Université Alioune Diop de Bambey

UFR : Sciences Appliquées et Technologies de l'Information et de la Communication (SATIC)

DEPARTEMENT : Mathématiques

SPECIALITE : Statistique et Informatique Décisionnelle (SID)

PROJET DE DATA MINING AVEC SAS

Présenté par :

Ablaye Sow Sidibe, Mouhamadou KANE et Birane Seye

Master Statistique et Informatique Décisionnelle

SUJET :

Analyse les facteurs clés qui conduisent à la réclamation d'une compagnie d'assurance.

Encadreur académique : Pr Souleymane SAM

Année 2020-2021

AVANT-PROPOS

L'Université d'Alioune Diop de Bambey (UADB) est l'héritière du Centre Universitaire Régional (CUR) de Bambey, dont la création remonte en 2004 (cf. décret 2004-916 portant « création et organisation d'un CUR à Bambey »). Depuis 2009, le CUR de Bambey a été transformé en université de plein exercice, suite à la promulgation du décret 2009-1221, portant « création, organisation et fonctionnement de l'Université de Bambey ». L'UADB comporte en son sein 3 unités de formation et de recherche (UFR) qui sont les suivants :

- L'UFR Sciences Appliquées et Technologies de l'Information et de la Communication (SATIC)
- L'UFR Economie - Management et Ingénierie Juridique (ECOMIJ)
- L'UFR Santé et Développement Durable (SDD) Et un Institut de Formation à Distance-UADB (IFOAD).

L'Université de Bambey a pour mission :

- De dispenser des formations de pointe, et parfaitement conformes aux besoins exprimés sur les marchés de l'emploi ;
- De contribuer à la diversification au niveau national de l'offre de formation universitaire ;
- De contribuer à la mise en place des conditions de qualification permanente des citoyens ;
- De contribuer tant au développement local (notamment par la promotion et le développement des atouts de son site d'emplacement et des zones voisines) que national (notamment par la formation d'une main d'œuvre qualifiée en mesure de contribuer à la création de richesses).

Agenda du projet

INTRODUCTION GENERALE

Selon Emile de Girardin « Le calcul des probabilités, appliqué à la mortalité humaine a donné naissance à une science nouvelle : celle des assurances.**».**

Effectuer un projet sur des données d'une compagnie d'assurance représente l'occasion d'enrichir ses connaissances à travers des travaux permettant de les compléter ou d'en acquérir de nouvelles. De notre point de vue, ils nous paraissaient nécessaires, sinon indispensable, d'être confronté à des situations et des nouvelles bases de données présentant un grand intérêt dans le secteur de Data mining.

Ainsi, la mission proposée et les éléments mis en œuvre dans ce projet ont été des éléments prépondérants dans le choix de celle-ci.

I. Contexte

Une **assurance** est un service qui fournit une prestation lors de la survenance d'un événement incertain et aléatoire souvent appelé "risque". En résumé c'est une souscription à un contrat qui permet de couvrir un bien contre le tiers et aux aléas pour une durée limitée et renouvelable

Etant au XXI^{ème} siècle avec la multiplicité des risques, il est devenu nécessaire de recourir aux services des assureurs pour une protection de soi, de ses activités et de ses biens. Pour ce faire, il est quasi nécessaire qu'il ait une interaction entre les assureurs, les assurés, les tiers et la règlementation (le code des assurances & le code Cima).

II. Problématique

Afin de pouvoir apporter des solutions aux différents problèmes liés aux réclamations pour accident des clients, plusieurs facteurs interviennent dans le processus de résolution.

Ainsi, il est important de se poser certaines questions et essayer d'en trouver des solutions.

III. Objectifs

Notre objectif dans ce projet est de mettre en place un type de model. Ce dernièr permettra de prévoir le risque de réclamation pour permettre à l'entreprise de prendre des décisions pour l'avenir afin de mieux gérer ses clients.

IV. Annonce du plan

Dans le cadre de bien restituer le travail, nous avons réparti notre plan en deux chapitres à savoir :

- Présenter nos données
- Construire notre model

A. REVU DE LA LITTERATURE

I. SCIENCE DES DONNEES ET AIDE A LA DECISION

1. Historique

Le terme science des données (data science en anglais) a été forgé lors du 2e colloque franco-japonais de statistique tenu à l'Université Montpellier II (France). Les participants ont reconnu l'émergence d'une nouvelle discipline au cœur de laquelle se trouvent des données de toutes origines, tailles, types et structures. Cette activité doit s'appuyer sur des concepts et des principes reconnus de la statistique et de l'analyse des données tout en exploitant pleinement la puissance croissante des outils informatiques. En 2001, William Cleveland reprenait essentiellement les mêmes idées dans un article programmatique paru en 2001 « Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics » qui précise les contours de cette discipline émergente.
Cette discipline est issue de l'apparition et du développement des bases de données et de l'Internet et répond aussi à la complexité croissante et au volume en croissance exponentielle du nombre de données numériques disponibles dans le monde (infobésité).
Elle a reçu beaucoup d'attention dernièrement grâce à l'intérêt grandissant pour les "données massives". Cependant, la science des données ne se limite pas à l'étude de bases de données pouvant être qualifiées de "données massives".
Par ailleurs, l'essor de techniques d'apprentissage automatique (en anglais machine learning) et d'intelligence artificielle a également participé à la croissance de cette discipline et à son ouverture vers de nouveaux champs en passant, par exemple, de l'analyse statistique pure de données fortement structurées à l'analyse de données semi-structurées (XML par exemple) pour notamment mettre « en correspondance des bases de données et de données textuelles ».

2. Contexte

Avec un nombre de plus en plus élevé, d'organisations, la prise de décisions est effectuée en fonction d'informations obtenues via le traitement et l'analyse de grands volumes de données qui ont aussi les caractéristiques d'être hétérogènes.
Dans ce contexte, les décideurs doivent être capables d'apprendre, de synthétiser l'information, de la rendre disponible sous formes accessibles à des utilisateurs de nature variée en considérant la nature et la collocation particulière de l'organisation, qui peut être de type privée/économique ou publique.

3. Science des données

La mission principale du scientifique des données, est de contribuer à l'identification, la formalisation, la conception et la validation expérimentale d'algorithmes et de méthodes formelles pour supporter les décideurs dans ces deux phases critiques.

4. Aide à la décision

L'aide à la Décision se caractérise par une vision transversale de sa thématique, en mettant l'accent d'une part sur le développement de modèles et d'algorithmes qui visent à apporter des éléments de réponse à de nombreux problèmes de décision (à la fois dans un cadre de processus de décision « humain » et dans un cadre de processus de décision « automatique ») ;

d’autre part sur l’élaboration d’une démarche constructive d’aide à la décision, fondée à la fois sur les « théories de la décision et du raisonnement » et sur les expériences d’aide à la décision réelles, qui permet d’élaborer des recommandations pratiques sur la conduite des processus d’aide à la décision et le développement des systèmes d’information et d’aide à la décision.

B. DONNEES

Dans ce module nous allons faire une brève présentation de notre base de données, pour ensuite faire une analyse descriptive de nos variables et enfin faire un traitement de données.

I. La base de données

Nous allons afficher un aperçu de notre base de données :

ContractId	Age	Gender	Children	Profession	Customer_Type	Multiple_cars	Driving_Licence_Years	Car_category	Annual_Kilometers	Gearbox	Fuel	Claim
1	55	Woman	4+	Retired	Agency	No	14	Saloon	24074	Automatic	Diesel	No
2	58	Man	1	Unemployed	Agency	No	13	Sport	19328	Manual	Diesel	No
3	26	Woman	4+	Independant	Agency	No	3	Sport	21141	Automatic	Diesel	No
4	47	Woman	1	Unemployed	Agency	No	22	Estate	41040	Manual	Petrol	No
5	54	Woman	2	Private Sector - Manager	On-line	No	18	Estate	31620	Manual	Diesel	No
6	78	Man	4+	Public Sector - Manager	Agency	No	18	Estate	17837	Manual	Diesel	No
7	39	Man	2	Independant	On-line	No	21	Saloon	20065	Automatic	Diesel	No
8	43	Woman	2	Retired	Agency	No	11	Saloon	35688	Automatic	Diesel	No
9	51	Woman	0	Private Sector - Employee	Agency	Yes	20	Sport	24503	Manual	Diesel	No
10	42	Woman	0	Public Sector - Manager	Agency	Yes	18	Estate	24053	Manual	Diesel	No
11	52	Woman	1	Student	Agency	Yes	18	Estate	27859	Automatic	Diesel	No
13	38	Man	3	Public Sector - Director	Agency	No	20	Sport	21000	Automatic	Diesel	No
15	20	Woman	2	Independant	Agency	No	2	Saloon	19898	Automatic	Diesel	No
16	36	Man	0	Independant	Agency	No	18	Sport	22955	Manual	Diesel	No
17	46	Man	2	Independant	On-line	No	7	Saloon	18803	Manual	Petrol	No

La base de données comporte plusieurs variables mais on a besoin de connaitre les dimension de notre base et aussi la nature de nos variables.

La procédure CONTENTS

Nom de la table	MOUHAMED.ASSURANCE	Observations	8221
Type de membre	DATA	Variables	13
Moteur	V9	Index	0
Créée	25/09/1997 21:54:55	Longueur d'observation	112
Dernière modification	25/09/1997 21:54:55	Observations supprimées	0
Protection		Compressée	NON
Type de table		Triée	NON
Libellé			
Représentation des données	WINDOWS_32		
Codage	wlatin1 Western (Windows)		

En résumé on a 8221 observations et 13 variables

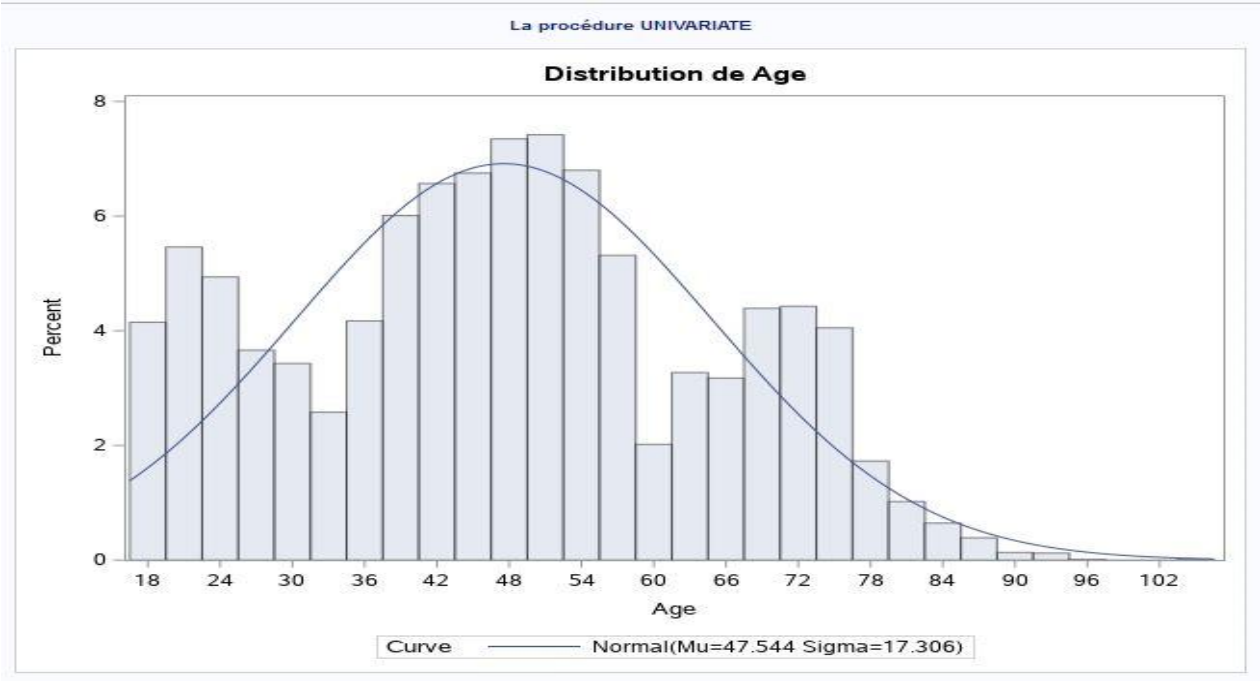
Liste alphabétique des variables et des attributs					
#	Variable	Type	Long.	Format	Informat
2	Age	Num.	8	BEST12.	BEST32.
10	Annual_Kilometers	Num.	8	BEST12.	BEST32.
9	Car_category	Texte	6	$6.	$6.
4	Children	Texte	2	$2.	$2.
13	Claim	Texte	2	$2.	$2.
1	ContractId	Num.	8	BEST12.	BEST32.
6	Customer_Type	Texte	8	$8.	$8.
8	Driving_Licence_Years	Num.	8	BEST12.	BEST32.
12	Fuel	Texte	6	$6.	$6.
11	Gearbox	Texte	9	$9.	$9.
3	Gender	Texte	5	$5.	$5.
7	Multiple_cars	Texte	4	$4.	$4.
5	Profession	Texte	31	$31.	$31.

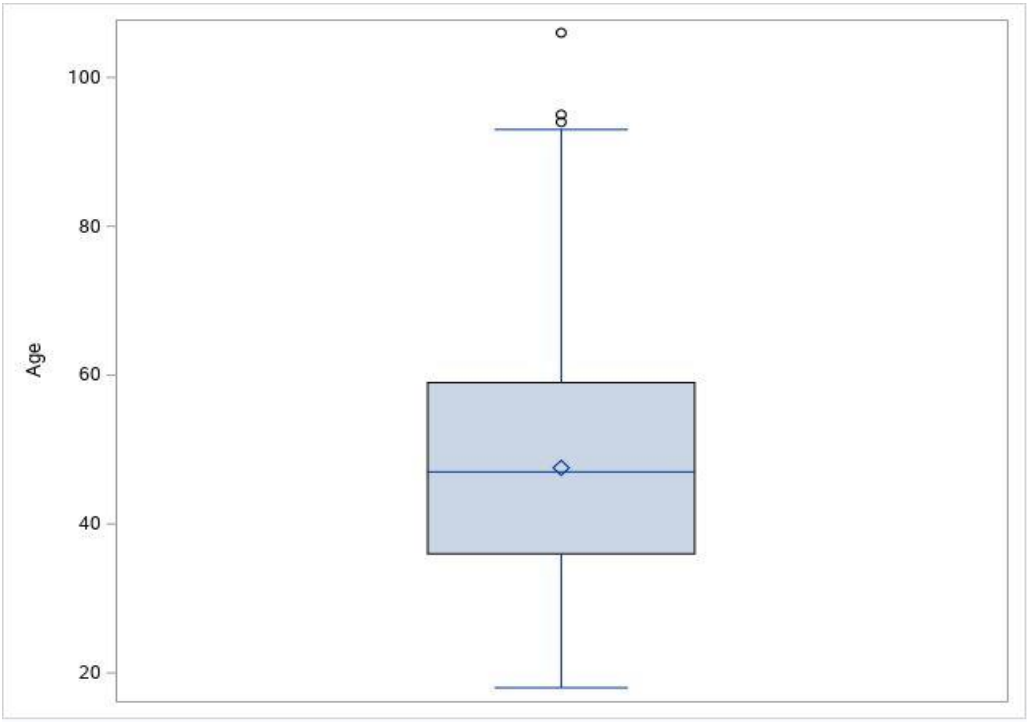
Parmi nos 13 variables la majorité sont des qualitatifs.

II. Analyse descriptive

Dans cette partie nous allons faire une étude descriptive de certaines de nos variables pour comprendre nos données.

- Pour la variables **Age**



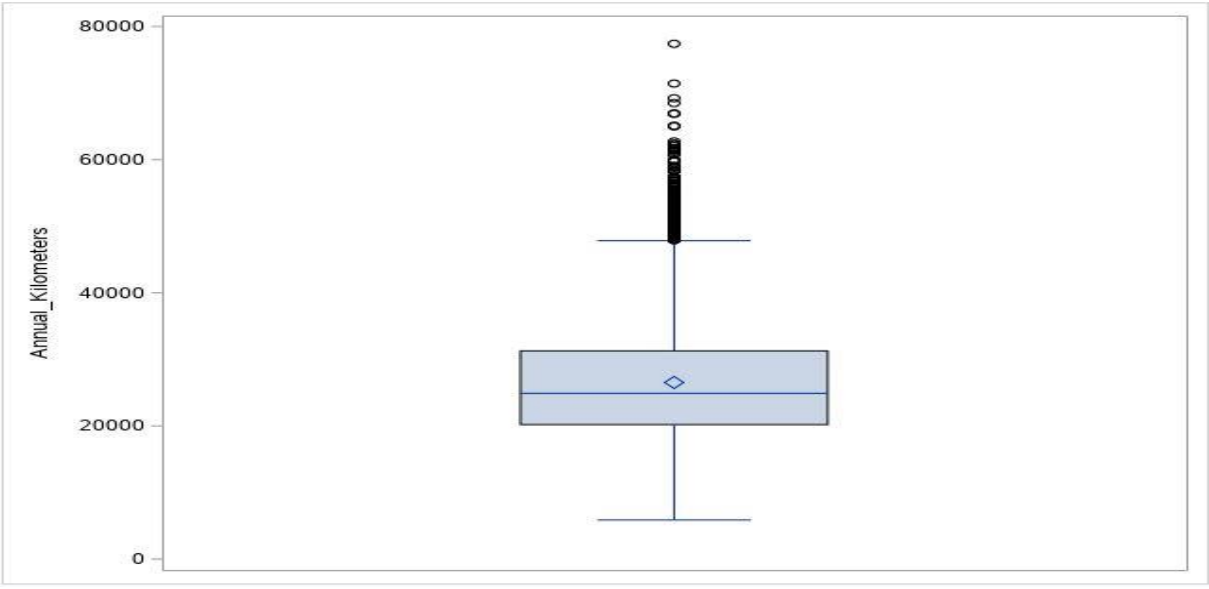


On voit que notre variable Age est bien distribue est aussi dans la base y'a beaucoup plus des clients qui ont un âge compris [42-54].

Le boxplot nous donne 3 observations qui sont des potentielles valeurs aberrantes. Dans notre suite on ne va pas les supprimer on va les conserver.

- Pour la variable **Annual_Kilometers**

Nous allons faire une analyse sur la variable Annual_Kilometers

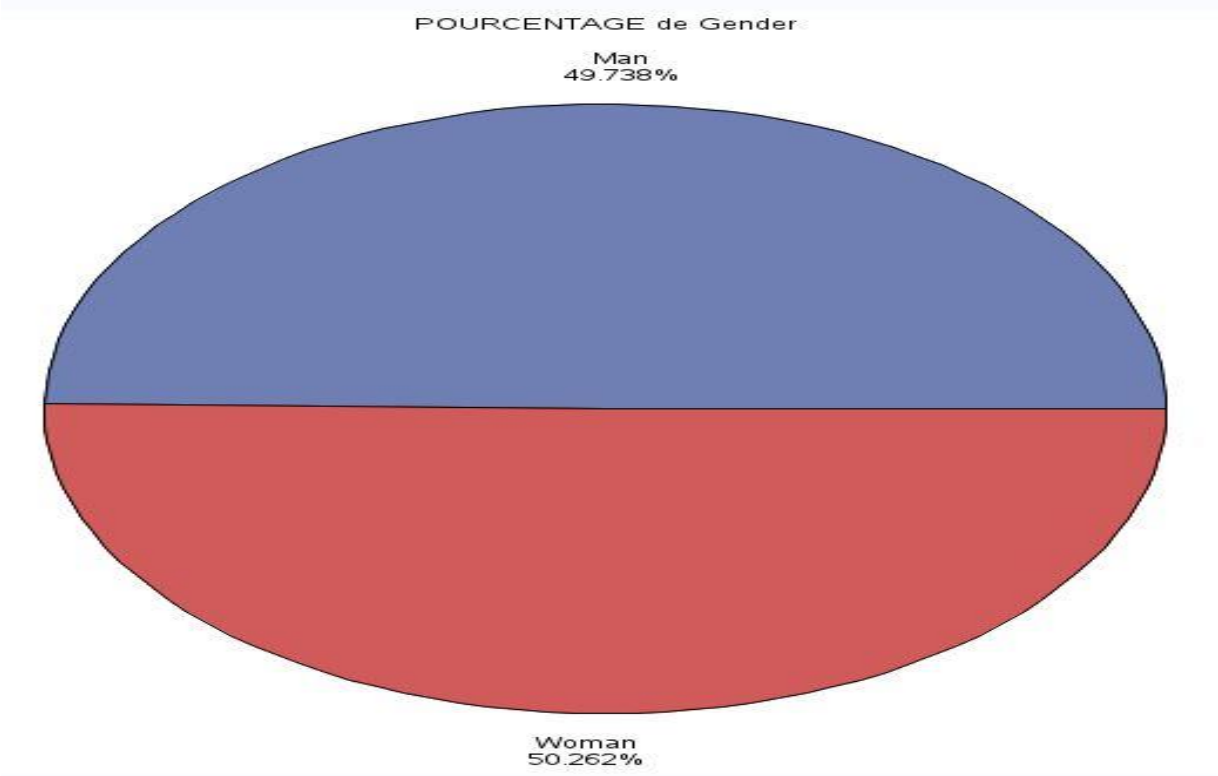


On constate que la variable **Annual_Kilometers** comportes excessivement des valeurs aberrantes donc on ne va pas conserver cette variable dans notre analyse**.**

- Pour la variable **ContractId**

Cette variable n'a pas une grande importance dans l'analyse de nos données.

- Pour la variable **Gender**



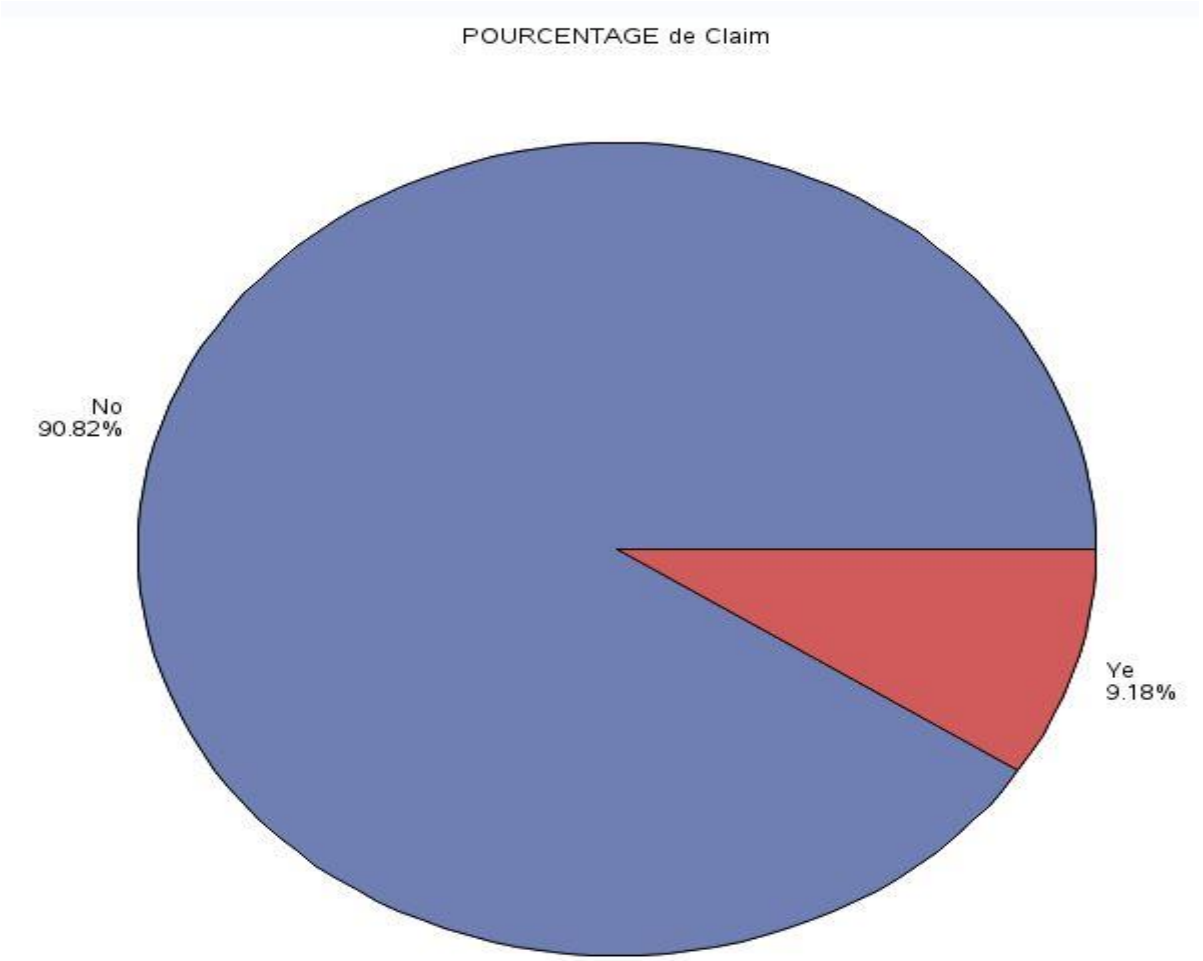
On a une repartition de notre variable Gender qui est presque equitable c'est-à-dire :

49.74 % d'homme et 50.26 % de femme.

- Pour la variable **Claim**

Cette variable constitue notre « target » c'est-à-dire notre variable à expliquer.

C'est une variable qualitative bimodale



On une répartition de 90.82 % des clients qui n'ont pas fait de réclamations par contre seulement 9.18 % on fait une réclamation.

On va aussi faire une analyse descriptive de nos variables qualitative par rapport à la variable a expliquer. Mais dans cette analyse on s'intéressera uniquement qu'à la modalité « **Ye** » de la variable « **Claim** »

- ✓ Gender et Claim

La procédure FREQ

Fréquence Pourcentage Pct de ligne Pct de col.	Table de Gender par Claim		
	Claim		
Gender	No	Ye	Total
Man	3589 43.66 87.77 48.07	500 6.08 12.23 66.23	4089 49.74
Woman	3877 47.16 93.83 51.93	255 3.10 6.17 33.77	4132 50.26
Total	7466 90.82	755 9.18	8221 100.00

Dans cette répartition on constate que les hommes qui font des réclamations sont au double des femmes qui le fond.

- ✓ Customer_Type et Claim

La procédure FREQ

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table de Customer_Type par Claim

Customer_Type	Claim		
	No	Ye	Total
Agency	5508 67.00 90.65 73.77	568 6.91 9.35 75.23	6076 73.91
On-line	1958 23.82 91.28 26.23	187 2.27 8.72 24.77	2145 26.09
Total	7466 90.82	755 9.18	8221 100.00

Le nombre de réclamation enregistrer par la modalité « Agency » et le triple de la modalité « On-line ».

- ✓ Children et Claim

La procédure FREQ

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table de Children par Claim

Children	Claim		
	No	Ye	Total
0	1116 13.57 83.72 14.95	217 2.64 16.28 28.74	1333 16.21
1	1392 16.93 90.98 18.64	138 1.68 9.02 18.28	1530 18.61
2	2184 26.57 90.77 29.25	222 2.70 9.23 29.40	2406 29.27
3	293 3.56 90.71 3.92	30 0.36 9.29 3.97	323 3.93
4+	2481 30.18 94.37 33.23	148 1.80 5.63 19.60	2629 31.98
Total	7466 90.82	755 9.18	8221 100.00

On constate que la variable « Children » a un bon pourcentage de réclamation. Seul les clients qui ont 3 enfants qui ont fait moins de réclamation.

- ✓ Multiple_cars et Claim

La procédure FREQ

Fréquence Pourcentage Pct de ligne Pct de col.	Table de Multiple_cars par Claim		
	Claim		
Multiple_cars	No	Ye	Total
No	4965 60.39 90.64 66.50	513 6.24 9.36 67.95	5478 66.63
Yes	2501 30.42 91.18 33.50	242 2.94 8.82 32.05	2743 33.37
Total	7466 90.82	755 9.18	8221 100.00

Les clients qui n'ont qu'une seule voiture font 3 de plus de réclamations qu'aux clients qui en ont plusieurs voitures.

- ✓ Car_category et Claim

La procédure FREQ

Fréquence Pourcentage Pct de ligne Pct de col.	Table de Car_category par Claim		
	Claim		
Car_category	No	Ye	Total
Estate	2695 32.78 94.79 36.10	148 1.80 5.21 19.60	2843 34.58
Saloon	2435 29.62 90.59 32.61	253 3.08 9.41 33.51	2688 32.70
Sport	2336 28.42 86.84 31.29	354 4.31 13.16 46.89	2690 32.72
Total	7466 90.82	755 9.18	8221 100.00

Les clients qui ont une des voitures de catégories « Estate » font moins de réclamations que les autres catégories.

- ✓ Profession et Claim

La procédure FREQ

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table de Profession par Claim

Profession	Claim		
	No	Ye	Total
Government	619 7.53 88.81 8.29	78 0.95 11.19 10.33	697 8.48
Independant	1234 15.01 91.20 16.53	119 1.45 8.80 15.76	1353 16.46
Private Sector - Director	600 7.30 91.60 8.04	55 0.67 8.40 7.28	655 7.97
Private Sector - Employee	653 7.94 94.64 8.75	37 0.45 5.36 4.90	690 8.39
Private Sector - Manager	596 7.25 90.85 7.98	60 0.73 9.15 7.95	656 7.98
Public Sector - Director	679 8.26 93.78 9.09	45 0.55 6.22 5.96	724 8.81
Public Sector - Employee	620 7.54 90.78 8.30	63 0.77 9.22 8.34	683 8.31
Public Sector - Manager	582 7.08 84.96 7.80	103 1.25 15.04 13.64	685 8.33
Retired	625 7.60 90.58 8.37	65 0.79 9.42 8.61	690 8.39
Student	652 7.93 91.32 8.73	62 0.75 8.68 8.21	714 8.69
Unemployed	606 7.37 89.91 8.12	68 0.83 10.09 9.01	674 8.20
Total	7466 90.82	755 9.18	8221 100.00

Les clients de profession « Independant » et « Public sector-Manager » font plus de réclamations que les autres.

- ✓ Gearbox et Claim

La procédure FREQ

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table de Gearbox par Claim

Gearbox	Claim		
	No	Ye	Total
Automatic	5249 63.85 90.55 70.31	548 6.67 9.45 72.58	5797 70.51
Manual	2217 26.97 91.46 29.69	207 2.52 8.54 27.42	2424 29.49
Total	7466 90.82	755 9.18	8221 100.00

Les clients dont leurs voitures ont un levier de vitesse « Automatic » font 3 fois de plus de réclamation qu'aux voitures avec un levier de vitesse « Manual ».

- ✓ Fuel et Claim

La procédure FREQ

Fréquence Pourcentage Pct de ligne Pct de col.	Table de Fuel par Claim		
	Claim		
Fuel	No	Ye	Total
Diesel	5160 62.77 89.58 69.11	600 7.30 10.42 79.47	5760 70.06
Petrol	2306 28.05 93.70 30.89	155 1.89 6.30 20.53	2461 29.94
Total	7466 90.82	755 9.18	8221 100.00

Les clients qui utilisent « Diesel » comme carburant font beaucoup plus de réclamations que ceux qui utilisent « Petrole » comme carburant.

III. Traitement des données

Dans cette section on va essayer de préparer notre base de données :

- ✓ Recoder certains de variables

On va recoder notre variable « Claim » par « 0 » à la place « No » et « 1 » à la place de « Ye »

On va aussi recoder notre variable « Children ».

- ✓ Dicredisation la variables « Age »

On va créer des classes d'Age pour mieux faire une bonne analyse :

De 18ans a 35ans : Ce sont les jeunes adultes et c'est la classe « 1 »

De 36ans a 65ans : Ce sont les adultes et c'est la classe « 2 »

65 ans et plus: Ce sont les vieux et c'est la classe « 3 »

- ✓ Selectionnner les variables à retenir

Dans cette partie on va supprimer la variable **Annual_Kilometers** car il comporte trop de variables aberrantes et la variable **ContractID** pour la modélisation.

C. CONSTRUCTION DE MODEL

Pour mieux construire un bon model, on va proceder ainsi :

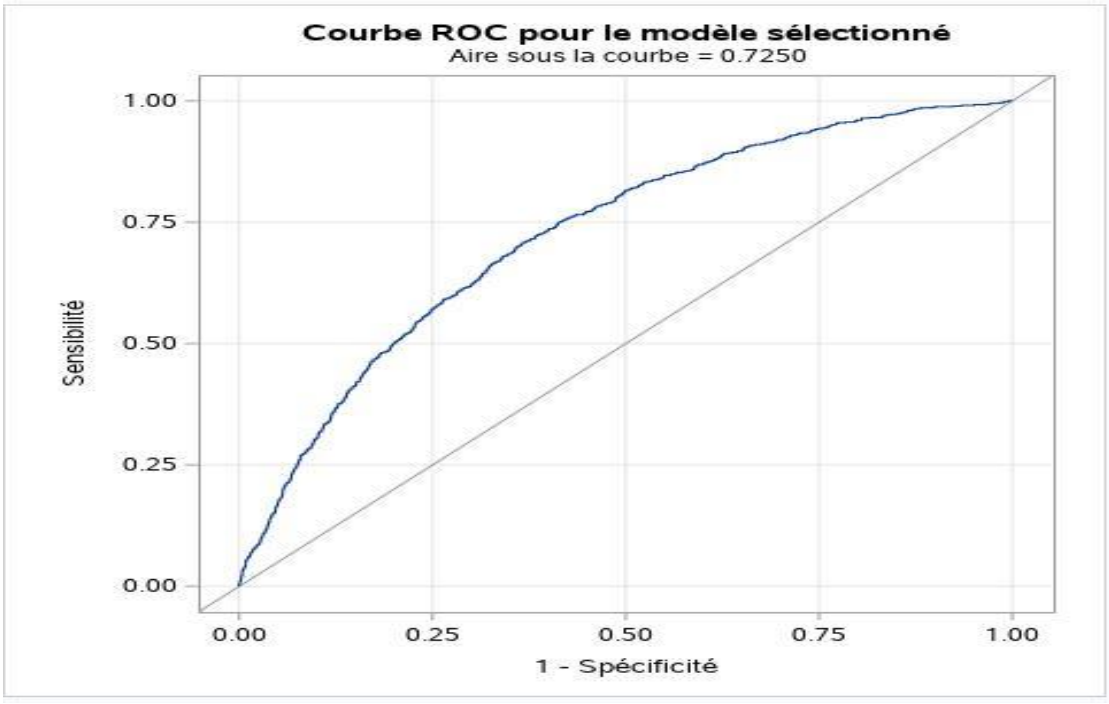
I. Selection de variable avec STEPWISE sous SAS

Vue qu'on une variable target binaire on va donc faire une regression logistique binaire et annalyser nos resultats :

Récapitulatif sur la sélection Stepwise							
Etape	Effet		DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2
	Saisi	Supprimé					
1	Children		4	1	120.3402		<.0001
2	Car_category		2	2	106.8791		<.0001
3	Gender		1	3	94.1053		<.0001
4	Fuel		1	4	38.4548		<.0001
5	Profession		10	5	48.1634		<.0001
6	Age		2	6	23.2410		<.0001
7	Driving_Licence_Year		1	7	13.9584		0.0002

Ce tableau nous résume les étapes pour la sélection des variables qui est aux nombre de « 7 » étapes. On constate que toutes les variables sont significatives.

Maintenant on va voir la courbe de ROC et son score :



Nous avons un score de 72.50 % qui est acceptable comme un bon model donc on peut proceder a l'échantillonnage de notre base.

II. Echantillonnage de notre base de données

Nous allons scinder notre base en deux :

- ✓ ASSURANCE_Apprentissage : qui sera notre base d'apprentissage ou on va construire le model.
- ✓ ASSURANCE_Validation : qui sera une base pour permettre de valider notre model.

Pour cela nous allons appliquer la regle 70% de la base pour l'apprentissage et 30% pour la validation.

La procédure CONTENTS

Nom de la table	MOUHAMED.ASSURANCE_APPRENTISSAGE	Observations	5821
Type de membre	DATA	Variables	14
Moteur	V9	Index	0
Créée	31/01/2022 19:42:55	Longueur d'observation	112
Dernière modification	31/01/2022 19:42:55	Observations supprimées	0
Protection		Compressée	NON
Type de table		Triée	NON
Libellé			
Représentation des données	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Codage	utf-8 Unicode (UTF-8)		

Claim	ContractId	Gender	Children	Profession	Customer_Type	Multiple_cars	Driving_Licence_Years	Car_category	Annual_Kilometers	Gearbox	Fuel	Age	Echantillon
0	2	Man	1	Unemployed	Agency	No	13	Sport	19328	Manual	Diesel	2	1
0	3	Woman	4+	Independant	Agency	No	3	Sport	21141	Automatic	Diesel	1	1
0	5	Woman	2	Private Sector - Manager	On-line	No	18	Estate	31620	Manual	Diesel	2	1
0	6	Man	4+	Public Sector - Manager	Agency	No	18	Estate	17837	Manual	Diesel	3	1
0	8	Woman	2	Retired	Agency	No	11	Saloon	35688	Automatic	Diesel	2	1
0	9	Woman	0	Private Sector - Employee	Agency	Yes	20	Sport	24503	Manual	Diesel	2	1
0	10	Woman	0	Public Sector - Manager	Agency	Yes	18	Estate	24053	Manual	Diesel	2	1
0	13	Man	3	Public Sector - Director	Agency	No	20	Sport	21000	Automatic	Diesel	2	1
0	16	Man	0	Independant	Agency	No	18	Sport	22955	Manual	Diesel	2	1
0	18	Man	0	Private Sector - Employee	Agency	Yes	13	Saloon	21278	Automatic	Diesel	2	1

La base d'apprentissage comporte 5821 d'observation qui représente exactement les 70 %

La procédure CONTENTS

Nom de la table	MOUHAMED.ASSURANCE_VALIDATION	Observations	2400
Type de membre	DATA	Variables	14
Moteur	V9	Index	0
Créée	31/01/2022 18:33:48	Longueur d'observation	112
Dernière modification	31/01/2022 18:33:48	Observations supprimées	0
Protection		Compressée	NON
Type de table		Triée	NON
Libellé			
Représentation des données	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Codage	utf-8 Unicode (UTF-8)		

Claim	ContractId	Gender	Children	Profession	Customer_Type	Multiple_cars	Driving_Licence_Years	Car_category	Annual_Kilometers	Gearbox	Fuel	Age	Echantillon
0	1	Woman	4+	Retired	Agency	No	14	Saloon	24074	Automatic	Diesel	2	2
0	4	Woman	1	Unemployed	Agency	No	22	Estate	41040	Manual	Petrol	2	2
0	7	Man	2	Independant	On-line	No	21	Saloon	20065	Automatic	Diesel	2	2
0	11	Woman	1	Student	Agency	Yes	18	Estate	27859	Automatic	Diesel	2	2
0	15	Woman	2	Independant	Agency	No	2	Saloon	19898	Automatic	Diesel	1	2
0	17	Man	2	Independant	On-line	No	7	Saloon	18803	Manual	Petrol	2	2
0	24	Man	1	Public Sector - Director	Agency	No	15	Sport	19236	Manual	Petrol	3	2
0	26	Man	1	Private Sector - Employee	On-line	No	8	Saloon	14524	Automatic	Petrol	1	2
0	29	Man	2	Independant	Agency	No	8	Saloon	21365	Automatic	Petrol	1	2
0	31	Woman	4+	Public Sector - Manager	Agency	No	21	Sport	26608	Manual	Diesel	3	2

La base de validation comporte 2400 d'observation qui représente exactement les 30 %

Ainsi on va enfin afficher notre base de prevision et l'interpreter.

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	48.8	D de Somers	0.022
Pourcentage discordant	46.6	Gamma	0.023
Pourcentage lié	4.6	Tau-a	0.004
Paires	5636830	c	0.511

On a un score égal à 50.1 qui n'est pas excellent mais aussi acceptable du fait qu'on a supprimé une variable à cause de ses variables aberrantes.

Donc on va valider ce model comme notre model final.

Claim	ContractId	Gender	Children	Profession	Customer_Type	Multiple_cars	Driving_Licence_Years	Car_category	Annual_Kilometers	Gearbox	Fuel	Age	Echantillon	F_Claim	I_Claim	P_0	P_1
0	1	Woman	4+	Retired	Agency	No	14	Saloon	24074	Automatic	Diesel	2	2	0	0	0.90156	0.098437
0	4	Woman	1	Unemployed	Agency	No	22	Estate	41040	Manual	Petrol	2	2	0	0	0.90156	0.098437
0	7	Man	2	Independant	On-line	No	21	Saloon	20065	Automatic	Diesel	2	2	0	0	0.90156	0.098437
0	11	Woman	1	Student	Agency	Yes	18	Estate	27859	Automatic	Diesel	2	2	0	0	0.90156	0.098437
0	15	Woman	2	Independant	Agency	No	2	Saloon	19898	Automatic	Diesel	1	2	0	0	0.90156	0.098437
0	17	Man	2	Independant	On-line	No	7	Saloon	18803	Manual	Petrol	2	2	0	0	0.90156	0.098437
0	24	Man	1	Public Sector - Director	Agency	No	15	Sport	19236	Manual	Petrol	3	2	0	0	0.90156	0.098437
0	26	Man	1	Private Sector - Employee	On-line	No	8	Saloon	14524	Automatic	Petrol	1	2	0	0	0.90156	0.098437
0	29	Man	2	Independant	Agency	No	8	Saloon	21365	Automatic	Petrol	1	2	0	0	0.90156	0.098437
0	31	Woman	4+	Public Sector - Manager	Agency	No	21	Sport	26608	Manual	Diesel	3	2	0	0	0.90156	0.098437
0	34	Man	2	Private Sector - Manager	On-line	No	12	Estate	20428	Manual	Petrol	2	2	0	0	0.90156	0.098437
0	53	Man	1	Student	Agency	Yes	21	Saloon	18550	Manual	Diesel	2	2	0	0	0.90156	0.098437
0	55	Man	1	Private Sector - Director	On-line	No	2	Saloon	16763	Automatic	Diesel	1	2	0	0	0.90156	0.098437
0	56	Man	1	Independant	On-line	No	7	Estate	21494	Automatic	Petrol	2	2	0	0	0.90156	0.098437
0	57	Woman	2	Retired	Agency	No	9	Estate	20177	Manual	Petrol	1	2	0	0	0.90156	0.098437
0	60	Man	3	Independant	Agency	No	11	Sport	18844	Automatic	Diesel	2	2	0	0	0.90156	0.098437
0	63	Woman	0	Private Sector - Employee	Agency	No	21	Estate	10306	Automatic	Diesel	2	2	0	0	0.90156	0.098437
0	64	Woman	2	Retired	On-line	No	1	Estate	28924	Manual	Petrol	1	2	0	0	0.90156	0.098437
0	66	Man	4+	Private Sector - Director	Agency	No	10	Sport	21559	Manual	Diesel	3	2	0	0	0.90156	0.098437
0	68	Woman	2	Public Sector - Director	Agency	Yes	7	Sport	23797	Automatic	Petrol	2	2	0	0	0.90156	0.098437

CONCLUSION GENERALE

L'objectif de notre projet était de concevoir model statistique pour permettre à une compagnie d'assurance de pouvoir prédire les risques de réclamations.

Pour mener à bien notre mission, nous avons mis en œuvre toutes nos connaissances acquises durant notre cursus universitaire, ajouté à l'expérience bénéfique lors des cours de Data mining pratique effectué avec notre professeur.

D'un côté, nous nous trouvons face à des difficultés nouvelles. Il nous a fallu du temps pour comprendre la base de données et aussi le SAS en ligne avec lequel on a fait notre projet.

En outre, ce projet nous a offert la possibilité de revoir et de découvrir avec beaucoup d'attention plusieurs notions statistique, grâce auxquelles le projet a été mise au point.

Enfin, un des avantages majeurs de ce projet a été le travail en groupe. Ce n'est bien sûr pas un travail en grande équipe mais nous apprend déjà à bien répartir le travail, à se concerter régulièrement pour organiser le travail et à s'enrichir mutuellement en partageant des idées. Cela a été très formateur tout au long du travail.

Certes, le travail n'est pas encore arrivé à son terme ; il manque encore à refaire certains travaux sur l'analyse descriptive mais aussi le déploiement. Cela nous permettra de se fixer des objectifs dans le futur afin d'obtenir un projet digne de son nom.