Classe : Magistère 2
Cours : SAS avancé
Année académique : 2022-2023

Projet SAS / Magistère 2

Consignes

- Ce projet doit être rendu au plus tard le **Vendredi 31 Mars 2023 à 23h59**, suivant des groupes de 2 personnes maximum à l'adresse omores.ronald@gmail.com .
- Le rendu se composera de **votre programme SAS** et **un fichier pdf portant les résultats des questions 3 et 7 de la partie 1**. Ces deux documents doivent être archivés dans un fichier (.zip) portant votre (vos) noms et prénoms, avant l'envoi .
- **Le programme a restitué doit être clair, 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 BASE

1- Construire une macro fonction nommée << ***file_import*** >> qui permet d'importer une table il doit contenir en paramètre :
- une macro variable qui contient le lien vers le dossier de stockage du fichier
- le nom du fichier et son extension
- le nom de la table en sortie
- le délimiter si nécessaire

2- Utilisez la macro fonction créée dans la question précédente pour importer les 06 fichiers du projet.

3- 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. Enregistrer le sous format *« PDF »* et le joindre au dossier d'envoi.

4- Dans une étape Data, créez une table nommée « customers1 » à partir de « customers ».

- Ajoutez-y une nouvelle colonne nommée *« anciennete »* qui donne l'écart en mois entre la date de résiliation de la carte "*cancellation_date*" et celle de suscription.
- une nouvelle colonne nommée "*state_groupe*" qui regroupe les modalité de la variable "*customer_state*" en 3 groupes comme suit :
  ==> Groupe1 : les modalités commençants par A, B, C, D, E, F, G
  ==> Groupe2 : les modalités commençants par M, N, O, P, Q
  ==> Groupe3 : les modalités commençants par R, S, T, U, V
- Triez la table par *« customer_state »* et *« anciennete »*.

5-

a- A l'aide d'une « étape PROC FREQ », donnez l'effectif des clients par "*state_groupe*" et "*anciennete*". Sauvegarder le résultat dans une table nommée *« customers21 »*.

Obs.	state_groupe	anciennete	n_contrat
1	Groupe1	2	10
2	Groupe1	3	27
3	Groupe1	4	31
4	Groupe1	5	41
5	Groupe1	6	78

b- A l'aide d'une « étape PROC FREQ », donnez par "*state_groupe*" et "*anciennete*" le nombre de clients ayant résilié (*cancellation=1*). Sauvegarder le résultat dans une table nommée *« customers22 »*.

Obs.	state_groupe	anciennete	n_resiliation
11	Groupe1	12	302
12	Groupe1	13	345
13	Groupe1	14	368
14	Groupe1	15	402
15	Groupe1	16	405

c- A l'aide d'une « étape PROC FREQ », donnez l'effectif total des clients par "*state_groupe*". Sauvegarder le résultat dans une table nommée *« customers23 »*

Obs.	state_groupe	n_cohorte
1	Groupe1	11619
2	Groupe2	22707
3	Groupe3	65115

6-

a- A l'aide d'une « étape **DATA / merge** » créez une table *« customers31 »* qui fusionne les tables *« customers21 »* et *« customers22 »*. Dans cette même étape « DATA » créer une nouvelle colonne qui cumule le nombre de contrats par "*state_groupe*" et par "*anciennete*"

Obs.	state_groupe	anciennete	n_resiliation	n_contrat	cum_n_contrat
1	Groupe1	2	10	10	10
2	Groupe1	3	27	27	37
3	Groupe1	4	31	31	68
4	Groupe1	5	41	41	109
5	Groupe1	6	78	78	187
6	Groupe1	7	96	96	283
7	Groupe1	8	157	157	440
8	Groupe1	9	177	177	617
9	Groupe1	10	230	230	847
10	Groupe1	11	260	260	1107

b- A l'aide d'une « étape **DATA / merge** », créez une table « customers32 » qui fusionne les tables « customers23 » et « customers31 ». Dans la même étape DATA, par "*state_groupe*", calculez :

- Créez une colonne nommée "*n_risque*" par le calcul suivant
  n_risque = n_cohorte + n_contrat - cum_n_contrat
- Créez une colonne nommée "*tx_survie*" par le calcul suivant
  *t**x_survie = LOG(1 - (n_resiliation/n_risque))***
- Créer une colonne nommée "*estim*", initialisée à 0 pour chaque "*state_groupe*" et qui donne la somme cumulée du "*tx_survie"*
- Créez une colonne nommée " *estimateur_survie*" qui applique la fonction exponentielle à la colonne "*estim*"

Obs.	state_groupe	anciennete	n_resiliation	n_contrat	cum_n_contrat	n_cohorte	n_risque	tx_survie	estim	estimateur_survie
1	Groupe1	2	10	10	10	11619	11619	-0.00086	-0.00086	0.99914
2	Groupe1	3	27	27	37	11619	11609	-0.00233	-0.00319	0.99682
3	Groupe1	4	31	31	68	11619	11582	-0.00268	-0.00587	0.99415
4	Groupe1	5	41	41	109	11619	11551	-0.00356	-0.00943	0.99062
5	Groupe1	6	78	78	187	11619	11510	-0.00680	-0.01623	0.98391
6	Groupe1	7	96	96	283	11619	11432	-0.00843	-0.02466	0.97564
7	Groupe1	8	157	157	440	11619	11336	-0.01395	-0.03860	0.96213
8	Groupe1	9	177	177	617	11619	11179	-0.01596	-0.05456	0.94690
9	Groupe1	10	230	230	847	11619	11002	-0.02113	-0.07569	0.92710
10	Groupe1	11	260	260	1107	11619	10772	-0.02443	-0.10012	0.90473

7- A l'aide de Microsoft Excel, tracez sur un même graphique, les courbes qui donnent l'estimateur de survie par ancienneté pour chaque groupe. La restitution doit se faire sur le même fichier que la question 3.

<u>PARTIE 2</u> : SAS SQL

1- Écrivez le programme SAS qui permet d'obtenir le nombre distinct de clients par groupe d'État et par type de carte. Ordonnez les résultats dans l'ordre décroissant suivant le nombre de clients. Utiliser la variable « *customer_id* ».

2- En partant de la requête précédente, écrivez la requête qui permet d'obtenir le nombre de commandes qui ont été passées au mois de juin 2017. Précisez également dans la même requête par combien de clients ont-elles été passées. Affichez les résultats par état du consommateur et type de carte de fidélité. Ordonnez les résultats dans l'ordre décroissant suivant le nombre de clients.

3- Écrivez le programme SAS qui permet d'obtenir pour les produits de poids supérieur à 29000, le nombre de commandes concernés. Afficher le résultat suivant le nom des produits en anglais.

4- Par type de carte de fidélité, affichez le nombre de commandes associées, chiffre d'affaires total des commandes, le minimum, moyen et maximum et l'écart type des montants de commandes.

loyalty_card_type	NB_commandes	CA_total	CA_min	CA_moy	CA_max	CA_std
basic	44096	7136227	0	154.8997	13664.08	225.0744
premium	22173	3540303	0	152.731	4764.34	205.7547
silver	33171	5332342	0	153.9537	6081.54	214.8654

5- Écrivez la requête SQL qui permet de déterminer pour chaque de État, le chiffre d'affaires total, le nombre de commandes réalisées, le panier moyen et le nombre moyen d'UVC.
NB : Panier moyen = Chiffre d'affaires / nombre de commandes
Nombre moyen d'UVC = Nombre total de produits / nombre de commandes

customer_state	CA_total	NB_commandes	NB_clients	NB_produits	Panier_moyen	NB_uvc
SP	7597210	41374	41374	49566	183.6228	1.197999
RJ	2769347	12762	12762	15327	216.9995	1.200987
MG	2326152	11544	11544	13638	201.5031	1.181393
RS	1147277	5432	5432	6486	211.2071	1.194035

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

Le principe de cette partie est de créer un programme automatisé qui réalise un échantillonnage 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 pour chaque valeur de la variable de stratification. Ensuite un sous échantillon aléatoire simple 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 la table « *customers* » créée dans la partie SQL.

1- Programme AS1
Créez un programme SAS à l'aide d'une étape **« Data »**, sans aucun macro-langage, qui : créez une variable aléatoire nommée « **i** » en utilisant la fonction ***« ranuni (0) »*** de SAS.
Triez par cette variable « **i** » et créez un échantillon avec les 5000 premières observations.

2- Programme AS2
Reprenez le programme AS1, 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, ainsi que la taille de l'échantillon (nombre d'observations).

3- Programme AS3
Reprenez le programme AS2 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. Modifiez le programme en conséquence.
Attention : il vous faudra utiliser l'instruction « call symputx » et la macro-fonction « %sysevalf ».

4- Programme AS4
Transformez le programme AS3 en un macro-programme « **%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 (le type de carte par exemple)

1- Programme ASTR1
Créez un macro-programme « **%ASTR1** » qui permet de collecter dans des macro variables les valeurs prises par la variable de stratification choisie, ainsi que leur effectif respectif.
Ce macro-programme aura comme paramètres : la table en entrée ainsi que la variable de stratification.
Attention : Ne pas prendre en compte les valeurs manquantes pour la variable de stratification.

2- Programme ASTR2
Reprenez le macro-programme « **%ASTR1** » et ajoutez-y une partie qui éclate la table en entrée en plusieurs strates : c'est-à-dire qu'il faut créer une table par strate.

3- Programme ASTR3
Reprenez le programme « **%ASTR2** » et adaptez le en ajoutant une partie qui crée les sous échantillons (un échantillon pour chaque strate). Utilisez la fonction ***« ranuni (0) »*** de SAS en vous inspirant du A-/
Attention : rajoutez dans les paramètres du macro-programme le taux d'échantillonnage et adaptez votre programme en conséquence.

4- Programme ASTR4
Reprenez le programme « **%ASTR3** » et ajoutez une partie qui joint les sous échantillons en une seule table SAS.