Équipe 8

Charles Comeau
(111 185 421)

Nicholas Langevin
(111 184 631)

Andréanne Larouche
(111 190 518)

Apprentissage statistique en actuariat
ACT-3114

Analyse des données de renouvellement d'assurance

présenté à
Marie-Pier Côté

École d'actuariat
Université Laval
27 février 2020

Contents

Introduction

Les données qui seront analysées dans ce rapport proviennent du jeu de données “eudirectlapse” du paquetage “CASdatasets” de R. Dans le but de modéliser le statut de renouvellement de polices d’assurance, représenté par la variable “lapse” dans ce cas-ci, il sera d’abord nécessaire de visualiser et de pré-traiter les données observées des 23 060 polices d’assurance. Il est à noter que la durée d’observation est de un an et que l’année visée et la compagnie demeurent inconnue. On pourra distinguer les statuts de renouvellement comme étant affiché à résigné (“Resignation”) ou à renouvellement (“Renew”) selon le cas approprié. Ce jeu de données est intéressant du fait qu’il permettera au fur de l’analyse de nous indiquer les variables types ayant un impact sur la décision de renouvellement de police des assurés d’une compagnie d’assurance X. En plus d’être un problème de nature actuarielle, le jeu de données choisi pourra nous permettre d’entammer une ouverture des réflexions possibles lorsque nous aurons à travailler dans une compagnie d’assurance. Étant trois personnes intéressés par l’assurance de dommages, ce problème nous semblait des plus appropriés et intéressant face à nos intérêts communs. Le nombre d’observations est également intéressant car il nous permettera de porter des conclusions précise avec assez de crédibilité sans toutefois être avoir à travailler avec un jeu de données inutilement trop volumineux. De plus, chaque variable explicative semble à prime à bord intéressante pour l’analyse et assez pertinente, ce que nous pourrons découvrir dans l’élaboration de ce travail pratique.

Analyse exploratoire des données

Variable réponse

La variable lapse indique si l'assuré à renouveller ou non sa police lors du renouvellement. Il s'agit de la variable exogène. Initialement, le choix du client était indiqué par une variable binaire. Si le client désirait résigner sa police lapse prenait la valeur 1, autrement elle prenait la valeur 0. À des fins de simplification et pour que la visualisation en soit amiliorée pour la suite, nous avons converti la variable en variable catégorielle à deux niveaux. La variable prendra maintenant la valeur **resignation** si le client résigne sa police et de **renouvellement** s'il la renouvelle.

On constate qu'il y a 23060 clients dont 20106 qui ont renouveller leur police d'assurance, ce qui représente une proportion de 87.19%. La variable réponse n'est donc pas symétrique et il sera important d'en tenir compte lors de la modélisation. L'analyse des variables explicatives contenus dans ce jeu de données nous permettera de mieux comprendre les causes de résignation et de créer des patrons pour ainsi arriver à bien modéliser la variable de renouvellement de police.

Variables catégorielles nominales

polholder_diffdriver

Cette variable représente la différence de statut qui pourrait avoir entre le propriétaire de la police et le conducteur principal.

Table 1:

Statut	Nombre d'observation
Conducteurs agée de 24+	1728
Commerciale	40
Conducteur apprenti de 17 ans	42
Partenaire de couple	8128
Utilisateur seul	11155
Jeunes utilisateurs	1955
Données manquantes	12

On constate que la plupart des voitures assurées est utilisée seulement par le détenteur de la police ou par l,assuré et son partenaire de couple puisque c'est deux cas représente 83.62% des observations. Il y a un pourcentage non négligeable de 8.48% pour lequel le véhicules est partagé par de jeunes conducteurs alors qu'il y a 1728% des cas ou le véhicule est plutôt partagé entre des personnes plus agées (24 ans et plus). À noté qu'il y a 12 observations pour lesquelles la variable est manquante. Cela sera traité dans la section traitement des valeurs manquantes.

polholder_gender

Cette variable représente le sexe du propriétaire de la police. Voici la répartition en pourcentage du sexe pour les propriétaires de police d'assurance.

Table 2:

Sexe	%
Homme	63.84
Femme	36.16

On voit qu'il y a significativement plus d'homme ayant une police d'assurance chez cet assureur que de femme.

polholder_job

Cette variable est, quant à elle, celle décrivant le travail du propriétaire du contrat. Deux valeurs sont possibles soit “medical” soit “normal”. On constate que 41.12% des assurés on un travail de type médical alors qu'il y en a 58.88% qui ont un autre type d'emploi.

policy_caruse

Cette variable représente les fins d'utilisation du véhicule.

Table 3:

Usage	Nombre d'observation
Commerciale	10
Privé ou aller travailler	19567
Données manquantes	3483

On constate qu'il y a un nombre considérable de données manquantes et très peu de véhicule pour un usage commerciale.

vehicl_garage

De sont côté, cette variable décrit le type de stationnement de la voiture. Voici la répartition des types de stationnement.

Table 4:

Moyen de stationnement	Nombre d'observation
Sous un abri d'auto	1413
Terrase de stationnement	2243
Stationnement privé	199
Garage	8863
Rue	5468
Garage sous-terrain	1056
Autre	2243
Données manquantes	1575

On voit que pour les moyens de stationnement les plus populaire sont le garage privé et la rue. Il y a des données manquantes, elles seront traitées plus loin dans le rapport.

La variable **polholder_BMCevol** indique si la prime de renouvellement à connue une hausse, une baisse ou est demeuré stable par rapport à la prime payée lors du dernier renouvellement.

Table 5:

Prime de renouvellement	Nombre d'observation
Hausse	869
Inchangée	12036
Baisse	10155

On constate que la plupart des contrats sont demeurés stables ou ont connus des baisses au niveau des primes. **(C'est un résultat étonnant, faudrait commenter la -dessus ???)**

La variable **vehicl_region** représente une région de l'union européenne. Il y a 14 régions et elles sont numérotés mais nous savons pas à quel emplacement géographique cela correspond. La figure 1 suivante permet de visualiser la disperssion des contrats dans les diverse régions. Il est ainsi possible d'observer que certaines régions sont prédominantes comme la région 4, 7 et 8 par exemple.

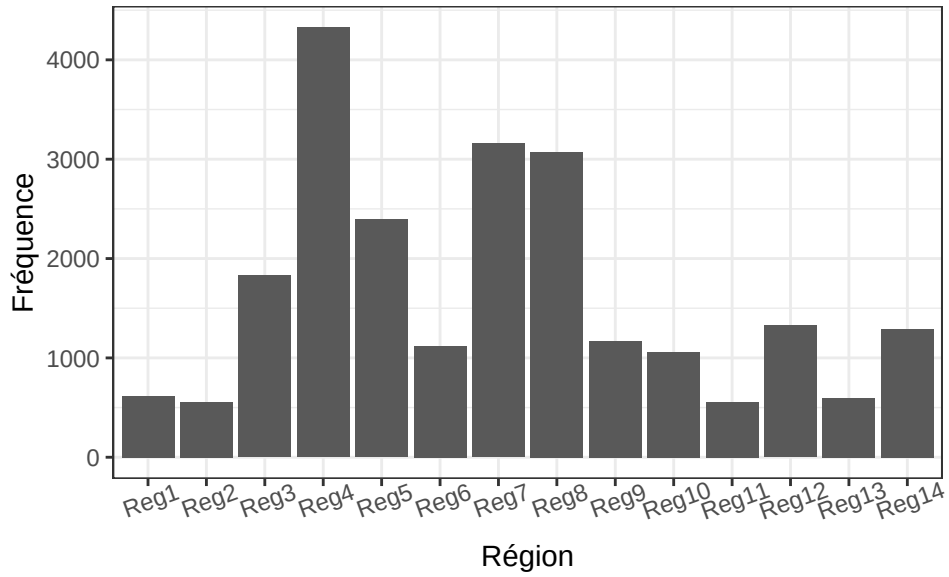


Figure 1: Distribution des régions des détenteurs de police, représenté par la variable **polholder_age**

Variables catégorielles ordinales

La variable catégorielle ordinale **prem_freqperyear** représente la fréquence par année à laquelle la prime est payable. Les fréquence possible est mensuelle, trimestrielle, semestrielle ou annuelle.

On voit qu'un peu moins de la moitié des clients paient la prime en un seul versement, environ un quart des clients paient trimestriellement, et le dernier quart est partagé par la prime payable semestriellement et mensuellement.

La variable **vehicl_powerkw** représente la puissance du moteur de la voiture conduit exprimé en chevaux moteurs. Initialement, cette variable contient 11 niveaux possibles. Cependant, en regardant les niveaux, nous avons constaté que les données pour cette variable n'ont pas été collectées uniformément puisque certaines catégories sont comprises dans une autre catégorie. Un des niveaux correspond aux véhicules d'une puissance se situant entre 125 et 300. Il y a aussi des groupes pour lequels la puissance se trouve entre l'interval du groupe présenté précédement, certains avec très peu d'observations d'autres avec un peu plus d'observations. Or, puisqu'on ne sait pas la puissance des voitures se trouvant dans le groupe de puissance 125-300 et que celui-ci comprend un grand nombre d'observation nous avons opté pour l'option d'ajouter les groupes pour lesquels leur puissance se situait entre 125 et 300 chevaux. Pour facilité la représentation du traitement effectué sur la variable **vehicl_powerkw** voici un tableau de fréquence avant traitement et tableau après traitement.

Variables numériques discrètes

La variable **polholder_age** est une variable numérique discrète représentant l'âge du propriétaire de la police d'assurance. La Figure 3 représente la distribution de âges des assurés.

L'âge minimal parmis les assuré est de 19 ans et l'âge maximal est de 85 ans. On constate qu'il y a une forte proportion d'assuré entre 30 et 45 ans. Il pourra être pertinent d'analyser si les assurés de plus de 45 ans sont présent en moins grand nombre dû au fait que les primes sont trop élevé et font d'avantage d'appel pour comparer les primes ailleurs ce qui les mènent vers la résignation.

Le nombre d'année sans résignation de la police d'assurance depuis la première année assurée est représenté par la variable umérique discrète **policy_age**. Avec l'aide de la Figure 4 on constate une forte décroissance du nombre d'assurée pour le nombre d'années depuis l'entré en vigueur pour les 3 premières années pour

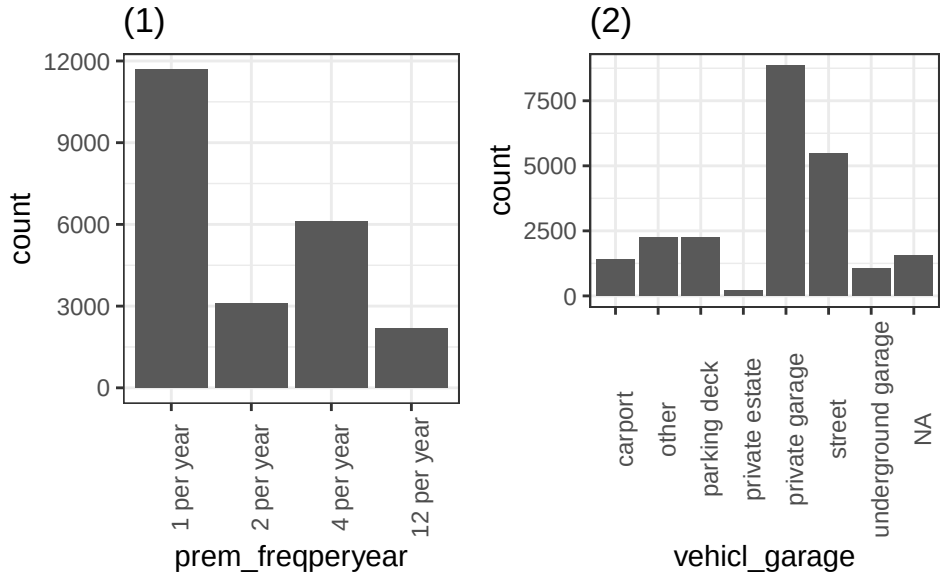


Figure 2: todo

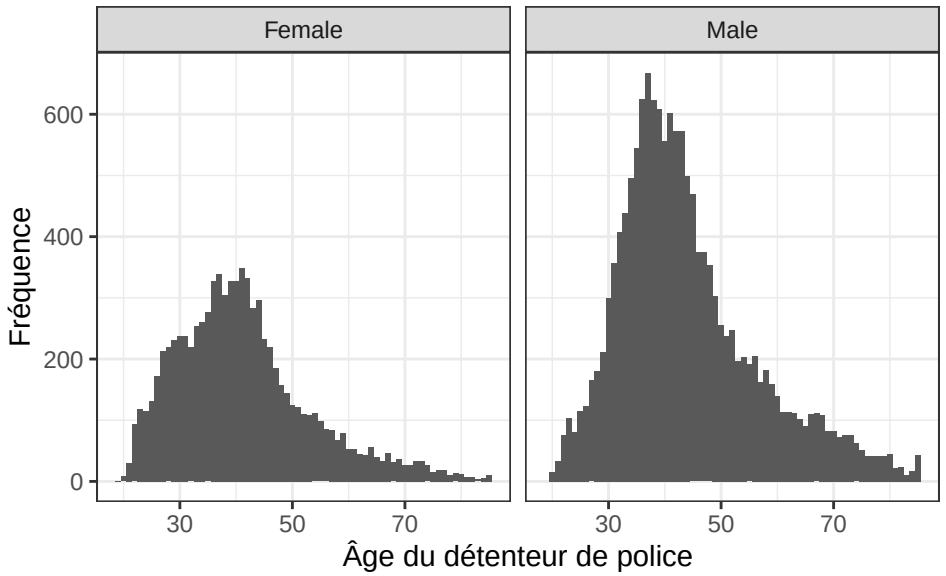


Figure 3: Distribution de l'âge des détenteurs de polices dans la base de données, représenter par la variable **polholder_age**

Table 6:

Puissance (kW)	Nombre d'observation
100	5116
125-300	1720
150	580
175	206
200	32
225	77
25-50	4968
250	16
275	4
300	2
75	10339

Table 7:

Puissance (kW)	Nombre d'observation
25-50	4968
75	10339
100	5116
125-300	2637

ensuite ce stabilisé par la suite.

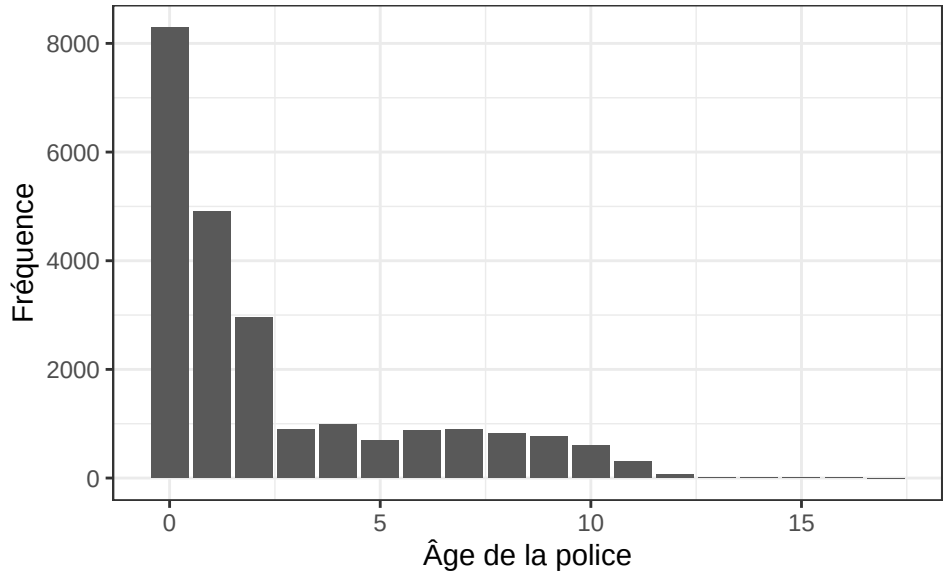


Figure 4: Distribution de l'âge pour laquelle une police est en vigeur, représenter par la variable **policy_age**

En ce qui concerne la variable discrète **policy_nbcontract** représentant le nombre de contrat, ou risque, que l'assuré possède chez l'assureur. L'histogramme illustré à la Figure 5 fait resortir le fait qu'il y a une forte concentration d'assuré pour lesquels le nombre de contrat qu'ils ont chez l'assureur est inférieur à 5. On peut aussi voir que certains assurés ont jusqu'à 15 contrats.

Il y a plusieurs variables numériques continues relatives à la prime. La variable **prem_final** représente le montant de la prime proposé pour le renouvellement par l'assureur alors que **prem_last** représente le montant payé lors du dernier renouvellement. La variable prem_market est la prime qui serait chargée selon le marché. La variable prem_pure est la prime qui représente les coûts espérés pour un assuré. Le **??** montre les distributions de chaque des primes. Par contre, une prime seul peut difficilement expliquer pourquoi un assuré voudrais résigné car si l'assuré mérite réellement sa prime, il n'aurait pas intérêt à aller chez un concurant. Par contre, si lors de sont renouvellement, il voit sont montant d'assurance augmenter d'un grand poucentage, il serat tenté d'aller voir ayeur. C'est pourquoi la variable **prem_index** à été crée et ajouté à notre jeu de données. celle-ci représente le pourcentage d'augmentation de la prime, sois la prime final divisé

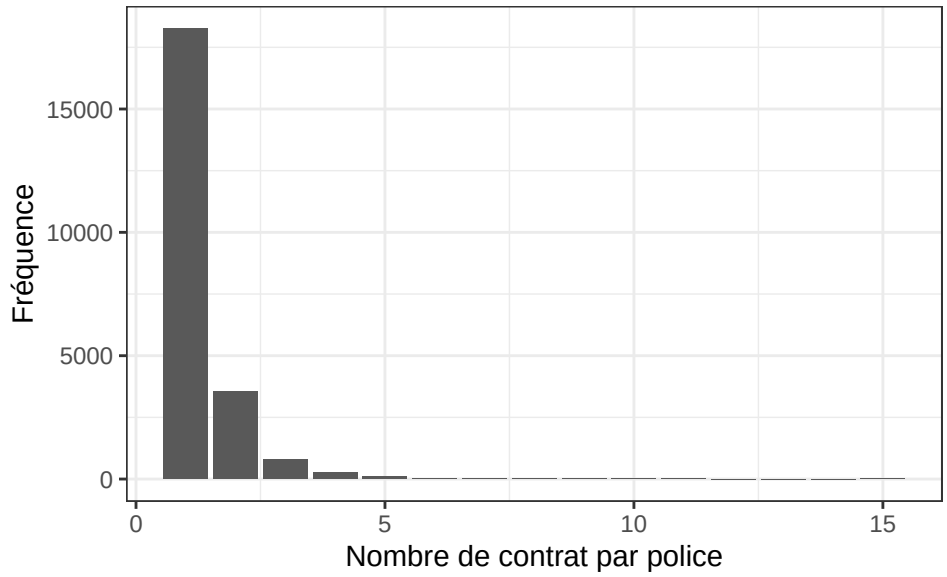


Figure 5: Distribution du nombre de contrats par police, représenter par la variable **policy_nbcontract**

par la prime du dernier terme.

Table 8:

Prime ($)	Minimum	Médiane	Moyenne	Maximum	Écart-type
Final	46.55	312.25	374.12	2948.05	212.9
Last	46.56	311	380.51	3362.07	227.94
Market	50.11	316.83	373.53	2416.84	201.92
Pure	45.55	301.45	355.88	2716.08	197.14
Index (%)	-0.58	0	0	2.27	0.1

Les deux prochaines variables sont en lien avec l'aĝe du véhicule, il s'agit de variables numériques discrètes. La variable vehicl_agepurchase représente l'âge du véhicule lorsque l'assuré a acheté le véhicule. La variable vehicl_age représente l'aĝe du véhicule actuellement.

Beaucoup de véhicule ont été achetés lorsqu'il était neuf (vehicl_purchase = 0). En examinant les véhicules conduits par es assurés on remarque qu'il y a peu de véhicule neuf et le nombre de véhicule est croissant en fonction de l'utilisation jusqu'à 13 ans puis décroit par la suite. On retrouve un grand nombre de véhicule assuré avec 18 ans d'usage, il est fort probable que cela correspondre aux véhicules de plus de 18 ans.

Variables numériques continues

Il est intéressant de voir comment se comporte la variable endogène en fonction des différentes variables explicatives. Bien sûr, nos variables explicatives comme décrit plus tôt ne sont pas toutes de même type il en résulte donc qu'elles n'auront pas toutes la même représentation visuelle par rapport à la variable endogène.

Pour variable continues **(voir lab_partionnement2)**

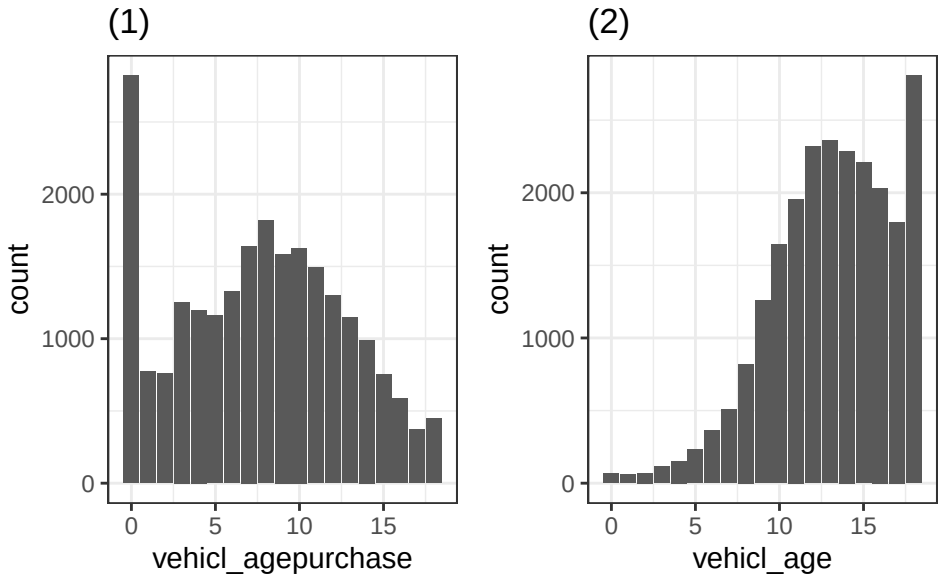


Figure 6: todo

Traitement des valeurs manquantes

La base de données contenait seulement trois variables avec des valeurs manquante. La variable indiquant la difference d'âge entre le détenteur de police et le conducteur est manquante à 0.05%, celle indiquant l'utilité du vehicule est manquante à 15.1% et la variable indiquant le type de garage où est entreposé le vehicule est manquante à 6.83%. La Figure 7 montre le patron de non réponse. On remarque que la variable *polholder_diffdriver* semble avoir un patron de non réponse monotone avec les deux autres. Par contre, puisqu'il y a seulement 12 cas, nous allons pas tenir compte de ce lien lors de l'imputation des données. Pour ce qui est des variables *policy_caruse* et *vehicl_garage*, on remarque qu'il sont parfois manquante en même temps, mais seulement pour une minorité de cas.

Premièrement, dans le but déterminer si le données manquantes sont MCAR, le test d'hypothese suivant à été effectué

$$H_0 : \text{Le données sont MCAR}$$
$$H_1 : \text{Le données ne sont pas MCAR}$$

Pour conclure que les données sont MCAR, il est nécéssaire d'accepter H_0 pour toutes les variables. Par contre, un seul refus de cette hypothèse sera néséssaire permetra de conclure l'hypothèse alternative, c'est a dire que les données ne sont pas MCAR. Pour effectuer le test avec une variable catégorielle, il sera nécéssaire d'utilisé une statistique khi-carré alors que pour une variable numérique, une statistique student sera utilisée.

```
## Warning in chisq.test(data[[var.data]], is.na(data[[na_var.data]]), correct =
## FALSE): Chi-squared approximation may be incorrect

## Warning in chisq.test(data[[var.data]], is.na(data[[na_var.data]]), correct =
## FALSE): Chi-squared approximation may be incorrect

## Warning in chisq.test(data[[var.data]], is.na(data[[na_var.data]]), correct =
## FALSE): Chi-squared approximation may be incorrect

## Warning in chisq.test(data[[var.data]], is.na(data[[na_var.data]]), correct =
## FALSE): Chi-squared approximation may be incorrect
```

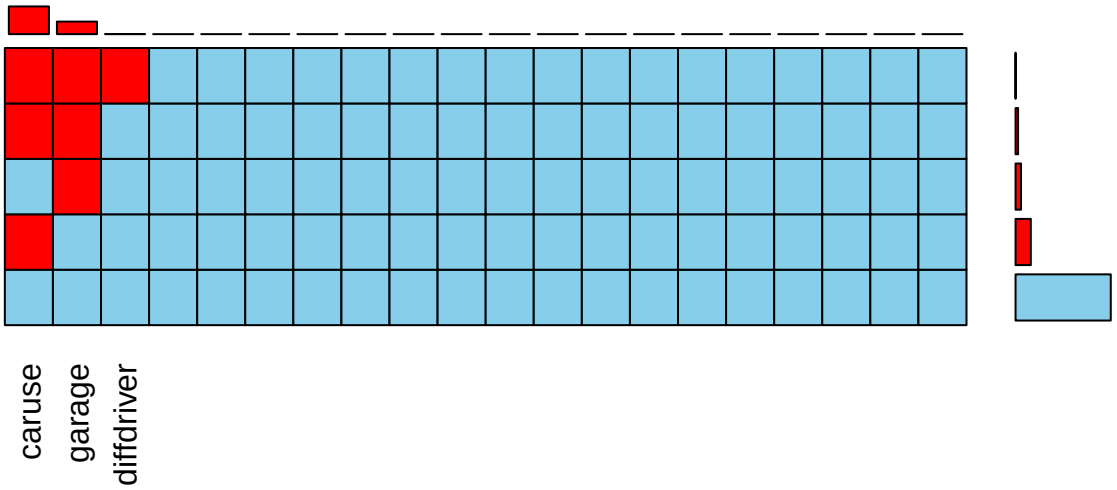


Figure 7: todo

```
## Warning in chisq.test(data[[var.data]], is.na(data[[na_var.data]]), correct =
## FALSE): Chi-squared approximation may be incorrect

## Warning in chisq.test(data[[var.data]], is.na(data[[na_var.data]]), correct =
## FALSE): Chi-squared approximation may be incorrect

## Warning in chisq.test(data[[var.data]], is.na(data[[na_var.data]]), correct =
## FALSE): Chi-squared approximation may be incorrect

## Warning in chisq.test(data[[var.data]], is.na(data[[na_var.data]]), correct =
## FALSE): Chi-squared approximation may be incorrect

## Warning in chisq.test(data[[var.data]], is.na(data[[na_var.data]]), correct =
## FALSE): Chi-squared approximation may be incorrect
```

En ce qui concerne les variables *vehicl_garage* et *policy_caruse*, plusieurs statistiques observé permette de rejeté l'hypothèse null en faveur de l'hypothèse alternative à un niveau significatif de 0.001. Par contre, dans le cas de *polholder_diffdriver*, seulement la variables *polholder_job* permet de rejeter H_0, c'est à dire que les données manquantes sont completement aléatoire.

Il est à noter qu'il n'est pas possible de vérifier avec certitude si les données sont MAR ou NMAR. Cela est dû au fait que puisque les données proviennent d'un compagnie inconnue, nous n'avons pas d'information sur la méthode de récolte de données et il nous est impossible de trouver des patrons qui pourraient provoquer des données de type NMAR. En conséquence, nous considérerons que nos données sont MAR. De ce sens, en effectuant des tests khi-carré pour la variable *polholder_diffdriver*, il a été remarqué que l'information sur la différence entre le détenteur de police et le conducteur nous indique que les variables sont toujours manquantes dans le cas ou le travail du détenteur de la police est dans le domaine de la médecine. Ceci renforce l'idée que le patrons de non réponse pour cette variable dépend des variables observés dans le jeux de données.

Pour l'imputation des données, la méthode d'imputation multiples à été choisie. Pour des restrictions de temps de calcul, cinq itérations de régression stochastique ont été fait. Pour la variable *policy_caruse*, une régression logistique a été effectué puisque la variable catégorielle comporte deux niveaux. Pour les

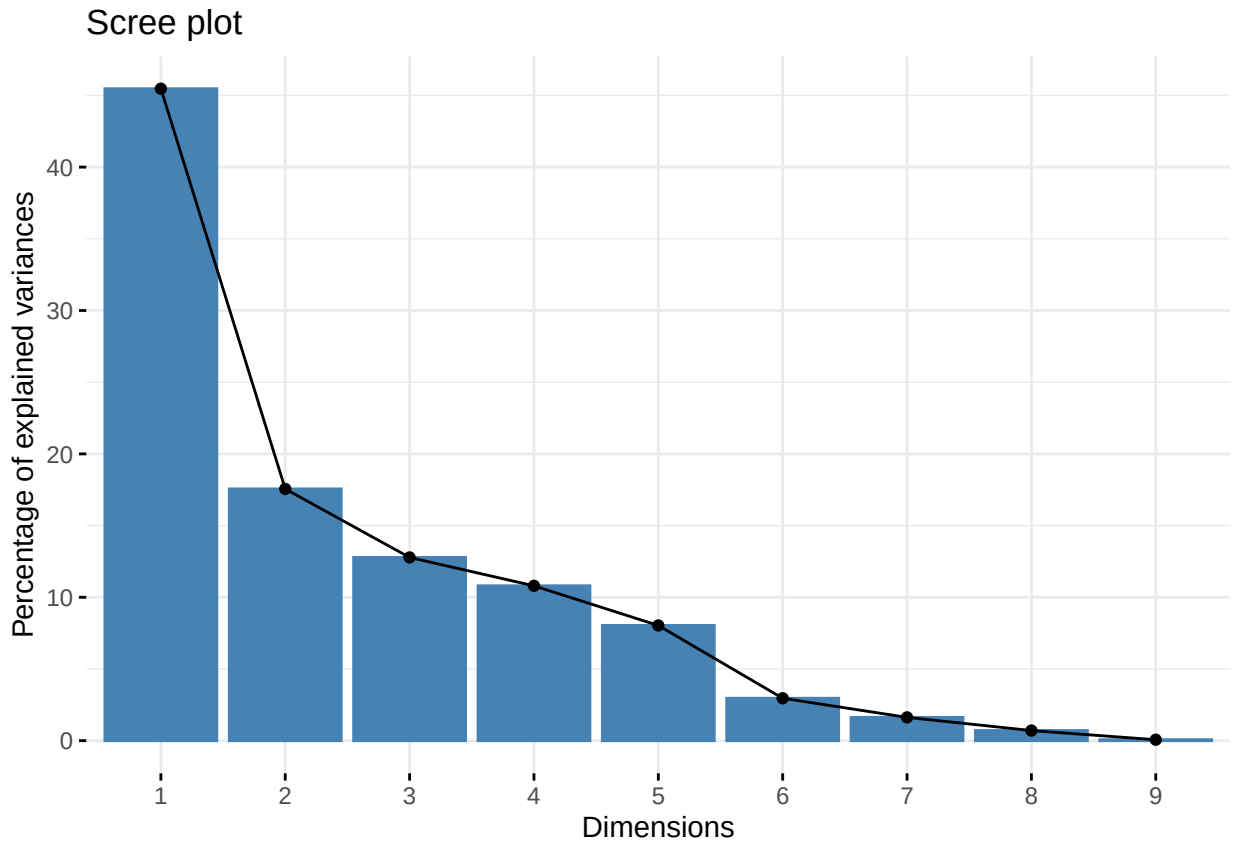
variables *vehicl_garage* et *polholder_diffdriver*, qui sont des variables catégorielles non-ordonées, une régression polynomiale a été utilisée.

Analyse en composantes principales

Étant donné que notre jeu de données contient `R nrow(Donnees_tempo)` observations, il peut être utile de visualiser les données à l'aide de l'analyse en composantes principales, appelé ACP. En effet, ce type d'analyse permet de mieux visualiser un jeu de données lorsque celui-ci est de grande dimension. Il sera ainsi possible de voir quelles variables explicatives sont plus intéressantes par leur impact sur la variance des composantes principales. Il est à noter qu'en général, on garde assez de composantes pour représenter entre 80 et 90 % de la variance totale.

Pour que cette méthode de visualisation puisse être utilisée, il sera nécessaire de prendre seulement les variables explicatives numériques de ce jeu de donnée. Les variables catégorielles ne seront pas analysées dans cette section car même en les transformant en variables numériques, elles ne seront pas représentative des valeurs leur qui leur aurait été attribuée en faisant la modification de type.

On doit ensuite choisir le nombre de composantes principales. Cette étape peut être complétée en ayant deja un pourcentage de variance expliquée en tête et en choisissant le nombre de composantes à partir des valeurs propres ou en analysant directement le diagramme d'ébouli. Dans ce cas, la méthode du coude ne sera pas utilisée, on priviligie d'avantage le choix selon le premier plateau observée. Le nombre de composantes choisit seront celle ne faisant pas partie du premier plateau observée.



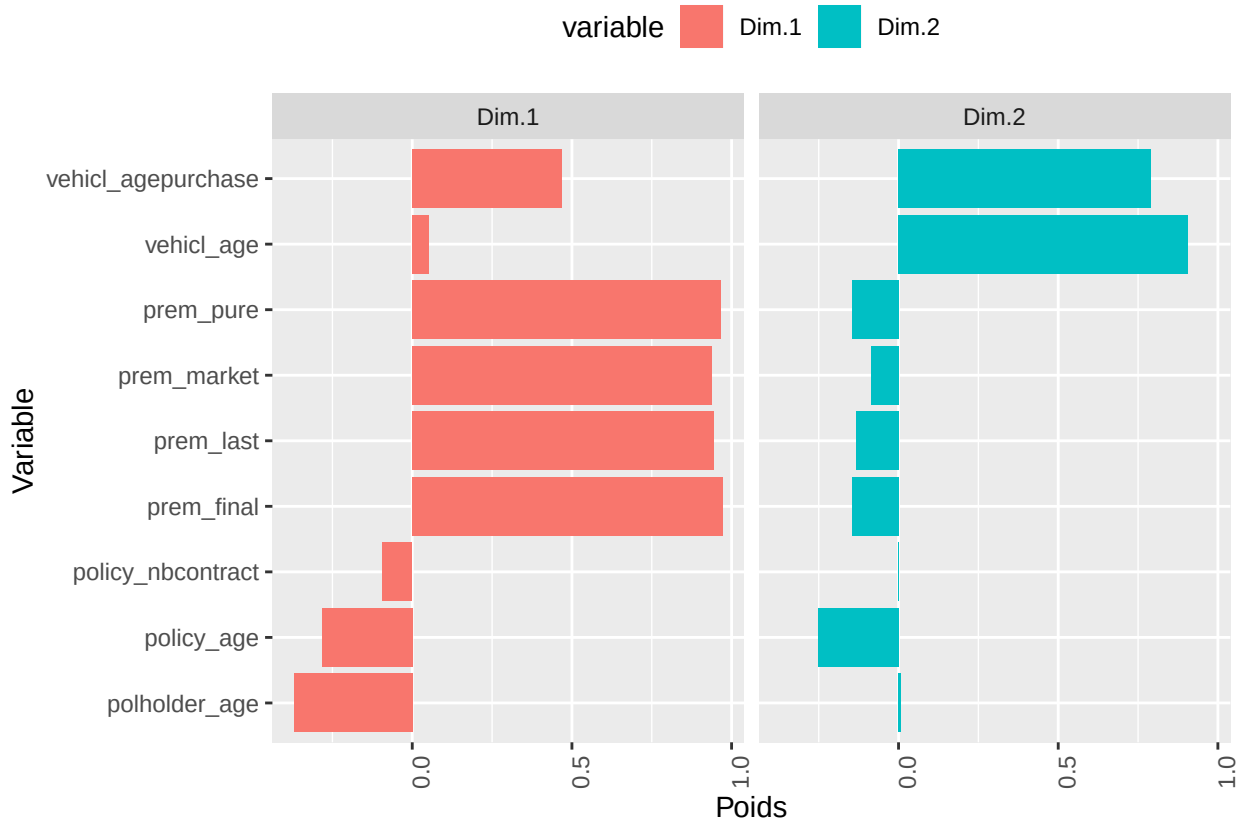
Selon le diagramme d'éboulis, il sera nécessaire de conserver 2 composantes principales et on observe, à l'aide des valeurs propres de la matrice de corélation, que 2 composantes principales permettent d'expliquer 63% de la variance totale.

À l'aide du graphique ACP des variables, on peut voir la gravité des contributions pour chacune des variables sur chaque composantes principales retenu. Ainsi, on peut observer que pour la première composante principale, un score élevé indique un contrat ayant une prime élevée, que ce soit la prime du marché, la prime pure, la prime finale ou la prime chargée lors du dernier renouvellement. Par contre, un assuré âgé

qui renouvelle depuis plusieurs année aura un score plus faible qu'un assuré en bas âge ayant une police d'assurance récente. Un score élevé représente donc un assuré en bas âge ayant une police récente et une prime élevé tandis qu'un score faible représente une personne plus âgée avec une faible prime d'assurance.

La deuxième composante principale représente, quant à elle, l'âge du véhicule assuré. Un score élevé est associé à des véhicules de moindres valeurs mais risquant d'avantage un bris de veillesse. Plus les polices d'assurance sont récentes et plus le score en sera augmenté. Ainsi, les polices d'assurances récente ayant des véhicules de l'année représenteront les scores les plus faibles pour cette composante.

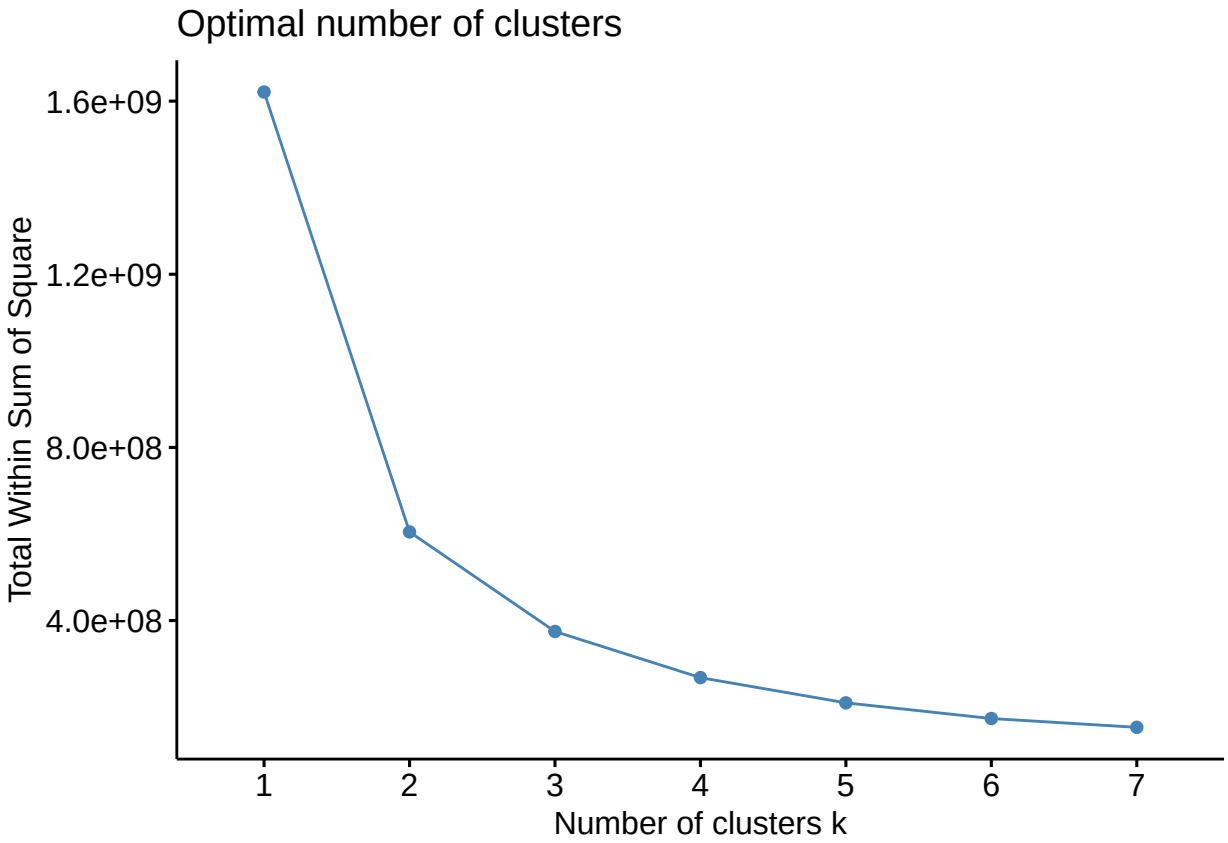
En illustrant les contributions des variables pour les deux premières composantes principales, il est plus facile de visualiser les conclusions mentionnées précédemment.



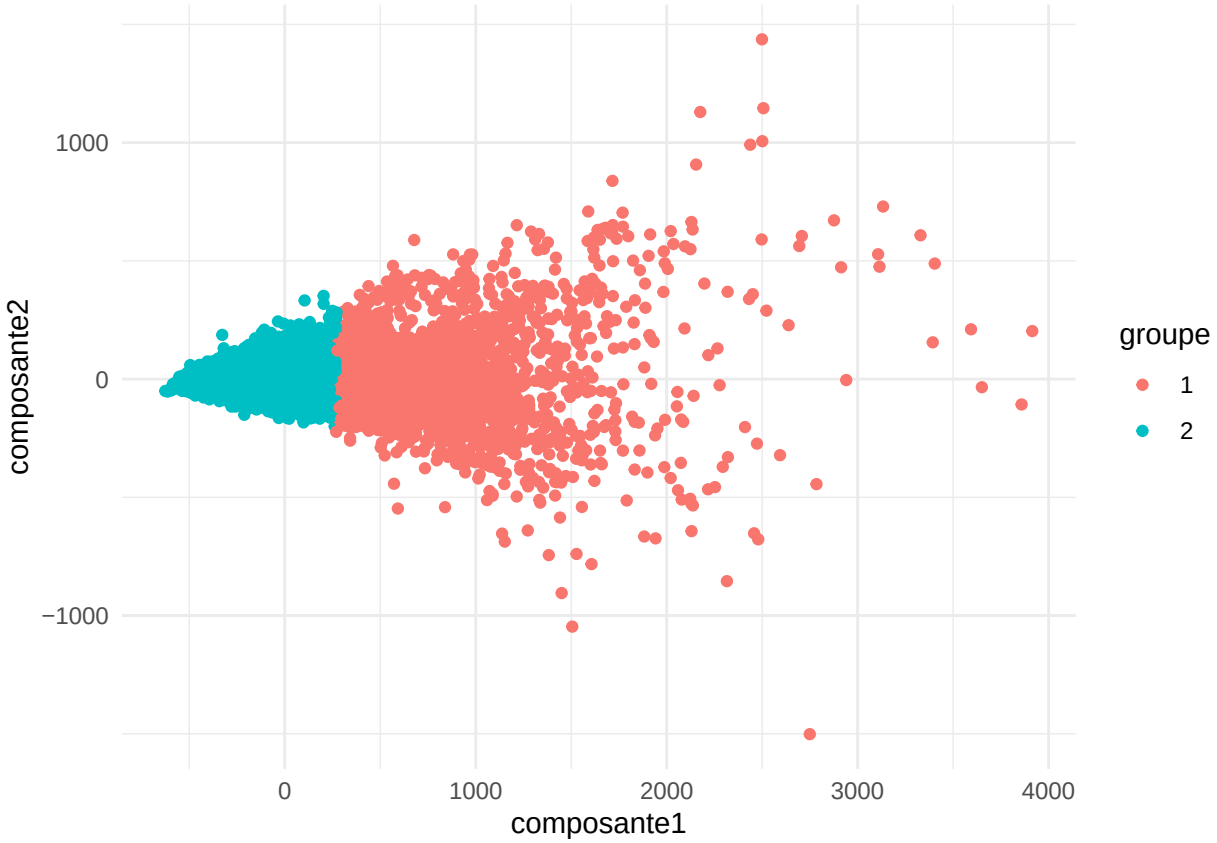
Partitionnement en k moyennes

Le partitionnement en k moyennes est utilisé pour classifier les observations en k groupes distincts. La valeur de k est une valeur qu'on transmet pour indiquer le nombre de partitions désirées. Chaque observation sera ensuite assigné à un seul groupe. L'algorithme utilisée pour ce type de partitionnement a pour objectif de minimiser la variance intra-groupe.

Le choix du nombre de groupe peut être choisit à l'aide de la méthode du coude. Ainsi en ce réferant au graphique suivant, on devrait faire le partitionnement sur 2 groupes distincts. On s'arrête la valeur de k qui ce situe dans le "pli de coude", soit juste avant le dernier plateau du diagramme d'éboulis. Il est à noter que le nombre d'observations a été réduit pour pouvoir faire le diagramme d'éboulis. Notre jeu de données étant trop volumineux, ce qui engendrait des erreurs d'exécution. L'échantillion utilisé a été extrait aléatoirement et sans remise pour avoir une représentation adéquate et la moins biaisé possible.



En ayant en tête le nombre de groupe nécessaire pour la classification, on effectue le partitionnement et on obtient le graphique suivant :



De ce graphique, on peut conclure que le partionnement c'est fait sur la première composante principale. Les assurés représentant moins de risque ce retrouve dans le groupe 2 tandis que les assurés plus risqués ce retrouve dans le groupe 1. Ainsi, le montant des primes typiques seraient d'environ 300$ ou moins pour le deuxième groupe et de plus de 300$

Conclusion

Comme mentionné précedemment, le jeu de données analysées dans ce travail pratique provient du paquetage "CASdatasets". Nous avons choisi ce jeu de données dans le but de modéliser le statut de renouvellement de polices d'assurance pour une compagnie et une année d'observation inconnu. Les variables explicatives touchent les caractéristiques liés aux primes payés, à l'assuré visé par la police d'assurance et au véhicule assuré.

PARLER DE L'ANALYSE EXPLORATOIRE ET DU PRÉTRAITEMENT

Puisque la variable réponse **lapse** est une variable catégorielle pour laquelle deux valeurs sont possibles , soit renouvellement ou résignation, il sera intéressant pour la suite de modéliser la probabilité qu'un assuré renouvelle ou résigne pour la prochaine année. Un modèle linéaire avec régression logistique sera ainsi a élaborer. La prédiction de la régression corresponderait, dans ce cas, à la probabilité désirée. Dans le cas ou on s'intéressait d'avantage à une prédiction de cette variable réponse, il sera possible d'utiliser un modèle linéaire avec régression logistique ou bien un autre modèle de classification supervisé.

Bibliographie

Annexe

Description du jeu de données soumis sur le forum :

Notre jeu de données représente le statut de renouvellement pour 23 060 polices d'assurance basées sur un an d'observation. Les données receuillies proviennent d'une compagnie d'assurance inconnue dont l'année d'observation est également inconnue.