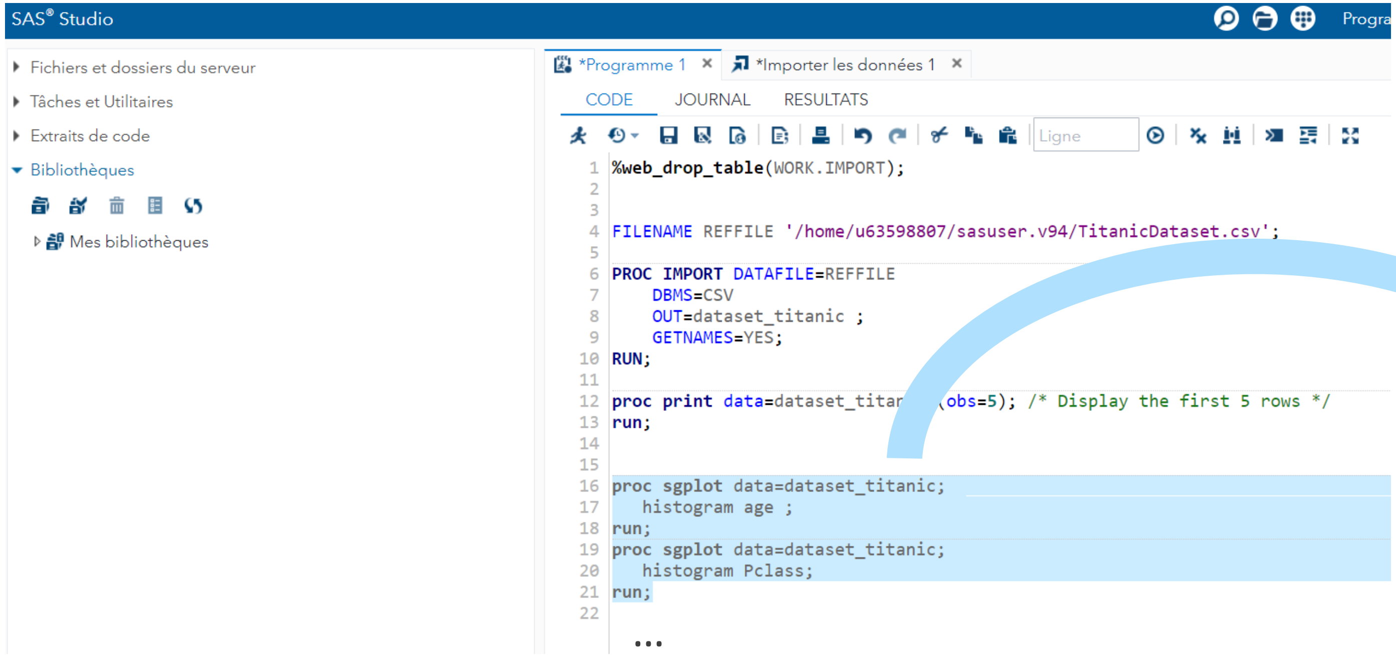
```
PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=Dataset_Titanic;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=Dataset_Titanic; RUN;

/*First Few Rows of the Dataset*/
proc print data=dataset_titanic (obs=5); /* Display the first 5 r    */
run;

/* Summary Statistics*/
proc means data=dataset_titanic;
   var age fare Pclass ;
run;
```



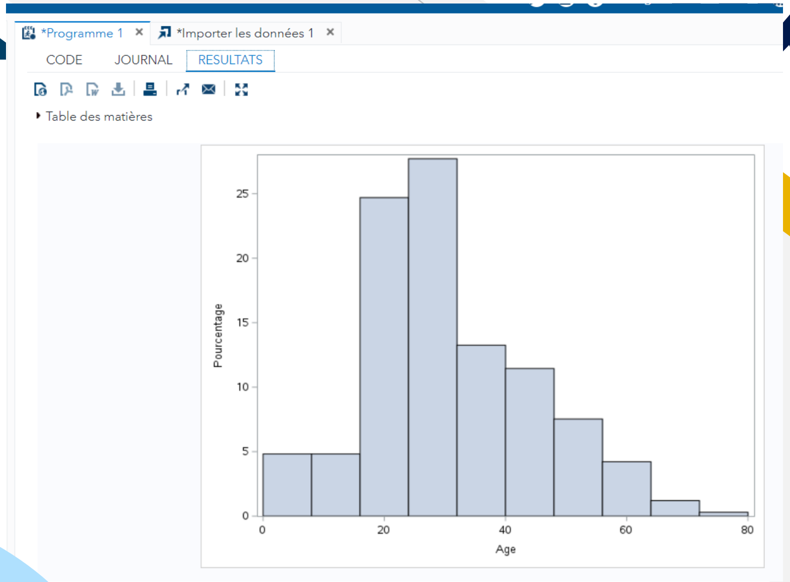
La procédure MEANS

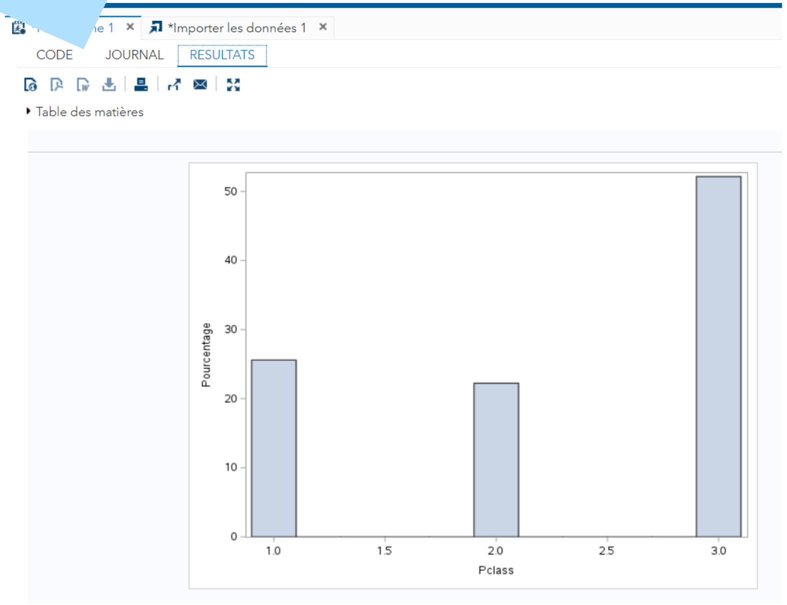
Variable	N	Moyenne	Ec-type	Minimum	Maximum
Age	332	30.2725904	14.1812092	0.1700000	76.0000000
Fare	417	35.6271885	55.9075762	0	512.3292000
Pclass	418	2.2655502	0.8418376	1.0000000	3.0000000

Étude des données – Distribution (2/3)

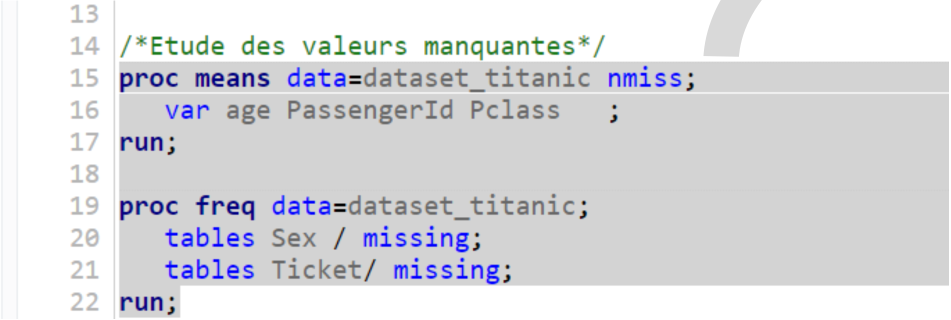


```
%web_drop_table(WORK.IMPORT);


FILENAME REFFILE '/home/u63598807/sasuser.v94/TitanicDataset.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=dataset_titanic ;
    GETNAMES=YES;
RUN;

proc print data=dataset_titanic (obs=5); /* Display the first 5 rows */
run;


proc sgplot data=dataset_titanic;
   histogram age ;
run;
proc sgplot data=dataset_titanic;
   histogram Pclass;
run;

  ...
```

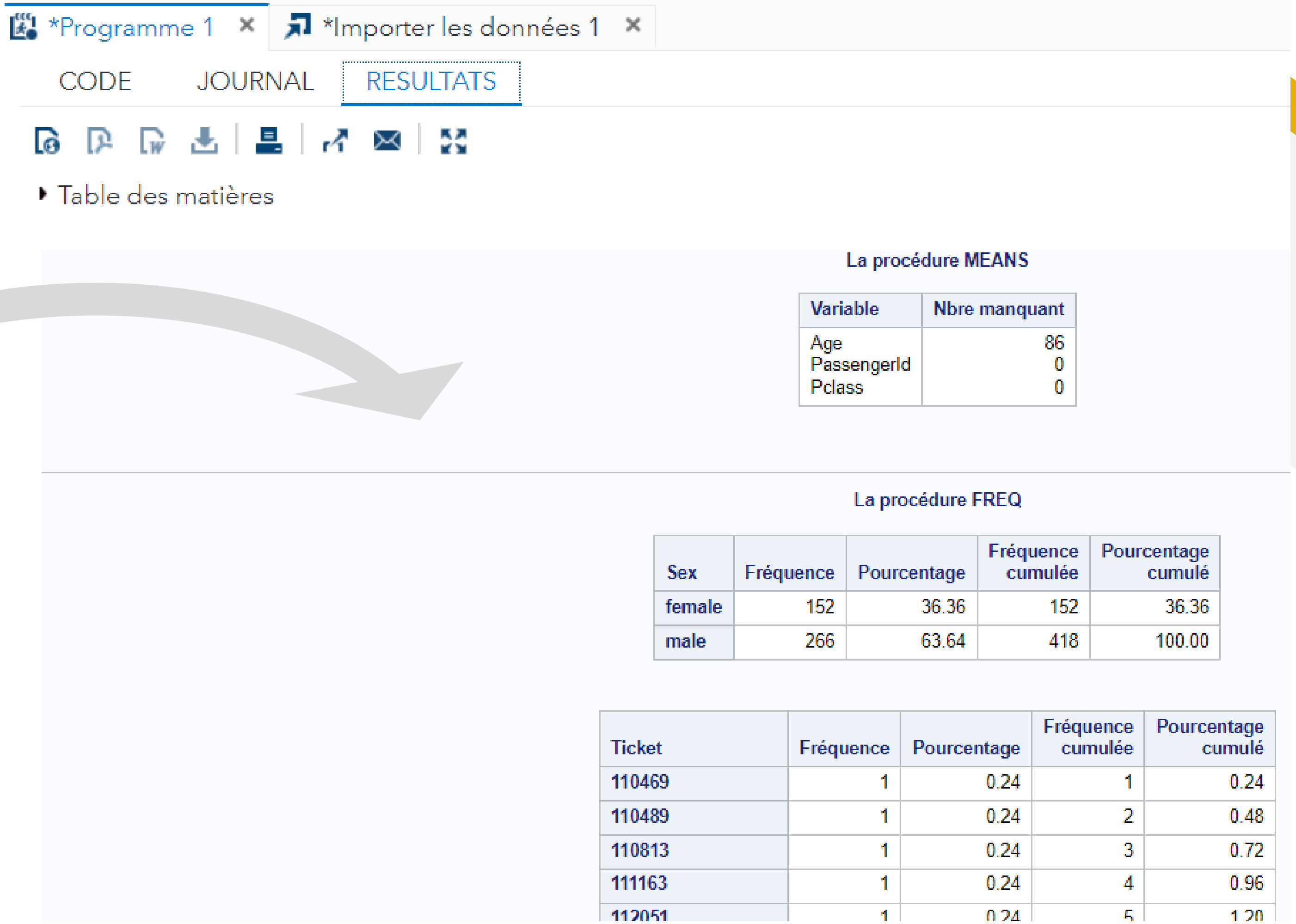




Étude des données – Valeurs manquantes (2/3)



```
/*Etude des valeurs manquantes*/
proc means data=dataset_titanic nmiss;
   var age PassengerId Pclass   ;
run;

proc freq data=dataset_titanic;
   tables Sex / missing;
   tables Ticket/ missing;
run;
```



La procédure MEANS

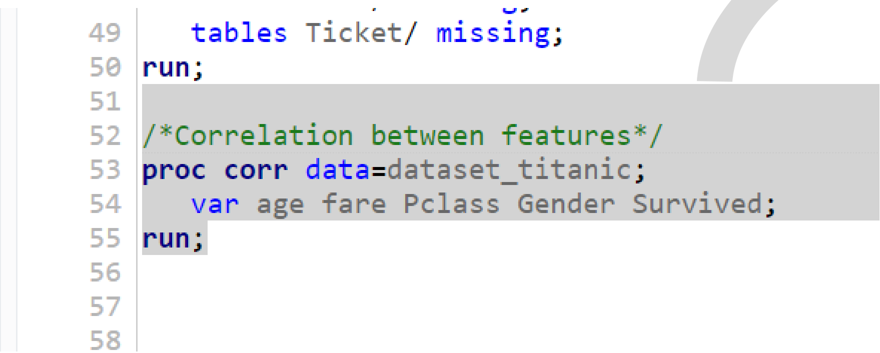
Variable	Nbre manquant
Age PassengerId Pclass	86 0 0

La procédure FREQ

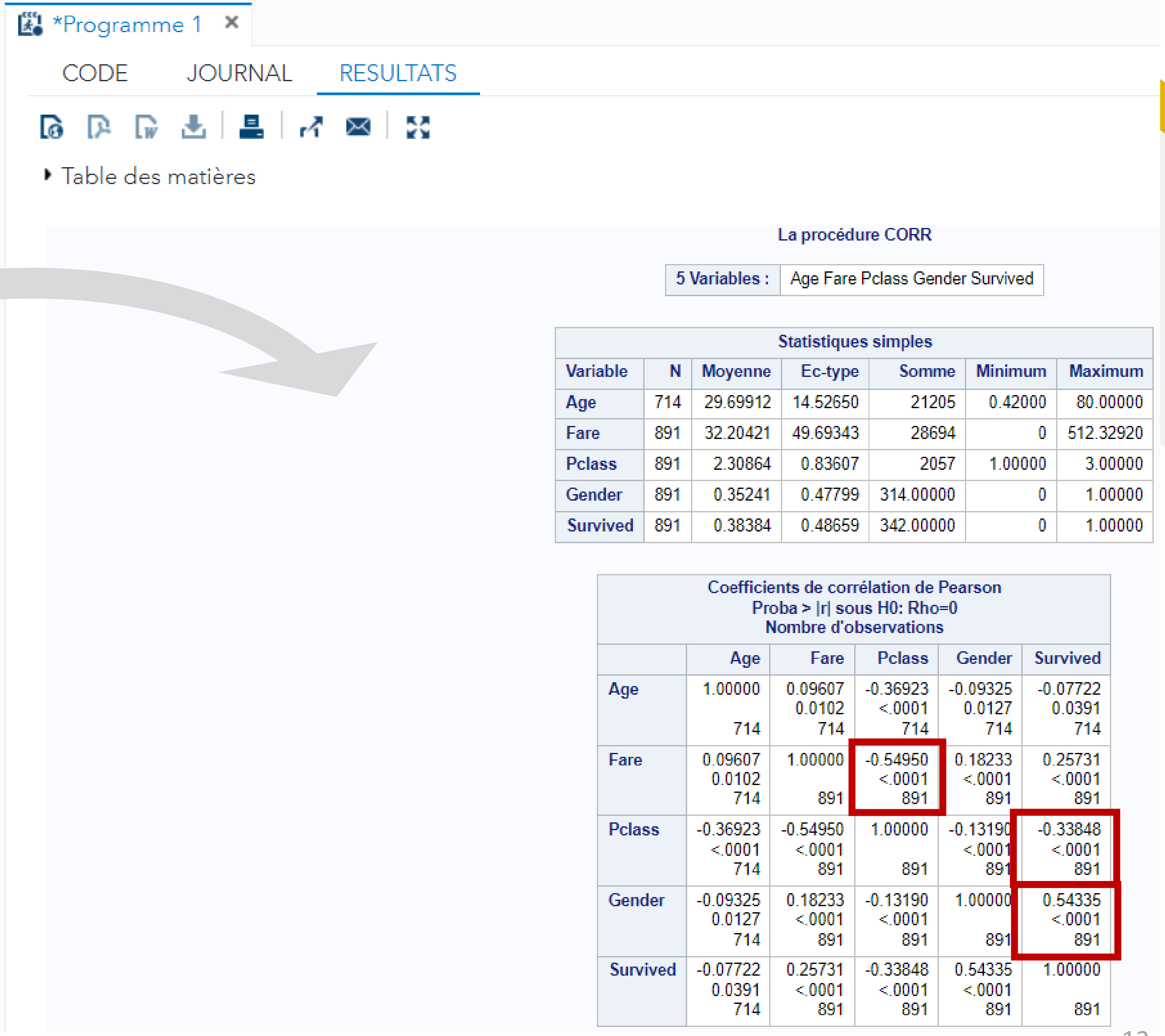
Sex	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
female	152	36.36	152	36.36
male	266	63.64	418	100.00

Ticket	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
110469	1	0.24	1	0.24
110489	1	0.24	2	0.48
110813	1	0.24	3	0.72
111163	1	0.24	4	0.96
112051	1	0.24	5	1.20

Étude des données – Corrélation des variables (3/3)



```
   tables Ticket/ missing;
run;

/*Correlation between features*/
proc corr data=dataset_titanic;
   var age fare Pclass Gender Survived;
run;
```



La procédure CORR

5 Variables :	Age Fare Pclass Gender Survived

Statistiques simples

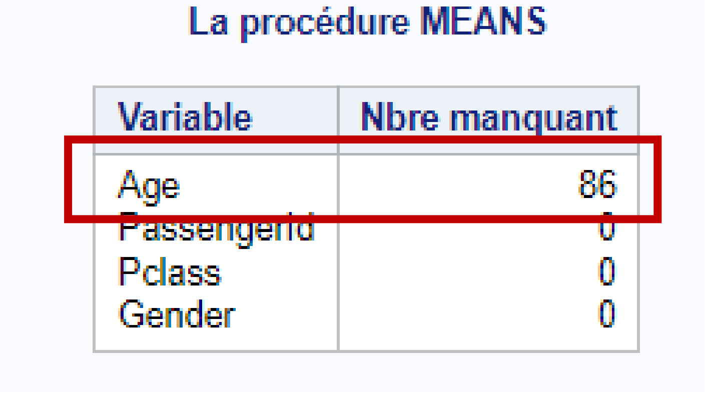
Variable	N	Moyenne	Ec-type	Somme	Minimum	Maximum
Age	714	29.69912	14.52650	21205	0.42000	80.00000
Fare	891	32.20421	49.69343	28694	0	512.32920
Pclass	891	2.30864	0.83607	2057	1.00000	3.00000
Gender	891	0.35241	0.47799	314.00000	0	1.00000
Survived	891	0.38384	0.48659	342.00000	0	1.00000

Coefficients de corrélation de Pearson
Proba > |r| sous H0: Rho=0
Nombre d'observations

	Age	Fare	Pclass	Gender	Survived
Age	1.00000 714	0.09607 0.0102 714	-0.36923 <.0001 714	-0.09325 0.0127 714	-0.07722 0.0391 714
Fare	0.09607 0.0102 714	1.00000 891	-0.54950 <.0001 891	0.18233 <.0001 891	0.25731 <.0001 891
Pclass	-0.36923 <.0001 714	-0.54950 <.0001 891	1.00000 891	-0.13190 <.0001 891	-0.33848 <.0001 891
Gender	-0.09325 0.0127 714	0.18233 <.0001 891	-0.13190 <.0001 891	1.00000 891	0.54335 <.0001 891
Survived	-0.07722 0.0391 714	0.25731 <.0001 891	-0.33848 <.0001 891	0.54335 <.0001 891	1.00000 891

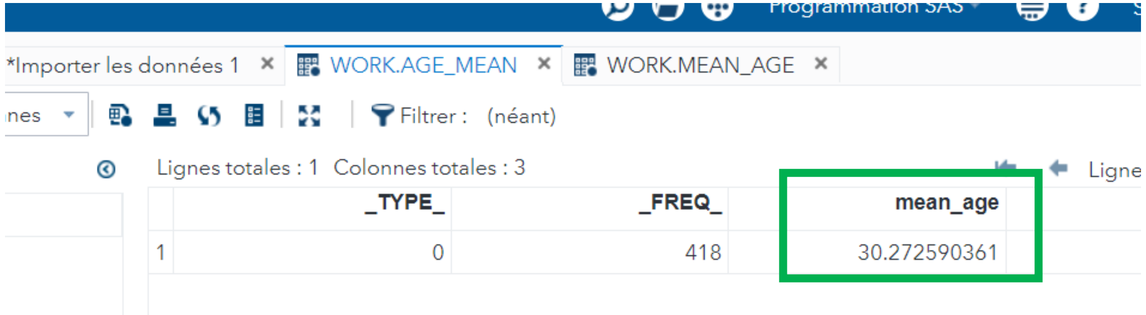
Data Processing

Data Processing – Gestion des Valeurs manquantes (méthode de la moyenne)



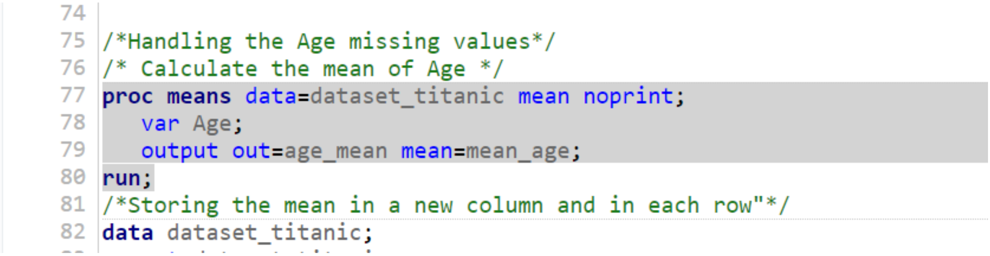
La procédure MEANS

Variable	Nbre manquant
Age	86
Passengerid	0
Pclass	0
Gender	0

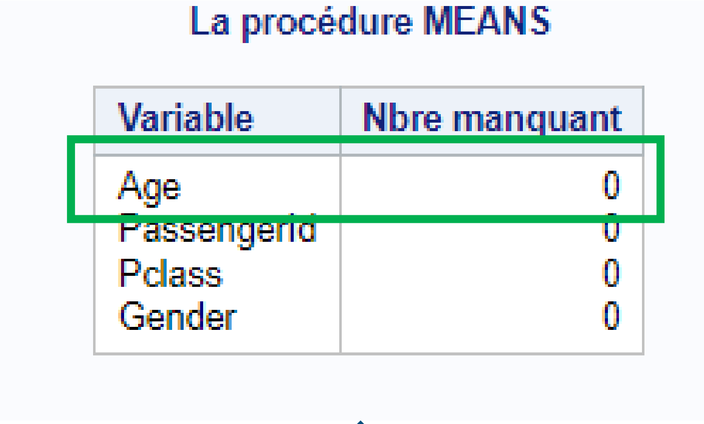


Lignes totales : 1 Colonnes totales : 3

	TYPE	_FREQ_	mean_age
1	0	418	30.272590361



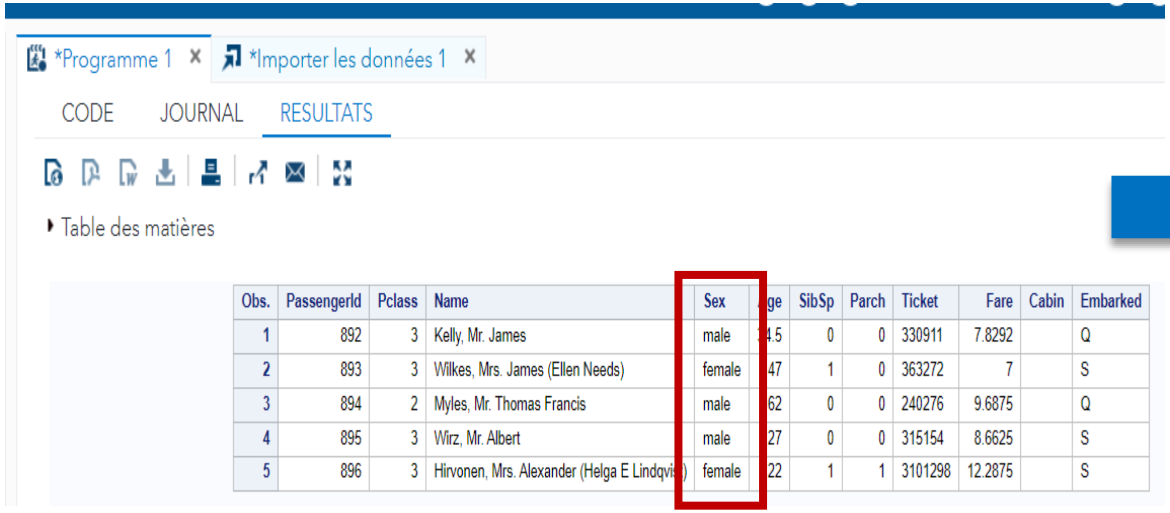
```
/*Handling the Age missing values*/
/* Calculate the mean of Age */
proc means data=dataset_titanic mean noprint;
   var Age;
   output out=age_mean mean=mean_age;
run;
/*Storing the mean in a new column and in each row"*/
data dataset_titanic;
```



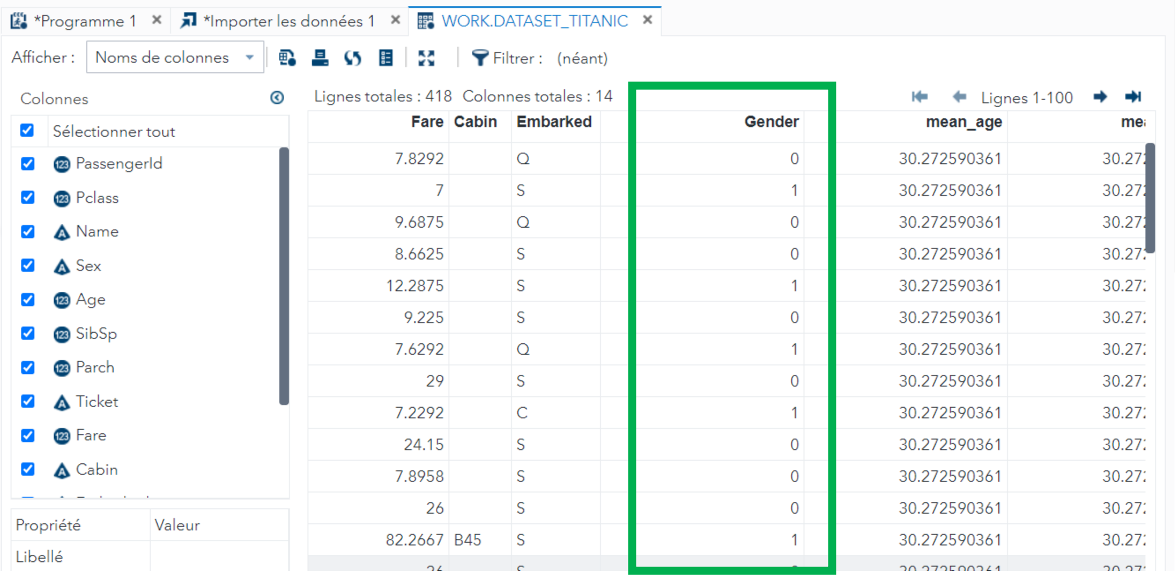
La procédure MEANS

Variable	Nbre manquant
Age	0
Passengerid	0
Pclass	0
Gender	0

Data Processing – Gestion des Valeurs (Encodage)



Obs.	PassengerId	Pclass	Name	Sex	ge	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	892	3	Kelly, Mr. James	male	4.5	0	0	330911	7.8292		Q
2	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47	1	0	363272	7		S
3	894	2	Myles, Mr. Thomas Francis	male	62	0	0	240276	9.6875		Q
4	895	3	Wirz, Mr. Albert	male	27	0	0	315154	8.6625		S
5	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvi	female	22	1	1	3101298	12.2875		S

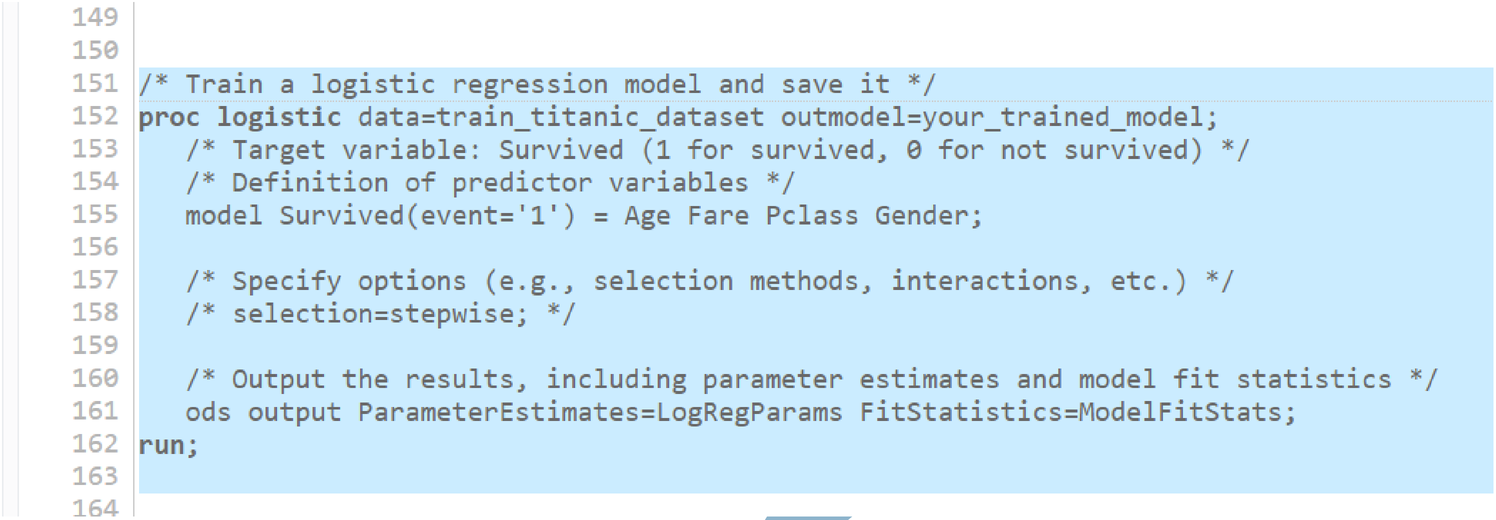


Lignes totales : 418 Colonnes totales : 14

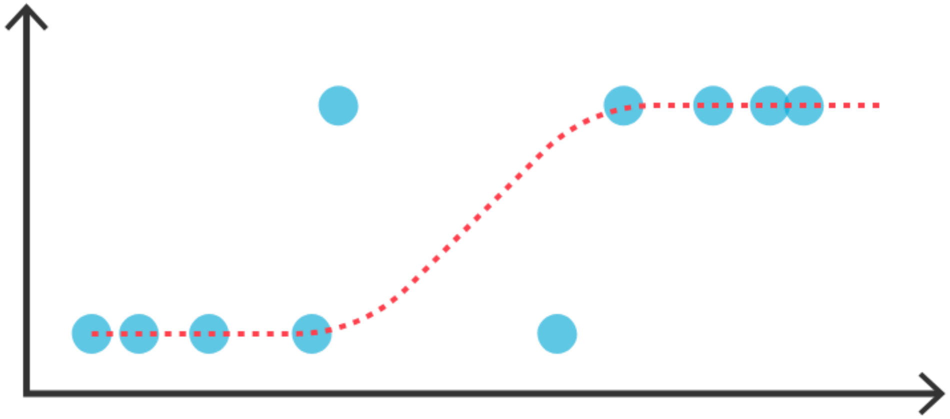
Fare	Cabin	Embarked	Gender	mean_age	me
7.8292		Q	0	30.272590361	30.27
7		S	1	30.272590361	30.27
9.6875		Q	0	30.272590361	30.27
8.6625		S	0	30.272590361	30.27
12.2875		S	1	30.272590361	30.27
9.225		S	0	30.272590361	30.27
7.6292		Q	1	30.272590361	30.27
29		S	0	30.272590361	30.27
7.2292		C	1	30.272590361	30.27
24.15		S	0	30.272590361	30.27
7.8958		S	0	30.272590361	30.27
26		S	0	30.272590361	30.27
82.2667	B45	S	1	30.272590361	30.27

Data Science - Exemple

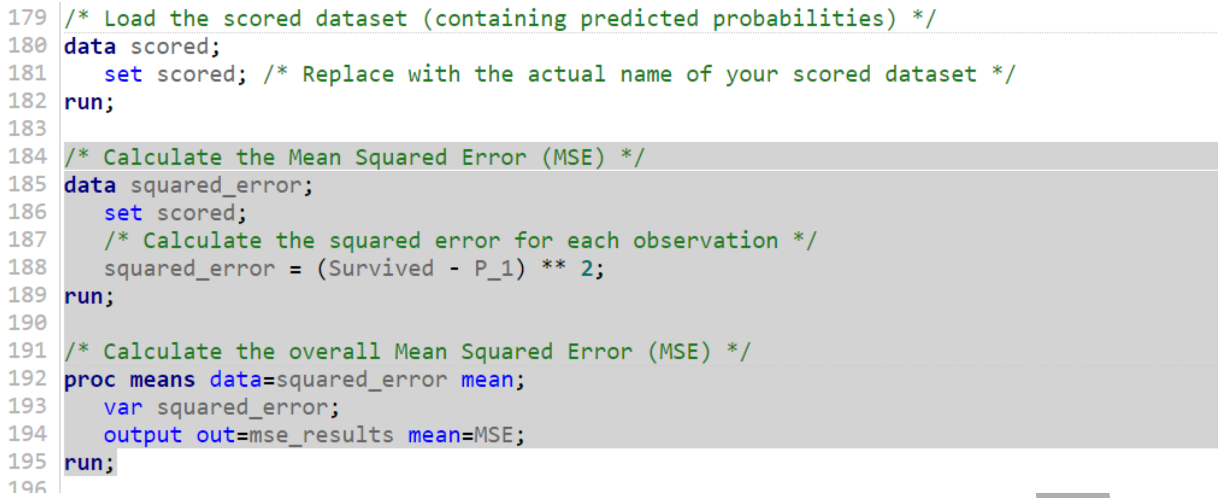
RÉGRESSION LOGISTIQUE – Implémentation du modèle



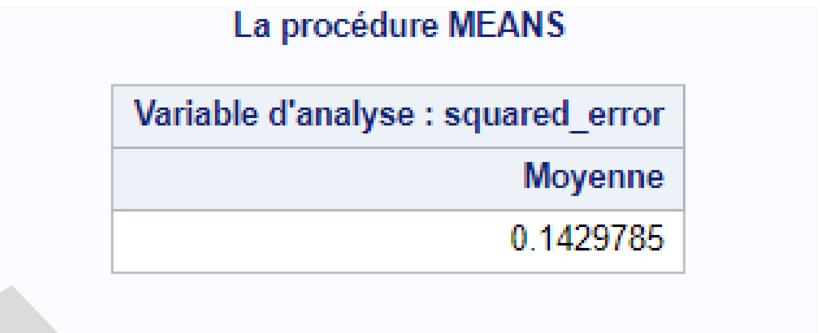
```


/* Train a logistic regression model and save it */
proc logistic data=train_titanic_dataset outmodel=your_trained_model;
   /* Target variable: Survived (1 for survived, 0 for not survived) */
   /* Definition of predictor variables */
   model Survived(event='1') = Age Fare Pclass Gender;

   /* Specify options (e.g., selection methods, interactions, etc.) */
   /* selection=stepwise; */

   /* Output the results, including parameter estimates and model fit statistics */
   ods output ParameterEstimates=LogRegParams FitStatistics=ModelFitStats;
run;


```



RÉGRESSION LOGISTIQUE – Evaluation du modèle



```
/* Load the scored dataset (containing predicted probabilities) */
data scored;
   set scored; /* Replace with the actual name of your scored dataset */
run;

/* Calculate the Mean Squared Error (MSE) */
data squared_error;
   set scored;
   /* Calculate the squared error for each observation */
   squared_error = (Survived - P_1) ** 2;
run;

/* Calculate the overall Mean Squared Error (MSE) */
proc means data=squared_error mean;
   var squared_error;
   output out=mse_results mean=MSE;
run;
```



La procédure MEANS

Variable d'analyse : squared_error
Moyenne
0.1429785