Classe : Magistère 2
Cours : SAS avancé
Année académique : 2020-2021

Projet SAS 2020-2021 / Magistère 2

Consignes :

- Ce projet est à rendre au plus tard le **31 Mars 2021 à 23h59**, suivant les groupes formés en classe.
- Le rendu doit être un dossier zippé portant votre (vos) noms. Il comportera deux sous dossiers : un premier sous dossier comprenant le programme des différentes parties ci-dessous, et un deuxième sous dossier comportant toutes les tables SAS en entrée et en sortie de votre projet.
- **Le programme a restitué doit être claire, limpide et bien commenté**.

NB : **Tout retard ou manquement à ces consignes sera sanctionné pas 5 points en moins sur la note du projet**.

Attention : Certaines des captures en exemple ci-dessous ne sont qu'une partie des résultats à obtenir et non leur intégralité.

PARTIE 1 : SAS SQL

1- Dans une librairie nommée « projet », créez les tables suivantes en vous servant des fichiers du projet. Utilisez « l'étape DATA » combinée avec les instructions INFILE et INPUT.

NB :

- Utilisez le format de date YYYY-MM-DD
- Utilisez l'année « 1900 » comme référence des années pour les dates. L'option « YEARCUTOFF » permet de l'indiquer à SAS.

« **ACCOUNT** »

account_id	district_id	frequency	date
576	55	POPLATEK MESICNE	1993-01-01
3818	74	POPLATEK MESICNE	1993-01-01
704	55	POPLATEK MESICNE	1993-01-01

« **CARD** »

card_id	disp_id	type	issued	issued_corr
1005	9285	classic	931107 00:00:00	1993-11-07
104	588	classic	940119 00:00:00	1994-01-19
747	4915	classic	940205 00:00:00	1994-02-05

« **CLIENT** »

client_id	birth_number	district_id	sex	birth_corr
1	706213	18	F	1970-12-13
2	450204	1	M	1945-02-04
3	406009	1	F	1940-10-09

« DISP »

disp_id	client_id	account_id	type
1	1	1	OWNER
2	2	2	OWNER
3	3	2	DISPONENT

« DISTRICT »

district_id	district_name	region	A4	A5
1	Hl.m. Praha	Prague	1204953	0
2	Benesov	central Bohemia	88884	80

« LOAN »

loan_id	account_id	date	amount	duration	payments	status
5314	1787	1993-07-05	96396	12	8033	B
5316	1801	1993-07-11	165960	36	4610	A

« ORDER »

order_id	account_id	bank_to	account_to	amount	k_symbol
29401	1	YZ	87144583	2452	SIPO
29402	2	ST	89597016	3372.7	UVER

« TRANS »

trans_id	account_id	date	type	operation	amount	balance	k_symbol	bank	account
695247	2378	1993-01-01	PRIJEM	VKLAD	700	700			
171812	576	1993-01-01	PRIJEM	VKLAD	900	900			

2- A l'aide de Microsoft Excel ou de tout autre logiciel (indiquez le logiciel), construisez le diagramme de la base de données qui relie toutes ces tables entre elles.

3- Ecrivez le programme SAS qui permet d'obtenir le nombre de client par district et sexe. Ordonnez par l'identifiant du district.

4- Réécrivez le programme précédent en y rajoutant le nom et la région du district, exactement comme ci-dessous.

Identifiant du distict	Nom du district	Region	Sexe du client	Nombre de client
1	Hl.m. Praha	Prague	F	324
1	Hl.m. Praha	Prague	M	339

5- Ecrivez le programme SAS qui affiche les districts qui ont un nombre total de clients supérieur à 100, en précisant les caractéristiques du district, le nombre de femmes et le nombre d'hommes.

Identifiant du distict	Nom du district	Region	Nombre total de clients	Nombre de clients Hommes	Nombre de clients Femmes
1	Hl.m. Praha	Prague	663	339	324
54	Brno - mesto	south Moravia	155	80	75

6- Ecrivez le programme SAS qui affiche le nombre d'ordres pour les clients qui possèdent au moins un compte. Affichez par ordre croissant d'âge des clients au 01-01-2010.

Age des clients	Nombre de clients	Nombre d'ordres
23	1	6471
24	1	12942
25	3	45297

7- Par type de carte de crédit, affichez le nombre de compte ayant bénéficié d'un prêt, le montant et durée minimum, moyen et maximum du prêt; ainsi que le nombre de d'emprunt par statut.

type	Nombre de compte avec un emprunt	Montant minimum des emprunts	Montant moyen des emprunts	Montant maximum des emprunts	Durée minimum des emprunts	Durée moyen des emprunts	Durée maximum des emprunts	Nombre d'emprunt catégorie A	Nombre d'emprunt catégorie B	Nombre d'emprunt catégorie C	Nombre d'emprunt catégorie D
classic	133	$12,540	$148,809	$475,680	12	35	60	46	1	84	2
gold	16	$17,508	$161,881	$417,600	12	34	60	8	0	7	1
junior	21	$39,576	$221,970	$495,180	12	40	60	6	1	14	0

8- Par type de carte de crédit et par catégorie d'emprunt, affichez le nombre de compte ayant bénéficié d'un prêt, ainsi que les statistiques quantitatives sur le montant et la durée du prêt.

status	type	Nombre de compte avec un emprunt	Montant moyen des emprunts	Montant minimum des emprunts	Montant maximum des emprunts	Variance des montants	Ecart moyen des montants	Durée moyen des emprunts	Durée minimum des emprunts	Durée maximum des emprunts	Variance des durées	Ecart moyen des durées
A	classic	46	$93,673	$12,540	$323,472	4.0557E9	63684.25	22	12	48	99.96522	9.998261
A	gold	8	$80,544	$17,508	$151,560	2.0332E9	45091.01	17	12	24	38.57143	6.21059
A	junior	6	$129,726	$39,576	$274,740	9.1313E9	95557.95	24	12	60	403.2	20.07984
B	classic	1	$208,128	$208,128	$208,128	.	.	48	48	48	.	.
B	junior	1	$174,744	$174,744	$174,744	.	.	24	24	24	.	.
C	classic	84	$175,902	$18,720	$475,680	1.382E10	117544.8	43	12	60	230.4165	15.17948
C	gold	7	$235,980	$41,256	$417,600	2.178E10	147579.3	50	12	60	308.5714	17.5662
C	junior	14	$264,877	$71,460	$495,180	1.466E10	121066.1	48	24	60	110.7692	10.5247
D	classic	2	$249,360	$177,744	$320,976	1.026E10	101280.3	48	48	48	0	0
D	gold	1	$293,880	$293,880	$293,880	.	.	60	60	60	.	.

9- Créez une table nommée « *client_macro* » qui regroupe les différentes informations des tables « *client* », « *disp* » et « *card* » en y rajoutant une colonne qui calcule l'âge du client au 01 Janvier 2010. On souhaite conserver toutes les lignes de la table des clients. Attention aux doublons de nom de colonne, les supprimer ou les renommer si besoin.

<u>Partie 1</u> : SAS Macro

Le principe de cette partie est de créer un échantillon à partir d'une table complète.

Cette partie se décompose en deux sous parties :

- Sondage aléatoire simple (AS) : les individus sont tirés au hasard dans la table, mais chacune des valeurs ne sera pas forcément représentée dans l'échantillon ; (problème des valeurs rares)
- Sondage aléatoire stratifié (ASTR), on échantillonne de manière stratifiée. Une variable de stratification est utilisée pour décomposer la table en strates : une strate par valeur de la variable. Ensuite un sous échantillon est tiré sur chaque strate (AS). Enfin on concatène les sous échantillons pour obtenir l'échantillon final.

A-/ Sondage aléatoire simple (AS)

Chaque programme utilisera par exemple la table « *client_macro* » créée dans la partie SQL.

1- Programme ASV1
Créez un programme SAS sans macro langage qui : crée une variable aléatoire (utiliser la fonction *ranuni(0)* de SAS), triez par cette variable et crée un échantillon avec les 200 premières observations.

2- Programme ASV2
Reprenez le programme AVS1, toujours sans créer de macro programme, ajouter en paramètre (utilisez « %let ») le nom de la table en entrée et le nom de la table en sortie, la taille de l'échantillon en nombres d'observations.

3- Programme ASV3
Reprenez le programme ASV2 en remplaçant le paramètre nombre d'observation par le pourcentage d'observation. Ainsi la valeur 20 de ce paramètre permettra d'obtenir un échantillon avec 20% des observations de la table d'entrée.

4- Programme ASV4
Transformez le programme ASV3 en un macro-programme appelé « **%AS** », avec les trois paramètres : table en entrée, table en sortie et taux d'échantillonnage.

B-/ Sondage aléatoire stratifié (ASTR)

Choisir dans la table « *client_macro* » une variable de stratification de type caractère (type de compte ou le type de carte …)

1- Programme ASTRV01
Créez un macro programme « **%ASTR** » qui permet de collecter dans des macro variables le nombre de valeurs prises et les valeurs correspondantes à la variable de stratification choisie. Attention à ne pas prendre en compte les valeurs manquantes

2- Programme ASTRV02
   Reprenez macro programme « **%ASTR** » et y ajoutez une partie qui éclate la table en entrée en strates. Il créera donc une table par strate.

3- Programme ASTRV03
   Reprenez le programme ASTRV02 et ajoutez une partie qui crée les sous échantillons (un échantillon pour chaque strate). Utiliser la fonction *ranuni(0)* de SAS en vous inspirant du A-/

4- Programme ASTRV04
   Reprenez le programme ASTRV03 et ajoutez une partie qui recolle les sous échantillons en une seule table SAS.

5- Programme ASTRV05
   Reprenez le programme ASTRV04 et ajoutez en paramètre le type de variable de stratification, c'est-à-dire une paramètre qui précise si la variable de stratification est une caractère ou numérique. Modifiez le macro programme en conséquence.

6- Programme ASTRV06
   Reprenez le programme ASTRV05 en informant l'utilisateur dans le journal des tailles des différents échantillons et de la taille de l'échantillon final.