PROJET D'EXPLORATION DES DONNÉES EN ASSURANCE

Auteur:

Ibrahima LY

BIG DATA, DATA SCIENCE ET ANALYSE DES RISQUES:

Résumé

Le processus d'ETL (Extraction, Transformation, Chargement) est un aspect fondamental dans tout workflow d'analyse de données, pouvant constituer jusqu'à 50% du temps investi par un analyste. Cette démarche commence par l'acquisition et l'exploration du dataset, incluant des procédures telles que l'ingestion des données, la création d'aperçus visuels initiaux et la détection d'anomalies, telles que des valeurs manquantes ou aberrantes. Ensuite, le processus avance vers le data viz, une étape essentielle pour faciliter la compréhension et l'interprétation des données. En fusionnant ces deux aspects, nous sommes en mesure d'extrait des insights précieux et pertinents à partir de notre analyse.

Table des matières

1 Introduction

Dans le secteur de l'assurance, qui est en constante évolution, l'analyse de données joue un rôle central. Elle permet d'éclairer les décisions, d'améliorer l'efficacité opérationnelle et de créer de nouvelles opportunités de croissance. Dans ce projet, nous aborderons l'analyse de données dans le domaine de l'assurance en trois étapes majeures :

La première partie se focalise sur l'analyse et le prétraitement des données. Cette étape cruciale permet de comprendre les données en notre possession et de les préparer pour les analyses ultérieures. L'importation de données, la gestion des valeurs manquantes, la détection d'anomalies et l'identification de valeurs aberrantes sont quelques-unes des tâches importantes réalisées à ce stade. C'est également lors de cette phase que nous effectuons les joitures entre dataFrames et les premières visualisations pour mieux appréhender la nature et la structure des données.

La deuxième partie est consacrée à l'analyse économétrique. En utilisant des méthodes statistiques, nous cherchons à découvrir les relations entre les différentes variables de notre jeu de données. Cela nous permet de comprendre comment ces variables interagissent et peuvent influencer les résultats dans le contexte de l'assurance. Les modèles économétriques nous permettent de faire des prédictions précises et de formuler des stratégies efficaces.

Enfin, la troisième partie se concentre sur l'application du machine learning. Ces techniques modernes d'analyse de données nous permettent de créer des modèles prédictifs puissants qui peuvent apprendre à partir des données et s'améliorer avec le temps. Que ce soit pour la prédiction des risques, l'optimisation des prix ou la détection de fraudes, le machine learning offre des outils précieux pour naviguer dans le paysage complexe de l'assurance.

En combinant ces trois phases, prétraitement et analyse des données, analyse économétrique et machine learning, nous visons à fournir une vision complète et approfondie de l'analyse de données dans le domaine de l'assurance.

Définition et Description des Variables de l'Étude :

IDpol : identifiant de la police d'assurance
Yearx : année de souscription de la police d'assurance
DrivAge : âge du conducteur (en années)
DrivGender : sexe du conducteur
BonusMalus : coefficient de bonus-malus
LicenceNb : nombre de permis de conduire associés à la police d'assurance
PayFreq : fréquence de paiement de la prime d'assurance
VehAge : âge du véhicule (en années)
VehClass : classe du véhicule
VehPower : puissance du véhicule
VehGas : type de carburant du véhicule
VehUsage : utilisation du véhicule
Garage : type de garage où est garé le véhicule
Area : zone géographique où est situé le véhicule
Region : région géographique où est situé le véhicule
Marketing : stratégie marketing associée à la police d'assurance
PremWindscreen : prime pour le pare-brise
PremDamAll : prime pour les dommages
PremFire : prime pour les incendies
PremAcc1 : prime pour les accidents corporels
PremAcc2 : prime pour les accidents matériels
PremLegal : prime pour la protection juridique
PremTPLM : prime pour la responsabilité civile (matérielle)
PremTPLV : prime pour la responsabilité civile (corporelle)
PremServ : prime pour les services d'assistance
PremTheft : prime pour le vol
PremTot : prime totale de l'assurance
Year : année de survenance du sinistre
Damage : montant du sinistre pour les dommages matériels
Fire : montant du sinistre pour les incendies
Other : montant du sinistre pour d'autres types de dommages
Theft : montant du sinistre pour les vols
TPL : montant du sinistre pour la responsabilité civile (corporelle)
Windscreen : montant du sinistre pour le pare-brise
OccurDate : date de survenance du sinistre
Payment : montant total des paiements effectués pour le sinistre
IDclaim : identifiant du sinistre
Guarantee : type de garantie de la police d'assurance

2 Chapitre I : Prétraitement et Manipulation des Données

2.1 Élimination des Doublons :

Dans la colonne 'IDpol', il existe 20 255 doublons, indiquant que chaque 'IDpol' peut correspondre à plusieurs entrées. Cela pourrait être dû à plusieurs enregistrements pour une même police d'assurance sur différentes années ou différents incidents.

Lorsqu'on prend en compte à la fois 'IDpol' et 'Year', le nombre de doublons augmente, suggérant que certains enregistrements sont identiques tant pour 'IDpol' que pour 'Year'. Ceci pourrait être le signe d'erreurs de saisie ou de véritables doublons dans les données.

Il n'y a que 2 doublons purs (sans compter 'Year') dans la base de données, ce qui pourrait également indiquer des erreurs de saisie.

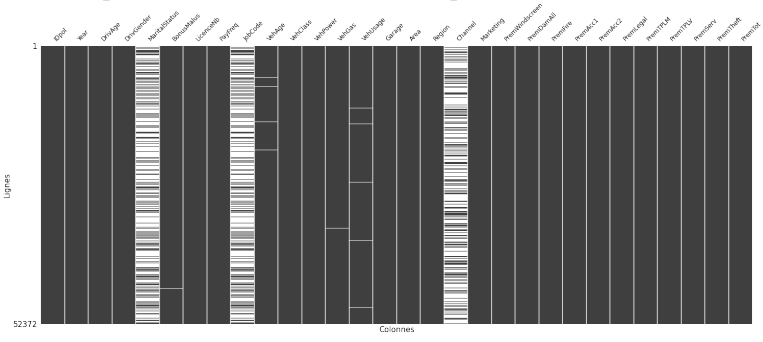
Cependant, il n'y a aucun doublon pur dans l'ensemble de la base de données, signifiant qu'il n'y a pas deux lignes entièrement identiques.

Le nombre de doublons 'IDpol' est confirmé par un autre test, soulignant la récurrence de 'IDpol' dans les données.

Ces doublons pourraient être dus à plusieurs facteurs, comme des erreurs de saisie, des mises à jour des enregistrements, ou la nature même des données (par exemple, une police d'assurance pouvant concerner plusieurs incidents). Il est essentiel d'analyser et de comprendre la source de ces doublons avant de décider de la manière de les traiter.

2.2 Traitement des Valeurs Manquantes :

Figure 1 : Répartition des valeurs manquantes dans le DataFrame



Pourcentage de Valeurs Manquantes (NaN) par Colonnes :

Channel : 68.65%
MaritalStatus : 67.31%
JobCode : 67.31%
VehAge : 0.93%
BonusMalus : 0.34%
VehGas : 0.31%
VehUsage : 0.24%
VehPower : 0.02%

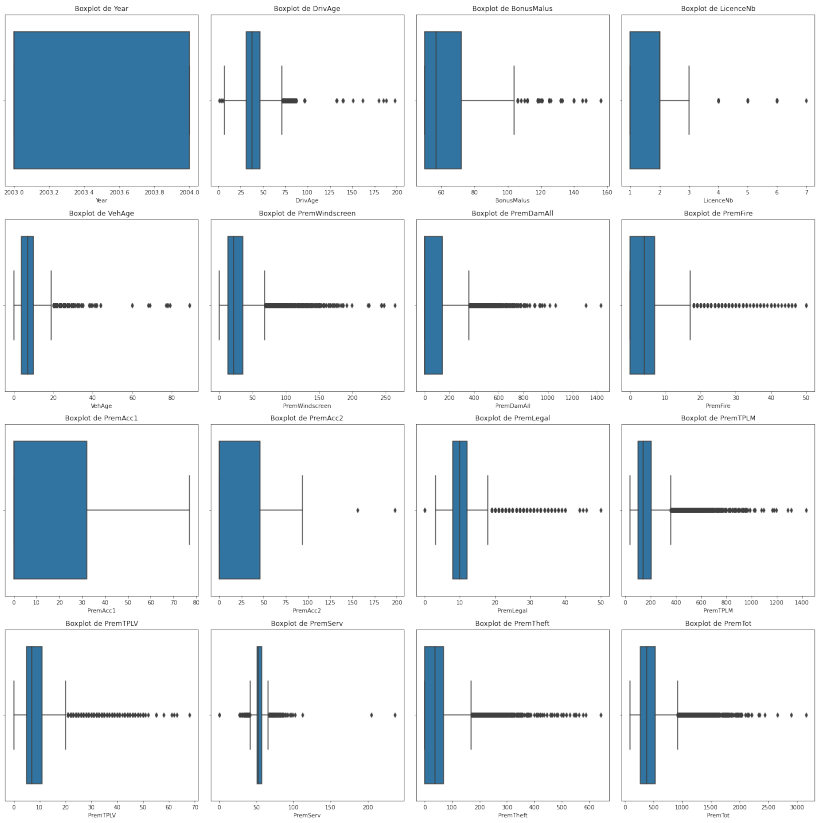
MaritalStatus et JobCode ont tous deux 67.31% de données manquantes. C'est une proportion importante, ce qui peut rendre ces colonnes moins utiles pour l'analyse. Il serait important de comprendre pourquoi ces valeurs sont manquantes. Cela pourrait être dû à une erreur de collecte de données, à des données non pertinentes pour certains individus (par exemple, le code de travail pour une personne sans emploi), ou à des réponses manquantes pour d'autres raisons. Pour résoudre ce problème, on pourrait envisager des méthodes pour les imputer, comme l'utilisation de la médiane ou de la moyenne, ou un modèle de prédiction, ou on pourrait envisager de supprimer ces colonnes si elles ne sont pas essentielles.

Channel présente 68.65% de données manquantes, ce qui est encore plus élevé. Les mêmes considérations s'appliquent ici. Si cette information est cruciale pour notre analyse, il faudrait chercher à comprendre pourquoi ces données manquent et comment on peut les remplacer ou les traiter.

Les colonnes BonusMalus, VehAge, VehPower, VehGas et VehUsage ont toutes moins de 1% de données manquantes. C'est une quantité relativement faible de données manquantes, et cela pourrait être dû à des erreurs aléatoires ou à des omissions lors de la collecte de données. Dans ce cas, il serait généralement acceptable d'imputer ces valeurs manquantes en utilisant une méthode appropriée, ou de supprimer les lignes comportant des valeurs manquantes si elles ne représentent pas une proportion importante de nos données.

2.3 Correction des Données Incohérentes :

Figure 2 : Statistiques descriptives des variables



En examinant les statistiques descriptives, on peut observer les éléments suivants :

DrivAge : L'âge maximal du conducteur est de 198 ans, ce qui semble improbable compte tenu de l'espérance de vie humaine. De plus, l'âge minimum est de 1 an, ce qui est également impossible pour un conducteur.

BonusMalus : La valeur maximale est de 156, ce qui est significativement supérieure à la médiane (57) et pourrait être considérée comme une valeur aberrante.

VehAge : L'âge maximal du véhicule est de 89 ans, ce qui est possible mais rare, donc cela pourrait être considéré comme une valeur aberrante.

PremDamAll, PremAcc1, PremAcc2, PremTheft : Les valeurs maximales de ces colonnes sont nettement plus élevées que leurs 75e percentiles, ce qui suggère la présence de valeurs aberrantes.

PremTot : La valeur maximale est de 3163.3, ce qui est considérablement supérieur à la médiane (381.55) et pourrait être considéré comme une valeur aberrante.

3 Chapitre II : Analyse Économétrique

Dans cette partie, nous allons formuler des hypothèses théoriques que nous vérifierons ensuite grâce à une analyse des données univariées et multivariées.

Hypothèses théoriques :

Première hypothèse : Plus une personne est âgée, plus elle a de chances d'avoir un bonus plutôt qu'un malus, car cette personne est expérimentée. L'expérience accumulée au fil des années permet aux conducteurs d'adopter des comportements plus prudents, ce qui réduit la probabilité d'être impliqué dans des accidents. En conclusion, une personne expérimentée est plus susceptible de bénéficier de bonus sur sa prime d'assurance plutôt que de malus.

Deuxième hypothèse : Plus une voiture est ancienne, plus elle a de chances d'être soumise à des tentatives de vol. Les voitures plus anciennes attirent les voleurs, notamment grâce à leur système de sécurité moins sophistiqué que celui des voitures récentes, mais aussi parce que plus une voiture est ancienne, plus elle a de chances d'être considérée comme une voiture de collection.

Dernière hypothèse : Les régions à forte densité de population, comme l'Île-de-France, ont une probabilité plus élevée d'avoir des conducteurs avec des malus. Les régions avec une grande concentration de véhicules et de circulation sont exposées à un risque d'accident plus élevé.

3.1 Analyse Descriptive Univariée :

La variable **YearY** semble présenter des valeurs aberrantes puisque sa moyenne est de 4004, ce qui est très éloigné de l'année actuelle (2023). Cela pourrait résulter d'erreurs dans les données ou d'un codage non standard des années.

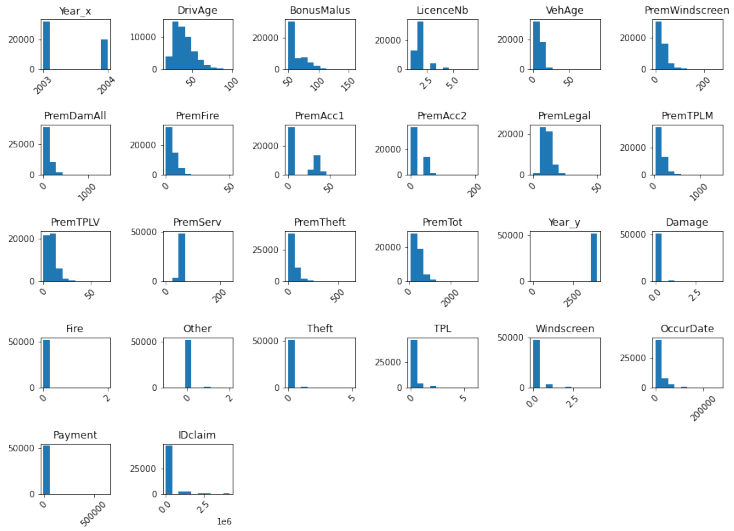
Dans cette section, nous allons nous concentrer sur la variable numérique **Bonus/Malus**. Un chiffre inférieur à 100 est considéré comme un bonus, tandis qu'un chiffre supérieur à 100 est considéré comme un malus. Il est important de noter que cette variable varie entre 50 et 350. D'après le tableau ci-dessus, nous observons que la valeur minimale de la variable BonusMalus est de 50 et la valeur maximale est de 156.

Nous continuons notre analyse des données univariées avec des histogrammes. Ces derniers nous permettent de comparer les distributions, de vérifier l'exis-

tence de valeurs aberrantes, ou simplement de déterminer si les valeurs numériques non nulles de notre base de données sont normalement distribuées.

3.1.1 Distribution des variables numériques

Figure 3 : Distribution des variables numériques



Nous constatons que la distribution est normale pour certaines variables, ce qui signifie qu'il y a autant de valeurs au-dessus de la moyenne que de valeurs en dessous.

Certaines variables présentent une distribution asymétrique. Elles sont décalées vers la droite, ce qui signifie que la majorité des valeurs sont à gauche de la moyenne, avec quelques valeurs élevées à droite. Certaines variables présentent une distribution uniforme, ce qui signifie qu'elles ont approximativement la même fréquence.

3.1.2 Statistiques descriptives des variables catégorielles :

IDpol : Il semble y avoir de nombreuses valeurs uniques, ce qui suggère que cette variable pourrait être un identifiant unique pour chaque contrat d'assurance.

DrivGender : Il y a plus d'hommes (34424) que de femmes (17938) dans cet ensemble de données.

PayFreq : La plupart des clients paient semestriellement (29203), suivis de ceux qui paient annuellement (17717).

VehClass : La majorité des véhicules dans cet ensemble de données sont classés comme "Cheapest" (18040), ce qui pourrait suggérer que la plupart des clients possèdent des véhicules moins coûteux.

VehPower : Il semble y avoir une distribution relativement équilibrée entre différentes catégories de puissance de véhicules, la catégorie P10 étant la plus courante (9218).

VehGas : La plupart des véhicules utilisent de l'essence régulière (31640), suivis par ceux qui utilisent le diesel (19643).

VehUsage : La plupart des véhicules sont utilisés pour des déplacements privés et pour se rendre au travail (50841), avec une petite proportion utilisée à des fins professionnelles.

Garage : La majorité des véhicules sont garés dans un "Closed zbox" (26531).

Area : La région la plus courante est A5 (15234), suivie par A3 (12797).

Il semble que les données du jeu de données soient bien distribuées.

Cependant, il y a quelques points à considérer sur certaines variables :

La variable **IDpol** semble être un identifiant unique pour chaque contrat d'assurance. Elle présente de nombreuses valeurs uniques et pourrait ne pas être utile pour une tâche de modélisation prédictive, car elle ne contient pas d'informations pertinentes sur les autres variables ou la variable cible.

La variable **Payment** contient également des valeurs négatives, qui pourraient ne pas avoir de sens dans le contexte des paiements d'assurance. Il serait nécessaire de comprendre pourquoi il y a des valeurs négatives et de traiter ces valeurs en conséquence.

Other : Cette variable contient également une valeur négative. Encore une fois, il serait utile de comprendre pourquoi.

VehUsage : Cette variable semble déséquilibrée, avec la majorité des voitures utilisées pour des déplacements "Privés + trajets pour se rendre au travail". Cela pourrait potentiellement biaiser le modèle si cette variable est utilisée pour la modélisation.

3.2 Analyse Descriptive Bivariée :

Figure 4 : pairplot de distributions :

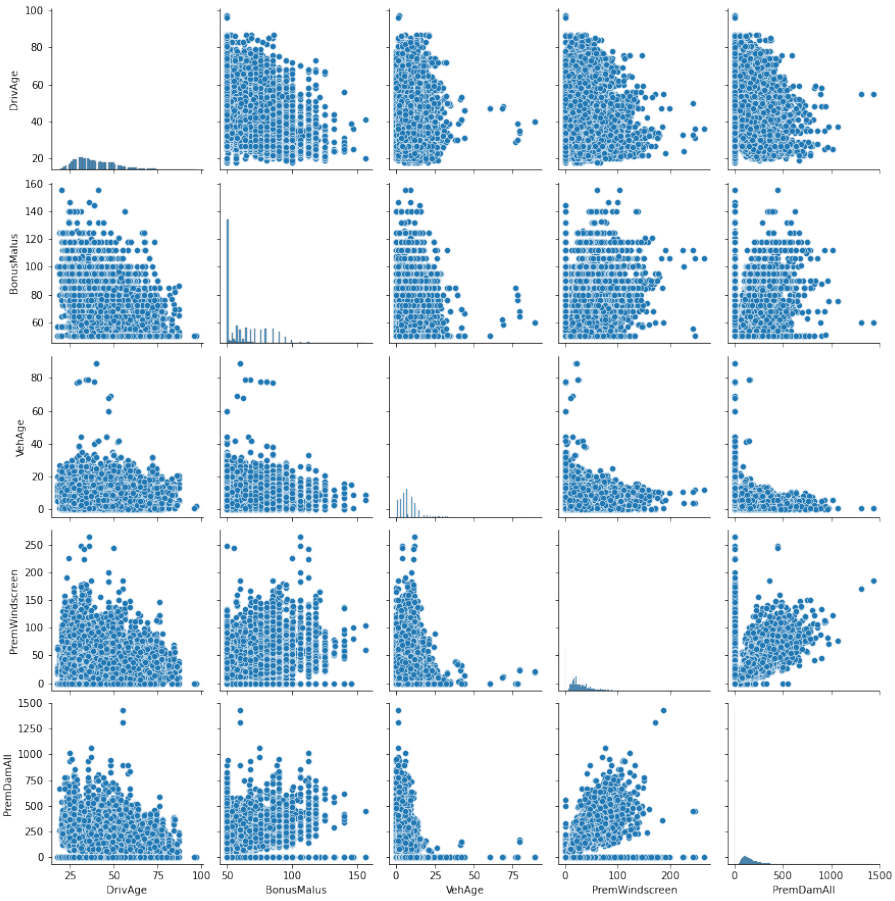
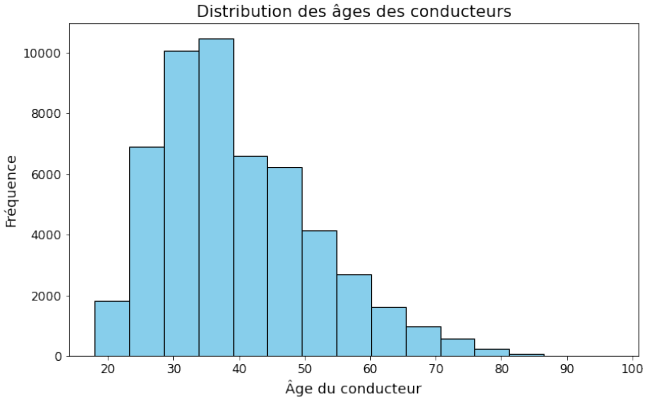


Figure 5 : Visualisations de la distributions :



Afin de réaliser une analyse de données multivariée pertinente, nous allons commencer avec un tableau croisé, mettant en relation une variable catégorielle, à savoir 'DrivGender', et une variable numérique, 'BonusMalus'.

Tableau croisé de variable entre une variable catégorielle et numérique

BonusMalus	50.0	51.0	52.0	53.0	54.0	55.0	56.0	57.0	58.0	\
DrivGender										
F	12132	870	432	224	1152	528	220	1742	482	
M	27582	1744	746	462	2298	862	548	2900	982	

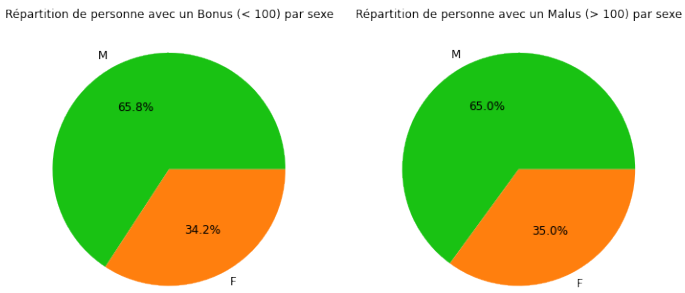
BonusMalus	59.0	...	120.0	121.0	125.0	126.0	132.0	133.0	140.0	\
DrivGender		...								
F	134	...	0	0	42	2	4	4	14	
M	388	...	4	4	54	0	12	0	18	

BonusMalus	145.0	147.0	156.0
DrivGender			
F	0	2	0
M	4	2	4

Pour bien comprendre ce tableau croisé, nous allons nous pencher sur les deux variables. Concernant la variable 'DrivGender', nous avons recensé le nombre d'individus de sexe féminin et masculin. De surcroît, nous avons dénombré combien d'hommes et de femmes se trouvent respectivement dans la catégorie Malus et dans la catégorie Bonus.
Dans le tableau de contingence, nous constatons que notre échantillon comprend 34 424 individus de sexe masculin et 17 938 individus de sexe féminin. On remarque une différence d'environ 48% entre le nombre d'hommes et de femmes, une différence loin d'être négligeable.

D'un point de vue graphique, il est immédiatement évident que les hommes ont davantage de bonus et de malus que les femmes, une observation qui s'explique par le fait que notre échantillon comporte une majorité d'hommes.

Figure 6 : Répartition de personne avec un Bonus $>$ ou $<$ à 100 par sexe

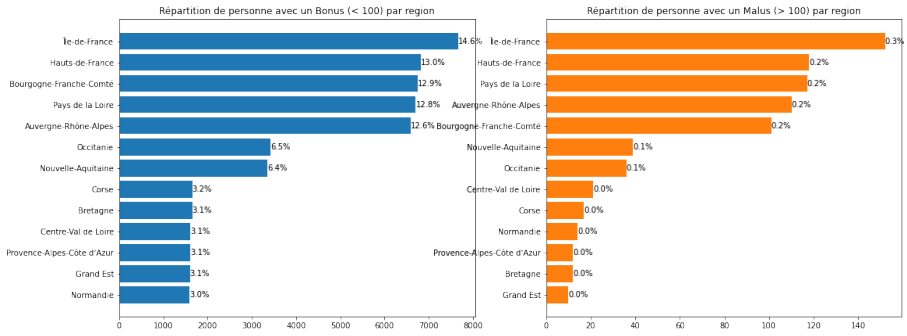


Cependant, nous constatons que la tendance à la répartition des bonus et des malus entre hommes et femmes est similaire. Pour approfondir notre analyse, nous avons comparé la répartition des malus par régions :

```
Île-de-France                7663
Hauts-de-France              6827
Bourgogne-Franche-Comté      6755
Pays de la Loire             6705
Auvergne-Rhône-Alpes         6583
Occitanie                    3426
Nouvelle-Aquitaine           3348
Corse                        1650
Bretagne                     1649
Centre-Val de Loire          1608
Provence-Alpes-Côte d'Azur   1607
Grand Est                    1599
Normandie                    1588
Name: Region, dtype: int64
Île-de-France                 152
Hauts-de-France               118
Pays de la Loire              117
Auvergne-Rhône-Alpes          110
Bourgogne-Franche-Comté       101
Nouvelle-Aquitaine             39
Occitanie                      36
Centre-Val de Loire            21
Corse                          17
Normandie                      14
Provence-Alpes-Côte d'Azur     12
Bretagne                       12
Grand Est                      10
```

Nous constatons que les trois régions où notre échantillon est le plus représenté sont : l'Île-de-France, les Hauts-de-France et la Bourgogne-Franche-Comté.

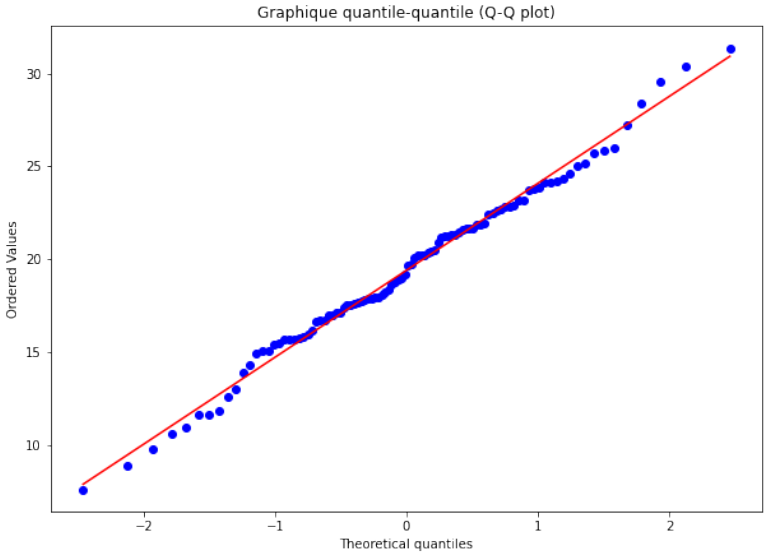
Figure 7 :Repartition de personne avec Malus plus de 100 par region.



En définitive, nous observons que la région où le nombre de personnes bénéficiant d'un Bonus est le plus élevé est l'Île-de-France, avec 14% des individus, suivie de la région des Hauts-de-France. En ce qui concerne les Malus, l'Île-de-France compte également le plus grand nombre d'individus dans notre échantillon avec un Malus, représentant environ 0.3% de l'échantillon.

Nous allons maintenant analyser la distribution à l'aide d'un graphique Q-Q pour déterminer si les données suivent une distribution normale. Cette étape est cruciale, car même si l'analyse univariée a confirmé que l'échantillon suit une loi normale, des tests statistiques pourraient révéler des écarts. Il est donc essentiel de vérifier cela avec un graphique Q-Q et une analyse de corrélation multiple.

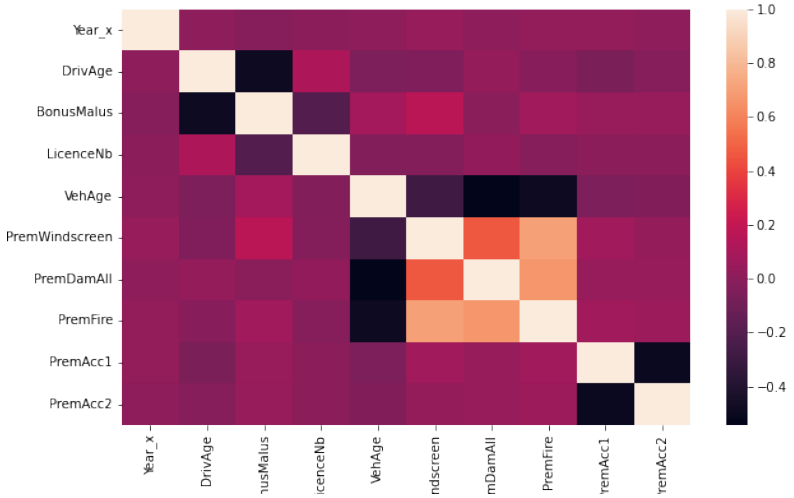
Figure 8 : qq plot.



Pour l'analyse de corrélation multiple, qui consiste à vérifier s'il existe une relation linéaire entre plusieurs variables, il est important d'identifier le type de base de données. Dans notre cas, notre base de données est considérée comme paramétrique. Pour ce faire, nous avons affiché une matrice de corrélation.

3.2.1 Matrice des corrélations :

Figure 9 : Corrélation entre les caractéristiques :



On observe que les variables **DrivAge et BonusMalus** sont négativement

corrélées avec un coefficient de **-0.49**, ce qui signifie qu'à mesure que l'âge du conducteur augmente, le BonusMalus diminue généralement.

De même, **VehAge** est négativement corrélé avec **PremWindscreen, PremDamAll, et PremFire,** avec des coefficients respectifs de **-0.29, -0.54, et -0.49.** Cela signifie que plus l'âge du véhicule est élevé, moins ces primes sont élevées.

PremWindscreen, PremDamAll, et PremFire sont positivement corrélés entre eux (avec un coefficient de corrélation variant entre **0.46 et 0.70**). Cela suggère que les véhicules avec une prime élevée pour le bris de glace ont tendance à avoir également une prime élevée pour les dommages tous accidents et l'incendie.

PremAcc1 et PremAcc2 sont négativement corrélés avec un coefficient de **-0.50,** ce qui signifie que lorsque la prime de l'accident 1 est élevée, la prime de l'accident 2 est généralement plus basse.

Comme décrit un pleu plus haut, on constate que les variables "DrivAge" et "BonusMalus" sont négativement corrélées à hauteur de -0.49%. L'augmentation de l'âge d'une personne est associée à une baisse de la variable "BonusMalus". Autrement dit, plus une personne vieillit, plus la probabilité d'obtenir un Bonus plutôt qu'un Malus augmente. **Cela pourrait s'expliquer par plusieurs hypothèses :**

1- Plus une personne est âgée, plus on estime qu'elle a de l'expérience sur la route et, de ce fait, est moins sujette aux accidents de la route.
2- Plus une personne est âgée, moins elle conduit, ce qui réduit le temps passé sur la route et, par conséquent, les occasions d'avoir un accident ou d'endommager sa voiture.

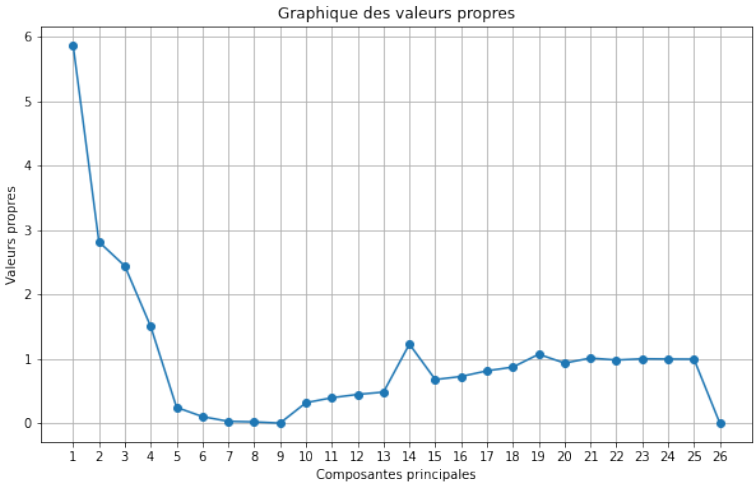
Prenons un autre exemple : nous observons une relation négative entre la variable "VehAge" et la variable "PremTheft", à hauteur de -0.48%. L'augmentation de l'âge du véhicule utilisé dans notre échantillon est associée à une baisse de la variable "PremTheft". Autrement dit, plus le véhicule vieillit, plus la probabilité de se le faire voler diminue. Cela pourrait s'expliquer par le fait qu'une personne possédant une voiture plus âgée a tendance à souscrire moins à une assurance antivol que celle possédant une voiture plus récente.

Prenons un troisième exemple : nous avons une relation positive entre la variable "BonusMalus" et la variable "PremTPLV" à hauteur de 0.66%. L'augmentation de la variable "BonusMalus" est associée à une hausse de la variable

"PremTPLV" à hauteur de 0.66%. Autrement dit, plus une personne a un malus important, plus la probabilité d'avoir une prime de garantie volontaire de responsabilité civile envers les tiers augmente. Cela suit une logique, lorsqu'une personne a tendance à être impliquée dans des accidents de la route, sa prime de souscription sera plus élevée.

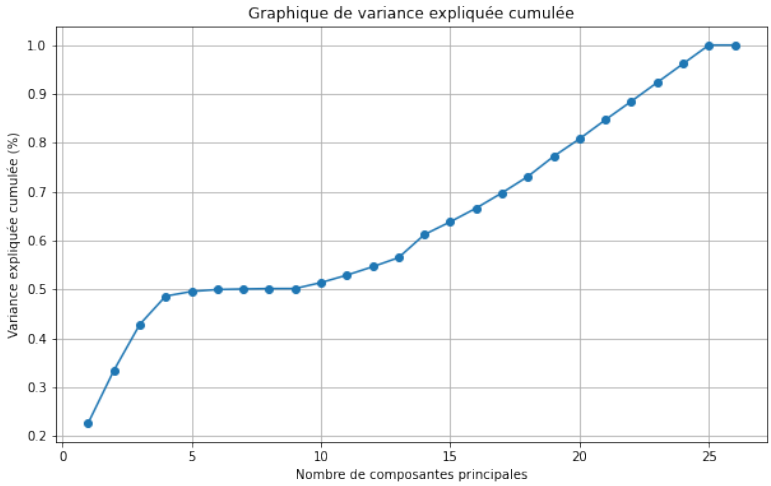
Nous avons formulé trois grandes hypothèses qui ont toutes été vérifiées. Enfin, nous allons expliquer en détail l'interprétation des valeurs propres et des variances expliquées.

Figure 10 : Graphique des valeurs propres :



Dans ce graphique, on s'aperçoit que les quatre premières composantes principales du modèle expliquent à elles seules presque plus de la moitié de la variance totale des données. De plus, nous constatons que les composantes 7, 8, 9 et 26 sont non significatives.

Figure 11 : Variance cumulée :



Enfin, nous allons examiner le graphique de la variance expliquée cumulée. On constate que la courbe commence à s'aplatir à partir de la cinquième composante principale jusqu'à la dixième. Cela signifie que les composantes de 5 à 10 n'expliquent qu'une proportion marginale de la variance. De plus, si l'on considère un seuil acceptable de 80% de la variance expliquée pour déterminer un nombre optimal de facteurs, on constate que 20 est le nombre de composantes principales acceptables. En effet, comme nous avons pu le voir dans le graphique des valeurs propres, à partir de la variable 20, on observe une stagnation, puis une baisse, indiquant ainsi la présence de variables peu significatives.

4 Chapitre III : Apprentissage Automatique (Machine Learning)

4.1 Comparaison des Modèles : Régression Linéaire vs Forêt Aléatoire

4.1.1 A : Calcul de la racine carrée de l'erreur quadratique moyenne (RMSE) :

La racine carrée de l'erreur quadratique moyenne (RMSE) est de : 2010.28

Cela signifie que, ce modèle de régression linéaire prédit les paiements avec une erreur de 2010.28. Cela pourrait être considéré comme bon ou mauvais en fonction du contexte.

Pour bien interpréter cette valeur, nous devons prendre en compte la plage de notre variable cible "Payment". Par exemple, si "Payment" varie généralement entre 0 et 10 000, une erreur de 2010.28 pourrait être acceptable. En revanche, si "Payment" varie généralement entre 0 et 2000, une erreur de 2010.28 serait très grande.

En outre, le RMSE seul ne donne pas une image complète de la performance du modèle. Il serait également utile de calculer d'autres mesures de performance, comme le R-carré (R^2), qui donne la proportion de la variance de la variable dépendante qui est prévisible à partir des variables indépendantes.

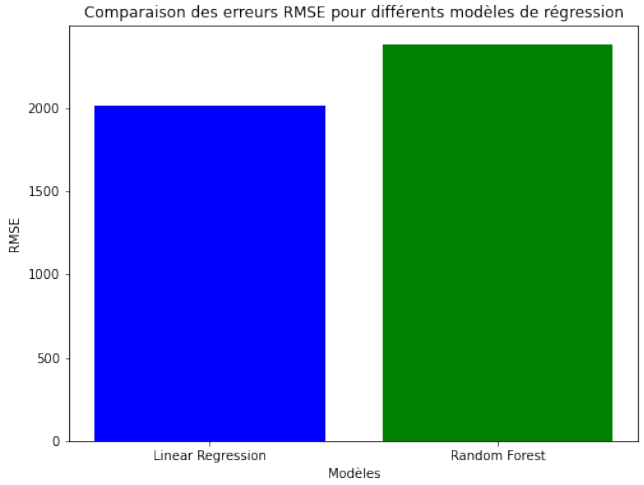
Enfin, il serait utile de comparer la performance de notre modèle de régression linéaire à celle d'autres modèles de régression pour voir si un autre modèle pourrait faire de meilleures prédictions.

4.1.2 B : Calcul de l'erreur du modèle de forêt aléatoire :

Le modèle de forêt aléatoire est de : 2378.85

Le modèle de régression linéaire a un RMSE de 2010.2807283220316, tandis que **le modèle de forêt aléatoire a un RMSE de 2378.852240326708**. Cela signifie que le modèle de régression linéaire a, en moyenne, des erreurs de prédiction plus faibles que le modèle de forêt aléatoire pour cette tâche de prédiction particulière. En d'autres termes, le modèle de régression linéaire semble mieux performer que le modèle de forêt aléatoire dans ce cas spécifique.

B : Figure 12 : Le modèle de la régression linéaire VS Le modèle de forêt aléatoire



Dans ce graphique, on constate que le modèle linéaire a une RMSE de 2010.28, tandis que la forêt aléatoire a une RMSE de 2378.85. Cela signifie que le modèle linéaire a, en moyenne, des erreurs de prédiction plus petites que la forêt aléatoire sur votre ensemble de données.

Pour mesurer la performance du modèle, en plus de la RMSE, on pout également utiliser d'autres mesures, telles que le coefficient de détermination R2, l'erreur absolue moyenne (MAE), l'erreur relative, et une validation croisée pour obtenir une estimation plus robuste de la performance du modèle.

4.1.3 Analyse de la performance du modèle :

- L'erreur quadratique moyenne du modèle MSE est de : 2165977.6070006425
- L'erreur absolue moyenne du modèle MAE est de : 380.1243339376991
- Le coefficient de détermination R2 est de : 0.19746541901422965

On constate que **Le MSE est assez élevée avec 2165977.6.** Ce qui indique que le modèle fait des erreurs de prédiction assez importantes. Cela pourrait suggérer que le modèle n'est pas très précis.

Le MAE est de 380.1 cela signifie qu'en moyenne, le modèle se trompe d'environ 380.1 unités lors de la prédiction du montant du paiement. Selon le

contexte, cela peut être considéré comme acceptable ou non.

Le R2 est de 0.197 . Cela signifie que le modèle ne peut expliquer que 19.7% de la variance de la variable de paiement. C'est relativement faible, ce qui suggère que le modèle n'est pas très performant.

Pour améliorer la performance de du modèle, on pout envisager à optimiser les hyperparamètres du modèle, par exemple en utilisant la validation croisée ou une recherche sur grille. Ou recueillir plus de données, si possible. Un ensemble de données plus grand pourrait aider le modèle à mieux apprendre.

4.1.4 validation croisée :

Le Score de validation croisée est de : [0.10165043 0.02835731 0.0276716 0.0733152 0.19718284]

Le Score moyen de validation croisée est de : 0.08563547509903184

Les scores de validation croisée obtenus représentent le coefficient de détermination R2 pour chaque fold de la validation croisée. Le R2 est une mesure de la quantité de variance dans la variable cible qui est prédite à partir des variables indépendantes. Dans le contexte de la régression, un score R2 plus élevé est généralement meilleur.

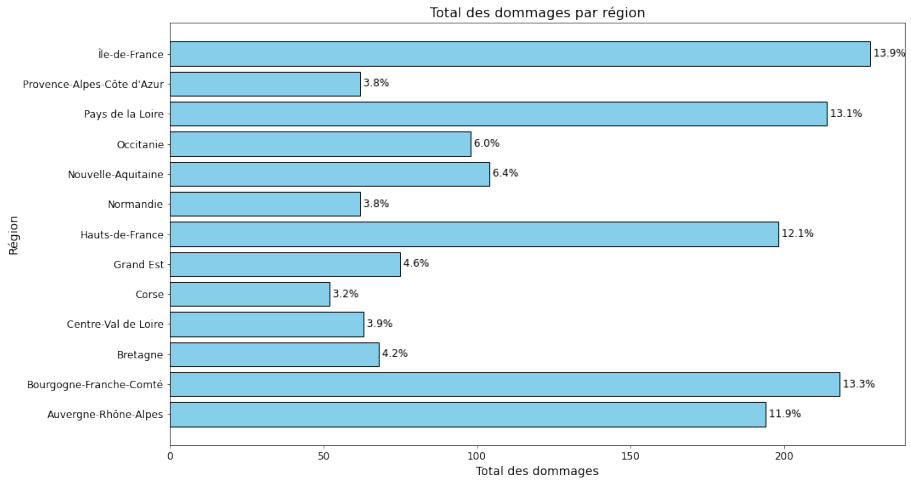
Ici, les scores de validation croisée varient beaucoup, **avec un minimum de 0.028 et un maximum de 0.197.** Cela pourrait indiquer que le modèle est instable ou qu'il ne généralise pas bien à de nouvelles données. Lorsque les scores de validation croisée varient beaucoup, cela signifie souvent que le modèle est peut-être en train de surajuster à une partie spécifique des données.

En outre, **le score moyen de validation croisée est d'environ 0.086**. Ce n'est pas un score particulièrement élevé, ce qui indique que le modèle ne prédit qu'une petite partie de la variance de la variable cible.

4.2 Analyse Exploratoire Approfondie

Voyons une analyse exploratoire des données pour de découvrir des tendances intéressantes, comme les régions où les dommages sont les plus importantspar exemple.

Figure 13 : Total des dommages par région



Ce résultat obtenu indique la somme totale des dommages pour chaque région. Il semble que la région Île-de-France ait subi le plus de dommages, avec un total de 13.90%, suivi de près par la région Bourgogne-Franche-Comté avec un total de 13.3%. À l'autre bout du spectre, la Corse a le plus faible total de dommages avec 3.2%

Cependant, il est important de noter que ces chiffres ne tiennent pas compte du nombre de polices d'assurance dans chaque région. Si une région a beaucoup plus de polices d'assurance que les autres, elle pourrait naturellement avoir un total de dommages plus élevé. Pour obtenir une image plus précise de la situation, il serait utile de normaliser ces chiffres par le nombre de polices dans chaque région pour obtenir le dommage moyen par police.

De plus, ces chiffres sont simplement des agrégations et ne fournissent pas d'informations sur la cause des dommages ou sur d'autres facteurs qui pourraient influencer le montant des dommages dans chaque région. Des analyses supplémentaires pourraient être nécessaires pour identifier ces facteurs.

Enfin, il est également important de rappeler que ces chiffres sont suscep-

tibles de varier d'une année à l'autre. Par conséquent, il pourrait être utile d'examiner les tendances au fil du temps pour voir si certaines régions ont constamment des totaux de dommages plus élevés ou si les chiffres fluctuent d'une année à l'autre.

Calcul des dommages moyens par police par région :

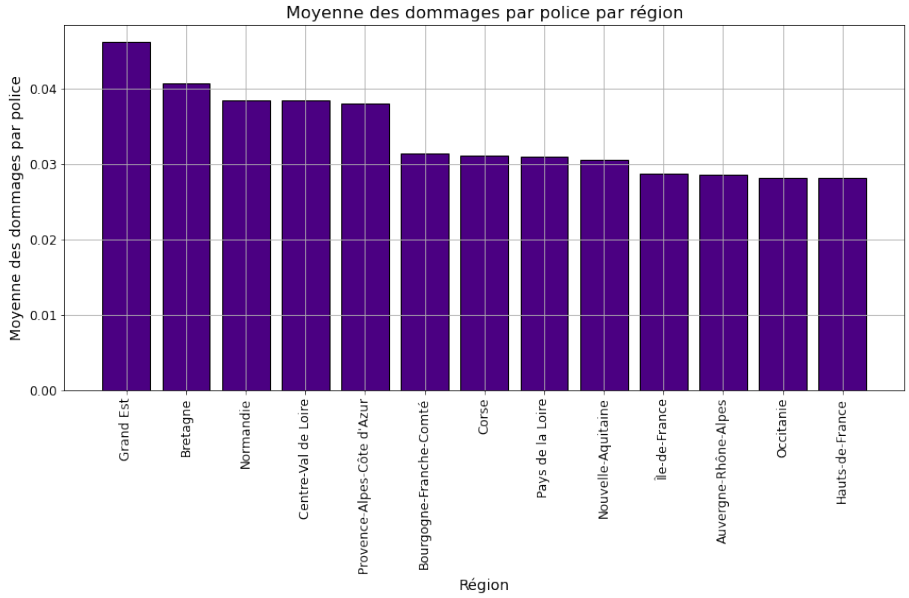
```
Grand Est                   0.046182
Bretagne                    0.040694
Normandie                   0.038438
Centre-Val de Loire         0.038391
Provence-Alpes-Côte d'Azur  0.037990
Bourgogne-Franche-Comté     0.031371
Corse                       0.031082
Pays de la Loire            0.030952
Nouvelle-Aquitaine          0.030570
Île-de-France               0.028759
Auvergne-Rhône-Alpes        0.028563
Occitanie                   0.028145
Hauts-de-France             0.028121
dtype: float64
```

On peut remarquer que la région du "Grand Est" a la plus haute moyenne de dommages par police, suivie de près par la "Bretagne" et la "Normandie". Cela signifie que, en moyenne, les polices d'assurance dans ces régions ont tendance à avoir des dommages plus élevés que dans les autres régions.

Il est important de noter que ces moyennes ne représentent pas nécessairement le risque global dans chaque région. Par exemple, une région avec une moyenne élevée de dommages par police pourrait aussi avoir un grand nombre de polices avec des dommages très faibles ou nuls. De même, une région avec une moyenne de dommages faible par police pourrait avoir un petit nombre de polices avec des dommages très élevés.

Ces résultats peuvent aider à identifier les régions où les dommages sont généralement plus élevés, ce qui peut être utile pour l'établissement des primes d'assurance, la gestion des risques, ou l'élaboration de stratégies pour réduire les dommages. Cependant, ils ne devraient pas être utilisés isolément pour prendre des décisions, et il serait utile de les combiner avec d'autres analyses et informations pour obtenir une image plus complète.

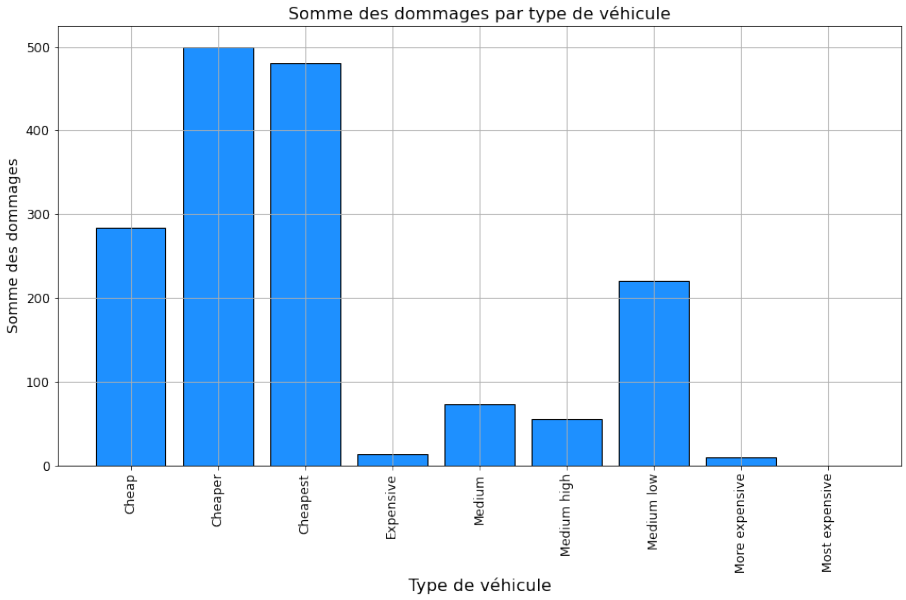
Figure 14 : Moyenne des dommages par police par région.png



Voyons en comment les dommages varient en fonction du type de véhicule. Pour ce faire, on peut regrouper les données par VehClass et calculer la somme des dommages pour chaque type de véhicule :

```
VehClass
Cheap             284.0
Cheaper           500.0
Cheapest          480.0
Expensive          14.0
Medium             73.0
Medium high        55.0
Medium low        220.0
More expensive     10.0
Most expensive      0.0
```

Figure 15 : Somme des dommages par type de véhicule.png

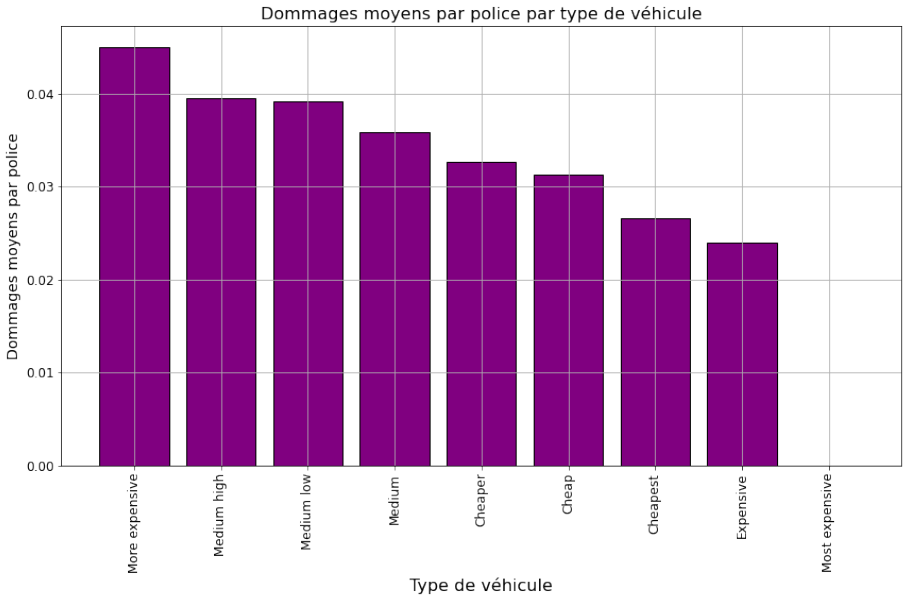


Il semble que les véhicules classés comme "Cheaper" et "Cheapest" ont subi le plus de dommages, avec respectivement 500 et 480 unités de dommages. Au contraire, les véhicules classés comme "Most expensive" n'ont subi aucun dommage, et ceux classés comme "More expensive" et "Expensive" ont subi très peu de dommages.

On pourrait interpréter cela comme indiquant que les véhicules moins chers sont plus susceptibles de subir des dommages. Cependant, il serait nécessaire de prendre en compte d'autres facteurs avant de tirer des conclusions définitives. Par exemple, il se peut qu'il y ait simplement plus de véhicules moins chers dans l'ensemble de données.

Et on peut faire de même pour le nombre de polices par type de véhicule, puis calculer les dommages moyens par police pour chaque type de véhicule :

Figure 16 : Dommages moyens par police par type de véhicule

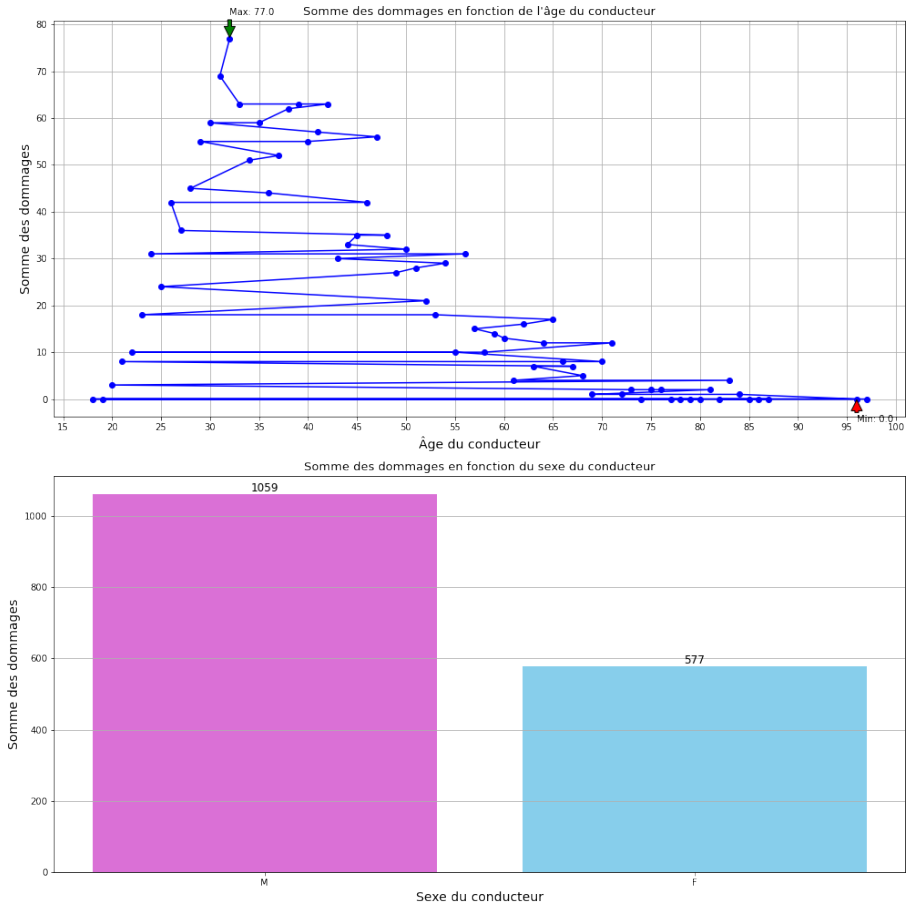


Ce graphique représent la moyenne des dommages par police d'assurance pour chaque type de véhicule. Il est intéressant de noter que bien que les véhicules "Cheaper" et "Cheapest" aient eu la plus grande somme de dommages, quand on prend en compte le nombre de polices d'assurance pour chaque type de véhicule, on observe que les véhicules "More expensive", "Medium high" et "Medium low" ont en moyenne plus de dommages par police.

Cela pourrait indiquer que même si les véhicules plus chers sont moins nombreux, ils subissent en moyenne plus de dommages par police d'assurance. Encore une fois, il faudrait prendre en compte d'autres facteurs pour tirer des conclusions définitives.

Enfin, analysons l'impact de l'âge et du sexe du conducteur sur le total des dommages. Pour cela, nous pouvons calculer la somme des dommages par âge et par sexe du conducteur.

Figure 17 : Somme des dommages en fonction de l'âge et du sexe du conducteur.



En ce qui concerne l'âge du conducteur, il semble que les conducteurs de 32 ans ont subi les dommages les plus élevés, suivis de près par ceux de 31 ans et de 33, 39 et 42 ans. Les conducteurs les plus âgés, de 74 à 78 ans et de 97 ans, n'ont pas subi de dommages. Cela pourrait suggérer que les conducteurs plus jeunes sont plus susceptibles de subir des dommages, mais il serait utile d'avoir plus d'informations pour tirer des conclusions définitives.

En ce qui concerne le sexe du conducteur, les hommes ont subi presque le double des dommages par rapport aux femmes. Cela pourrait suggérer que les hommes sont plus susceptibles de subir des dommages, mais encore une fois, des informations supplémentaires seraient utiles pour tirer des conclusions.

5 Conclusion

Après une analyse détaillée des données, voici quelques conclusions générales que nous pouvons tirer :

Région : Les données montrent que certaines régions ont plus de sinistres que d'autres. Par exemple, l'Île-de-France et la Bourgogne-Franche-Comté sont en tête en termes de dommages totaux. Cela pourrait être dû à divers facteurs comme la densité de la population, la météo, la qualité des routes, etc.

Type de véhicule : Il existe une relation apparente entre le type de véhicule assuré et le montant des dommages. Les véhicules de catégorie "Cheaper" et "Cheap" ont le plus haut total de dommages, ce qui pourrait suggérer que ces véhicules sont plus nombreux ou bien qu'ils sont plus susceptibles d'être impliqués dans des sinistres.

Âge du conducteur : L'âge du conducteur semble également avoir un impact sur le montant total des dommages. Les conducteurs autour de la trentaine semblent causer le plus de dommages, ce qui pourrait être lié à la démographie des conducteurs, leur comportement au volant ou leur expérience.

Sexe du conducteur : Les conducteurs masculins semblent causer plus de dommages que les conductrices. Cela pourrait être dû à des facteurs tels que le nombre de conducteurs hommes vs femmes, les habitudes de conduite, etc.

En conclusion, l'analyse de ces données peut aider la compagnie d'assurance à mieux comprendre les facteurs qui influencent les sinistres. Cela peut permettre de mieux évaluer les risques associés à chaque police d'assurance et de fixer les primes d'assurance de manière plus précise. Cependant, il est important de noter que ces résultats sont basés sur l'analyse des données disponibles et pourraient nécessiter des analyses plus approfondies pour confirmer ces tendances.

De plus, il serait bénéfique de réaliser une analyse plus approfondie en utilisant des méthodes statistiques plus avancées ou le machine learning pour mieux comprendre les relations entre les variables et leur impact sur les montants des sinistres.

6 Références :

Package‘CASdatasets’
http ://cas.uqam.ca/
http ://dutangc.perso.math.cnrs.fr/RRepository